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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 February 2009 (19.02.2009)

PCT

(10) International Publication Number
WO 2009/023067 A1

(51) International Patent Classification:
*G06F 15/16* (2006.01)

(21) International Application Number:
PCT/US2008/008137

(22) International Filing Date: 30 June 2008 (30.06.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/893,797 16 August 2007 (16.08.2007) US

(71) Applicant *(for all designated States except US)*: **FACEBOOK, INC.** [US/US]; 151 University Avenue, Palo Alto, California 94301 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **JUAN, Yun-fang** [—/US]; C/o Facebook, Inc., 151 University Avenue, Palo Alto, California 94301 (US). **JIN, Kang-xing** [US/US]; C/o Facebook, Inc., 151 University Avenue, Palo Alto, California 94301 (US).

(74) Agents: **BATHURST, Brian** et al.; 2200 Geng Road, Palo Alto, California 94303 (US).

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

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

**Published:**
— *with international search report*

(54) **Title:** SYSTEM AND METHOD FOR INVITATION TARGETING IN A WEB-BASED SOCIAL NETWORK

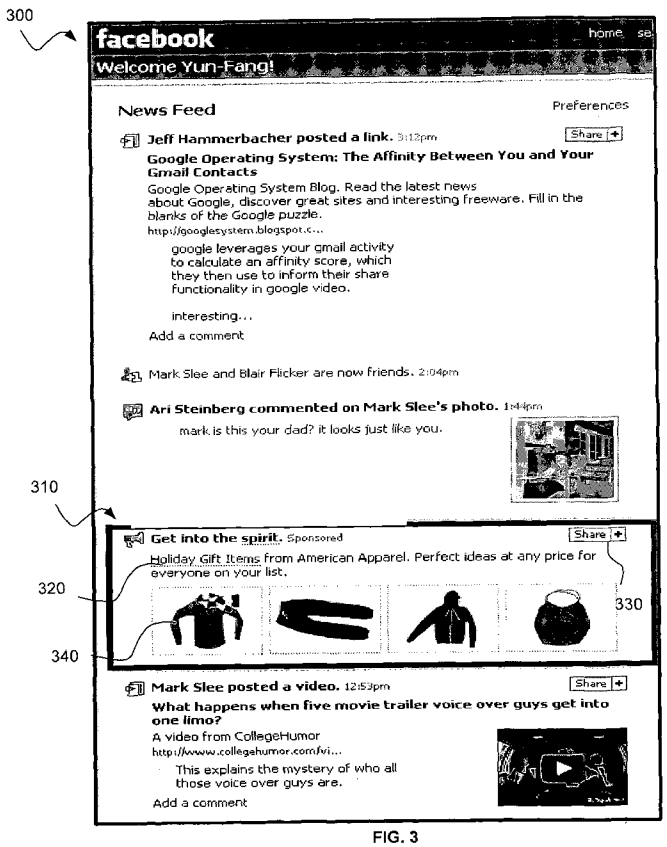


FIG. 3

(57) **Abstract:** A system and method for selecting users of a web-based social network who are likely to respond to an invitation, each of the users having associated profile information is disclosed. The method includes selecting pilot users and a reduced set of keywords from the profile information. The method further includes sending the invitation to the pilot users, receiving responses from the pilot users, and classifying the responses as either positive or negative. A training set of vector pairs is created each vector pair representing a pilot user and including data representing a classified response and training keywords selected from the reduced set of keywords and associated profile information for the pilot user. A function is determined based on the vectors and used to calculate a likelihood that each of one or more users of the web based social network will respond to the invitation.

WO 2009/023067 A1

# System and Method for Invitation Targeting in a Web-Based Social Network

## CROSS-REFERENCE TO RELATED APPLICATIONS

**[0001]** The present application incorporates by reference:

U.S. Patent Application Serial No. 11/639,655 filed on December 14, 2006 entitled "Systems and Methods for Social Mapping," which in turn claims the benefit and priority of U.S. Provisional Patent Application Serial No. 60/750,844 filed on December 14, 2005 entitled "Systems and Methods for Social Mapping,"

U.S. Patent Application Serial No. 11/499,093 filed on August 2, 2006 entitled "Systems and Methods for Dynamically Generating Segmented Community Flyers,"

U.S. Patent Application Serial No. 11/503,242 filed on August 11, 2006 entitled "System and Method for Dynamically Providing a News Feed About a User of a Social Network,"

U.S. Patent Application Serial No. 11/580,210 filed on October 11, 2006, entitled "System and Method for Tagging Digital Media,"

U.S. Patent Application Serial No. 11/893,820 filed on August 16, 2007, entitled "System and Method for Keyword Selection in a Web-Based Social Network," and

U.S. Patent Application Serial No. 11/796,184 filed on April 27, 2007 entitled "System and Method for Giving Gifts and Displaying Assets in a Social Network Environment."

# BACKGROUND OF THE INVENTION

## Field of the Invention

[0002] The present invention relates generally to social networks, and more particularly to invitations in a social network.

## Description of Related Art

[0003] Social network environments allow users to send many types of invitations to other users. Examples of an invitation include an advertisement, a request to join a group, a request for an information exchange, a survey, a request to write a blog entry, a request to verify a photo tag, and so forth.

[0004] An invitation may be personalized or targeted to a particular user in the social network environment. Targeting may include predicting a likelihood that the user will respond to an invitation and presenting the invitation to the user if the likelihood is high. Targeting may also be useful for determining that the user has a low probability of responding to certain invitations because the invitations are not interesting to the user.

[0005] There are several approaches to personalizing or targeting an invitation to a particular user. One approach is to track buying patterns. For example, after a customer purchases a book via an internet store, the store may tell the customer about products in stock that the customer might like such as other books by the same author, or books purchased by other people who also bought the book that the customer purchased. This approach, however, is limited to customers who purchase items.

[0006] Another approach to targeting is to present invitations to a user who is a member of a particular group. Groups may be based on gender, school, age,

residence, club membership, political affiliation, and so on. However, not all groups are well defined within the social network environment and determining that a person is a member of a group may be cumbersome and require skill and an understanding of the group dynamics and common interests. Unfortunately, none of these approaches automatically select users of a social network environment who have an increased probability of responding positively to an invitation.

## SUMMARY OF THE INVENTION

[0007] The invention provides a method for selecting users of a web-based social network, each having associated profile information, who are likely to respond to an invitation. In one embodiment, the method generates a probability function that will predict the likelihood of a user in a social network environment responding to an invitation. A pilot group of users is selected, as is a reduced set of keywords based on profiles of the pilot group. The method further includes sending the invitation to the pilot group and creating a training set of vectors based on responses to the invitation, the pilot group profiles, and the reduced set of keywords. The probability function may be determined from the training set and applied to the users in the social network environment to predict which users are more likely to respond to the invitation.

[0008] In another embodiment, the method comprises selecting a plurality of pilot users from the users in the web based social network, selecting a reduced set of keywords from the profile information for the pilot users, sending the invitation to the pilot users, and receiving responses to the invitation from the pilot users. The responses are classified as either positive or negative and a training set of vector pairs is created, each vector pair representing a pilot user and including data representing the classified response received from the pilot user and a set of training keywords selected from the reduced set of keywords and based at least in part on the associated profile information for the pilot user. The method further includes determining a function based on the training set of vector pairs that calculates from a user's profile information a likelihood that the user will respond to the invitation and calculating from the function a likelihood that each of one or more of the users in the web based social network will respond to the invitation.

## BRIEF DESCRIPTION OF THE DRAWINGS

**[0009]** FIG. 1 illustrates an exemplary social network that may be used with various embodiments of the invention.

**[0010]** FIG. 2 illustrates one embodiment an architecture of the invitation engine of FIG. 1.

**[0011]** FIG. 3 is screen a shot illustrating an example of an invitation embedded in a news feed for a user.

**[0012]** FIG. 4 is an illustration of an exemplary iPod download invitation according to one embodiment.

**[0013]** FIG. 5 is a flow chart of an exemplary method for selecting target users of a web-based social network.

## DETAILED DESCRIPTION OF THE INVENTION

**[0014]** The present invention provides a method for selecting users in a web-based social network who are likely to respond to an invitation. In one embodiment, the invitation is first sent to a pilot group of users selected at random. Positive and negative responses are recorded. A set of the pilot group profiles containing a reduced set of keywords may be correlated with the positive and negative responses to the invitation and the correlations may be used to determine a probability function that indicates the likelihood of responses based on the profiles. The profiles of other the users in the social network may be analyzed using the probability function, and target users may be selected to receive the invitation based on the likelihood of responding to the invitation.

**[0015]** FIG. 1 illustrates an exemplary social network environment 100 that may be used with embodiments of the invention. One or more users 102 at user devices 110, are coupled to a social network provider 130 via a communications network 120. In various embodiments, the user devices 110 may include a computer terminal, a personal digital assistant (PDA), a wireless telephone, a digital camera, a mobile device, a mobile phone, a cell-phone, a smart-phone, a notebook computer, a laptop computer, a hand-held game console, and so forth. The communications network 120 may include a local area network (LAN) such as an intranet, a wide area network (WAN) such as the Internet, a wireless network, etc.

**[0016]** The social network provider 130 is typically a server that provides social networking services, communication services, dating services, company intranets, and/or online games, etc. The social network provider 130 may assemble and store profiles of the users 102 for use in providing the social networking services. In some embodiments, the social network environment 100 includes one or more segmented communities, which are separate, exclusive or semi-exclusive subsets of

the social network environment 100, wherein users 102 who are segmented community members may access and interact with other members of their respective segmented community. Examples of such groupings are set forth in further detail in co-pending U.S. Patent Application Serial No. 11/369,655.

[0017] The users 102 may include various types of users 102A, 102B . . . 102N, (hereinafter users 102A-102N). For example, a user 102A may be a pilot user who is selected to receive an invitation as a part of the pilot study, while a user 102B is a target user selected to receive the invitation based on a probability function. A probability function is a function that returns a probability that a user 102 will respond positively (or negatively) to the invitation. It may, for example, be based on one or more keywords in the profile of the user 102.

[0018] The social network environment 100 further includes an invitation engine 140 coupled to the social network provider 130. The invitation engine 140 is configured to select a group of pilot users 102A for a pilot study, send invitations to the pilot group, determine a probability function based on results of the pilot study, select the target users 102B from the users 102 using the probability function, and send invitations the target users 102B.

