

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
7 May 2009 (07.05.2009)

PCT

(10) International Publication Number
WO 2009/057103 A2

(51) International Patent Classification:
G06Q 30/00 (2006.01)

(21) International Application Number:
PCT/IL2008/001416

(22) International Filing Date: 28 October 2008 (28.10.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/001,052 31 October 2007 (31.10.2007) US

(71) Applicant (for all designated States except US): **MUSICGENOME.COM INC.** [US/US]; Suite 1600, 1210 Market Street, Wilmington, Delaware 19801 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **GANG, Dan** [IL/IL]; 13 Derech HaYam, Apt. # 13, 55900 Ganna Tiqwa (IL). **LEHMANN, Daniel** [IL/IL]; 12/15 Tzeelim Street, Apt. # 15, 93896 Jerusalem (IL).

(74) Agents: **G.E. EHRLICH (1995) LTD.** et al.; 11 Menachem Begin Street, 52521 Ramat Gan (IL).

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

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

Published:

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

(54) Title: CONTENT-BASED RECOMMENDATIONS ACROSS DOMAINS

(57) Abstract: A method for content-based recommendation across domains, comprises: obtaining first and second sets of content items in two respective domains, getting a grading team to grade the items, then for each pair of first set item - second set item, calculating an individual pseudo-distance value as a distance function of a respective grading user's rankings for the pair of items. Then the pseudo-distances are averaged over all of the graders in the team to provide overall pseudo-distances for each pair, thus setting up a cross content database between the first and second domains. Then, an end user is present with a sample of content items from the first set. The end user grades the items and the grades are used in a function that weights them with the pseudo-distances in the database. The weighting function gives forecast grades to the items in the second set and those with the highest grades can be output to the end user, thereby providing content-based recommendation from said first domain to said second domain.



WO 2009/057103 A2

CONTENT-BASED RECOMMENDATIONS ACROSS DOMAINS

FIELD AND BACKGROUND OF THE INVENTION

The present invention, in some embodiments thereof, relates to a method and
5 apparatus for providing content based recommendations across domains and, more particularly, but not exclusively, to a method of providing such recommendations based on minimal levels of input by the end user.

Being able to recommend a specific item or a few items that will, with high probability, catch the attention of a potential user and convince him/her to buy it or use
10 it, out of a very large inventory, is nowadays a crucial ability in many current commercial situations. Five examples are given:

- helping a shopper to find a CD that fits his/her personal taste in a large CD store or in an internet virtual shop,
- suggesting a DVD to take home to customers of DVD rental outlets,
- 15 • guiding subscribers to premium services offered by cellular operators to entertainment items that they will be interested in, to be downloaded to their phones,
- deciding which records of potential dates will be presented to a subscriber to a dating web site,
- 20 • or presenting the descriptions and pictures of a small sample of the houses on the market or of car models to a potential buyer.

In any of the many situations, described for example in Chris Anderson's book "The Long Tail: Why the Future of Business is Selling Less of More" (2006), in which the distribution of requests for items has what has been called a long tail, i.e., items
25 that are very seldom sold make the bigger part of the sales, it is imperative to be able to guide each consumer personally to the items that he and almost nobody else will like.

State of the art techniques for recommendations can be classified in two categories:

- 30 • Content-based techniques that rely on some analysis of the items in the inventory: musical analysis of tracks, description of the movies, other subscribers to the dating site, cars, or houses, and

- Collaborative filtering methods that bypass the analysis of the content by accumulating information on a large customer base and extrapolating from the behavior of customers similar (in behavior) to the person one wants to recommend to.

5 Collaborative filtering is an attractive technology since it does not require professional analysis of the content and relies only on information gathered at no cost from users of the system. But, in many of the domains mentioned above, collaborative filtering methods have been found to be inferior to content based methods, and content-based techniques, have been found to provide better recommendations. The
10 reason for this is that the information one can gather from occasional users concerning the inventory is not very reliable. Furthermore any given individual may buy something and subsequently not like it, a point that such techniques are unable to take into account. For the automatic systems 'a person bought' is treated as being synonymous with 'a person liked' but this is plainly not always the case. A similar
15 point that automatic systems are unable to take into account is the possibility that a particular purchaser may be buying for someone else.

Furthermore it is generally the case that each occasional user interacts only with a relatively small number of items in the inventory. Such is the state of affairs in any of The Long Tail situations referred to above. The difficulty is that the information one
20 can gather even from a very large number of users each of whom interacts with just a very small fraction of the catalog is in no way comparable to what content-analysis can give. US Patent 7,102,067 teaches a system and a method for predicting the musical taste and/or preferences of the user and its integration into services provided by a wireless network provider. Although directed toward implementations with
25 wireless providers, the teaching can also be implemented on a regular, i.e., wireline network. The core of the teaching is a system capable of predicting whether a given user, i.e., customer, likes or does not like a specific song from a pre-analyzed catalog. Once such a prediction has been performed, those items that are predicted to be liked best by the user may be forwarded to the mobile device of the user on the cellular (or
30 other wireless) network. The system maintains a database containing propriety information about the songs in the catalog and, most important, a description (profile) of the musical taste of each of its customers, identified by their cellular telephone number. US 7,075,000 teaches a system and a method for predicting the musical taste

and/or preferences of the user. The described system receives, on one hand, ratings of a plurality of songs from the user and/or other information about the taste of the user, and on the other hand information about the songs in the catalog from which recommendations are to be given. The method then combines both types of information in order to determine the musical preferences of the user. These preferences are then matched to at least one musical selection, which is predicted to be preferred by the user. In these two teachings, the present inventors proposed a method, known as the songmap, in which a small group of experts or paid workers interacts with a significant part (or all) of the inventory to provide an indirect analysis of the content and we showed that very accurate recommendations can be obtained, at a reasonable expense.

The state of the art of content-based recommendation systems requires content-analysis that is specific to the domain concerned: musical analysis for musical tracks, classification in genres and description of the plot for movies, and vector of features for potential dating partners, cars, houses or pictures. Therefore the state of the art does not allow for cross-content recommendations such as recommendations of movies or dates based on musical taste, or recommendations of cars based on taste about movies.

The state of the art allows, though, for recommendations that may look like they are cross-content recommendations whereas they are not. First, collaborative filtering techniques are indeed cross-content in nature. Then, some dating sites, for example, ask their customers to describe their musical taste, either in words or by listing favorite songs or artists. They then match customers that have similar musical tastes. This is a technique of limited use since one may ask potential dates for their musical tastes, but one may not ask movies for their musical tastes. The technique may also miss some promising matches: one may appreciate a date who has very different musical tastes from one's own. Many well-matched couples do not share tastes, for example, one partner likes ballet and the other partner prefers boxing. The technology proposed below is very different and immune to the problems above.

The potential uses of such cross-content recommendations are far-reaching and varied. Such a technology may allow for deeply personal relations between the vendor using such technology and its customers.

SUMMARY OF THE INVENTION

The present invention provides a system for allowing choices made by a user in a first field to be used as a prediction of items of interest in a second field.

