

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2016 年 6 月 9 日 (09.06.2016)



(10) 国际公布号
WO 2016/086802 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2015/095834
- (22) 国际申请日: 2015 年 11 月 27 日 (27.11.2015)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201410736666.3 2014 年 12 月 4 日 (04.12.2014) CN
- (71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [GB/CN]; 英国开曼群岛大开曼乔治城资本大厦一座四层 847 号邮箱, Grand Cayman (GB)。
- (72) 发明人: 及
- (71) 申请人 (仅对美国): 潘晓彤 (PAN, Xiaotong) [CN/CN]; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。金柯 (JIN, Ke) [CN/CN]; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。刘忠义 (LIU,

Zhongyi) [CN/CN]; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。魏虎 (WEI, Hu) [CN/CN]; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

(74) 代理人: 北京国昊天诚知识产权代理有限公司 (CO-HORIZON INTELLECTUAL PROPERTY INC.); 中国北京市朝阳区西坝河西里 28 号英特公寓 C 座 104, Beijing 100028 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA,

[见续页]

(54) Title: COMPUTER SYSTEM-BASED RECOMMENDATION METHOD AND DEVICE THEREOF

(54) 发明名称: 基于计算机系统的推荐方法及其装置

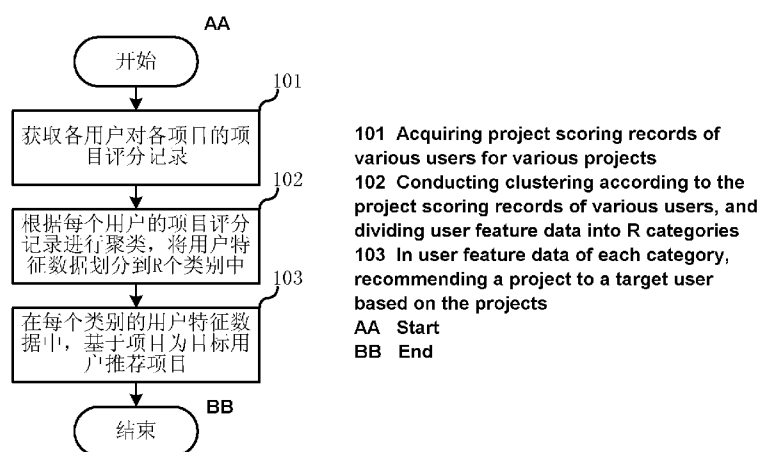


图 1 /FIG. 1

(57) Abstract: The present invention relates to recommendation technology of computer system implementation. Disclosed are a computer system-based recommendation method and a device thereof. The recommendation method of the present invention comprises: first, performing clustering according to a project scoring record of each user, so as to divide user characteristic data into a plurality of categories; then, in the user characteristic data of each category, based on a project, recommending the project to a target user, so that an efficient recommendation method can be realized with big data, thereby guaranteeing the stability of the system and the diversity of the recommendations. In addition, each calculation node does not need to save user characteristic data of all the categories, thereby avoiding the problem of an insufficient memory.

(57) 摘要: 本发明涉及以计算机系统实现的推荐技术, 公开了一种基于计算机系统的推荐方法及其装置。在本发明的推荐方法中, 先根据每个用户的项目评分记录进行聚类, 将用户特征数据划分到多个类别中, 再在每个类别的用户特征数据中基于项目为目标用户推荐项目, 可以在大数据下实现高效的推荐方法, 保证了系统的稳定性和推荐的多样性。此外, 每个计算节点不需要保存所有类别的用户特征数据, 避免了内存不足的问题。



WO 2016/086802 A1



RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG,

CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

基于计算机系统的推荐方法及其装置

技术领域

本发明涉及以计算机系统实现的推荐技术，特别涉及基于计算机系统的推荐方法及其装置。

5

背景技术

推荐算法通常分为基于内容的推荐，基于关联规则的推荐，基于协同过滤推荐，以及一些基本方法的组合。然而，本发明的发明人发现，当前 CF (Collaborative Filtering, 协同过滤) 算法存在一些问题，特别是在分布式环境下，有些问题更加明显，从 CF 运行逻辑来理解，算法瓶颈主要在以下三个地方：

第一点存在于数据规模中，无论哪次推荐，分布式框架的每个计算节点都要保留全局数据，因为每个 reducer 不能提前得知当前节点被分配的是哪些用户，所以只存储局部数据会影响数据精度。这时每个 reducer 就被实例化为一个小型的推荐场景。假设共有 t 单位的计算资源，则全局数据被冗余存储了 t-1 份，同时每个 reducer 在真正的推荐过程中只会遇到小部分数据计算，其它数据也会造成极大的资源浪费。因此当数据规模较大时，无论从时间上还是存储上，对每个计算节点都是巨大的负担。在我们的实验过程中，由于编程语言以及编译器的本地设计，当用户或者项目任一数据量超过千万级时，必然会出现数组过大越界问题，当用户或者项目任一数据量在千万级别时，则由于集群中各个计算节点的配置参差不齐，有些低配节点就会出现内存不足问题。

第二点为数据倾斜问题。从 CF 算法过程来看，无论是基于项目还是基于用户，我们都需要计算项目之间的相似度。这里存在一个隐蔽的问题：实际应用场景中，有些项目属于“活跃份子”，有些属于“不活跃份子”，例如在

25

使用 MapReduce framework 时，在 $\langle \text{key}, \text{value} \rangle$ 数据 schema (模式) 下，有些 key 对应的 value 会很多，有些会很少，这种数量不一致，参差不齐的情况，称为数据倾斜(data skew)。当 value 数量在不同 key 之间相差 3 个以上数量级时，在计算项目之间相似度过程中就会造成严重的数据倾斜，“活跃份子”导致计算时间长尾。同理，在推荐过程中，有些用户之前积累的行为多，有些用户之前积累的行为少，这时“活跃用户”就会拖累整体计算过程。

第三点为数据稀疏问题。在对象集合中，产生关系的对象对很少；可以理解把所有对象划分为一个矩阵，其中 (i,j) 表示第 i 个用户和第 j 个项目之间的关系，如果大多数点均为 0 (表示没有关系)，则定义为数据稀疏。数据稠密与之相反。特别是初始数据往往是不完全的，这时在计算项目之间相似度时就很容易出现数据稀疏问题，即用户项目矩阵的大部分位置都是 0。

发明内容

15 本发明的目的在于提供一种基于计算机系统的推荐方法及其装置，可以在大数据下实现高效的推荐方法，保证了系统的稳定性和推荐的多样性。

为解决上述技术问题，本发明的实施方式公开了一种基于计算机系统的推荐方法，该方法包括以下步骤：

获取各用户对各项目的项目评分记录；

20 根据每个用户的项目评分记录进行聚类，将用户特征数据划分到 R 个类别中， R 是大于 1 的整数；

在每个类别的用户特征数据中，基于项目为目标用户推荐项目。

本发明的实施方式还公开了一种基于计算机系统的推荐装置，装置包括：

25 用户项目初始关系计算模块，用于获取各用户对各项目的项目评分记录；

聚类模块，用于根据用户项目初始关系计算模块获取的每个用户的项目评分记录进行聚类，将用户特征数据划分到 R 个类别中， R 是大于 1 的整数；以及

5 推荐模块，用于在聚类模块所划分的每个类别的用户特征数据中，基于项目为目标用户推荐项目。

本发明实施方式与现有技术相比，主要区别及其效果在于：

在本发明的推荐方法中，先根据每个用户的项目评分记录进行聚类，将用户特征数据划分到多个类别中，再在每个类别的用户特征数据中基于项目为目标用户推荐项目，可以在大数据下实现高效的推荐方法，保证了系统的
10 稳定性和推荐的多样性。

进一步地，每个计算节点不需要保存所有类别的用户特征数据，避免了内存不足的问题。

进一步地，对于每个类别中的每个项目或每个用户，只选取与其关系最强的几个项目，而不是保留与其有关系的所有项目，可以避免关系较弱的项目产生的数据倾斜问题。
15

进一步地，采用数据稀疏度对数据稀疏问题进行检测，并在发现数据稀疏问题后，通过项目间的二度关系进行相似度补全，以避免数据稀疏对推荐准确度的影响。

进一步地，根据用户数量来选择是否要对用户进行聚类，以更好地适应于小数据下和大数据下的项目推荐。
20

附图说明

图 1 是本发明第一实施方式中一种基于计算机系统的推荐方法的流程示意图；

25 图 2 本发明第一实施方式中一种基于计算机系统的推荐方法中聚类判断的流程示意图；

图 3 是本发明第二实施方式中一种基于计算机系统的推荐方法中推荐步骤的流程示意图；

图 4 是本发明第二实施方式中一种基于计算机系统的推荐方法中推荐步骤的流程示意图；

5 图 5 是本发明第二实施方式中一种基于计算机系统的推荐方法中推荐步骤的流程示意图；

图 6 是本发明第二实施方式中一种基于计算机系统的推荐方法中数据补全的流程示意图；

图 7 是现有的计算用户相似度的示意图；

10 图 8 和图 9 是现有的基于用户的协同过滤的示意图；

图 10 和图 11 是现有的基于项目的协同过滤的示意图；

图 12 是现有的实现分布式 CF 算法的 MapReduce 框架图；

图 13 是本发明第二实施方式中一种基于计算机系统的推荐方法的流程示意图；

15 图 14 是本发明第二实施方式中一种基于计算机系统的推荐方法的流程示意图；

图 15 是本发明第三实施方式中一种基于计算机系统的推荐装置的结构示意图；

20 图 16 是本发明第四实施方式中一种基于计算机系统的推荐装置中推荐模块的结构示意图。

具体实施方式

在以下的叙述中，为了使读者更好地理解本申请而提出了许多技术细节。但是，本领域的普通技术人员可以理解，即使没有这些技术细节和基于
25 以下各实施方式的种种变化和修改，也可以实现本申请各权利要求所要求保护的技术方案。

为使本发明的目的、技术方案和优点更加清楚，下面将结合附图对本发明的实施方式作进一步地详细描述。

本发明第一实施方式涉及一种基于计算机系统的推荐方法。图 1 是该基于计算机系统的推荐方法的流程示意图。如图 1 所示，该方法包括以下步骤：
5 步骤：

在步骤 101 中，获取各用户对各项目的项目评分记录。可以理解，在本发明的各个实施方式中，项目可以为商品、服务或其它推荐对象。

此后进入步骤 102，根据每个用户的项目评分记录进行聚类，将用户特征数据划分到 R 个类别中，R 是大于 1 的整数。可以理解，在本发明的各个
10 实施方式中，可以采用 K-means 算法直接对用户特征数据进行聚类，也可以先采用 Canopy 算法进行粗聚类，再采用 K-means 算法进行细聚类。

