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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 November 2001 (29.11.2001)

PCT

(10) International Publication Number
WO 01/90926 A2

(51) International Patent Classification[7]: G06F 17/00

(21) International Application Number: PCT/US01/40760

(22) International Filing Date: 17 May 2001 (17.05.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/574,108 19 May 2000 (19.05.2000) US

(71) Applicant: **NAPSTER, INC.** [—/US]; 1475 Veterans Blvd., Redwood City, CA 94063 (US).

(72) Inventors: **JANNINK, Jan, F.**; 2041 Sharon Road, Menlo Park, CA 94025 (US). **SCHIRMER, Thomas, E.**; 2700 Del Medio Court, #216, Mountain View, CA 94040 (US). **SHIVAKUMAR, Narayanan**; 1035 Aster Avenue, Apt. 1113B, Sunnyvale, CA 94086 (US).

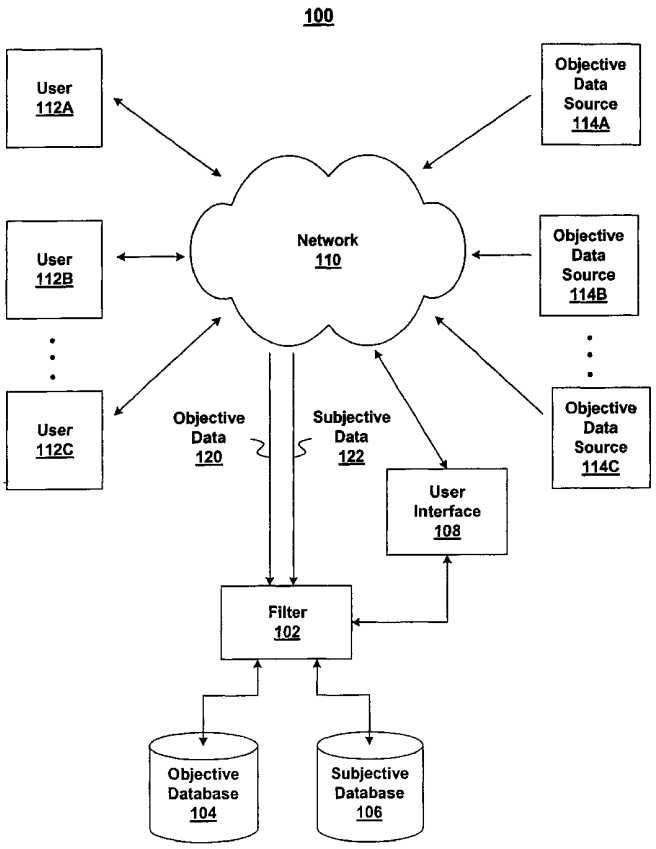
(74) Agents: **HAVERSTOCK, Thomas, B.** et al.; Haverstock & Owens LLP, Suite 420, 260 Sheridan Avenue, Palo Alto, CA 94306 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European

[Continued on next page]

(54) Title: SYSTEM AND METHOD FOR DETERMINING AFFINITY USING OBJECTIVE AND SUBJECTIVE DATA



(57) Abstract: A system and method for determining affinity between database items using objective and subjective data, including receiving a search item, computing an affinity between the search item and each of a plurality of items in an objective database, adjusting the affinities based on subjective data, and outputting a ranked result based on the adjusted affinities.

WO 01/90926 A2

patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

SYSTEM AND METHOD FOR DETERMINING AFFINITY USING OBJECTIVE AND SUBJECTIVE DATA

Background

Field of the Invention

The present invention relates generally to information retrieval techniques and more particularly to determining affinity between items using objective and subjective data.

Related Art

The Internet has dramatically changed the manner in which we access, gather, and collect information. Vast amounts of information are now available on-line. Various tools are available that aid users in searching this information, such as a variety of different search engines. However, it can be difficult and time consuming for on-line users to sift through the mountains of data that are available. There are many instances where users are interested in gathering information that is similar in some respect to a particular topic, but that may not lend itself to being found by a search engine. For example, users can search the Internet to find information on just about any song ever written by any artist by searching on the artist name or song title. A user may, however, be interested in finding new artists that might be similar in some respect to an artist that the user knows and enjoys. In other words, the user may be interested in finding those artists that have a high degree of affinity to the known artist. The term affinity is used herein to refer to a measure of similarity between two items. Unfortunately, known search engines are not particularly useful for such a search.

Tools are available today that allow a user to learn more about a topic of interest, where related topics are searched based on objective properties related to the topic. For example, many sites allow users to search for songs in a particular genre, or for newspaper

articles having a particular search term in the headline or body of the article text. However, this type of search often produces far too many results to be useful, or results which are of minimal relevance to the sought after topic. Many of these search facilities also fail to incorporate any subjective data into the search process, such as taking into account the opinion of other users who have sought similar information in the past.

Therefore, what is needed is an improved system and method for determining affinity between items of data using both objective and subjective data.

Summary of the Invention

The present invention is directed to a system and method for determining affinity between database items using objective and subjective data, including receiving a search item, computing an affinity between the search item and each of a plurality of items in an objective database, adjusting the affinities based on subjective data, and outputting a ranked result based on the adjusted affinities.

Brief Description of the Drawings

The present invention is described with reference to the accompanying drawings. In the drawings, like reference numbers indicate identical or functionally similar elements. Additionally, the left-most digit(s) of a reference number identifies the drawing in which the reference number first appears.

FIG. 1 is a block diagram depicting a network computing environment within which an example embodiment of the present invention operates.

FIGs. 2A and 2B depict example items stored in an objective and subjective database, respectively.

FIG. 3 is a flowchart that describes determining affinity between items using objective and subjective data according to an example embodiment of the present invention.

FIG. 4 is a flowchart that describes in greater detail computing an affinity between a search item and other items in the objective database based on objective properties of the items according to an example embodiment of the present invention.

FIG. 5 is a flowchart that describes in greater detail adjusting objective affinity calculations based on subjective data according to an example embodiment of the present invention.

FIG. 6 is a data flow diagram that illustrates the iterative nature of determining affinity between database items according to an example embodiment of the present invention.

Detailed Description

The present invention is directed to a system and method for determining affinity between database items using objective and subjective data, including receiving a search item, computing an affinity between the search item and each of a plurality of items in an objective database, adjusting the affinities based on subjective data, and outputting a ranked result based on the adjusted affinities.