[0019] Keywords include words or phrases relating to information entered by users and stored in the respective profiles of the users 102A-102N. Keywords may also be words or phrases entered by the social network provider 130 to characterize the users 102A-102N. Keywords may include words relating to demographics, interests, usage, actions, or other information that may describe each of the users 102A-102N. A particular user profile may include multiple occurrences of one or more keywords. The profile information for the users 102A-102N while typically stored with the social network provider may also be found in profile databases in the invitation engine 140.

[0020] FIG. 2 illustrates one embodiment an architecture of the invitation engine 140 of FIG. 1. As shown, the invitation engine 140 includes a profile database 200, an invitation module 210, a pilot group module 220, a dimension reduction module 230, a training set module 240, and a probability function module 250.

[0021] The profile database 200 manages profile information that is provided by users 102 of the social network. As discussed above, the profile information includes keywords relating to demographics, interests, usage, actions, and/or other information that may describe the users 102. The profile database 200 may store values to represent various types of keywords, including numerical values, binary values, and/or categorical values. For example, a numerical value may represent an age or phone number. A binary number may indicate whether a keyword occurs or does not occur in the profile of a user 102. For example, if the keyword is "football," a "1" may indicate that the word "football" occurs at least once in the profile of the user 102 and a "0" that the word "football" does not occur in the profile of the user 102. In other embodiments, a "1" may indicate that the word "football" occurs more than a predetermined number of times in the profile for the user 102. A categorical value may represent a selection from a list. For example, political views may be categorized as 1 = liberal, 2 = conservative, 3 = independent, etc.

[0022] Keywords relating to demographics may include information regarding age, gender, relationship status, home state, and school. Demographic keywords may be represented by numerical values, binary values, and/or categorical values. Keywords relating to interests include book titles, authors, movies, television programs, and music and may be represented by binary values. Examples of keywords relating to usage include information regarding friendships, blog posts, online gifts given and received via the social network provider 130, online purchases via the social network provider 130, photo uploads, photo downloads, photo tags,

and photo tag confirmations and may be represented by numerical values, binary values, and/or categorical values.

[0023] Table 1 illustrates an example of various keyword names, keyword types, and keyword values that may be stored in the profile database 200. For example, the keyword "Birth Year" in the Keyword Names column of Table 1 is a demographic keyword type and may be represented by a numerical value. The keyword, "Political Views" is also a demographic keyword type, but one that may be represented by a categorical value (e.g., 1=liberal, 2=conservative, 3=independent, etc.). The entry "Top 5000 Favorite Movies" in the Keyword Names column represents 5000 different keywords each associated with one of 5000 different movie titles, respectively. For example, the movie title "Gone with the Wind" may be a keyword. Each of the 5000 keywords is an Interest keyword and is represented by a binary value in the illustrated embodiment to indicate that the movie title occurs or does not occur in the profile of a user 102. While Demographic and Interest keyword types are illustrated in Table 1, other keyword types (e.g., contacts, skills, etc.) may also be included.

Table 1: Keywords

| Keyword Names | Keyword Type | Keyword Value |
|---|---|---|
| Gender | Demographic | Categorical |
| Birth Year | Demographic | Numerical |
| Political Views | Demographic | Categorical |
| Relationship Status | Demographic | Categorical |
| User Type | Demographic | Categorical |
| Top 5000 Favorite Movies | Interests | 5000 Binary |
| Top 5000 Favorite Books | Interest | 5000 Binary |
| Top 5000 Favorite Music | Interest | 5000 Binary |
| Top 5000 Favorite Activities | Interest | 5000 Binary |
| Top 5000 Favorite TV shows | Interest | 5000 Binary |

[0024] The profile for each user 102A-102N may be represented as a vector and each keyword that occurs in the profile may be represented as a dimension or an element of the vector. Dimensions may include entries other than keywords and some keywords may not be represented by a dimension. In some embodiments, dimensions may represent multiple keywords. Each dimension may include a numerical value, a binary value, or a categorical value. In various embodiments, a numerical value may represent the number of occurrences of a particular keyword in the profile of the user 102, an age of the user 102, income, the number of friends of the user 102, etc. A binary value may represent at least one occurrence (e.g., "1") or non-occurrence (e.g., "0") of the keyword in the profile of the user 102. A categorical value may represent a political view, gender, religion, etc. A profile database containing all the keywords for all the users 102 may include as many as 10,000 to 100,000 or more keywords i.e., dimensions. On the other hand, a reduced set of keywords (discussed below) may include many fewer keywords, for example 100 to 200 keywords. In some embodiments, the profile database 200 and/or the social network provider 130 includes a reduced set of keywords.

[0025] The invitation module 210 is configured to send an invitation to users 102A-102N of the social network environment 100 and receive responses to the invitation from the users 102A-102N. In various embodiments, the invitation module 210 may send invitations and receive responses from pilot users 102A and/or target users 102B. Examples of an invitation include an advertisement, a survey, a request to provide information to the social network provider 130, a request to send information to another user 102, a suggestion to form a group, a request to join a group, a request to confirm a photo tag, an offer to purchase a real, digital, or virtual asset, and so on. In some embodiments, an invitation may include an opportunity for the user to respond by taking an action.

[0026] In various embodiments, a response includes accepting the invitation by clicking on a link within the invitation, rejecting the invitation, requesting more information about the invitation, requesting to be reminded later of the invitation, and so forth. In some embodiments, ignoring the invitation may be a default response. A positive response may include clicking on a button associated with the invitation. Clicking on a link in an invitation is known as a "click through." Examples of a "click through" response include clicking on a link to purchase a product, view a webpage, download information, and upload information. A click-through rate may be calculated by dividing a number of "click-through" responses by a number of users who received the invitation. A response may further include taking other actions, such as joining a group, posting a photo, tagging a photo, answering a survey, forwarding a message, forming a group, posting a blog, and so forth.

[0027] The invitation module 210 may be configured to receive responses for a predetermined period of time. For example, the invitation module 210 may send an invitation to 50,000 pilot users 102A and receive responses to the invitation via the invitation module 210 for one hour. In some embodiments, the invitation module 210 may receive a predetermined number of responses. For example, the invitation module 210 may send an invitation to 50,000 pilot users 102A and stop accepting responses after receiving the first 10,000 responses.

[0028] The pilot group module 220 is configured to select the pilot users 102A from the users 102 and provide a list of the pilot users 102A to the invitation module 210. The pilot group module 220 may randomly select the pilot users 102A from all of the users 102 or from a subset of the users 102. Alternatively, the pilot group module 220 may select pilot users 102A based on various criteria, for example, age, gender, location, and so on.

[0029] The pilot group module 220 is further configured to receive the responses from the invitation module 210. The pilot group module 220 may provide the invitation module 210 with a time period for accepting responses from the pilot users 102A. Alternatively, the pilot group module 220 may direct the invitation module 210 to receive a predetermined number of responses from the pilot users 102A. For example, the pilot group module 220 may provide the invitation module with directions to accept only the first 10,000 responses.

[0030] In some embodiments, the pilot group module 220 may subdivide the pilot group into a plurality of subgroups randomly or according to one or more characteristics of the pilot users 102A. For example, a pilot group of about 50,000 pilot users 102A may be subdivided into 10 subgroups of about 5,000 pilot users 102A based on some characteristic or combination of characteristics, for example, geographical region, age bracket, occupation, membership in a social group, and so on. The pilot group module 220 may count the number of pilot users 102A who respond positively in each of the 10 separate segmented communities and direct the invitation module 210 to send the invitation to all of the users 102 in the network who share the characteristics of the pilot group that had the highest number of positive responses. Alternatively, the pilot group module 220 may divide the social network community 100 into subgroups based on characteristics of the users 102 and select a plurality of pilot users 102A at random from each of the subgroups. For example, 10 separate segmented communities may be selected from the social network community 100 and the pilot group module 220 may select 5,000 pilot users 102A at random from each of the segmented communities. The positive responses may be counted as above for each of the 10 separate segmented communities. This may save computation time in generating new probability functions for related invitations.

[0031] The dimension reduction module 230 is configured to reduce the number of keywords (i.e., dimensions) used in the profiles associated with the pilot group. The number of different keywords in the various profiles for all the users 102 can result in a very large set of keywords before dimension reduction. For example, a total of about 10,000 to 100,000 keywords might be found in the profiles for all or a large number of the users 102. Thus, 10,000 to 100,000 keywords may be available for correlation with responses. The memory space and computing resources required to process correlations with such a large number of keywords can be very large.

[0032] The dimension reduction module 230 reduces the 10,000 to 100,000 keywords to a reduced set of, for example, about 100 to 200 keywords using dimensional reduction techniques that are known in the art. The reduced keyword set may be based on the keywords collectively found in the profiles associated with the group of pilot users 102A. A simple, intuitive example of a keyword reduction technique includes keeping all the keywords found in all the profiles of the pilot group and discarding all keywords not found in their profiles. However, the number of remaining keywords might be to numerous. Techniques that may be useful for reducing the number of dimensions while minimizing information loss include singular vector decomposition (SVD), probabilistic latent semantic indexing (PLSI), linear discriminant analysis (LDA), feature selection, and so forth. The keyword reduction may be performed before or after sending the invitation to the pilot users 102A. In some embodiments, keyword reduction may produce new keywords that are based on combinations of keywords in the data set before reduction. For example, the keyword reduction module 230 may group several movie keywords (e.g., "spider man 1," "spider man 2," and "spider man 3") into one reduced keyword "spider man" representing spider man in general.

[0033] The training set module 240 is configured to classify the responses, correlate the classified response from each pilot user 102A with keywords in the

profile database 200 for the pilot user 102A, and create a training set of data pairs from the correlations. In some embodiments, the training set may not include data pairs from all of the pilot users 102A and the training set module 240 may select the pilot users 102A to be included the training set as discussed below.

[0034] The training set module 240 may classify each response for each pilot user 102A. Classification of a response includes determining if a response is a positive response or negative response. For example, the responses from the pilot users 102A may include clicking on the invitation (a positive response) or taking no action (a negative response). In various embodiments, positive responses include accepting an invitation by clicking on a link within the invitation, requesting more information about the invitation, requesting to be reminded later about the invitation, joining a group, posting a photo, tagging a photo, and so forth. Negative responses may include affirmatively rejecting the invitation (e.g., clicking on a "no" button), ignoring the invitation, abstaining from responding, and so forth. In some embodiments, classification includes assigning a value of "1" to a positive response and a value of "0" to a negative response. The training set module 240 may store the classifications ("1" or "0") in the profile database with the profile information associated with the respective pilot users 102A.