先采用 Canopy 算法进行粗聚类，再采用 K-means 算法进行细聚类，在保证准确性的同时，提高了聚类速度。

此外，可以理解，用户特征数据是由用户信息、项目信息和用户对项目的评分记录组成的数据。
15

此后进入步骤 103，在每个类别的用户特征数据中，基于项目为目标用户推荐项目。可以理解，在本发明的各个实施方式中，可以采用基于协同过滤、基于关联规则或基于效用的推荐算法来为目标用户推荐项目。

此后结束本流程。

当然，在本发明的其他实施方式中，也可以以项目为对象进行聚类，再在每个类别的用户特征数据中基于用户来为目标用户推荐项目，或是聚类和推荐都基于用户或都基于项目。
20

在本实施方式的推荐方法中，先根据每个用户的项目评分记录进行聚类，将用户特征数据划分到多个类别中，再在每个类别的用户特征数据中基
25 于项目为目标用户推荐项目，可以在大数据下实现高效的推荐方法，保证了系统的稳定性和推荐的多样性。

优选地，上述计算机系统为分布式系统。该计算机系统包括至少两个计算节点。

在步骤 103 中，将各类别的用户特征数据分配给多个计算节点，每个计算节点最多保存 $R-1$ 个类别的用户特征数据，每个计算节点在所保存的每个类别的用户特征数据中基于项目为目标用户推荐项目。每个计算节点不需要保存所有类别的用户特征数据，避免了内存不足的问题。

优选地，每个计算节点保存一个类别的用户特征数据并进行处理。此外，可以理解，在本发明的各实施方式中，可以根据各计算节点的配置将两个或两个以上类别的用户特征数据分配给高配置的计算节点进行处理。当然，在用户特征数据量不是很大的时候，也可以由一个计算节点进行处理。

作为可选实施方式，如图 2 所示，在步骤 102 前还包括以下步骤：

在步骤 201 中，判断用户数量是否大于用户规模阈值。若用户数量小于用户规模阈值，则进入步骤 202；若用户数量大于用户规模阈值，则进入步骤 102。

在步骤 202 中，直接在所有用户特征数据中基于项目为目标用户推荐项目。

此后结束本流程。

根据用户数量来选择是否要对用户进行聚类，以更好地适应于小数据下和大数据下的项目推荐。

此外，可以理解，在本发明的其他实施方式中，也可以不对用户数量进行判断，直接对用户特征数据进行聚类。

本发明第二实施方式涉及一种基于计算机系统的推荐方法。图 3 是该基于计算机系统的推荐方法中推荐步骤的流程示意图。

第二实施方式在第一实施方式的基础上主要进行了以下两个改进：

第一个改进为对于每个类别中的每个项目或每个用户，只选取与其关系最强的几个项目，而不是保留与其有关系的所有项目，可以避免关系较弱

的项目产生的数据倾斜问题。具体地说：

在步骤 103 中，采用基于项目的协同过滤算法为目标用户推荐项目。如图 3 所示，该步骤 103 中包括以下子步骤：

在子步骤 301 中，根据上述类别中每个用户的项目评分记录，计算上述
5 类别中所有项目间相似度，并为每个项目选取相似度最高的 M 个项目，M 为预定义的整数。

此后进入子步骤 302，根据上述类别中目标用户的项目评分记录，为目标用户选取评分最高的 T 个项目，T 为预定义的整数。

此后进入子步骤 303，将为目标用户选取的 T 个项目与为 T 个项目中每
10 个项目选取的 M 个项目相结合，并从中去除目标用户的项目列表中的项目，形成初始推荐结果。例如，为目标用户选取的 T 个项目为 A、B、C，而为项目 A、B、C 选取的 M 个项目分别为 (D, E)、(C, F) 和 (B, H)，则形成的初始推荐结果为 (D, E, F, H)。

此后结束本流程。

15 优选地，如图 4 所示，在子步骤 303 后还包括以下子步骤：

在子步骤 401 中，判断初始推荐结果中的项目数量是否大于 N，N 为预定义的整数。若初始推荐结果中的项目数量大于 N，则进入子步骤 402；若初始推荐结果中的项目数量小于 N，则进入子步骤 403。

在子步骤 402 中，从初始推荐结果中选取相似度最高的 N 个项目推荐给
20 目标用户。

此后结束本流程。

在子步骤 403 中，将目标用户的项目列表中的所有项目与为目标用户的项目列表中每个项目选取的 M 个项目相结合，并从中去除目标用户的项目列表中的所有项目，形成用户数据补全推荐结果。可以理解，用户数据补全
25 推荐结果的形成与初始推荐结果的形成类似，在此不再赘述。

此后结束本流程。

更优地，如图 5 所示，在子步骤 403 后还包括以下子步骤：

在子步骤 501 中，判断用户数据补全推荐结果中的项目数量是否大于 N，N 为预定义的整数。若用户数据补全推荐结果中的项目数量大于 N，则进入子步骤 502；若用户数据补全推荐结果中的项目数量小于 N，则进入子步骤 503。

在子步骤 502 中，从用户数据补全推荐结果中选取相似度最高的 N 个项目推荐给目标用户。

此后结束本流程。

在子步骤 503 中，将目标用户的项目列表中的所有项目与和目标用户的项目列表中每个项目具有相似度关系的所有项目相结合，并从中去除目标用户的项目列表中的所有项目，形成项目数据补全推荐结果。可以理解，项目数据补全推荐结果的形成与初始推荐结果的形成类似，在此不再赘述。

此后结束本流程。

第二个改进为采用数据稀疏度对数据稀疏问题进行检测，并在发现数据稀疏问题后，通过项目间的二度关系进行相似度补全，以避免数据稀疏对推荐准确度的影响。具体地说：

如图 6 所示，在步骤 103 后还包括以下步骤：

在步骤 601 中，判断数据稀疏度是否大于数据稀疏度阈值，数据稀疏度 $DSP = \frac{k}{P_l^2}$ ，其中 k 为计算得到的类别中具有相似度关系的项目对的数量，l 为类别中项目的数量， $P_l^2 = \frac{l(l-1)}{2}$ 。若数据稀疏度小于数据稀疏度阈值，则进入步骤 602；若数据稀疏度大于数据稀疏度阈值，则进入步骤 603。

在步骤 602 中，以第一项目、第二项目和第三项目为一组，第一项目与第二项目间、第二项目与第三项目间具有相似度关系，通过第二项目为第一项目和第三项目建立相似度关系，并根据补充的项目间相似度关系在类别中再次基于项目为目标用户推荐项目。

此后结束本流程。

在步骤 603 中，将基于项目计算得到的推荐项目推荐给目标用户。

此后结束本流程。

此外，可以理解，若通过项目间的二度关系对数据进行相似度补全后
5 仍然存在数据稀疏问题，可以继续通过项目间的三度、四度或更高度关系对数据进行相似度补全，以避免数据稀疏对推荐准确度的影响。

通常推荐算法分为基于内容的推荐、基于关联规则的推荐、基于协同过滤的推荐，以及一些基本方法的组合。基于内容的推荐根据用户（user）和项目（item）在某些属性上的相似性程度进行推荐，典型的如向量空间模型；
10 基于关联规则的推荐以关联规则为基础，把已购项目作为规则头，规则体为推荐对象；基于协同过滤的推荐会挖掘项目之间或者用户之间的深度关系，根据用户的群体行为规律（即购买了该项目的人群会倾向于什么别的项目？）为用户做推荐，例如推荐强关系项目。当两个用户(项目)之间具有强关系时，是指两者具有较高相似度，弱关系与之相反。

15 上述协同过滤有两种实现方法，第一种是基于用户（user-based），第二种是基于项目（item-based）。

1. 基于用户的协同过滤

顾名思义，user-based 协同过滤会先计算出当前用户最相似的 n 个相邻用户，在推荐过程中选定 n 个邻居用户的偏好项目，在计算用户之间相似度
20 时，需要根据两个用户的项目偏好来计算，如图 7 所示。

整个过程通过用户之间的关系来建立联系，用户之间的具体关系通过项目作为中介介质来计算。如图 8 和图 9 所示，具体步骤可以如下：

（1）计算出当前用户（即目标用户）的邻居列表，在计算过程中要利用当前用户与任一邻居的项目偏好列表，把项目之间的关系作为用户之间关系的桥梁。
25

（2）取 Top n 个邻居用户，作为推荐候选。

(3)在 Top n 个邻居用户中找出未在当前用户偏好列表中出现的项
目，建立推荐候选列表 (candidate list)。

(4)对 candidate list 中的每个 item i，计算其与当前用户的偏好列表
中每个项目的偏好，并得出最终得分 (final score)。

5 (5)对 candidate list 中的每个 item i，按照 final score 排序，取 Top m
个项目作为推荐结果。

2. 基于项目的协同过滤

item-based 协同过滤根据用户-项目关系，先计算出项目之间相似度，
根据当前用户已有行为，推荐其最相似的 n 个项目，如图 10 和图 11 所示。

10 整个流程通过项目之间的相似度来建立联系，具体步骤可以如下：

(1)通过用户作为桥梁，计算出 item i 与 item j 之间的相似度。

(2)构造一个矩阵，点(i, j)表示 item i 与 item j 之间的相似度。

(3)对当前用户的偏好列表中的每个 item，计算其 Top n 相似的
items。

15 (4)把所有相似的 items 按照 score 排序，将 Top n items 作为推荐结
果。

在这两种 CF 算法中，都需要进行相似度计算，但整体算法框架并不局
限于某种特定的相似度计算方法，系统只是把相似度计算设计为开放式接
口，实际上我们可以使用多种相似度算法，例如 Euclidean distance (欧几
里得距离)、jaccard coefficient (杰卡德系数) 等等。

20 在应用场景中，我们很难说清哪种算法更加优秀，算法性能取决于实
际数据分布：

1.当 item-item 矩阵比较稠密，大部分 item 之间的关系都可以通过一个
score 表达时，并且这种关系具有较好区分度时(即 score 分布较均匀，而不
是局限于某个区间)，item-based 算法往往能表现得更好。

2. 另外一个选择 item-based 算法的场景是 item 数量明显小于 user 数

量；反之，如果 user 数量小于 item 数量，则选择 user-based 算法。

3. 数据稳定性也是选择算法的一个参考因素，item 和 user 哪个更加稳定，哪种算法往往就会获得较好效果。

4. 如果我们追求推荐的多样性而非准确度，user-based 算法会表现更好。

以上一些经验并非总是有效，在实际应用中，要通过大量实验来找出较好的推荐方案。

如何评价一个推荐系统的推荐效果，业界也没有统一的标准，除了 machine learning（机器学习）中常用的 precision/recall（查准/召回）等指标外，我们有时也会关注推荐的多样性，即一个用户的推荐结果是否足够丰富。

