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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
10 August 2017 (10.08.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/133007 A1

(51) International Patent Classification:
G06Q 30/00 (2012.01)

(21) International Application Number:
PCT/CN2016/073690

(22) International Filing Date:
5 February 2016 (05.02.2016)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **HEWLETT PACKARD ENTERPRISE DEVELOPMENT LP** [US/US]; 3404 E. Harmony Road, Mail Stop 79, Fort Collins, Colorado 80528 (US).

(72) Inventors; and
(71) Applicants *(for US only)*: **YU, Xiao-Feng** [CN/CN]; Building 20, Universal Business Park, 10 Jiu XianQiao Road, Chaoyang District, Beijing 100015 (CN). **XIE, Jun-Qing** [CN/CN]; Building 20, Universal Business Park, 10 Jiu XianQiao Road, Chaoyang District, Beijing 100015 (CN). **GUO, Meng** [CN/CN]; 2557 Jinke Road, Shanghai 201203 (CN).

(74) Agent: **DEQI INTELLECTUAL PROPERTY LAW CORPORATION**; 7/F, Xueyuan International Tower, No. 1 Zhichun Road, Haidian District, Beijing 100083 (CN).

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

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

Published:

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

(54) **Title**: USER INTEREST AND RELATIONSHIP DETERMINATION

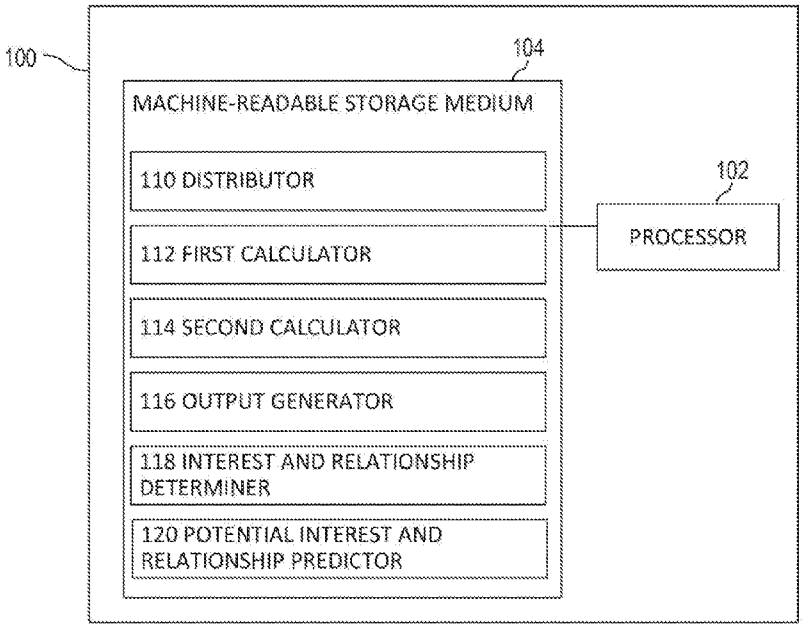

FIG. 1

(57) **Abstract**: A method for user interest and relationship determination may include distributing a first and a second set of pairs to a plurality of data nodes. The method may also include calculating, on a first data node, a probability of a user's interest in a product based on an observable factor and a latent factor and calculating, on a second data node, a probability of a likelihood of a relationship between the user and a second user, based on an observable factor and a latent factor. The method may also include determining a most likely interest and a most likely relationship of the user and predicting a potential interest of the user based on the most likely interest and the most likely relationship.

WO 2017/133007 A1

USER INTEREST AND RELATIONSHIP DETERMINATION

BACKGROUND

[0001] The advent of social networking sites on the Internet has led an unprecedented number of users registered with social networking sites to engage in interesting user activities such as commenting on, liking, and re-sharing content as well as interacting with each other to share thoughts. The exponential growth of information repositories and the diversity of users on these social networking sites provide great challenges.

BRIEF DESCRIPTION OF THE DRAWINGS

[0002] The following detailed description references the drawings, wherein:

[0003] FIG. 1 is a block diagram of an example system for user interest and relationship determination;

[0004] FIG. 2 is a flowchart of an example method for user interest and relationship determination;

[0005] FIG. 3 is a block diagram of an example system for user interest and relationship determination; and

[0006] FIG. 4 is a block diagram of an example system for user interest and relationship determination.

DETAILED DESCRIPTION

[0007] A user of a social network may have certain interests, such as products, events, items, etc. as well as connections to other people. These connections may be formally established though a direct connection or informally established. An informally established connection may between users that are connected through a third user, connected through a similar interest, connected though an action such as commenting on the same page, etc. A mutual bidirectional interaction is an action by the user that is influenced by both the user's individual interests and the user's connections.

[0008] For example, a first user may make a decision with respect to a first product based on her own interest in the first product and/or based on a second user's opinion. The opinion of the second user may be expressed as a comment on the social network, a message from the second user to the first user, an endorsement of the second user (a like, a thumbs up, etc.), etc. The first and second user may also be connected on the social network. Accordingly, the connection between the first user and the second user may be a mixture of their prior impressions to each other and their similar interests in product(s), such as the first product. The widespread social phenomenon of homophily suggests that socially acquainted users tend to behave similarly. The homophily social effect is also called the theory of "birds of a feather flock together" – people tend to follow the behaviors of their friends, and people tend to create relationships with other people who are already similar to them.

[0009] Determining the likelihood of a connection between the first user and the second user may be helpful in discovering similar interests for product recommendation. Moreover, if two users have similar interests, there may be a high likelihood of a connection between them. With the dramatically rapid growth and great success of many large-scale online social networking

services, social media establishes connections between companies and users. Tracking the data created by users on social networks may allow companies to gain feedback and insight in understanding the users' interests.

[0010] Recommending products to consumers could not only enhance revenue and profit, but also help commercial companies to understand consumers' interests and market demand. Moreover, discovering potentially valuable consumers though the connections of users on social media can aid companies in better decision making, and benefit product recommendation ultimately. The system for user interest and relationship determination leverages the bidirectional interactions between users' preferences and user-user connections in big social media and performs simultaneous user interest recommendation and connection discovery.