FIG. 1 illustrates a network computing environment 100 within which an example embodiment of the present invention operates, including a network 110 that is accessed by one or more users 112 (shown as 112A, 112B, and 112C). According to an example embodiment of the present invention, a filter 102 determines an affinity between a search item and other items within an objective database 104, where the affinity determination uses data stored in both objective database 104 and a subjective database 106. Objective data 120 can be collected from one or more objective data sources 114 (shown as 114A, 114B, and 114C) accessible via network 110, whereas subjective data 122 can be collected from users 112. Objective data 120 and subjective data 122 can be collected by filter 102, as described below, or by separate data collection software (not shown). Users 112 interact with filter 102 via a user interface 108. For example, users 112 can enter a search item and the resulting affinity relationships can be displayed, all via user interface 108. Filter 102 and user interface 108 can be implemented as one or more lines of computer code using any appropriate computer language.

The present invention can be applied to many different applications wherein it is advantageous to determine affinity relationships between items in objective database 104, and wherein subjective data 122 is available related to these affinity relationships. According to an example embodiment, filter 102 can be used to provide users 112 interested in music with additional information related to a favorite artist or song. For example, user 112 hears a song on the radio by an artist A1, and is interested in finding the names of other artists that are in some way similar to A1. Filter 102 can be used to determine other artists that are similar to Jane Doe, i.e., artists that have a high affinity relationship to A1 relative to other artists.

Though the present invention is described below in terms of this example music embodiment, the principles described herein can also be applied to many applications

involving other types of data. For example, filter 102 can be used to determine affinity relationships between many different types of media including, but not limited to, books, compact disks (CDs), digital video disks (DVDs), and newspaper articles.

Network 110 can represent any network, or two or more interconnected networks, that provides a communications pathway between users 112, objective data sources 114, and filter 102. For example, network 110 can represent the Internet. Alternatively, network 110 can represent a wireless telecommunications network interconnected with the Internet, whereby users 112 employing mobile handheld devices access filter 102 via a wireless connection.

Objective database 104 can represent any database (or multiple databases) that includes two or more items (otherwise referred to as records or entries) of a particular object class (e.g., artists, songs). Items can be described by various objective properties. FIG. 2A depicts example items 202 (shown as 202A through 202C) stored in objective database 104 according to an example embodiment of the present invention. Associated with each item 202 are one or more objective properties. According to the example music embodiment, an artist or song can be described by the following example properties: name or title (e.g., Jane Doe or "Song Title"), genre (e.g., rock, country, jazz), era (e.g., 1970's, big band), tempo (e.g., slow, fast), and popularity (e.g., number of albums sold, number of concert tickets sold). Other properties can include, but are not limited to, release date, length, energy, edginess, mood, imagery, and topic. Each property is preferably described as a quantitative value, though according to an alternative example embodiment properties can be described using textual descriptors.

According to an example embodiment of the present invention, objective data 120 is gathered from one or more objective data sources 114 via network 110 and used to populate objective database 104. For the example music application, objective data sources 114 can

include, but are not limited to, record company or other third party databases, music information sources, on-line dictionaries, artist web sites, and fan web sites. Objective data 120 can be gathered manually or automatically using typical web crawler technology known in the art.

Objective data 120 collected via network 110 must often be massaged into the format expected by objective database 104. Further, according to an example embodiment of the present invention, a normalization weight for each property is applied so that the relative contributions of each property to the affinity calculations described below are approximately equal. This normalization may be necessary, for example, where the range of values assigned to item properties have various magnitudes. Popularity could be expressed in terms of millions of records sold, whereas tempo could be expressed as a scalar quantity between 0 and 1. Given relatively small differences in both properties, the magnitude of a difference in popularity could vastly overshadow any difference in tempo, absent a normalization of both values. According to an example embodiment of the present invention, the normalization weights are chosen such that the weighted property values map down to a value between 0 and 1.

According to an example embodiment of the present invention, objective database 104 can be updated periodically as new objective data 120 becomes available. The rate at which this update occurs can depend, in part, on the desired freshness of the data within objective database 104, and on available memory and computational resources. For example, objective database 104 might be updated weekly, daily, or even hourly, depending upon the type of data, the database size and the available resources. Objective data 120 can be saved within objective database 104 and accessed as necessary for affinity (and other) calculations, as described in greater detail below.

Subjective database 106 can represent any database (or multiple databases) that includes data related to the opinions or actions of users, where the data bears some relationship to the affinity between items 202 stored in objective database 104. FIG. 2B depicts example subjective data records 204 (shown as 204A through 204C) stored in subjective database 106 according to an example embodiment of the present invention. Each entry 204 corresponds to a particular user 112, and has associated with it one or more rules. A rule indicates that a user's action or opinion suggests that an affinity relationship exists between two items 202 within objective database 104.

The data used to establish rules can be provided directly by a user. For example, a user can be asked to name artists having a particular set of properties, such as artists within a given genre and/or era. According to an example embodiment of the present invention, a rule can be established between each artist provided by the user indicating that, at least in the user's opinion, these artists are in some way related (i.e., there is an affinity relationship between the artists). The data used to establish rules can also be implied based on a user's actions. For example, a user browsing a web site might seek information on two or more artists sharing one or more properties. It might reasonably be inferred from the user's actions that these artists are in some way related and therefore that an affinity relationship exists. The collection and use of subjective data 122 will be described in greater detail below in conjunction with example embodiments of the present invention.

User interface 108 can represent, for example, a graphical user interface (GUI) implemented according to well known GUI techniques to perform the input/output (I/O) functionality described herein. According to an example embodiment of the present invention, user interface 108 can be implemented as described in co-pending U.S. Patent

Application Ser. No. 60/162,465, entitled "Systems and Methods For Visualization of Data Sets Containing Objects", which is incorporated by reference herein.

FIG. 3 is a flowchart that describes a process that determines affinity between database items using objective and subjective data according to an example embodiment of the present invention. In operation 302, a search item is received from a user 112. The search item represents an item for which user 112 wishes to find other similar items within objective database 104. For example, user 112 hears a particular song on the radio (song1), and wishes to find other similar songs. User 112 interacts with user interface 108 to input the title of song1.

According to an example embodiment of the present invention, a canonicalization technique is used to normalize the search item input by the user. A canonical label is associated with each item within objective database 104. A function is defined for each class of items that can be used to calculate the canonical label given a wide variety of typical variations of the label that are often used to refer to the item. In this way, users who misspell or use a shortened version of a label to refer to a particular item are mapped to the correct search item (i.e., the item intended by the user).

In operation 304, filter 102 computes an affinity between the search item and other items within objective database 104. According to an example embodiment of the present invention, filter 102 computes an affinity between the search item and each item within objective database 104 of the same object class. For example, if the search item is a song, then filter 102 computes an affinity between the search item and each of the other song items within objective database 104.