在大数据时代，单机版的推荐算法已经很难施展拳脚了，应用 MapReduce framework（框架）、hadoop framework 已经实现了一套完整的 CF 算法，算法包名是 Mahout，它不但实现了 item-based 与 user-based 算法，而且实现了多种相似度以及邻居算法。此外，在更高级的计算框架下也可以实现协同过滤，例如 Spark 框架。

User-based 算法：

（1）建立 data model（数据模型），初始化 user2item 和 item2user 数据结构

（2）根据 user-item-neighborhood relationship，利用某种相似度算法，对每个用户在全局(所有用户)计算 Top n neighborhood

（3）利用 user-neighborhood-item relationship，计算 possible items

（4）利用 item-possible item similarity，为当前用户推荐

Item-based 算法：

（1）建立 data model，初始化 user2item 和 item2user 数据结构

（2）根据 user-item-user-item relationship，计算每个用户的 possible

items

(3) 计算 possible item 与当前 user 的相关度:

$$pref_{i2i}(j) = \sum_{i=0}^n sim_{i2i}(i, j) * pref(i)$$

$$sim_{i2i}(j) = \sum_{i=0}^n sim_{i2i}(i, j)$$

$$preference(j) = \frac{pref_{i2i}(j)}{sim_{i2i}(j)}$$

(4) 按照 preference score 排序, 选择高分者作为 recommendation

5 items (推荐项目)。

上述 MapReduce framework 是一种分布式计算框架, 将一个任务解析为 map 过程与 reduce 过程, 其中 map 过程的输出为 <key,value> schema (模式), reduce 过程针对每个 key 对其所有 value 做具体算法。如图 12 所示, 为了实现分布式 CF 算法, 在 MapReduce framework 中, 我们需要在 map 过程中整理输入数据, 例如解析输入数据、加载初始数据 schema (模式)、将数据统一为 <key, value> 格式, 其中 key 为 userID(用户标识), value 为 itemID (项目标识) 和 score。并在 reduce 过程中初始化 Mahout data model 以及一些全局数据结构 (neighborhood 对象、recommender 对象、similarity 对象等等), 然后进行真正的推荐过程 (user-based 或 item-based 推荐)。

但是, 现有的 CF 算法还存在大数据问题、数据倾斜问题和数据稀疏问题。通过上述基于计算机系统的推荐方法可以解决以上这些问题。以下将从这三个方面对该基于计算机系统的推荐方法进行进一步的详细描述。

1. 聚类方法解决大数据问题

20 在数据规模较大的实际应用场景中, 例如亿级别数据量中, 我们使用聚类方法来把问题降解。聚类是一种无监督学习算法, 对于某一类对象, 比如用户或者项目, 根据对象属性将其划分到多个类别中, 不需要人工标注, 即在任何人工干预前提下, 我们把每个 item 表示为一个 feature list (特征

列表)，聚类算法会自动完成 cluster（集群）过程。

优选地，我们选取 user 作为聚类对象，即把从 feature 上来说相似的 user 聚到同一类中；这里为什么不选取 item 作为聚类对象呢？原因是如果我们选用 item 作为聚类对象，最终的聚类结果中某个类别的 items 只会局限在某几个 item 上，这样与推荐多样性指标相悖，影响推荐结果的多样性，所以我们以 user 作为聚类结果。另一个原因是我们采用 item-based 算法作为主体推荐算法，如果在聚类过程中还是使用 item 来聚类，在某种程度上会与 item-based 推荐产生重复，这样也会影响算法结果的多样性。当然，在本发明的其他实施例中，也可以采用 user-based 作为主体推荐算法，选取 item 作为聚类对象。

准备 Feature：我们把每个 user 当作一个 object（对象），然后将这个 user 特征化，该用户的每条历史记录被看作一个 feature，例如 user i 有一条记录 $\langle i, t, s \rangle$ ，表示用户 i 对 item t 的偏好是 s，则我们为它加一个 feature “t:s”，这样每个 user 都被特征化了。

可选地，聚类的规模是这样计算的，大约 10,000,000 users 会被聚到一个类别中，这可以保证在分布式计算平台上不会出现死机现象。当然，根据实际需要可以设置将其他数量的用户聚到一个类别中。

聚类算法的瓶颈在于计算 item 之间相似度上，优选地，我们使用 Canopy 算法来确定初始中心，然后用 Kmeans 来做最终聚类。Canopy 算法会先将全部数据划分到 r 个子集中，两个子集中可能出现数据重叠，然后在每个子集中用 Kmeans 算法聚类，不同子集中的数据之间不会进行相似度计算。聚类方法的流程图如图 13 所示。当然，在本发明的其他实施例中，也可以使用 Kmeans 算法或其他聚类算法对全部数据进行聚类。

其中，Canopy 算法过程具体如下：

（1）将数据集向量化得到一个 list（列表）后放入内存，选择两个距离阈值：T1 和 T2，其中 $T1 > T2$ ，T1 和 T2 的值可以用交叉校验来确定；

(2) 从 list 中任取一点 P, 用低计算成本方法快速计算点 P 与所有 Canopy 之间的距离 (如果当前不存在 Canopy, 则把点 P 作为一个 Canopy), 如果点 P 与某个 Canopy 距离在 T1 以内, 则将点 P 加入到这个 Canopy;

5 (3) 如果点 P 曾经与某个 Canopy 的距离在 T2 以内, 则需要把点 P 从 list 中删除, 这一步是认为点 P 此时与这个 Canopy 已经够近了, 因此它不可以再做其它 Canopy 的中心了;

(4) 重复步骤 2、3, 直到 list 为空结束。

2. 重构 CF 算法, 用 top N 方法解决数据倾斜问题

10 如图 14 所示, 重构的 CF 算法主要如下:

(1) 根据每个 user 的历史数据记录, 计算同一 user 下的不同 item 之间关系, 数据 schema 为 <item1, score1, item2, score2>。

(2) 以 item1_item2 为 key, 计算两个 item 之间的相似度。

15 (3) 每个 item 只保留 top M 个相似 items, 形成 toplItemList, 在为用
户推荐时也从这个 toplItemList 中取数据。

(4) 在 userItemList (即用户的项目列表) 中, 每个 user 只取 top T 个 items, 生成 betterItemList (即用户的偏好列表)。

20 (5) 从每个 user 的 betterItemList 中取出 items, 结合每个 item 的
toplItemList, 过滤掉已经有行为的 items, 生成 itemCandidateList (即初始
推荐结果)。

(6) 如果 itemCandidateList 中 item 个数少于 N, 则先还原 betterItemList 为 userItemList, 如果 itemCandidateList 中 item 个数仍不足, 则还原 toplItemList 为全部数据。

25 (7) 在 itemCandidateList 中按照相似度以及用户偏好计算出 top N
items 作为推荐结果。

3. 解决数据稀疏方法

在实验过程中，我们发现有些实验数据会出现严重数据稀疏问题，即在计算 item 之间相似度时，只有很少一部分 item pair (项目对) 有关系，大部分 item 之间无直接关系，因此我们定义了数据稀疏度： $DSP = \frac{k}{P_i^2}$ ，其中 P_i 为通过 CF 算法计算出的 i2i pair 数量，k 为不同的 item 数量，该度量值越小则数据越稀疏。可以理解，在本发明的其他实施例中，也可以使用其他数据稀疏度定义来检测数据稀疏问题。

优选地，解决数据稀疏的方法主要如下：

(1) 传统方法计算 CF

(2) 统计结果 DSP，如果 DSP 小于 threshold (即数据稀疏度阈值)，
10 则做 i2i 补全；具体 threshold 定义为 $DST = \alpha$ ，其中 α 自定义

(3) I2i 补全算法为 itemA->itemB->itemC，即利用中间 item 为两边的 item 建立联系，其中 itemA 与 itemB，itemB 与 itemC 为邻居。例如，itemA 与 itemB 具有相似度 S_{AB} ，itemB 与 itemC 具有相似度 S_{BC} ，则 itemA 与 itemC 具有相似度 $S_{AC} = S_{AB} * S_{BC}$ ，或 $e^{-(S_{AB} * S_{BC})}$ 。

15 实验证明，补全算法可以普遍增加 30% 新数据，为推荐做了有力的数据补充。

以上仅为本发明的一个优选实施例，各改进组合后形成本发明的较佳实施例，但各改进也可以分别使用。并且，在上述实施例中提到的各参数也可以根据需要进行相应设置。

20 本发明的各方法实施方式均可以以软件、硬件、固件等方式实现。不管本发明是以软件、硬件、还是固件方式实现，指令代码都可以存储在任何类型的计算机可访问的存储器中（例如永久的或者可修改的，易失性的或者非易失性的，固态的或者非固态的，固定的或者可更换的介质等等）。同样，存储器可以例如是可编程阵列逻辑（Programmable Array Logic，简称
25 “PAL”）、随机存取存储器（Random Access Memory，简称“RAM”）、可编程只读存储器（Programmable Read Only Memory，简

称“PROM”)、只读存储器(Read-Only Memory,简称“ROM”)、电可擦除可编程只读存储器(Electrically Erasable Programmable ROM,简称“EEPROM”)、磁盘、光盘、数字通用光盘(Digital Versatile Disc,简称“DVD”)等等。

5 本发明第三实施方式涉及一种基于计算机系统的推荐装置。图 15 是该基于计算机系统的推荐装置的结构示意图。如图 15 所示,该装置包括:

用户项目初始关系计算模块,用于获取各用户对各项目的项目评分记录。

聚类模块,用于根据用户项目初始关系计算模块获取的每个用户的项目评分记录进行聚类,将用户特征数据划分到 R 个类别中, R 是大于 1 的整数。以及

推荐模块,用于在聚类模块所划分的每个类别的用户特征数据中,基于项目为目标用户推荐项目。可以理解,在本发明的各个实施方式中,上述推荐模块可以采用基于协同过滤、基于关联规则或基于效用的推荐算法来为目标用户推荐项目。

此外,可以理解,在本发明的其他实施方式中,聚类模块也可以对项目进行聚类,推荐模块再在每个类别的用户特征数据中基于用户来为目标用户推荐项目,或是聚类和推荐都基于用户或都基于项目。

在本实施方式的推荐装置中,聚类模块先根据每个用户的项目评分记录进行聚类,将用户特征数据划分到多个类别中,推荐模块再在每个类别的用户特征数据中基于项目为目标用户推荐项目,可以在大数据下实现高效的推荐方法,保证了系统的稳定性和推荐的多样性。

优选地,上述计算机系统为分布式系统。该计算机系统包括至少两个计算节点。

25 上述推荐模块用于将各类别的用户特征数据分配给多个计算节点,每个计算节点最多保存 $R-1$ 个类别的用户特征数据,每个计算节点在所保存

的每个类别的用户特征数据中基于项目为目标用户推荐项目。每个计算节点不需要保存所有类别的用户特征数据，避免了内存不足的问题。

优选地，每个计算节点保存一个类别的用户特征数据并进行处理。此外，可以理解，在本发明的各实施方式中，可以根据各计算节点的配置将两个或两个以上类别的用户特征数据分配给高配置的计算节点进行处理。当然，在用户特征数据量不是很大的时候，也可以由一个计算节点进行处理。