[0035] The training set includes correlated pairs of data, each data pair representing a classified response and a profile of a pilot user 102A. The data pairs may be represented as vector pairs. Each vector pair may include a response vector representing a classified response by a pilot user 102A and a keyword vector representing keywords in the profile of the pilot user 102A. Each response vector may include a binary value as discussed above. Each keyword vector may include numerical, binary, or categorical values. For simplicity, only binary values are discussed below, thus, each dimension representing a keyword in the vector includes a "1" or "0" representing an occurrence or non-occurrence, respectively, of the

keyword. However, in general, dimensions including numerical and/or categorical values may also be included in the training set vectors.

[0036] In a simplified example, the reduced keyword set includes three keywords, namely ("Beatles," "hockey," "Murasaki") and the training set includes a first pilot user 102A and second pilot user 102A. A user profile for the first pilot user 102A may include the keywords ("Shakespeare," "Beatles," "hockey," "orange," "stargazing") and the keyword vector may be represented by a (1,1,0). The first pilot user 102A responds positively to an invitation for a football video and a "1" is entered in the training set response vector for the first pilot user 102A to indicate the positive response. Thus, the training set vector pair for the first pilot user 102A is (1), (1,1,0) representing: (response=1), ("Beatles"=1, "hockey"=1, "Murasaki"=0). The user profile for the second pilot user 102A may include the keywords ("Beatles," "red hot chili peppers," "pencil," "a bridge too far," "carpet cleaning," "rose"). The second pilot user 102A responds negatively to an invitation for the football video and a "0" is entered to indicate the negative response. Thus, the training set vector pair for the second pilot user 102A is (0), (1,0,0) representing: (response=0), ("Beatles"=1, "hockey"=0, "Murasaki"=0). This example is merely illustrative and the training set module 240 generally uses more complex methods known in the art for selecting keywords from the reduced keyword set for the keyword vector and correlating the response vector with the keyword vector. For example, some keywords common to both the reduced keyword set and a profile may not be represented in the keyword vector while some keywords not in common may be represented.

[0037] The training set may include vector pairs for all the pilot users 102A. Generally, the number of pilot users 102A who respond positively is much less than the number of pilot users 102A who respond negatively. In some embodiments, the training set module 240 may assign relative weights to the positive and/or negative pairs in the training set. The weights may be selected according to various weighting

schemes. In some embodiments, the relative weights of the positive and negative response may be selected to make the sum of the weighted positive pairs about equal to the sum of the weighted negative pairs. For example, if a pilot group returns 10,000 positive responses and 50,000 negative responses, the training set module 240 may assign a weight to the vector pairs in the positive responses that is five times the weight assigned to the vector pairs in the negative responses. Other weighting schemes may be applied to the vector pairs in the training set.

[0038] In some embodiments, the training set module 240 is configured to select a subset of the pilot users 102A to be included the training set. For example, the training set module 240 may stratify the pilot users into two groups of pilot users 102A based on whether the response vectors are positive or negative and include entries for all pilot users 102A who have responded positively and a random selection of about an equal number of entries for pilot users 102A who have responded negatively. When the training set is still too large, the training set module may select a smaller number of pilot users 102A randomly in about equal numbers from each of the two stratified groups.

[0039] The probability function module 250 is configured to generate a probability function based on the training set. The probability function module 250 may use the probability function to predict the likelihood that a user will respond positively (or negatively) to the invitation. In various embodiments, the probability function module 250 generates the probability function using a supervised learning procedure, or a machine learning technique such as a support vector machine (SVM), a neural network, or a boosted tree procedure. Boosted tree procedures may be used because boosted trees do not require normalization of attributes and output may be used to interpret results. More information about the probability function and supervised learning procedures is contained in a paper entitled "Personalization for

Online Social Networks" by Yun-Fang Juan, et al., presently unpublished and attached hereto as an appendix.

[0040] The probability function module 250 is further configured to select target users 102B to receive the invitation. The target users 102B may be selected from all the users 102 of the social network environment 100. For example, the probability function module 250 may rank all the users 102 from highest to lowest according to a calculated likelihood of responding positively to the invitation and select the 500,000 highest ranked users 102 to become target users 102B. In some embodiments, the target users 102B may be selected from less than all the users 102. For example, the probability function module 250 may rank a fraction of the users 102 and select target users 102B as above. Alternatively, the probability function module 250 may select target users 102B for whom the calculated likelihood of responding positively to an invitation exceeds a predetermined threshold value. The probability function module 250 may adjust the predetermined threshold value to select fewer or more target users 102B.

[0041] In some embodiments, a similar invitation may be sent to the selected target users 102B. A similar invitation may be any invitation that contains a similar content, message, or function as the invitation sent to the pilot users 102A. For example, an invitation to enter a blog about surfing may be similar to an invitation in the form of an advertisement to purchase snorkeling equipment via the social network provider 130 since both invitations relate to ocean sports.

[0042] The invitation module 210 may track the number of target users 102B who receive the invitation, the positive and negative responses to the invitation sent to the target users 102B, and/or the click-through rate. The response data tracked by the invitation module 210 may be used to perform keyword extraction. Please see co-pending U.S. Patent Application Serial No. 11/893,820 filed on August 16, 2007,

entitled "System and Method for Keyword Selection in a Web-Based Social Network."

[0043] While a response to an invitation is used as one type of response variable entered in the training set vectors, other types of response variables may be used to help segment the users 102 and allow vendors to target users 102. For example, a response variable may include a frequency of usage of a user interface element of the social network environment 100. Examples of such usage include number of blog posts, number of mobile photo uploads, etc. In some embodiments, the response variable may include a click through rate of a content element. A position of the content may be provided as a dimension to the training set module 240 and/or dimension reduction module 230 to account for positional effects. Group membership may be used as the response variable. For example, a response variable may have value of "1" if a user is a member of the interested group and "0" otherwise.

[0044] Although the invitation engine 140 is described as being comprised of various components (the profile database 200, the invitation module 210, the pilot group module 220, the dimension reduction module 230, the training set module 240, and the probability function module 250), fewer or more components may comprise the invitation engine 140 and still fall within the scope of various embodiments.

[0045] FIG. 3 is a screen shot illustrating an example of an invitation 310 embedded into a news feed 300 for a user 102. A news feed presents information about friends of a user 102 in a social network environment 100. The invitation 310 is contained in the form of a feed advertisement in the news feed 300 and includes links 320, 330, and 340. The link 320 is configured to direct the user 102 to a web based gift store. The link 330 is configured to enable the user 102 to forward or share the invitation 310 with another user 102 in the social network environment 100. The link

340 uses one or more clothing icons to direct the user 102 to particular pages within a web based gift store.

**[0046]** The invitation module 210 may embed the invitation 310 into a news feed 300 directed to pilot users 102A and monitor pilot users 102A for positive and negative responses to the invitation 310. For example, a positive response may include clicking on one or more of the links 320, 330 and 340. The invitation module 210 may send the same invitation 310 to target users 102B who are selected based on a probability function determined from results of the responses from pilot users 102A.

**[0047]** FIG. 4 is an illustration of an exemplary iPod download invitation 400 according to one embodiment. The iPod download invitation 400 contains download links 410 and 420, and share link 430. When selected, the download links 410 and 420 download a football game or information about the football game to the iPod. The share link 430 enables the user 102 to forward the invitation 400 to another user 102 of the social network. The invitation module 210 may embed the iPod download invitation 400 into a news feed 300 directed to pilot users 102A. The invitation module 210 may monitor the pilot users 102A for positive or negative responses. For example, the response to the iPod download invitation 400 may be considered positive if pilot user 102A clicks the share link 430 sending the iPod download invitation 400 to an acquaintance (another user 102) who selects the download link 410 or 420. In some embodiments, the training set may include profile information from users 102 who were not part of the pilot group, but responded positively to the iPod download invitation 400 that was received via the share link 430 from a pilot user 102A. The invitation module 210 may send the same iPod download invitation 400 to target users 102B who are selected based on a probability function determined from results of the responses of the pilot users 102A to the iPod download invitation 400.

[0048] FIG. 5 is a flow chart of an exemplary method 500 for selecting users of a web-based social network who are likely to respond to an invitation, each of the users having associated profile information. In step 502, a plurality of pilot users are selected from the users of the web based social network. In step 504, a reduced set of keywords is selected from the profile information of the pilot users.

[0049] In step 506, an invitation is sent to each of the pilot users. In step 508, responses are received from pilot users. In step 510, the received responses are classified as either positive or negative.

[0050] In step 512, a training set of vector pairs is created. Each of the vector pairs represents a pilot user and includes data representing the classified response received from the pilot user and a set of training keywords selected from the reduced set of keywords and based at least in part on the associated profile information for the pilot user.

[0051] In step 514, a function is determined based on the training set of vectors. In step 516, the function is used to calculate a likelihood that one or more of the users in the web based social network will respond to the invitation. In some embodiments, the likelihood of accepting the invitation is determined for every user of the social network.

[0052] In optional step 518, an invitation is sent to one or more target users who are selected to receive the invitation based on the calculated likelihood of responding.

[0053] While the method 500 is described as being comprised of various steps fewer or more steps may comprise the process and still fall within the scope of various embodiments. In some embodiments, the order of the steps of the method 500 may be varied and still fall within the scope the various embodiments. For

example, the step 504 of selecting a reduced set of keywords may be performed before or after the steps 506, 508, or 510.

**[0054]** The embodiments discussed herein are illustrative of the present invention. As these embodiments of the present invention are described with reference to illustrations, various modifications or adaptations of the methods and/or specific structures described may become apparent to those skilled in the art. All such modifications, adaptations, or variations that rely upon the teachings of the present invention, and through which these teachings have advanced the art, are considered to be within the spirit and scope of the present invention. Hence, these descriptions and drawings should not be considered in a limiting sense, as it is understood that the present invention is in no way limited to only the embodiments illustrated.

Appendix

# Personalization for Online Social Networks

Yun-Fang Juan
yunfang@facebook.com
+1-408-368-1357
Facebook.com
156 University Ave,
Palo Alto, CA, USA

Kang-Xing Jin
kxjin@facebook.com
+1-240-888-2082
Facebook.com
156 University Ave,
Palo Alto, CA, USA

## ABSTRACT

In this paper, we present two novel personalization algorithms for online social networks such as Myspace or Facebook. The algorithms use user profiles to predict individual users' responses to new features, UI changes, content, etc. as well as to extract the most relevant keywords from such predictive tasks. Dimension reduction models and supervised learning procedures are employed for such predictions. Then, a reconstruction procedure is performed to find the most relevant keywords. The predictions can be used to infer users' interest levels for various site introductions and can therefore be used for personalization. The keywords extracted from such tasks provide insights on the connections between users' interests and their preferences for site introductions. Our experiments on the predictive task of ad targeting have shown promising results, improving the ad response rate by over 50 percent in a live setting.