[0011] An example method for user interest and relationship determination may include distributing a first set of pairs and a second set of pairs to a plurality of data nodes, wherein each pair in the first set of pairs is of a user of a social network and a product on the social network and each pair in the second set of pairs defines a connection between users on the social network. The method may also include calculating, on a first data node belonging to the plurality, a first probability of a first user's interest in a first product based on a first observable factor and a first latent factor, wherein the first user and the first product belong to a first pair from the first set of pairs. The method may also include calculating, on a second data node, a second probability of a likelihood of a relationship between the first user and a second user of the social network, based on a second observable factor and a second latent factor, wherein the first user and the second user belong to a second pair from the second set of pairs. The method may also include determining, based on the first probability and the second probability, a most likely interest of the first user and a most likely relationship of the first user and predicting a potential interest of the first user based on the most likely interest and the most likely relationship.

[0012] FIG. 1 is a block diagram of an example system 100 for user interest and relationship determination. System 100 may include a processor

102 and a memory 104 that may be coupled to each other through a communication link (e.g., a bus). Processor 102 may include a Central Processing Unit (CPU) or another suitable hardware processor. In some examples, memory 104 stores machine readable instructions executed by processor 102 for system 100. Memory 104 may include any suitable combination of volatile and/or non-volatile memory, such as combinations of Random Access Memory (RAM), Read-Only Memory (ROM), flash memory, and/or other suitable memory. Memory 104 may also include a random access non-volatile memory that can retain content when the power is off.

[0013] Memory 104 stores instructions to be executed by processor 102 including instructions for and/or other components. According to various implementations, user interest and relationship determination system 100 may be implemented in hardware and/or a combination of hardware and programming that configures hardware. Furthermore, in FIG. 1 and other Figures described herein, different numbers of components or entities than depicted may be used.

[0014] Processor 102 may execute instructions of distributor 110 to distribute a first set of pairs and a second set of pairs to a plurality of data nodes. A data node stores data in the file system. The set of pairs includes any number of pairs. Each pair in the first set of pairs may be of a user of a social network and an interest of the user on the social network. Interests may include products, events, items, etc. Each pair in the second set of pairs may define a connection between users on the social network. The connection may be a direct connection or an indirect connection. An indirect connection may be between users that are connected through a third user, connected through a similar interest, connected though activities, such as commenting on the same page, etc.

[0015] The first pair and the second pair may be used as a first input key and a second input key, respectively, for a map function. A first observable factor and a first latent factor may be used as values for the first input key. A second observable factor and a second latent factor may be used as values for the second input key.

[0016] Distributor 110 may distribute the first and second set of pairs using a distributed data processing framework. Distributor 110 may distribute each pair in the first set of pairs and the second pairs to a plurality of data nodes. Each data node in the plurality of data nodes may processes a pair. One example framework is the Apache™ Hadoop® framework that allows for the scalable parallel and distributed computing of large data sets across clusters of computers using programming models such as MapReduce. Hadoop® consists of two layers: a data storage layer Hadoop Distributed File System and a data processing layer called MapReduce framework. The MapReduce framework adopts a master-slave architecture which consists of one master node and multiple slave nodes in the clusters. The master node is generally served as JobTracker and each slave node is generally served as TaskTracker.

[0017] Distributor 110 may also use a MapReduce programming technique. MapReduce is based on two functions: Map and Reduce. The Map function applies a user-defined function to each key-value pair < input key; input value > in the input data. The result of the map function may be a list of intermediate key-value pairs, sorted and grouped by key (i.e. list[< map key; map value >]), and passed as input to the Reduce function. The Reduce function applies a second user-defined function to the intermediate key and its associated values (i.e. < map key; list [map value] >), and produces the final aggregated result [< output key; output value >].

[0018] MapReduce may utilize a distributed file system from which the Map instances retrieve the input. An example distributed file system is the Hadoop Distributed File System (HDFS). HDFS is a chunk-based distributed file system that supports fault-tolerance by data partitioning and replication.

[0019] Processor 102 may execute instructions of first calculator 112 to calculate, on a first data node, a first probability of a first user's interest in a first interest based on a first observable factor and a first latent factor. An observable factor may be historical information corresponding to a user. For example, observable factors may include a user's registered data, user's behavioral data, etc. A latent factor is information corresponding to user

interactions between connections to interests. Latent factors are usually implicit and/or hidden and are thus unobservable. The first user and the first product may belong to a first pair from the first set of pairs (e.g. as discussed in reference to distributor 110). The first pair may be used as an input key for a map function. The first observable factor and the first latent factor may be used as values for the first input key. For example, the map key for the first data node may be the user-interest pair < i; j >. The value for the map key may be the product of observable and latent factors $\varphi\varphi_h$ for < i; j >.

[0020] Processor 102 may execute instructions of second calculator 114 to calculate, on a second data node, a second probability of a likelihood of a relationship between the first user and a second user based on a second observable factor and a second latent factor. The first user and the second user belong to a second pair from the second set of pairs (e.g. as discussed in reference to distributor 110). The second pair may be used as an input key for a map function. The second observable factor and the second latent factor may be used as values for the second input key. For example, the map key may be the product of the user-user pair < i; k >. The value for the map key may be the product of product of observable and latent factors $\varphi'\varphi'_h$ for < i; k >.

[0021] Processor 102 may execute instructions of output generator 116 to generate, based on the first probability and the second probability, a triplet. The triplet may be the output key of a map function. The value of the output key may be a product of probability distribution Y_{ij} S_{ik}. The triplet may be a user-interest-user triplet <i, j, k>. The triplet may include two users from the social network and a product that at least one of the two users has expressed interest in on the social network. Output generator 116 may determine a probability distribution of the first user's interest in the first product and the relationship between the first user and the second user.

[0022] Output generator 116 may incorporate a mutual latent random graphs (MLRGs) that incorporates the interactions between users' interests and users' connections. The MLRG may incorporate shared latent factors and coupled models to encode users' interests Y_{ij} (user i's interest in product j) and user-user connections S_{ik} (connection between user *i* and user *k*). Output

generator 116 may express the probability distribution of Y_{ij} as $Y_{ij} \sim p(\varphi\varphi_h,\theta)$, with θ representing any corresponding parameters. The expression may include an assumption that certain observable factors (φ) exist and certain latent factors (φ_h) exist. Output generator 116 may express the probability distribution of S_{ik} as $S_{ik} \sim p(\varphi'\varphi'_h,\Omega)$, with Ω representing any corresponding parameters. The expression may include an assumption that certain observable factors (φ') exist and certain latent factors (φ'_h) exist. Importantly, both φ_h and φ'_h may capture bidirectional interactions between interests and connections.

