US009270767B2

(12) **United States Patent**
Langlois et al.

(10) **Patent No.: US 9,270,767 B2**
(45) **Date of Patent: Feb. 23, 2016**

(54) **METHOD AND SYSTEM FOR DISCOVERY OF USER UNKNOWN INTERESTS BASED ON SUPPLEMENTAL CONTENT**

(71) Applicant: **YAHOO! INC.**, Sunnyvale, CA (US)

(72) Inventors: **Jean-Marc Langlois**, Menlo Park, CA (US); **Scott Gaffney**, Palo Alto, CA (US); **Choon Hui Teo**, Sunnyvale, CA (US); **Nathan Liu**, Sunnyvale, CA (US)

(73) Assignee: **YAHOO! INC.**, Sunnyvale, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 462 days.

(21) Appl. No.: **13/835,745**

(22) Filed: **Mar. 15, 2013**

(65) **Prior Publication Data**

US 2014/0280548 A1 Sep. 18, 2014

(51) **Int. Cl.**
G06F 15/16 (2006.01)
H04L 29/08 (2006.01)

(52) **U.S. Cl.**
CPC ***H04L 67/22*** (2013.01); ***H04L 67/306*** (2013.01)

(58) **Field of Classification Search**
CPC H04L 67/22; H04L 67/306
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

2005/0022239	A1	1/2005	Meuleman	
2006/0294084	A1	12/2006	Patel et al.	
2011/0196865	A1	8/2011	Eggink et al.	
2012/0117167	A1 *	5/2012	Sadja	G06F 17/3089 709/206
2014/0067702	A1 *	3/2014	Rathod	G06Q 10/10 705/319
2014/0095304	A1 *	4/2014	Ganesh	G06Q 30/02 705/14.49
2014/0156745	A1 *	6/2014	Hua	H04L 67/1095 709/204
2014/0278308	A1 *	9/2014	Liu	H04L 67/22 703/6
2014/0280550	A1 *	9/2014	Glass	H04L 67/22 709/204
2014/0280890	A1 *	9/2014	Yi	H04L 67/22 709/224
2015/0112918	A1 *	4/2015	Zheng	G06Q 30/02 706/48

FOREIGN PATENT DOCUMENTS

KR	1020080026952	A	3/2008
KR	10-1028810	B1	4/2011

OTHER PUBLICATIONS

International Search Report and Written Opinion issued on Jun. 17, 2014 in International Application PCT/US2014/021149.

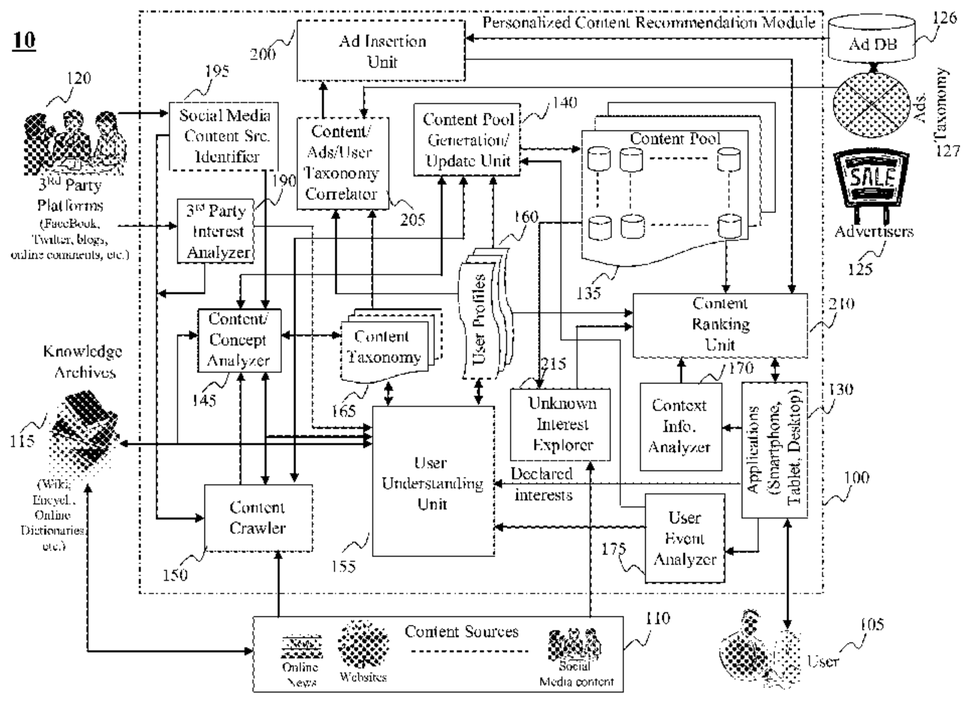
* cited by examiner

Primary Examiner — Mohamed Wasel
(74) *Attorney, Agent, or Firm* — Pillsbury Winthrop Shaw Pittman LLP

(57) **ABSTRACT**

The present teaching relates to discovery of user unknown interests. In one example, information related to a user is retrieved from a user profile. The information indicates one or more known interests of the user. At least one known interest of the user is identified based on the information. One or more supplemental interests with respect to each identified at least one known interest of the user are identified. The one or more supplemental interests do not overlap with the one or more known interests of the user. Supplemental content associated with the one or more supplemental interests are identified. Each piece of content in the supplemental content is ranked. At least one piece of content in the supplemental content is selected based on the ranking. The selected at least one piece of supplemental content is used to discover unknown interest of the user.

20 Claims, 24 Drawing Sheets



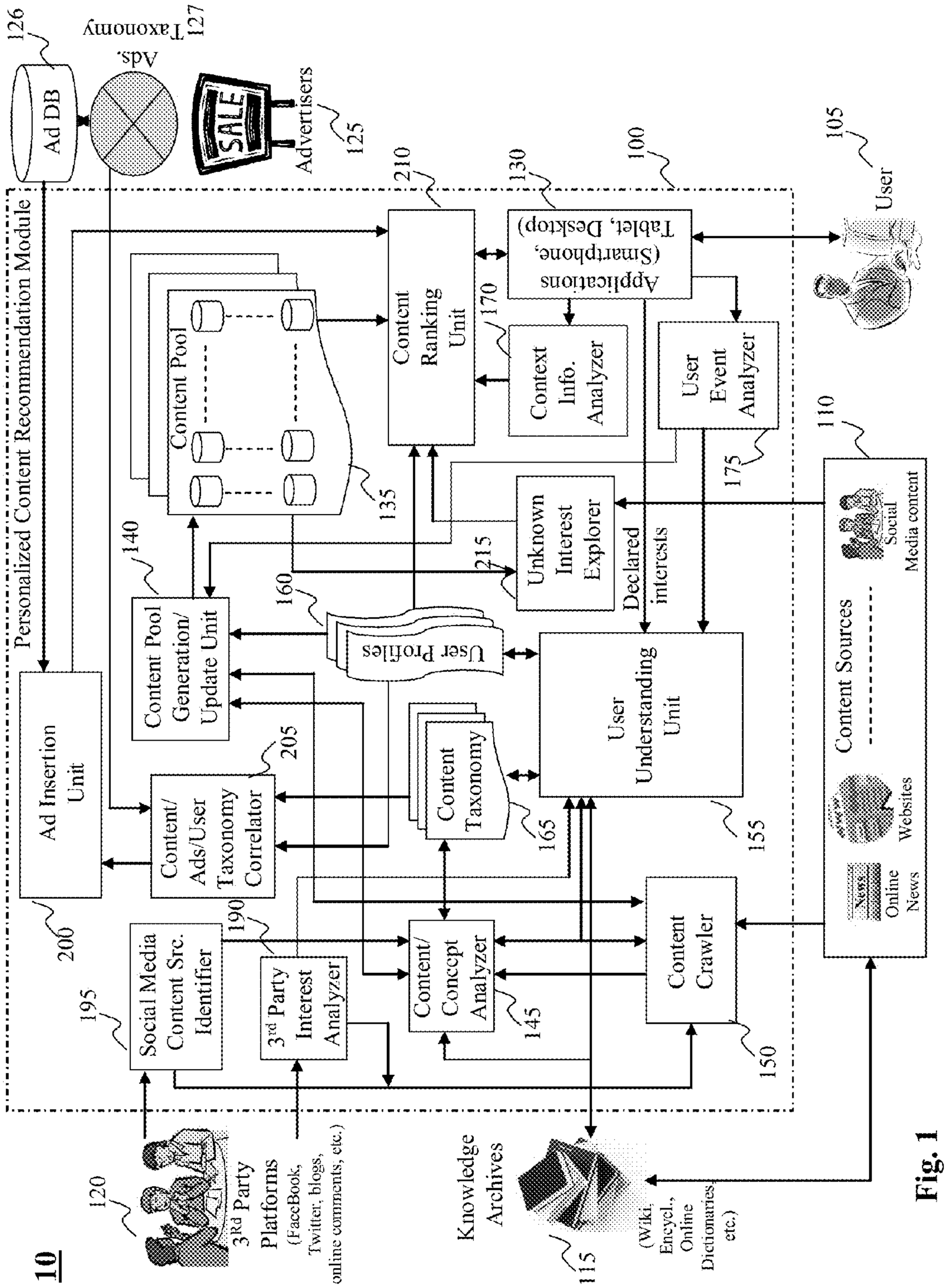
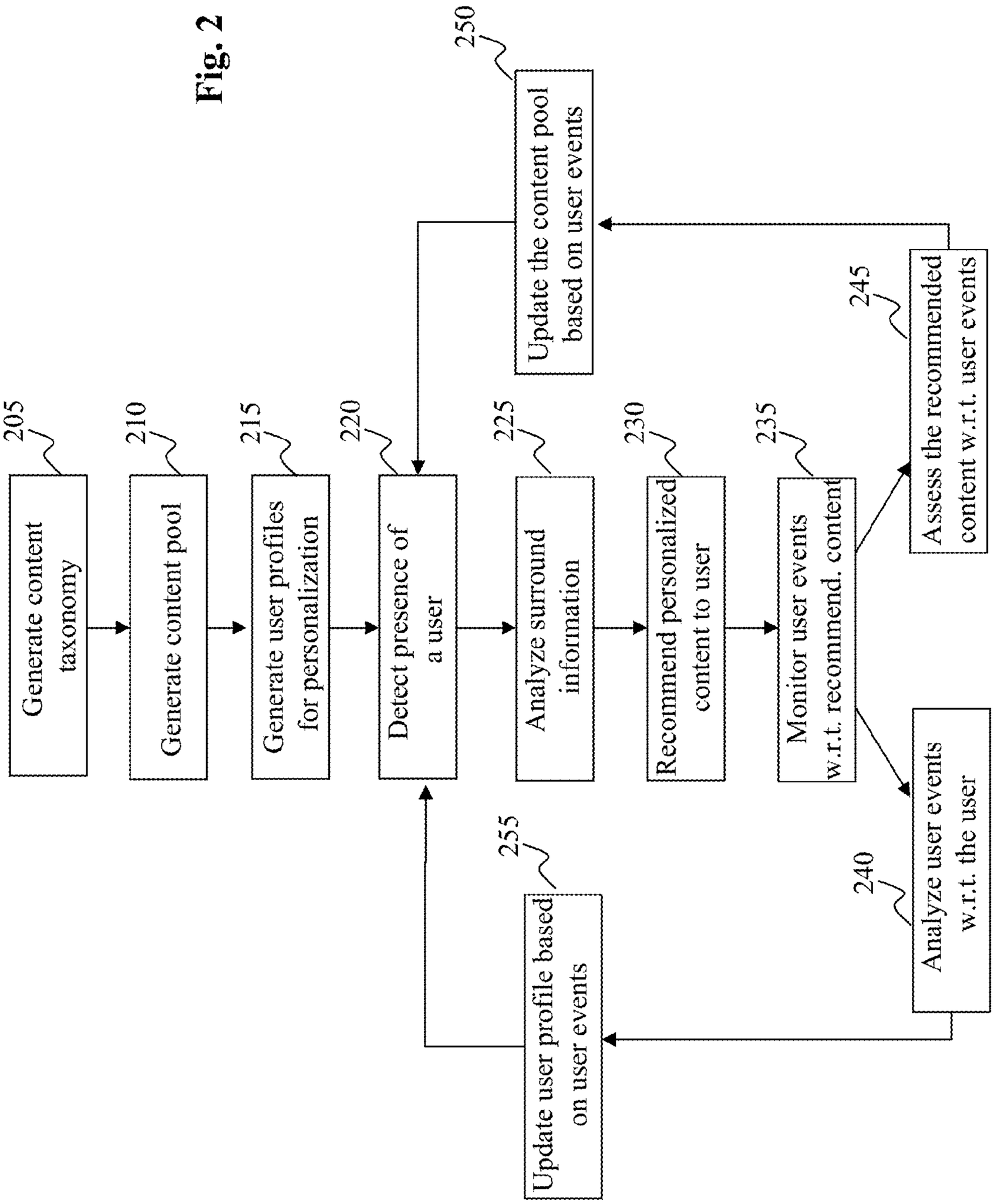

10
Personalized Content Recommendation Module
Ad DB
126
Ads, Taxonomy
127
SALE
Advertisers
125
105
User
100
130
Applications (Smartphone, Tablet, Desktop)
210
Content Ranking Unit
170
Context Info. Analyzer
User Event Analyzer
175
110
Content Pool
135
140
Content Pool Generation/ Update Unit
160
User Profiles
215
Unknown Interest Explorer
Declared interests
Content Sources
Social Media content
Online News
Websites
Ad Insertion Unit
200
Content/ Ads/User Taxonomy Correlator
205
Content Taxonomy
165
User Understanding Unit
155
190
Social Media Content Src. Identifier
195
3rd Party Interest Analyzer
Content/ Concept Analyzer
145
Content Crawler
150
120
3Rd Party Platforms (FaceBook, Twitter, blogs, online comments, etc.)
Knowledge Archives
115
(Wiki, Encycl., Online Dictionaries etc.)

Fig. 1

Fig. 2


205
Generate content taxonomy
210
Generate content pool
215
Generate user profiles for personalization
220
Detect presence of a user
225
Analyze surround information
230
Recommend personalized content to user
235
Monitor user events w.r.t. recommend. content
240
Analyze user events w.r.t. the user
245
Assess the recommended content w.r.t. user events
250
Update the content pool based on user events
255
Update user profile based on user events

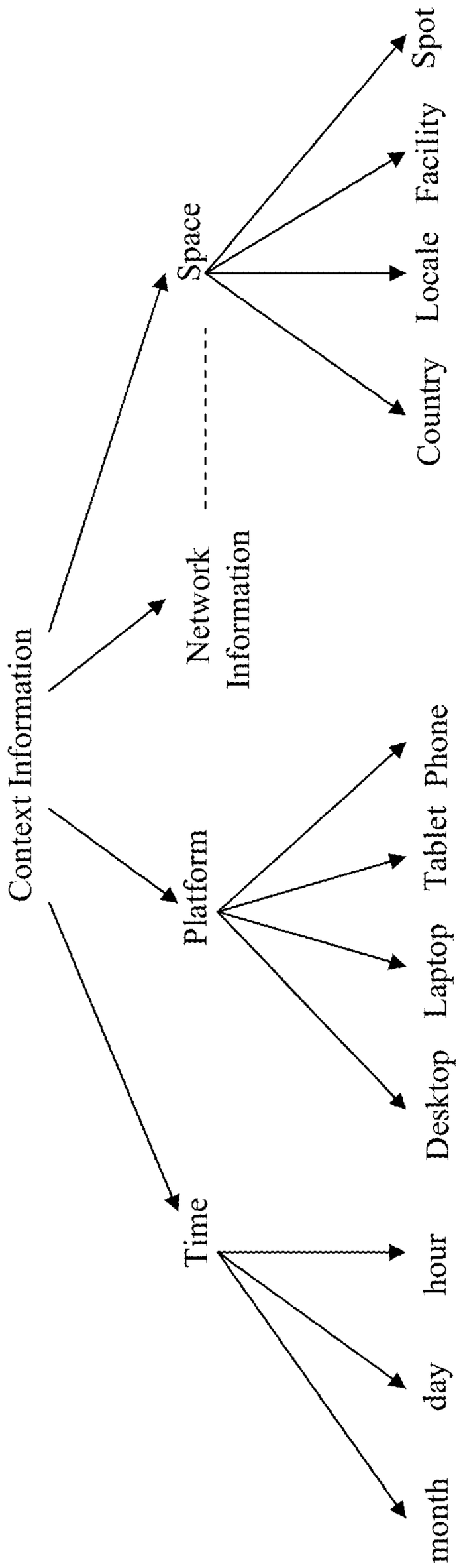

Context Information
Space
Spot
Facility
Locale
Country
Network Information
Platform
Desktop
Laptop
Tablet
Phone
Time
hour
day
month

Fig. 3

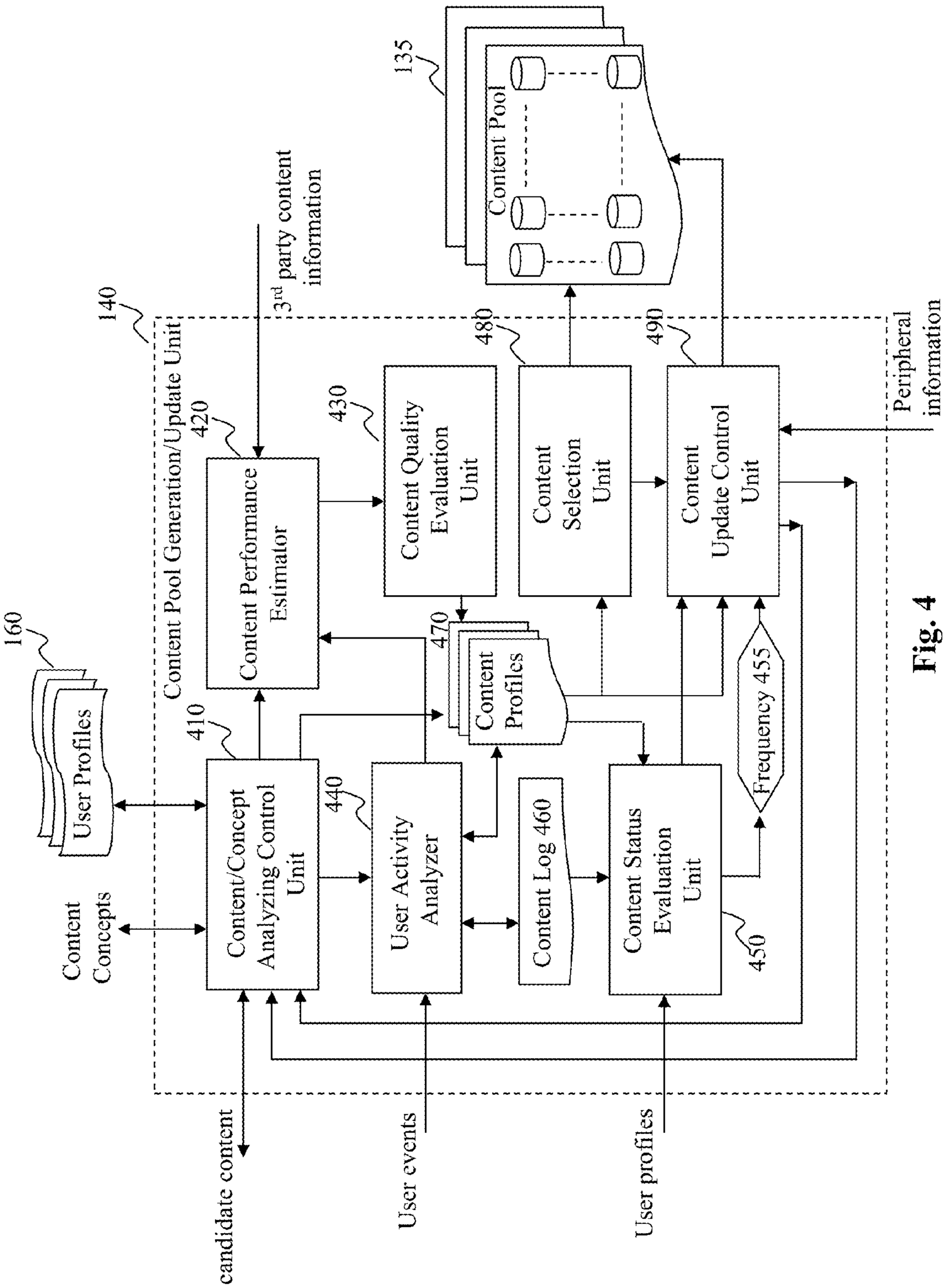

135
Content Pool
3rd party content information
140
Content Pool Generation/Update Unit
420
Content Performance Estimator
430
Content Quality Evaluation Unit
480
Content Selection Unit
490
Content Update Control Unit
Peripheral information
160
User Profiles
410
Content/Concept Analyzing Control Unit
440
User Activity Analyzer
470
Content Profiles
Content Log 460
Content Status Evaluation Unit
Frequency 455
450
Content Concepts
candidate content
User events
User profiles

