

(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
2020 年 11 月 19 日 (19.11.2020)



(10) 国际公布号
WO 2020/228514 A1

- (51) 国际专利分类号:
G06F 16/9535 (2019.01)
- (21) 国际申请号: PCT/CN2020/086946
- (22) 国际申请日: 2020 年 4 月 26 日 (26.04.2020)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201910395560.4 2019年5月13日 (13.05.2019) CN
- (71) 申请人: 腾讯科技(深圳)有限公司 (TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED) [CN/CN]; 中国广东省深圳市南山区高新区科技中一路腾讯大厦 35 层, Guangdong 518057 (CN)。
- (72) 发明人: 刘雨丹 (LIU, Yudan); 中国广东省深圳市南山区高新区科技中一路腾讯大厦 35 层,

Guangdong 518057 (CN)。葛凯凯 (GE, Kaikai); 中国广东省深圳市南山区高新区科技中一路腾讯大厦 35 层, Guangdong 518057 (CN)。张旭 (ZHANG, Xu); 中国广东省深圳市南山区高新区科技中一路腾讯大厦 35 层, Guangdong 518057 (CN)。林乐宇 (LIN, Leyu); 中国广东省深圳市南山区高新区科技中一路腾讯大厦 35 层, Guangdong 518057 (CN)。陈鑫 (CHEN, Xin); 中国广东省深圳市南山区高新区科技中一路腾讯大厦 35 层, Guangdong 518057 (CN)。郝晓波 (HAO, Xiaobo); 中国广东省深圳市南山区高新区科技中一路腾讯大厦 35 层, Guangdong 518057 (CN)。王伟 (WANG, Wei); 中国广东省深圳市南山区高新区科技中一路腾讯大厦 35 层, Guangdong 518057 (CN)。庄凯 (ZHUANG, Kai); 中国广东省深圳市南山区高新区科技中一路腾讯大厦 35 层, Guangdong 518057 (CN)。闫肃 (YAN, Su); 中国广东省深圳市南山区高新区科技中一

(54) Title: CONTENT RECOMMENDATION METHOD AND APPARATUS, AND DEVICE AND STORAGE MEDIUM

(54) 发明名称: 内容推荐方法、装置、设备及存储介质

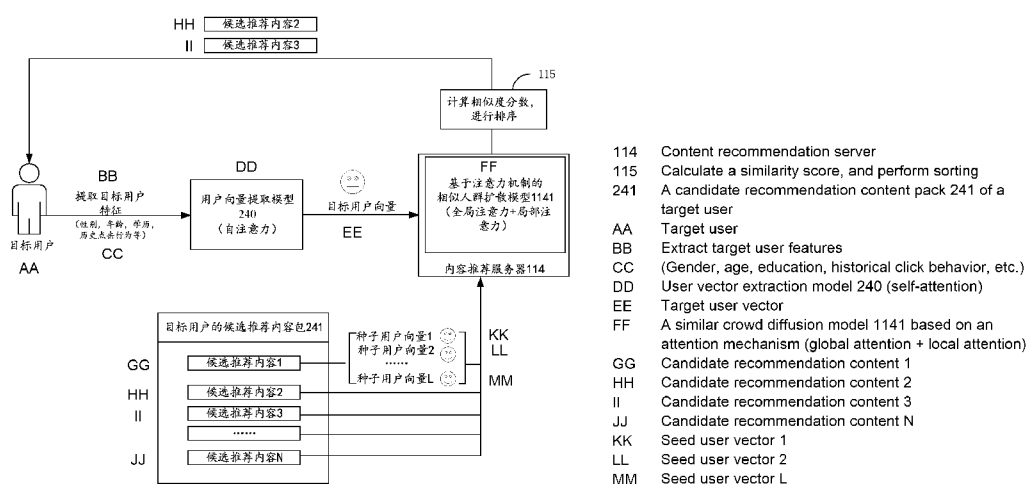


图 2

(57) Abstract: The present application relates to the field of artificial intelligence, and disclosed are a content recommendation method and apparatus, and a device and a storage medium. The method comprises: obtaining a target user vector of a target user; determining n groups of seed user vectors according to the target user vector, the n groups of seed user vectors corresponding to n candidate recommendation contents; calling a similar crowd diffusion model to calculate a similarity between the target user vector and each group of seed user vectors, the similar crowd diffusion model being used for calculating a similarity between user vectors on the basis of an attention mechanism; and determining a target content recommended to the target user among the n candidate recommendation contents according to the respective similarities of the n groups of seed user vectors. According to the present application, the problem of low accuracy of a recommendation method in the related art can be solved.

路腾讯大厦35层, Guangdong 518057 (CN)。 潘治达(PAN, Zhida); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。 唐琳瑶(TANG, Linyao); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。 张晶(ZHANG, Jing); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。

(74) 代理人: 深圳市深佳知识产权代理事务所(普通合伙)(SHENPAT INTELLECTUAL PROPERTY AGENCY); 中国广东省深圳市罗湖区南湖街道春风路庐山大厦B座18C2、18D、18E、18E2, Guangdong 518001 (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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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, 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))。

(57) 摘要: 本申请公开了内容推荐方法、装置、设备及存储介质, 属于人工智能领域。该方法包括: 获取目标用户的目标用户向量; 根据目标用户向量确定n组种子用户向量, n组种子用户向量与n条候选推荐内容对应; 调用相似人群扩散模型计算所述目标用户向量和每组所述种子用户向量的相似度, 所述相似人群扩散模型用于基于注意力机制计算用户向量之间的相似度; 根据n组种子用户向量各自的相似度, 确定n条候选推荐内容中向所述目标用户推荐的目标内容。本申请能够解决相关技术的推荐方法准确率较低的问题。

内容推荐方法、装置、设备及存储介质

本申请要求于 2019 年 05 月 13 日提交中国专利局、申请号为 2019103955604、申请名称为“内容推荐方法、训练方法、装置、设备及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5 技术领域

本申请涉及人工智能（Artificial Intelligence, AI）领域，特别涉及信息推荐技术。

背景技术

10 在内容推荐系统中，传统的推荐算法基于用户的历史行为和内容的语义特征进行内容推荐。

以基于点击率预估的推荐方法为例，服务器中设置有深度网络模型。针对候选内容集合中每一对“用户-内容”组合，该深度网络模型根据其中用户的历史点击行为、内容的语义特征和上下文特征，预测该用户对该内容的点击概率；然后，针对某用户的待推荐内容，按照点击概率由高到低的顺序，将
15 排序在前 n 位的内容作为信息流推荐给该用户。

在相关技术中，推荐算法往往只是根据目标用户的兴趣选择推送信息，通过分析信息内容判断用户对该信息的感兴趣程度，但这种推荐方法忽略了用户对于获知当前热点事件和阅读小众高质量内容等需求，往往存在准确率较低的问题。

20 发明内容

本申请实施例提供了内容推荐方法、装置、设备及存储介质，能够提高内容推荐的准确程度。所述技术方案如下：

根据本申请的一个方面，提供了一种内容推荐方法，由服务器执行，所述方法包括：

25 获取目标用户的目标用户向量；

根据所述目标用户向量确定 n 组种子用户向量，所述 n 组种子用户向量与 n 条候选推荐内容对应；所述 n 为正整数；

调用相似人群扩散模型计算所述目标用户向量和每组所述种子用户向量的相似度，所述相似人群扩散模型用于基于注意力机制计算用户向量之间的
30 相似度；

根据所述 n 组种子用户向量各自的所述相似度，确定所述 n 条候选推荐内容中向所述目标用户推荐的目标内容。

根据本申请的另一方面，提供了一种用户向量提取模型的训练方法，由服务器执行，所述用户向量提取模型包括嵌入层、合并层和全连接层，所述
35 合并层是基于自注意力机制的神经网络层，所述方法包括：

-2-

获取第一训练样本，所述第一训练样本包括多个样本用户的用户特征和对样本内容的互动记录；

将第一样本用户的用户特征输入所述用户向量提取模型，得到第一用户向量；所述第一样本用户是所述第一训练样本内多个样本用户中的任意一个；

5 计算所述第一用户向量与所述样本内容的嵌入向量的点积；

根据所述点积和所述第一样本用户对所述样本内容的互动记录，计算得到第一误差损失；

根据所述第一误差损失采用反向传播算法对所述用户向量提取模型进行训练。

10 根据本申请的另一方，提供了一种相似人群扩散模型的训练方法，由服务器执行，所述相似人群扩散模型包括第一全连接层、第二全连接层和注意力单元，所述方法包括：

获取第二训练样本，所述第二训练样本包括：样本用户向量和样本内容对应的种子用户向量；

15 调用所述第一全连接层对所述种子用户向量进行特征提取，得到种子相似扩散向量；以及，调用所述第二全连接层对所述样本用户向量进行特征提取，得到样本相似扩散向量；

调用所述注意力单元根据所述种子相似扩散向量和所述样本相似扩散向量计算得到所述种子加权向量；

20 计算所述种子加权向量和所述样本相似扩散向量之间的向量距离；

根据所述向量距离计算第二误差损失，根据所述第二误差损失采用反向传播算法对所述相似人群扩散模型进行训练。

根据本申请的另一方，提供了一种内容推荐装置，所述装置包括：

获取模块，用于获取目标用户的目标用户向量；

25 候选内容确定模块，用于根据所述目标用户向量确定 n 组种子用户向量，所述 n 组种子用户向量与 n 条候选推荐内容对应；所述 n 为正整数；

计算模块，用于调用相似人群扩散模型计算所述目标用户向量和每组所述种子用户向量的相似度，所述相似人群扩散模型用于基于注意力机制计算用户向量之间的相似度；

30 目标内容确定模块，用于根据所述 n 组种子用户向量各自的所述相似度，确定所述 n 条候选推荐内容中向所述目标用户推荐的目标内容。

根据本申请的另一方，提供了一种用户向量提取模型的训练装置，所述装置包括：

35 获取模块，用于获取第一训练样本，所述第一训练样本包括多个样本用户的用户特征和对样本内容的互动记录；

输入模块，用于将第一样本用户的用户特征输入所述用户向量提取模型，得到第一用户向量；所述第一样本用户是所述第一训练样本内多个样本用户

中的任意一个；

计算模块，用于计算所述第一用户向量与所述样本内容的嵌入向量的点积；

5 所述计算模块，还用于根据所述点积和所述第一样本用户对所述样本内容的互动记录，计算得到第一误差损失；

所述计算模块，还用于根据所述第一误差损失采用反向传播算法对所述用户向量提取模型进行训练。

根据本申请的另一方面，提供了一种相似人群扩散模型的训练装置，所述装置包括：

10 获取模块，用于获取第二训练样本，所述第二训练样本包括：样本用户向量和样本内容对应的种子用户向量；

调用模块，用于调用所述第一全连接层对所述种子用户向量进行特征提取，得到种子相似扩散向量；以及，调用所述第二全连接层对所述样本用户向量进行特征提取，得到样本相似扩散向量；

15 所述调用模块，还用于调用所述注意力单元根据所述种子相似扩散向量和所述样本相似扩散向量计算得到所述种子加权向量；

计算模块，用于计算所述种子加权向量和所述样本相似扩散向量之间的向量距离；

20 训练模块，用于根据所述向量距离计算第二误差损失，根据所述第二误差损失采用反向传播算法对所述相似人群扩散模型进行训练。

根据本申请的另一方面，提供了一种计算机设备，所述计算机设备包括：

处理器和存储器，所述存储器存储有计算机程序，所述计算机程序由所述处理器加载并执行以实现如上所述的内容推荐方法、或者用户向量提取模型的训练方法、或者相似人群扩散模型的训练方法。

25 根据本申请的另一方面，提供了一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序由所述处理器加载并执行以实现如上所述的内容推荐方法、或者用户向量提取模型的训练方法、或者相似人群扩散模型的训练方法。

30 根据本申请的另一方面，提供了一种计算机程序产品，包括指令，当其在计算机上运行时，使得计算机执行如上所述的内容推荐方法、或者用户向量提取模型的训练方法、或者相似人群扩散模型的训练方法。

本申请实施例提供的技术方案带来的有益效果至少包括：

35 采用 n 组种子用户向量来表征 n 条候选推荐内容在兴趣方面的特性，并在相似人群扩散模型中引入注意力机制，该注意力机制用于计算目标用户向量和候选推荐内容对应的多个种子用户向量之间的相似度，以确定目标用户对候选推荐内容的感兴趣程度，由于该注意力机制使得同一组种子用户向量在对不同的目标用户进行兴趣预测时提供不同的加权，从而提高了种子用户

向量的表达能力, 有效提高相似人群扩散模型在预测时的准确性和可靠程度。

附图说明

为了更清楚地说明本申请实施例中的技术方案, 下面将对实施例描述中所需要使用的附图作简单地介绍, 显而易见地, 下面描述中的附图仅仅是本申请的一些实施例, 对于本领域普通技术人员来讲, 在不付出创造性劳动的前提下, 还可以根据这些附图获得其他的附图。

图 1 是本申请一个示例性实施例提供的内容推荐系统的结构框图;

图 2 是本申请一个示例性实施例提供的内容推荐方法的示意图;

图 3 是本申请一个示例性实施例提供的内容推荐方法的流程图;

10 图 4 是本申请另一个示例性实施例提供的内容推荐方法的流程图;

图 5 是本申请另一个示例性实施例提供的内容推荐方法的流程图;

图 6 是本申请另一个示例性实施例提供的内容推荐方法的结构图;

图 7 是本申请另一示例性实施例提供的内容推荐方法的流程图;

15 图 8 是本申请一示例性实施例提供的内容推荐方法的用户向量的提取过程图

图 9 是本申请另一个示例性实施例提供的神经网络模型推荐后的界面内容图;

图 10 是本申请一个示例性实施例提供的用户向量的获取过程的流程图;

20 图 11 是本申请一个示例性实施例提供的神经网络模型中的用户向量获取方法的流程图;

图 12 是本申请一个示例性实施例提供的相似人群扩散模型的训练方法的流程图;

图 13 是本申请一个示例性实施例提供的注意力单元的结构图;

图 14 是本申请一个示例性实施例提供的内容推荐的界面图;

25 图 15 是本申请一个示例性实施例提供的内容推荐功能的显示流程图;

图 16 是本申请另一个示例性实施例提供的内容推荐的界面图;

图 17 是本申请一个示例性实施例提供的性能比较图;

图 18 是本申请另一个示例性实施例提供的性能比较图;

图 19 是本申请一个示例性实施例提供的内容推荐装置的框图;

30 图 20 是本申请另一个示例性实施例提供的内容推荐装置的框图;

图 21 是本申请另一个示例性实施例提供的内容推荐装置的框图;

图 22 是本申请一个实施例提供的一种服务器的结构框图。

具体实施方式

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

首先对本申请实施例涉及的若干个名词进行简介:

-5-

神经网络模型：是由多个处理单元（称为神经元）广泛地互相连接而形成的复杂网络系统。神经网络模型用于模拟和反映人脑功能的许多基本特征，是一个高度复杂的非线性动力学习系统。

5 相似人群扩散（look-alike）模型：是一种基于种子用户和目标用户之间的相似性，对目标用户进行内容推荐的神经网络模型。种子用户可以是一个用户或一群用户，比如，点击过网页 A 的 300 万个用户可被视为该网页 A 的种子用户人群。