- According to a first aspect of the present invention there is provided a method
- 5 for content-based recommendation across domains, comprising, using a processor:
- obtaining a first set of content items in a first domain;
- obtaining a second set of content items in a second domain;
- obtaining rankings for content items of the first set and the second set from a group of grading users such that individual grading users provide rankings for items in the first
- 10 set and items in the second set;
- for individual grading users calculating individual pseudo-distance values between content items in the first set and content items in the second set as a distance function of a respective grading user's rankings in the first set and the user's corresponding ranking in the second set;
- 15 averaging the individual pseudo-distance values over all of the grading users, to calculate an overall pseudo-distance value between any given content item in the first set and each content item in the second set, thereby setting up a cross content database between the first domain and the second domain;
- using the cross-content database:
- 20 presenting an end user with a sample of content items from the first set;
- obtaining from the end-user sample rankings for content items in the sample;
- finding items in the second set which are indicated via a grading function, the function comprising computing, for each item in the second set, an average of the end-user sample rankings weighted by corresponding pseudo-distances; and
- 25 returning the found items in the second set to the end user, thereby providing content-based recommendation from the first domain to the second domain.

In an embodiment, the content items comprise items downloadable to communication devices over cellular networks.

- In an embodiment, the distance function comprises calculating a difference
- 30 between the respective rankings.

In an embodiment, the grading function comprises applying a grading value in inverse proportion to a respective pseudo-distance.

The method may comprise dividing the second set into a plurality of separately configured zones, such that content of each zone can be returned independently.

The method may comprise calculating gradings for respective zones based on pseudo-distances weighted by a respective end-user's sample rankings, and returning
5 content items additionally based on resulting gradings of zones.

The method may comprise providing the end user with a user interface to allow selections of one or more of the plurality of zones.

The method may comprise dividing each zone into separately configured maps such that each zone is configurable with some or all of its maps at any given time, the
10 remaining maps being placed in an offline state.

The method may comprise selecting a number of maps to remain offline depending on available processing resources.

According to a second aspect of the present invention there is provided apparatus for content-based recommendation across domains, comprising:

- 15 an domain content item input unit configured for obtaining a first set of content items in a first domain and a second set of content items in a second domain;
a grading unit, associated with the domain content item input unit, configured for obtaining rankings for content items of the first set and the second set from a group of grading users such that individual grading users provide rankings for items in the first
20 set and items in the second set;
a pseudo-distance calculation unit, associated with the grading unit, configured for:
a) calculating individual pseudo-distance values between content items in the first set and content items in the second set as a distance function of a respective grading user's rankings in the first set and the user's corresponding ranking in the second set; and
25 b) averaging the individual pseudo-distance values over all of the grading users, to calculate an overall pseudo-distance value between any given content item in the first set and each content item in the second set, thereby setting up a cross content database between the first domain and the second domain;
a sample output, associated with the cross-content database, for presenting an end user
30 with a sample of content items from the first set;
a sample ranking input associated with the cross-content database configured for obtaining from the end user sample rankings for content items in the sample and applying the rankings to the cross content database;

a recommendation discovery unit associated with the cross-content database for finding items in the second set which are indicated via a grading function of sample rankings weighted by respective pseudo-distances; and

an end user output for returning the found items in the second set to the end user,
5 thereby providing content-based recommendation from the first domain to the second domain.

In an embodiment, the content items comprise one member of the group consisting of: items downloadable to communication devices over cellular networks, items for sale in stores, and items downloadable over a computer network.

10 According to a third aspect of the present invention there is provided a method of creating a cross-content database for content-based recommendation across domains, comprising, using a processor:

obtaining a first set of content items in a first domain;

obtaining a second set of content items in a second domain;

15 obtaining rankings for content items of the first set and the second set from a group of grading users such that individual grading users provide rankings for items in the first set and items in the second set;

for individual grading users calculating individual pseudo-distance values between content items in the first set and content items in the second set as a distance function
20 of a respective grading user's rankings in the first set and the user's corresponding ranking in the second set;

averaging the individual pseudo-distance values over all of the grading users, to calculate an overall pseudo-distance value between any given content item in the first set and each content item in the second set, thereby setting up the cross content
25 database between the first domain and the second domain.

According to a fourth aspect of the present invention there is provided a method for content-based recommendation across domains, comprising, using a processor:

obtaining access to a cross content database, the cross content database comprising:

30 a first set of content items in a first domain;

a second set of content items in a second domain;

and pseudo-distances linking content items of the first domain to content items of the second domain, the pseudo-distances representing distances between gradings

provided to the items by individual grading users averaged over a group of grading users, and

using the cross-content database:

presenting an end user with a sample of content items from the first set;

5 obtaining from the end user sample rankings for content items in the sample;

finding items in the second set which are indicated via a grading function of respective overall pseudo-distances weighted by the sample rankings; and

returning the found items in the second set to the end user, thereby providing content-based recommendation from the first domain to the second domain.

10 Unless otherwise defined, all technical and/or scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which the invention pertains. Although methods and materials similar or equivalent to those described herein can be used in the practice or testing of embodiments of the invention, exemplary methods and/or materials are described below. In case of
15 conflict, the patent specification, including definitions, will control. In addition, the materials, methods, and examples are illustrative only and are not intended to be necessarily limiting.

Implementation of the method and/or system of embodiments of the invention can involve performing or completing selected tasks manually, automatically, or a
20 combination thereof. Moreover, according to actual instrumentation and equipment of embodiments of the method and/or system of the invention, several selected tasks could be implemented by hardware, by software or by firmware or by a combination thereof using an operating system.

For example, hardware for performing selected tasks according to
25 embodiments of the invention could be implemented as a chip or a circuit. As software, selected tasks according to embodiments of the invention could be implemented as a plurality of software instructions being executed by a computer using any suitable operating system. In an exemplary embodiment of the invention, one or more tasks according to exemplary embodiments of method and/or system as
30 described herein are performed by a data processor, such as a computing platform for executing a plurality of instructions. Optionally, the data processor includes a volatile memory for storing instructions and/or data and/or a non-volatile storage, for example, a magnetic hard-disk and/or removable media, for storing instructions and/or data.

Optionally, a network connection is provided as well. A display and/or a user input device such as a keyboard or mouse are optionally provided as well.

BRIEF DESCRIPTION OF THE DRAWINGS

5 Some embodiments of the invention are herein described, by way of example only, with reference to the accompanying drawings. With specific reference now to the drawings in detail, it is stressed that the particulars shown are by way of example and for purposes of illustrative discussion of embodiments of the invention. In this regard, the description taken with the drawings makes apparent to those skilled in the art how embodiments of the invention may be practiced.

In the drawings:

FIG. 1 is a simplified flow chart showing a process for creating a cross-content database according to a first preferred embodiment of the present invention;

15 FIG. 2 is a simplified diagram showing a process for providing a cross-content recommendation using the cross-content database constructed according to the embodiment of FIG. 1;

FIG. 3 is a simplified diagram showing apparatus for managing construction and use of a cross-content database according to preferred embodiments of the present invention;

20 FIG. 4 is a simplified diagram showing aspects of the apparatus of FIG. 3 in greater detail;

FIG. 5 is a graph showing experimental results achieved using an embodiment according to the present invention, and

FIG. 6 is a graph showing further results of the experiment of FIG. 5.

DESCRIPTION OF EMBODIMENTS OF THE INVENTION

The present invention, in some embodiments thereof, relates to a cross-content recommendation system and method and, more particularly, but not exclusively, to a system that uses rankings applied to samples taken from a first domain to calculate preferences amongst items in a second domain.