As described above, affinity represents a measure of the distance, or similarity, between two items. Affinity between two items can be calculated as the normalized measure of the difference between the items' properties. Those items that are close, i.e., have a relatively small distance between them, are considered to have a stronger affinity than those items that are further apart. FIG. 4 is a flowchart that describes operation 304 in greater detail according to an example embodiment of the present invention, illustrating one approach to computing an affinity between a search item and other items in the objective database.

As described above, objective data 120 that is collected to populate objective database 104 can, in some instances, be normalized such that each property contributes in an approximately equal manner to the affinity calculation between two items. In operation 402, which is an optional step, these normalization weights can be adjusted according to the user's preferences for altering the relative importance of certain properties with respect to the affinity calculation. For example, a user 112 enters artist1 as a search item, but wishes to specify that popularity is the most important property when determining affinity. In other words, popularity contributes a larger component to the affinity calculation than do the other properties.

According to an example embodiment of the present invention, users 112 are allowed to specify an order of relative importance between objective properties, where the weights associated with each property are adjusted by a set amount according to the order. For example, a user can specify a preferred order of importance such as genre, era, popularity, and tempo. In this case, the normalization weight associated with genre will be adjusted to reflect an increased importance, whereas the weight associated with tempo will be adjusted to reflect a decreased importance, and the weights in the middle will be adjusted appropriately.

By predefining the user preferences in this manner, all possible orderings of the item properties can be pre-computed and stored for fast retrieval. This pre-computing can be reasonably performed for up to approximately six properties; greater than six properties can result in unrealistic computational and storage requirements.

As shown in FIG. 4, operations 404, 406, 408, and 410 are repeated for each item within objective database 104 for which an affinity value is calculated. In operation 404, the similarity is computed between each property of the search item and the corresponding property of the current item in objective database 104 for which the affinity is being calculated (the target item). According to an example embodiment of the present invention, the similarity between properties is calculated as the distance between the numerical property values. In operation 406, the distances calculated in operation 404 are scaled by the appropriate normalization weights (by the standard normalization weights, or if adjusted in operation 402, by the adjusted normalization weights). In operation 408, the normalized differences are combined to form an affinity measurement between the search item and the target item. These steps are then repeated to generate an affinity measurement for each target item in objective database 104.

Consider the following illustrative example. Objective database 104 includes four items (A1, A2, A3, and A4) each having three properties (x, y, z), given by:

$A1 = [x1, y1, z1]$

$A2 = [x2, y2, z2]$

$A3 = [x3, y3, z3]$

$A4 = [x4, y4, z4]$

where x1, y1, and z1 are the values for the three properties associated with item A1, and so on through A4. Assume that user 102 enters A1 as the search item. Operations 404 through 408 can be summarized as:

$$A12 = N_x|x1 - x2| + N_y|y1 - y2| + N_z|z1 - z2|$$

$$A13 = N_x|x1 - x3| + N_y|y1 - y3| + N_z|z1 - z3|$$

$$A14 = N_x|x1 - x4| + N_y|y1 - y4| + N_z|z1 - z4|$$

where A12 represents the affinity calculation between items A1 and A2, and so on for A13 and A14, where N_x, N_y, and N_z represent the normalization weights for properties x, y, and z, respectively, and where $|\cdot|$ denotes an absolute value operation. The values of N_x, N_y, and N_z can be adjusted to achieve a weighting of properties desired by the user. Those skilled in the art will recognize that the distance calculation described with respect to this example embodiment is equivalent to calculating an L1-distance if the properties associated with each item are treated as vectors. Other distance metrics can be used to calculate affinities including, but not limited to, Euclidean (L2)-distance, and dot product (cosine)-distance.

Items A2, A3, and A4 can then be ranked in order of their affinity, from smallest (the greatest affinity to search item A1) to largest (the least affinity to search item A1). Assume for purposes of illustration that item A4 has the smallest affinity value, followed by A2 and then A3 with the highest affinity value. The initial ranking is therefore A4, A2, A3, where affinity is calculated using objective data stored in objective database 104.

In operation 412, the items within objective database 104 are clustered according to the affinities calculated in operations 404 through 410. According to an example embodiment of the present invention, as a result of the affinity computation over the objective database, it is possible to group items according to their affinity to other items. This grouping or clustering of related items in the database indicates which items are

predisposed to have strong affinity. Some items will belong to more than one cluster. These clusters are used in the pre-processing of subjective data as described below. To continue our above example, let us assume that after the affinity computation, we find that A1, A2, and A4 fall in one cluster C1, based on their affinity scores, while A3 and A4 fall in another, C2. Note that A4 is both in C1 and C2.

Returning now to FIG. 3, in operation 306 the affinity values computed in operation 304 are adjusted based on subjective data. This type of operation is referred to within the relevant art as collaborative filtering. The collaborative filtering process allows for the injection of the subjective opinion of a consensus of users to reinforce the affinity computation and to make the results more relevant to the users' preferences. FIG. 5 is a flowchart that describes operation 306 in greater detail according to an example embodiment of the present invention. As described above, subjective data 122 can include data that is collected directly from users 112, such as explicitly querying users 112 to enter similar artists sharing one or more properties (e.g., "please enter your favorite jazz artists from the 1990's"). However, it is often difficult to collect statistically relevant quantities of explicitly produced data.

In contrast, significant amounts of relevant subjective data can be inferred from the actions of users 112. In operation 500, for example, user browsing activity is collected as subjective data 122. According to an example embodiment of the present invention, browsing activity data is categorized according to browser cookie values. As user 112 interacts with user interface 108 to request information on various items in objective database 104, browser cookie values are sent by the user's browser software along with the user's request and stored in user activity logs. Browser cookie values serve as a number identifying the active browser (user ID), the time at which the access took place, and the item requested

(e.g., artist name, song title). And since the browser cookie values are constant from one session to the next, a user's browsing activity can be collected and correlated over multiple browsing sessions. The log records are sorted by user ID, with the result being that the user requests are separated into bins corresponding to different users of the web site.

In operation 502, the user browsing activity collected in operation 500 is partitioned by the clusters determined in operation 412. According to an example embodiment of the present invention, each individual user's requests are partitioned by cluster, where each bin corresponds to a particular cluster. Because items can belong to multiple clusters, items can therefore be partitioned into multiple bins.

In operation 504, rules are assigned within each partition. According to an example embodiment of the present invention, the user requests within each bin partitioned in operation 502 are assumed to be potentially similar to one another, at least in the opinion of the user who made the requests. Each combination of items within a particular object class are therefore paired, forming a rule. For example, a bin might contain three artists (A1, A2, and A3) and four songs (S4, S5, S6, and S7). Three rules are created from three pairings of artists (R12, R13, R23, where R12 represents the rule created from pairing A1 and A2, and so on). Six rules are created from the six pairings of songs (R45, R46, R47, R56, R57, and R67, where R45 represents the rule created from pairing S4 and S5, and so on).