[0023] The four factors φ, φ_h, φ', φ'_h can be instantiated in different ways. Each factor may be defined as the exponential family of an inner product over sufficient statistics (feature functions) and corresponding parameters. Each factor may be a clique template whose parameters are tied. More specifically, the factors may be defined as:

[0024] Equation (1): $$\varphi = \exp\{\sum \bar{\alpha} f\}$$

[0025] Equation (2): $$\varphi_h = \exp\{\sum \bar{\beta} g\}$$

[0026] Equation (3): $$\varphi' = \exp\{\sum \bar{\gamma} h\}$$

[0027] Equation (4): $$\varphi'_h = \exp\{\sum \bar{\delta} q\}$$

[0028] $\bar{\alpha}$,$\bar{\beta}$, $\bar{\gamma}$, and $\bar{\delta}$ may be real-valued weighting vectors and f, g, h and q may be corresponding vectors of sufficient statistics (feature functions).

[0029] In other words, a map function may involve calculating probability distributions on data nodes in parallel (e.g. as discussed as discussed in reference to first calculator 112 and second calculator 114) and generating triplet product of probability distribution Y_{ij} S_{ik} (as discussed in reference to output generator 116). Each data node may calculate the probability distribution $Y_{ij} \sim p(\varphi\varphi_h, \theta)$ and the probability distribution $S_{jk} \sim p(\varphi'\varphi'_h, \Omega)$. This process may be repeated until a convergence occurs.

[0030] The probability distribution Y_{ij} may be calculated as:

[0031] Equation (5): $$Y_{ij} \sim p(\varphi\varphi_h, \theta) = \frac{1}{z_1}\exp\{\sum \bar{\alpha} f + \sum \bar{\beta} g\}$$

[0032] Similarly, the probability distribution may be calculated as:

[0033] Equation (6): $S_{jk} \sim p(\varphi'\varphi'_h, \Omega) = \frac{1}{Z_2}\exp\{\sum \bar{\gamma}h + \sum \bar{\delta}q\}$

[0034] In equation (5) above, $\theta = \{\bar{\alpha}, \bar{\beta}\}$ may be the parameter vector for Y_{ij}, and in equation (6), $\Omega = \{\bar{\gamma}, \bar{\delta}\}$ may be the parameter vector for S_{ik}. Both $Z1$ and $Z2$ are the normalization factors for Y_{ij} and S_{ik}, respectively. Thus the joint probability distribution of the mutual latent random graphs (MLRGs) can be formally defined as expressed in equation (7) below, where $Z = Z1 \cdot Z2$ is the normalization factor of MLRGs.

[0035] Equation (7):

$$(Y_{ij}, S_{jk}) \sim Y_{ij} \cdot S_{jk} \sim p(\varphi\varphi_h, \theta) \cdot p(\varphi'\varphi'_h, \Omega) = \frac{1}{Z}\exp\{\sum \bar{\alpha}f + \sum \bar{\beta}g + \sum \bar{\gamma}h + \sum \bar{\delta}q\}$$

[0036] Processor 102 may execute instructions of interest and relationship determiner 118 to determine, based on the first probability and the second probability, a most likely interest of the first user and/or a most likely relationship of the first user. A triplet (e.g. as discussed in reference to output generator 116) may be used as an input key for a reduce function. A probability distribution and/or a product of probability distribution Y_{ij} S_{jk} may be used as values for the input key for the reduce function. Interest and relationship determiner 118 may merge a result of processing by the plurality of data nodes (e.g. as discussed in reference to distributor 110) using the triplet (e.g. as discussed in reference to output generator 116) as a key so that all values using the same triplet are grouped together.

[0037] Interest and relationship determiner 118 may determine the most likely interest of the first user and the most likely relationship of the first user as an output of the reduce function. An output key for the output of the reduce function may be an objective function $\mathcal{L}(\theta, \Omega)$. The value for the output key may be updated and optimized parameters θ and Ω. Interest and relationship determiner 118 may maximize an objective function

corresponding to the triplet. A first parameter of the objective function may correspond to the most likely interest of the first user and a second parameter of the objective function may correspond to the most likely relationship of the first user. The objective function may be maximized using a data mining algorithm, such as stochastic gradient descent.

[0038] A data mining algorithm (such as a stochastic gradient descent) may be performed with respect to θ with Ω fixed and Ω may be updated. A data mining algorithm (such as a stochastic gradient descent) may be performed with respect to Ω with θ fixed and θ may be updated. This process may be repeated until a convergence occurs.

[0039] Stochastic gradient descent (SGD) may loop over all the observations and update the parameters θ and Ω by moving in the direction defined by negative gradient. Each data node (e.g. as discussed in reference to first calculator 112 and second calculator 114), may compute and optimize with respect to either Y_{ij} or S_{ik} in the Map phase, and the results may be combined in a reduce phase to optimize both parameters θ and Ω globally. After distributed SGD learning, the optimized parameters can be obtained and joint recommendation of interest and friendship can be achieved by computing the most likely Y_{ij} or S_{ik}, respectively.

[0040] In other words, the reduce function may include calculating the objective function $\mathcal{L}(\theta, \Omega)$ and updating all parameters on a master node. The master node may calculate and maximize the objective function $\mathcal{L}(\theta, \Omega)$. The master node may update and optimize the parameters (θ, Ω) such that $(\theta^*, \Omega^*) = \arg\max \mathcal{L}(\theta, \Omega)$.

[0041] After stochastic gradient descent (SGD) for distributed MapReduce learning, an optimized θ and Ω of MLRGs may be obtained. The optimized parameters θ and Ω may be used to discover user interest and infer user-user friendship. More specifically, given the testing social media data, the inference may find the most likely types of user interest and corresponding

user-user relationship labels that have the maximum posterior probability. This can be accomplished by performing the model inference of MLRGs. Performing the model inference may include predicting the labels of user interest and user-user friendship by finding the maximum a posterior (MAP) user interest labeling assignment and corresponding user-user friendship labeling assignment that have the largest marginal probability according to equations (5) and (6) described above.

[0042] The overall MapReduce processing of the user interest and relationship determination system may be summarized as follows. Each processing job in may be broken down to as many Map tasks as input data blocks and one or more Reduce tasks. A master node may select idle workers (data nodes) and may assigns each data node a map or a reduce task according to the stage. Before starting the Map task, an input file may be loaded on the distributed file system. At loading, the file may partitioned into multiple data blocks of the same size. One example size of a data block may be 64MB. Each block may be triplicated for fault-tolerance. Each block may also be assigned to a mapper, a worker which is assigned a map task, and the mapper may applies a map function (Map()) to each record in the data block.