30 More particularly, a method for content-based recommendation across domains, comprises: obtaining first and second sets of content items in two respective domains, getting a grading team to grade the items, then for each pair of first set item –

second set item, calculating an individual pseudo-distance value as a distance function of a respective grader's rankings for the pair of items. Then the pseudo-distances are averaged over all of the graders in the team to provide overall pseudo-distances for each pair, thus setting up a cross content database between the first and second domains. Then, an end user is presented with a sample of content items from the first set. The end user grades the items and the grades are used in a function that weights the pseudo-distances in the database. The weighting function gives forecast grades to the items in the second set and those with the highest grades can be output to the end user, thereby providing content-based recommendation from said first domain to said second domain.

Before explaining at least one embodiment of the invention in detail, it is to be understood that the invention is not necessarily limited in its application to the details of construction and the arrangement of the components and/or methods set forth in the following description and/or illustrated in the drawings and/or the Examples. The invention is capable of other embodiments or of being practiced or carried out in various ways.

Referring now to the drawings, Figure 1 is a simplified flow chart showing a method of constructing a cross-content database for use in making the recommendations.

The method begins in stage 10 and the first part of the process is to construct a cross-content database which links content items across domains. Using computing including a processing unit a first set of content items are obtained in stage 20 and entered into a database. Then in stage 30 a second set of content items are obtained. The second set differs from the first set in that the content items belong to a different domain. The items could be media items and the different domains could be a different media type, such as films and music. Alternatively the items could be catalog entries for real world items or services being made available. The two domains would be related in the sense that it is plausible that taste in one domain is predictive of taste in the other domain.

Then, in stage 40, a group of graders are shown each of the items in the two sets and asked to grade or rank the items according to the grading scale being used. Grading may be according to any suitable scale, but binary, and five level scales are suggested. Each item thus receives a ranking from each grader.

Subsequently, in stage 50 a pseudo-distance is obtained for each user between each item in the first set and each item in the second set. The pseudo-distance may simply be the absolute value of the difference in the ranking levels between each item. Thus let us take an example where the first domain is songs and the second domain is

5 films. The grading system is a five level grading system where 1 denotes dislike, and five denotes like very much. Let us say a certain grader gives a grade of 4 (from 5) to the Beatles song Yellow Submarine. Subsequently he gives a grade of 3 to the classic film Sting. Then the pseudo-distance for this user between the two items is 1.

Then, in stage 60 each pair of first set – second set items is taken. Following stage 50

10 each such pair has a large number of different pseudo-distances, one for each grader. Now the average is found over all the individual pseudo-distance values to give an average or overall pseudo-distance between the two items. This average pseudo-distance is now entered in stage 70 to form a cross content database.

Reference is now made to FIG. 2, which is a simplified flow chart illustrating

15 how the cross-content database produced in FIG. 1 can be used to provide recommendations to end users as to items from the second domain that may be of interest to them.

Initially access is made to the cross-content database in stage 100. A sample of items from the first set is selected in stage 110 and made available to the end user in

20 stage 120.

The end user is then requested to provide rankings for the content items in the sample. Alternatively, such rankings may be elicited from previous actions of the end user, such as purchases. The ranking or grading scale does not have to be the same as that used in the initial grading of the database items, although it would be expected in

25 most cases that the same scale would be used for convenience.

The rankings obtained from the end user are then weighted (inversely) by the pseudo-distances in the cross-content databases to obtain values (predicted rankings) for each of the items in the second set. It will be borne in mind that pseudo-distances are calculated for all pairs of items in the two zones so that rankings from just a

30 sample in the first set will return values for all of the items in the second set.

Possibilities for the precise function used to weight by the pseudo-distances and return a forecast grade for the items in the second set are discussed in greater detail below. However, typically the function provides that an overall grade is

provided for an item in the second set which is proportional to the rankings in the first set and inversely proportional to the pseudo-distances.

Finally, in stage 150, the highest graded items in the second set are returned to the end user. A predetermined number of such items may be returned or the end user may himself set a number of suggestions he wants to view. Alternatively a grade threshold may be used so that any number of items above a threshold grade may be sent as suggestions. A maximum of such items above the threshold may likewise be set so that the end user is not inundated with suggestions.

In one embodiment, the items are content items and the end user is a cellular telephone user, who downloads content to his cellular telephone. Such an embodiment provides a solution to the well-known problem that user interfacing is difficult on cellular devices and searching for content is particularly difficult. Here the end user is sent helpful suggestions based on his own taste and is thus saved the task of searching.

In one embodiment, the second set is not a single set, but rather is itself divided into a plurality of separately configured zones, such that content of each zone can be returned independently. Thus, in the above examples, films could be divided into zones of drama, comedy, thriller, horror, musical, children, etc. The user may then select the zones he is interested in through a user interface. As a further alternative, the results may be modified to ensure that the user gets at least one suggestion from each zone. As a yet further alternative, the rankings weighted by the pseudo-distances may be used to give a grade to each zone, and then all the results provided are from the highest graded zone or zones.

The zones themselves may be divided into separately configured maps. Each zone may have a large number of items, such that calculating a pseudo-distance to each item may be unduly consumptive of resources. At any given time not all of the maps are kept online, and thus the amount of resources such that each zone is configurable with some or all of its maps at any given time, the remaining maps being placed in an offline state.

Reference is now made to FIG. 3, which illustrates a computerized apparatus for carrying out the method detailed above.

A domain content item input unit 200 obtains the first set of content items in a first domain and the second set of content items in a second domain.

a grading unit 210 is connected to the domain content item input unit, and obtains rankings for content items of the two sets from the group of grading users referred to above. A pseudo-distance calculation unit 220 is connected to the grading unit and carries out the following functions:

- 5 a) calculating individual pseudo-distance values between content items in the first set and content items in the second set as a distance function of a respective grading user's rankings in the first set and the same user's corresponding ranking in the second set, as explained above;
- b) averaging the individual pseudo-distance values over all of the grading users, to
10 calculate overall pseudo-distance values between any given content item in the first set and each content item in the second set. The result is cross-content database 230 in which content items in the first set are linked via an overall pseudo-distance value to each content item in the second set.

Having set up the cross-content database 230, the end user is then provided
15 with a sample through a sample output unit or interface 240. The sample comprises content items from the first set, which the end user is intended to grade.

A sample ranking input 250 is connected to cross-content database 230 and obtains the grades or rankings that the end user has supplied to his sample. The pseudo-distances are now applied as weightings to the rankings to allow
20 recommendation discovery unit 260 to calculate forecast grades for the items in the second set, as explained above. Those items found to have the highest forecast grades are discovered and provided as recommendations through output unit 260.

The technique is a cross-domain technique but is derived from the single-domain mapping technique described in US Patents 7,075,000 and 7,102,067. The
25 cross-domain technique is now discussed in greater detail. Assume we want to make recommendations from a set S of items on the basis of the user's taste in another set T of items (typically of a different nature). A focus group, the grading group referred to above, is put together. This is a group of typically up to 60 persons, although the more the better, who do not have special training. They are paid to examine each of the
30 items in S and in T and give each one of those a grade corresponding to whether they like the item or not. Such grades are usually given on a scale from 1 to 5 (Hate, Do not like, So-So, Like, Love), but may be given also on a binary scale (Do not Like, Like) or any other scale.