Fig. 4

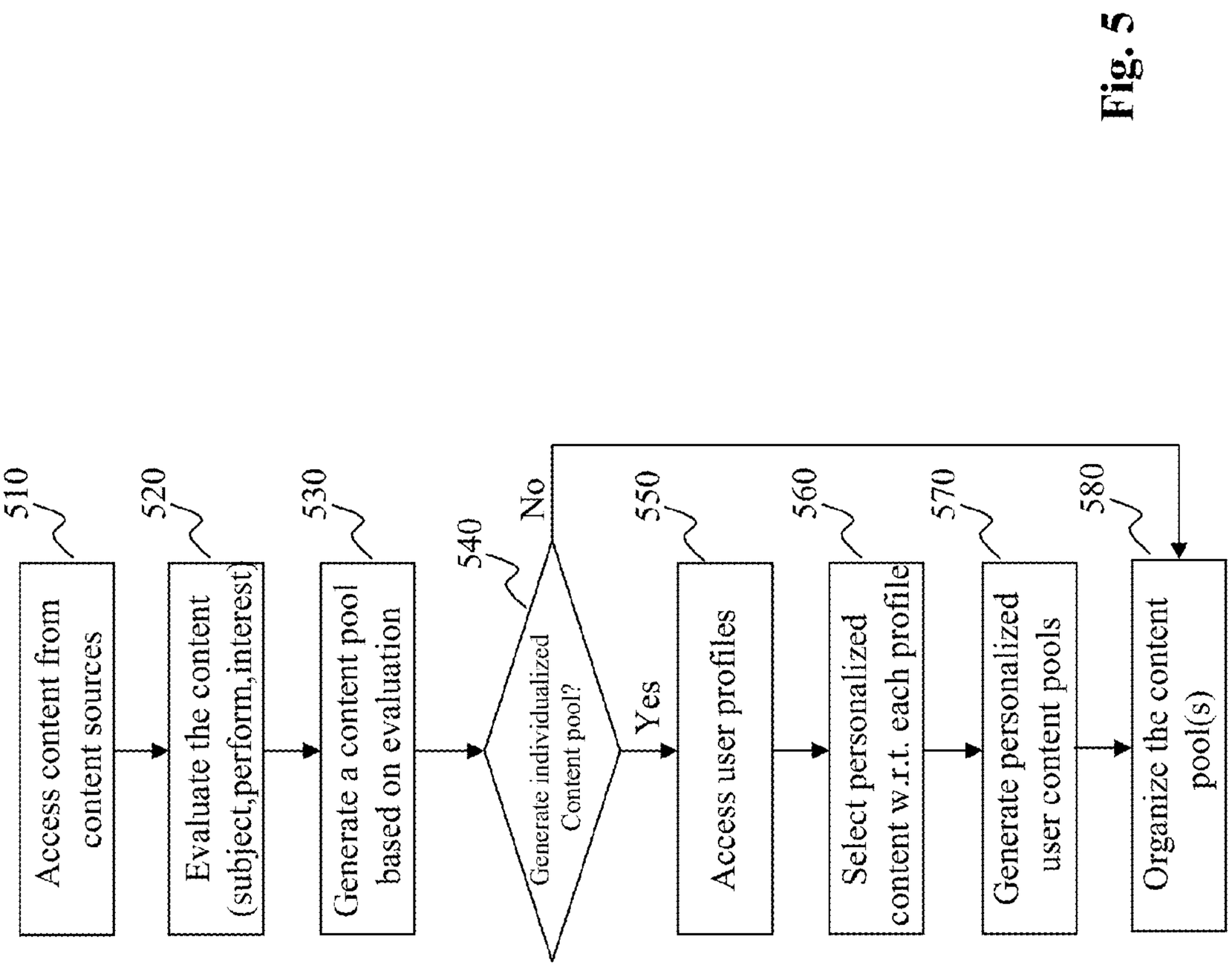

510
Access content from content sources
520
Evaluate the content (subject,perform,interest)
530
Generate a content pool based on evaluation
540
Generate individualized Content pool?
No
Yes
550
Access user profiles
560
Select personalized content w.r.t. each profile
570
Generate personalized user content pools
580
Organize the content pool(s)

Fig. 5

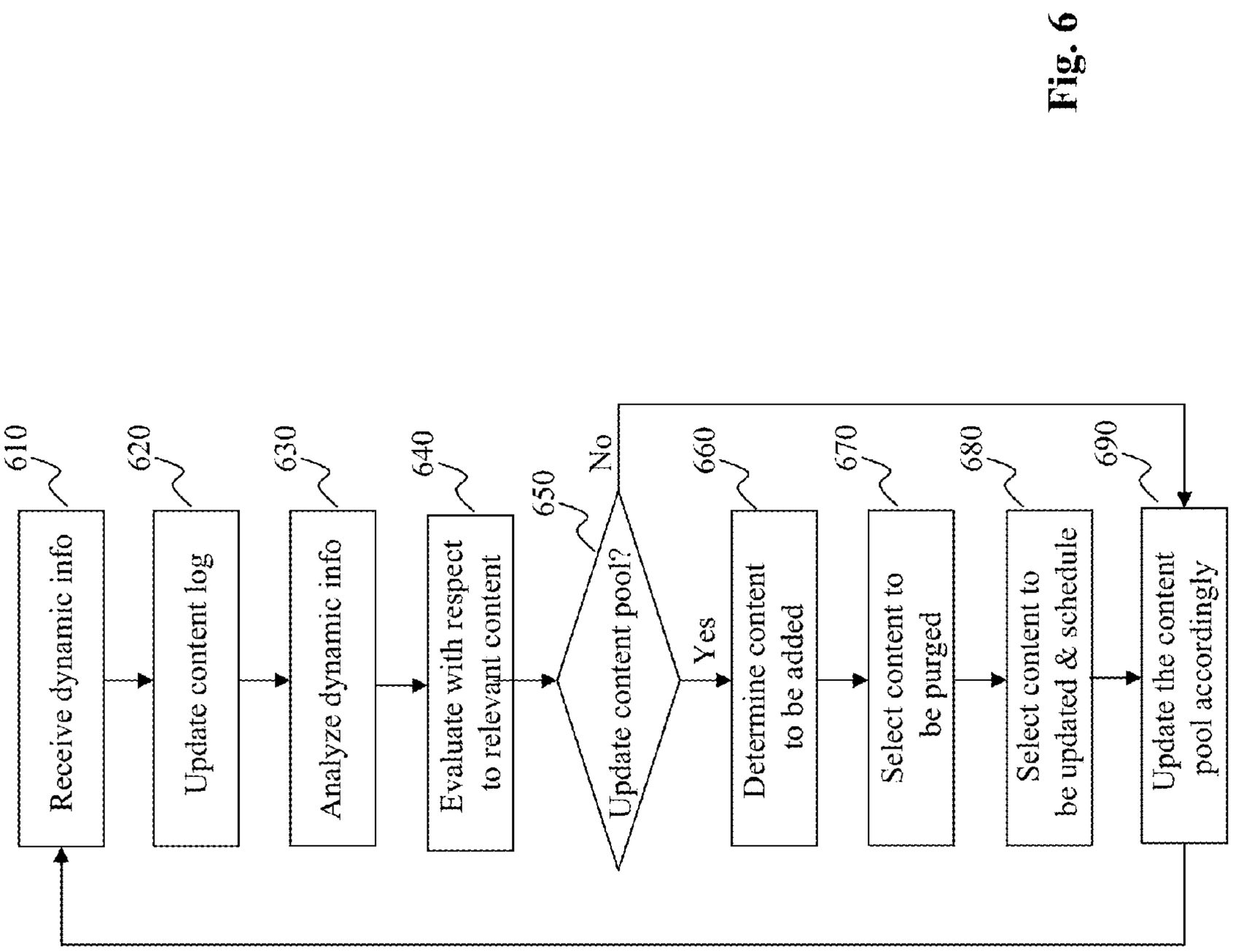

610
Receive dynamic info
620
Update content log
630
Analyze dynamic info
640
Evaluate with respect to relevant content
650
Update content pool?
No
Yes
660
Determine content to be added
670
Select content to be purged
680
Select content to be updated & schedule
690
Update the content pool accordingly

Fig. 6

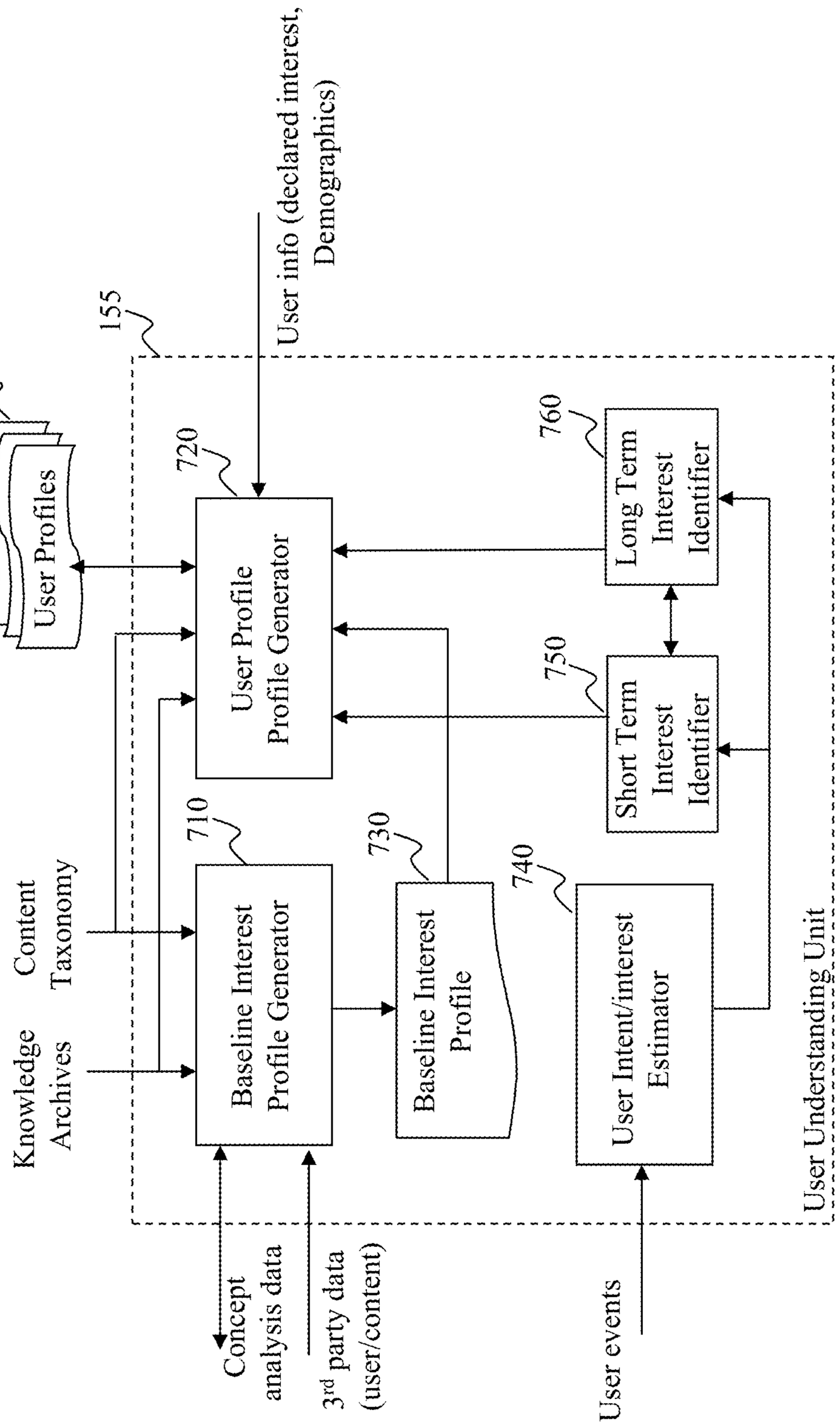

User info (declared interest, Demographics)
155
160
User Profiles
720
User Profile
Profile Generator
760
Long Term
Interest
Identifier
750
Short Term
Interest
Identifier
710
Baseline Interest
Profile Generator
730
Baseline Interest
Profile
740
User Intent/interest
Estimator
Content
Taxonomy
Knowledge
Archives
User Understanding Unit
Concept
analysis data
3rd party data
(user/content)
User events

Fig. 7

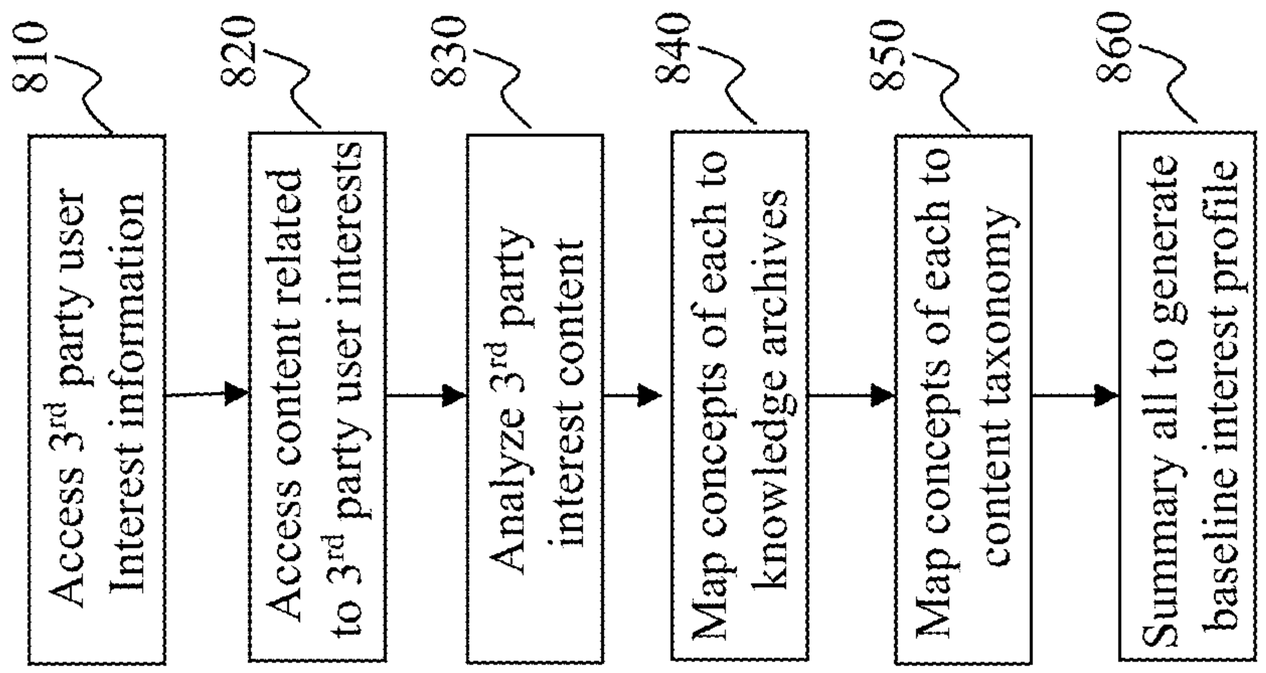

810
Access 3rd party user Interest information
820
Access content related to 3rd party user interests
830
Analyze 3rd party interest content
840
Map concepts of each to knowledge archives
850
Map concepts of each to content taxonomy
860
Summary all to generate baseline interest profile

Fig. 8

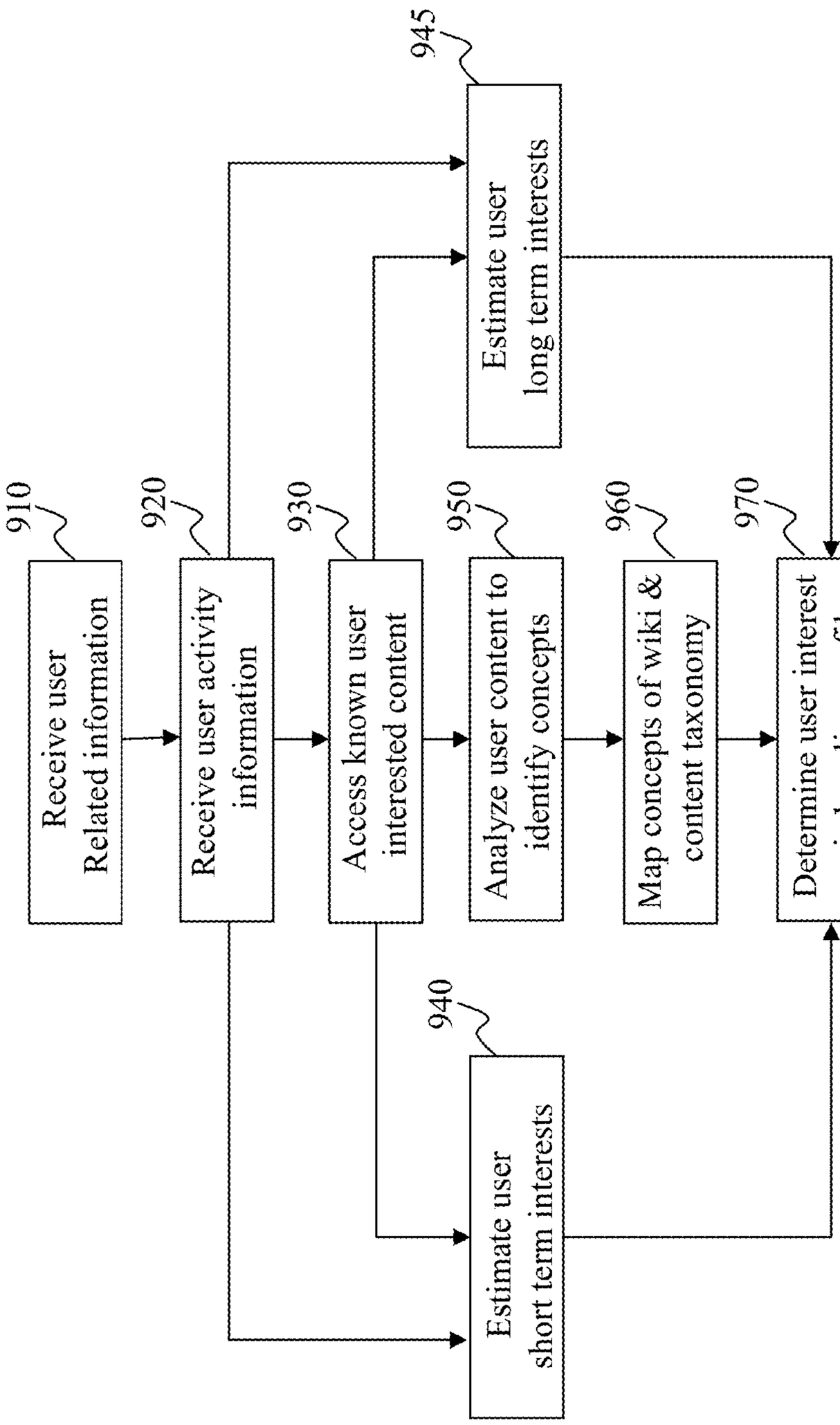

910
Receive user Related information
920
Receive user activity information
930
Access known user interested content
940
Estimate user short term interests
945
Estimate user long term interests
950
Analyze user content to identify concepts
960
Map concepts of wiki & content taxonomy
970
Determine user interest via baseline profile
980
Generate a user profile

Fig. 9

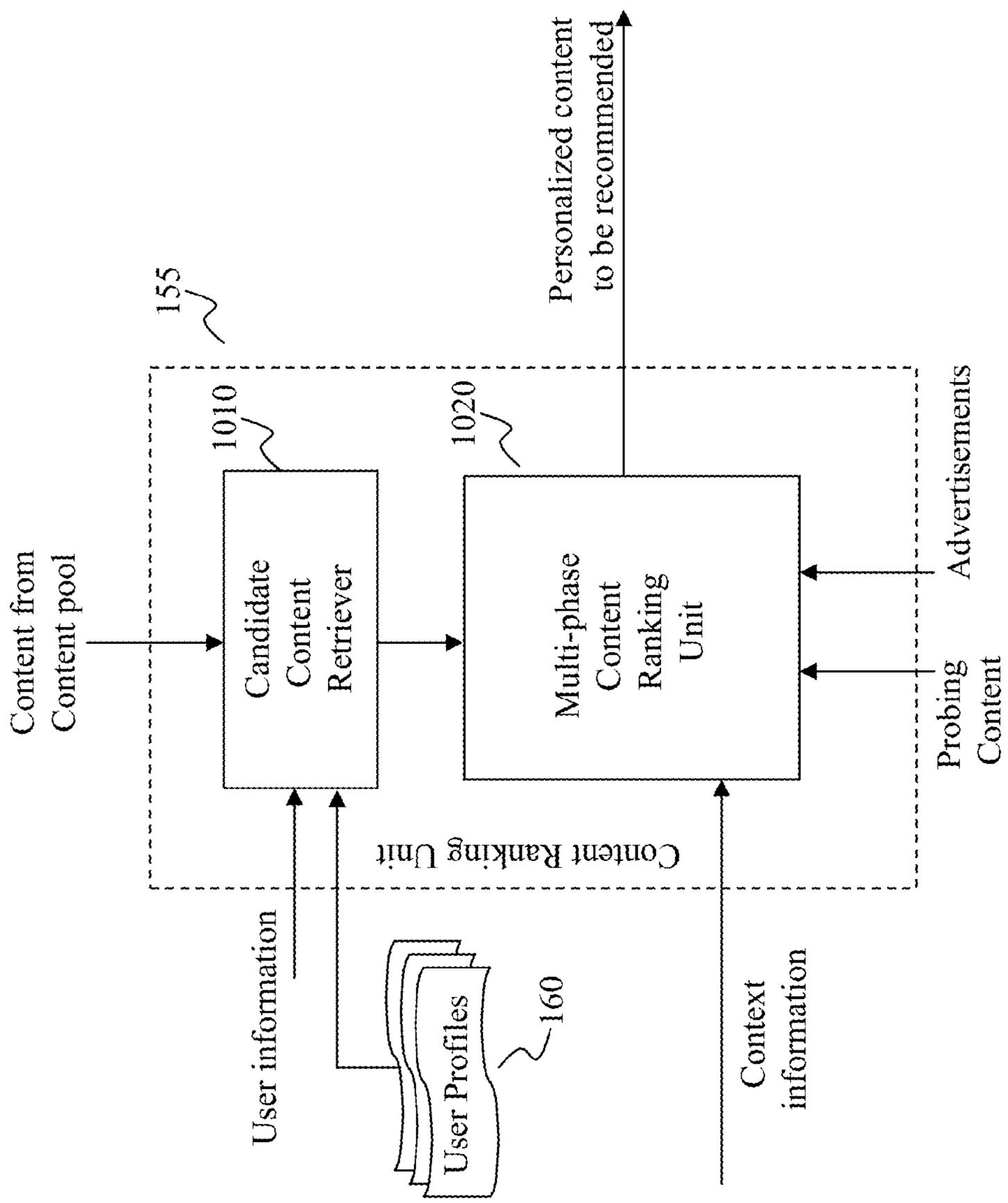

Content from Content pool
155
1010
Candidate Content Retriever
1020
Multi-phase Content Ranking Unit
Personalized content to be recommended
Content Ranking Unit
Advertisements
Probing Content
User information
User Profiles
160
Context information