10 注意力机制（Attention Mechanism）：是利用有限的注意力资源从大量信息中快速筛选出高价值信息的手段。注意力机制被广泛使用在自然语言处理、图像识别及语音识别等各种不同类型的深度学习任务中，是深度学习技术中最值得关注与深入了解的核心技术之一。

目标用户向量：用于表征目标用户的兴趣的向量。

种子用户向量：用于表征种子用户的兴趣的向量。

15 某个候选推荐内容对应的种子用户人群：包括与该候选推荐内容发生过互动行为的用户。发生过互动行为的用户包括但不限于：点击过候选推荐内容的用户、分享过候选推荐内容的用户、阅读停留时长超过阈值的用户、评论或点赞过候选推荐内容的用户中的至少一种。

内容：采用文字、图片、表情、音频和视频中的至少一项来组成的信息集合。比如，内容可以是网页、电视剧、音乐等。

20 在相似人群扩散模型中，通过计算目标用户和种子用户之间的相似度来进行内容推荐。种子用户通常为很多个，比如几十个、几百个或几千个。在这些种子用户中，可能存在一些种子用户与目标用户非常相似，而另外一些种子用户与目标用户之间的相似度较低。通常情况下，每个种子用户都是同等对待的，因此相似度较低的种子用户会引入较大的噪声，使得相似人群扩散模型的精度下降。

30 本申请实施例中，在相似人群扩散模型中引入了注意力机制，该注意力机制用于计算目标用户向量和候选推荐内容对应的种子用户向量之间的相似度。针对不同的目标用户，该注意力机制能够从多个种子用户向量中，选择出对当前的目标用户进行内容推荐更具参考价值的向量信息，从而提高向目标用户推荐内容的准确性。

图 1 示出了本申请一个示例性实施例提供的内容推荐系统 100 的结构框图。该内容推荐系统 100 包括：终端 110 和推送平台 140。

35 终端 110 通过无线网络或有线网络与推送平台 140 相连。终端 110 可以是智能手机、游戏主机、台式计算机、平板电脑、电子书阅读器、MP3 播放器、MP4 播放器和膝上型便携计算机等设备中的至少一种。终端 110 安装和

-6-

运行有支持内容推荐的应用程序。该应用程序可以是声音社交应用程序、即时通讯应用程序、团队语音应用程序、基于话题或频道或圈子进行人群聚合的社交类应用程序、基于购物的社交类应用程序、浏览器程序、视频程序等应用程序中的任意一种。示意性的，终端 110 是第一用户使用的终端，终端 5 110 中运行的应用程序登录第一用户帐号。

终端 110 通过无线网络或有线网络与推送平台 140 相连。

推送平台 140 包括一台服务器、多台服务器、云计算平台和虚拟化中心中的至少一种。推送平台 140 用于为支持内容推荐的应用程序提供后台服务。可选地，推送平台 140 承担主要计算工作，终端 110 承担次要计算工作；或者，10 推送平台 140 承担次要计算工作，终端 110 承担主要计算工作；或者，推送平台 140、终端 110 两者之间采用分布式计算架构进行协同计算。

在一些实施例中，推送平台 140 可以包括：接入服务器 141、内容推荐服务器 144 和数据库 146。接入服务器 141 用于为终端 110 提供接入服务。内容推荐服务器 144 用于提供内容推荐有关的后台服务。内容推荐服务器 144 15 可以是一台或多台。当内容推荐服务器 144 是多台时，存在至少两台内容推荐服务器 144 用于提供不同的服务，和/或，存在至少两台内容推荐服务器 144 用于提供相同的服务，比如以负载均衡方式提供同一种服务，本申请实施例对此不加以限定。内容推荐服务器 144 中设置有相似人群扩散模型。在本申请实施例中，将基于实时注意力机制的相似人群扩散模型简称为 RALM 20 (Realtime Attention based Look-alike Model)。

本领域技术人员可以知晓，在图 1 所示的内容推荐系统 110 中，终端 110 的数量可以更多或更少，比如上述终端 110 可以仅为一个，或者上述终端 110 为几十个或几百个，或者更多数量，此时上述内容推荐系统还包括其他终端 110。本申请实施例对终端 110 的数量和设备类型不加以限定。

图 2 示出了一个示例性实施例提供的内容推荐方法的示意图。在内容推荐服务器 114 中运行有基于注意力机制的相似人群扩散模型 1141。

内容推荐服务器 114 提取到目标用户的用户特征，并将该目标用户的用户特征输入到用于提取用户向量的用户向量提取模型 240 中，用户向量提取模型 240 是基于注意力机制的神经网络模型，该神经网络模型中包括基于自注意力的合并层，将用户特征以向量的形式输入至用户向量提取模型 240 中后，经由合并层计算各个用户特征的权重并进行加权融合，输出能够代表该用户兴趣的用户向量。与传统神经网络模型中广泛使用的连接层相比，该合并层的性能显著优于传统的连接层，能够解决强相关特征和弱相关特征带来的过渡拟合和噪声问题，尤其在多域特征上表现出显著的优越的性能。

每个目标用户对应有候选推荐内容包 241（或候选推荐内容池），候选推荐内容包 241 中包括多条候选推荐内容，例如包括：候选推荐内容 1、候

-7-

选推荐内容 2、候选推荐内容 3.....候选推荐内容 N，每条候选推荐内容对应一组与该候选推荐内容互动过的种子用户，采用向量的方式表示每个种子用户，相应地，与某条候选推荐内容对应的一组种子用户向量即可代表该条候选推荐内容在兴趣维度上的特性。

5 内容推荐服务器 114 将目标用户向量和用于代表某条候选推荐内容的一组种子用户向量，输入至基于注意力机制的相似人群扩散模型 1141 中。该相似人群扩散模型 1141 能够基于注意力机制，从一组种子用户向量中挑选出对于目标用户更有参考价值的向量进行相似度计算。

10 在一些实施例中，该相似人群扩散模型 1141 中的注意力单元包括全局注意力单元和局部注意力单元，其中，全局注意力单元用于捕捉多个种子用户向量的组成情况，对具有代表性的个体（也即与目标用户相近的种子用户向量）进行加权，并对噪声个体（也即与目标用户不太相近的种子用户向量）进行惩罚，相比于对每个种子用户向量进行同等处理的方式能够取得更优的效果；局部注意力单元用于捕捉与目标用户向量最接近的部分种子用户向量，
15 将相关个体加权到目标用户，使得每组种子用户向量的表现力在不同的目标用户之间发生变化，从而大大提高种子用户向量的表达能力。

内容推荐服务器 114 将每条候选推荐内容的种子用户向量与目标用户向量进行计算得到相似度，该相似度用于表示目标用户对该条候选推荐内容的感兴趣程度。将每一条候选推荐内容按照相似度分数进行排序，将排名靠前的
20 的候选推荐内容推荐给目标用户。

图 3 示出了一个示例性实施例提供的内容推荐方法的流程图。该方法可以由图 1 所示的内容推荐服务器 144 来执行，该方法包括以下步骤：

步骤 201，获取目标用户的目标用户向量。

25 目标用户是待推送推荐内容的用户帐号。目标用户向量是用来代表目标用户兴趣的向量。内容推荐服务器 144 根据目标用户向量可以确定推荐给该目标用户的候选推荐内容包，该候选推荐内容包包括 n 条候选推荐内容，n 为大于 1 的整数。

30 可选地，目标用户向量是根据用户在至少一个域中的特征获得的，其中，域是按照特征属性划分的逻辑范围。示意性的，域包括性别、年龄、学历中的至少一种。

步骤 202，根据目标用户向量确定 n 组种子用户向量。

n 组种子用户向量与 n 条候选推荐内容对应。其中，采用 n 组种子用户向量中的第 i 组种子用户向量来表征第 i 条候选推荐内容在兴趣方面的特性。

35 对于第 i 条候选推荐内容，第 i 组种子用户向量对应的各个种子用户，是在历史时间段内与第 i 条候选推荐内容产生过互动行为的用户。

在一个示意性的例子中，内容推荐服务器 144 获得每条候选推荐内容的

历史点击用户的用户特征，点击过该候选推荐内容的用户视为该候选推荐内容的种子用户。一组种子用户的数量可以为 L 个，根据每个种子用户的用户特征计算得到该种子用户的种子用户向量， L 个种子用户向量用于代表该条候选推荐内容在兴趣方面的特性。其中， L 为正整数。

5 步骤 203，调用相似人群扩散模型计算目标用户向量和每组种子用户向量的相似度。

其中，相似人群扩散模型用于基于注意力机制计算相似度。该相似人群扩散模型能够基于注意力机制，从一组种子用户向量中挑选出对于目标用户更有参考价值的种子用户向量进行相似度的计算。

10 内容推荐服务器 144 通过相似人群扩散模型计算出目标用户向量与每组种子用户向量的相似度，由于每组种子用户向量对应一条候选推荐内容，因此，可以利用每条候选推荐内容对应的一组种子用户向量的相似度，来评价目标用户对该条候选推荐内容的感兴趣程度。示意性的，相似度越高表示目标用户对该条候选推荐内容越感兴趣。

15 其中，注意力机制用于对一组中的多个种子用户向量赋予不同的权重，例如：对于与目标用户向量相似度较高的种子用户向量赋予较高的权重，对于相似度较低种子用户向量赋予较低的权重，尽可能减小低相似度的种子用户向量对计算结果的噪声影响。

20 可选地，注意力机制包括：自注意力机制、全局注意力机制、局部注意力机制中的至少一种。

步骤 204，根据 n 组种子用户向量各自的相似度，确定 n 条候选推荐内容中向目标用户推荐的目标内容。

25 在一些实施例中，可以根据相似度对 n 条候选推荐内容进行排序，将排序在前 m 位的候选推荐内容，确定为向目标用户推荐的目标内容， n 和 m 均为正整数， m 小于或等于 n 。在一些实施例中，也可以确定相似度高于预设阈值的候选推荐内容，作为向目标用户推荐的目标内容。本申请在此不对基于相似度确定目标内容的方式做任何限定。

30 需要说明的是，本实施例采用相似度高的前几条候选推荐内容作为向目标用户推荐的目标内容，在其他可能的实施方式中，也可以采用差异度分数对候选推荐内容进行排序，差异度分数用于表示目标用户对该条候选推荐内容的不感兴趣程度，根据目标用户的不感兴趣的候选推荐内容确定推荐策略。

35 综上所述，本实施例在相似人群扩散模型中引入注意力机制，该注意力机制用于计算目标用户向量和候选推荐内容对应的多个种子用户向量的相似度，并根据相似度确定目标用户对候选推荐内容的感兴趣程度，通过比较目标用户和多个种子用户之间的兴趣相似度，预测目标用户对一条候选推荐内容感兴趣的可能性，采用一组种子用户向量来表征一条候选推荐内容在兴趣方面的特性，能有效提高内容预测的准确性和可靠程度。

基于上述实施例中的内容，图 4 示出了另一个示例性实施例提供的内容推荐方法的流程图。上述步骤 202 包括以下步骤 202a 和步骤 202b：

步骤 202a，确定目标用户的 n 条候选推荐内容。

5 由于内容推荐系统中可供推荐的内容可能会非常多（比如上亿条），针对单个目标用户，内容推荐服务器 144 可以通过预设方式，初步筛选出目标用户对应的候选推荐内容包。

10 预设方式包括但不限于：采用其它神经网络模型进行筛选、基于目标用户的好友用户查看的内容进行筛选、基于最近需要投放的内容进行筛选等方式。本申请实施例对目标用户对应的 n 条候选推荐内容的确定方式不加以限定。

步骤 202b，对于 n 条候选推荐内容中的每条候选推荐内容，确定与该条候选推荐内容对应的一组种子用户向量。

15 对于 n 条候选推荐内容中的每条候选推荐内容，将与该候选推荐内容在预设时段内产生过互动行为的 L 个用户的用户向量，作为与候选推荐内容对应的一组种子用户向量，每组种子用户向量包括 L 个用户的用户向量。

20 示例性的，内容推荐系统确定一条候选推荐内容，将在预设时段内（如最近一周内、最近一个月内等）点击过（或阅读过、分享过、评论过、点赞过、采用虚拟资源或实体资源进行赞赏过等）该条候选推荐内容的 300 万个用户的用户向量作为种子用户向量。当然，也可以将与该条候选推荐内容以其他形式进行互动的用户，做为该条候选推荐内容的种子用户。

25 本技术方案通过在相似人群扩散模型中引入基于注意力机制的注意力单元，对不同的种子用户赋予不同的权重，从而提高相似人群扩散模型的推荐准确度。下面结合图 5 和图 6 对本方案做进一步详细说明，其中，图 5 示出了另一个示例性实施例提供的内容推荐方法的流程图，图 6 示出了另一个示例性实施例提供的内容推荐方法的流程图。

30 结合图 5，建立了塔状的学习模型。左侧的称为“种子塔”，以 n 个种子用户的种子用户向量作为输入量，记为 $R_{\text{seeds}} \in \mathbb{R}^{n \times m}$ ，其中 m 表示种子用户向量的维数。在第一层包括作为变换矩阵的第一全连接层 310，将种子用户向量的 $n \times m$ 的输入矩阵转换为 $n \times h$ 的矩阵大小，其中 h 表示变换后的种子 look-alike 向量的维数。然后，全局注意力单元根据多个种子 look-alike 向量计算得到全局向量，局部注意力单元将多个种子 look-alike 向量和目标 look-alike 向量计算得到局部向量，将全局向量和局部向量合并到一个维数为 h 的种子加权向量中。右侧的称为“目标塔”，一个维数为 m 的目标用户向量被转化为维数 h 。在这两个塔的顶部，计算出种子加权向量和目标 look-alike 向量的点积，用于表示种子加权向量和目标 look-alike 向量的相似度。关于推

35

荐方面，相似度本质上是目标用户单击候选推荐内容的可能性。

相似人群扩散模型包括：第一全连接层 310、第二全连接层 320、注意力单元 330 和相似度计算层 340；调用第一全连接层 310 对种子用户向量进行特征提取，得到种子 look-alike 向量，调用第二全连接层 320 对目标用户向量进行特征提取，得到目标 look-alike 向量；其中，第一全连接层 310 和第二全连接层 320 分别用于将种子用户向量和目标用户向量转换到神经网络能够识别的向量空间中。

示意性的，每条候选推荐内容在兴趣维度的特性，采用 300 万个点击过该条候选推荐内容的种子用户向量来进行表示。一方面，将该 300 万个种子用户向量输入至第一全连接层 310 中，提取 300 万个种子用户向量特征，得到种子 look-alike 向量。另一方面，将目标用户向量输入至第二全连接层 320 中，提取目标用户向量特征，得到目标 look-alike 向量。可选地，种子 look-alike 向量和目标 look-alike 向量的计算过程是互相独立的，本实施例对两者的执行先后顺序不加以限定。

在该过程中，通过注意力单元 330 根据种子 look-alike 向量和目标 look-alike 向量计算得到全局向量和局部向量，根据全局向量和局部向量计算得到种子加权向量；通过相似度计算层 340 计算种子加权向量和目标 look-alike 向量之间的向量距离，将向量距离确定为相似度。然后，内容推荐服务器 144 按照相似度由高到低的顺序，进行内容推荐。

综上所述，本实施例中通过注意力单元对点击过该条候选推荐内容的多个种子用户向量和目标用户向量进行计算，能够有效提高种子用户向量和目标用户向量之间的相似度的准确度，减少种子用户向量中的噪声影响，从而获得能够更准确地反映目标用户对候选推荐内容的感兴趣程度。