Each member of the focus group rates each of the items of S and T. From all those grades a pseudo-distance is computed for every pair of items (x, y) where x is an item of S and y is an item of T. Let g_i and h_i be the grades given by member I of the focus group to items x and y respectively. The quantity $|g_i - h_i|^c$ for some number $c > 1$ is the pseudo-distance between x and y in the eyes of focus group member i: $d_i(x, y)$. The pseudo-distance $d(x, y)$ is computed as the average over the members of the focus group of the $d_i(x, y)$.

The pseudo-distances $d(x, y)$ for x in S and y in T may then be used in the following way to give personal recommendations to a user. We gather information on the taste of the user on some small number of the items of T, the sample referred to above. A sample of five items is sufficient to give valuable recommendations, but ten or twenty can give more reliable results, and the larger the sample the better. The main limitation in sample size is the patience of the end-user.

The information provided by the end-user in response to the sample items either consists of or is translated into grades for each one of those few items. An extrapolation formula, using the pseudo-distances $d(x, y)$, described hereinabove, provides a good guess or forecast of the grade which would have been given by the present end user to each one of the items of S. More particularly, the grade guessed for x is influenced by the one given by the user to y in inverse proportion of $d(x, y)$, or some power (larger than 1) of $d(x, y)$. The items finally presented, i.e., recommended to the user are those items of S whose guessed grades are the highest.

One method for extrapolation as per the above uses a matrix whose entries are the pseudo-distances for each pair of items as derived above. The matrix may be the only information about the catalog, that is the overall set of items, that is used in solving the prediction and recommendation tasks. Variations on the method presented may differ in the type of information used in computing those pseudo-distances. The general idea is that the pseudo-distance between two items i and j is small if, typically, a user who likes i is also expected to like j, while a user who dislikes i is also expected to dislike j and vice versa.

Pseudo-distances generated by different methods may also be optionally combined with the method above and/or with each other, by some weighted linear calculation for example, to define a new pseudo-distance.

Given a matrix of pseudo-distances $d(i,j)$ and a vector of ratings of some of the items:

$$r(k_1), \dots, r(k_n)$$

where $r(k)$ is the rating given by the user to item k , the forecasted rating (for

5 the user) of song i is the weighted average of the ratings $r(k_1), \dots, r(k_n)$

where $r(k)$ is weighted by a quantity that is inversely proportional to some fixed power of $d(k,i)$. In other words, the forecast for the rating of item i is given by:

$$10 \quad c \left[\sum_{j=1}^n r(k_j) / d(k_j, i)^a \right]$$

where a is a positive real number and c is a normalizing factor:

$$c = 1 / \sum (1 / d(k_j, i)^a) .$$

The implementation of an integrated cross-content recommendation system
15 suggests that consideration be given from the start to the integration in a single database of the descriptions of the items from the many different domains, hence the cross-content database. Furthermore an integrated interface is useful for the focus group to carry out grading.

In one implementation, the catalog – that is the set of all the items- is broken
20 down into a number of sub-catalogs. The implementation stores a map of pseudo-distances for each sub-catalog, and is useful in cases where it would be too costly, in terms of memory and processor resources to store a single fully connected map for the full catalog. The sub-catalogs are not necessarily disjoint. In one variation pseudo-distances are initially calculated for the sub-catalogues, or zones, and only the zones
25 found to be closest are then investigated for the individual recommendations. The result is to reduce the amount of processing that needs to be carried out for individual requests.

Reference is now made to FIG. 4, in which the system is shown as divided into three main components, a database 300, a recommendation server 310 and an
30 application server 320.

1. **Database:**

The database holds all the required data of the catalogs and also user data. It is divided into sections according to data type:

- General tables – hold general data on types of items and constant data.
- 5 • Content type specific tables – hold the catalog data of the specific zone, for example music, music DVD, DVD, audio books, games, wallpapers, dating, houses etc. Each content type data is held in a different group of tables.
- User tables – hold general data on user and the user profiles.
- Common tables – hold data which connect items from all zones or content types to
10 items in the user tables.

2. **Application Server:**

The Application Server is used to communicate with the users and perform the non-algorithmic actions of the service such as textual search, musical DNA search,
15 searching according to item parameters, similar items search, retrieves charts, hit parades, similar users etc.

The Application Server parses the user requests, retrieves the necessary data on the user and the catalog from the database, builds the answer to the user request and sends it to the user. It also saves to the database changes in the user profile such as new
20 recommended items, new rated items by the user and also changes in the user data such as new address, phone or password as necessary. The latter may be required in an embodiment for mobile telephones, whereas they would be superfluous in a shop where anonymous users enter and use computerized service stations.

When ever there is a need for a service that requires the Recommendation Server
25 (the algorithmic engine), the Application server sends the requests with the appropriate data on the user profile, to the Recommendation engine and retrieves the required data from it.

The recommendation server is used to create prediction lists on specific content types (specific zones) or on some content types (some zones) or all content types (all
30 zones) depending on the required implementation in the application, using the methodologies outlined above.

The Recommendation Server then receives requests from the Application Server and returns a list of items that best fit the user's taste.

The data for predicting user's taste is held in a number of maps that are divided into groups according to specific content type (zones) and quality of the sub-catalogs. When a request is made by the user to retrieve a list of items according to his taste, the Recommendation Server runs the user profile thru some or all of the maps, according
5 to the request type, retrieves from each map a list of recommended items and builds one combined recommendation list. The list can contain items from one zone, several zones or all zones.

The number of online maps can vary for each of the content types (zones). If the number of maps is too large then some of the maps are set offline and are set back
10 to online when other currently online maps are set to offline.

In conclusion, a focus group of dedicated workers rates large catalogs of items from different domains: say musical tracks, movie trailers, records of potential dates, cars, houses and so on. Their ratings compute a matrix of item-to-item pseudo-distances and the pseudo-distances are then combined with an end-user's own ratings
15 of a sample of items to provide a forecast or guess as to which of a very large inventory of items will reliably be appreciated by a user on whom only minimal information has to be gathered beforehand.

The technique for cross-content recommendation as described above has been tested and found adequate in the following domains: recommending movies based on a
20 sample of the user's taste on musical tracks, recommending musical tracks based on a sample of the taste of the user about movies, and recommending potential female dates to a male customer based on a sample of his musical taste.

Experimental Results are shown in the graphs of Figure 5 and Figure 6. FIG. 5 is a graph that shows the level of satisfaction of an end user with the recommendations
25 made to him by the system of the present embodiments. FIG. 5 presents two curves. Both curves measure the level of satisfaction of an end user when he is recommended movie trailers. One of the curves (the upper one) represents the level of satisfaction obtained when the user samples and grades movie trailers. The other curve (the lower one) represents the quality of recommendations when the user samples and grades
30 musical items. This last curve represents cross-content recommendations. The system recommendations are based on gradings provided by a focus group of 31 graders or raters grading a catalog of 382 movie trailers and 1000 songs. The graph shows that as the end user himself rated a successively larger sample his results were improved,

although further improvement was slow after the end user had rated more than 30 sample items. The graph shows that the quality of cross-content recommendations is only slightly inferior to that of recommendations based on samples from the same medium.

5 FIG. 6 shows the same experiment as in FIG. 5, but here the end user is recommended musical items based on his sampling of musical items (upper curve) or movie trailers (lower curve). Again one sees that cross-content recommendations are only slightly inferior in quality to recommendations based on items of the same type.