作为可选实施方式，上述装置还包括用户规模判断模块，用于在聚类模块进行聚类前，判断用户数量是否大于用户规模阈值。

推荐模块用于若用户规模判断模块确认用户数量小于用户规模阈值，则直接在所有用户特征数据中基于项目为目标用户推荐项目。

聚类模块用于若用户规模判断模块确认用户数量大于用户规模阈值，则根据每个用户的项目评分记录进行聚类，将用户特征数据划分到 R 个类别中， R 是大于 1 的整数。

根据用户数量来选择是否要对用户进行聚类，以更好地适应于小数据下和大数据下的项目推荐。

此外，可以理解，在本发明的其他实施方式中，也可以不对用户数量进行判断，直接对用户进行聚类。

第一实施方式是与本实施方式相对应的方法实施方式，本实施方式可与第一实施方式互相配合实施。第一实施方式中提到的相关技术细节在本实施方式中依然有效，为了减少重复，这里不再赘述。相应地，本实施方式中提到的相关技术细节也可应用在第一实施方式中。

本发明第四实施方式涉及一种基于计算机系统的推荐装置。图 16 是该基于计算机系统的推荐装置中推荐模块的结构示意图。

第四实施方式在第三实施方式的基础上主要进行了以下两个改进：

第一个改进为对于每个类别中的每个项目或每个用户，只选取与其关系最强的几个项目，而不是保留与其有关系的所有项目，可以避免关系较弱

的项目产生的数据倾斜问题。具体地说：

上述推荐模块采用基于项目的协同过滤算法为目标用户推荐项目。如图 16 所示，该推荐模块包括：

项目相似度子模块，用于根据类别中每个用户的项目评分记录，计算类别中所有项目间相似度，并为每个项目选取相似度最高的 M 个项目， M 为预定义的整数。

用户推荐子模块，用于根据类别中目标用户的项目评分记录，为目标用户选取评分最高的 T 个项目， T 为预定义的整数。以及

初始推荐子模块，用于将用户推荐子模块为目标用户选取的 T 个项目与项目相似度子模块为 T 个项目中每个项目选取的 M 个项目相结合，并从中去除目标用户的项目列表中的项目，形成初始推荐结果。

优选地，上述推荐模块还包括：

初始推荐判断子模块，用于判断初始推荐子模块所形成的初始推荐结果中的项目数量是否大于 N ， N 为预定义的整数。

初始推荐筛选子模块，用于若初始推荐判断子模块确认初始推荐结果中的项目数量大于 N ，从初始推荐结果中选取相似度最高的 N 个项目推荐给目标用户。以及

用户数据规模还原子模块，用于若初始推荐判断子模块确认初始推荐结果中的项目数量小于 N ，则将目标用户的项目列表中的所有项目与为目标用户的项目列表中每个项目选取的 M 个项目相结合，并从中去除目标用户的项目列表中的所有项目，形成用户数据补全推荐结果。

更优地，上述推荐模块还包括：

补全推荐判断子模块，用于判断用户数据规模还原子模块所形成的用户数据补全推荐结果中的项目数量是否大于 N ， N 为预定义的整数。

补全推荐筛选子模块，用于若补全推荐判断子模块确认用户数据补全推荐结果中的项目数量大于 N ，从用户数据补全推荐结果中选取相似度最高

的 N 个项目推荐给目标用户。以及

项目数据规模还原子模块，用于若补全推荐判断子模块确认用户数据补全推荐结果中的项目数量小于 N，则将目标用户的项目列表中的所有项目与和目标用户的项目列表中每个项目具有相似度关系的所有项目相结合，并
5 从中去除目标用户的项目列表中的所有项目，形成项目数据补全推荐结果。

第二个改进为采用数据稀疏度对数据稀疏问题进行检测，并在发现数据稀疏问题后，通过项目间的二度关系进行相似度补全，以避免数据稀疏对推荐准确度的影响。具体地说：

上述装置还包括：

10 推荐结果数据稀疏度判断模块，用于判断数据稀疏度是否大于数据稀疏度阈值，数据稀疏度 $DSP = \frac{k}{P_l^2}$ ，其中 k 为计算得到的类别中具有相似度关系的项目对的数量，l 为类别中项目的数量， $P_l^2 = \frac{l(l-1)}{2}$ 。以及

数据稀疏补全模块，用于若推荐结果数据稀疏度判断模块确认数据稀疏度小于数据稀疏度阈值，则以第一项目、第二项目和第三项目为一组，第
15 一项目与第二项目间、第二项目与第三项目间具有相似度关系，通过第二项目为第一项目和第三项目建立相似度关系。

推荐模块用于根据数据稀疏补全模块所补充的项目间相似度关系在类别中再次基于项目为目标用户推荐项目，并且若推荐结果数据稀疏度判断模块确认数据稀疏度大于数据稀疏度阈值，则将基于项目计算得到的推荐项目
20 推荐给目标用户。

此外，可以理解，若通过项目间的二度关系对数据进行相似度补全后仍然存在数据稀疏问题，可以继续通过项目间的三度、四度或更高度关系对数据进行相似度补全，以避免数据稀疏对推荐准确度的影响。

以上各改进组合后形成本发明的较佳实施方式，但各改进也可以分别
25 使用。

第二实施方式是与本实施方式相对应的方法实施方式，本实施方式可与第二实施方式互相配合实施。第二实施方式中提到的相关技术细节在本实施方式中依然有效，为了减少重复，这里不再赘述。相应地，本实施方式中提到的相关技术细节也可应用在第二实施方式中。

5 综上，由于我们面临的应用场景是 user 数量和 item 数量都在亿级别，传统的算法已经无法满足我们的需求了，所以在上述基于计算机系统的推荐方法和装置中，使用聚类与重构 CF 算法两种方法可以解决这个问题。改进之后，在使用 600 个 reducer 情况下，可以在 90 分钟内实现亿级别数据量的推荐。并且通过定义数据稀疏的评价指标，当 item-item 相似度计算结束
10 后，如果结果小于评价指标的某一阈值，则计算 item 之间的更高度关系，做相似度补全，提高推荐准确性。

需要说明的是，本发明各设备实施方式中提到的各模块都是逻辑模块，在物理上，一个逻辑模块可以是一个物理模块，也可以是一个物理模块的一部分，还可以以多个物理模块的组合实现，这些逻辑模块本身的物理实现方式并不是最重要的，这些逻辑模块所实现的功能的组合才是解决本发明
15 所提出的技术问题的关键。此外，为了突出本发明的创新部分，本发明上述各设备实施方式并没有将与解决本发明所提出的技术问题关系不太密切的模块引入，这并不表明上述设备实施方式并不存在其它的模块。

需要说明的是，在本专利的权利要求和说明书中，诸如第一和第二等
20 之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过
25 程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个”限定的要素，并不排除在包括所述要素的过程、方法、物品或

者设备中还存在另外的相同要素。

虽然通过参照本发明的某些优选实施方式，已经对本发明进行了图示和描述，但本领域的普通技术人员应该明白，可以在形式上和细节上对其作各种改变，而不偏离本发明的精神和范围。

权 利 要 求 书

1. 一种基于计算机系统的推荐方法，其特征在于，该方法包括以下步骤：
骤：

5 获取各用户对各项目的项目评分记录；

根据每个用户的项目评分记录进行聚类，将用户特征数据划分到 R 个类别中， R 是大于 1 的整数；

在每个所述类别的用户特征数据中，基于项目为目标用户推荐项目。

2. 根据权利要求 1 所述的基于计算机系统的推荐方法，其特征在于，
10 所述计算机系统包括至少两个计算节点；

所述“在每个所述类别的用户特征数据中，基于项目为目标用户推荐项目”的步骤中，将各所述类别的用户特征数据分配给多个计算节点，每个计算节点最多保存 $R-1$ 个所述类别的用户特征数据，每个计算节点在所保存的每个所述类别的用户特征数据中基于项目为目标用户推荐项目。

15 3. 根据权利要求 1 所述的基于计算机系统的推荐方法，其特征在于，
所述“在每个所述类别的用户特征数据中，基于项目为目标用户推荐项目”的步骤中，采用基于项目的协同过滤算法为目标用户推荐项目；

所述“在每个所述类别的用户特征数据中，基于项目为目标用户推荐项目”的步骤中包括以下子步骤：

20 根据所述类别中每个用户的项目评分记录，计算所述类别中所有项目间相似度，并为每个项目选取相似度最高的 M 个项目， M 为预定义的整数；

根据所述类别中所述目标用户的项目评分记录，为所述目标用户选取评分最高的 T 个项目， T 为预定义的整数；

将为所述目标用户选取的 T 个项目与为所述 T 个项目中每个项目选取的

M个项目相结合，并从中去除所述目标用户的项目列表中的项目，形成初始推荐结果。

4. 根据权利要求 3 所述的基于计算机系统的推荐方法，其特征在于，在形成初始推荐结果的子步骤后还包括以下子步骤：

5 判断所述初始推荐结果中的项目数量是否大于 N，N 为预定义的整数；

若所述初始推荐结果中的项目数量大于 N，则从所述初始推荐结果中选取相似度最高的 N 个项目推荐给所述目标用户；

若所述初始推荐结果中的项目数量小于 N，则将所述目标用户的项目列表中的所有项目与为所述目标用户的项目列表中每个项目选取的 M 个项目相结合，并从中去除所述目标用户的项目列表中的所有项目，形成用户数据
10 补全推荐结果。

5. 根据权利要求 4 所述的基于计算机系统的推荐方法，其特征在于，在形成用户数据补全推荐结果的子步骤后还包括以下子步骤：

判断所述用户数据补全推荐结果中的项目数量是否大于 N，N 为预定义
15 的整数；

若所述用户数据补全推荐结果中的项目数量大于 N，则从所述用户数据补全推荐结果中选取相似度最高的 N 个项目推荐给所述目标用户；

若所述用户数据补全推荐结果中的项目数量小于 N，则将所述目标用户的项目列表中的所有项目与和所述目标用户的项目列表中每个项目具有相似
20 度关系的所有项目相结合，并从中去除所述目标用户的项目列表中的所有项目，形成项目数据补全推荐结果。

6. 根据权利要求 1 所述的基于计算机系统的推荐方法，其特征在于，所述“在每个所述类别的用户特征数据中，基于项目为目标用户推荐项目”的步骤后还包括以下步骤：

判断数据稀疏度是否大于数据稀疏度阈值，所述数据稀疏度 $DSP = \frac{k}{P_l^2}$ ，其中 k 为计算得到的所述类别中具有相似度关系的项目对的数量， l 为所述类别中项目的数量， $P_l^2 = \frac{l(l-1)}{2}$ ；