[0043] The intermediate outputs produced by the mappers may be sorted locally for grouping key-value pairs sharing the same key. After local sort, a combine function (Combine()) may be applied to perform pre-aggregation on the grouped key-value pairs so that the communication cost taken to transfer all the intermediate outputs to reducers is minimized. Then the mapped outputs may be stored in local disks of the mappers, partitioned into R, where R is the number of Reduce tasks in the MR job. This partitioning may be done by a hash function e.g. hash(key) mod R.

[0044] When all Map tasks are completed, the MapReduce scheduler may assign Reduce tasks to workers. The intermediate results

may be shuffled and assigned to reducers via HTTPS protocol. Since all mapped outputs may already be partitioned and stored in local disks, each reducer may perform the shuffling by simply pulling its partition of the mapped outputs from mappers. Put another way, each record of the mapped outputs may be assigned to only a single reducer by one-to-one shuffling strategy. Note that this data transfer may be performed by reducers' pulling intermediate results. A reducer may read the intermediate results and merge them by the intermediate keys, i.e. map key, so that all values of the same key are grouped together. The grouping may be done by external merge-sort. Each reducer may also apply a reduce function (Reduce()) to the intermediate values for each map key it encounters. The output of reducers may be stored and triplicated in the file system.

[0045] The number of Map tasks may not depend on the number of nodes, but may be based on the number of input blocks. Each block may be assigned to a single Map task. However, all Map tasks do not need to be executed simultaneously and neither do all Reduce tasks. The MapReduce framework may executes tasks based on runtime scheduling scheme. In other words, MapReduce may not build any execution plan that specifies which tasks will run on which nodes before execution.

[0046] With the runtime scheduling, MapReduce may achieve fault tolerance by detecting failures and reassigning tasks of failed nodes to other healthy nodes in the cluster. Nodes which have completed their tasks may be assigned another input block. This scheme naturally achieves load balancing in that faster nodes will process more input chunks and slower nodes process less inputs in the next wave of execution. Furthermore, a MapReduce scheduler may utilize a speculative and redundant execution. Tasks on straggling nodes may be redundantly executed on other idle nodes that have finished their assigned tasks, although the tasks are not guaranteed to end earlier on the new assigned nodes than on the straggling nodes. Map and

Reduce tasks may be executed with no communication between other tasks. Thus, there is no contention arisen by synchronization and no communication cost between tasks during a MR job execution.

[0047] An example architecture for the user interest and relationship determination system 100 may exploit Extraction-Transformation-Loading (ETL) technology for heterogeneous (structured and unstructured) big social data to the data storage layer. An example storage layer may include a relational database management system (RDBMS), a NoSQL database management system and logs of social media data. The architecture may also include server-based tool designed to transfer data between Hadoop and relational databases. Example tools may include the Sqoop2™ system (from Cloudera™), MongoDB connector™ (from MongoDB, Inc.) and Flume4™ (from Apache™) to transfer the RDBMS, NoSQL and Log data to the joint recommender layer for distributed analysis respectively. Sqoop2 is a tool designed for transferring bulk data between Hadoop and structured data stores such as relational databases. The MongoDB connector™ is a plugin for Hadoop™ that provides the ability to use MongoDB™ as an input source and/or an output destination. Flume™ is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It uses a simple extensible data model that allows for online analytic application. The joint recommender layer may consists of a data model storing rich social information and a joint recommender engine for MLRGs and advanced MapReduce learning.

[0048] Processor 102 may execute instructions of potential interest and relationship predictor 120 to predict a potential interest of the first user and/or a potential relationship between the first user and a user of the social network based on the most likely interest and the most likely relationship.

[0049] FIG. 2 is a flowchart of an example method 200 for user interest and relationship determination. Method 200 may be described below as being executed or performed by a system, for example, system 100 of FIG. 1, system 300 of FIG. 3 or system 400 of FIG. 4. Other suitable systems and/or computing devices may be used as well. Method 200 may be implemented in the form of executable instructions stored on at least one machine-readable storage medium of the system and executed by at least one processor of the system. The processor may include a Central Processing Unit (CPU) or another suitable hardware processor. The machine-readable storage medium may be non-transitory. Method 200 may be implemented in the form of electronic circuitry (e.g., hardware). At least one block of method 200 may be executed substantially concurrently or in a different order than shown in FIG. 2. Method 200 may include more or less blocks than are shown in FIG. 2. Some of the blocks of method 200 may, at certain times, be ongoing and/or may repeat.

[0050] Method 200 may start at block 202 and continue to block 204, where the method may include distributing a first set of pairs and a second set of pairs to a plurality of data nodes. Each pair in the first set of pairs may be of a user of a social network and a product on the social network. Each pair in the second set of pairs may define a connection between users on the social network. A first pair from the first set of pairs and a second pair from the second set of pairs may be used as a first input key and a second input key, respectively, for a map function. A first observable factor and a first latent factor may be used as values for the first input key. A second observable factor and a second latent factor may be used as values for the second input key. At block 206, the method may include calculating, on a first data node belonging to the plurality of data nodes, a first probability of a first user's interest in a first product based on a first observable factor and a first latent factor. The first user and the first product belong to a first pair from the first set of pairs.

[0051] At block 208, the method may include calculating, on a second data node, a second probability of a likelihood of a relationship between the first user and a second user, based on a second observable factor and a second latent factor. The first user and the second user belong to a second pair from the second set of pairs. At block 210, the method may include determining, based on the first probability and the second probability, a most likely interest of the first user and a most likely relationship of the first user. At block 212, the method may include predicting a potential interest of the first user based on the most likely interest and the most likely relationship. The method may also include predicating a potential relationship between the first user and another user of the social network based on the most likely interest and the most likely relationship. Method 200 may eventually continue to block 214, where method 200 may stop.

[0052] FIG. 3 is a block diagram of an example system 300 for user interest and relationship determination. System 300 may include a processor 302 and a memory 304 that may be coupled to each other through a communication link (e.g., a bus). Processor 302 may include a Central Processing Unit (CPU) or another suitable hardware processor. In some examples, memory 304 stores machine readable instructions executed by processor 302 for operating system 300. Memory 304 may include any suitable combination of volatile and/or non-volatile memory, such as combinations of Random Access Memory (RAM), Read-Only Memory (ROM), flash memory, and/or other suitable memory.

[0053] Memory 304 stores instructions to be executed by processor 302 including instructions for a first probability calculator 308, a second probability calculator 310, an interest and relationship determiner 312, a triplet generator 314 and an interest and relationship predictor 316. The components of system 300 may be implemented in the form of executable instructions stored on at least one machine-readable storage medium of system 300 and executed

by at least one processor of system 300. The machine-readable storage medium may be non-transitory. Each of the components of system 300 may be implemented in the form of at least one hardware device including electronic circuitry for implementing the functionality of the component.