图 6 示出了一个示例性实施例提供的内容推荐方法的流程图。示意性的，一条候选推荐内容的种子用户群由与该条候选推荐内容互动过的 L 个种子用户组成，种子用户群和目标用户通常受两方面影响，一方面是种子用户自身的组成情况，另一方面是目标用户与种子用户群中相似的用户信息。

结合图 5，本实施例中的注意力单元 330 包括局部注意力单元，用于提取种子用户向量中的局部表示，并将相关个体加权到目标用户中，使种子用户向量的表现力在不同的目标用户之间发生变化。本实施例中的注意力单元 330 还包括全局注意力单元，用于提取全部种子用户向量的自身组成情况，对全部种子用户向量中具有代表性的个体进行加权，并对噪声个体进行惩罚，从而获得更准确的种子向量表示。示意性的，注意力单元的执行步骤包括两个分支：

第一分支包括：

步骤 2031a，调用第一全连接层对种子用户向量进行特征提取，得到种子

look-alike 向量。

步骤 2031b, 调用局部注意力单元对目标 look-alike 向量和种子 look-alike 向量计算得到局部注意力向量。

$$E_{local_s} = E_s \text{softmax}(\tanh(E_s^T W_l E_u))$$

其中, E_s 表示种子用户向量矩阵, E_u 表示目标用户向量, W_l 表示局部注意力的转换矩阵, E_{local_s} 表示最终经过注意力计算后加权求和所得的局部注意力向量。

为了计算种子用户向量和目标用户向量之间的相似性, 本实施例必须将多个种子用户向量汇集成一个, 均值池化是一种常用的方法。一般来说, 在数百万种子用户向量中, 可能只有一小部分种子用户的兴趣与目标用户有关。因此, 本实施例在这里设置了一个局部注意力单元, 以激活对某个目标用户的局部兴趣并进行适应性调整, 并自适应地向目标用户学习种子的个性化表示

第二分支包括:

步骤 2032a, 调用第二全连接层对目标用户向量进行特征提取, 得到目标 look-alike 向量。

步骤 2032b, 调用全局注意力单元对种子 look-alike 向量计算得到全局注意力向量。

可选地, 全局注意力单元捕捉相应的信息生成对应向量, 计算公式如下:

$$E_{global_s} = E_s \text{softmax}(E_s^T \tanh(W_g E_s))$$

其中, E_s 表示种子用户向量矩阵, W_g 表示全局注意力的转换矩阵, E_{global_s} 表示最终经过注意力计算后加权求和所得的全局注意力向量。

步骤 2033, 对局部注意力向量和全局注意力向量进行加权计算, 得到种子加权向量。

步骤 2034, 计算种子加权向量和目标 look-alike 向量之间的向量距离, 将向量距离确定为相似度。

通过相似度算法得出两个向量的相似度。可选地, 向量的相似度采用两个向量之间的余弦距离来表征。该余弦距离采用步骤 2032 中的目标 look-alike 向量和步骤 2033 中的种子加权向量之间的乘积计算得到。

示意性的, 步骤 2033 和步骤 2034 可以通过如下公式计算:

$$\text{score}_{u,s} = \alpha \cdot \text{cosine}(E_u, E_{global_s}) + \beta \cdot \text{cosine}(E_u, E_{local_s})$$

其中, α 和 β 是加和权重。作为一个示例, $\alpha=0.3$, $\beta=0.7$ 。

综上所述, 本实施例通过采用全局注意力单元和局部注意力单元对种子用户向量进行计算。其中, 全局注意力单元用来学习种子用户向量的全局代表性, 对具有代表性的个体进行加权, 并对噪声个体惩罚, 从而获得更为准

确的全局表示；局部注意力单元针对特定的目标用户自适应地学习种子表示，使种子的表示因受众不同而不同，从而显著提高了种子用户向量的表达能力。

结合参考图 7，本申请实施例提供了一种基于实时注意力机制的相似人群扩散模型 RALM，该 RALM 包括有用户向量提取模型 240 和相似人群扩散模型 1141。该用户向量提取模型 240 能够提取目标用户的目标用户向量。相似人群扩散模型 1141 包括全局注意力单元 1141a 和局部注意力单元 1141b。

以候选推荐内容是最新消息或长尾内容为例，每个最新消息或长尾内容对应有一组种子用户向量，采用该组种子用户向量来代表该候选推荐内容在兴趣方向的特性，全局注意力单元 1141a 用来基于一组种子用户向量计算全局向量，局部注意力单元 1141b 用来基于一组种子用户向量和目标用户向量计算局部向量，然后根据全局向量和局部向量计算得到种子加权向量，计算种子加权向量和目标用户向量之间的距离作为相似度。RALM 按照相似度由高到低的顺序，向目标用户推送候选推荐内容。

用户向量提取模型 240 应该使用尽可能多的用户特征域来进行目标用户表示学习，以提高用户兴趣的多样性。基于深度学习的模型能够对多种特征进行建模。然而根据发明人的实验发现，类似 DNN 的深度模型在输入弱相关特征时表现出更差的性能；另外，像感兴趣的标签这样的强相关特征会导致过拟合，像购物兴趣这样的弱相关特征甚至会产生一定程度的负面影响。本申请实施例提供了一种采用注意力机制的用户向量提取模型 240，能够解决或改善这些问题。

图 8 示出了一示例性实施例提供的用户向量的生成方法的流程图。该方法可以用于生成每个用户的用户向量，比如目标用户的目标用户向量和种子用户的种子用户向量。

作为本实施例的一个示例，用户的用户特征包括：离散特征和连续特征中的至少一种。

可选地，用户特征包括但不限于：人口属性特征、设备特征、历史行为特征和标签特征中的至少一类。在一个示例中，离散特征包括人口属性特征、设备特征和标签特征；连续特征包括历史行为特征。

人口属性特征用于指示用户在人物基础属性上的特征。示意性的，人口属性特征包括但不限于：性别、年龄、学历、婚姻情况、地址、收入中的至少一种。其中，性别可视为一个域（或用户特征域）、年龄可视为一个域、学历可视为一个域等等。域是对一个训练集中样本子集产生整体性分布偏差的因素，比如对于人脸识别，不同用户的发色黑、黄和白可视为一种域的差异；不同用户是否戴有眼镜，也可视为一种域的差异；不同用户是否蓄有胡子，也视为一种域的差异。

设备特征用于指示用户使用的电子设备的特征。示意性的，设备特征包括但不限于：硬件型号、操作系统类型、生产厂商、国别版本、唯一硬件标

识码、网络地址中的至少一种。

历史行为特征用于表征用户在使用互联网服务时的各种行为产生的特征。示意性的，历史行为特征包括但不限于：阅读行为、点赞行为、回复行为、分享行为、购买行为、网络社交行为、加入会员行为中的至少一种。

5 标签特征是由系统或其他用户以标签形式为该用户标定的特征。标签特征可以是用户感兴趣的标签、购物兴趣标签、点击兴趣标签等。示意性的，标签特征包括但不限于：幽默、善良、90后、金牛座、美食、起床困难户、宅女、韩剧达人、旅游、完美主义等。

10 可选地，人口属性特征、设备特征、行为特征和标签特征均采用向量表示。

用户向量提取模型240包括：嵌入层(含均值池化层601和全连接层602)、合并层603和全连接层604。

15 用户向量提取模型240将离散特征经过均值池化层601转换为向量形式，将连续特征经过全连接层602转换为向量模式，将向量形式的用户特征输入至合并层603中。其中，合并层603是基于自注意力机制的神经网络层，由特征输入层提取的向量形式的用户特征，经由合并层603计算各用户特征的权重并进行加权融合，得到更能代表当前用户的用户兴趣的用户向量。

合并层603将输出结果输入至全连接层604中，其中，全连接层604用于捕获各向量的非线性关联关系，最终输出能够代表该用户兴趣的用户向量。

20 作为本实施例的一个示例，上述合并层603可以基于改进的多层感知器(Multi-Layer Perceptron, MLP)来实现。在图8的右下框中示出了基本的多层感知器和本实施例中改进的多层感知器的对比。

25 基本的多层感知器包括：连接层61和MLP层。多个域的用户特征的向量表示通过连接层61连接为一个宽的密集层向量(包括图8中的向量1和向量2)，该宽的密集层向量将输入至MLP层中进行处理。发明人通过对基本的多层感知器的训练过程进行观察发现，训练过程的权重优化总是对某些特定的域进行过拟合，这些特定的域与目标用户对内容的兴趣(比如感兴趣的标签)密切相关，导致推荐结果总是由强相关特征来决定，导致弱相关特征对推荐过程的贡献是不足的，比如用户的购物兴趣对推荐过程做出贡献不足。

30 因此，基本的多层感知器无法全面理解多域特征，导致最终的推荐结果缺乏多样性。

35 改进的多层感知器包括：注意力合并层62和MLP层。该注意力合并层62包括：并列的多个全连接层和自注意力单元。全连接层的数量与用户特征域的数量相同，每个域的用户特征向量通过全连接层进行连接后，经过自注意力单元根据上下文用户特征来学习权重(比如图中的a1和a2)的个性化分布，从而激活不同域对应的神经元的不同部位，确保强相关特征和弱相关特征都能有效地发挥作用。

综上所述，本实施例提供的方法，通过注意力合并层 62 能够确保强相关特征（比如感兴趣的标签）和弱相关特征（比如购物兴趣）均能够对推荐过程做出贡献，从而提高了用户向量提取模型所提取到的用户特征向量准确性，实现“千人千面”的内容推荐效果。

5

图 9 示出了一示例性实施例提供的内容推荐方法的流程图，该方法包括：在线服务阶段、在线异步处理阶段和离线训练阶段。

在线服务阶段

即为图 3、图 4 和图 6 实施例介绍的预测过程。该在线服务阶段包括
10 步骤 701，获取目标用户的目标用户向量；步骤 702，根据目标用户向量确定目标用户对应的 n 条候选推荐内容，遍历候选推荐内容；其中，每个目标用户都可以对应多条候选推荐内容，内容推荐服务器对每条候选推荐内容 713 对应的多个种子用户向量 714 与目标用户向量进行比较，在比较过程中采用局部注意力单元 703 和全局注意力单元 704 进行特征提取，根据目标用户向量的特性在不同维度对种子用户向量赋予不同权重，步骤 705，计算目标用户向量与代表该条候选推荐内容的多个种子用户向量的相似度分数；步骤
15 706，根据每条候选推荐内容的相似度分数进行排序服务，然后根据相似度分数由高到低的顺序向目标用户进行内容推荐。

在线异步处理阶段

20 为保证新闻资讯等内容能够被实时快速曝光，图 9 所示的推荐系统中还包括用户反馈监控模块，当一条推荐内容被推送到多个目标用户后，每分每秒都会产生很多互动的用户（比如点击该推荐内容的用户），用户反馈监控模块会实时获取用户点击日志 712，采集最近点击过该推荐内容的用户作为种子用户，因此每条候选推荐内容对应的种子用户的名单会实时更新。

25 示意性的，用户反馈监控模块实时更新库存的候选推荐内容对应的种子用户名单及种子用户向量，为防止短时间内候选推荐内容的点击用户过多，出现内容推荐服务器 114 在对种子用户向量进行后续聚类计算时负担过大的问题，每条候选推荐内容对应的种子用户群仅保留最近产生有互动行为的 300 万个用户。当然，每条候选推荐内容仅保留用户的数量可以是 300 万，也可以是其他数量，具体参数由技术人员设定。互动行为包括点击、阅读、点赞、
30 评论、转发中的至少一种。

另一方面，对于一条候选推荐内容 713，上述实施例中以选取最近点击过该候选推荐内容的 300 万个用户作为种子用户，基于注意力机制的相似人群扩散模型 1141 计算 300 万个种子用户的种子用户向量与目标用户向量之间的相似度分数。由于对于每条候选推荐内容计算 300 万个种子用户向量的计算量非常大，因此为了减少相似人群扩散模型 1141 的计算量和计算时间，本
35 实施例可选地采用聚类的方式减轻相似人群扩散模型 1141 的运算负担。

可选地，本实施例采用目标聚类算法 711（比如 K-means 聚类）对同一组中的种子用户向量进行聚类处理，对于 n 组种子用户向量中的第 i 组种子用户向量，将第 i 组种子用户向量中的 L 个种子用户向量聚类为 K 个种子用户向量， K 小于 L ， i 小于或等于 n 。

5 示例性的，聚类方法包括：将 L 个种子用户向量采用 K-means 聚类算法聚类为 K 个集合；将同一个集合中的至少两个种子用户向量进行平均，得到与该集合对应的种子用户向量。以 L 为 300 万， K 为 300 个为例，该 K-means 聚类过程能够将一条候选推荐内容对应的 300 万个种子用户向量，聚类为 300 个种子用户向量。或者，对于所述 K 个集合中的每个集合，将该集合中位于
10 聚类质心的种子用户向量，确定为与该集合对应的种子用户向量。

当候选推荐内容对应 L 个种子用户向量时，将 L 个种子用户向量异步聚类为 K 个种子用户向量，由于 K 小于 L ，使计算次数缩小至原来的 K/L ，能有效减少神经网络模型的运算量。

作为本申请的一个示例，当种子用户包括两个以上时：种子用户 1、种子用户 3、...、种子用户 x ，种子用户 1 的种子用户向量为 $[0.1, 0.2, 0.8, 0.9, 0.2, 0.4, 0.6]$ ，种子用户 3 的种子用户向量为 $[0.8, 0.2, 0.2, 0.2, 0.0, 0.6, 0.6]$ ，...，种子用户 x 的种子用户向量为 $[0.5, 0.7, 0.2, 0.5, 0.6, 0.7, 0.8]$ 。每个种子用户的用户向量为 M 维向量，将每个 M 维的用户向量中的第 i 维度值进行平均， $1 \leq i \leq M$ ，合并得到一个新的 M 维向量作为该种子用户的用户特征。
15 比如，将每个种子用户向量中的第 1 维度值 0.1、0.8、...、0.5 进行平均，从而得到聚类后的种子用户向量中的第 1 维度值 0.45，以此类推。

作为本申请的另一个示例，虽然种子用户向量是不断实时更新的，但每次插入新的种子用户向量后，不需要立即对 K-means 聚类的种子用户向量簇进行更新。推荐系统可以每 5 分钟（或其他时间间隔）来运行一次 K-means
25 算法，并将聚类得到的每个种子用户向量簇中的聚类质心向量作为聚类后的种子用户向量保存至数据库中，用于在线预测过程中。

聚类后的种子用户向量定义为：

$$R_{\text{seeds}} = \{E_{\text{centroid1}}, E_{\text{centroid2}}, \dots, E_{\text{centroidk}}\};$$

其中， $E_{\text{centroidk}}$ 是 K-means 聚类的第 k 个种子用户向量簇的聚类质心向量。

30 需要说明的是，K-means 聚类算法只是示例性的举例说明，其他聚类算法如均值漂移聚类、基于密度的聚类方法(Density-Based Spatial Clustering of Applications with Noise, DBSCAN)、用高斯混合模型 (Gaussian Mixture Model, GMM) 聚类等其他算法也可以达到减少神经网络模型计算量的目的，本实施例对此不做限定。