5 若所述数据稀疏度小于数据稀疏度阈值，则以第一项目、第二项目和第三项目为一组，所述第一项目与所述第二项目间、所述第二项目与所述第三项目间具有相似度关系，通过所述第二项目为所述第一项目和所述第三项目建立相似度关系，并根据补充的项目间相似度关系在所述类别中再次基于项目为所述目标用户推荐项目；

10 若所述数据稀疏度大于数据稀疏度阈值，则将基于项目计算得到的推荐项目推荐给所述目标用户。

7. 根据权利要求 1 至 6 中任一项所述的基于计算机系统的推荐方法，其特征在于，所述“根据每个用户的项目评分记录进行聚类，将用户特征数据划分到 R 个类别中， R 是大于 1 的整数”的步骤前还包括以下步骤：

判断用户数量是否大于用户规模阈值；

15 若所述用户数量小于用户规模阈值，则直接在所有用户特征数据中基于项目为目标用户推荐项目；

若所述用户数量大于用户规模阈值，则进入“根据每个用户的项目评分记录进行聚类，将用户特征数据划分到 R 个类别中， R 是大于 1 的整数”的步骤。

20 8. 一种基于计算机系统的推荐装置，其特征在于，所述装置包括：

用户项目初始关系计算模块，用于获取各用户对各项目的项目评分记录；

聚类模块，用于根据所述用户项目初始关系计算模块获取的每个用户

的项目评分记录进行聚类，将用户特征数据划分到 R 个类别中， R 是大于 1 的整数；以及

推荐模块，用于在所述聚类模块所划分的每个所述类别的用户特征数据中，基于项目为目标用户推荐项目。

- 5 9. 根据权利要求 8 所述的基于计算机系统的推荐装置，其特征在于，所述计算机系统包括至少两个计算节点；

所述推荐模块用于将各所述类别的用户特征数据分配给多个计算节点，每个计算节点最多保存 $R-1$ 个所述类别的用户特征数据，每个计算节点在所保存的每个所述类别的用户特征数据中基于项目为目标用户推荐项目。

10 目。

10. 根据权利要求 8 所述的基于计算机系统的推荐装置，其特征在于，所述推荐模块采用基于项目的协同过滤算法为目标用户推荐项目；

所述推荐模块包括：

项目相似度子模块，用于根据所述类别中每个用户的项目评分记录，计算所述类别中所有项目间相似度，并为每个项目选取相似度最高的 M 个项目， M 为预定义的整数；

15 项目， M 为预定义的整数；

用户推荐子模块，用于根据所述类别中所述目标用户的项目评分记录，为所述目标用户选取评分最高的 T 个项目， T 为预定义的整数；以及

初始推荐子模块，用于将所述用户推荐子模块为所述目标用户选取的 T 个项目与所述项目相似度子模块为所述 T 个项目中每个项目选取的 M 个项目相结合，并从中去除所述目标用户的项目列表中的项目，形成初始推荐结果。

20 个项目与所述项目相似度子模块为所述 T 个项目中每个项目选取的 M 个项目相结合，并从中去除所述目标用户的项目列表中的项目，形成初始推荐结果。

11. 根据权利要求 10 所述的基于计算机系统的推荐装置，其特征在于，所述推荐模块还包括：

初始推荐判断子模块，用于判断所述初始推荐子模块所形成的初始推荐结果中的项目数量是否大于 N ， N 为预定义的整数；

初始推荐筛选子模块，用于若所述初始推荐判断子模块确认所述初始推荐结果中的项目数量大于 N ，从所述初始推荐结果中选取相似度最高的 N 个项目推荐给所述目标用户；以及

用户数据规模还原子模块，用于若所述初始推荐判断子模块确认所述初始推荐结果中的项目数量小于 N ，则将所述目标用户的项目列表中的所有项目与为所述目标用户的项目列表中每个项目选取的 M 个项目相结合，并从中去除所述目标用户的项目列表中的所有项目，形成用户数据补全推荐结果。

12. 根据权利要求 11 所述的基于计算机系统的推荐装置，其特征在于，所述推荐模块还包括：

补全推荐判断子模块，用于判断所述用户数据规模还原子模块所形成的用户数据补全推荐结果中的项目数量是否大于 N ， N 为预定义的整数；

15 补全推荐筛选子模块，用于若所述补全推荐判断子模块确认所述用户数据补全推荐结果中的项目数量大于 N ，从所述用户数据补全推荐结果中选取相似度最高的 N 个项目推荐给所述目标用户；以及

项目数据规模还原子模块，用于若所述补全推荐判断子模块确认所述用户数据补全推荐结果中的项目数量小于 N ，则将所述目标用户的项目列表中的所有项目与和所述目标用户的项目列表中每个项目具有相似度关系的所有项目相结合，并从中去除所述目标用户的项目列表中的所有项目，形成项目数据补全推荐结果。

13. 根据权利要求 8 所述的基于计算机系统的推荐装置，其特征在于，所述装置还包括：

推荐结果数据稀疏度判断模块，用于判断数据稀疏度是否大于数据稀疏度阈值，所述数据稀疏度 $DSP = \frac{k}{P_l^2}$ ，其中 k 为计算得到的所述类别中具有相似关系的项目对的数量，l 为所述类别中项目的数量， $P_l^2 = \frac{l(l-1)}{2}$ ；以及

5 数据稀疏补全模块，用于若所述推荐结果数据稀疏度判断模块确认所述数据稀疏度小于数据稀疏度阈值，则以第一项目、第二项目和第三项目为一组，所述第一项目与所述第二项目间、所述第二项目与所述第三项目间具有相似关系，通过所述第二项目为所述第一项目和所述第三项目建立相似关系；

10 所述推荐模块用于根据所述数据稀疏补全模块所补充的项目间相似关系在所述类别中再次基于项目为所述目标用户推荐项目，并且若所述推荐结果数据稀疏度判断模块确认所述数据稀疏度大于数据稀疏度阈值，则将基于项目计算得到的推荐项目推荐给所述目标用户。

14. 根据权利要求 8 至 13 中任一项所述的基于计算机系统的推荐装置，其特征在于，所述装置还包括用户规模判断模块，用于在所述聚类模块
15 聚类前，判断用户数量是否大于用户规模阈值；

所述推荐模块用于若所述用户规模判断模块确认所述用户数量小于用户规模阈值，则直接在所有用户特征数据中基于项目为目标用户推荐项目；

所述聚类模块用于若所述用户规模判断模块确认所述用户数量大于用户规模阈值，则根据每个用户的项目评分记录进行聚类，将用户特征数据划
20 分到 R 个类别中，R 是大于 1 的整数。