[0054] Processor 302 may execute instructions of first probability calculator 308 to calculate, on a first data node, a first probability of a first user's interest in a first product based on a first observable factor and a first latent factor. The first user and the first product may be used as a first input key. The first user and the second user may be used as a second input key for a map function. A first observable factor and a first latent factor may be used as values for the first input key. A second observable factor and a second latent factor are used as values for the second input key. Processor 302 may execute instructions of second probability calculator 310 to calculate, on a second data node, a second probability of a likelihood of a relationship between the first user and a second user based on a second observable factor and a second latent factor. Processor 302 may execute instructions of interest and relationship determiner 312 to determine, based on the first probability and the second probability, a most likely interest of the first user and a most likely relationship of the first user.

[0055] Processor 302 may execute instructions of triplet generator 314 to generate, based on the first probability and the second probability, a triplet including two users from the social network and a product that at least one of the two users has expressed interest in on the social network. Processor 302 may execute instructions of an interest and relationship predictor 316 predict a potential interest of the first user and/or a potential relationship of the first user to another user on the social network based on the most likely interest and the most likely relationship.

[0056] FIG. 4 is a block diagram of an example system 400 for user interest and relationship determination. System 400 may be similar to system

100 of FIG. 1, for example. In the example illustrated in FIG. 4, system 400 includes a processor 402 and a machine-readable storage medium 404. Although the following descriptions refer to a single processor and a single machine-readable storage medium, the descriptions may also apply to a system with multiple processors and multiple machine-readable storage mediums. In such examples, the instructions may be distributed (e.g., stored) across multiple machine-readable storage mediums and the instructions may be distributed (e.g., executed by) across multiple processors.

[0057] Processor 402 may be at least one central processing unit (CPU), microprocessor, and/or other hardware devices suitable for retrieval and execution of instructions stored in machine-readable storage medium 404. In the example illustrated in FIG. 5, processor 402 may fetch, decode, and execute instructions 406, 408, 410, 412 and 414 to perform user interest and relationship determination. Processor 402 may include at least one electronic circuit comprising a number of electronic components for performing the functionality of at least one of the instructions in machine-readable storage medium 404. With respect to the executable instruction representations (e.g., boxes) described and shown herein, it should be understood that part or all of the executable instructions and/or electronic circuits included within one box may be included in a different box shown in the figures or in a different box not shown.

[0058] Machine-readable storage medium 404 may be any electronic, magnetic, optical, or other physical storage device that stores executable instructions. Thus, machine-readable storage medium 404 may be, for example, Random Access Memory (RAM), an Electrically-Erasable Programmable Read-Only Memory (EEPROM), a storage drive, an optical disc, and the like. Machine-readable storage medium 404 may be disposed within system 400, as shown in FIG. 4. In this situation, the executable instructions may be "installed" on the system 400. Machine-readable storage medium 404

may be a portable, external or remote storage medium, for example, that allows system 400 to download the instructions from the portable/external/remote storage medium. In this situation, the executable instructions may be part of an "installation package". As described herein, machine-readable storage medium 404 may be encoded with executable instructions for context aware data backup. The machine-readable storage medium may be non-transitory.

[0059] Referring to FIG. 4, pair distribute instructions 406, when executed by a processor (e.g., 402), may cause system 400 to distribute a first set of pairs and a second set of pairs to a plurality of data nodes. Each pair in the first set of pairs may be of a user of a social network and a product on the social network. Each pair in the second set of pairs may define a connection between users on the social network. A first pair from the first set of pairs and a second pair from the second set of pairs may be used as a first input key and a second input key, respectively, for a map function. A first observable factor and a first latent factor may be used as values for the first input key. A second observable factor and a second latent factor are used as values for the second input key.

[0060] Probability determine instructions 408, when executed by a processor (e.g., 402), may cause system 400 to determine, on the plurality of data nodes, a probability distribution of a first user's interest in a first product and a relationship between the first user and a second user. The probability may be based on an observable factor and a latent factor. Triplet generate instructions 410, when executed by a processor (e.g., 402), may cause system 400 to generate, based on the probability distribution, a triplet including two users from the social network and an interest product that at least one of the two users has expressed interest in on the social network. Most likely interest and relationship determine instructions 412, when executed by a processor (e.g., 402), may cause system 400 to determine, based on the probability distribution, a most likely interest of the first user and a most likely relationship

of the first user. Potential interest and relationship predict instructions 414, when executed by a processor (e.g., 402), may cause system 400 to predict a potential interest of the first user and/or a potential relationship between the first user and another user of the social network based on the most likely interest and the most likely relationship.

[0061] The foregoing disclosure describes a number of examples for user interest and relationship determination. The disclosed examples may include systems, devices, computer-readable storage media, and methods for user interest and relationship determination. For purposes of explanation, certain examples are described with reference to the components illustrated in FIGS. 1-4. The functionality of the illustrated components may overlap, however, and may be present in a fewer or greater number of elements and components. Further, all or part of the functionality of illustrated elements may co-exist or be distributed among several geographically dispersed locations. Further, the disclosed examples may be implemented in various environments and are not limited to the illustrated examples.

[0062] Further, the sequence of operations described in connection with FIGS. 1-4 are examples and are not intended to be limiting. Additional or fewer operations or combinations of operations may be used or may vary without departing from the scope of the disclosed examples. Furthermore, implementations consistent with the disclosed examples need not perform the sequence of operations in any particular order. Thus, the present disclosure merely sets forth possible examples of implementations, and many variations and modifications may be made to the described examples.

CLAIMS

1. A method comprising:

distributing a first set of pairs and a second set of pairs to a plurality of data nodes, wherein each pair in the first set of pairs is of a user of a social network and a product on the social network and each pair in the second set of pairs defines a connection between users on the social network;

calculating, on a first data node belonging to the plurality of data nodes, a first probability of a first user's interest in a first product based on a first observable factor and a first latent factor, wherein the first user and the first product belong to a first pair from the first set of pairs;

calculating, on a second data node, a second probability of a likelihood of a relationship between the first user and a second user, based on a second observable factor and a second latent factor, wherein the first user and the second user belong to a second pair from the second set of pairs;

determining, based on the first probability and the second probability, a most likely interest of the first user and a most likely relationship of the first user; and

predicting a potential interest of the first user based on the most likely interest and the most likely relationship.