10 It is appreciated that certain features of the invention, which are, for clarity, described in the context of separate embodiments, may also be provided in combination in a single embodiment. Conversely, various features of the invention, which are, for brevity, described in the context of a single embodiment, may also be provided separately or in any suitable subcombination or as suitable in any other described embodiment of the invention. Certain features described in the context of
15 various embodiments are not to be considered essential features of those embodiments, unless the embodiment is inoperative without those elements.

Although the invention has been described in conjunction with specific embodiments thereof, it is evident that many alternatives, modifications and variations will be apparent to those skilled in the art. Accordingly, it is intended to embrace all
20 such alternatives, modifications and variations that fall within the spirit and broad scope of the appended claims.

25 All publications, patents and patent applications mentioned in this specification are herein incorporated in their entirety by reference into the specification, to the same extent as if each individual publication, patent or patent application was specifically and individually indicated to be incorporated herein by reference. In addition, citation or identification of any reference in this application shall not be construed as an admission that such reference is available as prior art to the present invention. To the extent that section headings are used, they should not be construed as necessarily limiting.

WHAT IS CLAIMED IS:

1. A method for content-based recommendation across domains, comprising, using a processor:

obtaining a first set of content items in a first domain;

obtaining a second set of content items in a second domain;

obtaining rankings for content items of said first set and said second set from a group of grading users such that individual grading users provide rankings for items in said first set and items in said second set;

for individual grading users calculating individual pseudo-distance values between content items in said first set and content items in said second set as a distance function of a respective grading user's rankings in said first set and said user's corresponding ranking in said second set;

averaging said individual pseudo-distance values over all of said grading users, to calculate an overall pseudo-distance value between any given content item in said first set and each content item in said second set, thereby setting up a cross content database between said first domain and said second domain;

using said cross-content database:

presenting an end user with a sample of content items from said first set;

obtaining from said end-user sample rankings for content items in said sample;

finding items in said second set which are indicated via a grading function, said function comprising computing, for each item in the second set, an average of the end-user sample rankings weighted by corresponding pseudo-distances; and

returning said found items in said second set to said end user, thereby providing content-based recommendation from said first domain to said second domain.

2. Method according to claim 1, wherein said content items comprise one member of the group consisting of: items downloadable to communication devices

over cellular networks, items for sale in stores, and items downloadable over a computer network.

3. Method according to claim 1, wherein said distance function comprises calculating a difference between said respective rankings.

4. Method according to claim 1, wherein said grading function comprises applying a grading value in inverse proportion to a respective pseudo-distance.

5. Method according to claim 1, comprising dividing said second set into a plurality of separately configured zones, such that content of each zone can be returned independently.

6. Method according to claim 5, further comprising calculating gradings for respective zones based on pseudo-distances weighted by a respective end-user's sample rankings, and returning content items additionally based on resulting gradings of zones.

7. Method according to claim 5, further comprising providing said end user with a user interface to allow selections of one or more of said plurality of zones.

8. Method according to claim 5, comprising dividing each zone into separately configured maps such that each zone is configurable with some or all of its maps at any given time, the remaining maps being placed in an offline state.

9. Method according to claim 8, further comprising selecting a number of maps to remain offline depending on available processing resources.

10. Apparatus for content-based recommendation across domains, comprising:

an domain content item input unit configured for obtaining a first set of content items in a first domain and a second set of content items in a second domain;

a grading unit, associated with said domain content item input unit, configured for obtaining rankings for content items of said first set and said second set from a group of grading users such that individual grading users provide rankings for items in said first set and items in said second set;

a pseudo-distance calculation unit, associated with said grading unit, configured for:

- a) calculating individual pseudo-distance values between content items in said first set and content items in said second set as a distance function of a respective grading user's rankings in said first set and said user's corresponding ranking in said second set; and
- b) averaging said individual pseudo-distance values over all of said grading users, to calculate an overall pseudo-distance value between any given content item in said first set and each content item in said second set, thereby setting up a cross content database between said first domain and said second domain;

a sample output, associated with said cross-content database, for presenting an end user with a sample of content items from said first set;

a sample ranking input associated with said cross-content database configured for obtaining from said end user sample rankings for content items in said sample and applying said rankings to said cross content database;

a recommendation discovery unit associated with said cross-content database for finding items in said second set which are indicated via a grading function of sample rankings weighted by respective pseudo-distances; and

an end user output for returning said found items in said second set to said end user, thereby providing content-based recommendation from said first domain to said second domain.

11. Apparatus according to claim 10, wherein said content items comprise one member of the group consisting of: items downloadable to communication devices over cellular networks, items for sale in stores, and items downloadable over a computer network.

12. Apparatus according to claim 10, wherein said distance function comprises a difference between said respective rankings.

13. Apparatus according to claim 10, wherein said grading function comprises applying a grading value in inverse proportion to a respective pseudo-distance.

14. Apparatus according to claim 10, wherein said cross-content database is modifiable by dividing said second set into a plurality of separately configured zones, such that content of each zone can be returned independently.

15. Apparatus according to claim 14, further comprising a user interface to allow end-user selections of one or more of said plurality of zones.

16. Apparatus according to claim 14, wherein said cross-content database is configured to allow division of each zone into separately configured maps such that each zone is configurable with some or all of its maps at any given time, the remaining maps being placed in an offline state.

17. Apparatus according to claim 16, wherein said cross-content database comprises a map selection unit configured for selecting a number of said maps to remain offline depending on available processing resources.

18. A method of creating a cross-content database for content-based recommendation across domains, comprising, using a processor:

obtaining a first set of content items in a first domain;

obtaining a second set of content items in a second domain;

obtaining rankings for content items of said first set and said second set from a group of grading users such that individual grading users provide rankings for items in said first set and items in said second set;

for individual grading users calculating individual pseudo-distance values between content items in said first set and content items in said second

set as a distance function of a respective grading user's rankings in said first set and said user's corresponding ranking in said second set;

averaging said individual pseudo-distance values over all of said grading users, to calculate an overall pseudo-distance value between any given content item in said first set and each content item in said second set, thereby setting up said cross content database between said first domain and said second domain.

19. A method for content-based recommendation across domains, comprising, using a processor:

obtaining access to a cross content database, said cross content database comprising:

a first set of content items in a first domain;

a second set of content items in a second domain;

and pseudo-distances linking content items of said first domain to content items of said second domain, said pseudo-distances representing distances between gradings provided to said items by individual grading users averaged over a group of grading users, and

using said cross-content database:

presenting an end user with a sample of content items from said first set;

obtaining from said end user sample rankings for content items in said sample;

finding items in said second set which are indicated via a grading function of respective overall pseudo-distances weighted by said sample rankings; and

returning said found items in said second set to said end user, thereby providing content-based recommendation from said first domain to said second domain.