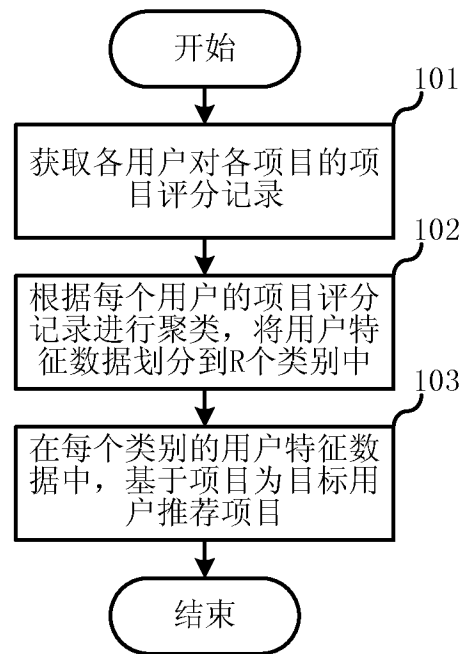


图 1

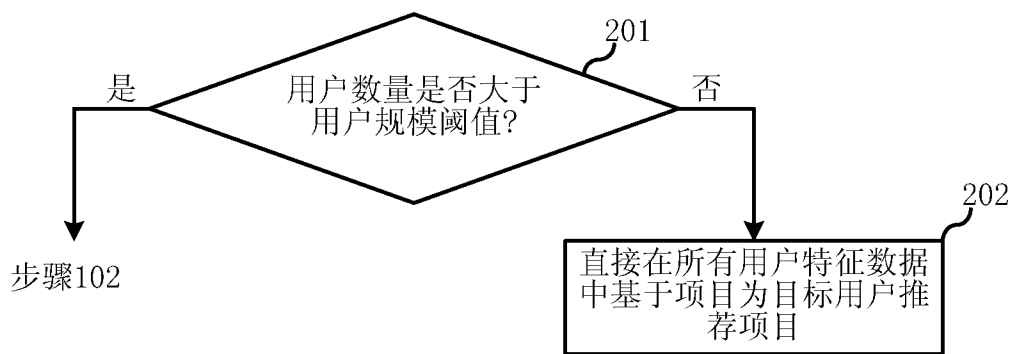


图 2

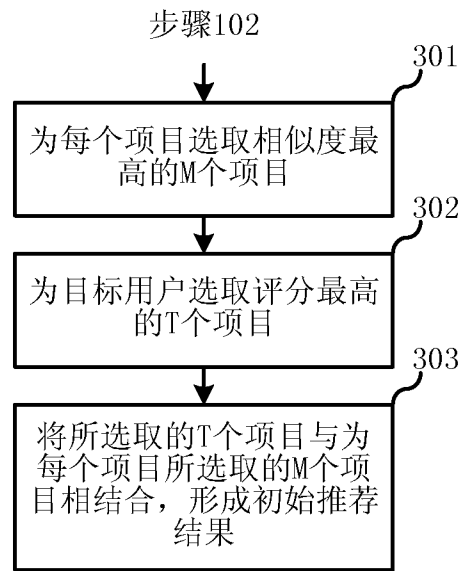


图 3

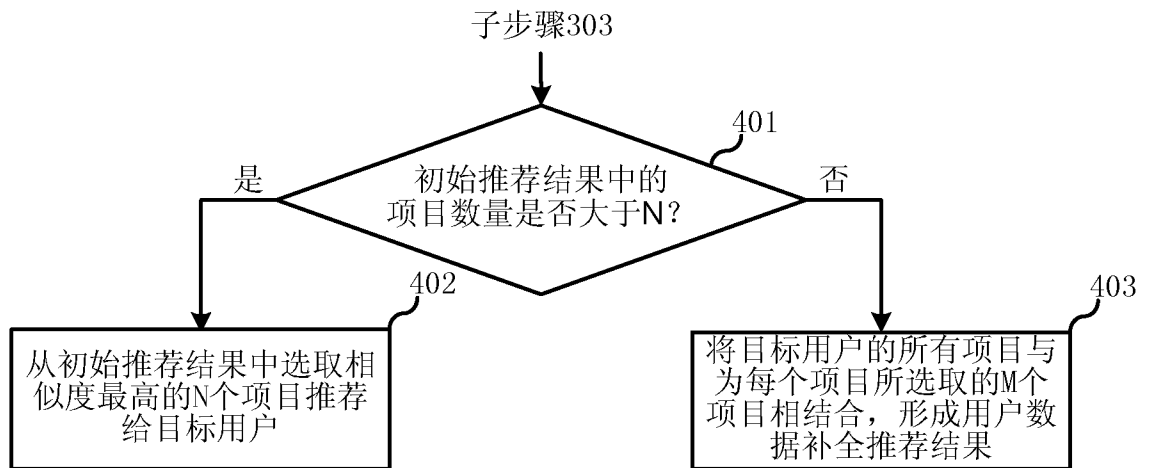


图 4

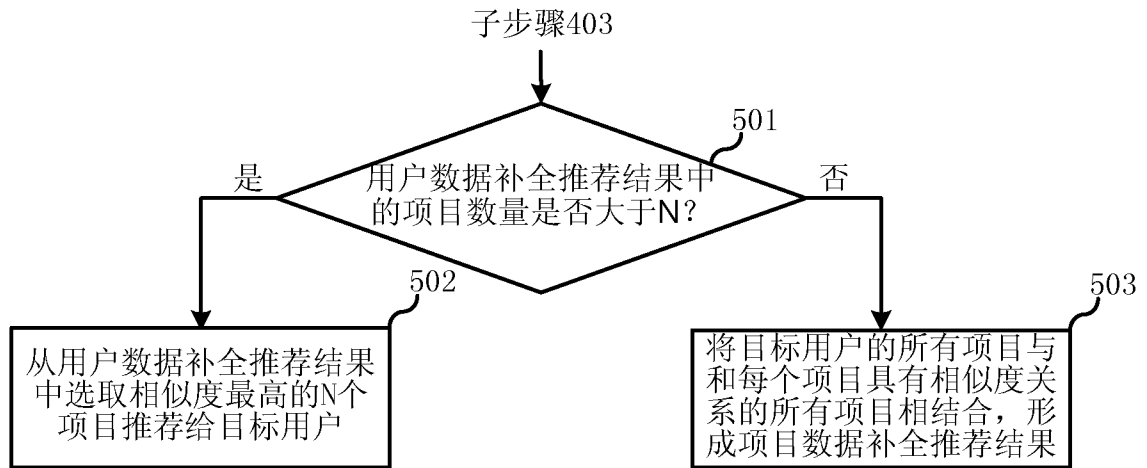


图 5

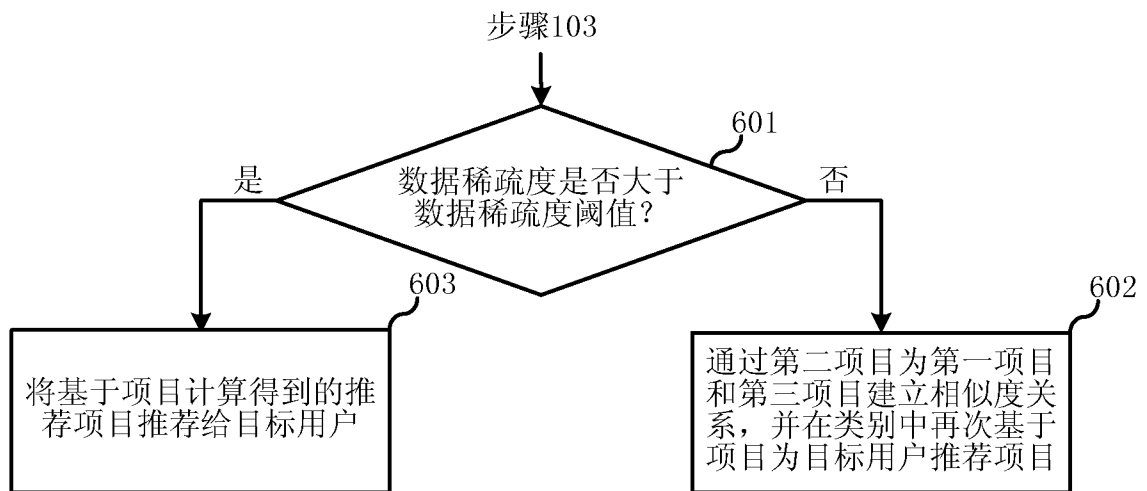


图 6

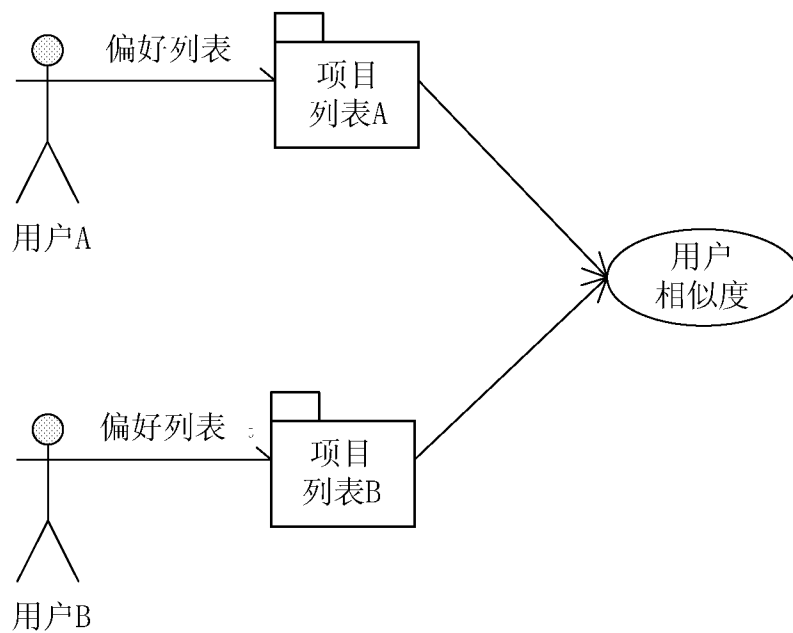


图 7

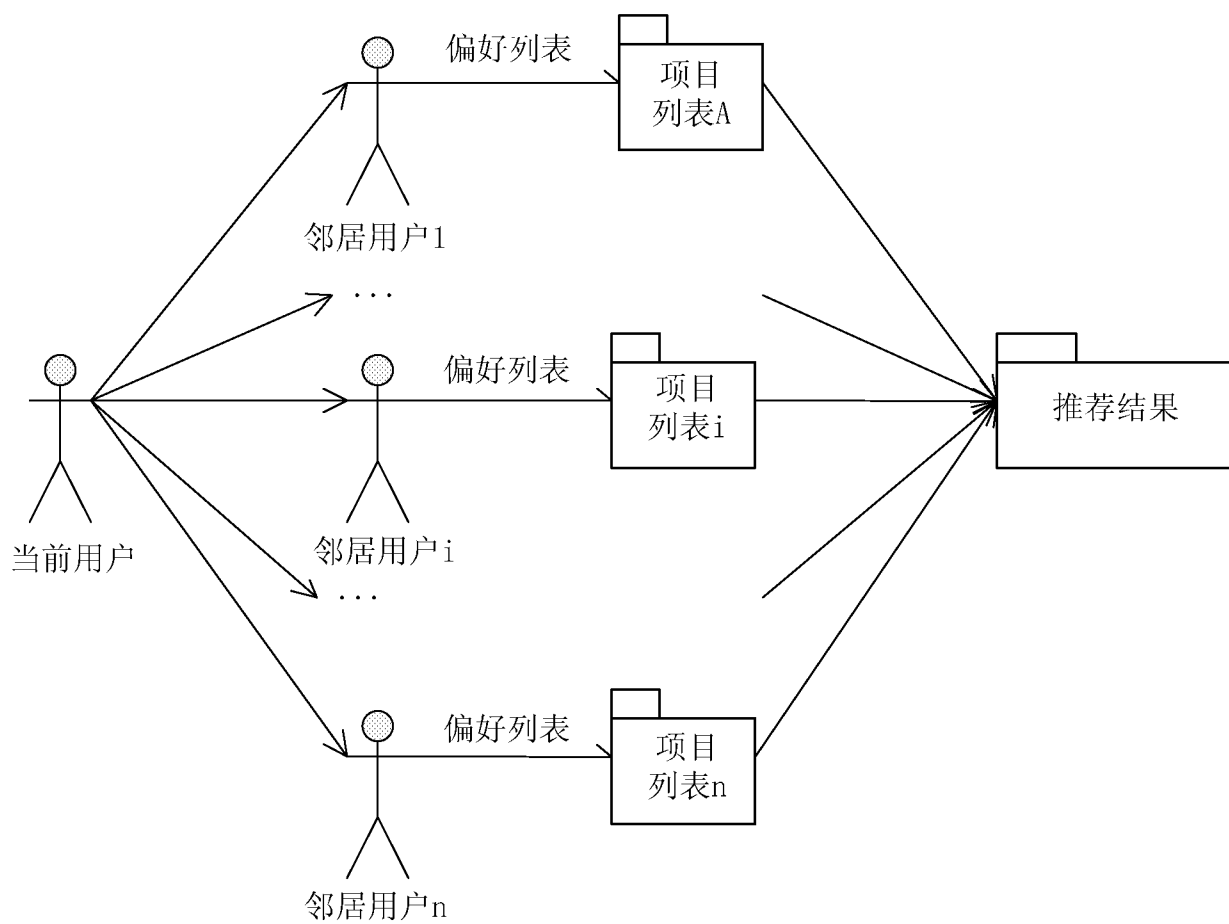


图 8

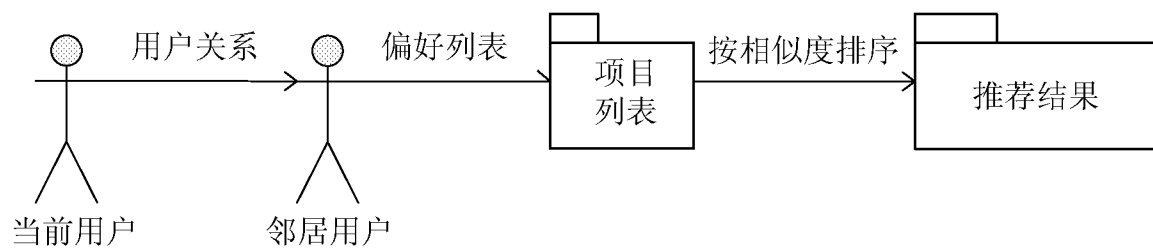


图 9

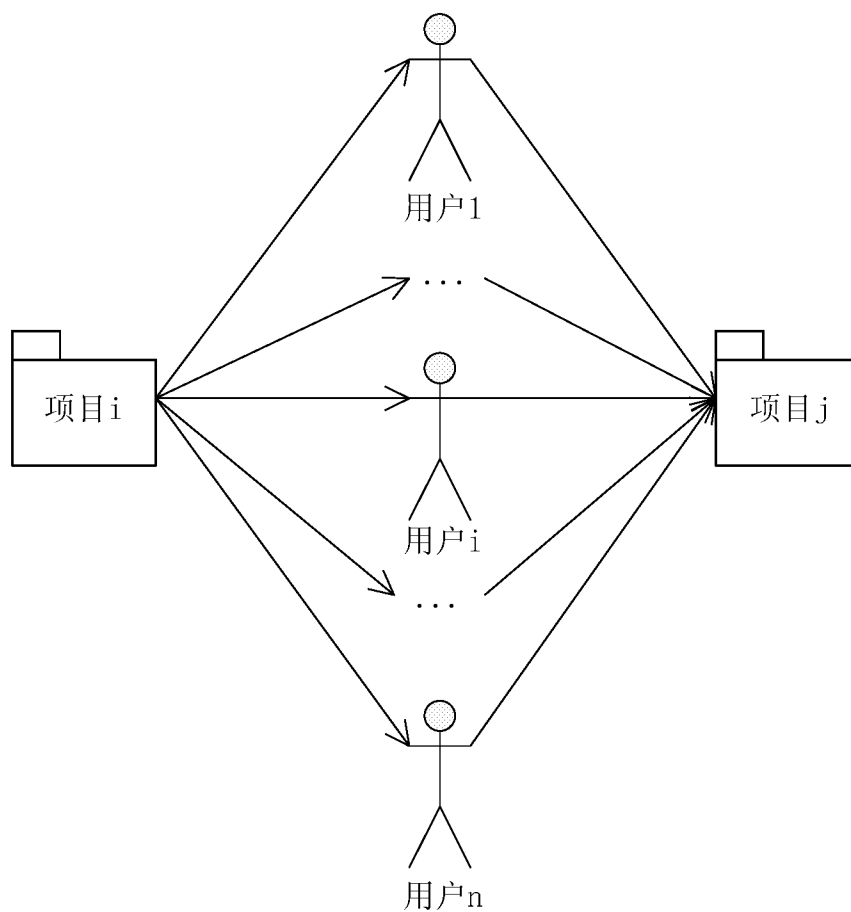


图 10

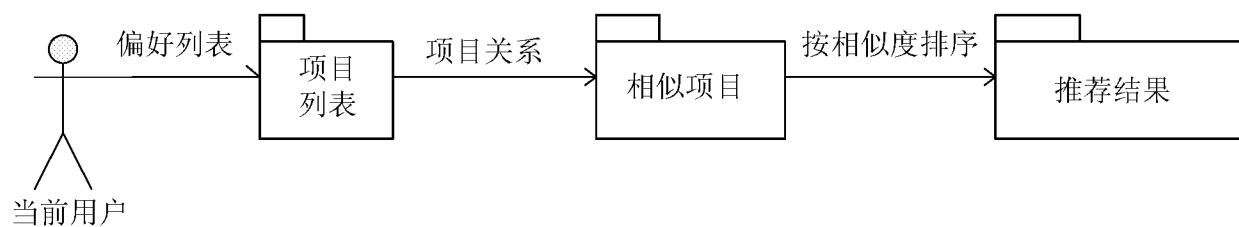


图 11

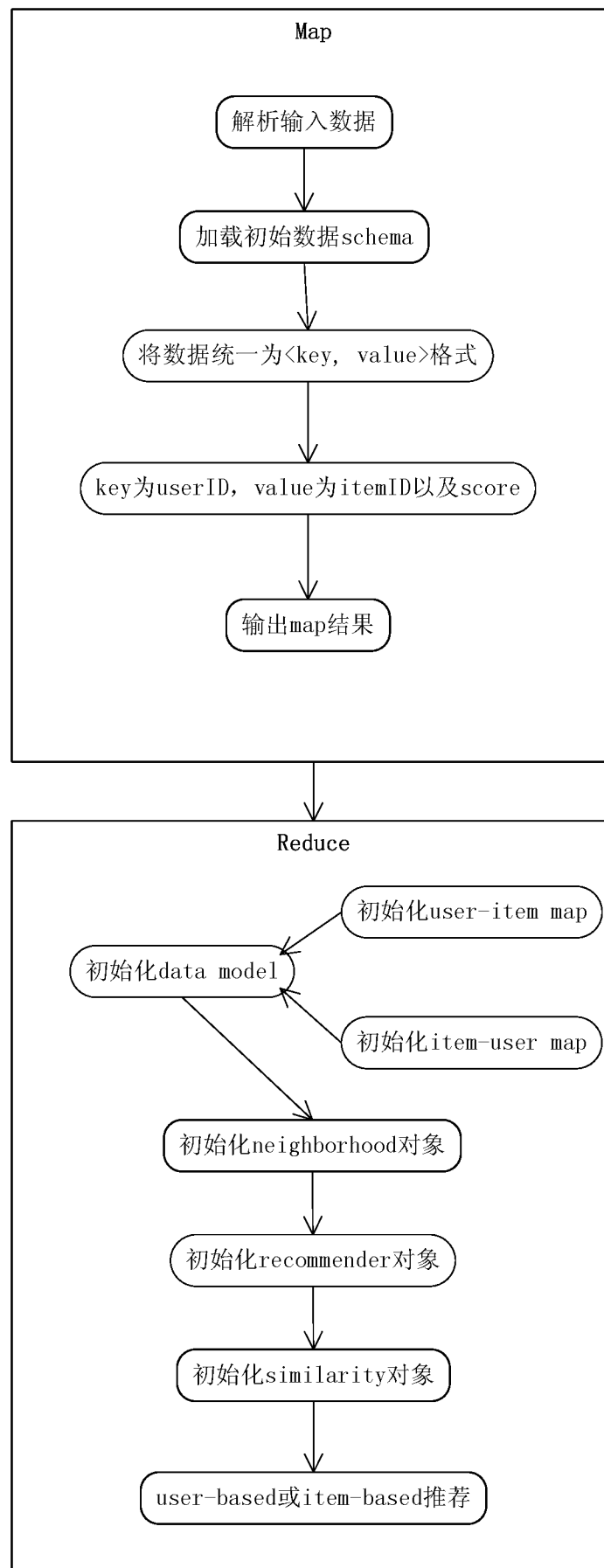


图 12

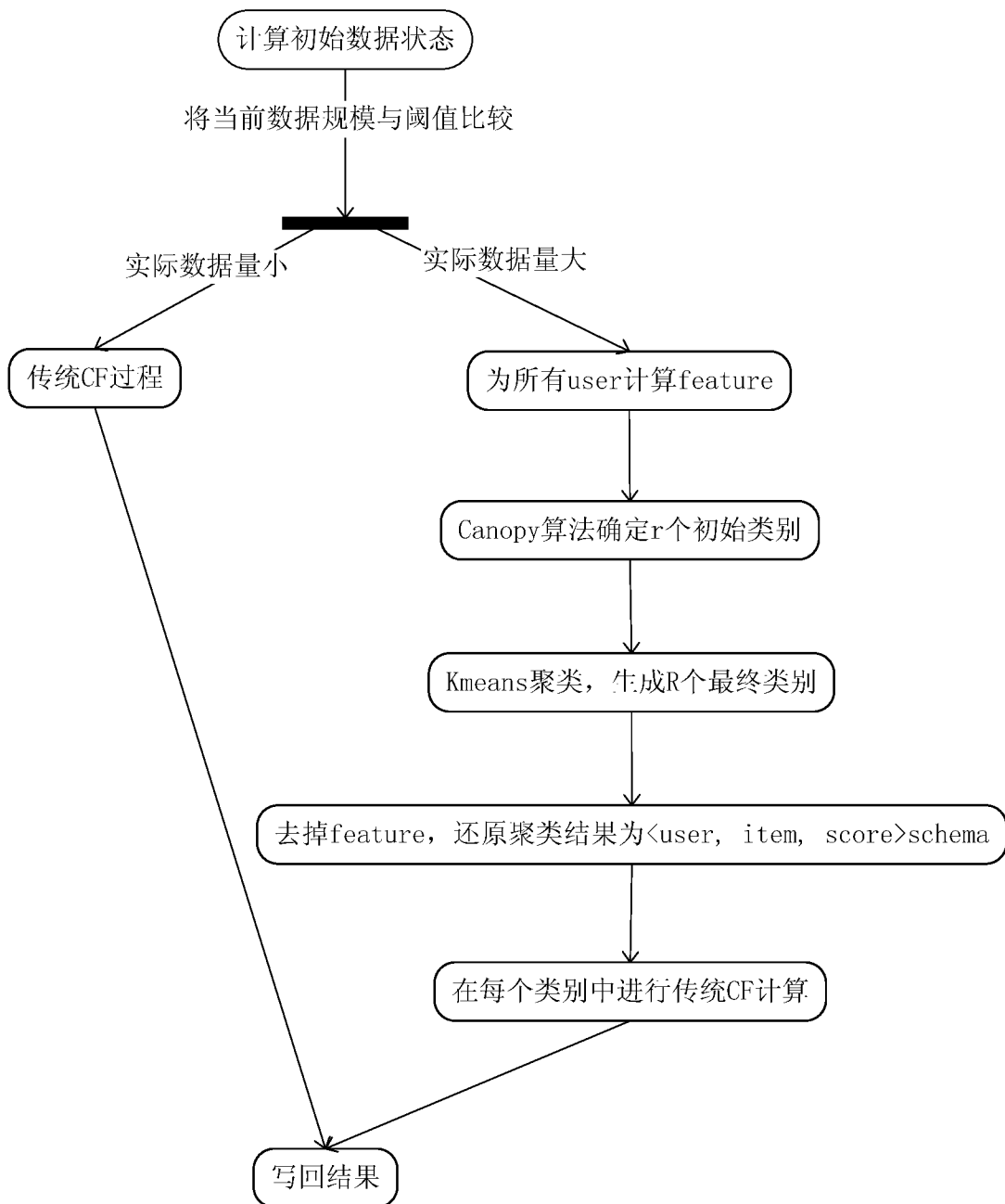


图 13

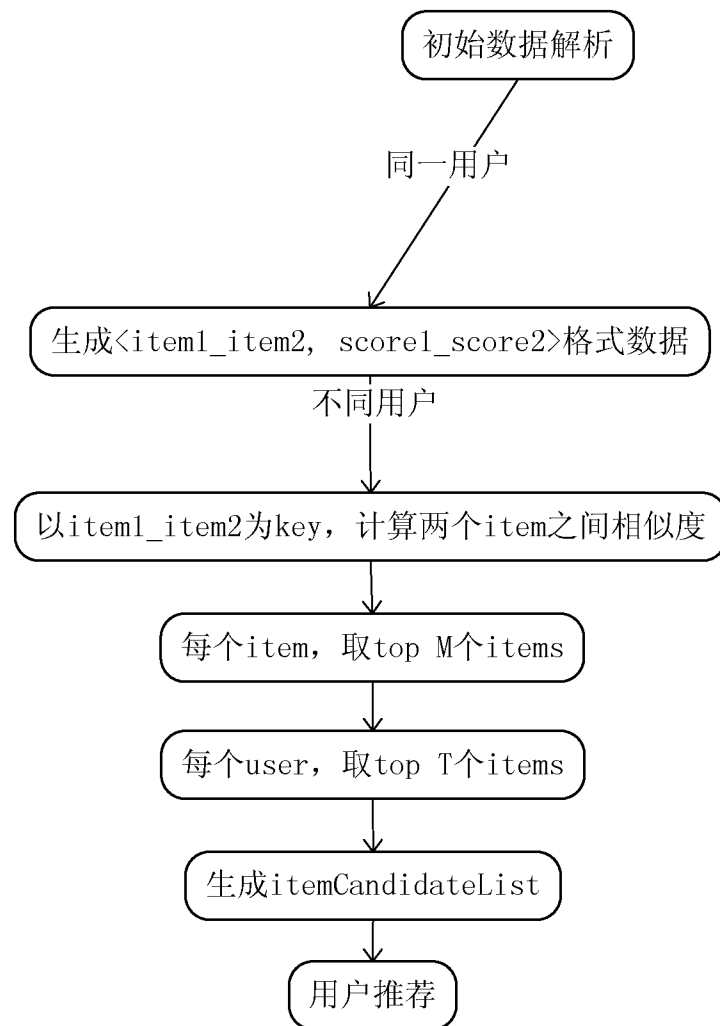


图 14

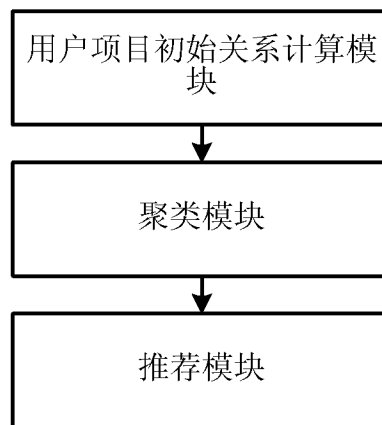


图 15

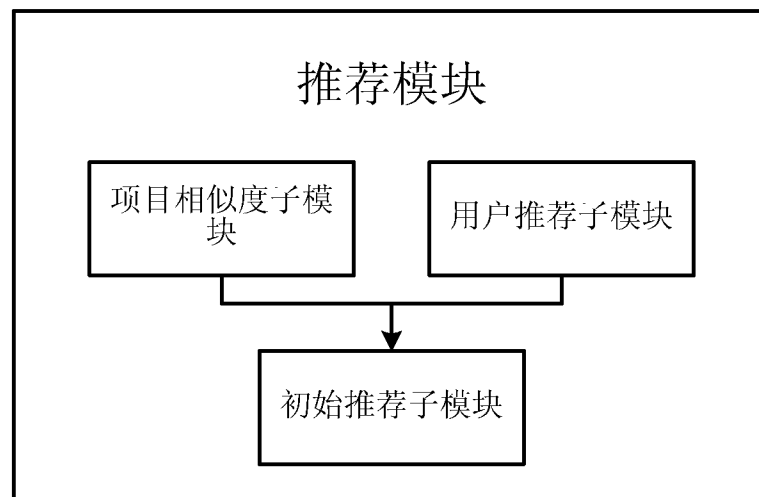


图 16

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2015/095834

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F; G06Q

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

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

CPRSABS; CNTXT; VEN; CNKI; IEEE: recommend+, collaborative filter+, cluster+, group?, class+, category, user?, spars+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 101685458 A (HUAWEI TECHNOLOGIES CO., LTD.), 31 March 2010 (31.03.2010), description, pages 2-14	1-5, 7-12, 14
Y	CN 101685458 A (HUAWEI TECHNOLOGIES CO., LTD.), 31 March 2010 (31.03.2010), description, pages 2-14	6, 13
Y	CN 102640141 A (TOYOTA MOTOR CORPORATION), 15 August 2012 (15.08.2012), description, paragraphs 51-83	6, 13
A	CN 103678672 A (BEIJING ZHONGXINGTONG SOFTWARE TECHNOLOGY CO., LTD.), 26 March 2014 (26.03.2014), the whole document	1-14
A	CN 103412948 A (BEIJING JIAOTONG UNIVERSITY), 27 November 2013 (27.11.2013), the whole document	1-14
A	CN 103049488 A (BEIJING QIHOO TECHNOLOGY CO., LTD. et al.), 17 April 2013 (17.04.2013), the whole document	1-14
A	CN 103389966 A (ALIBABA GROUP HOLDING LIMITED), 13 November 2013 (13.11.2013), the whole document	1-14

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

☒ See patent family annex.

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