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Extraction, Predictive Models, Personalization

## General Terms

Algorithms, Measurement, Internet.

## Keywords

Information Extraction, Social Networks, Machine Learning, User Targeting

## 1. INTRODUCTION

Online social networks such as Myspace and Facebook have experienced phenomenal growth in the past two years. In November 2006, 57.2 million Internet users visited Myspace and 16.7 million visited Facebook [Comscore]. People use these sites to connect with friends and to express themselves. A typical user puts up a profile, which contains personal information ranging from gender, hometown, and political views to favorite TV shows, interests, and recent activities. In this paper, our focus is to use the profile information to predict individual users' preferences and extract relevant keywords.

We define a *predictive task* as the process of predicting an individual user's tendency to take a particular action such as responding to an advertisement, joining a group, or writing a blog entry. Let us begin with a brief motivating example of a predictive task. Figure 1 shows an example of an advertisement presented to users of Facebook. There are two questions we would like to answer with regard to this advertisement. First, who is the most likely to respond to this advertisement? Second, what are the most relevant keywords and demographics for this advertisement?

If the two questions presented above can be answered sufficiently for all such predictive tasks, a wide array of personalized applications can be realized. From the solution to the first question, we will be able to estimate users' preferences for any new feature, content or message introduced to the site. From the solution to the second question, we can match the new site introduction with users using relevant demographic information and keywords in their profiles. Additionally, we can begin to characterize the types of users who are inclined to respond. This paper aims to provide sufficient solutions to both questions. We apply our solutions empirically to advertisements, but the solutions are easily generalizable to other applications, which we discuss in the concluding section of this paper.

In our solution to the problems, we assume that some response data have been collected in the past. We have some *positive examples* from users who responded as well as some *negative examples* from users who were given the opportunity but did not respond. We refer to this data as the *training set*.

For the advertisement example, we can answer the first question by formulating it as a supervised learning problem where a binary target variable is defined by encoding positive and negative examples in the training set as value 1 and 0 respectively. The learning task aims to use all information obtained from users, primarily profile information, to predict individual users' probability of responding to the ad (i.e. having value 1). In practice, the accuracy of the probability is less of a concern than the relative ranking of the users. As long as we can identify the users who are most likely to respond to the ad and show the ad to them instead of users who are unlikely to respond, we can improve both the effectiveness of the ad as measured by click through rate and the overall user experience, since we are showing users content in which they are interested. The algorithm of identifying the top users will be detailed in Section 2.

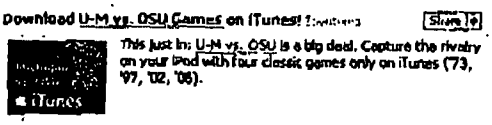


**Figure 1: An Ad Example**

To answer the second question of what the most relevant keywords or demographics are, an intuitive approach is to compute keyword histograms for both the positive and negative

Appendix

examples and then to find the keywords or demographics that demonstrate the greatest difference between these two groups. Two technical issues arise from this approach. First, how is the difference defined? Second, since the number of positive examples is usually small, the difference is often not statistically significant. We aim to address these two issues in the current paper. The details will be presented in Section 3.

The rest of this paper is organized as follows. In Section 2, we give an overview of the user-ranking algorithm. In Section 3, we present two different methods of extracting relevant keywords and briefly compare several difference measures. We then illustrate our algorithms by showing some empirical results in ad targeting in Section 4. We present our conclusions in Section 5.

## 2. THE USER-RANKING ALGORITHM

The user-ranking algorithm employs a black box model to estimate user preferences by correlating historical response data with user data such as demographics, profile information, site navigation behavior, etc. For the rest of the paper, we will refer to user data as *profile information*.

Let N denote the number of users in the training set and i denote the user index, i = 1..N. Let $y_i$ denote the target value for user i where $y_i = 1$ if user i is one of the positive examples. Otherwise, $y_i = 0$. Let $P_i$ denote the profile information for user i, which is represented as a very high dimensional vector. Typically $P_i$ will contain a small number of demographic attributes such as gender, age, etc. as well as a large number of indicator variables that indicate the presence of particular keywords. For example, an element of $P_i$ might indicate if user i has movie "Fight Club" as one of his/her favorites. In order to have a good representation of the user set, $P_i$ typically must include tens of thousands such indicator variables. Our goal is to use the attributes in $P_i$ to estimate the probability that $y_i = 1$.

With the N pairs of $(y_i, P_i)$, we would like to find a function $f$ of $P_i$ to estimate the probability that $y_i = 1$ using off-the-shelf supervised learning procedures such as SVM[], neural networks[], or boosted trees[]. No matter which learning method we choose, we will have to fit our data set into computer memory. It is not unusual that we have millions of users in the training set and tens of thousands of attribute values for each user. This results in a very high memory requirement. For example, if there are two million users in the training set, in which every user has thirty thousand attribute values in double precision floating numbers (8 bytes each), we will need 2,000,000x30,000x8 ≈ 480 gigabytes just to load the training set into memory, which is virtually infeasible in a single PC server environment. One can argue that the attributes need not to be double precision floats or the data can be represented in a sparse format to lower the memory requirement. However, in most off-the-shelf learning procedures, it is out of our control to decide how the data can be fit into memory. Therefore, we have to assume the most general conditions. There are two ways to address the memory issue, and we use both in this paper. First, in most applications we usually have far more negative examples than positive examples. We can stratify the sample, picking all the positive examples and randomly sampling a portion of the negative examples. If the size of the stratified sample is still too large, we can randomly sample from the stratified sample to get a sample of desirable sample size.

The second way to address the memory issue is through dimension reduction. Again, some off-the-shelf software such as SVD[], pLSI[], or LDA[] can be used to reduce tens of thousands of dimensions in $P_i$ to just a few hundred. The loss of information from the dimension reduction procedure is usually not a big issue as the indicator variables in $P_i$ are quite sparse in most cases. Also, it is reported in [] that a reduced dimension between 200 and 300 is sufficiently effective in most IR applications. In this paper, we use SVD for dimension reduction, although other methods would also work well.

Let $X_i$ and M denote the reduced attribute vector of user i and the sample size of the new data set after stratification, respectively. We can now apply a supervised learning procedure on these M pairs of $(y_i, X_i)$. Before proceeding, we need to determine the relative weights of the positive and negative examples. The weight parameter is often subjective, as the decision maker has to weigh the cost of false negatives against the cost of false positives. If the goal is to optimize the area under the ROC[] curve (AUC), it has been proven [] that we should make the sum of weights of positive examples equal to the sum of weights of negative examples.

We choose boosted trees as our model training method. Boosted trees have been a popular choice in supervised learning procedures since their debut in late 1990s. Compared to SVM or neural networks, boosted trees are fast and do not require normalization of attributes []. Boosted trees are also a universal function [] that can approximate any function. Finally, the output from boosted trees gives what the most important variables are for predicting the target variable. This can help us interpret the results.

After training the predictive function $f$ on $(y_i, X_i)$ using boosted trees, we can use this function to score all the registered users of the social network site. The higher the score, the more likely the user will respond to a similar message in the future. We can also rank the users according to the output scores and target future messages to the highest ranked users.

The user ranking algorithm can be summarized in the following 5 steps:

1. Convert the training set into N pairs of $(y_i, P_i)$
2. Apply dimension reduction procedures on $P_i$ to get N pairs of $(y_i, X_i)$
3. Select M pairs from (2) through stratification and random sampling
4. Train the predictive function $f$ on the M pairs of $(y_i, X_i)$ from (3) with boosted trees.
5. Score all the register users using $f$ and rank the users accordingly.

Appendix

## 3. THE KEYWORD EXTRACTION ALGORITHM

In section 2, we presented an algorithm to estimate individual users' probabilities of responding to particular messages. In addition to being able to predict the users most likely to respond to a message, it is also desirable to be able to extract insights into who these users are (e.g. what interests distinguish them from users who did not respond, and how do the users who responded differ demographically). The variable importance profile from the output model in section 2 provides a quick overview of the importance of individual variables. However, the variable importance does not give details about the correlation between user preferences and the actual attribute values. In addition, due to dimension reduction, the attributes used in the model building process are essentially un-interpretable. It is difficult to extract insights directly from the model, especially for the indicator variables.

Instead of extracting insights from the model, we can achieve the same goal by examining the data itself. One general approach is to divide the data into two sets, A and B, compute set-specific histograms of all the attributes in $P_j$, and then find the attributes with the largest differences between the two sets. The performance of this approach clearly depends on how the two sets are defined and how the difference measure is defined. We discuss each of these in turn.

One way to construct A and B is to have A be the set of positive examples and B be the set of negative examples. Another way would be to have A be the set of highest scored users from the user-ranking algorithm and B be a random set of users. Using positive and negative examples makes the most sense, intuitively. However, due to the typically low number of positive examples, we might not have enough top keywords that yield a statistically significant difference. That is why we propose the second alternative. By doing this, we can extract keywords from a bigger sample to alleviate the statistical significance problem. This approach can also be seen as a way to expand the set of relevant keywords, as the model will not only extract keywords seen frequently in the positive group but it will also extract keywords which co-occur with the these top keywords. This property is attributed to the fact that the dimension reduction procedure usually maps highly correlated variables to the same variable in the reduced variable set. More results and comparisons will be shown in section 4.4.

Before proceeding, it is worth noting that there are many other plausible ways to define A and B. For example, we could let A be a random set of users and B be the negative examples or we could let A be the set of highest scored users and B be the set of lowest scored users.]

Let $A_j$ and $B_j$ denote the percentage of users whose jth keyword indicator variable is 1 for group A and B, respectively. The difference between groups can be defined in multiple ways. Table 1 list outs some examples of the difference measures. In practice, the arithmetic difference tends to rank popular attributes too high, while the ratio difference tends to rank rare attributes too high. Empirically, the information gain measure achieves a good balance and generally considered the best difference measure [citation]. As mentioned earlier, we can use these difference measures to extract the top keywords from the computed histograms. Some results will be presented in section 4.4.