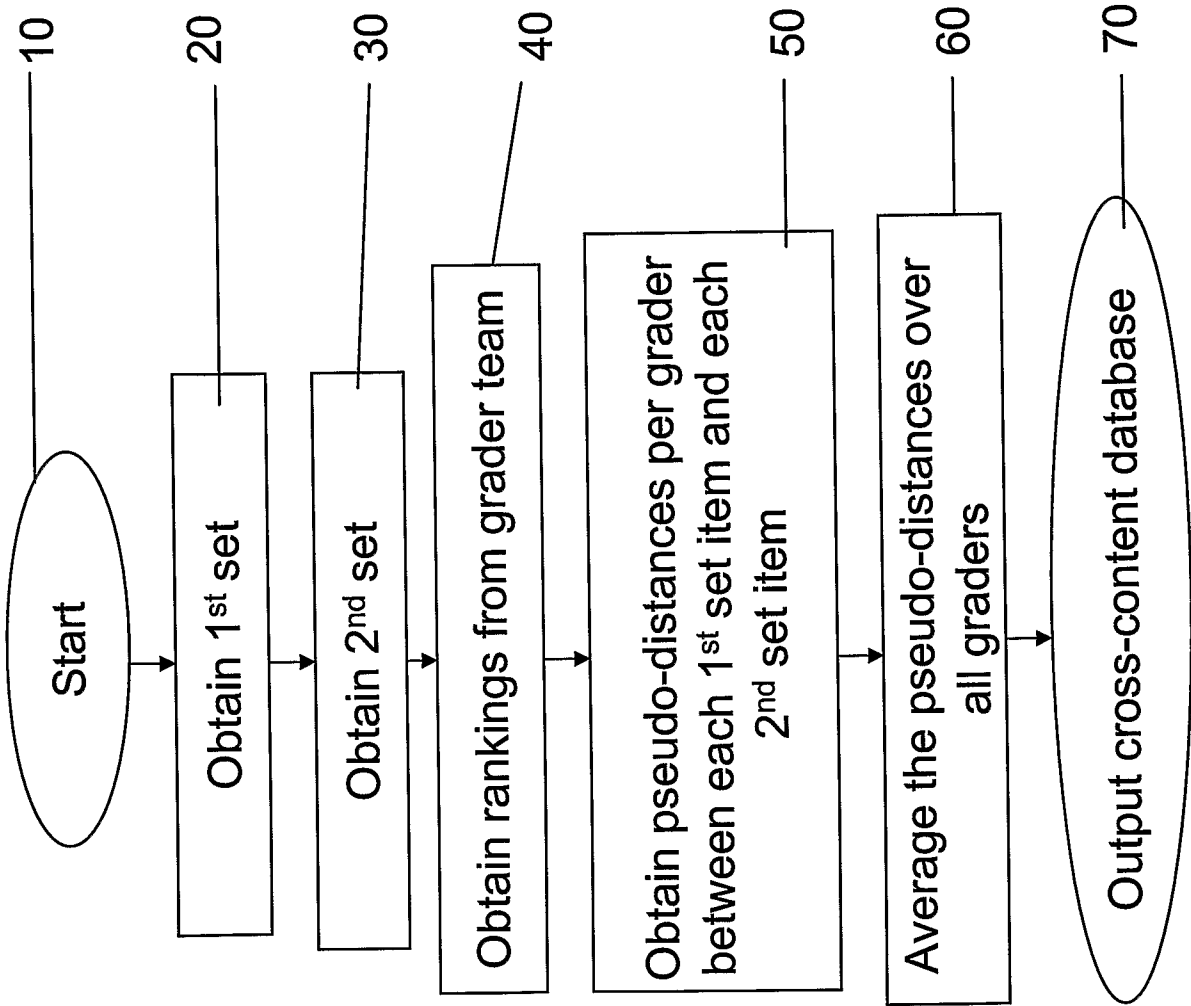


FIG. 1

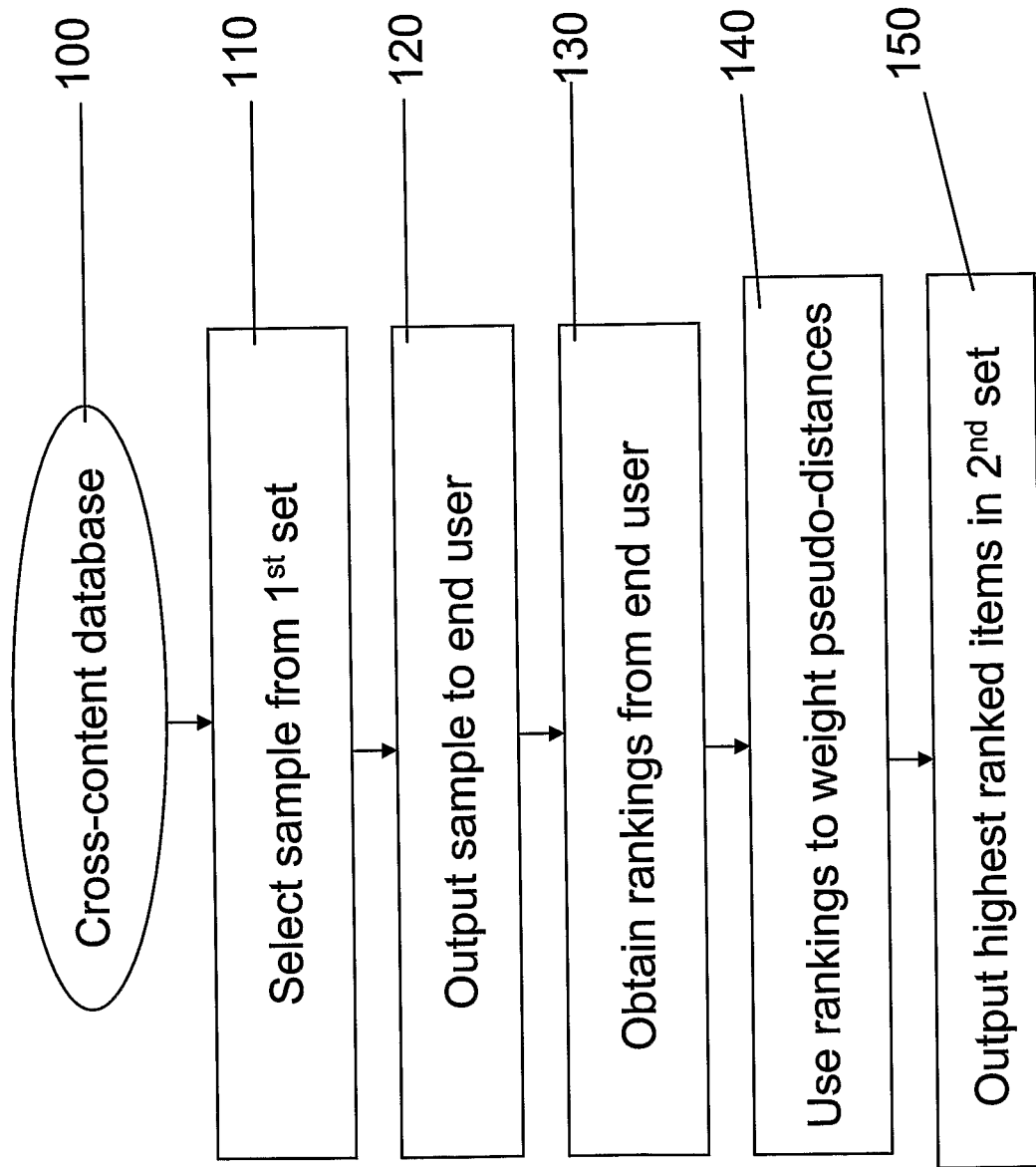


FIG. 2

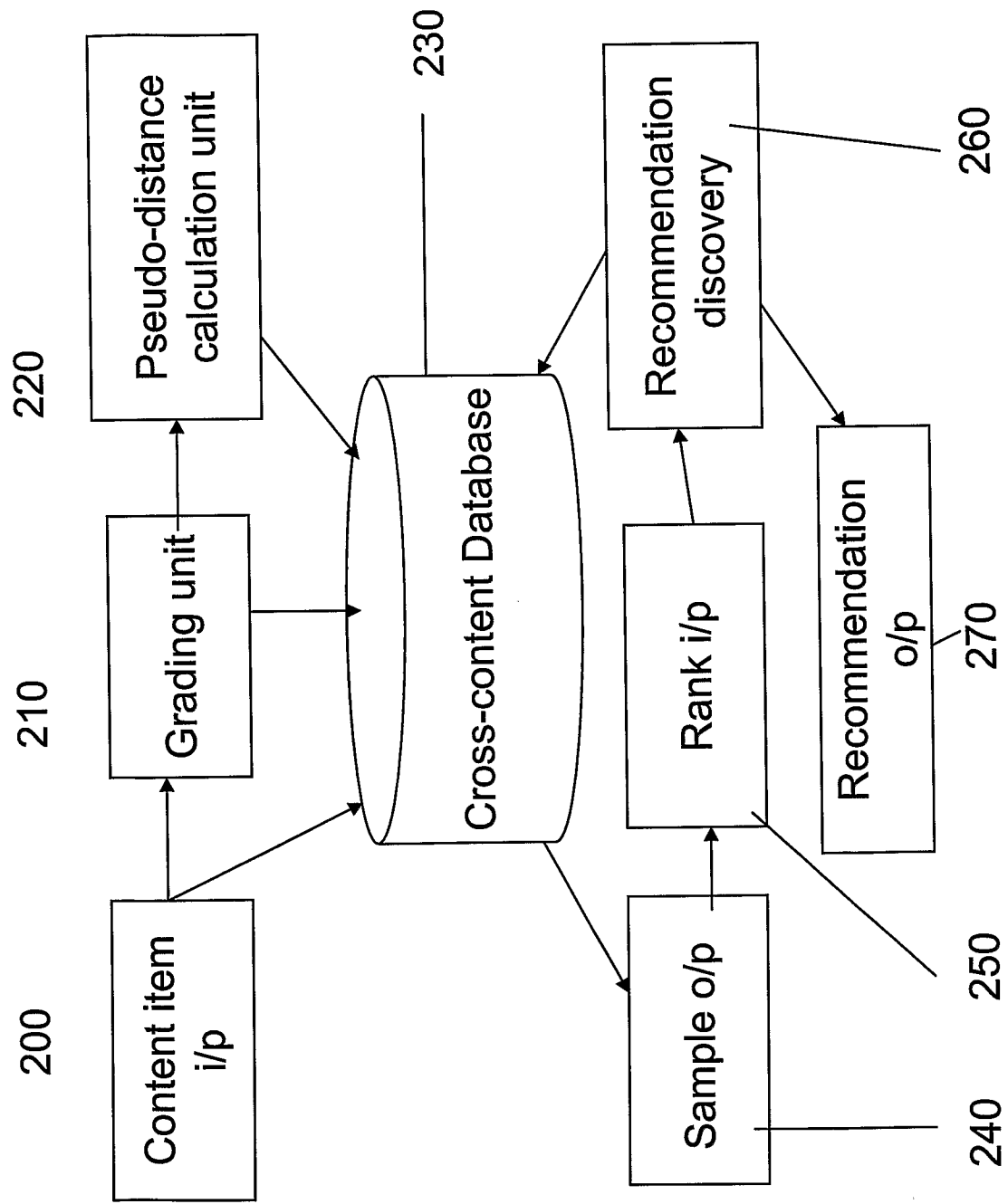


FIG. 3