Fig. 10

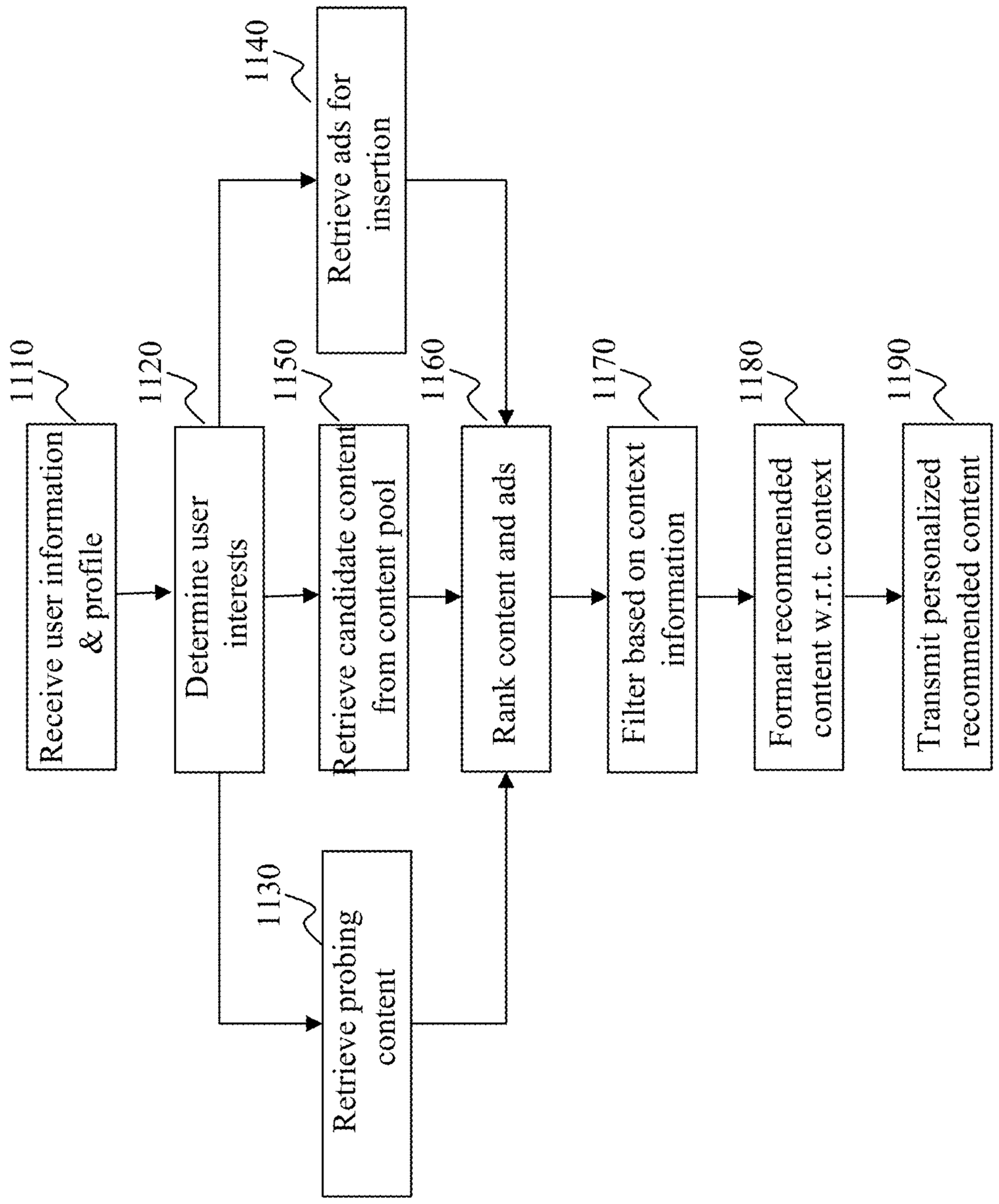

1110
Receive user information & profile
1120
Determine user interests
1130
Retrieve probing content
1140
Retrieve ads for insertion
1150
Retrieve candidate content from content pool
1160
Rank content and ads
1170
Filter based on context information
1180
Format recommended content w.r.t. context
1190
Transmit personalized recommended content

Fig. 11

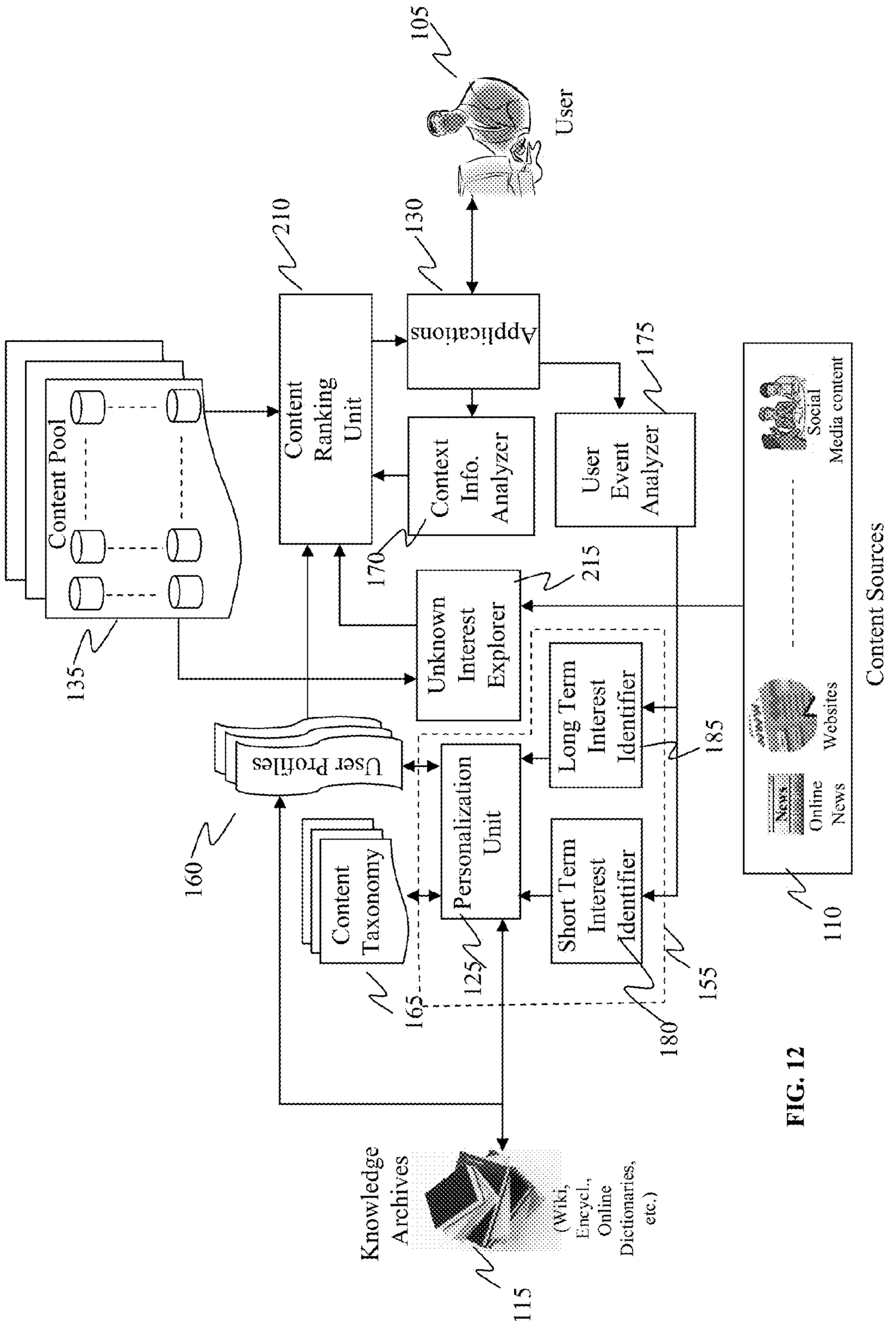

105
User
210
130
Content Pool
135
Content Ranking Unit
Applications
Context Info. Analyzer
170
User Event Analyzer
175
215
Unknown Interest Explorer
Long Term Interest Identifier
185
User Profiles
160
Personalization Unit
Content Taxonomy
165
125
Short Term Interest Identifier
180
155
Social Media content
Online News
Websites
Content Sources
110
Knowledge Archives
(Wiki, Encycl., Online Dictionaries, etc.)
115

FIG. 12

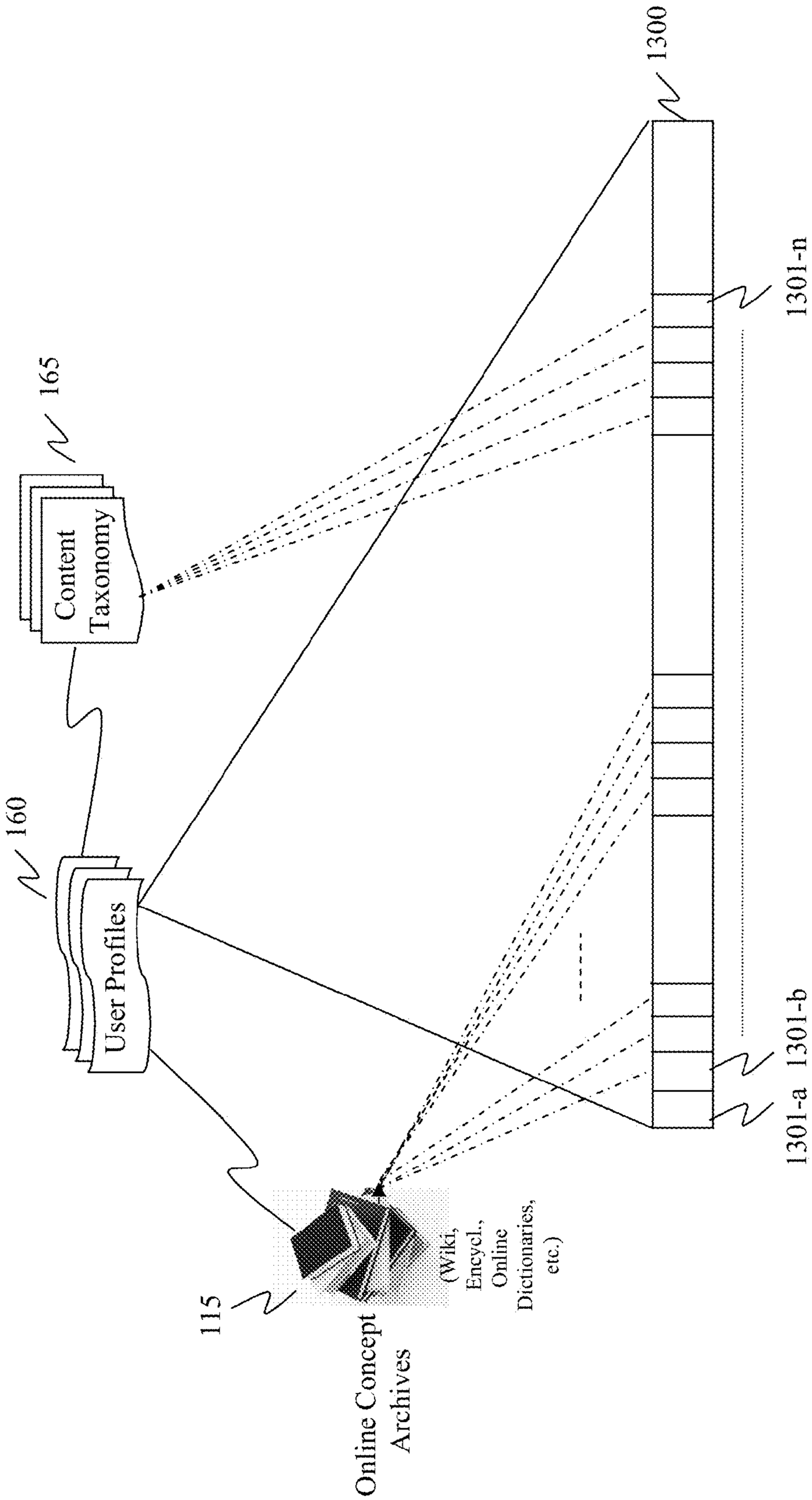

1300
1301-n
165
Content Taxonomy
160
User Profiles
115
Online Concept Archives
(Wiki, Encycl., Online Dictionaries, etc.)
1301-b
1301-a

FIG. 13

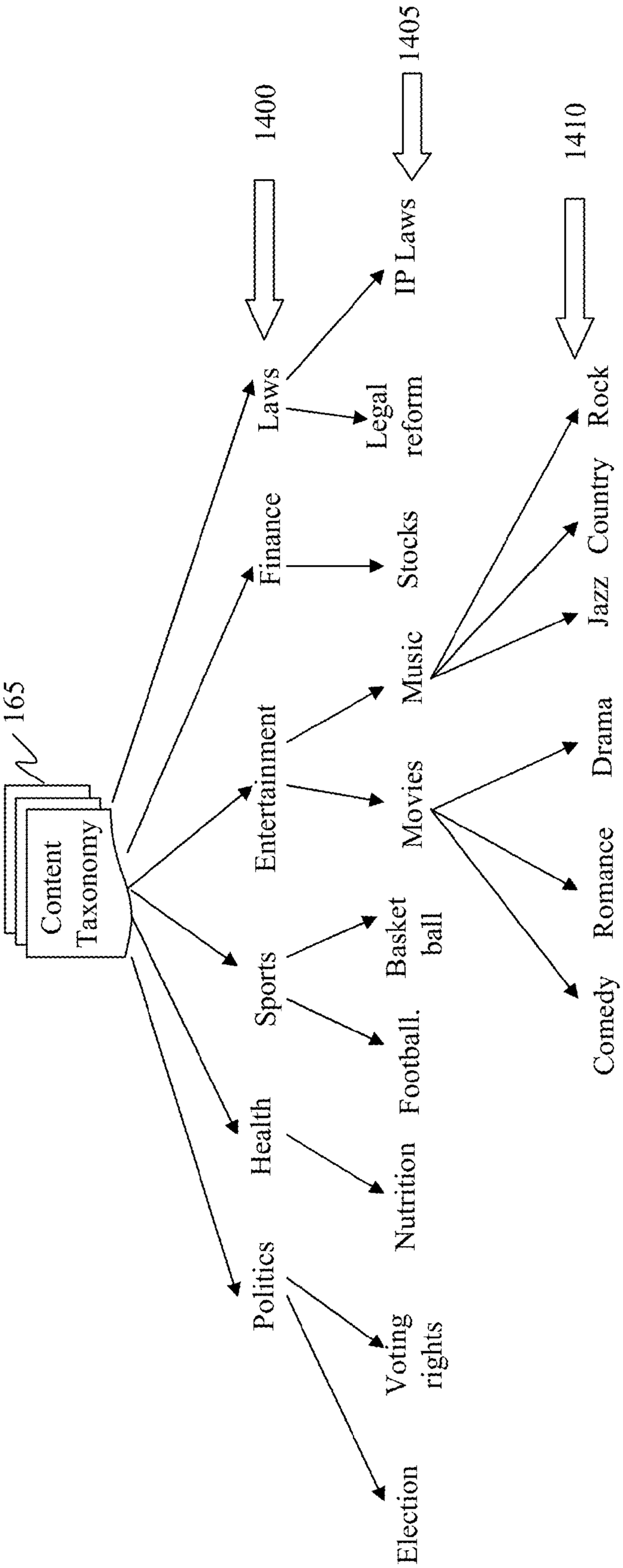

1400
1405
1410
165
Content Taxonomy
Sports
Entertainment
Health
Politics
Finance
Laws
Basket ball
Football.
Nutrition
Voting rights
Election
Legal reform
Stocks
Movies
Music
Comedy
Romance
Drama
Jazz
Country
Rock
IP Laws

FIG. 14

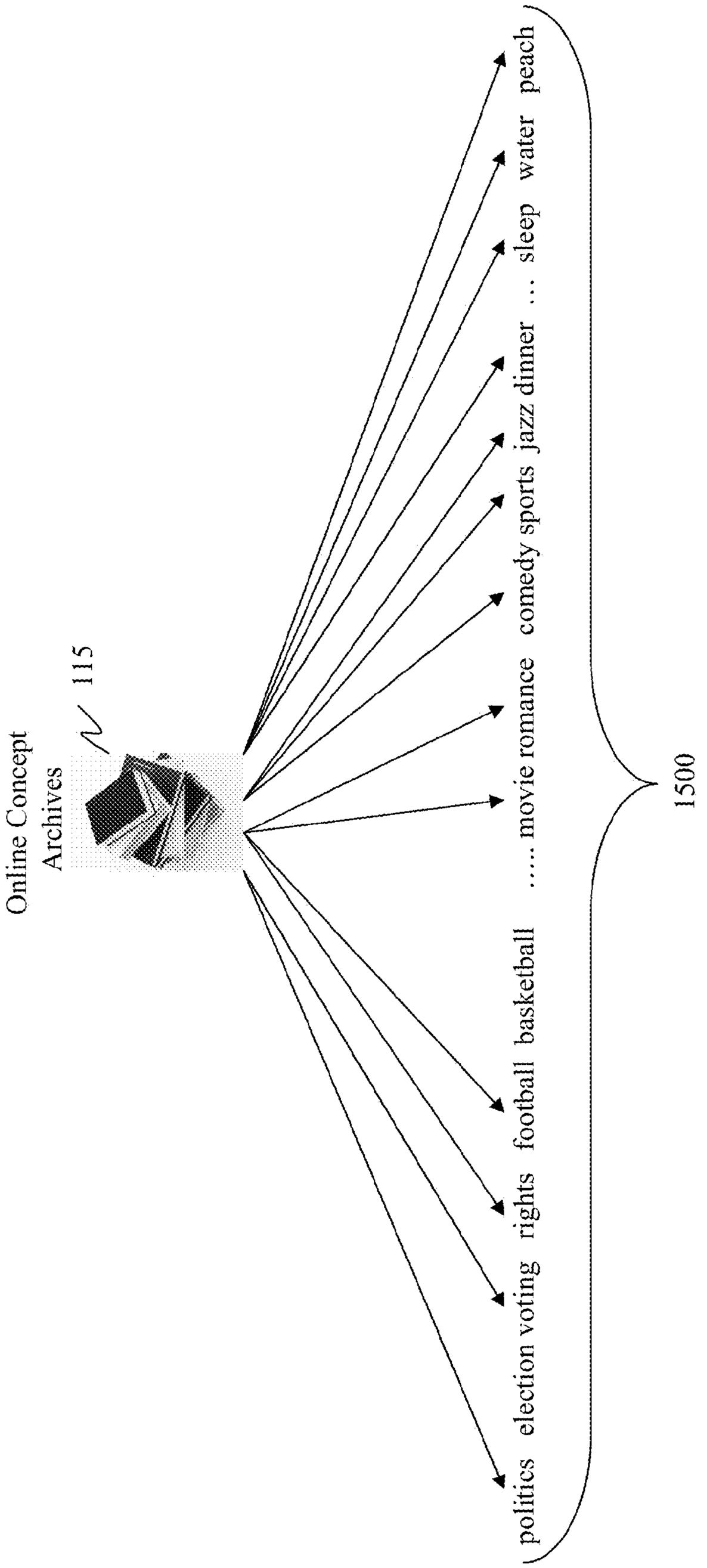

Online Concept Archives
115
politics election voting rights football basketball ..... movie romance comedy sports jazz dinner ... sleep water peach
1500

FIG. 15

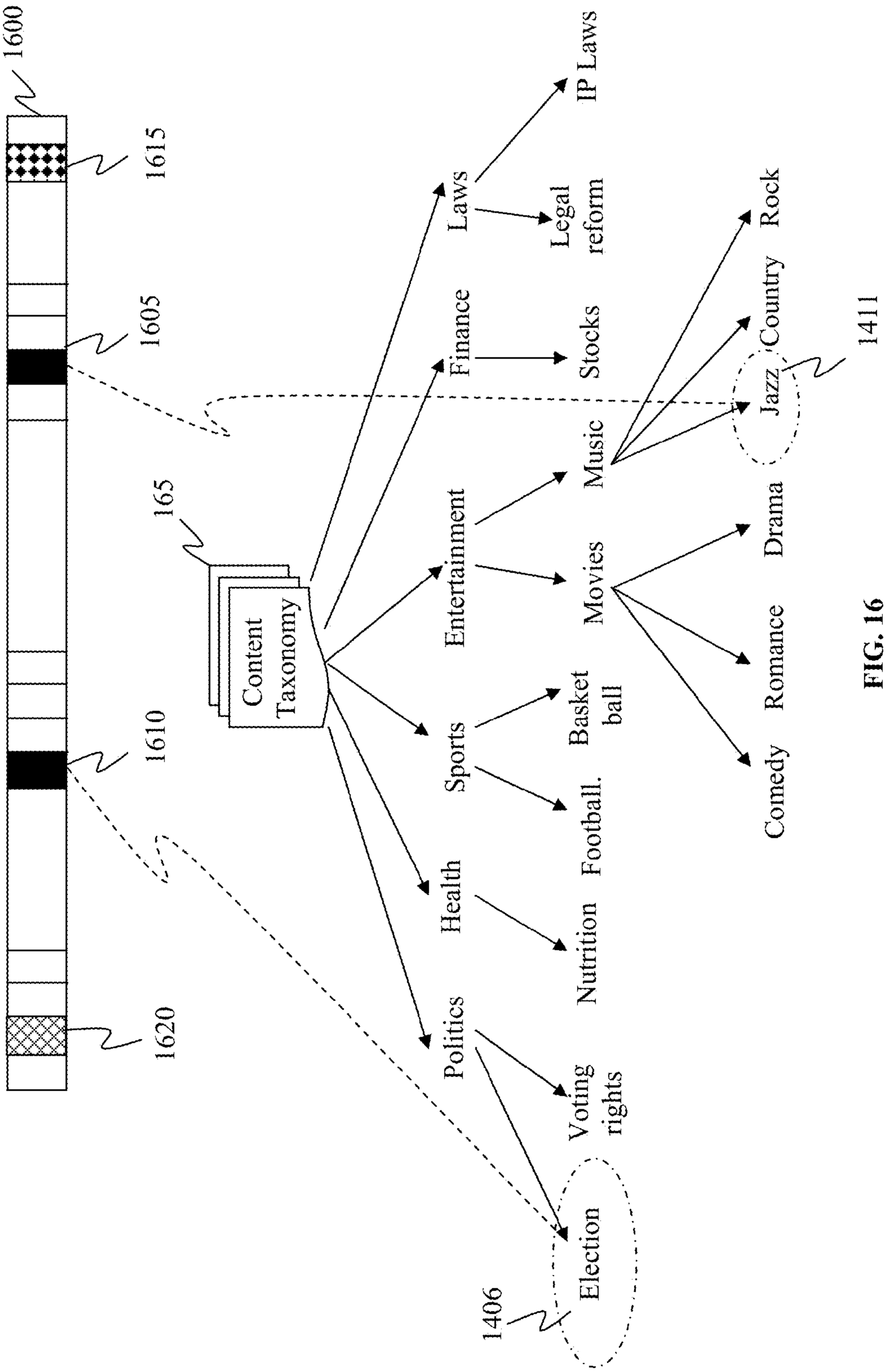

1600
1615
1605
1610
1620
165
Content Taxonomy
Entertainment
Sports
Health
Politics
Finance
Laws
Legal reform
IP Laws
Movies
Music
Basket ball
Football.
Nutrition
Voting rights
Stocks
Drama
Romance
Comedy
Jazz
Country
Rock
1411
1406
Election