**Table 1: Examples of Difference Measures**

| Difference Name | Formula |
|---|---|
| Arithmetic Difference | $(A_j - B_j)$ |
| Ratio Difference | $(A_j / B_j)$ |
| Information Gain | $H(A_j)$-$H(B_j)$<br>H(*) denotes entropy of * |
| Odds Ratio | $(A_j(1-B_j))/((1-A_j)B_j)$ |
| Relevancy Score | $((A_j+D)/(B_j+D))$<br>D is the Laplace succession parameter |

The keyword extraction algorithm can be summarized as follows:

1. Compute histograms of each keyword indicator variable for the positive and negative groups respectively
2. For each keyword, calculate the difference between the positive and negative groups
3. Rank keywords by difference from high to low. Select top keywords as the most relevant
4. Select the highest scored users from the boosted tree model.
5. Get a random sample from the set of registered users
6. repeat 1 to 3 for the highest score group and the random group

Appendix

## 4. AD TARGETING RESULTS

### 4.1 Data

In this section, we present empirical results of the user ranking and keyword extraction algorithms. We applied our algorithms to three sponsored story ads on Facebook. The ads are embedded in the News Feed section of user home pages, as shown in Figure 2.

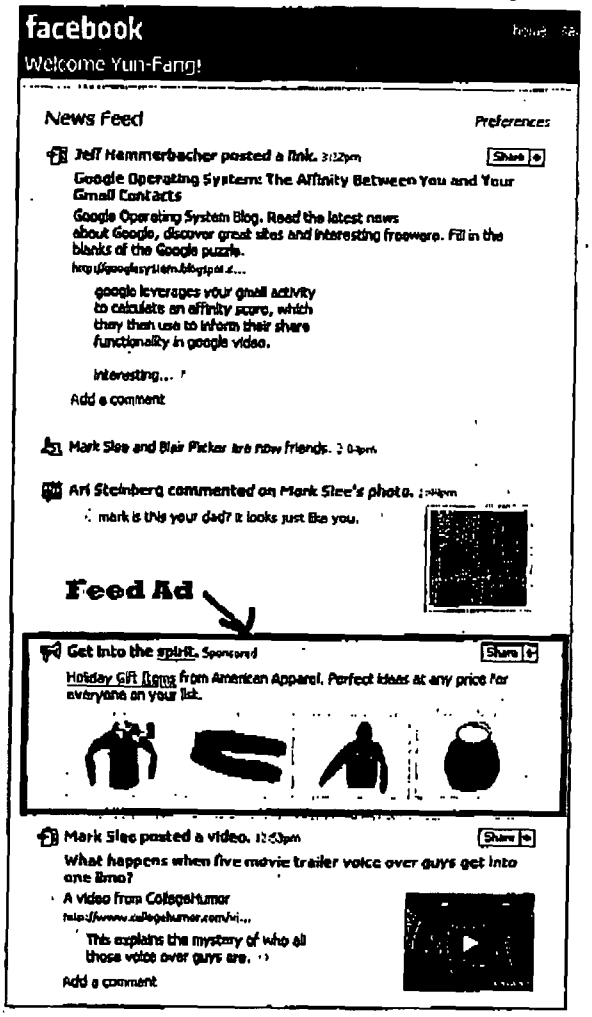


Figure 2: A Feed Ad Example

The three ads are shown in Figure 3. These ads cover three distinct interest categories—football, the movie Van Wilder 2, and digital photography. Each of these three ads is inventory constrained, so it is not feasible to show them to every user. Our goal is to target the ads to users who are most likely to respond and to be able to characterize these users. One can make conjectures about what keywords and demographic attributes will be relevant and target the users accordingly. However, this approach is not objective and might not generate enough users to target on.

The user ranking algorithm address these two issues by objectively generating an arbitrary number of users who are most likely to respond. The prerequisite for applying the algorithm to ad targeting is that we have to run the ad campaign in two stages. The first stage is the *pilot stage* where we show the ad to a random sample of users and collect the responses for our training set. Then we go through the modeling exercise, rank all the registered users from the site, and generate a list of users who are most likely to respond to the same ad in the future. The targeted user list will be used in the second *production stage* where the ad is targeted to the users ranked most highly by the model. In Section 4.2, we present the results from running our model on all three ads, including the predicted ROC curves. In Section 4.3, we present actual response rates from a live test of targeted and untargeted digital photography ads. In Section 4.4, we present the results of keyword extraction for all three ads.

In the stratified samples, there are 1534, 29417, and 2531 positive examples and 99572, 470268, and 44699 negative examples for the football, Van Wilder 2 and digital photography ads, respectively.

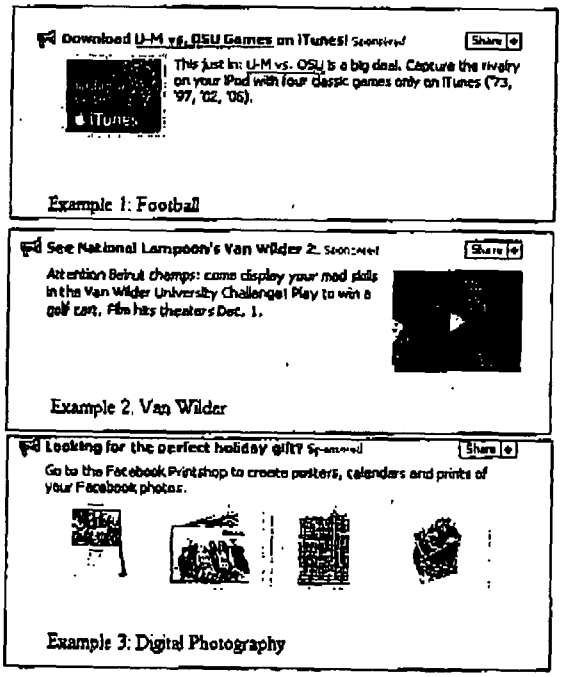


Figure 3: Ad Messages

## Appendix

**Table 2: Raw Attribute List**

| Name | Category | Type |
|---|---|---|
| Gender | Demographic | Categorical |
| Birth Year | Demographic | Numerical |
| Political Views | Demographic | Categorical |
| Relationship Status | Demographic | Categorical |
| User Type | Demographic | Categorical |
| Top 5000 Favorite Movies | Keyword Indicator | 5000 Binary |
| Top 5000 Favorite Books | Keyword Indicator | 5000 Binary |
| Top 5000 Favorite Music | Keyword Indicator | 5000 Binary |
| Top 5000 Interests | Keyword Indicator | 5000 Binary |
| Top 5000 Favorite TV shows | Keyword Indicator | 5000 Binary |

Table 2 summarizes the raw attributes ( $P_j$ ) used to predict ad responses. There are 25,000+ attributes, and most of them are indicator variables. SVD (singular vector decomposition)[] reduced these 25,000 indicator variables to 100 numerical variables. The 100 numerical SVD variables along with the demographic variables are used as predictors for the boosted trees that fit the response data we collected in the pilot phase.

For evaluation purposes, we divided the historical pilot stage data into the training set and the test set. We fit the model with the training set and back tested our model using the test set to predict future performance.

### 4.2 Model Summary

During the modeling process, we adjusted the case weights to make the sum of weights of all positive examples equivalent to sum of weights of all negative examples. As mentioned in section 2, this strategy optimizes the area under the ROC curve. This strategy does not necessarily optimize every single application with different targeting percentages. However, it is the common denominator, since the modeling process is fully automated. We set our learning rate = 0.008 and number of trees = 500 for the boosted tree procedure after some experimentation.

Table 3 lists the model performance summary of the ads from the test set. The lift score represents the ratio of the response rate with targeting to the response rate without targeting. The ROC curves are shown in Figure 4. The AUC (area under ROC curve) should be a number between 0.5 and 1. If the users are ranked randomly, the AUC score will be close to 0.5. If the users are ranked perfectly (all the positive examples are on the top), the AUC will be 1. Our goal is to maximize AUC. As shown in Table 3, the AUC scores are well above 0.5 and the responding rates for all three ads can be dramatically improved by the models with improvements ranging from 45% to 309%.

# Appendix

Table 3: Model Performance Summary

| | Lift | | |
|---|---|---|---|
| Targeting Percent | Example1 : Football | Example 2: Van Wilder | Example 3: Digital Photography |
| 5% | 4.0925 | 2.5656 | 1.4563 |
| 10% | 3.4091 | 2.3824 | 1.4401 |
| 15% | 2.9802 | 2.2805 | 1.5102 |
| 20% | 2.6515 | 2.1284 | 1.4401 |
| 25% | 2.3840 | 1.9927 | 1.4628 |
| 30% | 2.1549 | 1.8271 | 1.3484 |
| 35% | 1.9915 | 1.7022 | 1.3407 |
| 40% | 1.8876 | 1.5984 | 1.3066 |
| 50% | 1.6313 | 1.4407 | 1.2460 |
| 60% | 1.4562 | 1.3101 | 1.1893 |
| 70% | 1.2915 | 1.2145 | 1.1096 |
| 80% | 1.1648 | 1.1387 | 1.0841 |
| 90% | 1.0774 | 1.0707 | 1.0374 |
| 100% | 1.0000 | 1.0000 | 1.0000 |
| AUC | 0.7343 | 0.6708 | 0.5846 |

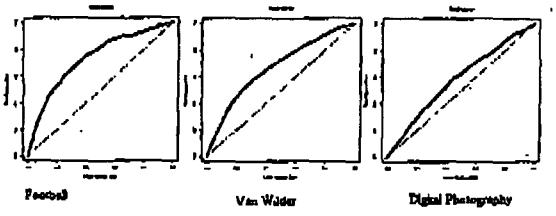

*Figure 4: ROC Curves for the Ads*

Table 4 summarizes the top 10 variables and their importance for models generated for all 3 ads. The variables with names starting with svd are variables after dimension reduction. From the results, we can see that gender and birth year are pretty effective predictors across the board while the SVD variables also play a significant role, contributing as much as 85% in the digital photography case.