In operation 506, a subjective property is computed for each rule created in operation 504. According to an example embodiment of the present invention, the subjective property is a value that is indicative of the relative number of occurrences of a particular rule within the bins of users 112. Those rules which appear in the bins of multiple users 112 are identified as being statistically significant. A rule found in the bin of a single user 112 has little statistical significance, but when many users 112 have the same rule (i.e., the same pairs

of accesses) the rule becomes a more significant indicator of popular opinion linking the two items. As they are calculated, subjective property values can be stored in a square matrix having a dimension equal to the number of items in objective database 104 (actually, only one-half of the matrix is required, since the matrix is symmetric), where each element in the matrix represents the subjective property value between item-X and item-Y.

In operation 508, the objective affinities calculated in operation 304 between the search item and other items in objective database 104 are adjusted using the subjective properties calculated in operation 506. The objective affinity calculation can be expressed as a function $AC_0(A1, A2)$ where A1, A2 are the two items for which the affinity is being calculated. The collaborative filtering function produces rules of the form $CF_n(A1, A2)$, where n represents an integer index. According to an example embodiment of the present invention, the rules generated by collaborative filtering can be used to adjust the affinity calculation in operation 510 as given by:

$$AC_1(A1, A2) = f(AC_0(A1,A2)) + g(CF_0(A1,A2))$$

where $AC_1(A1, A2)$ represents the affinity calculation adjusted by the collaborative filtering component, and f and g are weighting functions that adjust the output of AC_0 and CF_0, respectively. The weighting functions f and g can be adjusted to achieve a desired balance between the objective and subjective components of the combined affinity calculation. The results of AC_1 are presented to users via the user interface. Further user activity is logged, allowing CF_1 to be computed. Further adjustments to AC, resulting in AC_n, are computed according to the following equation:

$$AC_n(A1, A2) = f(AC_{n-1}(A1,A2)) + g(CF_{n-1}(A1,A2)) + h(CF_{n-2}(A1,A2))$$

where weighting function h adjusts not the most recent collaborative filtering values from CF_{n-1}, but the previously computed values from CF_{n-2}.

Adjusting the objective affinities using collaborative filtering can change the order of the ranked results. In the example described above, the initial ranking was determined to be A4, A2, A3, using objective affinity values. However, for example, the subjective data could indicate a strong similarity between A1 and A2 (e.g., A1 and A2 share a common objective property and many users 112 requested both A1 and A2). This could impact the distance calculation to the point where A12 has a stronger affinity than A14, with the adjusted ranked result given by A2, A4, A3.

In operation 510, the clusters determined in operation 412 are updated based on the adjusted affinities calculated in operation 508. For example, referring back to example clusters C1 and C2 defined above, their contents may change based on the adjusted affinities. C1 and C2 were determined after the initial affinity computation described above. After incorporating the collaborative filtering adjustments to the affinity scores we find that A1 and A2 still belong to C1, while A4 no longer does, and similarly A2 now also belongs to C2, alongside A3, and A4, the original members of C2. For the example music application, the adjustment of clusters can signify that artist A4 is considered to no longer be similar to A1, while artist A2 is no producing music more similar to A3 than before.

According to an example embodiment of the present invention, operations 500 through 506 can be performed "off-line" rather than being performed with each new search item. Subjective data 122 can be collected and processed into subjective properties for rules on a periodic basis, as new subjective data become available, and stored in subjective database 106. For example, the user activity logs can be queried at the end of each day, and the subjective properties updated based on the new information.

Further, according to another example embodiment of the present invention, subjective data 122 over a given window of time is used to calculate subjective properties, rather than using all available subjective data. By shortening this time window, the subjective properties can be more reflective of user opinions at a particular moment or period in time, though the calculation may be somewhat less statistically reliable depending on the amount of data available for the collaborative filtering process. For example, user opinion shifts fairly rapidly over time regarding which artists or songs are in vogue, and whether particular songs or artists are considered similar to one another. The length of the time window can therefore be varied to trade-off statistical reliability with response time to changing user opinion. In order to maintain continuity with prior filtering results, the prior results, when available, can be incorporated into the affinity calculations, as a second separate subjective measure alongside the most recent results, albeit with a reduced weighting factor. This technique is related to cache aging techniques pioneered for networking protocols. The filtering result aging strategy permits the gradual introduction of new items into the user interface, and prevents the sudden disappearance of previously existing content.

The computational and memory requirements associated with the subjective processing represented by operations 500 through 508 can be reduced without significant reduction in the accuracy of the processing. According to an example embodiment of the present invention, only those items in objective database 104 having an objective affinity greater than a given threshold are put through subjective processing. Eliminating those items below the affinity threshold tends to eliminate statistical outliers, items which are unlikely to be similar to the search item, but which can significantly reduce computational and memory burdens. According to an alternative example embodiment, only the top N_{out} items in terms of the objective affinity calculations are put through subjective processing. This can have a

similar effect to applying an affinity threshold, but may be less reliable since the top N_{out} items can vary tremendously in precision from one item to another.

Further, according to another example embodiment of the present invention, various techniques can be used to pre-select those users 112 who are more likely to request similar items, i.e., are more likely to share common rules. Subjective properties are therefore only calculated between users 112 that have been pre-selected as being more likely to request similar items. As a result, the processing and memory requirements associated with subjective processing can be greatly decreased without significantly impacting the reliability of the subjective properties, because those users 112 eliminated from the calculation are less likely to have contributed common rules. These techniques can include, but are not limited to, min-hashing algorithms and iceberg algorithms. By using these techniques, the present invention can scalably handle larger amounts of data by using ever more stringent pre-selection criteria (thereby limiting the number of user for which subjective properties are calculated).

FIG. 6 is a data flow diagram that illustrates the iterative nature of determining affinity between database items according to an example embodiment of the present invention. The affinity computation represented by operation 304 uses objective data from objective database 104 to generate an initial objective affinity calculation for a search item. The results of the affinity calculation are presented to the user via user interface 108, for example, as a list of similar items ranked according to their affinity to the search item. The user can then select one or more of the similar items for additional searching, and the browsing results are stored in subjective database 106. The user can potentially create a rule between the search item and the selected similar item as a result of selecting the similar item. If a significant number of users also select the similar item, this can result in the affinity

between the two items being increased as a result of the collaborative filtering operation represented by operation 306.