2. The method of claim 2 wherein the first pair and the second pair are used as a first input key and a second input key, respectively, for a map function, the first observable factor and the first latent factor are used as values for the first input key and the second observable factor and the second latent factor are used as values for the second input key.

3. The method of claim 2, further comprising

generating, based on the first probability and the second probability, a triplet including two users from the social network and a product that at least one of the two users has expressed interest in on the social network.

4. The method of claim 3, further comprising:

maximizing an objective function corresponding to the triplet, wherein a first parameter of the objective function corresponds to the most likely interest of the first user and a second parameter of the objective function corresponds to the most likely relationship of the first user.

5. The method of claim 4, wherein the objective function is maximized using a stochastic gradient descent.

6. The method of claim 1 further comprising:

determining a probability distribution of the first user's interest in the first product and the relationship between the first user and the second user.

7. The method of claim 6, wherein a user-interest-user triplet is used as an input key for a reduce function and the probability distribution is used as a value for the input key.

8. The method of claim 7, further comprising:

distributing each pair in the first set of pairs and the second pairs to the plurality of data nodes, wherein each data node in the plurality of data nodes processes a pair; and

merging a result of processing by the plurality of data nodes using the triplet as a key so that all values using the same triplet are grouped together.

9. A system comprising:

a first probability calculator to calculate, on a first data node, a first probability of a first user's interest in a first product based on a first observable factor and a first latent factor;

a second probability calculator to calculate, on a second data node, a second probability of a likelihood of a relationship between the first user and a second user based on a second observable factor and a second latent factor;

an interest and relationship determiner to determine, based on the first probability and the second probability, a most likely interest of the first user and a most likely relationship of the first user;

a triplet generator to generate, based on the first probability and the second probability, a triplet including two users from the social network and a product that at least one of the two users has expressed interest in on the social network; and

a relationship predictor to predict a potential relationship of the first user based on the most likely interest and the most likely relationship.

10. The system of claim 9 wherein the first user and the first product are used as a first input key and the first user and the second user are used as a second input key for a map function, the first observable factor and the first latent factor are used as values for the first input key and the second observable factor and the second latent factor are used as values for the second input key.

11. The system of claim 9 wherein the triplet is used as an input key for a reduce function and a value for the input key is a probability distribution of the first user's interest in the first product and the relationship between the first user and the second user.

12. A non-transitory machine-readable storage medium encoded with instructions, the instructions executable by a processor of a system to cause the system to:

distribute a first set of pairs and a second set of pairs to a plurality of data nodes, wherein each pair in the first set of pairs is of a user of a social network and a product on the social network and each pair in the second set of pairs defines a connection between users on the social network;

determine, on the plurality of data nodes, a probability distribution of a first user's interest in a first product and a relationship between the first user and a second user, wherein the probability is based on an observable factor and a latent factor;

generate, based on the probability distribution, a triplet including two users from the social network and an interest product that at least one of the two users has expressed interest in on the social network.

determine, based on the probability distribution, a most likely interest of the first user and a most likely relationship of the first user; and

predict a potential interest of the first user based on the most likely interest and the most likely relationship.

13. The non-transitory machine-readable storage medium of claim 12 wherein the triplet is used as an input key for a reduce function and the probability distribution is used as a value for the input key.

14. The non-transitory machine-readable storage medium of claim 12, wherein the instructions executable by the processor of the system further cause the system to:

maximize an objective function corresponding to the triplet, wherein a first parameter of the objective function corresponds to the most likely interest of

the first user and a second parameter of the objective function corresponds to the most likely relationship of the first user.

15. The non-transitory machine-readable storage medium of claim 12, wherein the instructions executable by the processor of the system further cause the system to:

distribute each pair in the first set of pairs and the second pairs to the plurality of data nodes, wherein each data node in the plurality of data nodes processes a pair; and

merge a result of processing by the plurality of data nodes using the triplet as a key so that all values using the same triplet are grouped together.

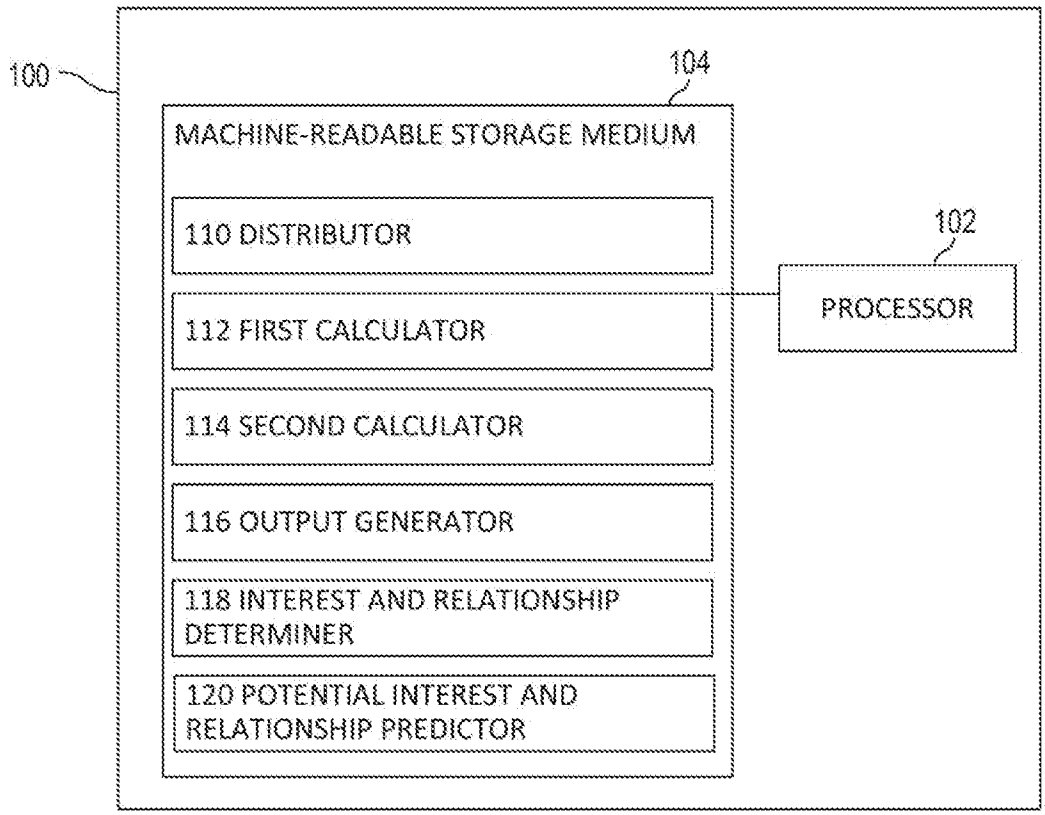
100
104
MACHINE-READABLE STORAGE MEDIUM
110 DISTRIBUTOR
112 FIRST CALCULATOR
114 SECOND CALCULATOR
116 OUTPUT GENERATOR
118 INTEREST AND RELATIONSHIP DETERMINER
120 POTENTIAL INTEREST AND RELATIONSHIP PREDICTOR
102
PROCESSOR