Date of the actual completion of the international search 28 January 2016 (28.01.2016)	Date of mailing of the international search report 04 February 2016 (04.02.2016)
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer XU, Lin Telephone No.: (86-10) 62089885

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2015/095834

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 101685458 A	31 March 2010	CN 101685458 B	19 September 2012
		US 2011184977 A1	28 July 2011
		WO 2010037286 A1	08 April 2010
CN 102640141 A	15 August 2012	EP 2515244 A1	24 October 2012
		WO 2011074113 A1	23 June 2011
		US 2012239604 A1	20 September 2012
		US 9087123 B2	21 July 2015
		CN 102640141 B	13 May 2015
		JP 5403069 B2	29 January 2014
CN 103678672 A	26 March 2014	None	
CN 103412948 A	27 November 2013	None	
CN 103049488 A	17 April 2013	CN 103049488 B	25 November 2015
CN 103389966 A	13 November 2013	None	

A. 主题的分类 G06F 17/30 (2006.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类																										
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F; G06Q 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) CPRSABS; CNTXT; VEN; CNKI; IEEE: 推荐, 聚类, 簇, 群, 类, 组, 用户, 协同过滤, 稀疏, recommend+, collaborative filter+, cluster+, group?, class+, category, user?, spars+																										
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>CN 101685458 A (华为技术有限公司) 2010年 3月 31日 (2010 - 03 - 31) 说明书第2-14页</td> <td>1-5, 7-12, 14</td> </tr> <tr> <td>Y</td> <td>CN 101685458 A (华为技术有限公司) 2010年 3月 31日 (2010 - 03 - 31) 说明书第2-14页</td> <td>6, 13</td> </tr> <tr> <td>Y</td> <td>CN 102640141 A (丰田自动车株式会社) 2012年 8月 15日 (2012 - 08 - 15) 说明书第51-83段</td> <td>6, 13</td> </tr> <tr> <td>A</td> <td>CN 103678672 A (北京中兴通软件科技股份有限公司) 2014年 3月 26日 (2014 - 03 - 26) 全文</td> <td>1-14</td> </tr> <tr> <td>A</td> <td>CN 103412948 A (北京交通大学) 2013年 11月 27日 (2013 - 11 - 27) 全文</td> <td>1-14</td> </tr> <tr> <td>A</td> <td>CN 103049488 A (北京奇虎科技有限公司 等) 2013年 4月 17日 (2013 - 04 - 17) 全文</td> <td>1-14</td> </tr> <tr> <td>A</td> <td>CN 103389966 A (阿里巴巴集团控股有限公司) 2013年 11月 13日 (2013 - 11 - 13) 全文</td> <td>1-14</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	CN 101685458 A (华为技术有限公司) 2010年 3月 31日 (2010 - 03 - 31) 说明书第2-14页	1-5, 