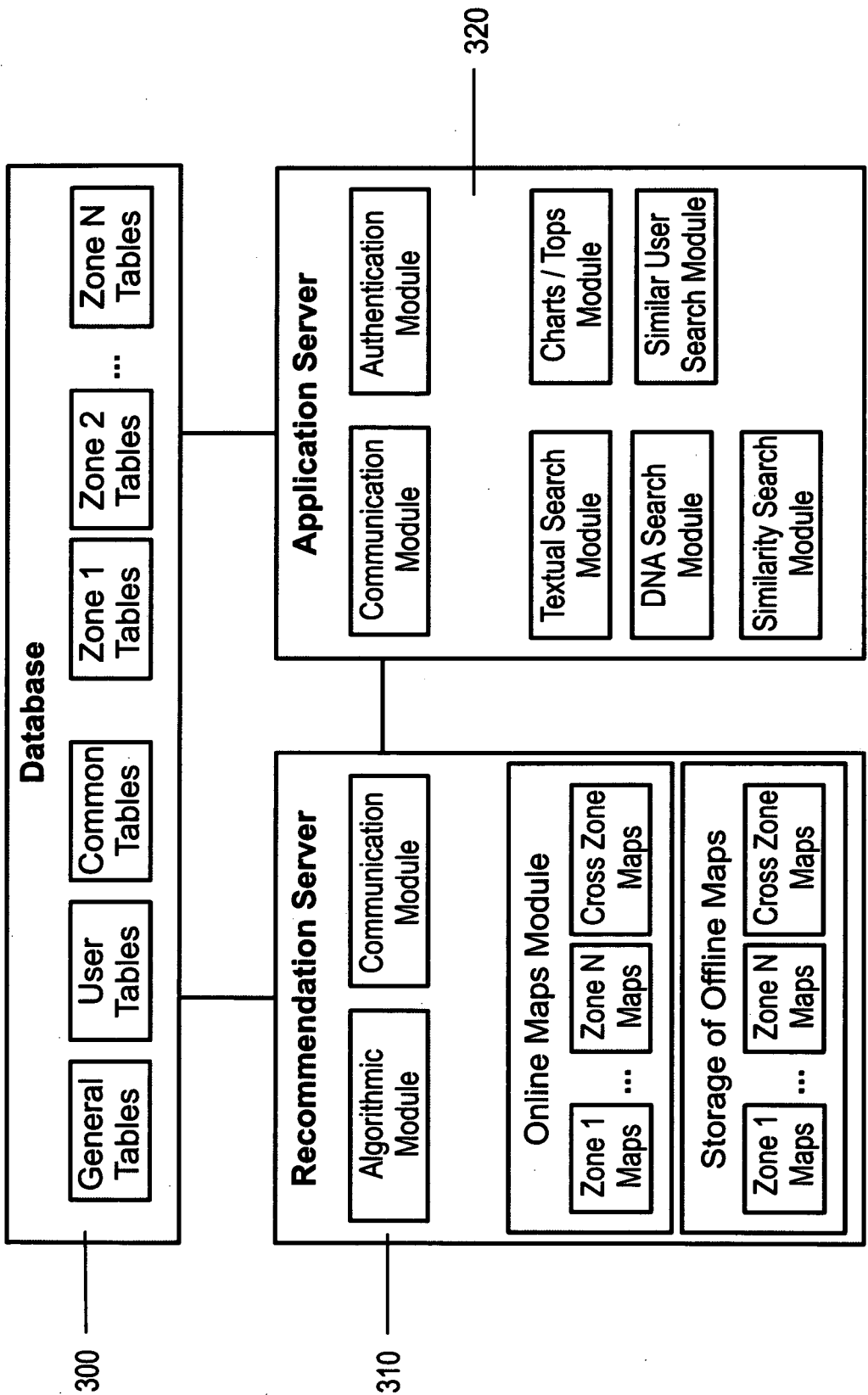


FIG.4

Catalog: 382
movies trailers
and 1000 songs
Focus group: 31
raters

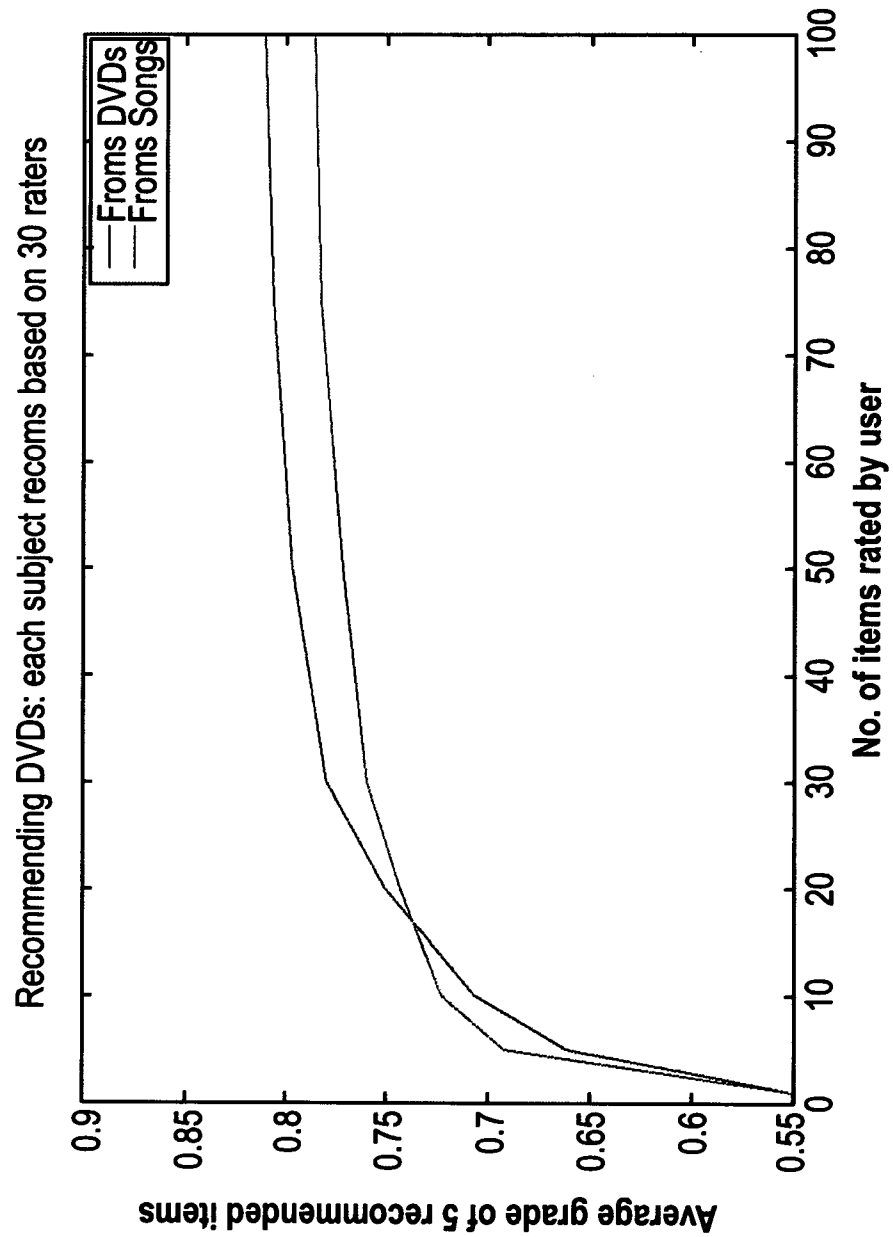