By using an affinity threshold to eliminate outliers, the affinity computation can effectively make the collaborative filtering computation more efficient, whereas the collaborative filtering computation can increase the relevance of the results that the affinity computation produces. With each iteration of the loop depicted in FIG. 6, this mutually reinforcing process produces ever more efficient and relevant affinity calculations as more current subjective and objective data become available.

While the present invention has been described in terms of a preferred embodiment, other embodiments and variations are within the scope of the following claims.

What is claimed is:

1. A method comprising:

receiving a search item;

computing an affinity between said search item and each of a plurality of items in an objective database; and

adjusting said affinities based on subjective data.

2. The method of claim 1, further comprising outputting a ranked result based on said adjusted affinities.

3. The method of claim 1, wherein said search item and said objective database items each include a plurality of features, and wherein said computing an affinity comprises:

for each objective database item, determining a similarity between each property of said search item and the corresponding property of said objective database item, scaling each of said similarities by a normalization weight corresponding to said property, and combining said similarities to form said affinity.

4. The method of claim 3, wherein said normalization weights are adjusted according to user preference.

5. The method of claim 3, wherein said similarities are determined by a distance measure.

6. The method of claim 1, further comprising calculating one or more clusters of said items based on said affinities.

7. The method of claim 6, wherein said search item and said objective database items each include a plurality of features, said subjective data comprises user requests collected from user activity logs, and wherein said adjusting said affinities comprises:

partitioning said user requests according to said clusters;

assigning a rule to each pair of said user requests within said partitions;

computing a subjective property for each of said rules; and

adjusting said affinities based on said subjective properties.

8. The method of claim 7, wherein said adjusting said affinities further comprises updating said clusters based on said updated affinities.

9. The method of claim 7, wherein said user requests are weighted according to the time at which said user requests are entered into said user activity logs.

10. The method of claim 7, wherein said user requests are collected for a fixed time window.

11. A system comprising:

means for receiving a search item;

means for computing an affinity between said search item and each of a plurality of items in an objective database; and

means for adjusting said affinities based on subjective data.

12. The system of claim 11, further comprising means for outputting a ranked result based on said adjusted affinities.

13. The system of claim 11, wherein said search item and said objective database items each include a plurality of features, and wherein said means for computing an affinity comprises:

means for determining, for each objective database item, a similarity between each property of said search item and the corresponding property of said objective database item, scaling each of said similarities by a normalization weight corresponding to said property, and combining said similarities to form said affinity.

14. The system of claim 13, wherein said normalization weights are adjusted according to user preference.

15. The system of claim 13, wherein said similarities are determined by a distance measure.

16. The method of claim 11, further comprising means for calculating one or more clusters of said items based on said affinities.

17. The system of claim 16, wherein said search item and said objective database items each include a plurality of features, said subjective data comprises user requests collected from user activity logs, and wherein said means for adjusting said affinities comprises:

means for partitioning said user requests according to said clusters;

means for assigning a rule to each pair of said user requests within said partitions;

means for computing a subjective property for each of said rules; and

means for adjusting said affinities based on said subjective properties.

18. The system of claim 17, wherein said means for adjusting said affinities further comprises means for updating said clusters based on said updated affinities.

19. The system of claim 17, wherein said user requests are weighted according to the time at which said user requests are entered into said user activity logs.

20. The system of claim 17, wherein said user requests are collected for a fixed time window.

100

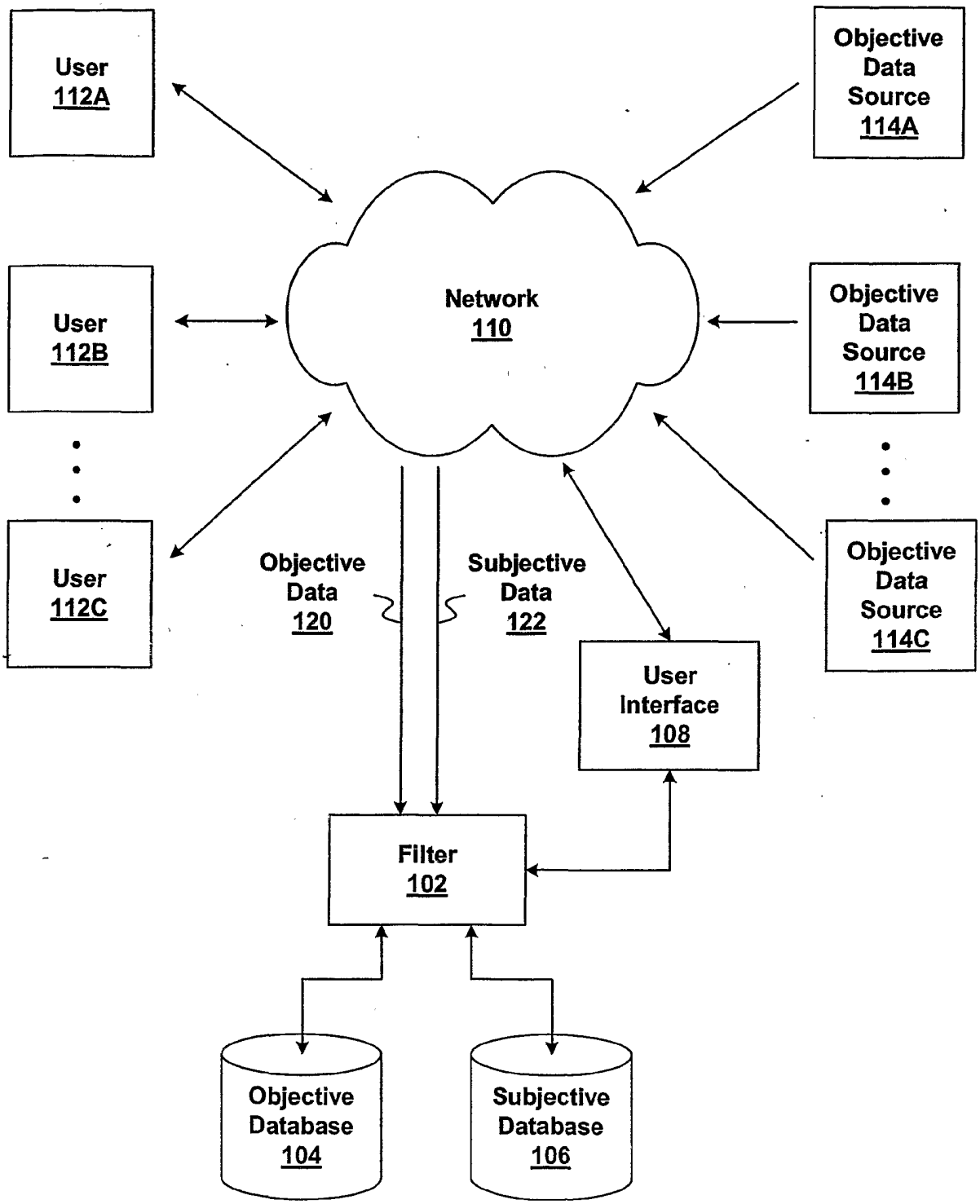

User
112A
User
112B
User
112C
Network
110
Objective
Data
120
Subjective
Data
122
Objective
Data
Source
114A
Objective
Data
Source
114B
Objective
Data
Source
114C
User
Interface
108
Filter
102
Objective
Database
104
Subjective
Database
106