7-12, 14	Y	CN 101685458 A (华为技术有限公司) 2010年 3月 31日 (2010 - 03 - 31) 说明书第2-14页	6, 13	Y	CN 102640141 A (丰田自动车株式会社) 2012年 8月 15日 (2012 - 08 - 15) 说明书第51-83段	6, 13	A	CN 103678672 A (北京中兴通软件科技股份有限公司) 2014年 3月 26日 (2014 - 03 - 26) 全文	1-14	A	CN 103412948 A (北京交通大学) 2013年 11月 27日 (2013 - 11 - 27) 全文	1-14	A	CN 103049488 A (北京奇虎科技有限公司 等) 2013年 4月 17日 (2013 - 04 - 17) 全文	1-14	A	CN 103389966 A (阿里巴巴集团控股有限公司) 2013年 11月 13日 (2013 - 11 - 13) 全文	1-14
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																								
X	CN 101685458 A (华为技术有限公司) 2010年 3月 31日 (2010 - 03 - 31) 说明书第2-14页	1-5, 7-12, 14																								
Y	CN 101685458 A (华为技术有限公司) 2010年 3月 31日 (2010 - 03 - 31) 说明书第2-14页	6, 13																								
Y	CN 102640141 A (丰田自动车株式会社) 2012年 8月 15日 (2012 - 08 - 15) 说明书第51-83段	6, 13																								
A	CN 103678672 A (北京中兴通软件科技股份有限公司) 2014年 3月 26日 (2014 - 03 - 26) 全文	1-14																								
A	CN 103412948 A (北京交通大学) 2013年 11月 27日 (2013 - 11 - 27) 全文	1-14																								
A	CN 103049488 A (北京奇虎科技有限公司 等) 2013年 4月 17日 (2013 - 04 - 17) 全文	1-14																								
A	CN 103389966 A (阿里巴巴集团控股有限公司) 2013年 11月 13日 (2013 - 11 - 13) 全文	1-14																								
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																										
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																										
国际检索实际完成的日期 2016年 1月 28日		国际检索报告邮寄日期 2016年 2月 4日																								
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10) 62019451		授权官员 徐琳 电话号码 (86-10) 62089885																								

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2015/095834

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	101685458	A	2010年 3月 31日	CN	101685458	B	2012年 9月 19日
				US	2011184977	A1	2011年 7月 28日
				WO	2010037286	A1	2010年 4月 8日
CN	102640141	A	2012年 8月 15日	EP	2515244	A1	2012年 10月 24日
				WO	2011074113	A1	2011年 6月 23日
				US	2012239604	A1	2012年 9月 20日
				US	9087123	B2	2015年 7月 21日
				CN	102640141	B	2015年 5月 13日
				JP	5403069	B2	2014年 1月 29日
CN	103678672	A	2014年 3月 26日	无			
CN	103412948	A	2013年 11月 27日	无			
CN	103049488	A	2013年 4月 17日	CN	103049488	B	2015年 11月 25日
CN	103389966	A	2013年 11月 13日	无			