FIG. 1

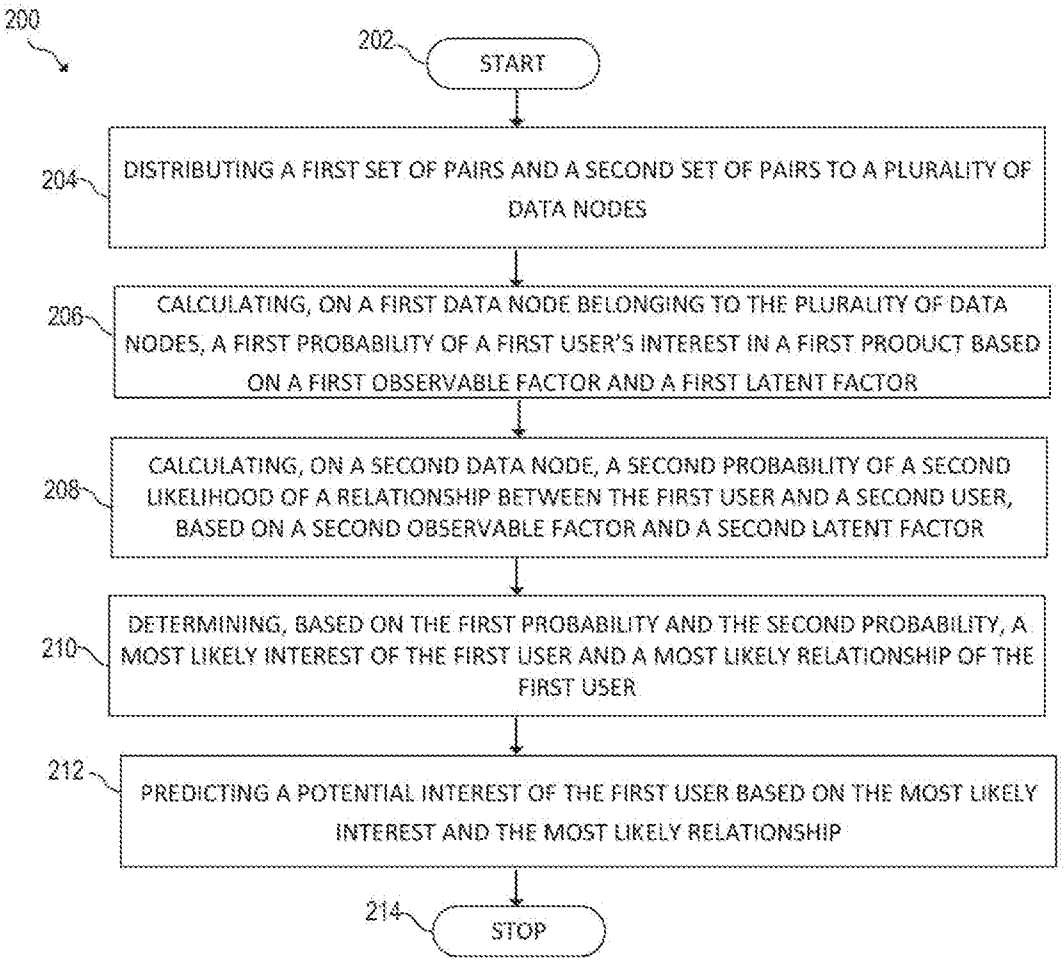
200
202
START
204
DISTRIBUTING A FIRST SET OF PAIRS AND A SECOND SET OF PAIRS TO A PLURALITY OF DATA NODES
206
CALCULATING, ON A FIRST DATA NODE BELONGING TO THE PLURALITY OF DATA NODES, A FIRST PROBABILITY OF A FIRST USER'S INTEREST IN A FIRST PRODUCT BASED ON A FIRST OBSERVABLE FACTOR AND A FIRST LATENT FACTOR
208
CALCULATING, ON A SECOND DATA NODE, A SECOND PROBABILITY OF A SECOND LIKELIHOOD OF A RELATIONSHIP BETWEEN THE FIRST USER AND A SECOND USER, BASED ON A SECOND OBSERVABLE FACTOR AND A SECOND LATENT FACTOR
210
DETERMINING, BASED ON THE FIRST PROBABILITY AND THE SECOND PROBABILITY, A MOST LIKELY INTEREST OF THE FIRST USER AND A MOST LIKELY RELATIONSHIP OF THE FIRST USER
212
PREDICTING A POTENTIAL INTEREST OF THE FIRST USER BASED ON THE MOST LIKELY INTEREST AND THE MOST LIKELY RELATIONSHIP
214
STOP

FIG. 2

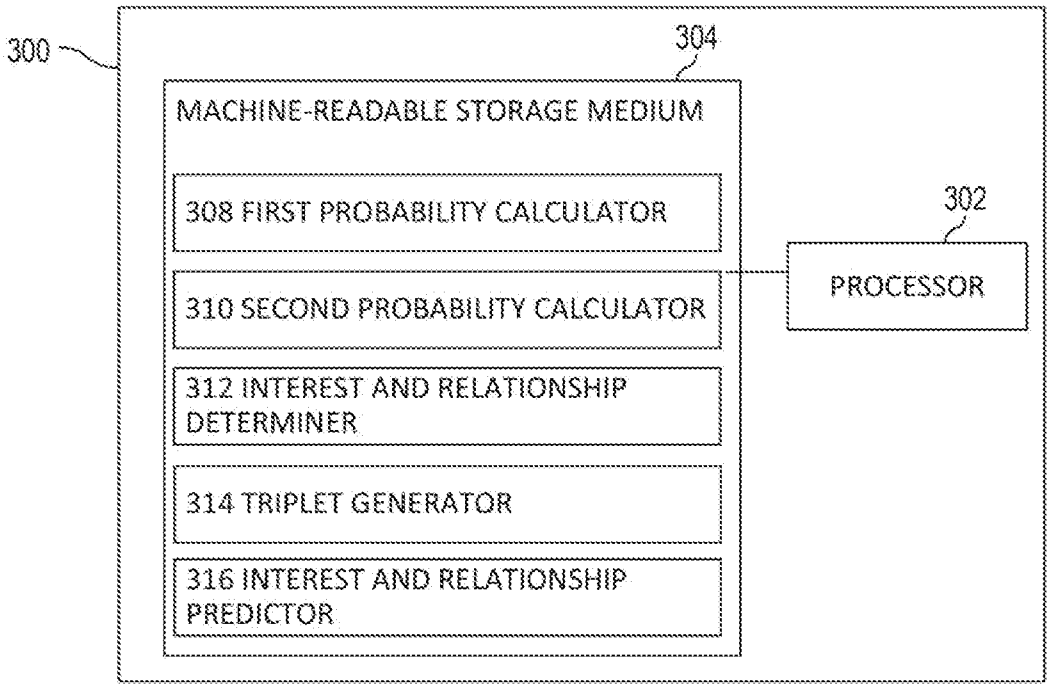
300
304
MACHINE-READABLE STORAGE MEDIUM
308 FIRST PROBABILITY CALCULATOR
310 SECOND PROBABILITY CALCULATOR
312 INTEREST AND RELATIONSHIP DETERMINER
314 TRIPLET GENERATOR
316 INTEREST AND RELATIONSHIP PREDICTOR
302
PROCESSOR