FIG. 1

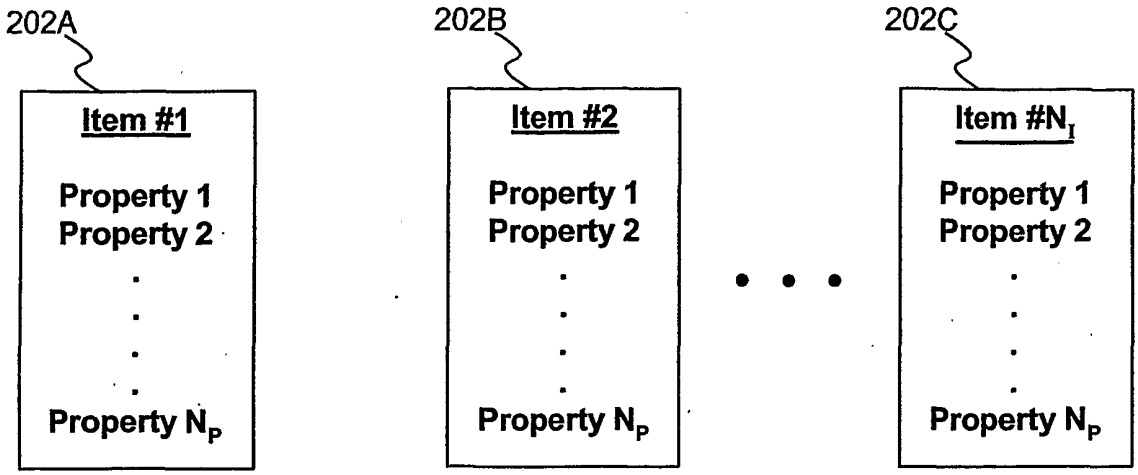

202A
Item #1
Property 1
Property 2
Property N_P
202B
Item #2
Property 1
Property 2
Property N_P
202C
Item #N_I
Property 1
Property 2
Property N_P

FIG. 2A

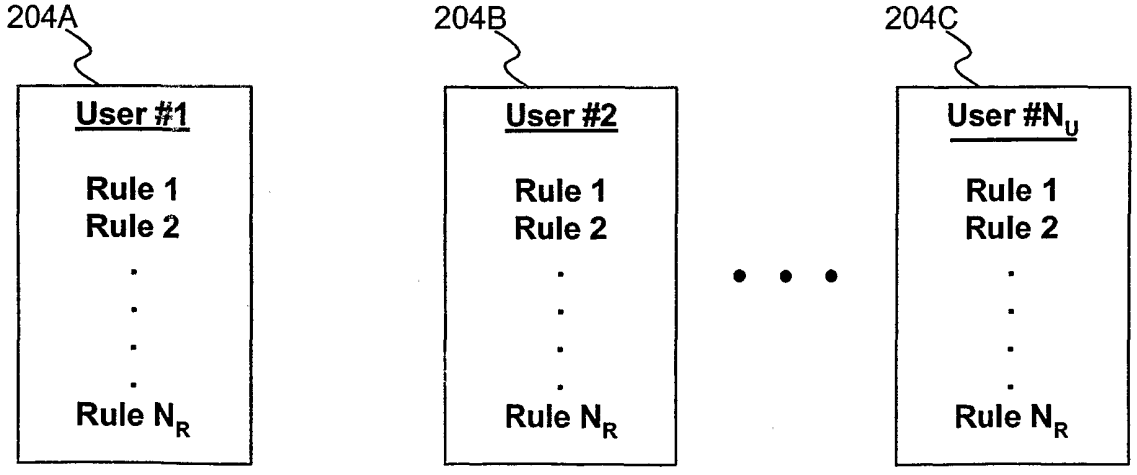

204A
User #1
Rule 1
Rule 2
Rule N_R
204B
User #2
Rule 1
Rule 2
Rule N_R
204C
User #N_U
Rule 1
Rule 2
Rule N_R

FIG. 2B

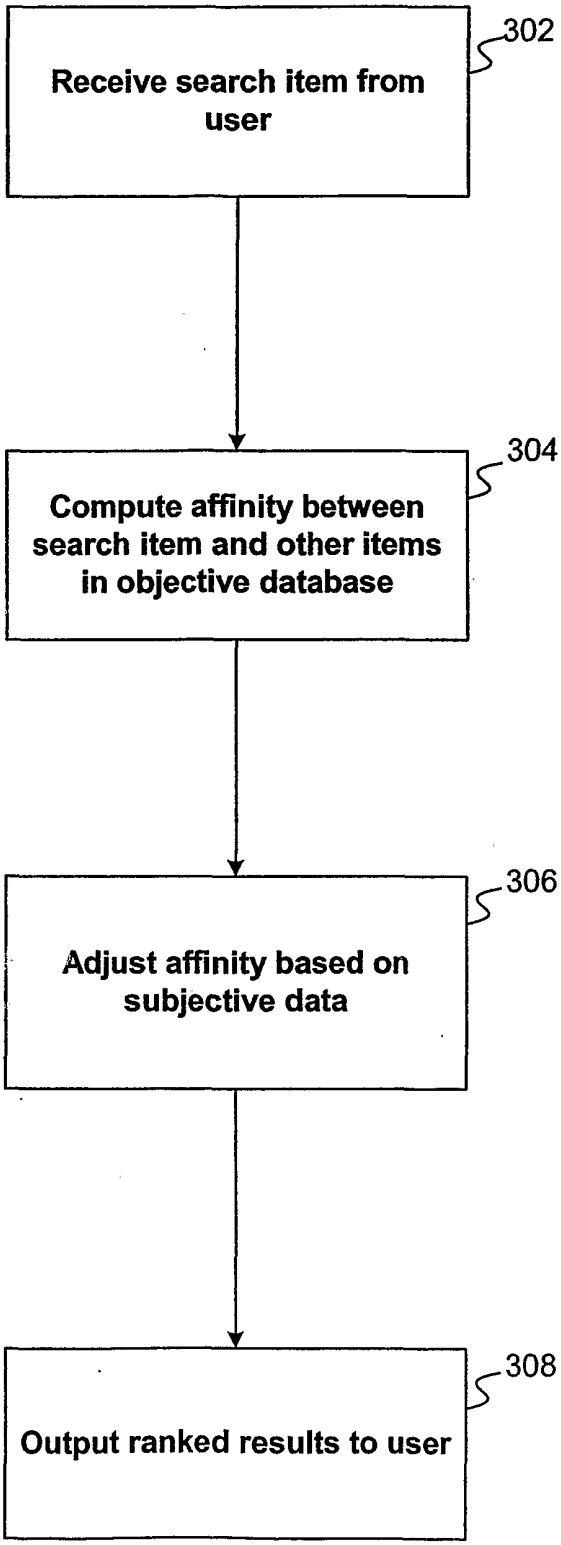

Receive search item from user
302
Compute affinity between search item and other items in objective database
304
Adjust affinity based on subjective data
306
Output ranked results to user
308