FIG. 16

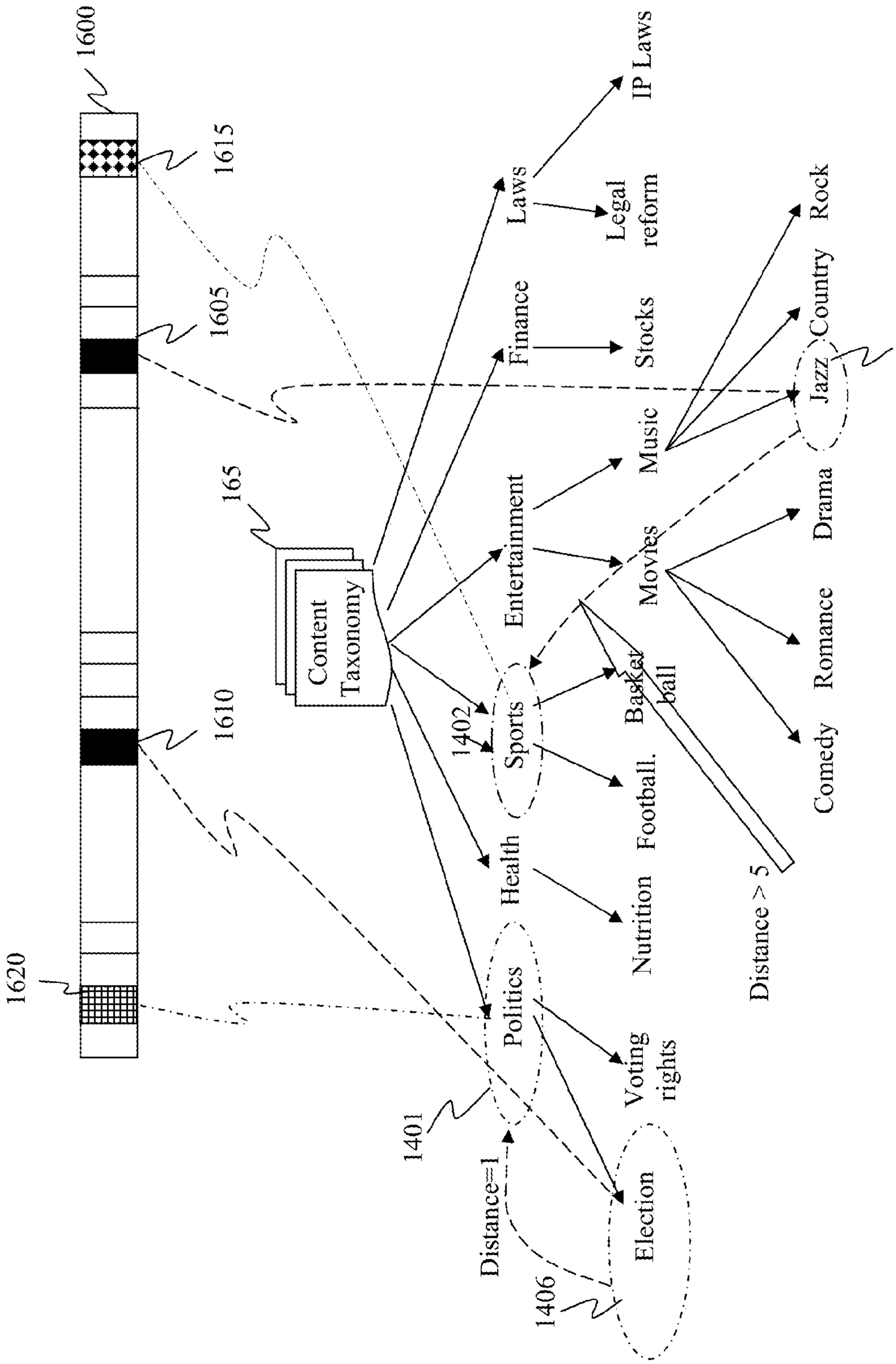

1600
1615
1605
1610
1620
165
Content Taxonomy
1402
1401
1406
1411
Laws
IP Laws
Legal reform
Finance
Stocks
Entertainment
Music
Jazz
Country
Movies
Drama
Romance
Comedy
Sports
Basket ball
Football.
Health
Nutrition
Politics
Voting rights
Election
Distance=1
Distance > 5

FIG. 16a

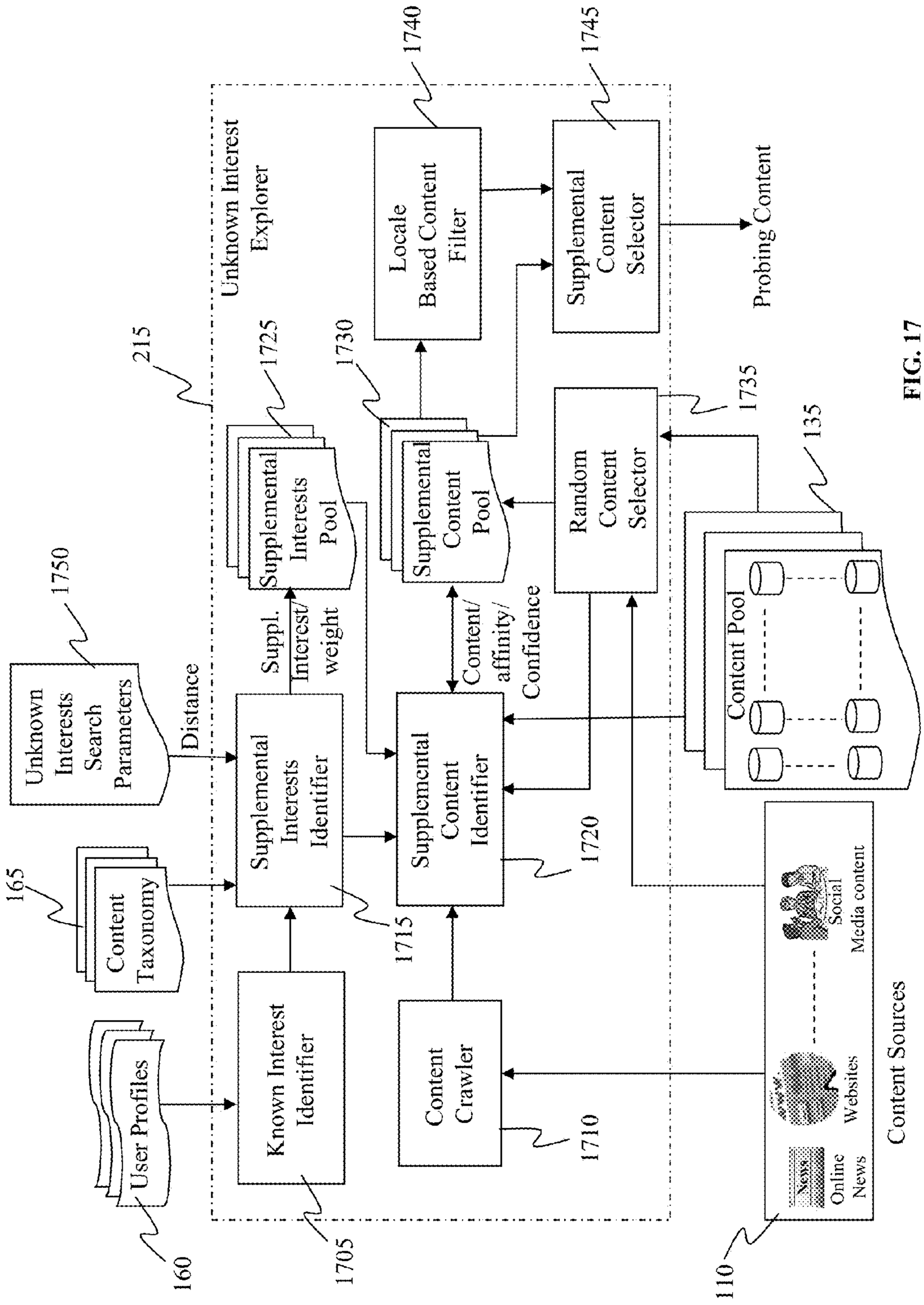

Unknown Interest Explorer
215
Unknown Interests Search Parameters
1750
Distance
Content Taxonomy
165
Supplemental Interests Identifier
1715
Suppl. Interest/ weight
Supplemental Interests Pool
1725
Supplemental Content Identifier
1720
Content/ affinity/ Confidence
Supplemental Content Pool
1730
Locale Based Content Filter
1740
Supplemental Content Selector
1745
Probing Content
Random Content Selector
1735
Content Pool
135
Social Media content
Online News Websites
Content Sources
110
User Profiles
160
Known Interest Identifier
1705
Content Crawler
1710

FIG. 17

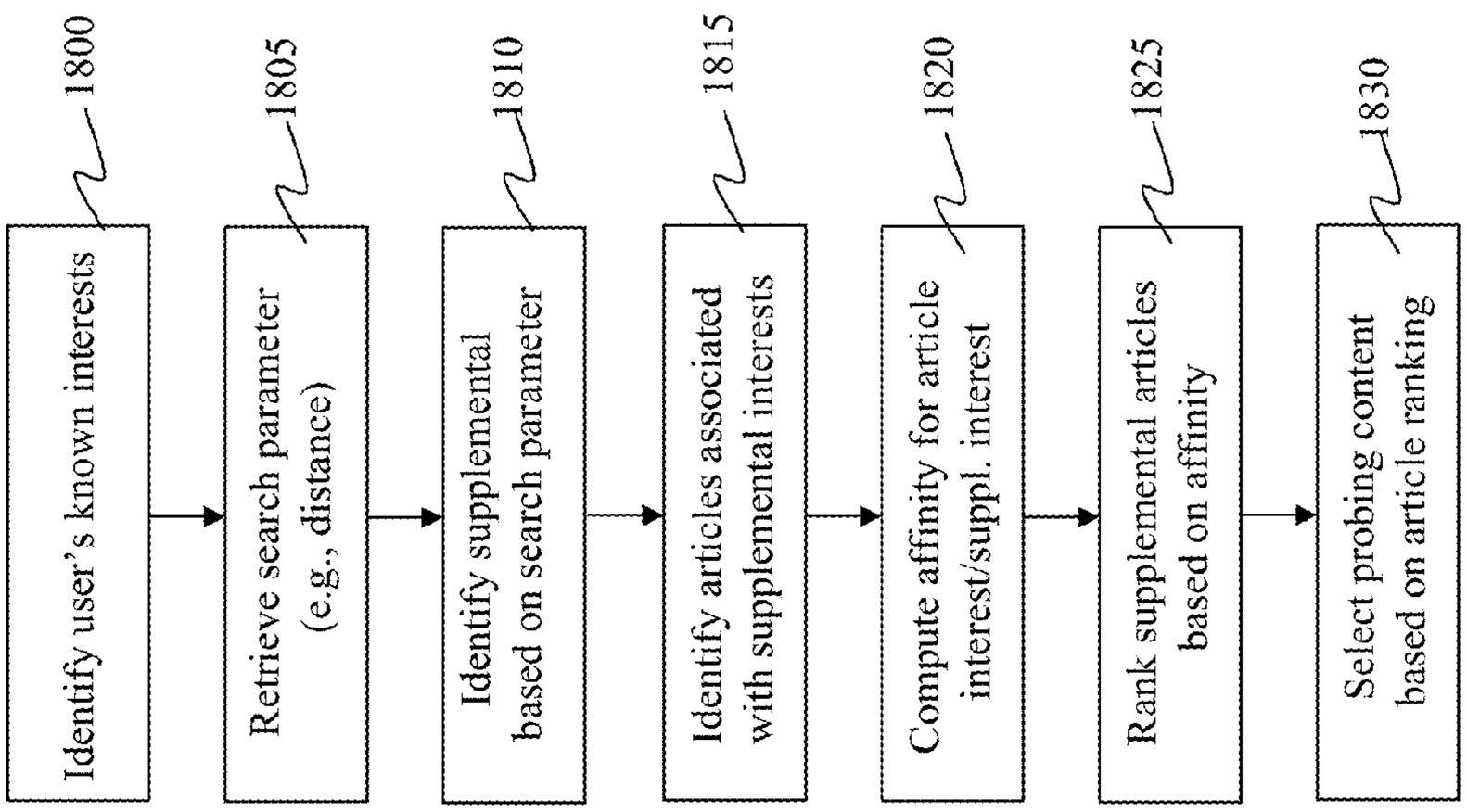

Identify user's known interests
1800
Retrieve search parameter (e.g., distance)
1805
Identify supplemental based on search parameter
1810
Identify articles associated with supplemental interests
1815
Compute affinity for article interest/suppl. interest
1820
Rank supplemental articles based on affinity
1825
Select probing content based on article ranking
1830

FIG. 18

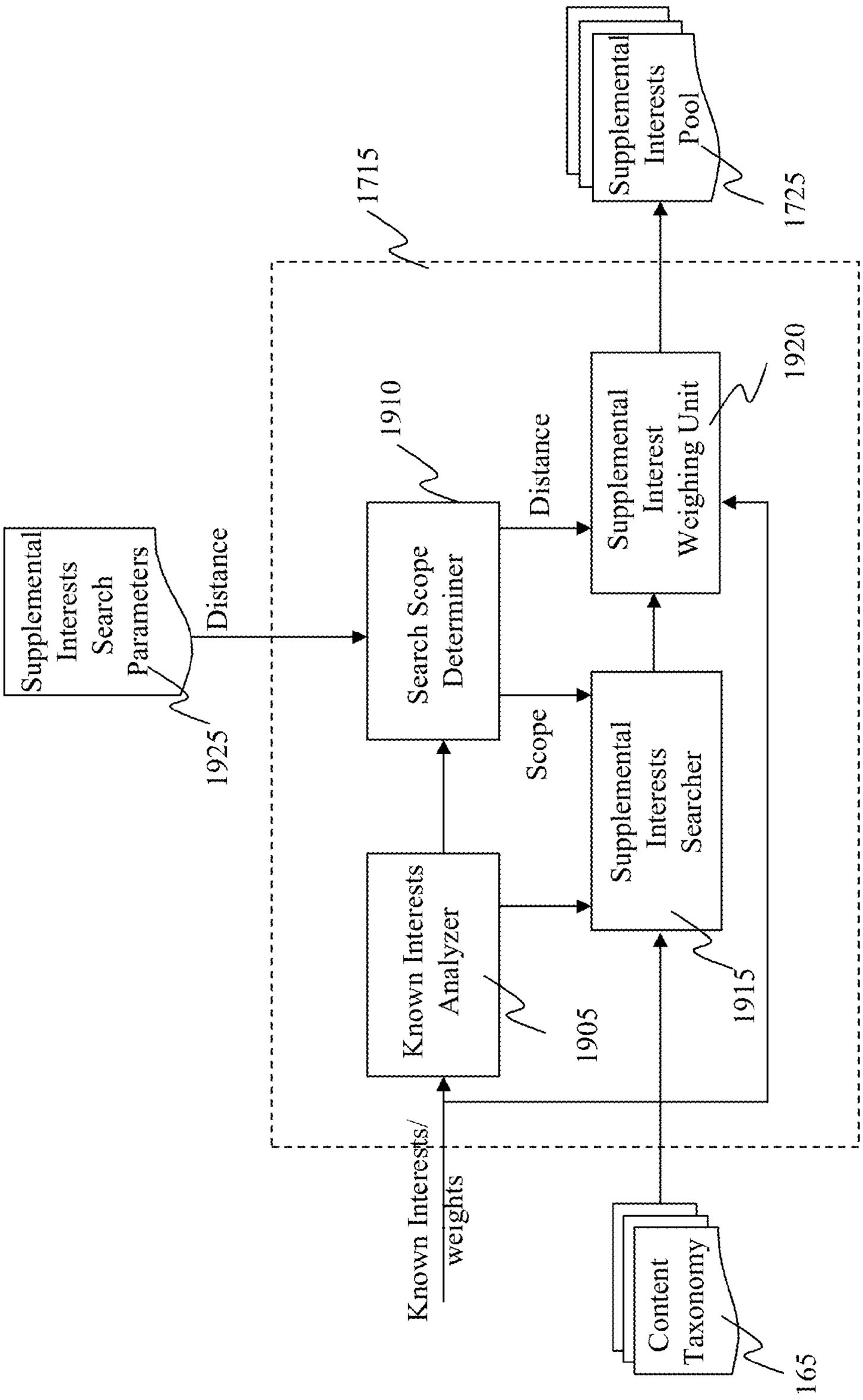

Supplemental Interests Pool
1725
1715
1920
Supplemental Interest Weighing Unit
Distance
1910
Search Scope Determiner
Supplemental Interests Search Parameters
Distance
1925
Scope
Supplemental Interests Searcher
1915
1905
Known Interests Analyzer
Known Interests/ weights
Content Taxonomy
165

FIG. 19

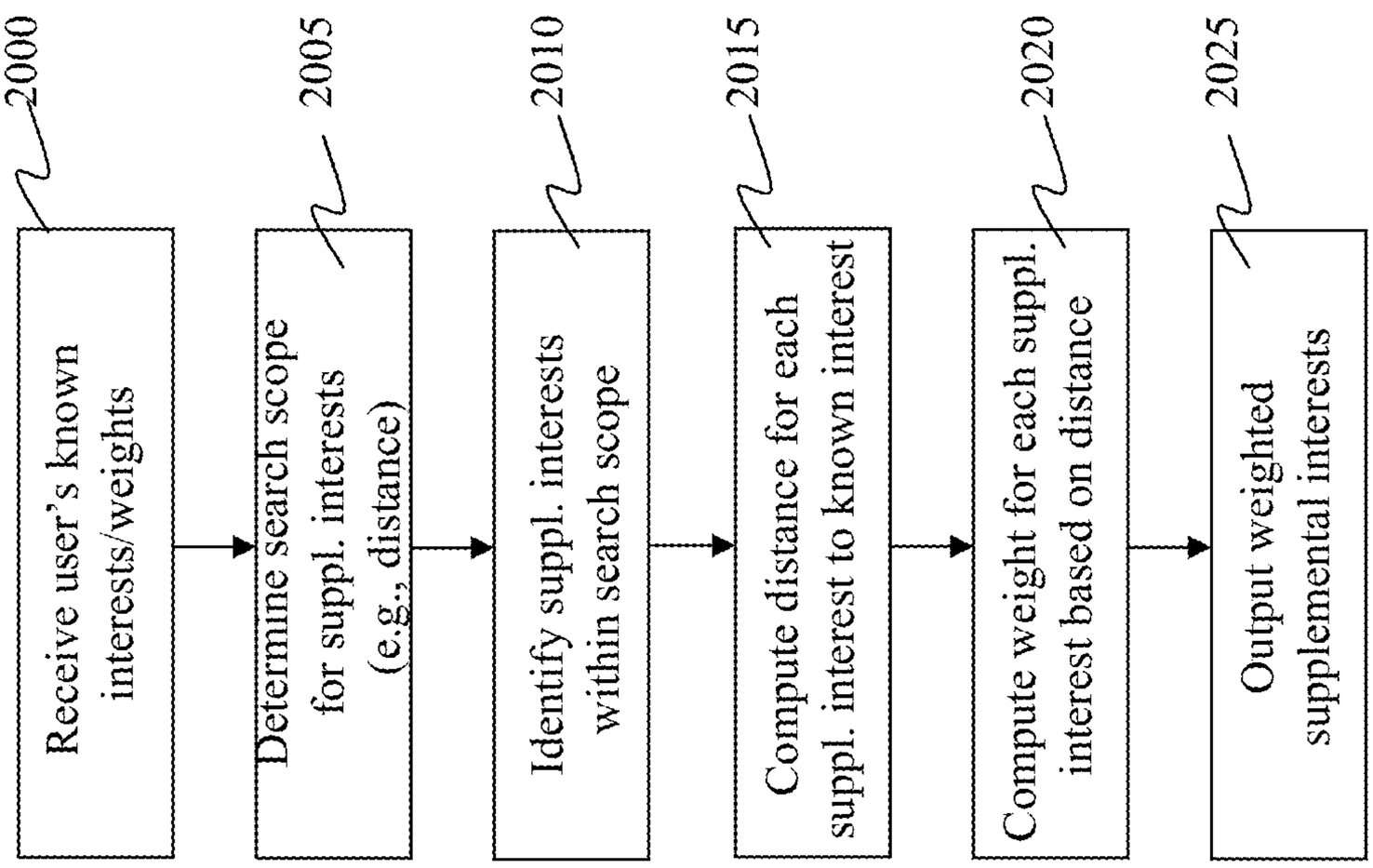

2000
Receive user's known interests/weights
2005
Determine search scope for suppl. interests (e.g., distance)
2010
Identify suppl. interests within search scope
2015
Compute distance for each suppl. interest to known interest
2020
Compute weight for each suppl. interest based on distance
2025
Output weighted supplemental interests

FIG. 20

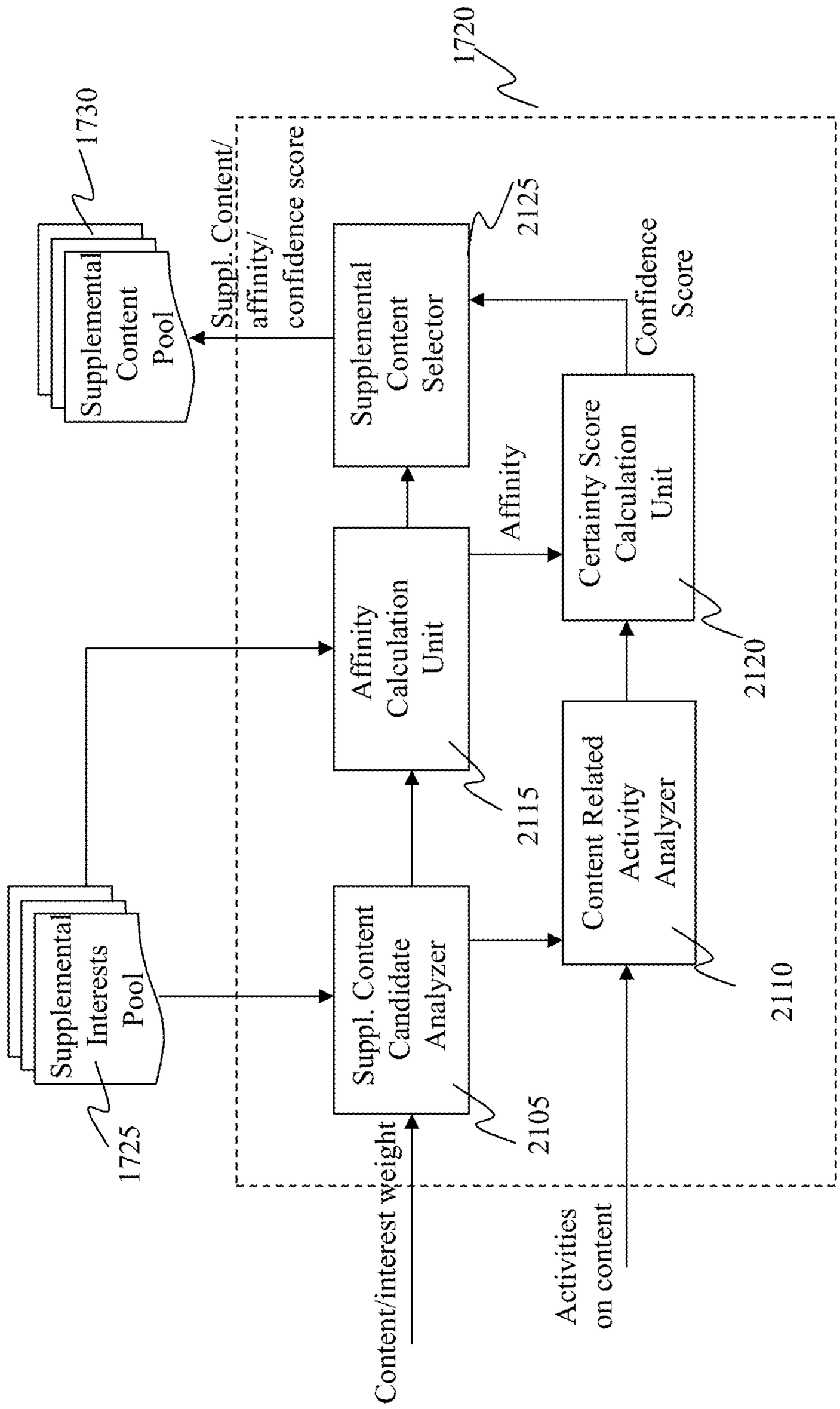

1720
1730
Suppl. Content/ affinity/ confidence score
Supplemental Content Pool
2125
Supplemental Content Selector
Confidence Score
Affinity Calculation Unit
Affinity
Certainty Score Calculation Unit
2120
2115
Content Related Activity Analyzer
Supplemental Interests Pool
Suppl. Content Candidate Analyzer
2110
1725
2105
Content/interest weight
Activities on content