Table 4: Variable Importance Profiles

| | Example1: Football | |
|---|---|---|
| Rank | Name | Importance (%) |
| 1 | Gender | 42.3305 |
| 2 | Birth Year | 5.2199 |
| 3 | svd-nl.49 | 4.3341 |
| 4 | User Type | 2.7169 |
| 5 | svd-nl.37 | 2.0727 |
| 6 | svd-nl.10 | 1.9999 |
| 7 | svd-nl.6 | 1.7548 |
| 8 | svd-nl.17 | 1.7277 |
| 9 | svd-nl.26 | 1.5955 |
| 10 | svd-nl.23 | 1.5904 |
| | Example 2: Van Wilder | |
| Rank | Name | Importance (%) |
| 1 | Gender | 73.6656 |
| 2 | svd-nl.17 | 4.1079 |
| 3 | svd-nl.10 | 3.7775 |
| 4 | svd-nl.37 | 3.2012 |
| 5 | Birth Year | 2.1210 |
| 6 | svd-nl.36 | 2.0921 |
| 7 | svd-nl.19 | 1.9901 |
| 8 | svd-nl.16 | 1.5685 |
| 9 | Relationship Status | 1.2708 |
| 10 | svd-nl.41 | 1.1262 |
| | Example 3: Digital Photography | |
| Rank | Name | Importance (%) |
| 1 | Birth Year | 6.3643 |
| 2 | svd-nl.7 | 5.3351 |
| 3 | Gender | 4.9933 |
| 4 | Relationship Status | 3.4577 |
| 5 | svd-nl.6 | 2.7879 |
| 6 | svd-nl.66 | 2.7300 |
| 7 | svd-nl.13 | 2.6770 |
| 8 | User Type | 2.6360 |
| 9 | svd-nl.24 | 2.5274 |
| 10 | svd-nl.5 | 2.4393 |

# Appendix

If we only use the demographic variables and apply decision tree modeling on the data, the model performance is consistently worse. The AUC for football, Van Wilder 2 and digital photography are 0.723, 0.645, and 0.568 respectively and lift scores at 5% are 2.93, 2.06, and 1.39, respectively. This shows that the combination of SVD and boosted trees increases performance.

## 4.3 Live Test Results

We tested the model for the digital photography ad in a live setting by showing the ad to two sets of users. The first set was the 'random' set, consisting of users randomly sampled from all the registered users on the site. The second set was the 'target' set, consisting of users ranked in the top 2.5% by the boosted tree model described in Section 4.2 The two sets of users were shown the same ad message over the same time period. The response data is shown in Table 5. Both the click through rate and the unique user click through rate on the ad were dramatically higher in the target set.

**Table 5: Live Test Result**

| | Random CTR(%) | Target CTR(%) | Relative Diff(%) | Z-statistic | p-value |
|---|---|---|---|---|---|
| CTR | 1.9761 | 3.1795 | 60.899 | 19.986 | <6.22096E-16 |
| UU CTR | 1.8082 | 2.8509 | 57.661 | 18.205 | <6.22096E-16 |

Appendix

## 4.4 Keyword Extraction

Tables 6 – 8 show the top keywords extracted from the training set and the models using information gain as the difference measure. As the results have shown, the keywords extracted are highly relevant. "Sportscenter" and "Football" are extracted from the football example while "Van Wilder" is extracted from the Van Wilder 2 example. The digital photography example doesn't generate apparently relevant examples on the top 20. However, "photography" is ranked 44 from the training set and 98 from the model, which is much higher than the original rank of 7958. There is definitely room for improvement for the ranking of the keywords. A composite rank score based on editorially judged data should give us a more objective view on how much weight we should put on each difference measure.

From the results, it seems that the keywords extracted from the training set are better than the model. This is not surprising, since the keywords are extracted directly from the observed examples. However, due to the low number of the positive examples (usually in the order of thousands), the extracted keywords are no longer statistically significant after rank 200. In applications where high coverage is important, the model will be able provide a statistically significant keyword list of any desirable size.

Table 6: Top 20 Keywords for Football

| | From Training Set | | From Model | |
|---|---|---|---|---|
| Rank | Type | Keyword | Type | Keyword |
| 1 | tv | sportscenter | tv | sportscenter |
| 2 | interests | football | interests | girls |
| 3 | interests | girls | movies | anchorman |
| 4 | movies | happy gilmore | movies | wedding crashers |
| 5 | tv | family guy | movies | friday night lights |
| 6 | movies | anchorman | tv | family guy |
| 7 | interests | basketball | movies | happy gilmore |
| 8 | music | red hot chili peppers | interests | football |
| 9 | movies | friday night lights | movies | Old |
| 10 | interests | baseball | interests | baseball |
| 11 | interests | sports | interests | sports |
| 12 | movies | wedding crashers | movies | 40 year old virgin |
| 13 | tv | south park | movies | billy madison |
| 14 | tv | 24 | tv | south park |
| 15 | tv | seinfeld | movies | remember titans |
| 16 | movies | old | interests | basketball |
| 17 | tv | lost | tv | simpsons |
| 18 | movies | remember titans | movies | dodgeball |
| 19 | music | foo fighters | tv | seinfeld |
| 20 | movies | gladiator | movies | longest yard |

Table 7: Top Keywords for Van Wilder

| | From Training Set | | From Model | |
|---|---|---|---|---|
| Rank | Type | Keyword | Type | Keyword |
| 1 | tv | family guy | Tv | family guy |
| 2 | tv | simpsons | Tv | simpsons |
| 3 | tv | scrubs | Music | Rock |
| 4 | movies | boondock saints | interests | girls |
| 5 | movies | van wilder | tv | south park |
| 6 | tv | south park | interests | football |
| 7 | tv | prison break | interests | cars |
| 8 | tv | 24 | tv | 24 |
| 9 | Tv | entourage | interests | sports |
| 10 | tv | lost | movies | boondock saints |
| 11 | interests | football | tv | lost |
| 12 | movies | gladiator | Tv | prison break |
| 13 | interests | girls | movies | wedding crashers |
| 14 | movies | snatch | movies | anchorman |
| 15 | movies | fight club | tv | futurama |
| 16 | tv | futurama | movies | gladiator |
| 17 | interests | sports | music | country |
| 18 | movies | old | movies | old |
| 19 | tv | sportscenter | movies | super troopers |
| 20 | movies | super troopers | music | rap |

Appendix

**Table 8: Top Keywords for Digital Photography**

| | From Training Set | | From Model | |
|---|---|---|---|---|
| Rank | Type | Keyword | Type | Keyword |
| 1 | tv | grey's anatomy | tv | gilmore girls |
| 2 | tv | project runway | movies | notebook |
| 3 | music | black eyed peas | music | Fray |
| 4 | tv | friends | tv | grey's anatomy |
| 5 | tv | will grace | movies | Mean girls |
| 6 | tv | desperate housewives | tv | friends |
| 7 | tv | 24 | music | jack johnson |
| 8 | tv | laguna beach | tv | laguna beach |
| 9 | music | jack johnson | music | john mayer |
| 10 | tv | hills | music | coldplay |
| 11 | tv | gilmore girls | tv | project runway |
| 12 | movies | notebook | movies | Rent |
| 13 | music | kelly clarkson | movies | Walk remember |
| 14 | music | fray | tv | desperate housewives |
| 15 | interest s | boys | music | Panic disco |
| 16 | movies | 10 things hate | books | Harry potter |
| 17 | music | beatles | music | rascal flatts |
| 18 | movies | love actually | music | goo goo dolls |
| 19 | music | death cab cutie | tv | one tree hill |
| 20 | movies | mean girls | interests | shopping |

## 5. CONCLUSIONS

This paper presents two novel personalization algorithms for online social networks. The first algorithm uses profile information to learn a function to predict the future response probabilities for individual users based on the historical response data. The second algorithm helps characterize the users who are most likely to respond to a message by extracting the most important demographics and profile keywords associated with these users. Our empirical results have shown that the algorithms are effective in improving ad response rates and can extract highly relevant keywords. As long as the response variable is well defined, the same algorithms can be used for any personalization applications in online social networks. Some other relevant applications include:

- **UI Design.** The response variable can be defined as the amount of usage of a UI element such as number of blog posts, number of mobile photo uploads, etc.
- **Content Ranking.** The response variable can be defined as the click through rate of a content element. The position of the content should be supplied as a feature in the modeling process to account for positional effects.
- **Group Recommendation.** The response variable will have value 1 if a user is a member of the interested group and 0 otherwise.

The modeling exercises on the above personalization applications will help segment users and allow sites to target existing products to fit user interests. Keyword characterization also allows us to extract insights from the models, which could give guidance for future product offerings.

## 6. REFERENCES

[1] Matrix Computations

[2] Stochastic Gradient Boosting

[3] Comscore. http://www.comscore.com/press/release.asp?id=1152

# CLAIMS

What is claimed is:

1. A method for selecting users of a web-based social network who are likely to respond to an invitation, each of the users having associated profile information, the method comprising:

selecting a plurality of pilot users from the users in the web based social network;

selecting a reduced set of keywords from the profile information for the pilot users;

sending the invitation to the pilot users;

receiving responses to the invitation from the pilot users;

classifying the received responses as either positive or negative;

creating a training set of vector pairs, each vector pair representing a pilot user and including data representing the classified response received from the pilot user and a set of training keywords selected from the reduced set of keywords and based at least in part on the associated profile information for the pilot user;

determining a function based on the training set of vector pairs that calculates from a user's profile information a likelihood that the user will respond to the invitation; and

calculating from the function a likelihood that each of one or more of the users in the web based social network will respond to the invitation.

2. The method of claim 1, further comprising ranking the one or more users based on the calculated likelihood of each user responding and sending the invitation to a predetermined number of users having the highest calculated likelihood of responding.

3. The method of claim 1, further comprising sending the invitation to users having a calculated likelihood of responding that is higher than a threshold value.

4. The method of claim 1, wherein selecting a plurality of pilot users comprises randomly selecting a predetermined number of pilot users from the users of the web based social network.

5. The method of claim 1, wherein selecting a reduced set of keywords comprises using singular vector decomposition (SVD).

6. The method of claim 1, wherein selecting a reduced set of keywords comprises using latent dirichlet allocation.

7. The method of claim 1, wherein selecting a reduced set of keywords comprises using probabilistic latent semantic indexing (PLSI).

8. The method of claim 1, wherein selecting a reduced set of keywords comprises using feature selection.

9. The method of claim 1, wherein determining a function based on the training set of vector pairs comprises using a boosted trees method.

10. The method of claim 1, wherein determining a function based on the training set of vector pairs comprises using support vector machine (SVM) analysis.

11. The method of claim 1, wherein determining a function based on the training set of vector pairs comprises using neural network analysis.

12. The method of claim 1, wherein classifying the received responses comprises classifying the received responses as positive if the user clicks on the invitation.

13. The method of claim 1, wherein classifying the received responses comprises classifying the received responses as positive if the user requests information about the invitation.

14. The method of claim 1, wherein classifying the received responses comprises classifying the received responses as positive if the user makes a purchase.