FIG. 3

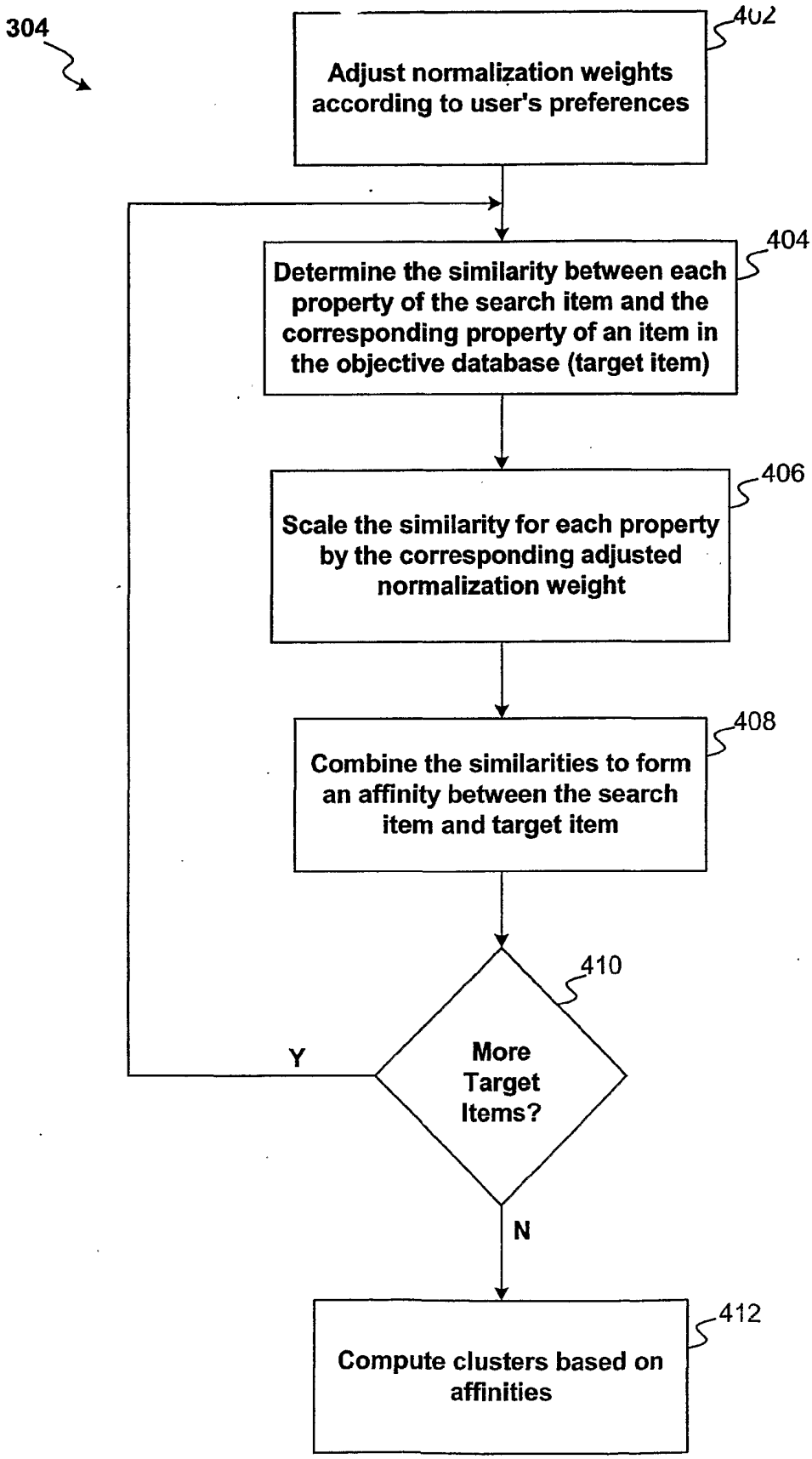

304
402
Adjust normalization weights according to user's preferences
404
Determine the similarity between each property of the search item and the corresponding property of an item in the objective database (target item)
406
Scale the similarity for each property by the corresponding adjusted normalization weight
408
Combine the similarities to form an affinity between the search item and target item
410
More Target Items?
Y
N
412
Compute clusters based on affinities