FIG. 21

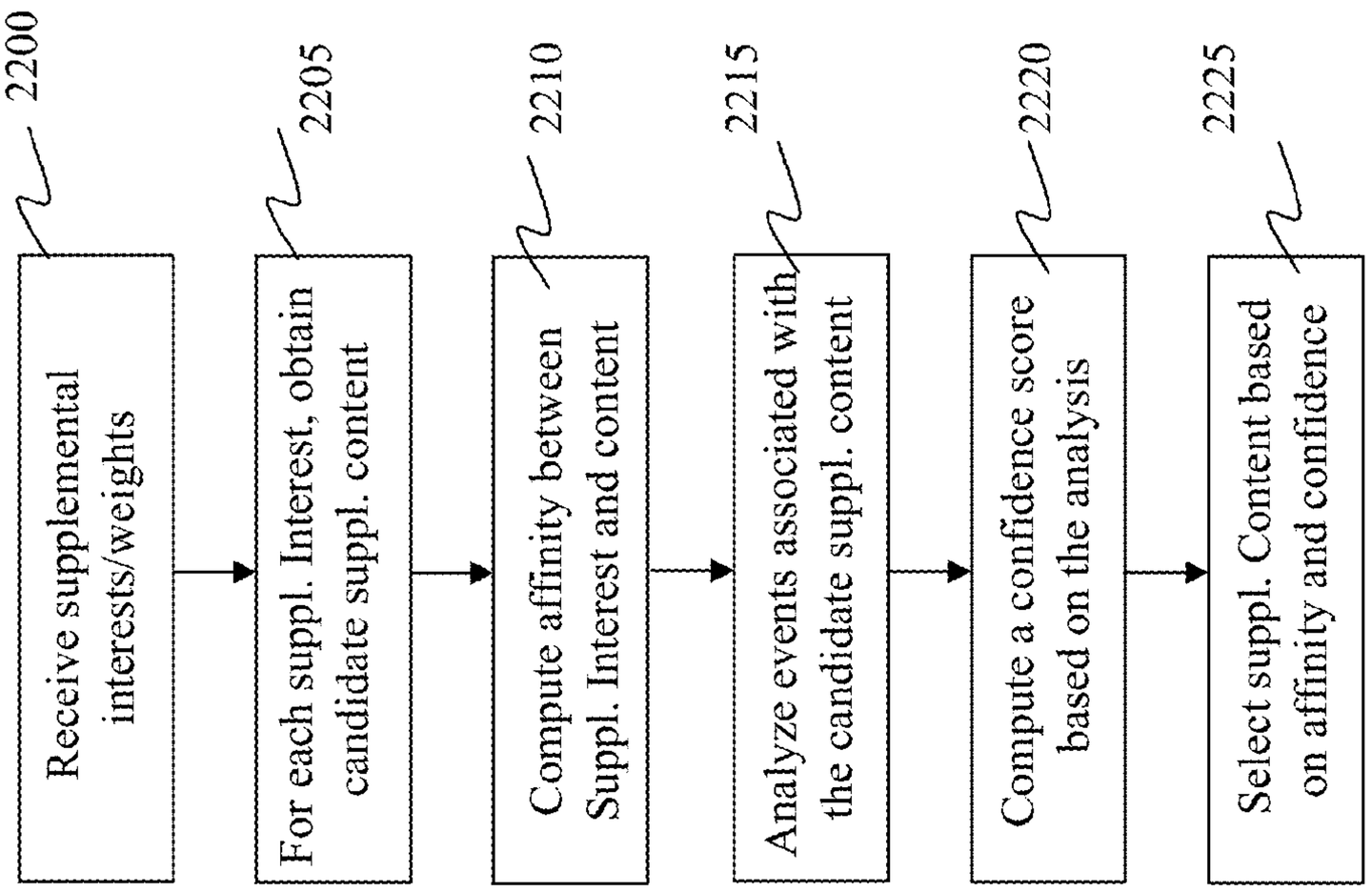

2200
Receive supplemental interests/weights
2205
For each suppl. Interest, obtain candidate suppl. content
2210
Compute affinity between Suppl. Interest and content
2215
Analyze events associated with the candidate suppl. content
2220
Compute a confidence score based on the analysis
2225
Select suppl. Content based on affinity and confidence

FIG. 22

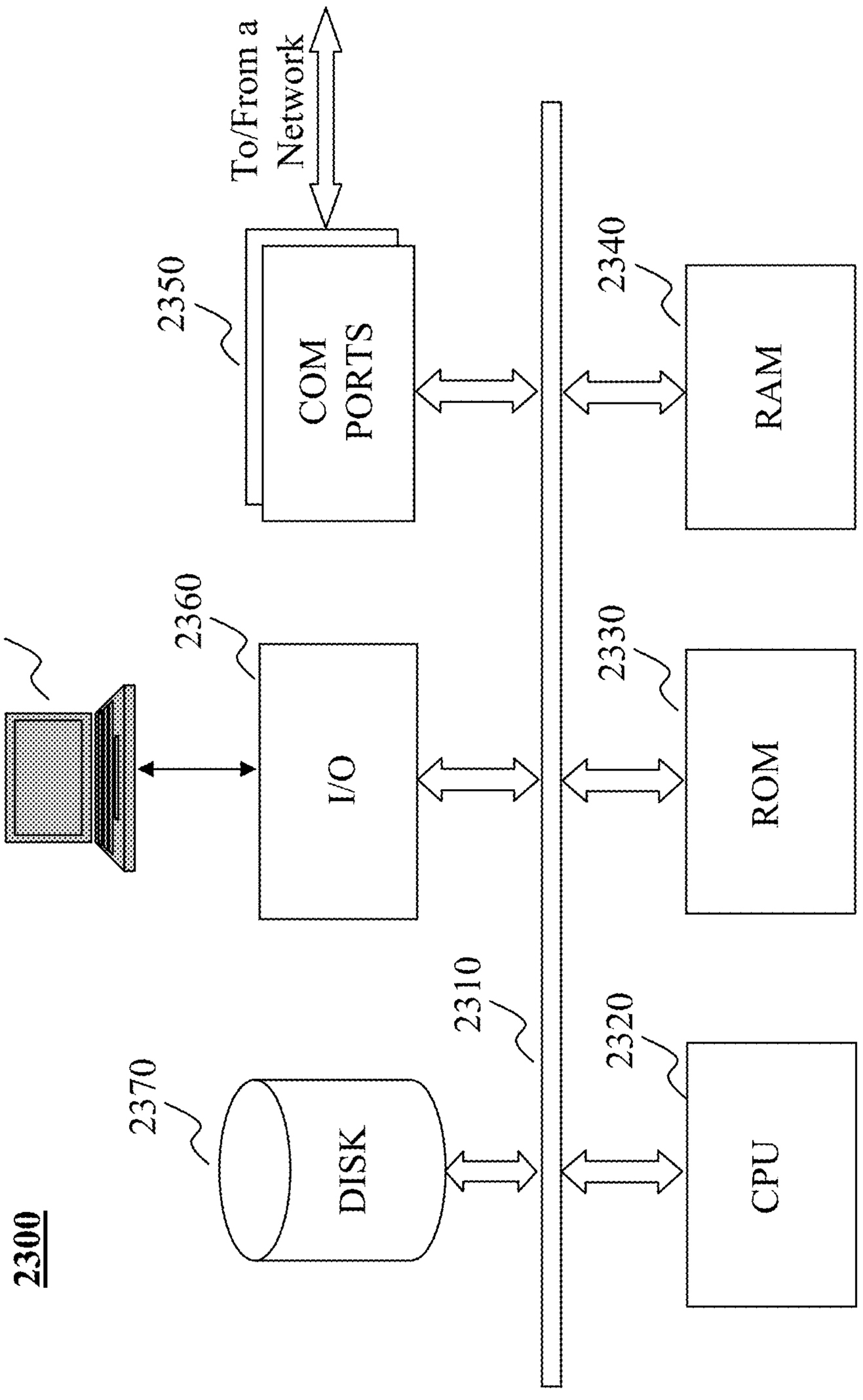

2300
2370
DISK
2310
2320
CPU
2380
2360
I/O
2330
ROM
2340
RAM
2350
COM PORTS
To/From a Network

Fig. 23

METHOD AND SYSTEM FOR DISCOVERY OF USER UNKNOWN INTERESTS BASED ON SUPPLEMENTAL CONTENT

BACKGROUND

1. Technical Field

The present teaching relates to methods and systems for providing content. Specifically, the present teaching relates to methods and systems for providing online content.

2. Discussion of Technical Background

The Internet has made it possible for a user to electronically access virtually any content at anytime and from any location. With the explosion of information, it has become more and more important to provide users with information that is relevant to the user and not just information in general. Further, as users of today's society rely on the Internet as their source of information, entertainment, and/or social connections, e.g., news, social interaction, movies, music, etc, it is critical to provide users with information they find valuable.

Efforts have been made to attempt to allow users to readily access relevant and on the point content. For example, topical portals have been developed that are more subject matter oriented as compared to generic content gathering systems such as traditional search engines. Example topical portals include portals on finance, sports, news, weather, shopping, music, art, film, etc. Such topical portals allow users to access information related to subject matters that these portals are directed to. Users have to go to different portals to access content of certain subject matter, which is not convenient and not user centric.

Another line of efforts in attempting to enable users to easily access relevant content is via personalization, which aims at understanding each user's individual likings/interests/preferences so that an individualized user profile for each user can be set up and can be used to select content that matches a user's interests. The underlying goal is to meet the minds of users in terms of content consumption. User profiles traditionally are constructed based on users' declared interests and/or inferred from, e.g., users' demographics. There have also been systems that identify users' interests based on observations made on users' interactions with content. A typical example of such user interaction with content is click through rate (CTR).

These traditional approaches have various shortcomings. For example, users' interests are profiled without any reference to a baseline so that the level of interest can be more accurately estimated. User interests are detected in isolated application settings so that user profiling in individual applications cannot capture a broad range of the overall interests of a user. Such traditional approach to user profiling lead to fragmented representation of user interests without a coherent understanding of the users' preferences. Because profiles of the same user derived from different application settings are often grounded with respect to the specifics of the applications, it is also difficult to integrate them to generate a more coherent profile that better represent the user's interests.

User activities directed to content are traditionally observed and used to estimate or infer users' interests. CTR is the most commonly used measure to estimate users' interests. However, CTR is no longer adequate to capture users' interests particularly given that different types of activities that a user may perform on different types of devices may also reflect or implicate user's interests. In addition, user reactions to content usually represent users' short term interests. Such observed short term interests, when acquired piece meal, as traditional approaches often do, can only lead to reactive,

rather than proactive, services to users. Although short term interests are important, they are not adequate to enable understanding of the more persistent long term interests of a user, which are crucial in terms of user retention. Most user interactions with content represent short term interests of the user so that relying on such short term interest behavior makes it difficult to expand the understanding of the increasing range of interests of the user. When this is in combination with the fact that such collected data is always the past behavior and collected passively, it creates a personalization bubble, making it difficult, if not impossible, to discover other interests of a user unless the user initiates some action to reveal new interests.

Yet another line of effort to allow users to access relevant content is to pooling content that may be interested by users in accordance with their interests. Given the explosion of information on the Internet, it is not likely, even if possible, to evaluate all content accessible via the Internet whenever there is a need to select content relevant to a particular user. Thus, realistically, it is needed to identify a subset or a pool of the Internet content based on some criteria so that content can be selected from this pool and recommended to users based on their interests for consumption.

Conventional approaches to creating such a subset of content are application centric. Each application carves out its own subset of content in a manner that is specific to the application. For example, Amazon.com may have a content pool related to products and information associated thereof created/updated based on information related to its own users and/or interests of such users exhibited when they interact with Amazon.com. Facebook also has its own subset of content, generated in a manner not only specific to Facebook but also based on user interests exhibited while they are active on Facebook. As a user may be active in different applications (e.g., Amazon.com and Facebook) and with each application, they likely exhibit only part of their overall interests in connection with the nature of the application. Given that, each application can usually gain understanding, at best, of partial interests of users, making it difficult to develop a subset of content that can be used to serve a broader range of users' interests.

Another line of effort is directed to personalized content recommendation, i.e., selecting content from a content pool based on the user's personalized profiles and recommending such identified content to the user. Conventional solutions focus on relevance, i.e., the relevance between the content and the user. Although relevance is important, there are other factors that also impact how recommendation content should be selected in order to satisfy a user's interests. Most content recommendation systems insert advertisement to content identified for a user for recommendation. Some traditional systems that are used to identify insertion advertisements match content with advertisement or user's query (also content) with advertisement, without considering matching based on demographics of the user with features of the target audience defined by advertisers. Some traditional systems match user profiles with the specified demographics of the target audience defined by advertisers but without matching the content to be provided to the user and the advertisement. The reason is that content is often classified into taxonomy based on subject matters covered in the content yet advertisement taxonomy is often based on desired target audience groups. This makes it less effective in terms of selecting the most relevant advertisement to be inserted into content to be recommended to a specific user.

There is a need for improvements over the conventional approaches to personalizing content recommendation.

SUMMARY

The teachings disclosed herein relate to methods, systems, and programming for providing personalized web page layouts. In an embodiment a method for identifying content for a user is disclosed, the method is implemented on a computing device having at least one processor, storage, and a communication interface connected to a network. The method comprising retrieving user information related to a user, wherein the information indicates one or more interests of the user, identifying at least one interest of the user, determining one or more supplemental interests with respect to each of the at least one interest of the user, where the one or more supplemental interests do not overlap with the one or more interests of the user, and identifying supplemental content associated with the one or more supplemental interests with respect to each of the at least one interest of the user, wherein the supplemental content associated with the one or more supplemental interests is used to discover unknown interest of the user.

In another embodiment, the method further comprises identifying relatedness between each piece of the supplemental content and its corresponding supplemental interest, ranking each piece of the supplemental content based on the relatedness, selecting at least some of the supplemental content based on the ranking, and outputting the selected supplemental content.

In another embodiment, the method further comprises retrieving random content from a content pool, adding the random content to the supplemental content, selecting the random content, and outputting the random content. In still another embodiment, the method further comprises filtering the ranked supplemental content based on a criteria. In still another embodiment, the criteria is demographics. In an embodiment, a system for identifying unknown user content is disclosed. The system comprises a retrieval unit for retrieving user information related to a user, wherein the information indicates one or more interests of the user, an interest analyzer for identifying at least one interest of the user, a supplemental interest identifier for determining one or more supplemental interests with respect to each of the at least one interest of the user, where the one or more supplemental interests do not overlap with the one or more interests of the user, and a supplemental content identifier for identifying supplemental content associated with the one or more supplemental interests with respect to each of the at least one interest of the user, wherein the supplemental content associated with the one or more supplemental interests is used to discover unknown interest of the user.

In another embodiment the system further comprises a supplemental weighting unit for identifying relatedness between each piece of the supplemental content and its corresponding supplemental interest, a ranking unit for ranking each piece of the supplemental content based on the relatedness, a selector for selecting at least some of the supplemental content based on the ranking, and an output for outputting the selected supplemental content.

In an embodiment, a non-transitory computer readable medium having recorded thereon information for identifying unknown user interest is disclosed. The medium, when read by a computer, causes the computer to perform the steps of retrieving user information related to a user, wherein the information indicates one or more interests of the user, identifying at least one interest of the user, determining one or more supplemental interests with respect to each of the at least one interest of the user, where the one or more supplemental interests do not overlap with the one or more interests of the user, and, identifying supplemental content associated with the one or more supplemental interests with respect to each of the at least one interest of the user, wherein the supplemental content associated with the one or more supplemental interests is used to discover unknown interest of the user.

In another embodiment, the medium when read by the computer, further causes the computer to perform the steps of identifying relatedness between each piece of the supplemental content and its corresponding supplemental interest, ranking each piece of the supplemental content based on the relatedness, selecting at least some of the supplemental content based on the ranking and outputting the selected supplemental content.

BRIEF DESCRIPTION OF THE DRAWINGS

The methods, systems and/or programming described herein are further described in terms of exemplary embodiments. These exemplary embodiments are described in detail with reference to the drawings. These embodiments are non-limiting exemplary embodiments, in which like reference numerals represent similar structures throughout the several views of the drawings, and wherein:

FIG. **1** depicts an exemplary system diagram for personalized content recommendation, according to an embodiment of the present teaching;

FIG. **2** is a flowchart of an exemplary process for personalized content recommendation, according to an embodiment of the present teaching;

FIG. **3** illustrates exemplary types of context information;

FIG. **4** depicts an exemplary diagram of a content pool generation/update unit, according to an embodiment of the present teaching;

FIG. **5** is a flowchart of an exemplary process of creating a content pool, according to an embodiment of the present teaching;

FIG. **6** is a flowchart of an exemplary process for updating a content pool, according to an embodiment of the present teaching;

FIG. **7** depicts an exemplary diagram of a user understanding unit, according to an embodiment of the present teaching;

FIG. **8** is a flowchart of an exemplary process for generating a baseline interest profile, according to an embodiment of the present teaching;

FIG. **9** is a flowchart of an exemplary process for generating a personalized user profile, according to an embodiment of the present teaching;

FIG. **10** depicts an exemplary system diagram for a content ranking unit, according to an embodiment of the present teaching;

FIG. **11** is a flowchart of an exemplary process for the content ranking unit, according to an embodiment of the present teaching;

FIG. **12** is a diagram illustrating a portion of a personalization system utilized to find and deliver content related to a user's unknown interests, in accordance with one embodiment of the present teaching;

FIG. **13** is a diagram illustrating a high dimensional vector of user interest, in accordance with another embodiment of the present teaching;

FIG. **14** is a diagram illustrating a typical structured content taxonomy in an embodiment of the present teaching;

FIG. **15** is a diagram illustrating an on-line concept archive or index according to embodiments of the present teaching;

FIG. **16** is a diagram illustrating a high dimensional vector of user interest mapped to a content taxonomy according to one embodiment of the present teaching;

FIG. **16***a* is a diagram illustrating a high dimensional vector of user interest mapped to a content taxonomy and indicating potentially other relevant interests;

FIG. **17** is a diagram illustrating an unknown interest explorer in accordance with an embodiment of the present teaching;

FIG. **18** is a flow diagram illustrating a method of implementing an unknown interest explorer in accordance with an embodiment of the present teaching.

FIG. **19** is a diagram illustrating a supplemental interest identifier in accordance with an embodiment of the present teaching;

FIG. **20** is flow diagram illustrating a method of implementing a supplemental interest identifier in accordance with an embodiment of the present teaching;

FIG. **21** is a diagram illustrating a supplemental content identifier in accordance with an embodiment of the present teaching;

FIG. **22** is a flow diagram illustrating a method of implementing a supplemental content identifier in accordance with an embodiment of the present teaching; and

FIG. **23** depicts a general computer architecture on which the present teaching can be implemented.

DETAILED DESCRIPTION

In the following detailed description, numerous specific details are set forth by way of examples in order to provide a thorough understanding of the relevant teachings. However, it should be apparent to those skilled in the art that the present teachings may be practiced without such details. In other instances, well known methods, procedures, components, and/or circuitry have been described at a relatively high-level, without detail, in order to avoid unnecessarily obscuring aspects of the present teachings.

The present teaching relates to personalizing on-line content recommendations to a user. Particularly, the present teaching relates to a system, method, and/or programs for personalized content recommendation that addresses the shortcomings associated the conventional content recommendation solutions in personalization, content pooling, and recommending personalized content.

With regard to personalization, the present teaching identifies a user's interests with respect to a universal interest space, defined via known concept archives such as Wikipedia and/or content taxonomy. Using such a universal interest space, interests of users, exhibited in different applications and via different platforms, can be used to establish a general population's profile as a baseline against which individual user's interests and levels thereof can be determined. For example, users active in a third party application such as Facebook or Twitter and the interests that such users exhibited in these third party applications can be all mapped to the universal interest space and then used to compute a baseline interest profile of the general population. Specifically, each user's interests observed with respect to each document covering certain subject matters or concepts can be mapped to, e.g., Wikipedia or certain content taxonomy. A high dimensional vector can be constructed based on the universal interest space in which each attribute of the vector corresponds to a concept in the universal space and the value of the attribute may corresponds to an evaluation of the user's interest in this particular concept. The general baseline interest profile can be derived based on all vectors represent the population. Each vector representing an individual can be normalized against the baseline interest profile so that the relative level of interests of the user with respect to the concepts in the universal interest space can be determined. This enables better understanding of the level of interests of the user in different subject matters with respect to a more general population and result in enhanced personalization for content recommendation. Rather than characterizing users' interests merely according to proprietary content taxonomy, as is often done in the prior art, the present teaching leverages public concept archives, such as Wikipedia or online encyclopedia, to define a universal interest space in order to profile a user's interests in a more coherent manner. Such a high dimensional vector captures the entire interest space of every user, making person-to-person comparison as to personal interests more effective. Profiling a user and in this manner also leads to efficient identification of users who share similar interests. In addition, content may also be characterized in the same universal interest space, e.g., a high dimensional vector against the concepts in the universal interest space can also be constructed with values in the vector indicating whether the content covers each of the concepts in the universal interest space. By characterizing users and content in the same space in a coherent way, the affinity between a user and a piece of content can be determined via, e.g., a dot product of the vector for the user and the vector for the content.