15. The method of claim 1, wherein classifying the received responses comprises classifying the received response as negative if the user does not respond to the invitation within a certain time.

16. The method of claim 1, wherein creating a training set of vectors for one or more pilot users further comprises:
stratifying the pilot users based on classified responses; and
randomly sampling the stratified pilot users.

17. The method of claim 1, wherein creating a training set of vector pairs further comprises:
selecting a first set of pilot users;
selecting a second set of pilot users;
determining a scalar based on a number of pilot users in the first set and a number of pilot users in the second set; and
multiplying each vector pair representing pilot users in the first set by the scalar.

18. The method of claim 17, wherein the scalar is about equal to a number of pilot users in the second set divided by a number of pilot users in the first set.

19. The method of claim 17, wherein the first set of pilot users is selected from pilot users providing a positive response and the second set of pilot users is selected from pilot users providing a negative response.

20. The method of claim 17, wherein the first set of pilot users is selected from pilot users providing a positive response and the second set of pilot users is selected from a random sample of all pilot users.

21. The method of claim 17, wherein the first set of pilot users is selected from pilot users providing a positive response and the second set of pilot users is selected from a random sample of pilot users providing a negative response.

22. The method of claim 1, wherein sending the invitation to a pilot user comprises embedding the invitation into a personal web page.

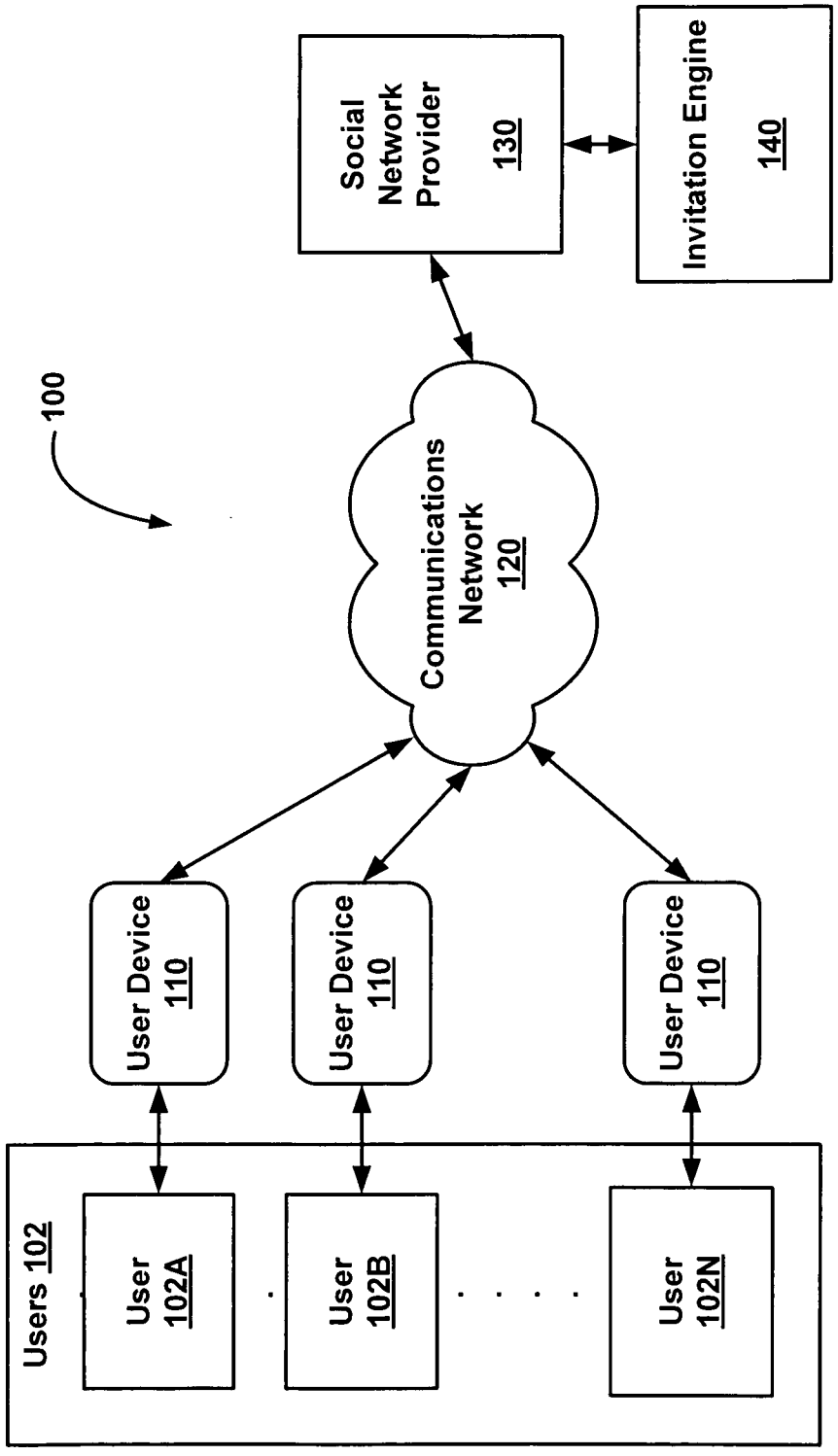
100
Social Network Provider 130
Invitation Engine 140
Communications Network 120
User Device 110
User Device 110
User Device 110
Users 102
User 102A
User 102B
User 102N

FIG. 1

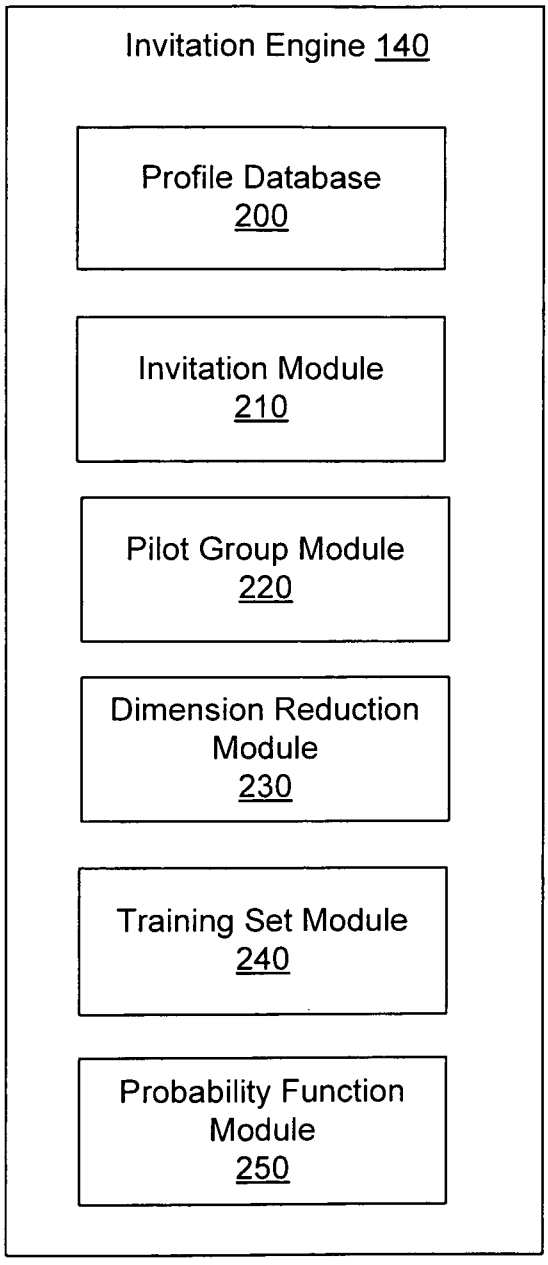
Invitation Engine 140
Profile Database 200
Invitation Module 210
Pilot Group Module 220
Dimension Reduction Module 230
Training Set Module 240
Probability Function Module 250

FIG. 2

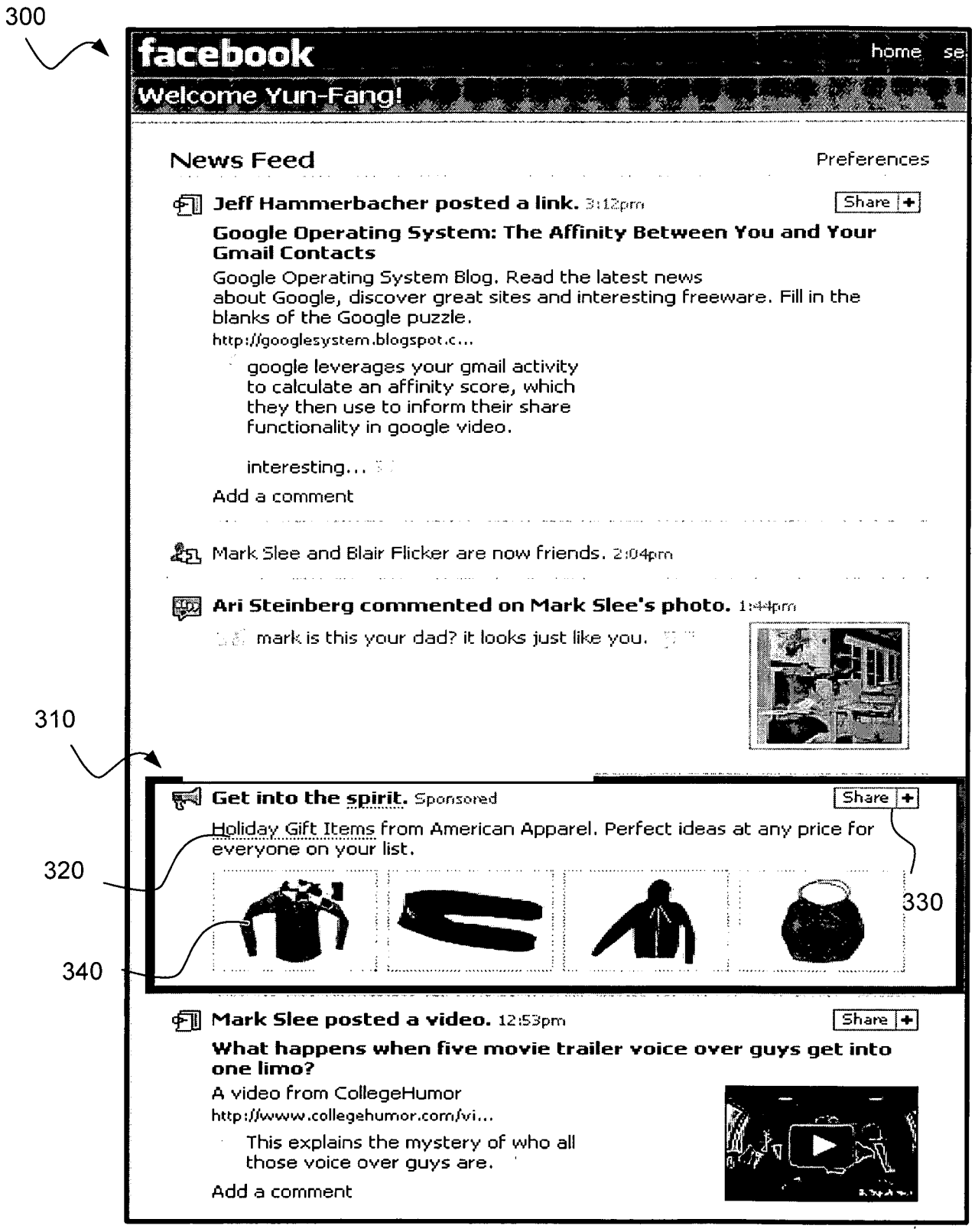
300
facebook
home
Welcome Yun-Fang!
News Feed
Preferences
Jeff Hammerbacher posted a link. 3:12pm
Share
Google Operating System: The Affinity Between You and Your Gmail Contacts
Google Operating System Blog. Read the latest news about Google, discover great sites and interesting freeware. Fill in the blanks of the Google puzzle.
http://googlesystem.blogspot.c...
google leverages your gmail activity to calculate an affinity score, which they then use to inform their share functionality in google video.
interesting...
Add a comment
Mark Slee and Blair Flicker are now friends. 2:04pm
Ari Steinberg commented on Mark Slee's photo. 1:44pm
mark is this your dad? it looks just like you.
310
Get into the spirit. Sponsored
Share
Holiday Gift Items from American Apparel. Perfect ideas at any price for everyone on your list.
320
330
340
Mark Slee posted a video. 12:53pm
Share
What happens when five movie trailer voice over guys get into one limo?
A video from CollegeHumor
http://www.collegehumor.com/vi...
This explains the mystery of who all those voice over guys are.
Add a comment