FIG. 5

Music Recommendations - Cross

Catalog: 382 movies trailers and 1000 songs
Focus group: 31 raters

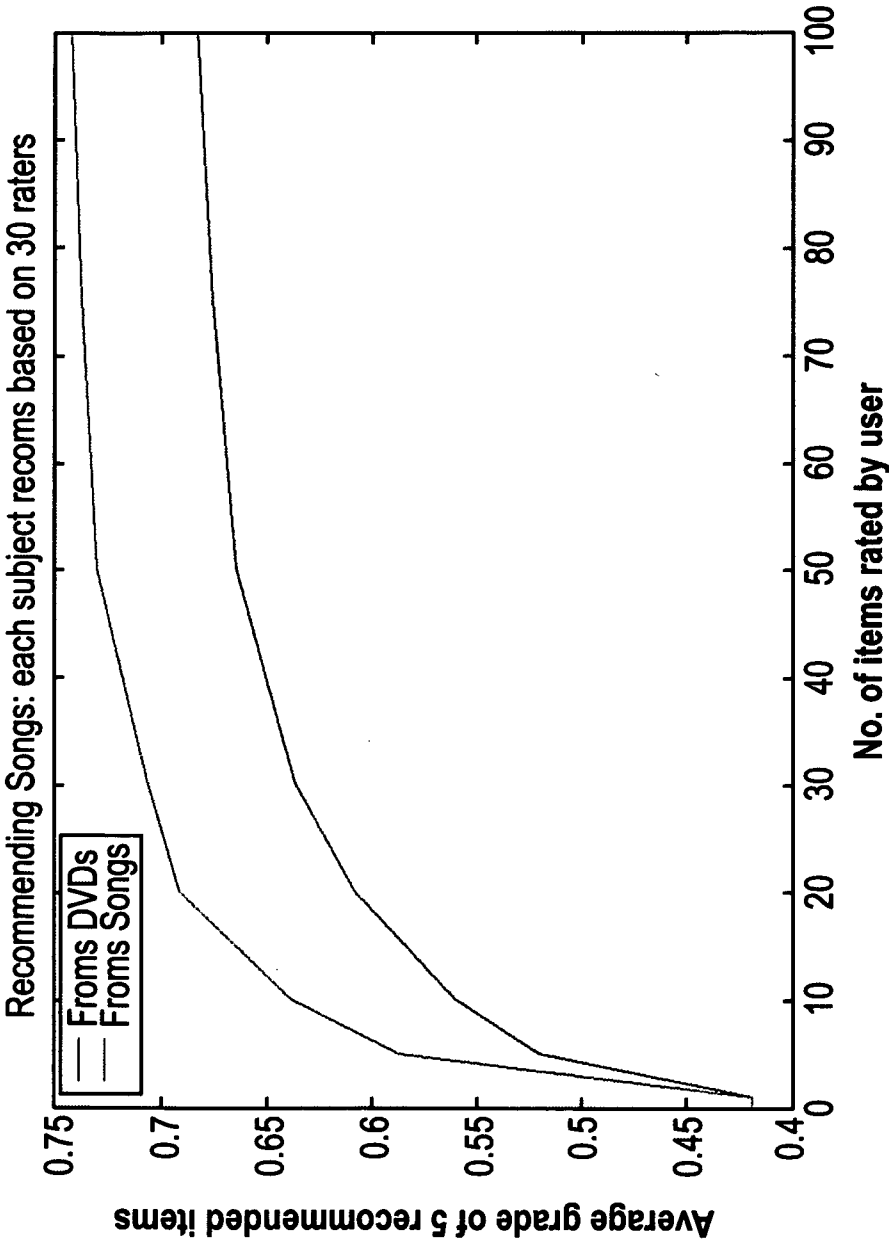


FIG. 6