The present teaching also leverages short term interests to better understand long term interests of users. Short term interests can be observed via user online activities and used in online content recommendation, the more persistent long term interests of a user can help to improve content recommendation quality in a more robust manner and, hence, user retention rate. The present teaching discloses discovery of long term interests as well as short term interests.

To improve personalization, the present teaching also discloses ways to improve the ability to estimate a user's interest based on a variety of user activities. This is especially useful because meaningful user activities often occur in different settings, on different devices, and in different operation modes. Through such different user activities, user engagement to content can be measured to infer users' interests. Traditionally, clicks and click through rate (CTR) have been used to estimate users' intent and infer users' interests. CTR is simply not adequate in today's world. Users may dwell on a certain portion of the content, the dwelling may be for different lengths of time, users may scroll along the content and may dwell on a specific portion of the content for some length of time, users may scroll down at different speeds, users may change such speed near certain portions of content, users may skip certain portion of content, etc. All such activities may have implications as to users' engagement to content. Such engagement can be utilized to infer or estimate a user's interests. The present teaching leverages a variety of user activities that may occur across different device types in different settings to achieve better estimation of users' engagement in order to enhance the ability of capturing a user's interests in a more reliable manner.

Another aspect of the present teaching with regard to personalization is its ability to explore unknown interests of a user by generating probing content. Traditionally, user profiling is based on either user provided information (e.g., declared interests) or passively observed past information such as the content that the user has viewed, reactions to such content, etc. Such prior art schemes can lead to a personalization bubble where only interests that the user revealed can be used for content recommendation. Because of that, the only user activities that can be observed are directed to such

known interests, impeding the ability to understand the overall interest of a user. This is especially so considering the fact that users often exhibit different interests (mostly partial interests) in different application settings. The present teaching discloses ways to generate probing content with concepts that is currently not recognized as one of the user's interests in order to explore the user's unknown interests. Such probing content is selected and recommended to the user and user activities directed to the probing content can then be analyzed to estimate whether the user has other interests. The selection of such probing content may be based on a user's current known interests by, e.g., extrapolating the user's current interests. For example, for some known interests of the user (e.g., the short term interests at the moment), some probing concepts in the universal interest space, for which the user has not exhibited interests in the past, may be selected according to some criteria (e.g., within a certain distance from the user's current known interest in a taxonomy tree) and content related to such probing concepts may then be selected and recommended to the user. Another way to identify probing concept (corresponding to unknown interest of the user) may be through the user's cohorts. For instance, a user may share certain interests with his/her cohorts but some members of the circle may have some interests that the user has never exhibited before. Such un-shared interests with cohorts may be selected as probing unknown interests for the user and content related to such probing unknown interests may then be selected as probing content to be recommended to the user. In this manner, the present teaching discloses a scheme by which a user's interests can be continually probed and understood to improve the quality of personalization. Such managed probing can also be combined with random selection of probing content to allow discovery of unknown interests of the user that are far removed from the user's current known interests.

A second aspect of recommending quality personalized content is to build a content pool with quality content that covers subject matters interesting to users. Content in the content pool can be rated in terms of the subject and/or the performance of the content itself. For example, content can be characterized in terms of concepts it discloses and such a characterization may be generated with respect to the universal interest space, e.g., defined via concept archive(s) such as content taxonomy and/or Wikipedia and/or online encyclopedia, as discussed above. For example, each piece of content can be characterized via a high dimensional vector with each attribute of the vector corresponding to a concept in the interest universe and the value of the attribute indicates whether and/or to what degree the content covers the concept. When a piece of content is characterized in the same universal interest space as that for user's profile, the affinity between the content and a user profile can be efficiently determined.

Each piece of content in the content pool can also be individually characterized in terms of other criteria. For example, performance related measures, such as popularity of the content, may be used to describe the content. Performance related characterizations of content may be used in both selecting content to be incorporated into the content pool as well as selecting content already in the content pool for recommendation of personalized content for specific users. Such performance oriented characterizations of each piece of content may change over time and can be assessed periodically and can be done based on users' activities. Content pool also changes over time based on various reasons, such as content performance, change in users' interests, etc. Dynamically changed performance characterization of content in the content pool may also be evaluated periodically or dynamically based on performance measures of the content so that the content pool can be adjusted over time, i.e., by removing low performance content pieces, adding new content with good performance, or updating content.

To grow the content pool, the present teaching discloses ways to continually discover both new content and new content sources from which interesting content may be accessed, evaluated, and incorporated into the content pool. New content may be discovered dynamically via accessing information from third party applications which users use and exhibit various interests. Examples of such third party applications include Facebook, Twitter, Microblogs, or YouTube. New content may also be added to the content pool when some new interest or an increased level of interests in some subject matter emerges or is predicted based on the occurrence of certain (spontaneous) events. One example is the content about the life of Pope Benedict, which in general may not be a topic of interests to most users but likely will be in light of the surprising announcement of Pope Benedict's resignation. Such dynamic adjustment to the content pool aims at covering a dynamic (and likely growing) range of interests of users, including those that are, e.g., exhibited by users in different settings or applications or predicted in light of context information. Such newly discovered content may then be evaluated before it can be selected to be added to the content pool.

Certain content in the content pool, e.g., journals or news, need to be updated over time. Conventional solutions usually update such content periodically based on a fixed schedule. The present teaching discloses the scheme of dynamically determining the pace of updating content in the content pool based on a variety of factors. Content update may be affected by context information. For example, the frequency at which a piece of content scheduled to be updated may be every 2 hours, but this frequency can be dynamically adjusted according to, e.g., an explosive event such as an earthquake. As another example, content from a social group on Facebook devoted to Catholicism may normally be updated daily. When Pope Benedict's resignation made the news, the content from that social group may be updated every hour so that interested users can keep track of discussions from members of this social group. In addition, whenever there are newly identified content sources, it can be scheduled to update the content pool by, e.g., crawling the content from the new sources, processing the crawled content, evaluating the crawled content, and selecting quality new content to be incorporated into the content pool. Such a dynamically updated content pool aims at growing in compatible with the dynamically changing users' interests in order to facilitate quality personalized content recommendation.

Another key to quality personalized content recommendation is the aspect of identifying quality content that meets the interests of a user for recommendation. Previous solutions often emphasize mere relevance of the content to the user when selecting content for recommendation. In addition, traditional relevance based content recommendation was mostly based on short term interests of the user. This not only leads to a content recommendation bubble, i.e., known short interests cause recommendations limited to the short term interests and reactions to such short term interests centric recommendations cycle back to the short term interests that start the process. This bubble makes it difficult to come out of the circle to recommend content that can serve not only the overall interests but also long term interests of users. The present teaching combines relevance with performance of the content so that not only relevant but also quality content can be selected and recommended to users in a multi-stage ranking system.

In addition, to identify recommended content that can serve a broad range of interests of a user, the present teaching relies on both short term and long term interests of the user to identify user-content affinity in order to select content that meets a broader range of users' interests to be recommended to the user.

In content recommendation, monetizing content such as advertisements are usually also selected as part of the recommended content to a user. Traditional approaches often select ads based on content in which the ads are to be inserted. Some traditional approaches also rely on user input such as queries to estimate what ads likely can maximize the economic return. These approaches select ads by matching the taxonomy of the query or the content retrieved based on the query with the content taxonomy of the ads. However, content taxonomy is commonly known not to correspond with advertisement taxonomy, which advertisers use to target at certain audience. As such, selecting ads based on content taxonomy does not serve to maximize the economic return of the ads to be inserted into content and recommended to users. The present teaching discloses method and system to build a linkage between content taxonomy and advertisement taxonomy so that ads that are not only relevant to a user's interests but also the interests of advertisers can be selected. In this way, the recommended content with ads to a user can both serve the user's interests and at the same time to allow the content operator to enhance monetization via ads.

Yet another aspect of personalized content recommendation of the present teaching relates to recommending probing content that is identified by extrapolating the currently known user interests. Traditional approaches rely on selecting either random content beyond the currently known user interests or content that has certain performance such as a high level of click activities. Random selection of probing content presents a low possibility to discover a user's unknown interests. Identifying probing content by choosing content for which a higher level of activities are observed is also problematic because there can be many pieces of content that a user may potentially be interested but there is a low level of activities associated therewith. The present teaching discloses ways to identify probing content by extrapolating the currently known interest with the flexibility of how far removed from the currently known interests. This approach also incorporates the mechanism to identify quality probing content so that there is an enhanced likelihood to discover a user's unknown interests. The focus of interests at any moment can be used as an anchor interest based on which probing interests (which are not known to be interests of the user) can be extrapolated from the anchor interests and probing content can be selected based on the probing interests and recommended to the user together with the content of the anchor interests. Probing interests/content may also be determined based on other considerations such as locale, time, or device type. In this way, the disclosed personalized content recommendation system can continually explore and discover unknown interests of a user to understand better the overall interests of the user in order to expand the scope of service.

Additional novel features will be set forth in part in the description which follows, and in part will become apparent to those skilled in the art upon examination of the following and the accompanying drawings or may be learned by production or operation of the examples. The advantages of the present teachings may be realized and attained by practice or use of various aspects of the methodologies, instrumentalities and combinations set forth in the detailed examples discussed below.

FIG. **1** depicts an exemplary system diagram **10** for personalized content recommendation to a user **105**, according to an embodiment of the present teaching. System **10** comprises a personalized content recommendation module **100**, which comprises numerous sub modules, content sources **110**, knowledge archives **115**, third party platforms **120**, and advertisers **125** with advertisement taxonomy **127** and advertisement database **126**. Content sources **110** may be any source of on-line content such as on-line news, published papers, blogs, on-line tabloids, magazines, audio content, image content, and video content. It may be content from content provider such as Yahoo! Finance, Yahoo! Sports, CNN, and ESPN. It may be multi-media content or text or any other form of content comprised of website content, social media content, such as Facebook, twitter, Reddit, etc, or any other content rich provider. It may be licensed content from providers such AP and Reuters. It may also be content crawled and indexed from various sources on the Internet. Content sources **110** provide a vast array of content to the personalized content recommendation module **100** of system **10**.

Knowledge archives **115** may be an on-line encyclopedia such as Wikipedia or indexing system such as an on-line dictionary. On-line concept archives **115** may be used for its content as well as its categorization or indexing systems. Knowledge archives **115** provide extensive classification system to assist with the classification of both the user's **105** preferences as well as classification of content. Knowledge concept archives, such as Wikipedia may have hundreds of thousands to millions of classifications and sub-classifications. A classification is used to show the hierarchy of the category. Classifications serve two main purposes. First they help the system understand how one category relates to another category and second, they help the system maneuver between higher levels on the hierarchy without having to move up and down the subcategories. The categories or classification structure found in knowledge archives **115** is used for multidimensional content vectors as well as multidimensional user profile vectors which are utilized by personalized content recommendation module **100** to match personalized content to a user **105**. Third party platforms **120** maybe any third party applications including but not limited to social networking sites like Facebook, Twitter, LinkedIn, Google+. It may include third party mail servers such as GMail or Bing Search. Third party platforms **120** provide both a source of content as well as insight into a user's personal preferences and behaviors.

Advertisers **125** are coupled with the ad content database **126** as well as an ads classification system or ad. taxonomy **127** intended for classified advertisement content. Advertisers **125** may provide streaming content, static content, and sponsored content. Advertising content may be placed at any location on a personalized content page and may be presented both as part of a content stream as well as a standalone advertisement, placed strategically around or within the content stream.

Personalized content recommendation module **100** comprises applications **130**, content pool **135**, content pool generation/update unit **140**, concept/content analyzer **145**, content crawler **150**, unknown interest explorer **215**, user understanding unit **155**, user profiles **160**, content taxonomy **165**, context information analyzer **170**, user event analyzer **175**, third party interest analyzer **190**, social media content source identifier **195**, advertisement insertion unit **200** and content/advertisement/taxonomy correlator **205**. These components are connected to achieve personalization, content pooling, and recommending personalized content to a user.

For example, the content ranking unit **210** works in connection with context information analyzer **170**, the unknown interest explorer **215**, and the ad insertion unit **200** to generate personalized content to be recommended to a user with personalized ads or probing content inserted. To achieve personalization, the user understanding unit **155** works in connection with a variety of components to dynamically and continuously update the user profiles **160**, including content taxonomy **165**, the knowledge archives **115**, user event analyzer **175**, and the third party interest analyzer **190**. Various components are connected to continuously maintain a content pool, including the content pool generation/update unit **140**, user event analyzer **175**, social media content source identifier **195**, content/concept analyzer **145**, content crawler **150**, the content taxonomy **165**, as well as user profiles **160**.

Personalized content recommendation module **100** is triggered when user **105** engages with system **10** through applications **130**. Applications **130** may receive information in the form of a user id, cookies, log in information from user **105** via some form of computing device. User **105** may access system **10** via a wired or wireless device and may be stationary or mobile. User **105** may interface with the applications **130** on a tablet, a Smartphone, a laptop, a desktop or any other computing device which may be embedded in devices such as watches, eyeglasses, or vehicles. In addition to receiving insights from the user **105** about what information the user **105** might be interested, applications **130** provides information to user **105** in the form of personalized content stream. User insights might be user search terms entered to the system, declared interests, user clicks on a particular article or subject, user dwell time or scroll over of particular content, user skips with respect to some content, etc. User insights may be a user indication of a like, a share, or a forward action on a social networking site, such as Facebook, or even peripheral activities such as print or scan of certain content. All of these user insights or events are utilized by the personalized content recommendation module **100** to locate and customize content to be presented to user **105**. User insights received via applications **130** are used to update personalized profiles for users which may be stored in user profiles **160**. User profiles **160** may be database or a series of databases used to store personalized user information on all the users of system **10**. User profiles **160** may be a flat or relational database and may be stored in one or more locations. Such user insights may also be used to determine how to dynamically update the content in the content pool **135**.

A specific user event received via applications **130** is passed along to user event analyzer **175**, which analyzes the user event information and feeds the analysis result with event data to the user understanding unit **155** and/or the content pool generation/update unit **140**. Based on such user event information, the user understanding unit **155** estimates short term interests of the user and/or infer user's long term interests based on behaviors exhibited by user **105** over long or repetitive periods. For example, a long term interest may be a general interest in sports, where as a short term interest may be related to a unique sports event, such as the Super Bowl at a particular time. Over time, a user's long term interest may be estimated by analyzing repeated user events. A user who, during every engagement with system **10**, regularly selects content related to the stock market may be considered as having a long term interest in finances. In this case, system **10** accordingly, may determine that personalized content for user **105** should contain content related to finance. Contrastingly, short term interest may be determined based on user events which may occur frequently over a short period, but which is not something the user **105** is interested in the long term. For example, a short term interest may reflect the momentary interest of a user which may be triggered by something the user saw in the content but such an interest may not persist over time. Both short and long term interest are important in terms of identifying content that meets the desire of the user **105**, but need to be managed separately because of the difference in their nature as well as how they influence the user.

In some embodiments, short term interests of a user may be analyzed to predict the user's long term interests. To retain a user, it is important to understand the user's persistent or long term interests. By identifying user **105**'s short term interest and providing him/her with a quality personalized experience, system **10** may convert an occasional user into a long term user. Additionally, short term interest may trend into long term interest and vice versa. The user understanding unit **155** provides the capability of estimating both short and long term interests.

The user understanding unit **155** gathers user information from multiple sources, including all the user's events, and creates one or more multidimensional personalization vectors. In some embodiments, the user understanding unit **155** receives inferred characteristics about the user **105** based on the user events, such as the content he/she views, self declared interests, attributes or characteristics, user activities, and/or events from third party platforms. In an embodiment, the user understanding unit **155** receives inputs from social media content source identifier **195**. Social media content source identifier **195** relies on user **105**'s social media content to personalize the user's profile. By analyzing the user's social media pages, likes, shares, etc, social media content source identifier **195** provides information for user understanding unit **155**. The social media content source identifier **195** is capable of recognizing new content sources by identifying, e.g., quality curators on social media platforms such as Twitter, Facebook, or blogs, and enables the personalized content recommendation module **100** to discover new content sources from where quality content can be added to the content pool **135**. The information generated by social media content source identifier **195** may be sent to a content/concept analyzer **145** and then mapped to specific category or classification based on content taxonomy **165** as well as a knowledge archives **115** classification system.

The third party interest analyzer **190** leverages information from other third party platforms about users active on such third party platforms, their interests, as well as content these third party users to enhance the performance of the user understanding unit **155**. For example, when information about a large user population can be accessed from one or more third party platforms, the user understanding unit **155** can rely on data about a large population to establish a baseline interest profile to make the estimation of the interests of individual users more precise and reliable, e.g., by comparing interest data with respect to a particular user with the baseline interest profile which will capture the user's interests with a high level of certainty.

When new content is identified from content source **110** or third party platforms **120**, it is processed and its concepts are analyzed. The concepts can be mapped to one or more categories in the content taxonomy **165** and the knowledge archives **115**. The content taxonomy **165** is an organized structure of concepts or categories of concepts and it may contain a few hundred classifications of a few thousand. The knowledge archives **115** may provide millions of concepts, which may or may not be structures in a similar manner as the content taxonomy **165**. Such content taxonomy and knowledge archives may serve as a universal interest space. Concepts estimated from the content can be mapped to a universal

interest space and a high dimensional vector can be constructed for each piece of content and used to characterize the content. Similarly, for each user, a personal interest profile may also be constructed, mapping the user's interests, characterized as concepts, to the universal interest space so that a high dimensional vector can be constructed with the user's interests levels populated in the vector.

Content pool **135** may be a general content pool with content to be used to serve all users. The content pool **135** may also be structured so that it may have personalized content pool for each user. In this case, content in the content pool is generated and retained with respect to each individual user. The content pool may also be organized as a tiered system with both the general content pool and personalized individual content pools for different users. For example, in each content pool for a user, the content itself may not be physically present but is operational via links, pointers, or indices which provide references to where the actual content is stored in the general content pool.

Content pool **135** is dynamically updated by content pool generation/update module **140**. Content in the content pool comes and go and decisions are made based on the dynamic information of the users, the content itself, as well as other types of information. For example, when the performance of content deteriorates, e.g., low level of interests exhibited from users, the content pool generation/update unit **140** may decide to purge it from the content pool. When content becomes stale or outdated, it may also be removed from the content pool. When there is a newly detected interest from a user, the content pool generation/update unit **140** may fetch new content aligning with the newly discovered interests. User events may be an important source of making observations as to content performance and user interest dynamics. User activities are analyzed by the user event analyzer **175** and such Information is sent to the content pool generation/update unit **140**. When fetching new content, the content pool generation/update unit **140** invokes the content crawler **150** to gather new content, which is then analyzed by the content/concept analyzer **145**, then evaluated by the content pool generation/update unit **140** as to its quality and performance before it is decided whether it will be included in the content pool or not. Content may be removed from content pool **135** because it is no longer relevant, because other users are not considering it to be of high quality or because it is no longer timely. As content is constantly changing and updating content pool **135** is constantly changing and updating providing user **105** with a potential source for high quality, timely personalized content.

In addition to content, personalized content recommendation module **100** provides for targeted or personalized advertisement content from advertisers **125**. Advertisement database **126** houses advertising content to be inserted into a user's content stream. Advertising content from ad database **126** is inserted into the content stream via Content ranking unit **210**. The personalized selection of advertising content can be based on the user's profile. Content/advertisement/user taxonomy correlator **205** may re-project or map a separate advertisement taxonomy **127** to the taxonomy associated with the user profiles **160**. Content/advertisement/user taxonomy correlator **205** may apply a straight mapping or may apply some intelligent algorithm to the re-projection to determine which of the users may have a similar or related interest based on similar or overlapping taxonomy categories.

Content ranking unit **210** generates the content stream to be recommended to user **105** based on content, selected from content pool **135** based on the user's profile, as well as advertisement, selected by the advertisement insertion unit **200**. The content to be recommended to the user **105** may also be determined, by the content ranking unit **210**, based on information from the context information analyzer **170**. For example, if a user is currently located in a beach town which differs from the zip code in the user's profile, it can be inferred that the user may be on vacation. In this case, information related to the locale where the user is currently in may be forwarded from the context information analyzer to the Content ranking unit **210** so that it can select content that not only fit the user's interests but also is customized to the locale. Other context information include day, time, and device type. The context information can also include an event detected on the device that the user is currently using such as a browsing event of a website devoted to fishing. Based on such a detected event, the momentary interest of the user may be estimated by the context information analyzer **170**, which may then direct the Content ranking unit **210** to gather content related to fishing amenities in the locale the user is in for recommendation.

The personalized content recommendation module **100** can also be configured to allow probing content to be included in the content to be recommended to the user **105**, even though the probing content does not represent subject matter that matches the current known interests of the user. Such probing content is selected by the unknown interest explorer **215**. Once the probing content is incorporated in the content to be recommended to the user, information related to user activities directed to the probing content (including no action) is collected and analyzed by the user event analyzer **175**, which subsequently forwards the analysis result to long/short term interest identifiers **180** and **185**. If an analysis of user activities directed to the probing content reveals that the user is or is not interested in the probing content, the user understanding unit **155** may then update the user profile associated with the probed user accordingly. This is how unknown interests may be discovered. In some embodiments, the probing content is generated based on the current focus of user interest (e.g., short term) by extrapolating the current focus of interests. In some embodiments, the probing content can be identified via a random selection from the general content, either from the content pool **135** or from the content sources **110**, so that an additional probing can be performed to discover unknown interests.

To identify personalized content for recommendation to a user, the content ranking unit **210** takes all these inputs and identify content based on a comparison between the user profile vector and the content vector in a multiphase ranking approach. The selection may also be filtered using context information. Advertisement to be inserted as well as possibly probing content can then be merged with the selected personalized content.

FIG. **2** is a flowchart of an exemplary process for personalized content recommendation, according to an embodiment of the present teaching. Content taxonomy is generated at **205**. Content is accessed from different content sources and analyzed and classified into different categories, which can be pre-defined. Each category is given some labels and then different categories are organized into some structure, e.g., a hierarchical structure. A content pool is generated at **210**. Different criteria may be applied when the content pool is created. Examples of such criteria include topics covered by the content in the content pool, the performance of the content in the content pool, etc. Sources from which content can be obtained to populate the content pool include content sources **110** or third party platforms **120** such as Facebook, Twitter, blogs, etc. FIG. **3** provides a more detailed exemplary flowchart related to content pool creation, according to an

embodiment of the present teaching. User profiles are generated at **215** based on, e.g., user information, user activities, identified short/long term interests of the user, etc. The user profiles may be generated with respect to a baseline population interest profile, established based on, e.g., information about third party interest, knowledge archives, and content taxonomies.