35 综上所述，本实施例提供的方法，通过对在线异步处理过程中的种子用户向量进行实时更新，使得突发性的最新新闻能够尽快收集足够多的用户历

史数据，从而尽快地作为候选推荐内容推荐给目标用户；还能够基于小众高质量文章的最近点击用户的种子用户向量，将小众高质量文章尽快推荐给潜在感兴趣的目标用户。

- 5 本实施例提供的方法，通过 K-means 算法对一组的数量庞大的种子用户向量进行聚类，从而得到精简数量后的种子用户向量，从而极大程度上减轻内容推荐服务器在计算相似度分数时的计算量。

离线训练阶段

该阶段是采用训练集对用户向量提取模型 240、以及相似人群扩散模型 1141 进行训练的阶段。

- 10 对用户向量提取模型 240 进行训练的过程，可称为用户表示学习 707，在该阶段中，离线训练服务器中输入各个用户的用户特征、以及每个用户对推荐内容的互动行为（如阅读文章、观看视频、订阅公众号等）作为样本，采用误差反向传播算法对用户向量提取模型 240 进行训练。

- 15 对相似人群扩散模型 1141 进行训练的过程，可称为 look-alike 学习 708。look-alike 学习 708 以用户表示学习 707 生成的能够代表用户兴趣的用户向量做为输入内容，将该用户的历史点击作为样本对用户向量进行训练，经过以点击作为正样本，不点击作为负样本，根据误差反向传播来调整相似人群扩散模型 1141 中各个神经网络层的权重。

- 20 下面采用两个不同的实施例来对用户向量提取模型 240 的训练过程，以及相似人群扩散模型 1141 的训练过程进行分别阐述。

图 10 示出了本申请一个示例性实施例提供的用户向量提取模型的训练方法的流程图。该方法可以由内容推荐服务器 114 或者额外的离线训练服务器来执行。用户向量提取模型包括嵌入层、合并层和全连接层，其中合并层是基于自注意力机制的神经网络层。该训练方法包括以下步骤：

- 25 步骤 1001，获取第一训练样本，第一训练样本包括多个样本用户的用户特征和对样本内容的互动记录；

第一训练样本包括：样本用户的用户特征、样本内容的嵌入向量、样本用户对样本内容的互动记录。其中，互动记录包括：点击、点赞、评论、转发中的至少一种。

- 30 步骤 1002，将第一样本用户的用户特征输入用户向量提取模型，得到第一用户向量；

结合图 11 可知，用户向量提取模型包括：嵌入层（含均值池化层 601 和全连接层 602）、合并层 603 和全连接层 604。在一些实施例中，合并层 603 包括：注意力合并层和 MLP 层。

- 35 将第一样本用户的离散特征经过映射(embedding)后输入均值池化层 601 转换为向量形式，将连续特征经过全连接层 602 转换为向量形式，将向量形式的用户特征输入至合并层 603 中，其中，合并层 603 是基于自注意力机制

的神经网络模型。

向量形式的用户特征经由合并层 603 计算各个用户特征的权重并进行加权融合，得到能够代表第一样本用户的用户兴趣的第一用户向量。第一样本用户是第一训练样本内多个样本用户中的任意一个。

5 步骤 1003，计算第一用户向量与样本内容的嵌入向量的点积；

步骤 1004，根据点积和第一样本用户对样本内容的互动记录，计算得到第一误差损失；

10 其中，在已有的多个候选内容中选择样本内容的方法可以视为：将用户表示学习作为多分类学习，从数百万候选内容中选择一个感兴趣的内容。为了提高训练效率，在计算第一误差损失时，本申请采用负采样代替传统的归一化指数 softmax 函数。显然，如果随机选取候选内容作为负样本，抽样分布就会偏离实际。因此，本实施例开发了一个类似噪声对比估计的损失函数，即谷歌在 Word2vec 中指出的损失函数。为了模拟一个无偏分布，本实施例首先根据所有候选内容的出现频率对它们进行排序，然后根据其排名计算每个
15 候选内容作为样本内容的可能性：

$$p(x_i) = [\log(k+2) - \log(k+1)] / \log(D+1);$$

20 其中 x_i 表示第 i 项候选内容， k 表示第 i 项候选内容的秩， D 表示所有候选内容的最大秩， $p(X_i)$ 表示选择第 i 项候选内容作为负样本的可能性。当采用最活跃的用户行为主导训练损失时，本实施例将每个候选用户的最大正样本限制在 50 个，样本的正负比例为 1/10。然后用 Softmax 函数，根据候选用户的特征 U 和第 i 项候选内容的特征，对第 i 项候选内容上选择 c 的可能性进行规范化：

$$P(c = i | U, X_i) = \frac{e^{x_i u}}{\sum_{j \in X} e^{x_j u}}$$

25 其中， $u \in \mathbb{R}^N$ 表示第一样本用户的高维嵌入向量， $x_j \in \mathbb{R}^N$ 表示所有样本内容的嵌入向量。作为本申请的一个示例，样本内容的嵌入向量是由一个深层次的预测神经网络来产生的， e 是自然常数，该预测神经网络可以是相关技术中用来进行预测的神经网络。此外，本实施例还以显式反馈和隐式反馈为例，确保推荐结果的多样性。所有类型的内容如文章、视频和网站上的行为都将被考虑。

30 设第一样本用户对样本内容产生了互动行为作为正样本，第一样本用户对样本内容未产生互动行为作为负样本。离线训练服务器计算点积与正样本（或负样本）之间的第一误差损失。

设向量 u 为第一样本用户的高维嵌入向量， x_i 表示第 i 个样本内容的嵌入向量，本实施例计算了 $P(c=i | U, X_i)$ 和交叉熵损失：

$$L = - \sum_{j \in X} y_i \log P(c = i | U, X_i)$$

其中 $y_i \in \{0, 1\}$ ，表示标签，0 代表未点击行为，1 代表点击行为。使用 ADAM 优化器来最小化误差损失。当误差损失收敛时，用最后一层的输出作为用户的用户向量（目标用户的目标用户向量或种子用户的种子用户向量）。

5 步骤 1005，根据第一误差损失采用反向传播算法对用户向量提取模型进行训练。

示意性的，离线训练服务器采用基于梯度下降的反向传播（Back propagation）算法对预测误差进行反向传播，从而更新神经网络模型中的权重。

10 将一组训练样本输入用户向量提取模型进行训练，视为一次训练过程。当一次训练过程结束后，离线训练服务器判定本次训练过程是否满足训练结束条件。训练结束条件包括但不限于如下两种情况：

一、预测误差小于设定阈值；

该设定阈值由技术人员预先设定，或根据训练过程进行随时调整。

15 二、训练次数大于设定次数。

由于某些情况下，长时间训练也无法使得预测误差小于某个设定阈值。因此也可按照训练次数来设定训练结束条件，比如该训练次数达到 10000 次，则停止训练。

20 当本次训练过程未满足训练结束条件时，根据反向传播算法对预测误差进行反向传播，更新用户向量提取模型中的权重；当本次训练过程满足训练结束条件时，结束用户向量提取模型的训练。

综上所述，本实施例提供的用户向量提取模型的训练方法中，由于采用合并层来基于自注意力机制对多个域的用户特征进行特征提取，通过合并层能够确保强相关特征（比如感兴趣的标签）和弱相关特征（比如购物兴趣）
25 均能够对推荐过程做出贡献，从而大幅度提高用户向量提取模型提取到的用户向量的表示准确性。

图 12 示出了本申请一个示例性实施例提供的相似人群扩散模型 1141 的训练方法的流程图。该方法可以由内容推荐服务器 1141 或者额外的离线训练
30 服务器来执行。相似人群扩散模型包括：第一全连接层、第二全连接层和注意力单元。该训练方法包括以下步骤：

步骤 1201，获取第二训练样本，第二训练样本包括：样本用户向量和样本内容对应的种子用户向量；

35 样本内容可以是文章、广告、新闻等，用于训练相似人群扩散模型的第二训练样本的样本数量根据技术人员确定，本实施例不做限定。

每个样本内容对应的种子用户向量，是对该样本内容产生过互动行为的用户的用户向量。

步骤 1202，调用第一全连接层对种子用户向量进行特征提取，得到种子 look-alike 向量，以及调用第二全连接层对样本用户向量进行特征提取，得到样本 look-alike 向量；

步骤 1203，调用注意力单元根据种子 look-alike 向量和样本 look-alike 向量计算得到种子加权向量；

可选地，注意力单元包括：全局注意力单元和局部注意力单元。

结合参考图 13，本申请实施例中的全局注意力单元包括：双曲线正切函数 1311、线性函数 1312、归一化指数函数 1313。本申请实施例中的局部注意力单元包括：矩阵乘法函数 1321、双曲线正切函数 1322、线性函数 1323 和归一化指数函数 1324。

离线训练服务器通过全局注意力单元中的双曲线正切函数 1311、线性函数 1312、归一化指数函数 1313 依次对种子 look-alike 向量进行处理，得到全局向量。

离线训练服务器通过局部注意力单元中的矩阵乘法函数 1321、双曲线正切函数 1322、线性函数 1323 和归一化指数函数 1324 依次对种子 look-alike 向量和样本 look-alike 向量，得到局部向量。

离线训练服务器对全局向量和局部向量加权求和，得到种子加权向量

步骤 1204，计算种子加权向量和样本 look-alike 向量之间的向量距离；向量距离可以采用余弦距离、Jaccard 距离和点积距离中的至少一种。

步骤 1205，根据向量距离计算第二误差损失，根据第二误差损失采用反向传播算法对用户向量提取模型进行训练。

本实施例使用逻辑斯蒂函数 Sigmoid 交叉熵函数作为损失函数；

$$L = -\frac{1}{N} \sum_{x, y \in D} (y \log p(x) + (1 - y) \log(1 - p(x)))$$

其中 D 表示训练集，x 表示输入的样本用户向量，y 为 {0, 1} 中的标签，0 代表样本用户的未点击行为，1 代表样本用户的点击行为，p(x) 是种子用户向量和样本用户向量之间通过 Sigmoid 函数预测的相似度分数。N 代表训练集中的用户数量。

图 14 示出了本申请一个示例性实施例提供的内容推荐结果的界面图。上述实施方式中的内容推荐方法可以运用在文档推荐应用中，例如新闻推荐、文章推荐等，示意性的，该文档推荐应用可以为社交软件中的新闻子程序。文档推荐应用根据目标用户历史与文档的互动情况计算该目标用户的目标用户向量和候选推荐内容，采用上述内容推荐方法向其推荐相关文档。

上述实施方式中的内容推荐方法也可以运用在视频推荐应用中，该视频

推荐应用根据目标用户历史与视频的互动情况向其推荐相关视频内容。

可选地，互动方式包括点赞、收藏、评论、转发中的至少一种。

本申请各个实施例提供的内容推荐方法，可以应用于安装有目标应用程序的终端中，该目标应用程序是具有信息推送功能的应用程序，该终端可以是智能手机、平板电脑、个人计算机或便携式计算机等等。比如，该目标应用程序是游戏、社交类应用程序、即时通信应用程序、视频播放应用程序等等，本申请实施例对此不做限定。

下面结合几个示意性的应用场景，对本申请实施例提供内容推荐方法进行说明。

10 一、文章推荐场景

在阅读类应用程序中，本申请实施例提供的内容推荐方法可以实现成为应用程序中的一个内容推送模块。应用程序根据用户的日常阅读习惯，对点击阅读文章的时间长度、点赞、评论或转发等行为进行特征提取，得到该用户的用户特征。可选地，应用程序还可以获取用户在注册时上报的相关身份信息，示意性的，身份信息可以包括性别、年龄、学历、所在城市等。

15 二、视频推送场景

在视频推送场景下，本申请实施例提供的内容推荐方法可以实现成为应用程序中的一个内容推送模块。视频应用程序根据用户的观看内容以及观看时长确定该用户对该视频的感兴趣程度。可选地，应用程序根据用户观看该视频的时间长度与视频全长的占比判断用户对该视频的感兴趣程度，例如，用户观看时长为视频全长的 100%，则认为该用户对该视频感兴趣，若观看时长小于视频全长的 50%，则认为该用户对该视频不感兴趣。

当然，上述仅以几种可能的应用场景为例进行示意性说明，本申请实施例提供的方法还可以应用于其他需要内容推荐的应用场景，本申请实施例并不对具体应用场景进行限定。

在一个具体的例子中，上述方法可以应用于腾讯公司出品的即时通信应用“微信”中，微信是中国最受欢迎的即时通讯应用，拥有超过 14 亿的用户。如图 15 所示，用户的使用方法包括以下步骤：

步骤 207，第一终端显示功能入口界面，功能入口界面上显示有信息展示界面的功能入口控件，功能入口控件上显示有新信息通知提醒；

由于应用程序可以通过多种应用功能，因此在初始状态下第一终端可以先显示该应用程序的功能入口界面，该功能入口界面上显示有信息展示界面的功能入口控件。可选地，该功能入口界面上还显示有其它功能界面的功能入口控件。

在如图 16 所示的示意性例子中，该应用程序是一个即时通讯程序，该即时通信程序的功能入口界面 20 上显示有“朋友圈”功能入口、“扫一扫”功能入口、“摇一摇”功能入口、“看一看”功能入口和“小程序”功能入口。其中，“看

一看”功能入口是信息展示界面的功能入口控件。

步骤 208, 第一终端在接收到对信息展示界面的功能入口控件的第一触发信号时, 从功能入口界面跳转至信息展示界面进行显示。

当信息展示界面“看一看”的功能入口控件被点击后, 从功能入口界面 20 跳转至信息展示界面 30 进行显示。

在如图 16 所示的示意性例子中, 该信息展示界面 30 上包括“好看”标签页和“精选”标签页。该“好看”标签页用于显示上文中基于 RALM 推荐的信息, 而“精选”标签页用于显示基于其它推荐算法推荐的推荐信息。在信息展示界面 30 刚打开时, 可以默认显示“好看”标签页中的互动推荐信息。

10 在另一个示意性的例子中, 信息展示界面包括: 第一标签页和第二标签页, 第一标签页和第二标签页中的任意一个标签页用于显示互动推荐信息, 比如第二标签页用于显示感兴趣的信息, 第一标签页用于显示长尾信息和热点新闻信息。可选地, 第一终端从功能入口界面先跳转至信息展示界面的第一标签页进行显示, 该第一标签页上还显示有第二标签页的标签。当接收到
15 第二标签页的标签对应的触发信号时, 在信息展示界面中将第一标签页切换为第二标签页进行显示。

20 深度学习 (例如深度神经网络(DNNs)和递归神经网络(RNNs)) 等模型在推荐任务上越来越受到人们的青睐。这些基于深度学习的方法有效的给出了用户偏好、项目特征以及用户与项目之间的非线性关系, 与传统的推荐算法相比, 深度学习算法在大多数情况下都有较好的推荐性能。

25 在相关技术中, 许多具有竞争力的长尾内容, 包括推送的新颖内容和最新的新闻, 通常因为缺乏推荐模型必需的行为特征难以达到广泛和及时的曝光。本实施例称之为推荐系统中的马太效应, 导致推荐内容质量低下, 多样性差。除了性能外, 提高推荐结果的质量和多样性已经成为许多推荐系统面临的共同问题。一般情况下, 损失和曲线下与坐标轴围成的面积 (Area Under Curve, AUC) 是排序模型评价中常用的度量标准。然而, 竞争模型定义了不同的损失函数, 使得比较损失变得困难。而且, AUC 与样本中的正负比例是相关的。本实施例引入了一种名为 K 精度的度量, 标记为 prec@K , 表示前 K 个结果将成为用户的实际阅读内容。方程如下:

$$\text{prec@K} = \frac{1}{N} \sum_i^N \frac{\text{size}(R_{iK} \cap S_i)}{\min(K, \text{size}(S_i))}$$