FIG. 3

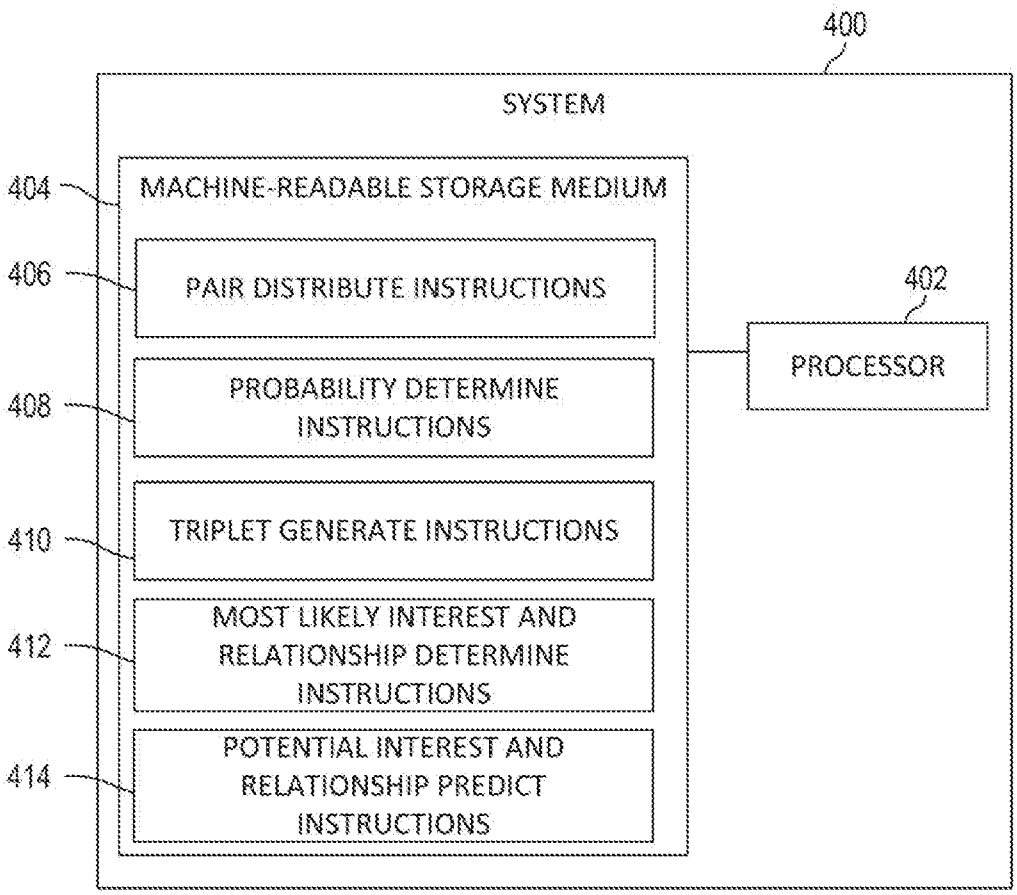
400
SYSTEM
404
MACHINE-READABLE STORAGE MEDIUM
406
PAIR DISTRIBUTE INSTRUCTIONS
402
PROCESSOR
408
PROBABILITY DETERMINE INSTRUCTIONS
410
TRIPLET GENERATE INSTRUCTIONS
412
MOST LIKELY INTEREST AND RELATIONSHIP DETERMINE INSTRUCTIONS
414
POTENTIAL INTEREST AND RELATIONSHIP PREDICT INSTRUCTIONS

FIG. 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/073690

A. CLASSIFICATION OF SUBJECT MATTER

G06Q 30/00(2012.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F, G06Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, CNKI, WPI, EPODOC: user, product, predict, profile, interest, probability, factor, observ+, latent+, relation+, potential, commend

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 104281956 A (NANJING UNIVERSITY OF INFORMATION SCIENCE & TECHNOLOGY) 14 January 2015 (2015-01-14) description, paragraphs 0025-0065, figures 1-5	1, 9, 12
X	US 2009077132 A1 (SONY CORPORATION) 19 March 2009 (2009-03-19) description, paragraphs 0057-0063, 0084-0104	1, 9, 12
A	CN 103996143 A (EAST CHINA NORMAL UNIVERSITY) 20 August 2014 (2014-08-20) the whole documents	1-15
A	US 6633852 B1 (MICROSOFT CORPORATION) 14 October 2003 (2003-10-14) the whole documents	1-15

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

☑ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
13 October 2016	26 October 2016

Name and mailing address of the ISA/CN	Authorized officer
STATE INTELLECTUAL PROPERTY OFFICE OF THE P.R.CHINA 6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088 China	LI,Nan
Facsimile No. (86-10)62019451	Telephone No. (86-10)62413690

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/073690

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 104281956 A	14 January 2015	None	
US 2009077132 A1	19 March 2009	WO 2007037139 A1	05 April 2007
		JP 2007122683 A	17 May 2007
		EP 1835419 A1	19 September 2007
		CN 101069184 A	07 November 2007
		KR 20080045659 A	23 May 2008
		IND ELNP200703809 E	24 August 2007
CN 103996143 A	20 August 2014	None	
US 6633852 B1	14 October 2003	None	