Once the user profiles and the content pool are created, when the system **10** detects the presence of a user, at **220**, the context information, such as locale, day, time, may be obtained and analyzed, at **225**. FIG. **4** illustrates exemplary types of context information. Based on the detected user's profile, optionally context information, personalized content is identified for recommendation. A high level exemplary flow for generating personalized content for recommendation is presented in FIG. **5**. Such gathered personalized content may be ranked and filtered to achieve a reasonable size as to the amount of content for recommendation. Optionally (not shown), advertisement as well as probing content may also be incorporated in the personalized content. Such content is then recommended to the user at **230**.

User reactions or activities with respect to the recommended content are monitored, at **235**, and analyzed at **240**. Such events or activities include clicks, skips, dwell time measured, scroll location and speed, position, time, sharing, forwarding, hovering, motions such as shaking, etc. It is understood that any other events or activities may be monitored and analyzed. For example, when the user moves the mouse cursor over the content, the title or summary of the content may be highlighted or slightly expanded. In anther example, when a user interacts with a touch screen by her/his finger[s], any known touch screen user gestures may be detected. In still another example, eye tracking on the user device may be another user activity that is pertinent to user behaviors and can be detected. The analysis of such user events includes assessment of long term interests of the user and how such exhibited short term interests may influence the system's understanding of the user's long term interests. Information related to such assessment is then forwarded to the user understanding unit **155** to guide how to update, at **255**, the user's profile. At the same time, based on the user's activities, the portion of the recommended content that the user showed interests are assessed, at **245**, and the result of the assessment is then used to update, at **250**, the content pool. For example, if the user shows interests on the probing content recommended, it may be appropriate to update the content pool to ensure that content related to the newly discovered interest of the user will be included in the content pool.

FIG. **3** illustrates different types of context information that may be detected and utilized in assisting to personalize content to be recommended to a user. In this illustration, context information may include several categories of data, including, but not limited to, time, space, platform, and network conditions. Time related information can be time of the year (e.g., a particular month from which season can be inferred), day of a week, specific time of the day, etc. Such information may provide insights as to what particular set of interests associated with a user may be more relevant. To infer the particular interests of a user at a specific moment may also depend on the locale that the user is in and this can be reflected in the space related context information, such as which country, what locale (e.g., tourist town), which facility the user is in (e.g., at a grocery store), or even the spot the user is standing at the moment (e.g., the user may be standing in an aisle of a grocery store where cereal is on display). Other types of context information includes the specific platform related to the user's device, e.g., Smartphone, Tablet, laptop, desktop, bandwidth/data rate allowed on the user's device, which will impact what types of content may be effectively presented to the user. In addition, the network related information such as state of the network where the user's device is connected to, the available bandwidth under that condition, etc. may also impact what content should be recommended to the user so that the user can receive or view the recommended content with reasonable quality.

FIG. **4** depicts an exemplary system diagram of the content pool generation/update unit **140**, according to an embodiment of the present teaching. The content pool **135** can be initially generated and then maintained according to the dynamics of the users, contents, and needs detected. In this illustration, the content pool generation/update unit **140** comprises a content/concept analyzing control unit **410**, a content performance estimator **420**, a content quality evaluation unit **430**, a content selection unit **480**, which will select appropriate content to place into the content pool **135**. In addition, to control how content is to be updated, the content pool generation/update unit **140** also includes a user activity analyzer **440**, a content status evaluation unit **450**, and a content update control unit **490**.

The content/concept analyzing control unit **410** interfaces with the content crawler **150** (FIG. **1**) to obtain candidate content that is to be analyzed to determine whether the new content is to be added to the content pool. The content/concept analyzing control unit **410** also interfaces with the content/concept analyzer **145** (see FIG. **1**) to get the content analyzed to extract concepts or subjects covered by the content. Based on the analysis of the new content, a high dimensional vector for the content profile can be computed via, e.g., by mapping the concepts extracted from the content to the universal interest space, e.g., defined via Wikipedia or other content taxonomies. Such a content profile vector can be compared with user profiles **160** to determine whether the content is of interest to users. In addition, content is also evaluated in terms of its performance by the content performance estimator **420** based on, e.g., third party information such as activities of users from third party platforms so that the new content, although not yet acted upon by users of the system, can be assessed as to its performance. The content performance information may be stored, together with the content's high dimensional vector related to the subject of the content, in the content profile **470**. The performance assessment is also sent to the content quality evaluation unit **430**, which, e.g., will rank the content in a manner consistent with other pieces of content in the content pool. Based on such rankings, the content selection unit **480** then determines whether the new content is to be incorporated into the content pool **135**.

To dynamically update the content pool **135**, the content pool generation/update unit **140** may keep a content log **460** with respect to all content presently in the content pool and dynamically update the log when more information related to the performance of the content is received. When the user activity analyzer **440** receives information related to user events, it may log such events in the content log **460** and perform analysis to estimate, e.g., any change to the performance or popularity of the relevant content over time. The result from the user activity analyzer **440** may also be utilized to update the content profiles, e.g., when there is a change in performance. The content status evaluation unit **450** monitors the content log and the content profile **470** to dynamically determine how each piece of content in the content pool **135** is to be updated. Depending on the status with respect to a piece of content, the content status evaluation unit **450** may decide to purge the content if its performance degrades below

a certain level. It may also decide to purge a piece of content when the overall interest level of users of the system drops below a certain level. For content that requires update, e.g., news or journals, the content status evaluation unit **450** may also control the frequency **455** of the updates based on the dynamic information it receives. The content update control unit **490** carries out the update jobs based on decisions from the content status evaluation unit **450** and the frequency at which certain content needs to be updated. The content update control unit **490** may also determine to add new content whenever there is peripheral information indicating the needs, e.g., there is an explosive event and the content in the content pool on that subject matter is not adequate. In this case, the content update control unit **490** analyzes the peripheral information and if new content is needed, it then sends a control signal to the content/concept analyzing control unit **410** so that it can interface with the content crawler **150** to obtain new content.

FIG. **5** is a flowchart of an exemplary process of creating the content pool, according to an embodiment of the present teaching. Content is accessed at **510** from content sources, which include content from content portals such as Yahoo!, general Internet sources such as web sites or FTP sites, social media platforms such as Twitter, or other third party platforms such as Facebook. Such accessed content is evaluated, at **520**, as to various considerations such as performance, subject matters covered by the content, and how it fit users' interests. Based on such evaluation, certain content is selected to generate, at **530**, the content pool **135**, which can be for the general population of the system or can also be further structured to create sub content pools, each of which may be designated to a particular user according to the user's particular interests. At **540**, it is determined whether user-specific content pools are to be created. If not, the general content pool **135** is organized (e.g., indexed or categorized) at **580**. If individual content pools for individual users are to be created, user profiles are obtained at **550**, and with respect to each user profile, a set of personalized content is selected at **560** that is then used to create a sub content pool for each such user at **570**. The overall content pool and the sub content pools are then organized at **580**.

FIG. **6** is a flowchart of an exemplary process for updating the content pool **135**, according to an embodiment of the present teaching. Dynamic information is received at **610** and such information includes user activities, peripheral information, user related information, etc. Based on the received dynamic information, the content log is updated at **620** and the dynamic information is analyzed at **630**. Based on the analysis of the received dynamic information, it is evaluated, at **640**, with respect to the content implicated by the dynamic information, as to the change of status of the content. For example, if received information is related to user activities directed to specific content pieces, the performance of the content piece may need to be updated to generate a new status of the content piece. It is then determined, at **650**, whether an update is needed. For instance, if the dynamic information from a peripheral source indicates that content of certain topic may have a high demand in the near future, it may be determined that new content on that topic may be fetched and added to the content pool. In this case, at **660**, content that needs to be added is determined. In addition, if the performance or popularity of a content piece has just dropped below an acceptable level, the content piece may need to be purged from the content pool **135**. Content to be purged is selected at **670**. Furthermore, when update is needed for regularly refreshed content such as journal or news, the schedule according to which update is made may also be changed if the dynamic information received indicates so. This is achieved at **680**.

FIG. **7** depicts an exemplary diagram of the user understanding unit **155**, according to an embodiment of the present teaching. In this exemplary construct, the user understanding unit **155** comprises a baseline interest profile generator **710**, a user profile generator **720**, a user intent/interest estimator **740**, a short term interest identifier **750** and a long term interest identifier **760**. In operation, the user understanding unit **155** takes various input and generates user profiles **160** as output. Its input includes third party data such as users' information from such third party platforms as well as content such users accessed and expressed interests, concepts covered in such third party data, concepts from the universal interest space (e.g., Wikipedia or content taxonomy), information about users for whom the personalized profiles are to be constructed, as well as information related to the activities of such users. Information from a user for whom a personalized profile is to be generated and updated includes demographics of the user, declared interests of the user, etc. Information related to user events includes the time, day, location at which a user conducted certain activities such as clicking on a content piece, long dwell time on a content piece, forwarding a content piece to a friend, etc.

In operation, the baseline interest profile generator **710** access information about a large user population including users' interests and content they are interested in from one or more third party sources (e.g., Facebook). Content from such sources is analyzed by the content/concept analyzer **145** (FIG. **1**), which identifies the concepts from such content. When such concepts are received by the baseline interest profile generator **710**, it maps such concepts to the knowledge archives **115** and content taxonomy **165** (FIG. **1**) and generate one or more high dimensional vectors which represent the baseline interest profile of the user population. Such generated baseline interest profile is stored at **730** in the user understanding unit **155**. When there is similar data from additional third party sources, the baseline interest profile **730** may be dynamically updated to reflect the baseline interest level of the growing population.

Once the baseline interest profile is established, when the user profile generator receives user information or information related to estimated short term and long term interests of the same user, it may then map the user's interests to the concepts defined by, e.g., the knowledge archives or content taxonomy, so that the user's interests are now mapped to the same space as the space in which the baseline interest profile is constructed. The user profile generator **720** then compares the user's interest level with respect to each concept with that of a larger user population represented by the baseline interest profile **730** to determine the level of interest of the user with respect to each concept in the universal interest space. This yields a high dimensional vector for each user. In combination with other additional information, such as user demographics, etc., a user profile can be generated and stored in **160**.

User profiles **160** are updated continuously based on newly received dynamic information. For example, a user may declare additional interests and such information, when received by the user profile generator **720**, may be used to update the corresponding user profile. In addition, the user may be active in different applications and such activities may be observed and information related to them may be gathered to determine how they impact the existing user profile and when needed, the user profile can be updated based on such new information. For instance, events related to each user

may be collected and received by the user intent/interest estimator **740**. Such events include that the user dwelled on some content of certain topic frequently, that the user recently went to a beach town for surfing competition, or that the user recently participated in discussions on gun control, etc. Such information can be analyzed to infer the user intent/interests. When the user activities relate to reaction to content when the user is online, such information may be used by the short term interest identifier **750** to determine the user's short term interests. Similarly, some information may be relevant to the user's long term interests. For example, the number of requests from the user to search for content related to diet information may provide the basis to infer that the user is interested in content related to diet. In some situations, estimating long term interest may be done by observing the frequency and regularity at which the user accesses certain type of information. For instance, if the user repeatedly and regularly accesses content related to certain topic, e.g., stocks, such repetitive and regular activities of the user may be used to infer his/her long term interests. The short term interest identifier **750** may work in connection with the long term interest identifier **760** to use observed short term interests to infer long term interests. Such estimated short/long term interests are also sent to the user profile generator **720** so that the personalization can be adapted to the changing dynamics.

FIG. **8** is a flowchart of an exemplary process for generating a baseline interest profile based on information related to a large user population, according to an embodiment of the present teaching. The third party information, including both user interest information as well as their interested content, is accessed at **810** and **820**. The content related to the third party user interests is analyzed at **830** and the concepts from such content are mapped, at **840** and **850**, to knowledge archives and/or content taxonomy. To build a baseline interest profile, the mapped vectors for third party users are then summarized to generate a baseline interest profile for the population. There can be a variety ways to summarize the vectors to generate an averaged interest profile with respect to the underlying population.

FIG. **9** is a flowchart of an exemplary process for generating/updating a user profile, according to an embodiment of the present teaching. User information is received first at **910**. Such user information includes user demographics, user declared interests, etc. Information related to user activities is also received at **920**. Content pieces that are known to be interested by the user are accessed at **930**, which are then analyzed, at **950**, to extract concepts covered by the content pieces. The extracted concepts are then mapped, at **960**, to the universal interest space and compared with, concept by concept, the baseline interest profile to determine, at **970**, the specific level of interest of the user given the population. In addition, the level of interests of each user may also be identified based on known or estimated short and long term interests that are estimated, at **940** and **945**, respectively, based on user activities or content known to be interested by the user. A personalized user profile can then be generated, at **980**, based on the interest level with respect to each concept in the universal interest space.

FIG. **10** depicts an exemplary system diagram for the content ranking unit **210**, according to an embodiment of the present teaching. The content ranking unit **210** takes variety of input and generates personalized content to be recommended to a user. The input to the content ranking unit **210** includes user information from the applications **130** with which a user is interfacing, user profiles **160**, context information surrounding the user at the time, content from the content pool **135**, advertisement selected by the ad insertion unit **200**, and optionally probing content from the unknown interest explorer **215**. The content ranking unit **210** comprises a candidate content retriever **1010** and a multi-phase content ranking unit **1020**. Based on user information from applications **130** and the relevant user profile, the candidate content retriever **1010** determines the content pieces to be retrieved from the content pool **135**. Such candidate content may be determined in a manner that is consistent with the user's interests or individualized. In general, there may be a large set of candidate content and it needs to be further determined which content pieces in this set are most appropriate given the context information. The multi-phase content ranking unit **1020** takes the candidate content from the candidate content retriever **1010**, the advertisement, and optionally may be the probing content, as a pool of content for recommendation and then performs multiple stages of ranking, e.g., relevance based ranking, performance based ranking, etc. as well as factors related to the context surrounding this recommendation process, and selects a subset of the content to be presented as the personalized content to be recommended to the user.

FIG. **11** is a flowchart of an exemplary process for the content ranking unit, according to an embodiment of the present teaching. User related information and user profile are received first at **1110**. Based on the received information, user's interests are determined at **1120**, which can then be used to retrieve, at **1150**, candidate content from the content pool **135**. The user's interests may also be utilized in retrieving advertisement and/or probing content at **1140** and **1130**, respectively. Such retrieved content is to be further ranked, at **1160**, in order to select a subset as the most appropriate for the user. As discussed above, the selection takes place in a multi-phase ranking process, each of the phases is directed to some or a combination of ranking criteria to yield a subset of content that is not only relevant to the user as to interests but also high quality content that likely will be interested by the user. The selected subset of content may also be further filtered, at **1170**, based on, e.g., context information. For example, even though a user is in general interested in content about politics and art, if the user is currently in Milan, Italy, it is likely that the user is on vacation. In this context, rather than choosing content related to politics, the content related to art museums in Milan may be more relevant. The multi-phase content ranking unit **1020** in this case may filter out the content related to politics based on this contextual information. This yields a final set of personalized content for the user. At **1180**, based on the contextual information associated with the surrounding of the user (e.g., device used, network bandwidth, etc.), the content ranking unit packages the selected personalized content, at **1180**, in accordance with the context information and then transmits, at **1190**, the personalized content to the user.

More detailed disclosures of various aspects of the system **10**, particularly the personalized content recommendation module **100**, are covered in different U.S. patent applications as well as PCT applications, entitled "Method and System For User Profiling Via Mapping Third Party Interests To A Universal Interest Space", "Method and System for Multi-Phase Ranking For Content Personalization", "Method and System for Measuring User Engagement Using Click/Skip In Content Stream", "Method and System for Dynamic Discovery And Adaptive Crawling of Content From the Internet", "Method and System For Dynamic Discovery of Interesting URLs From Social Media Data Stream", "Method and System for Discovery of User Unknown Interests", "Method and System for Efficient Matching of User Profiles with Audience

Segments", "Method and System For Mapping Short Term Ranking Optimization Objective to Long Term Engagement", "Social Media Based Content Selection System", "Method and System For Measuring User Engagement From Stream Depth", "Method and System For Measuring User Engagement Using Scroll Dwell Time", "Almost Online Large Scale Collaborative Based Recommendation System", and "Efficient and Fault-Tolerant Distributed Algorithm for Learning Latent Factor Models through Matrix Factorization". The present teaching is particularly directed to systems and methods for identifying personalized user interests from unknown interests. Specifically, the present disclosure relates to identifying user interests in content beyond the currently known user interests by inserting probe content into the personalized user stream.

Recommendation systems strive to present items that are highly personalized for a user. As a result the user interaction will be more and more limited to the list of interests that the recommendation system currently known for the user. In the long term this can lead to a personalization filter bubble where the user is recommended only items that represent a very narrow subset of the user interests. This bubble or bottleneck may be alleviated by presenting random items from the corpus of items every so often in order to discover new interests for the user, however such an approach is very haphazard.

Personalized content or recommendation systems have always strived to find a balance between exploiting the current known information about a user to present an optimal list versus exploring the space of possible unknown interests by presenting a sub-optimal list of content to a user and monitor the reaction. In systems where the corpus of articles is very large and the set of interests is also very large then a random exploration is very in-efficient at discovering new positive interests for a user. Many articles with interests of little or negative value will be presented to the user before an article with interest of positive value will be discovered.

In systems using collaborative filtering for example a list of recommended content may be a mixture of both strategies, i.e., content based on user preferences and random content, but the balance of exploration and exploitation is un-controlled. These filtering systems may work well if a large number user interactions can be represented by a relatively small latent subspace, however, such systems do not allow for fine control between exploration and exploitation. Some systems may use a multi-arm bandit or Thomspon sampling approach, which simultaneously attempt to acquire new knowledge and to optimize its decisions based on existing knowledge where the amount of exploration versus exploitation can be more carefully controlled. Multi-arm bandit and Thompson sampling however, are inefficient given that most articles will have few if any user interactions.

Accordingly, a need exists where a user's profile over a space of interests is created and generates distance metrics over that space so that they may be used in intelligently selecting the items used for exploration. The distance measured can be included on top of a user's actions in order to balance exploration with exploitation. Further, a need exists for a method and system to explore the list of user interests beyond the current known list by defining distance metrics in the interest space and by carefully leveraging observed user interactions to intelligently select likely content the user may be interested in. The present disclosure targets for exploration items with interests which are nearby the current set of user interests, such targeted interests greatly improve the chance that one of the exploration items will be liked by the user.

FIG. **12** is a diagram illustrating portion of a content personalization system **10**, as shown in FIG. **1** including an unknown interest explorer **215**. The other relevant portions of the content personalization system **10** in the embodiment includes applications **130**, user event analyzer **175**, user understanding unit **155**, knowledge archives **115**, content taxonomy **165**, user profiles **160**, content pool **135**, content ranking unit **210**, context information analyzer **170**, and content sources **110**. Unknown interest explorer **215** identifies probing content obtained from content pool **135** or from content sources **110** that would not otherwise be identified by the content ranking unit **210** based on information related to a user including the user profile **160**. Unknown interest explorer **215** feeds the probing content into content ranking unit **210** for recommendation to the user **105** via applications **130**. User **105** may select to view the content or not, but if user **105** does view the content, the user event analyzer **175** will analyze the user's behavior with respect to the probing content and attempt to determine whether the user's activity reflects any interest of the user on the subject matter represented by the probing content.

Such detected user activities directed to the probing content are sent from the user event analyzer **175** to the user understanding unit **155**, which may collect information related to the probing content and correlate with the user activities directed to the probing content to determine whether the user is interested in the concept or subject matter present in the probing content. If new user interest is discovered through the analysis, the user understanding unit **155** will update the user profile in **160** so that the newly discovered interest can be reflected in the user profile. In this way, the personalized content recommendation module **100** can continuously discover users' unknown interests in order to enhance the understanding of users' overall interests.

FIG. **13** depicts high dimensional vector **1300** of user's interest stored in user profiles **160**. High dimensional vector **1300** is built based on knowledge archives **115** and/or a content taxonomy **165**. Each entry in the vector **1301***a*, **1301***b* . . . **1301***n* maps to a concept in the knowledge archives or to a class in the content taxonomy **165** and the score recorded in each entry of this vector represents a level of estimated user interest in this particular concept. The vector may be built based on both the concepts in the knowledge archives and taxonomy. Multiple vectors may also be built, each of which corresponds to one source (e.g., one is to Wikipedia and the other is to a content taxonomy). In general, the knowledge archives and content taxonomy provide a wide range of coverage in terms of interests and forms a universal interest space.

FIG. **14** is an exemplary structure of content taxonomy **165**. First level entries **1400** represent first level categories, which are intended to be high level topics or subjects (i.e., politics, sports, entertainment, etc). Second level entries **1410** are subcategories of first level entries **1400** (politics→election & voting rights: Sports→football & basketball). Third level entries **1420** are sub categories of subcategories, i.e., subcategories of level 2, These may be further refinements (Entertainment→Movies→comedy & drama & romance). A user may be interested in the first level category or the third level category, but one does not necessarily imply the pother. For example a user who is interested in elections may not be interested in politics as a broad concept, and the user's vector in high dimensional vector **1300** would be weighted accordingly. However, closer relationships between category levels may be some indication of possible interesting or unknown categories of content that the user may be interested in.

FIG. **15** depicts an exemplary structure of knowledge archives **115** such as wikipedia. Although the knowledge

archive **115** may include similar content as in content taxonomy **165**, it may be organized in a flat structure in one dimensional space without sub-categories. For example, politics voting right and election are all categories but are not related as first level and second level. High dimensional vector **1300** may be built from the categories **1500** found in the knowledge archive as well. Generating a high dimensional vector **1300** from either or both concept taxonomy **165** and knowledge archive structure **115** will result in a vector representing user interest where each entry or interest is weighted based on past user behaviors.

FIG. **16** depicts a high dimensional vector **1600** built for a user **105** where there are certain estimated/identified user interests in particular subjects mapped to the content taxonomy **165**. High dimensional vector **1600** may contain identified interests **1605** and **1610** which have a high score (represented as solid black) indicating a strong user interests. Entries corresponding to **1615** and **1620** may indicate it is not known at this point whether the user is interested in the corresponding concepts. User interest **1605** for example corresponds to third level category jazz **1411** and interest **1610** corresponds to a first level interest election **1406**. Both of these weighted interest **1605** and **1610** indicate a user's interests in the topics for which personalized content would be collected from the content pool **135** and present to user **105** after going through content ranking unit **210** which utilizes the high dimensional vectors **1600** in the user profile **160** and the content vector to rank the content for personalization.