30 其中, K 表示推荐结果的数量, R_{iK} 表示向用户 i 推荐的目标内容中的前 K 项, S 表示用户已阅读的内容, N 表示测试集中的用户数, 有些用户的阅读量可能少于 K 项, 导致不规律的低 prec@K , 所以本实施例在 K 与 S_i 大小之间选取了一个最小值。在本实验中, 本实施例比较了不同模型中的 AUC 和

- 22 -

prec@K, $K \in \{10, 50\}$, 比较结果如表一所示

表一

模型	AUC	prec@10	prec@50
逻辑回归模型	0.5252	0.0811	0.0729
雅虎相似人群扩散型	0.5512	0.1023	0.0941
Youtube DNN	0.5903	0.1217	0.1087
含均值池化层的 RALM	0.5842	0.1108	0.0980
RALM	0.6012	0.1295	0.1099

本实施例可以看到逻辑回归模型的性能不佳, 验证了深度模型的优势。另外, 这也是由于它在获取簇信息能力的不足。雅虎公司提供的相似人群扩散模型表现比逻辑回归模型更好, 但比深度模型弱。这是因为内容数量较大, 并以一个非迭代的特征选择作为过滤器。Youtube DNN 比雅虎相似人群扩散模型获得了 0.0391 的绝对增益, 证明了深度模型的有效性。采用含均值池化层的 RALM 比雅虎相似模型获得了 0.033 的绝对 AUC 增益, 但比 Youtube DNN 弱。这意味着简单地平均用户嵌入向量并不像端到端的深度模型那样有效。带有注意力单元的 RALM 在所有模型中表现最好, 它比 Youtube DNN 获得了 0.0109 的绝对增益和 0.0078/0.0022 prec@10/prec@50 的增益。这一改进证明了注意力机制有助于提取种子用户的信息, 发现种子用户与目标用户之间的局部关系。

如图 17 所示, 注意力合并层在 AUC 中的性能要好于级联层, 在测试集中的损失也要好得多, 这要归功于注意力单元的设计。当各个域连在一起并向前传送时, 神经元的相同部分就会被激活给所有用户。当涉及到注意力合并层时, 不同的神经元将被激活到不同的用户面前, 这意味着学习到特征之间的个性化交互, 这对模型有很大的帮助。在 look-alike 学习中, 聚类数 k 是 K-means 聚类中的一个关键参数。

本实施例进行了一个实验来观察价值对性能的影响。在这个实验中, k 被设为 5, 10, 20, 50, 100, 每一度量是平均从测试集经过 5 期的训练。

结合图 18, 其中, K 表示推荐结果的数量, 不同 k 值的 RALM 的性能比较图。可以看出, 在 k=20 之后, AUC 和 prec@K 是稳定的。随着 k 的增加, 模型的性能得到了改善。结果表明, 聚类越多, 种子丢失信息越少。另一方面, 更大的 k 意味着更多的计算量。在图 18 中, 本实施例可以看到曲线中的弯头点, 其中 k=20 之后度量值几乎没有变高, 因此本实施例的在线模型中的 K 设置为 20。

为了验证 RALM 给本实施例的推荐系统带来的好处, 本实施例在实验环境中进行了 A/B 在线测试。本实施例按用户划分在线流量。对照组和实验组的 ID 和排列比例相同。为了在线评估性能, 涉及以下指标:

曝光: 曝光是指用户读取一个条目, 并且该条目只计算一次。随着观众

- 23 -

通过 look-alike 模式扩展，对于给定的项目，应该接触到更多的用户，并且曝光量预计会增加。

点击率(Click-through Rate, CTR): 随着受众的增加，许多新用户与种子用户分享相同的兴趣。因此，CTR 预计不会降低。

5 分类标签的多样性: 本实施例的目的之一是丰富用户对本实施例系统的兴趣，因此本实施例定义了一个名为多样性的度量。它由用户在一天中阅读的内容、类别或标记数表示。随着用户表示的更加全面，候选内容的种类将越来越多，类别和标签的多样性也会增加。

10 基尼系数: RALM 旨在缓解马太效应，因此本实施例使用基尼系数来度量推荐系统中所有候选内容的点击量分布。较高的基尼系数表明，该系统消耗的长尾物质较多，且具有较好的分配能力。

表二

项目	变化
曝光	+9.112%
点击率	+1.09%
类别的多样性	+8.435%
标签的多样性	+15.938%
基尼系数	+5.36%

15 从表二中本实施例可以看到，由于观众的扩大，曝光率得到了很大的提升。同时点击率也略有增加，这表明扩展用户对这些内容非常感兴趣，此外，将注意力合并层应用于用户表示中，挖掘出影响用户兴趣的更多隐藏因素。因此，观众可以通过更多方面的兴趣得到扩展，分类和标签的多样性也显著增加。总的来说，RALM 为种子提供了高质量和多样化的扩展。所有推荐的内容都可以到达合适的目标用户。相应地，基尼系数达到 5.36% 的增益。

20 以下为本申请的装置实施例，对于装置实施例中未详细阐述的细节，可参考上述对应的方法实施例。

图 19 示出了本申请一个示例性实施例提供的内容推荐装置的框图。该装置可以通过处理器、存储器以及存储器中的程序来实现。该装置包括：

获取模块 701，用于获取目标用户的目标用户向量；

25 候选内容确定模块 702，用于根据目标用户向量确定 n 组种子用户向量，n 组种子用户向量与 n 条候选推荐内容对应；所述 n 为正整数；

计算模块 703，用于调用相似人群扩散模型计算目标用户向量和每组种子用户向量的相似度，相似人群扩散模型用于基于注意力机制计算用户向量之间的相似度分数；

30 目标内容确定模块 704，用于根据所述 n 组种子用户向量各自的相似度，确定所述 n 条候选推荐内容中向目标用户推荐的目标内容。

-24-

可选的, 所述相似人群扩散模型包括: 第一全连接层、第二全连接层和注意力单元; 所述计算模块 703 用于:

调用第一全连接层对种子用户向量进行特征提取, 得到种子相似扩散向量; 以及, 调用第二全连接层对目标用户向量进行特征提取, 得到目标相似扩散向量;

调用注意力单元根据种子相似扩散向量和目标相似扩散向量计算得到种子加权向量;

用于计算种子加权向量和目标相似扩散向量之间的向量距离, 将向量距离确定为相似度。

可选的, 所述注意力单元包括: 局部注意力单元和全局注意力单元; 计算模块 703 用于:

调用局部注意力单元对目标相似扩散向量和种子相似扩散向量计算得到局部注意力向量;

调用全局注意力单元对种子相似扩散向量计算得到全局注意力向量;

对局部注意力向量和全局注意力向量进行加权计算, 得到种子加权向量。

可选的, 确定模块 702 用于:

确定目标用户的 n 条候选推荐内容;

对于 n 条候选推荐内容中的每条候选推荐内容, 确定与该条候选推荐内容对应的一组种子用户向量;

可选的, 确定模块 702 用于:

对于 n 条候选推荐内容中的每条候选推荐内容, 将与候选推荐内容产生过互动行为的 L 个用户的用户向量, 作为与该条候选推荐内容对应的一组种子用户向量。

可选的, 确定模块 702 用于:

对于 n 条候选推荐内容中的每条候选推荐内容, 确定与该条候选推荐内容在预设时段内产生过互动行为的 L 个用户的用户向量, 作为与该条候选推荐内容对应的一组种子用户向量。

可选的, 本申请实施例中的内容推荐装置还包括: 聚类模块 705。

聚类模块 705, 用于对于 n 组种子用户向量中的第 i 组种子用户向量, 将第 i 组种子用户向量中的 L 个种子用户向量聚类为 K 个种子用户向量, K 小于 L , i 小于或等于 n 。

可选的, 聚类模块 705 用于:

将 L 个种子用户向量采用目标聚类算法聚类为 K 个集合;

对于 K 个集合中的每个集合, 将该集合中的至少一个种子用户向量进行平均, 得到与该集合对应的种子用户向量; 或者, 对于 K 个集合中的每个集合, 将该集合中位于聚类质心的种子用户向量, 确定为与该集合对应的种子用户向量

-25-

可选的, 获取模块 701 用于:

调用用户向量提取模型中的嵌入层对所述目标用户在至少一个域上的用户特征进行特征提取, 得到每个域对应的特征向量, 域是按照特征属性进行划分的逻辑范围;

5 调用用户向量提取模型中的合并层计算每个域的域权重, 基于所述各个域的域权重将各个域各自对应的特征向量进行加权融合, 得到加权融合向量, 合并层是基于注意力机制的神经网络层;

调用用户向量提取模型中的全连接层对加权融合向量进行特征提取, 得到目标用户的目标用户向量。

10

图 20 示出了本申请一个示例性实施例提供的相似人群扩散模型的训练装置的框图。该装置可以通过处理器、存储器以及存储器中的程序来实现。该装置包括:

15 获取模块 711, 用于获取第二训练样本, 第二训练样本包括: 样本用户向量和样本内容对应的种子用户向量;

调用模块 712, 用于调用第一全连接层对种子用户向量进行特征提取, 得到种子相似扩散向量; 以及, 调用第二全连接层对样本用户向量进行特征提取, 得到样本相似扩散向量;

20 调用模块 712, 还用于调用注意力单元根据种子相似扩散向量和样本相似扩散向量计算得到种子加权向量;

计算模块 713, 用于计算种子加权向量和样本相似扩散向量之间的向量距离;

训练模块 714, 用于根据向量距离计算第二误差损失, 根据第二误差损失采用反向传播算法对相似人群扩散模型进行训练。

25

图 21 示出了本申请一个示例性实施例提供的用户向量提取模型的训练装置, 装置包括:

获取模块 721, 用于获取第一训练样本, 第一训练样本包括多个样本用户的用户特征和对样本内容的互动记录;

30 输入模块 722, 用于将第一样本用户的用户特征输入用户向量提取模型, 得到第一用户向量; 第一样本用户是第一训练样本内多个样本用户中的任意一个;

计算模块 723, 用于计算第一用户向量与样本内容的嵌入向量的点积;

35 计算模块 723, 还用于根据点积和第一样本用户对样本内容的互动记录, 计算得到第一误差损失;

计算模块 723, 还用于根据第一误差损失采用反向传播算法对用户向量提取模型进行训练。

本申请实施例还提供了一种计算机存储介质，该计算机可读存储介质存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或指令集由所述处理器加载并执行以实现如上所述的内容推荐方法、用户向量提取模型的训练方法或者相似人群扩散模型的训练方法。

本申请实施例还提供了一种计算机程序产品，该计算机程序产品存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或指令集由所述处理器加载并执行以实现如上所述的内容推荐方法、用户向量提取模型的训练方法或者相似人群扩散模型的训练方法。

图 22 示出了本申请一个实施例提供的服务器的结构示意图。该服务器用于实施上述实施例中提供的内容推荐方法、用户向量提取模型的训练方法或者相似人群扩散模型的训练方法。具体来讲：

服务器 800 包括中央处理单元 (CPU) 801、包括随机存取存储器 (RAM) 802 和只读存储器 (ROM) 803 的系统存储器 804，以及连接系统存储器 804 和中央处理单元 801 的系统总线 805。服务器 800 还包括帮助计算机内的各个器件之间传输信息的基本输入/输出系统 (I/O 系统) 806，和用于存储操作系统 813、应用程序 814 和其他程序模块 815 的大容量存储设备 807。

基本输入/输出系统 806 包括有用于显示信息的显示器 808 和用于用户输入信息的诸如鼠标、键盘之类的输入设备 809。其中显示器 808 和输入设备 809 都通过连接到系统总线 805 的输入输出控制器 810 连接到中央处理单元 801。基本输入/输出系统 806 还可以包括输入输出控制器 810 以用于接收和处理来自键盘、鼠标、或电子触控笔等多个其他设备的输入。类似地，输入输出控制器 810 还提供输出到显示屏、打印机或其他类型的输出设备。

大容量存储设备 807 通过连接到系统总线 805 的大容量存储控制器 (未示出) 连接到中央处理单元 801。大容量存储设备 807 及其相关联的计算机可读介质为服务器 800 提供非易失性存储。也就是说，大容量存储设备 807 可以包括诸如硬盘或者 CD-ROM 驱动器之类的计算机可读介质 (未示出)。

不失一般性，计算机可读介质可以包括计算机存储介质和通信介质。计算机存储介质包括以用于存储诸如计算机可读指令、数据结构、程序模块或其他数据等信息的任何方法或技术实现的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括 RAM、ROM、EPROM、EEPROM、闪存或其他固态存储其技术，CD-ROM、DVD 或其他光学存储、磁带盒、磁带、磁盘存储或其他磁性存储设备。当然，本领域技术人员可知计算机存储介质不局限于上述几种。上述的系统存储器 804 和大容量存储设备 807 可以统称为存储器。

根据本申请的各种实施例，服务器 800 还可以通过诸如因特网等网络连接到网络上的远程计算机运行。也即服务器 800 可以通过连接在系统总线 805 上的网络接口单元 811 连接到网络 812，或者说，也可以使用网络接口单元 811 来连接到其他类型的网络或远程计算机系统（未示出）。

5 本领域技术人员在考虑说明书及实践这里公开的发明后，将容易想到本申请的其它实施方案。本申请旨在涵盖本申请的任何变型、用途或者适应性变化，这些变型、用途或者适应性变化遵循本申请的一般性原理并包括本申请未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性地，本申请的真正范围和精神由下面的权利要求指出。

10 应当理解的是，本申请并不局限于上面已经描述并在附图中示出的精确结构，并且可以在不脱离其范围进行各种修改和改变。本申请的范围仅由所附的权利要求来限制。

应当理解的是，在本文中提及的“多个”是指两个或两个以上。“和/或”，描述关联对象的关联关系，表示可以存在三种关系，例如，A 和/或 B，可以表示：单独存在 A，同时存在 A 和 B，单独存在 B 这三种情况。字符“/”一般表示前后关联对象是一种“或”的关系。

15 本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成，也可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，上述提到的存储介质可以是只读存储器，
20 磁盘或光盘等。

以上所述仅为本申请的较佳实施例，并不用以限制本申请，凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

权 利 要 求

1、一种内容推荐方法，由服务器执行，所述方法包括：

获取目标用户的目标用户向量；

5 根据所述目标用户向量确定 n 组种子用户向量，所述 n 组种子用户向量与 n 条候选推荐内容对应；所述 n 为正整数；

调用相似人群扩散模型计算所述目标用户向量和每组所述种子用户向量的相似度，所述相似人群扩散模型用于基于注意力机制计算用户向量之间的相似度；

10 根据所述 n 组种子用户向量各自的所述相似度，确定所述 n 条候选推荐内容中向所述目标用户推荐的目标内容。

2、根据权利要求 1 所述的方法，所述相似人群扩散模型包括：第一全连接层、第二全连接层和注意力单元；

所述调用相似人群扩散模型计算所述目标用户向量和每组所述种子用户向量的相似度，包括：

15 调用所述第一全连接层对所述种子用户向量进行特征提取，得到种子相似扩散向量；以及，调用所述第二全连接层对所述目标用户向量进行特征提取，得到目标相似扩散向量；

调用所述注意力单元根据所述种子相似扩散向量和所述目标相似扩散向量计算得到所述种子加权向量；

20 计算所述种子加权向量和所述目标相似扩散向量之间的向量距离，将所述向量距离确定为所述相似度。

3、根据权利要求 2 所述的方法，所述注意力单元包括：局部注意力单元和全局注意力单元；

25 所述调用所述注意力单元根据所述种子相似扩散向量和所述目标相似扩散向量计算得到所述种子加权向量，包括：

调用所述局部注意力单元对所述目标相似扩散向量和所述种子相似扩散向量计算得到局部注意力向量；

调用所述全局注意力单元对所述种子相似扩散向量计算得到全局注意力向量；

30 对所述局部注意力向量和所述全局注意力向量进行加权计算，得到所述种子加权向量。

4、根据权利要求 1 至 3 任一所述的方法，所述根据所述目标用户向量确定 n 组种子用户向量，包括：

确定所述目标用户的所述 n 条候选推荐内容；

35 对于所述 n 条候选推荐内容中的每条候选推荐内容，确定与该条候选推荐内容对应的一组种子用户向量。

5、根据权利要求 4 所述的方法，所述对于所述 n 条候选推荐内容中的每

条候选推荐内容,确定与该条候选推荐内容对应的一组种子用户向量,包括:

对于所述 n 条候选推荐内容中的每条候选推荐内容,确定与该条候选推荐内容在预设时段内产生过互动行为的 L 个用户的用户向量,作为与该条候选推荐内容对应的一组种子用户向量。

5 6、根据权利要求 1 至 4 任一所述的方法,所述方法还包括:

对于所述 n 组种子用户向量中的第 i 组种子用户向量,将所述第 i 组种子用户向量中的 L 个种子用户向量聚类为 K 个种子用户向量, K 小于 L , i 小于或等于 n 。

10 7、根据权利要求 6 所述的方法,所述将所述第 i 组种子用户向量中的 L 个种子用户向量聚类为 K 个种子用户向量,包括:

将所述 L 个种子用户向量采用目标聚类算法聚类为 K 个集合;

15 对于所述 K 个集合中的每个集合,将该集合中的至少一个种子用户向量进行平均,得到与该集合对应的种子用户向量;或者,对于所述 K 个集合中的每个集合,将该集合中位于聚类质心的种子用户向量,确定为与该集合对应的种子用户向量。

8、根据权利要求 1 至 3 任一所述的方法,所述获取所述目标用户的目标用户向量,包括:

20 调用用户向量提取模型中的嵌入层对所述目标用户在至少一个域上的用户特征进行特征提取,得到每个域对应的特征向量,所述域是按照特征属性进行划分的逻辑范围;

调用所述用户向量提取模型中的合并层计算所述每个域的域权重,基于所述各个域的域权重将所述各个域各自对应的所述特征向量进行加权融合,得到加权融合向量;所述合并层是基于注意力机制的神经网络层;

25 调用所述用户向量提取模型中的全连接层对所述加权融合向量进行特征提取,得到所述目标用户的目标用户向量。

9、根据权利要求 8 所述的方法,所述合并层包括:注意力合并层和多层感知器 MLP;所述注意力合并层是基于自注意力机制的神经网络层;

30 所述调用所述用户向量提取模型中的合并层计算所述每个域的域权重,基于所述各个域的域权重将所述各个域各自对应的所述特征向量进行加权融合,得到加权融合向量,包括:

调用所述注意力合并层计算所述各个域的域权重,将所述各个域各自对应的所述特征向量进行加权融合后输入至所述 MLP;

调用所述 MLP 对加权融合后的所述特征向量进行处理,得到所述加权融合向量。

35 10、根据权利要求 8 所述的方法,通过以下方式训练所述用户向量提取模型包:

获取第一训练样本,所述第一训练样本包括多个样本用户的用户特征和

对样本内容的互动记录；

将第一样本用户的用户特征输入所述用户向量提取模型，得到第一用户向量；所述第一样本用户是所述第一训练样本内多个样本用户中的任意一个；
计算所述第一用户向量与所述样本内容的嵌入向量的点积；

5 根据所述点积和所述第一样本用户对所述样本内容的互动记录，计算得到第一误差损失；

根据所述第一误差损失采用反向传播算法对所述用户向量提取模型进行训练。

10 11、根据权利要求 2 所述的方法，通过以下方式训练所述相似人群扩散模型：

获取第二训练样本，所述第二训练样本包括：样本用户向量和样本内容对应的种子用户向量；

15 调用所述第一全连接层对所述种子用户向量进行特征提取，得到种子相似扩散向量；以及，调用所述第二全连接层对所述样本用户向量进行特征提取，得到样本相似扩散向量；

调用所述注意力单元根据所述种子相似扩散向量和所述样本相似扩散向量计算得到所述种子加权向量；

计算所述种子加权向量和所述样本相似扩散向量之间的向量距离；

20 根据所述向量距离计算第二误差损失，根据所述第二误差损失采用反向传播算法对所述相似人群扩散模型进行训练。

12、一种计算机设备，其特征在于，所述计算机设备包括处理器和存储器，所述存储器中存储有计算机程序，所述计算机程序由所述处理器加载并执行以实现如权利要求 1-11 任一项所述的内容推荐方法。

25 13、一种计算机可读存储介质，其特征在于，所述存储介质中存储有计算机程序，所述计算机程序由处理器加载并执行以实现如权利要求 1-11 任一项所述的内容推荐方法。