FIG. 3

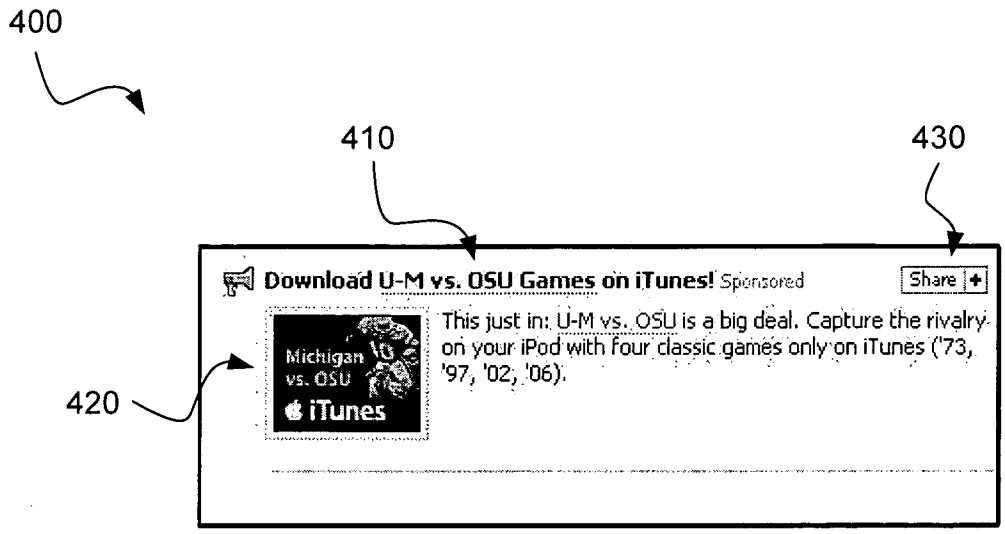
400
410
430
420
Download U-M vs. OSU Games on iTunes! Sponsored
Share
Michigan vs. OSU
iTunes
This just in: U-M vs. OSU is a big deal. Capture the rivalry on your iPod with four classic games only on iTunes ('73, '97, '02, '06).

FIG. 4

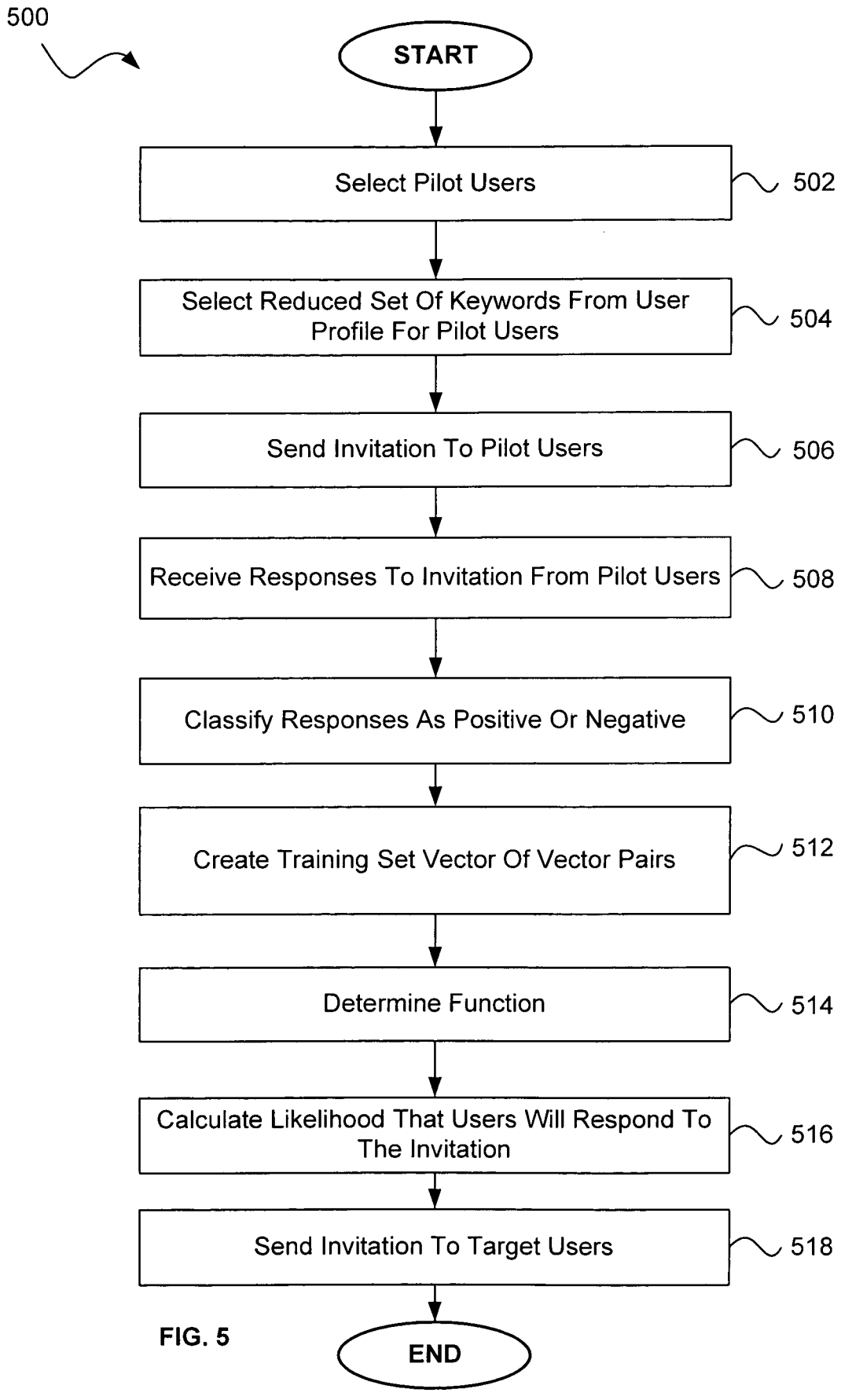
500
START
Select Pilot Users
502
Select Reduced Set Of Keywords From User Profile For Pilot Users
504
Send Invitation To Pilot Users
506
Receive Responses To Invitation From Pilot Users
508
Classify Responses As Positive Or Negative
510
Create Training Set Vector Of Vector Pairs
512
Determine Function
514
Calculate Likelihood That Users Will Respond To The Invitation
516
Send Invitation To Target Users
518
END

FIG. 5

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 08/08137

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 15/16 (2008.04)
USPC - 709/218
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC (8): G06F 15/16 (2008.04)
USPC: 709/218

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC: 706/16, 25, 45; 707/9, 10; 709/204, 219, 225, 227 (See keywords below)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Pub WEST (USPT, PGPB, JPAB, EPAB), Google Scholar.
Search Terms Used: likely, possibility, probability, predict, chance, respond, reply, react, invitation, advertisement, message, user, customer profile, pilot, target, candidate, keyword, positive, negative, classify, vector, rank, social network, singular vector.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| Y | WO 2007/070676 A2 (ZUCKERBERG et al.), 21 June 2007 (21.06.2007), entire document, especially para [0028]-[0031]; [0040]-[0042]; [0046]-[0047] and [0049]-[0050]. | 1-22 |
| Y | US 2004/0204973 A1 (WITTING et al.), 14 October 2004 (14.10.2004), entire document, especially para [0022]-[0025]; [0029]-[0033]; [0038]-[0041] and [0053]-[0055]. | 1-22 |
| Y | US 2006/0136405 A1 (DUCATEL et al.), 22 June 2006 (22.06.2006), entire document, especially para [0029]-[0030]; [0043]-[0049] and [0067]-[0070]. | 1-22 |
| Y | US 2006/0251339 A1 (GOKTURK et al.), 09 November 2006 (09.11.2006), entire document, especially para [0082]; [0085]; [0133]-[0135]; [0178]-[0180] and [0247]. | 5-7 and 10-11 |

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

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

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 26 September 2008 (26.09.2008) | 02 OCT 2008 |
| Name and mailing address of the ISA/US<br>Mail Stop PCT, Attn: ISA/US, Commissioner for Patents<br>P.O. Box 1450, Alexandria, Virginia 22313-1450<br>Facsimile No. 571-273-3201 | Authorized officer:<br>Lee W. Young<br>PCT Helpdesk: 571-272-4300<br>PCT OSP: 571-272-7774 |

Form PCT/ISA/210 (second sheet) (April 2007)