FIG. **16***a* depicts an exemplary scheme to identify currently unknown interests of a user in order to generate probing content. In this example, some known interests of the user may be identified from the high dimensional vector **1600** associated with the user. Such known interests have been mapped to a content taxonomy. Unknown interest of the same user can be identified, in accordance with the present disclosure, by extrapolating the user's current known interests based on content taxonomy tree. For example, in an embodiment, the system may explore the taxonomy tree to identify supplemental interest by traversing a taxonomy tree within a certain distance from the each node in the taxonomy where the user's known interest is mapped to. For example, in FIG. **16***a*, the user's interests are mapped to topics "election" **1406** and "Jazz" **1411**. From these two nodes, nearby topics such as "Politics" **1401** or "Sports" **1402** may be identified by traverse the taxonomy tree. In this way, user's unknown interests Politics and Sports can be extrapolated from the user's known interests. Based on such identified unknown interest, content related to such topics can be identified as probing content and recommended to the user to test whether it is a subject of interest of not.

In searching for unknown interests, there may be some limitations such as a distance may be provided to limit the scope of the search. The content taxonomy can be a very big tree and when the distance is set small, only nearby similar interests/topics can be explored. If the distance limitation is set large, the unknown interests that are allowed to be explored can be quite different from the user's current known interests. The actual distance between the user's known interest and an unknown interest to be explored may be measured in different ways. For example, each hop along the content taxonomy tree may be defined as a unit of distance. The number of hops between a known interest and the identified unknown interest may readily lead to a calculation of the actual distance between the two. When the limitation set via a distance is infinity, any unknown interests can be used to explore user's interests. There may be other limitations put in place to limit how to identify unknown interests. For example, the manner by which the taxonomy tree is traversed may be limited to going only certain directions, e.g., going up first before going horizontal, etc.

In the example illustrated in FIG. **16***a*, the distance between "election" and "politics" can be one (one hop) while the distance between "Jazz" and "Sports" may be five (2 hops up and horizontal hop may be counted as greater than 3). This can be viewed as interest relatedness distance metric, which is a valuation of the user's known interests and the potential to find the unknown interest to be the interest of the user. The unknown interest explorer may "walk through" the taxonomy based on the interest relatedness distance metric to identify currently unknown interest.

Unknown interest explorer **215** may have preset limitations as to how far the exploration can go. For example, the threshold could be set to 10 to allow for very unrelated topics to be used to probe a user or contrastingly it could be set to 3 to keep topics more closely related. Furthermore, unknown interest explorer **215** may occasionally randomly set the distance threshold to allow random topics to be injected in the hopes of identifying a completely unrelated unknown interest.

In an embodiment, other distances metrics may be used to identify unknown interests as well. Examples of such distances metrics include, but are not limited to: the co-occurrence of two interests in a corpus of articles, the co-occurrence of two interests in a large set of user profiles, and the co-occurrence of two interests in a large set of user sessions.

For the co-occurrence of two interests in a corpus of articles, the distance metric can be computed as follows:

For each pair of interests (labeled as X and Y), the system may compute a contingency table,

TABLE 01

	Y = 1	Y = 0
X = 1	η_{11}	η_{10}
X = 0	η_{01}	η_{00}

Where X=1 denotes when an interest is present in the article and X=0 denote s when an interest is not present in the article. Similarly for Y=1 and Y=0, the number count η_{10} represent the number of articles where X=1 and Y=0. Similarly for η_{11}, η_{01} and η_{00}. Once the matrix is compiled, a distance metric can be defined as the log odd ratio of $1/(1+(\eta_{11}*\eta)/(\eta_{01}*\eta_{01}))$ where $\eta=\eta_{00}+\eta_{01}+\eta_{10}+\eta_{11}$.

In another embodiment, a similarity co-occurrence can also be computed from looking at the interest profiles of a large set of users. For each pair of interests (X and Y), the system can compute a contingency table as before, except that η_{10} now represents represent the number of users having interest X (X=1) in his/her profile and not having Y (Y=0) in his/her profile at the same time. Similarly, η_{11}, η_{01} and η_{00} may be computed. Once all four are computed, the log-odd ratio is computed as in the distance metric.

In another embodiment, a similar co-occurrence may be computed by looking at the interests of a large set of user sessions. For each pair of interests (X and Y), one may compute a contingency table as before, except that η_{10} now represents the number of user sessions having interest X (X=1) present in the session and not having Y (Y=0) in the same session. In an embodiment, the session can be defined as a series of interactions of the user with the application. Sessions are delimited by long period s of inactivity (e.g. 30 minutes or more). The presence or absence of an interest in a user is computed by looking at the interests of the articles clicked by the user during the session.

Similarly values for η_{11}, η_{01} and η_{00} are computed. As with other embodiments, a log-odd ratio is computed as the distance metric.

Regardless of the computation method used, once multiple distance metrics are defined and the contingency table computed—they can be combined to produce a better distance metric.

In an embodiment, a plurality of distance metrics can be combined together to create a more predictive distance metric. The predictive power of a distance metric can be determined by looking at the number of supplemental contents that is clicked by the user in the application.

FIG. **17** illustrates an embodiment of the unknown interest explorer **215**. In the this embodiment, unknown interest explorer **215** receives inputs from user profile **160**, content taxonomy **165**, content sources **110**, content pool **135** and unknown interest search parameters **1750** to generate probing content which is sent to the content ranking unit **210**.

Unknown interest explorer **215** comprises known interest identifier **1705**, content crawler **150**, supplemental interest identifier **1715**, supplemental content identifier **1720**, supplemental interest pool **1725**, supplemental content pool **1730**, random content selector **1735**, local based content filter **1740** and supplemental content selector **1745**. Known interest identifier **1705** receives the high dimensional vector **1600** of a user's interest from user profiles **160** and identifies the known interests of the user **105**. Those interests are passed to the supplemental interests identifier **1715** which receives the unknown interest search parameters **1750** which will be the distance parameters on the content taxonomy tree, for example, from which supplemental interests will be identified. These may be simple numbers i.e., 1-5 or may be randomly generated numbers that fall below a max distance threshold. They may also be computed based on some other user indicators as described above. Using the input of content taxonomy **165**, a set of supplemental interests is identified with respect to each of one or more known interest and such supplemental interests are identified within the search parameters **1750**. Each of the identified supplemental interest can be weighed. For example, each unknown interest or supplemental interest can be weighed based on its distance from the known interest based on which the unknown interest is found.

One intuitive way to weigh a supplemental interest is to take the inverse of the distance, i.e., the short the distance between the known interest and the unknown interest, the higher weight is it and the longer the distance, the smaller weight is assigned. For example, a supplemental interest that has a distance **1** from a known interest will be weighed higher then a supplemental interest that has a distance **5** from a known interest. Once the supplemental interests are identified, they are passed along to the supplemental interests pool **1725** along with their weights. Supplemental content identifier **1720** may retrieve that information and gather content related to the supplemental interests identified by invoking content crawler **150** to fetch related content. The sources of the supplemental content may be the content pool or may be other general internet sources.

The supplemental content that is identified may be ranked based on a score such as an affinity score which measures the affinity or match between a supplemental or unknown interest and the content. The more related the content is to the supplemental interest, the higher the affinity score. Each piece of supplemental content may then be weighed with the affinity score or the weigh associated with the supplemental interest or both. The supplemental content may then be placed in supplemental content pool **1730** for introduction to the user **105**.

Additionally and/or alternatively, random content may be selected by random content selector **1735** from content pool **135** and added to the supplemental content pool for random presentment too user **105** with the hopes of identifying unknown interests. Supplemental content pool **1730** may rank the supplemental content based on the affinity/weighting and/or confidence score so that the supplemental content with the highest ranking will be presented in a higher priority to user **105**.

Supplemental content in content pool **1730** may also be filtered by locale based content filter **1740** for example or other criteria filters such as age, gender, etc., by removing unrelated content, i.e., geographically based content which may be of no interest to user **105** based on current demographics. The ranked supplemental content from content pool **1730** pre and post locale filtering will then be selected by supplemental content selector **1745** based on the ranking as probing content to be added to the content ranking unit **210** for presentment to the user **105** via application **130**.

FIG. **18** is a diagram of the flow of information performed by unknown interest explorer **215**. At step **1800** the user's interests are identified in the known interest identifier **1705** from the information stored in the user profile **160**. At step **1805** supplemental interests are identified by the supplemental interest identifier **1715**. Once user's interests are identified from the high dimensional vectors, the supplemental interest search parameters **1750** are received by the unknown interest explorer **215** and are used to identify a range of supplemental interests. At step **1815**, the supplemental interests identified in step **1810** by the supplemental interest identifier **1715**, are used to identify supplemental content utilizing the supplemental content identifier **1720** which receives content directly from the content pool **135** and content sources **110**. At step **1820**, an affinity score is computed on the content that is related to the supplemental interests.

Affinity may be based on the relationship between the identified supplemental interest topic and the content of the document. At step **1825**, the identified supplemental content is ranked based on the affinity score and or the weight of the supplemental interests. Each rank may be weighed with the interest weight from the supplemental set and the article interests weight. An uncertainty measure can also be added to each article—and a number of positive/negative interaction can be assigned. The ranked supplemental content is then passed to the supplemental content pool **1730**.

Ordering of the supplemental content pool can be any number of way. In an embodiment, it may be ordered by affinity used in constructing the pool of supplemental articles. In another embodiment, popularity of the article may be used to do the ordering. Randomly selected the articles can also be used since the supplemental pool is already pre-selected to contain supplemental interests candidates. At step **1830**, the ranked supplemental content is selected from the supplemental content pool **1730** by the supplemental content selector **1745** for placement into the personalized content stream. Once the pool of supplemental articles has been selected, it is then combined with the regular set of articles identified for the user. This combination can be done in many ways. In an embodiment, the supplemental content is selected and it is then inserted into content pool of articles for the user. In another embodiment, the score assigned to each article in the content pool of articles and the supplemental articles are ordered by this score across both set of articles and the top articles are returned to the user as recommended content. The score in an embodiment can be computed by combining popularity and affinity scores. The final score can also include a random factor computed from the distance in order to

explore the space of known and unknown interests. Articles with interests with large distances will have larger variation in final score. The user **105** is presented with the recommended list of articles and engages with the articles. Articles with more positive interactions will change the user profile **160** by increasing the weights with those article interests. Articles with more negative interactions will change the user profile **160** by decreasing the weights with those article interests. The more often an interest in the profile is presented in an article to the user, the smaller the uncertainty associated with that supplemental interest will be.

FIG. **19** depicts an embodiment of a supplemental interests identifier **1715**. Supplemental interest identifier **1715** may be comprised of a known interest analyzer **1905**, search scope determiner **1910**, supplemental interests searcher **1915** and supplemental interest weighing unit **1920**. Supplemental interest identifier **1715** receives a user's known interest and their associated weights from high dimensional vector **1600** and identifies a user's supplemental interests and their respective weights.

FIG. **20** is a flowchart of an exemplary process of the supplemental interest identifier **1715**. At step **2000**, known interest analyzer **1905** receives the user's high dimensional vector from the user's profile **160**. At step **2005**, the search scope determiner **1910** receives the supplemental interest search parameters **1925** which may include the distance from a known interest the supplemental interest identifier should search for interests. Next, at step **2010** the supplemental interest searcher **1915** relying on the interest parameters from the search scope determiner **1910** searches the known interests based on the parameters and identifies supplemental interest based on the content taxonomy **165**. For example, as seen in FIG. **16***a*, if the scope of the search parameters include a distance of **5**, then sports **1402** may be an identified supplemental interest based on the clear interest in jazz **1411** because it is within the defined distance parameter **5**. Similarly, politics **1401** which has a distance=1 will be a supplemental interest identified from interest elections **1406**.

Once identified, at step **2015** the distance for each supplemental interest is computed and at step **2020** the supplemental interest weight unit **1920** computes a weight for each supplemental interest based on the distance. Supplemental interest weights are inversely proportional to their distances, that is the greater the distance, the smaller the weight assigned to each supplemental interest. At step **2025** the weight of each supplemental interest may be outputted to for example, to the supplemental content identifier **1720** of supplemental interest pool **1725** for use in identifying supplemental content.

FIG. **21** is a diagram of an embodiment of the supplemental content identifier **1720**. supplemental content identifier **1720** comprises supplemental content candidate analyzer **2105**, content related activity analyzer **2110**, affinity calculation unit **2115**, certainty score calculation unit **2120** and supplemental content selector **2125**.

FIG. **22** describes the flow of supplemental content identifier **1720**. At step **2200**, supplemental content identifier **1720** receives the content interest weights from supplemental interest weighing unit **1920**. At step **2205** for each supplemental interest identified, supplemental content is obtained from the content pool **135** or from the content sources **110**. Once content is obtained, in step **2210** the affinity score between the proposed supplemental content and the supplemental interest is computed in affinity calculation unit **2115**. At step **2215** the supplemental content is analyzed in content related activity analyzer **2110** for quality events associated with that content indicating its broad quality. These events may include user dwell time, user click-through-rates, etc. At step **2220**, a confidence score of the potential supplemental content is calculated by the certainty score calculation unit **2120** which then passes the confidence score to the supplemental content selector **2125** at step **2225**. Based on the content affinity score and the content confidence score, i.e., quality of the content. At step **2225** supplemental content is selected and outputted the supplemental content pool **1730**.

To implement the present teaching, computer hardware platforms may be used as the hardware platform(s) for one or more of the elements described herein. The hardware elements, operating systems, and programming languages of such computers are conventional in nature, and it is presumed that those skilled in the art are adequately familiar therewith to adapt those technologies to implement the processing essentially as described herein. A computer with user interface elements may be used to implement a personal computer (PC) or other type of work station or terminal device, although a computer may also act as a server if appropriately programmed. It is believed that those skilled in the art are familiar with the structure, programming, and general operation of such computer equipment and as a result the drawings should be self-explanatory.

FIG. **23** depicts a general computer architecture on which the present teaching can be implemented and has a functional block diagram illustration of a computer hardware platform that includes user interface elements. The computer may be a general-purpose computer or a special purpose computer. This computer **2300** can be used to implement any components of the unknown interest identifier architecture as described herein. Different components of the system in the present teaching can all be implemented on one or more computers such as computer **2300**, via its hardware, software program, firmware, or a combination thereof. Although only one such computer is shown, for convenience, the computer functions relating to the target metric identification may be implemented in a distributed fashion on a number of similar platforms, to distribute the processing load.

The computer **2300**, for example, includes COM ports **2302** connected to and from a network connected thereto to facilitate data communications. The computer **2300** also includes a central processing unit (CPU) **2304**, in the form of one or more processors, for executing program instructions. The exemplary computer platform includes an internal communication bus **2306**, program storage and data storage of different forms, e.g., disk **2308**, read only memory (ROM) **2310**, or random access memory (RAM) **2312**, for various data files to be processed and/or communicated by the computer, as well as possibly program instructions to be executed by the CPU. The computer **2300** also includes an I/O component **2314**, supporting input/output flows between the computer and other components therein such as user interface elements **2316**. The computer **2300** may also receive programming and data via network communications.

Hence, aspects of the method of discovering user unknown interest from known interests, as outlined above, may be embodied in programming. Program aspects of the technology may be thought of as "products" or "articles of manufacture" typically in the form of executable code and/or associated data that is carried on or embodied in a type of machine readable medium. Tangible non-transitory "storage" type media include any or all of the memory or other storage for the computers, processors or the like, or associated modules thereof, such as various semiconductor memories, tape drives, disk drives and the like, which may provide storage at any time for the software programming.

All or portions of the software may at times be communicated through a network such as the Internet or various other

telecommunication networks. Such communications, for example, may enable loading of the software from one computer or processor into another. Thus, another type of media that may bear the software elements includes optical, electrical, and electromagnetic waves, such as used across physical interfaces between local devices, through wired and optical landline networks and over various air-links. The physical elements that carry such waves, such as wired or wireless links, optical links or the like, also may be considered as media bearing the software. As used herein, unless restricted to tangible "storage" media, terms such as computer or machine "readable medium" refer to any medium that participates in providing instructions to a processor for execution.

Hence, a machine readable medium may take many forms, including but not limited to, a tangible storage medium, a carrier wave medium or physical transmission medium. Non-volatile storage media include, for example, optical or magnetic disks, such as any of the storage devices in any computer(s) or the like, which may be used to implement the system or any of its components as shown in the drawings. Volatile storage media include dynamic memory, such as a main memory of such a computer platform. Tangible transmission media include coaxial cables; copper wire and fiber optics, including the wires that form a bus within a computer system. Carrier-wave transmission media can take the form of electric or electromagnetic signals, or acoustic or light waves such as those generated during radio frequency (RF) and infrared (IR) data communications. Common forms of computer-readable media therefore include for example: a floppy disk, a flexible disk, hard disk, magnetic tape, any other magnetic medium, a CD-ROM, DVD or DVD-ROM, any other optical medium, punch cards paper tape, any other physical storage medium with patterns of holes, a RAM, a PROM and EPROM, a FLASH-EPROM, any other memory chip or cartridge, a carrier wave transporting data or instructions, cables or links transporting such a carrier wave, or any other medium from which a computer can read programming code and/or data. Many of these forms of computer readable media may be involved in carrying one or more sequences of one or more instructions to a processor for execution.

Those skilled in the art will recognize that the present teachings are amenable to a variety of modifications and/or enhancements. For example, although the implementation of various components described above may be embodied in a hardware device, it can also be implemented as a software only solution. In addition, the components of the system as disclosed herein can be implemented as a firmware, firmware/software combination, firmware/hardware combination, or a hardware/firmware/software combination.

While the foregoing has described what are considered to be the best mode and/or other examples, it is understood that various modifications may be made therein and that the subject matter disclosed herein may be implemented in various forms and examples, and that the teachings may be applied in numerous applications, only some of which have been described herein. It is intended by the following claims to claim any and all applications, modifications and variations that fall within the true scope of the present teachings.

We claim:

1. A method for identifying content for a user, the method implemented on a machine having at least one processor, storage, and a communication interface connected to a network, the method comprising:

retrieving information related to a user from a user profile, wherein the information indicates one or more known interests of the user;

identifying at least one known interest of the user based on the information;

determining one or more supplemental interests with respect to each of the identified at least one known interest of the user, where the one or more supplemental interests do not overlap with the one or more known interests of the user;

identifying supplemental content associated with the one or more supplemental interests with respect to each of the identified at least one known interest of the user;

ranking each piece of content in the supplemental content; and

selecting at least one piece of content in the supplemental content based on the ranking, wherein

the selected at least one piece of supplemental content associated with the one or more supplemental interests is used to discover unknown interest of the user.

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

identifying relatedness between each piece of content in the supplemental content and its corresponding supplemental interest; and

outputting the selected content from the supplemental content.

3. The method of claim **1** further comprising:

randomly obtaining content; and

adding the randomly obtained content to the supplemental content.

4. The method of claim **1** further comprising filtering the ranked content in the supplemental content based on a criteria.

5. The method of claim **1**, wherein step of determining comprises:

estimating a metric for each of a plurality of candidate supplemental interests; and

selecting the one or more supplemental interests based on their respective metrics with respect to a threshold.

6. The method of claim **5**, wherein the metric includes at least one of:

a distance between two interests in a content taxonomy;

a co-occurrence of two interests in a collection of content;

a co-occurrence of two interests in a set of user profiles;

a co-occurrence of two interests in a set of user sessions; and

any combination thereof.

7. The method of claim **1**, wherein the unknown interest of the user is discovered based on interaction between the user and the selected at least one piece of supplemental content.

8. A system for identifying unknown user content, the system comprising:

a retrieval unit for retrieving information related to a user from a user profile, wherein the information indicates one or more known interests of the user;

an interest analyzer for identifying at least one known interest of the user based on the information;

a supplemental interest identifier for determining one or more supplemental interests with respect to each of the identified at least one known interest of the user, where the one or more supplemental interests do not overlap with the one or more known interests of the user;

a supplemental content identifier for identifying supplemental content associated with the one or more supplemental interests with respect to each of the identified at least one known interest of the user;

a ranking unit for ranking each piece of content in the supplemental content; and

a selector for selecting at least one piece of content in the supplemental content based on the ranking, wherein

the selected at least one piece of supplemental content associated with the one or more supplemental interests is used to discover unknown interest of the user.

9. The system of claim **8**, further comprising:

a supplemental weighting unit for identifying relatedness between each piece of content in the supplemental content and its corresponding supplemental interest; and

an output for outputting the selected content from the supplemental content.

10. The system of claim **8**, further comprising a random content selector configured for:

randomly obtaining content; and

adding the randomly obtained content to the supplemental content.

11. The system of claim **8**, wherein the supplemental interest identifier is further configured for:

estimating a metric for each of a plurality of candidate supplemental interests; and

selecting the one or more supplemental interests based on their respective metrics with respect to a threshold.

12. The system of claim **11**, wherein the metric includes at least one of:

a distance between two interests in a content taxonomy;

a co-occurrence of two interests in a collection of content;

a co-occurrence of two interests in a set of user profiles;

a co-occurrence of two interests in a set of user sessions; and

any combination thereof.

13. The system of claim **8**, wherein the unknown interest of the user is discovered based on interaction between the user and the selected at least one piece of supplemental content.

14. The system of claim **8**, wherein the ranked content in the supplemental content is filtered based on a criteria.

15. A non-transitory machine-readable medium having recorded thereon information for identifying unknown user interest, wherein the information, when read by a machine, causes the machine to perform the steps of:

retrieving information related to a user from a user profile, wherein the information indicates one or more known interests of the user;

identifying at least one known interest of the user based on the information;

determining one or more supplemental interests with respect to each of the identified at least one known interest of the user, where the one or more supplemental interests do not overlap with the one or more known interests of the user;

identifying supplemental content associated with the one or more supplemental interests with respect to each of the identified at least one known interest of the user;

ranking each piece of content in the supplemental content; and

selecting at least one piece of content in the supplemental content based on the ranking, wherein

the selected at least one piece of supplemental content associated with the one or more supplemental interests is used to discover unknown interest of the user.

16. The non-transitory machine-readable medium of claim **15**, wherein the information, when read by the machine, further causes the machine to perform the steps of:

identifying relatedness between each piece of content in the supplemental content and its corresponding supplemental interest; and

outputting the selected content from the supplemental content.

17. The non-transitory machine-readable medium of claim **15**, wherein the information, when read by the machine, further causes the machine to perform the steps of:

randomly obtaining content; and

adding the randomly obtained content to the supplemental content.

18. The non-transitory machine-readable medium of claim **15**, wherein step of determining comprises:

estimating a metric for each of a plurality of candidate supplemental interests; and

selecting the one or more supplemental interests based on their respective metrics with respect to a threshold.

19. The non-transitory machine-readable medium of claim **18**, wherein the metric includes at least one of:

a distance between two interests in a content taxonomy;

a co-occurrence of two interests in a collection of content;

a co-occurrence of two interests in a set of user profiles;

a co-occurrence of two interests in a set of user sessions; and

any combination thereof.

20. The non-transitory machine-readable medium of claim **15**, wherein the unknown interest of the user is discovered based on interaction between the user and the selected at least one piece of supplemental content.

* * * * *