14、一种计算机程序产品，包括指令，当其在计算机上运行时，使得计算机执行如权利要求 1-11 任一项所述的内容推荐方法。

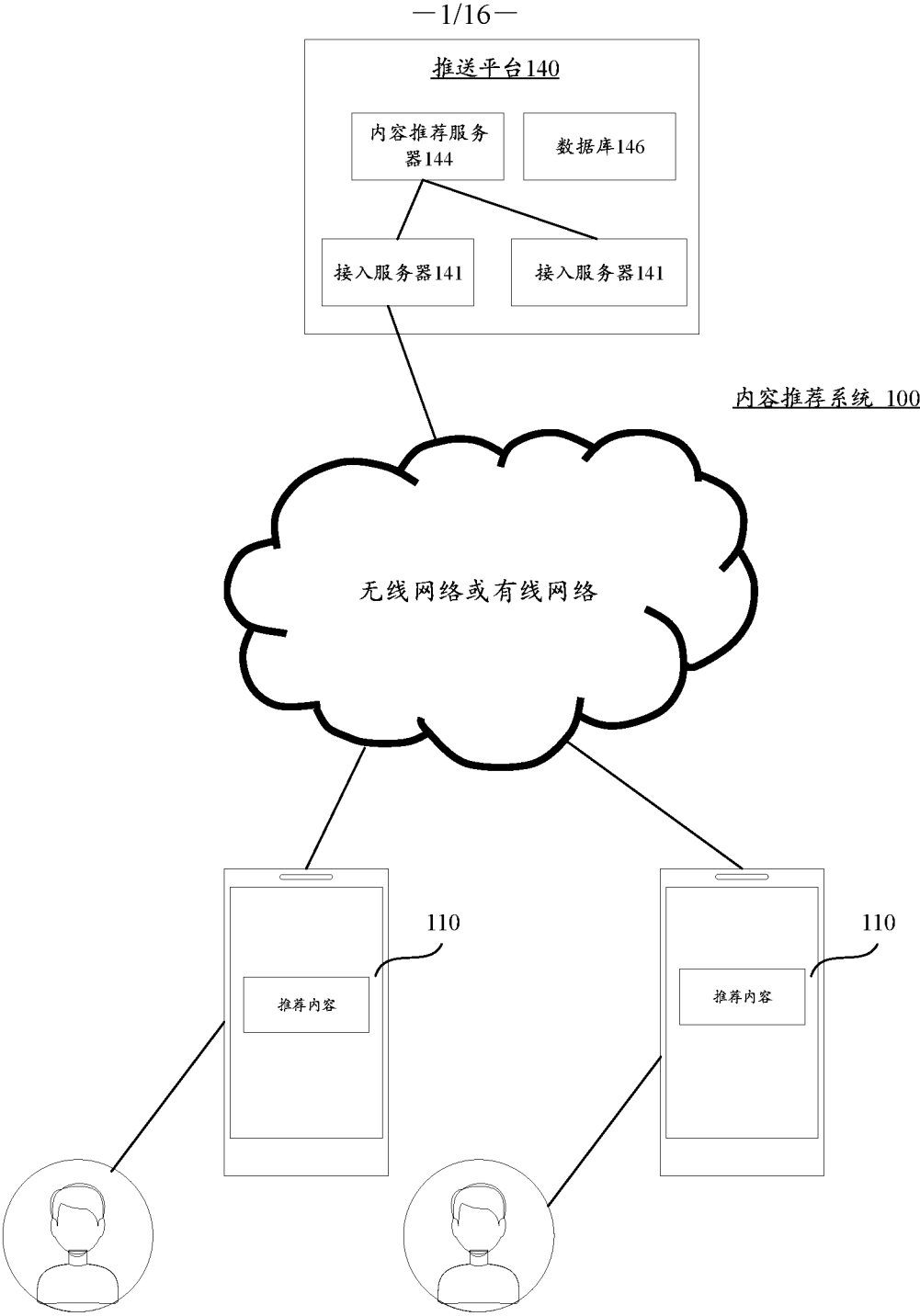


图 1

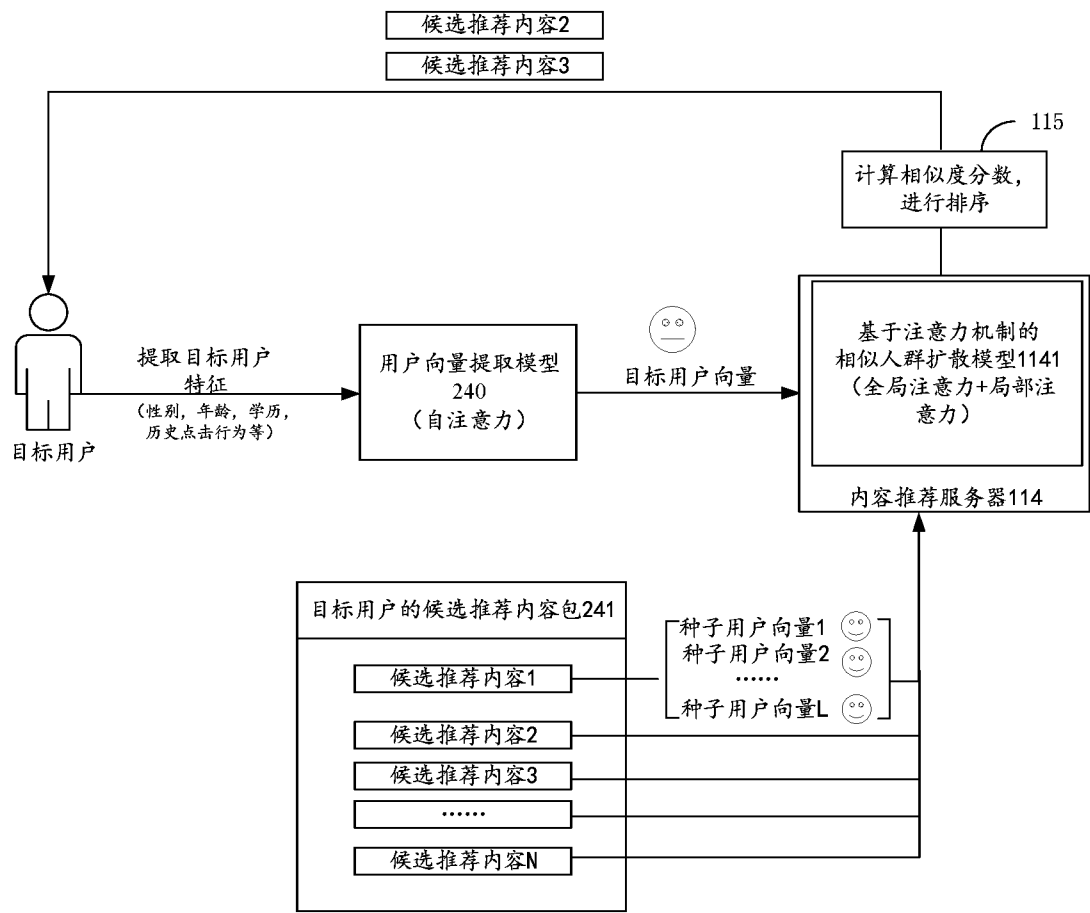


图 2

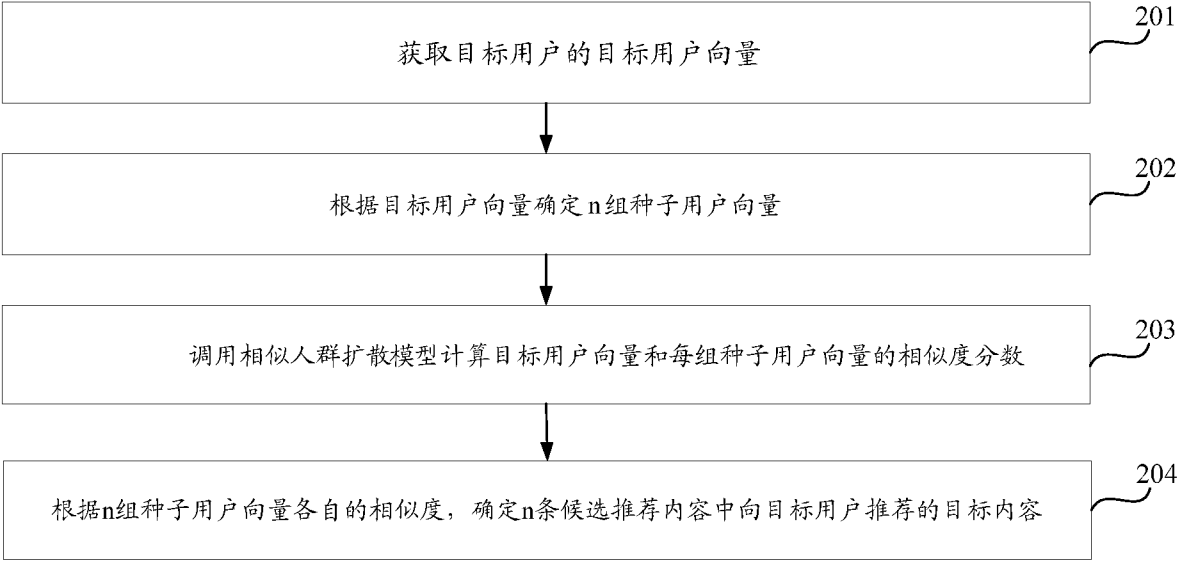


图 3

—3/16—

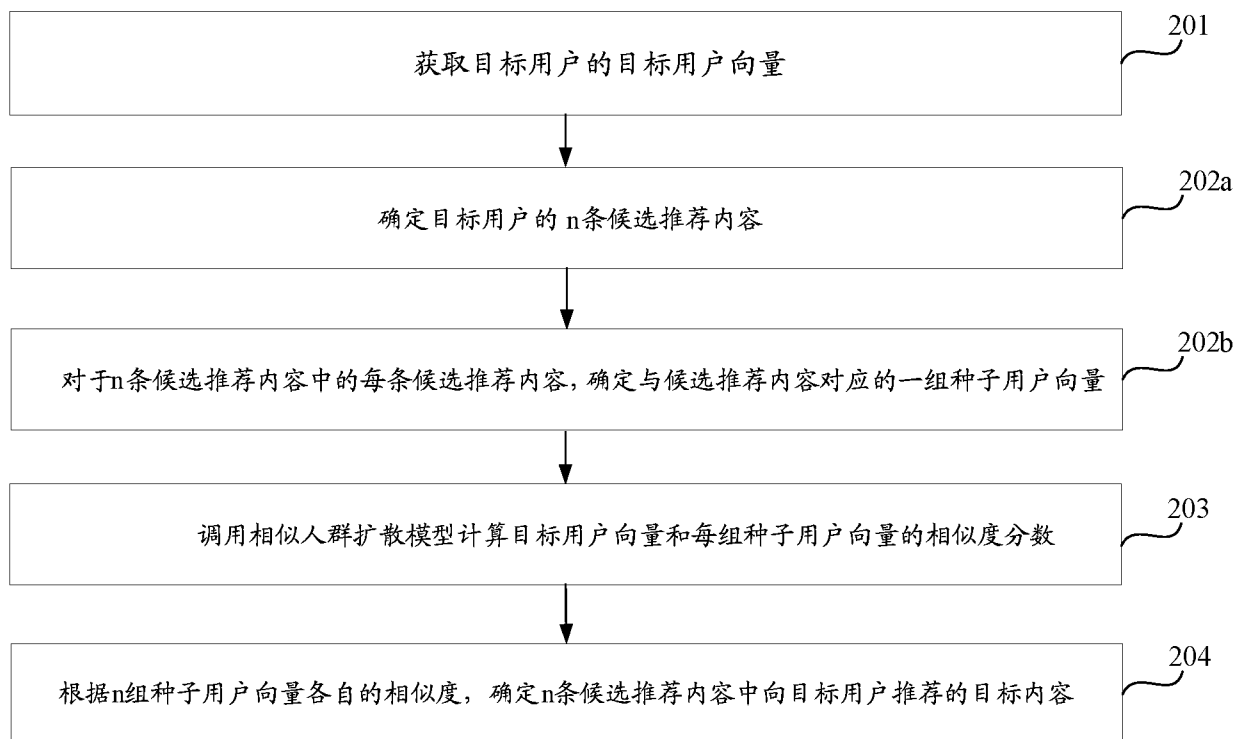


图 4

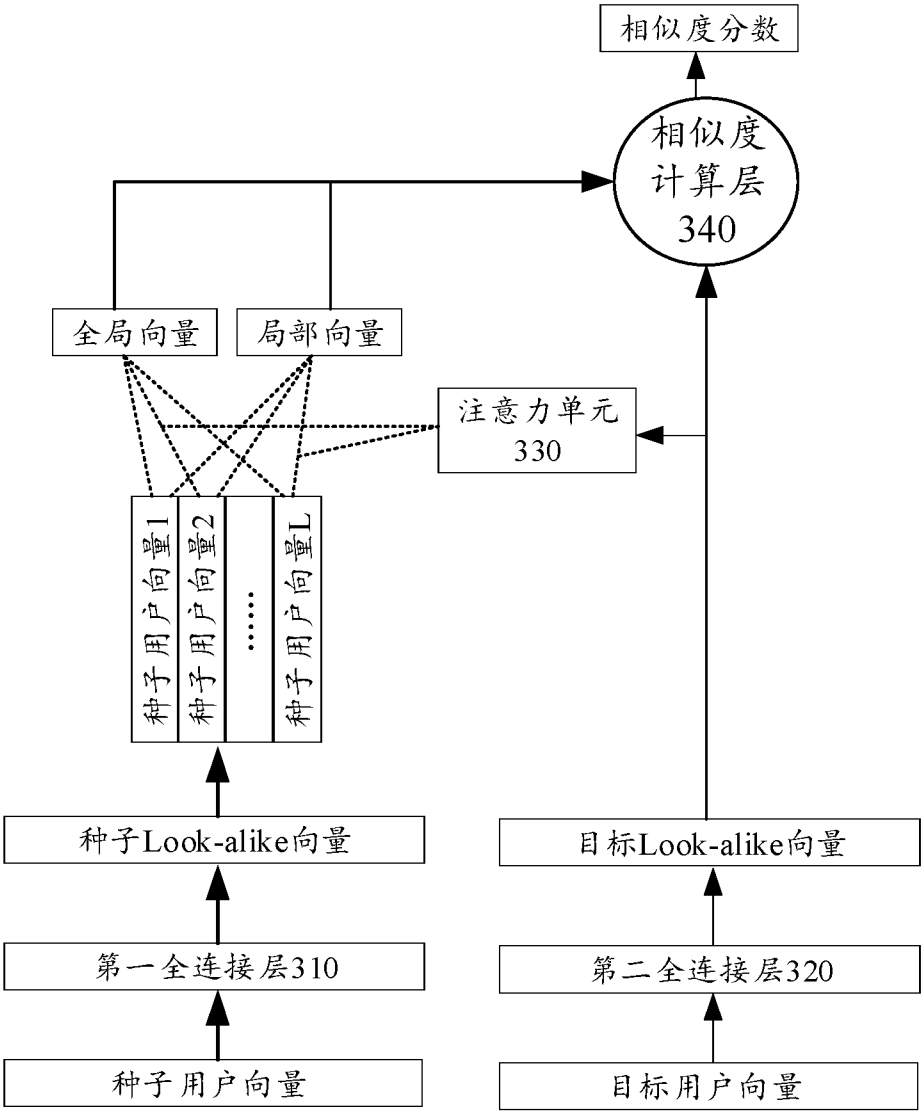


图 5

—5/16—

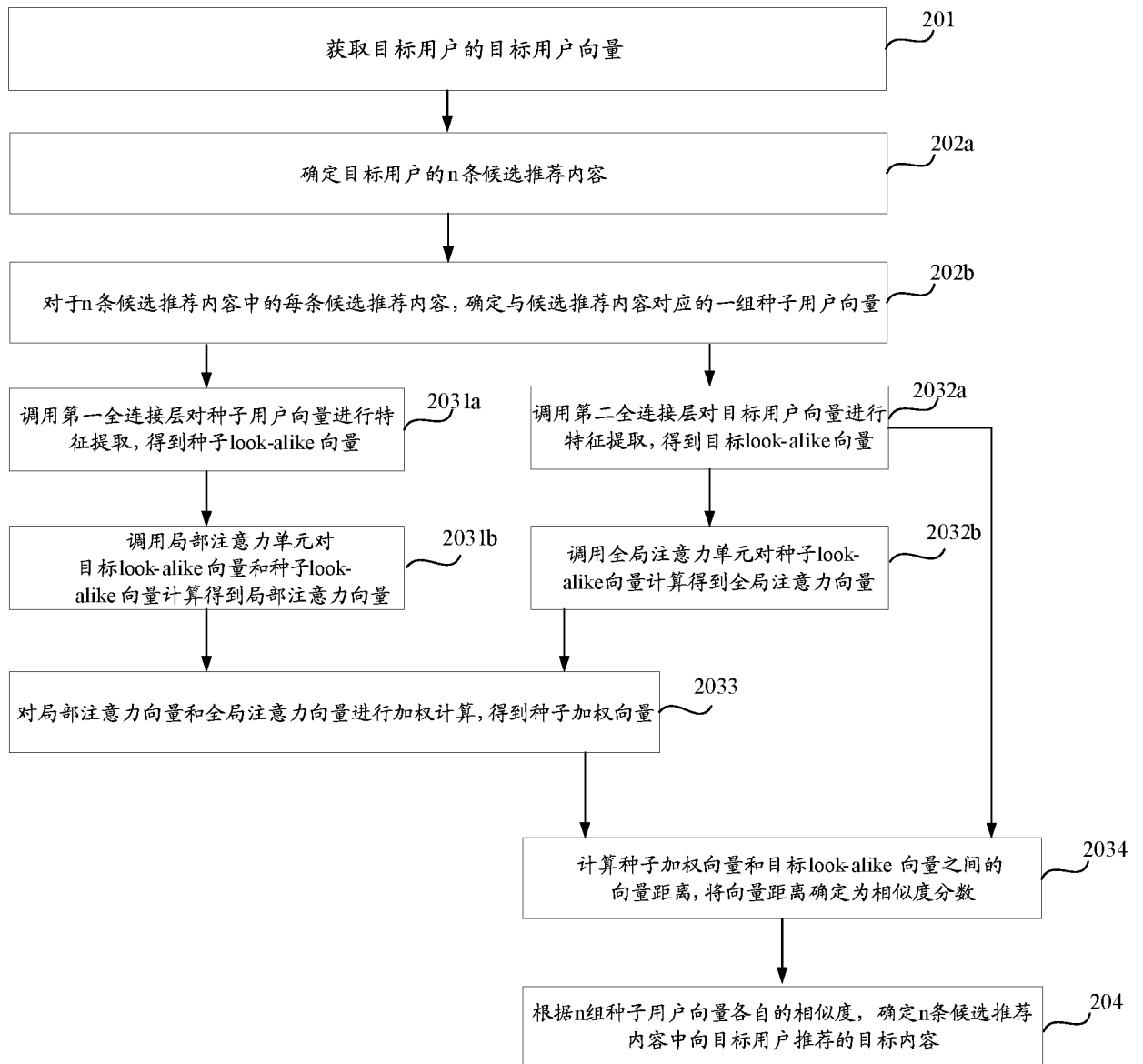


图 6