FIG. 4

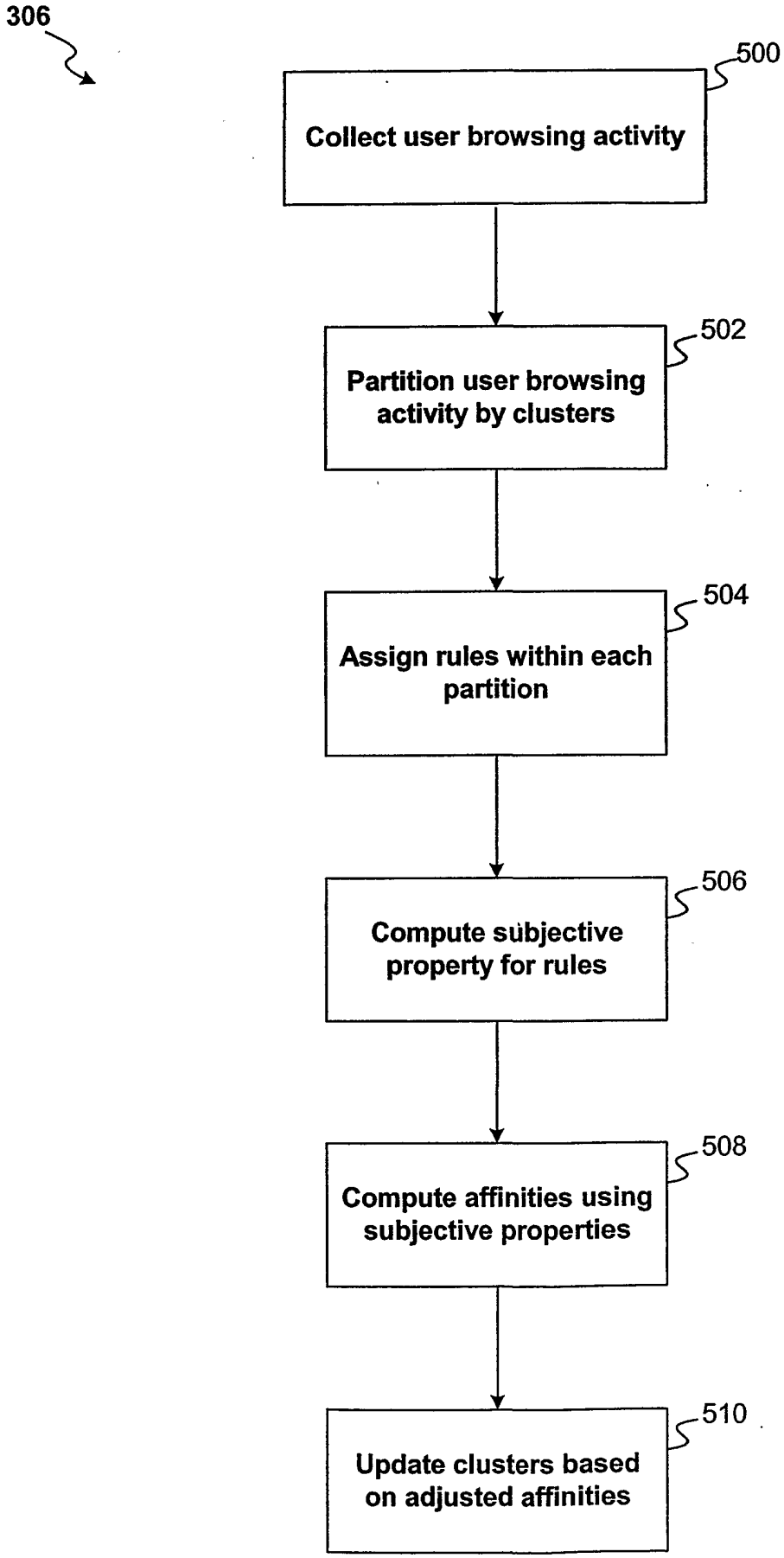

306
500
Collect user browsing activity
502
Partition user browsing activity by clusters
504
Assign rules within each partition
506
Compute subjective property for rules
508
Compute affinities using subjective properties
510
Update clusters based on adjusted affinities

FIG. 5

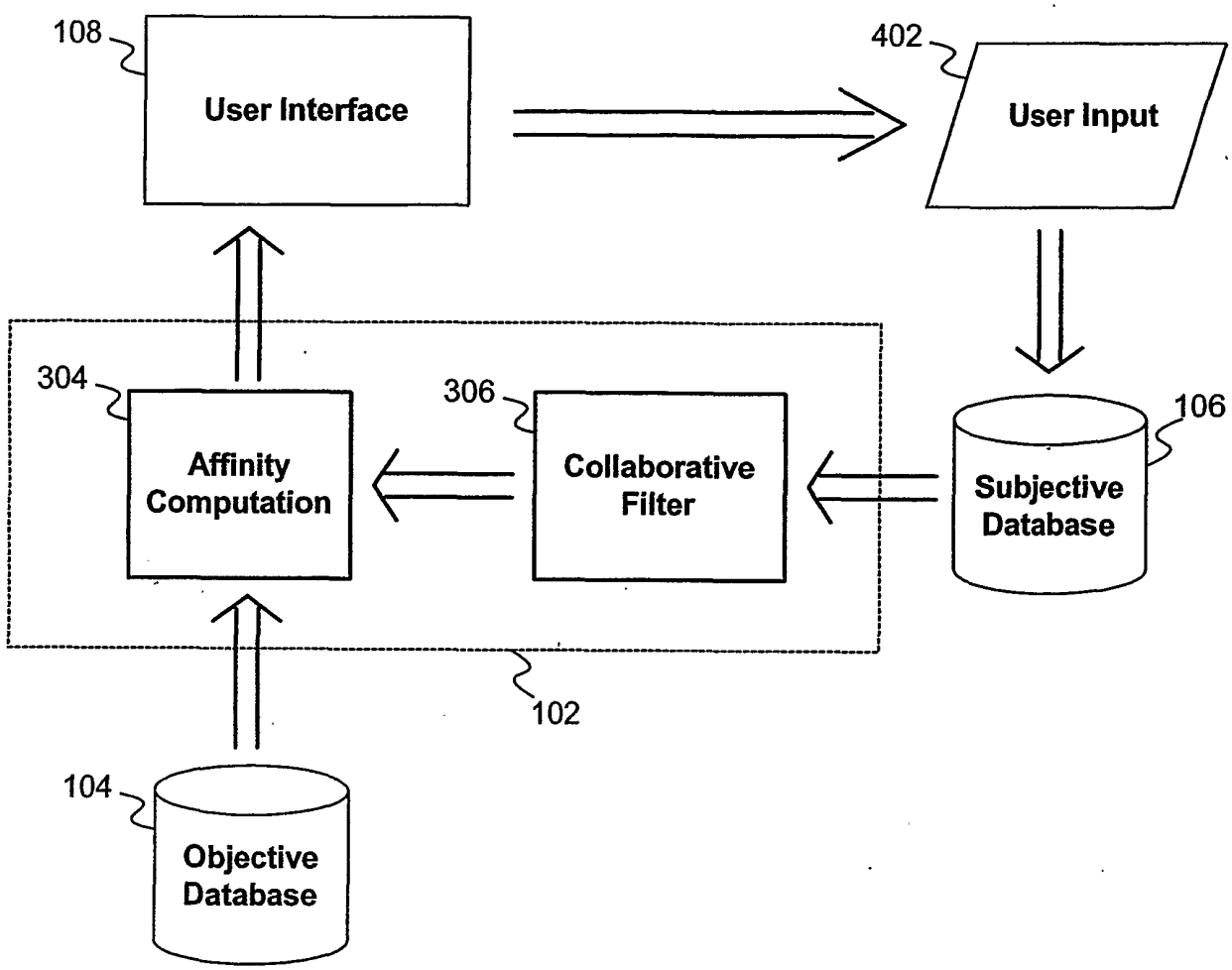

108
User Interface
402
User Input
304
Affinity Computation
306
Collaborative Filter
106
Subjective Database
102
104
Objective Database

FIG. 6