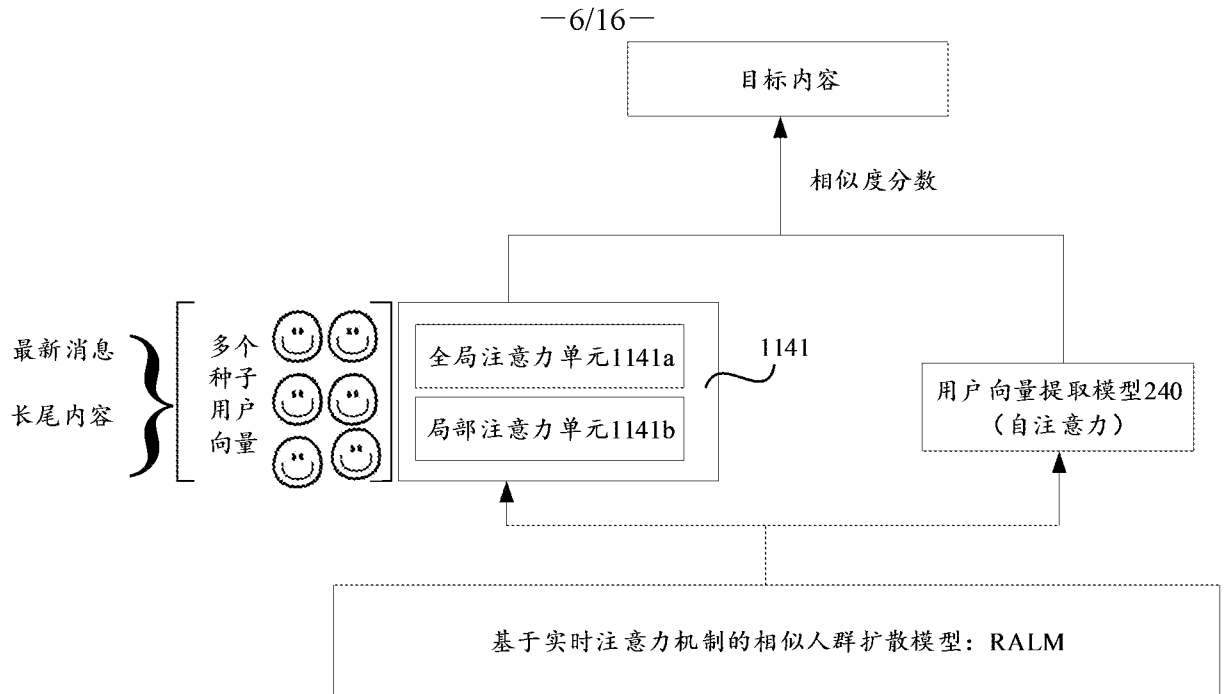


图 7

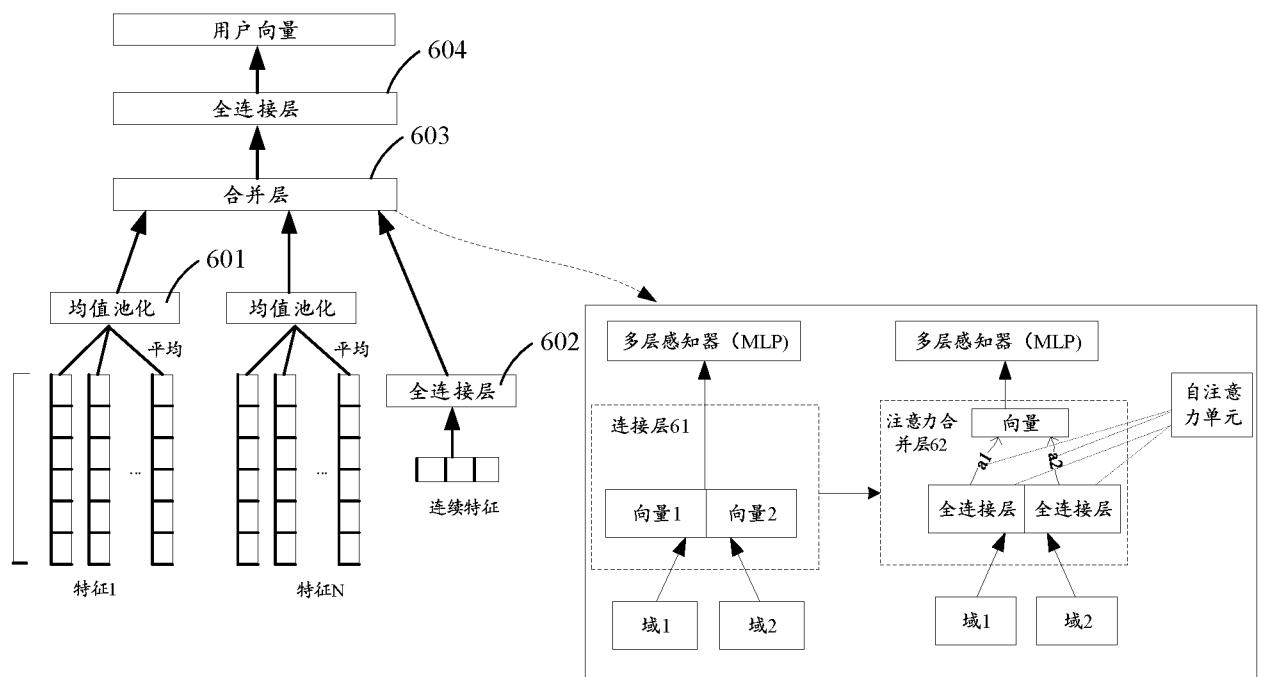


图 8

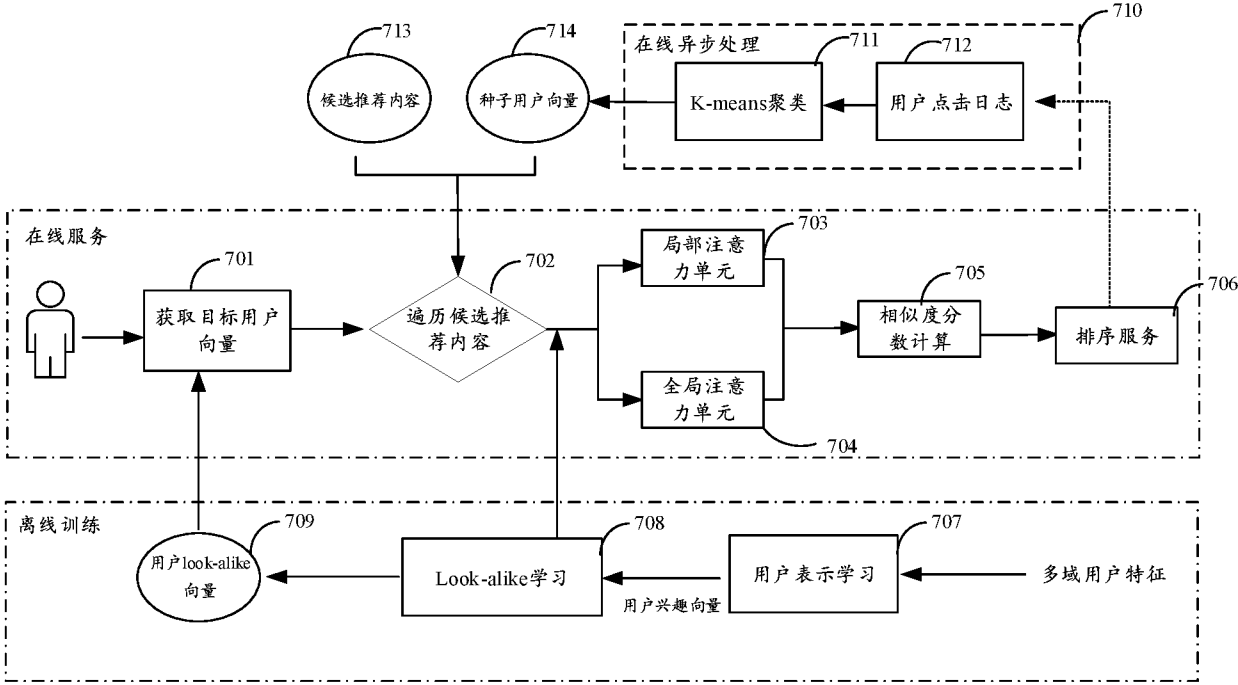


图 9

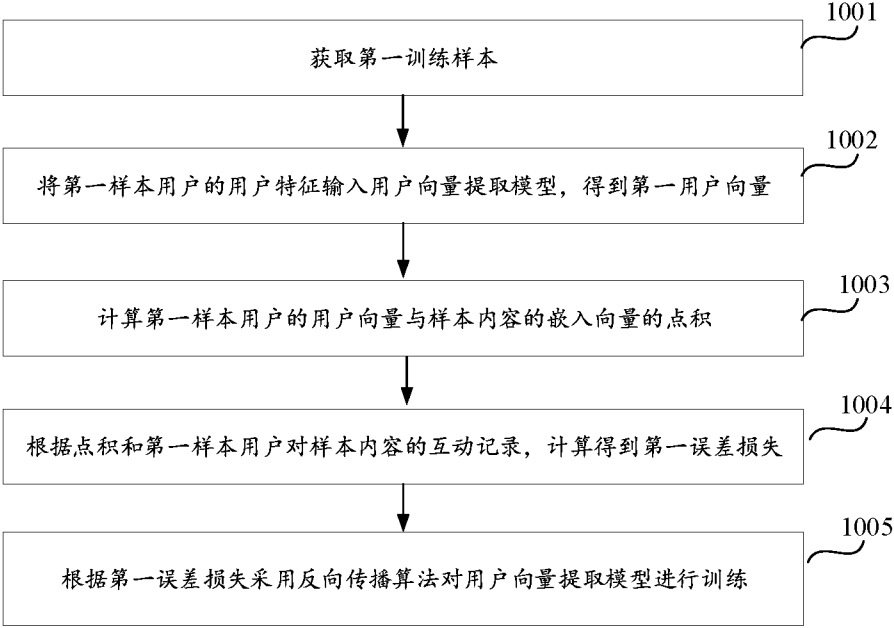


图 10

—8/16—

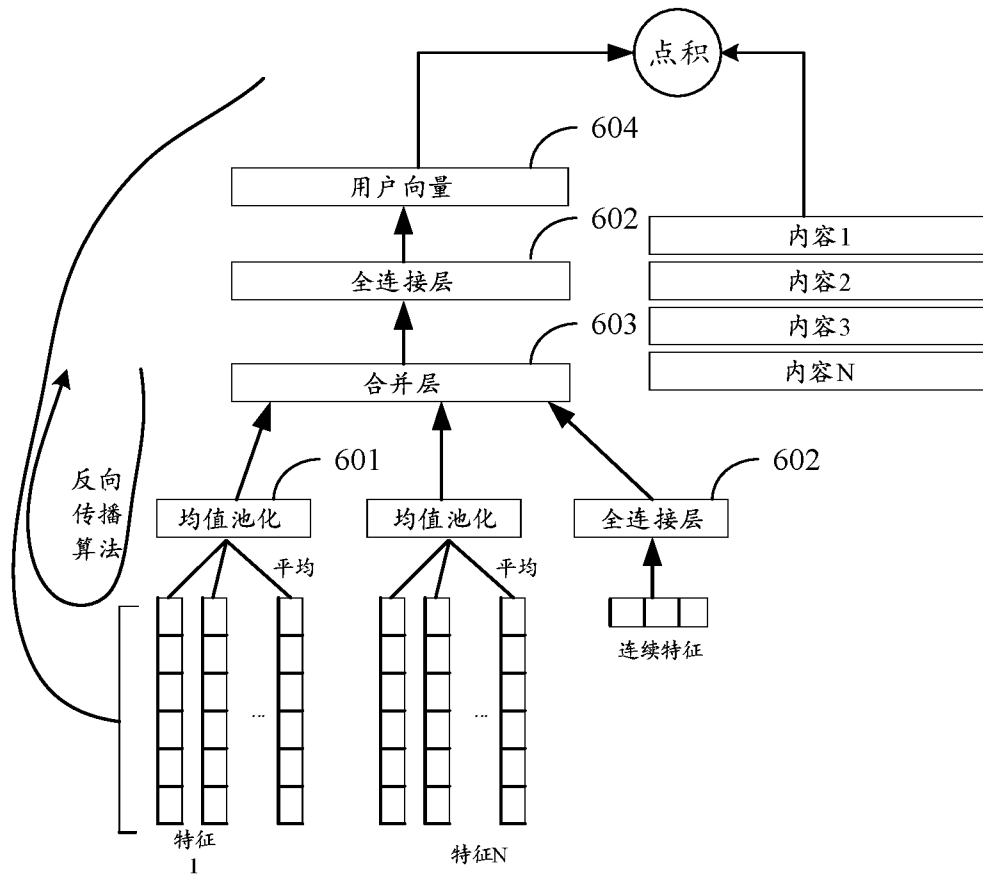


图 11

—9/16—

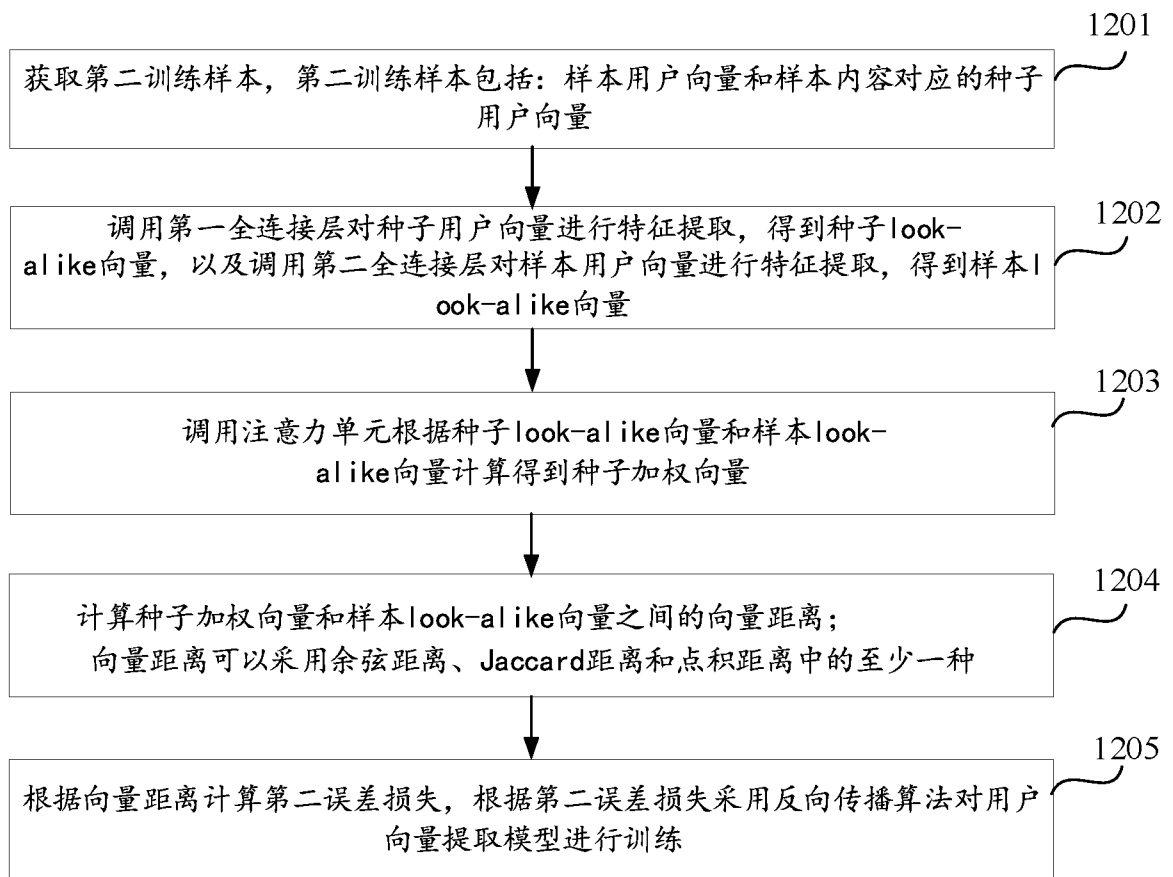


图 12

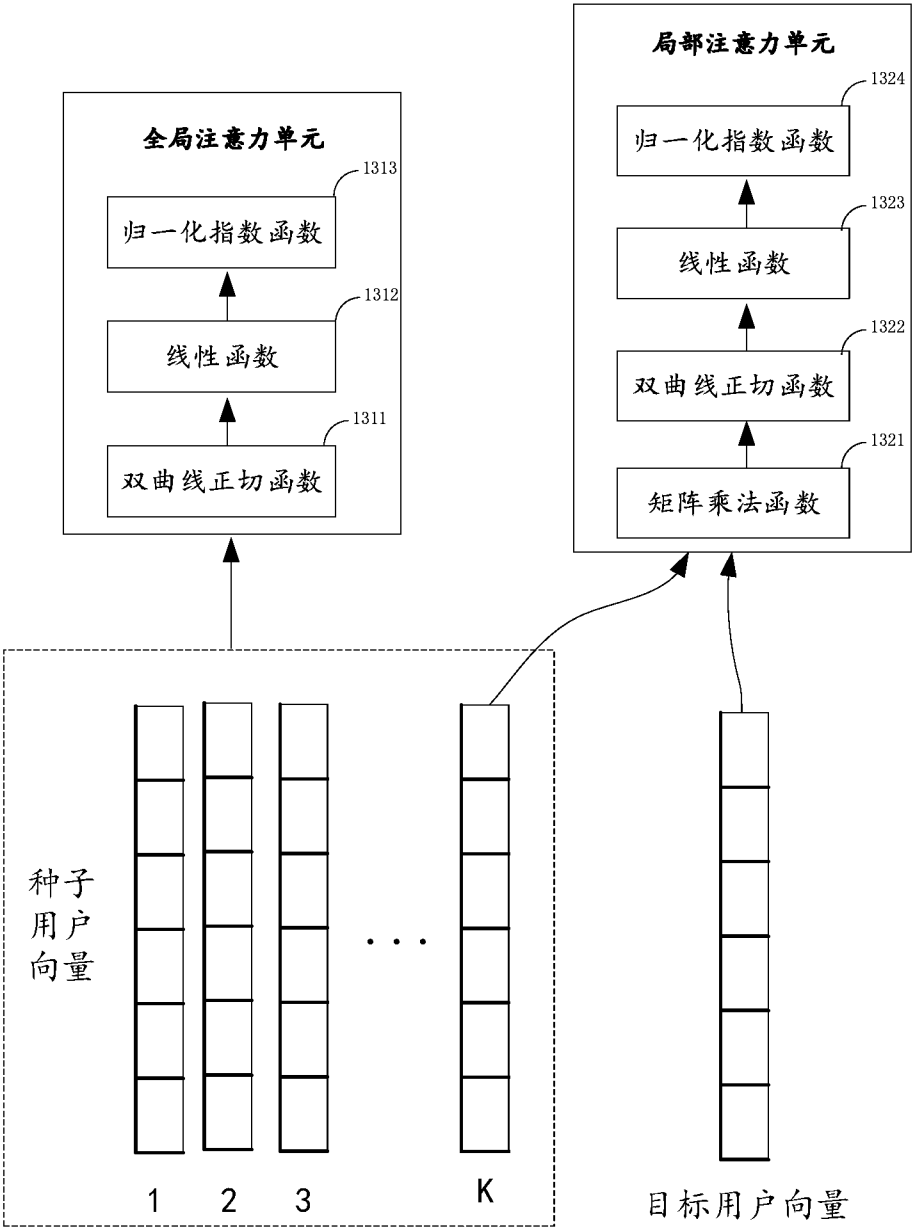


图 13

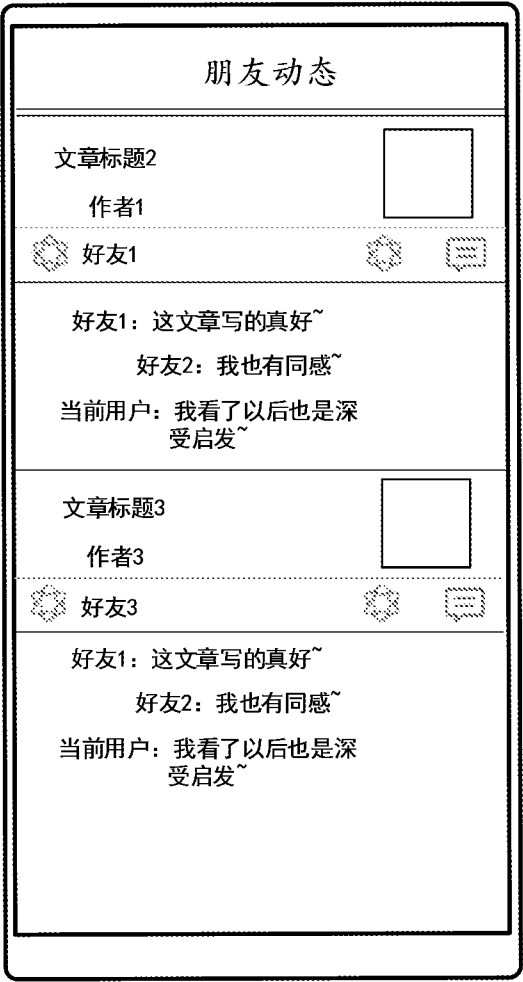


图 14

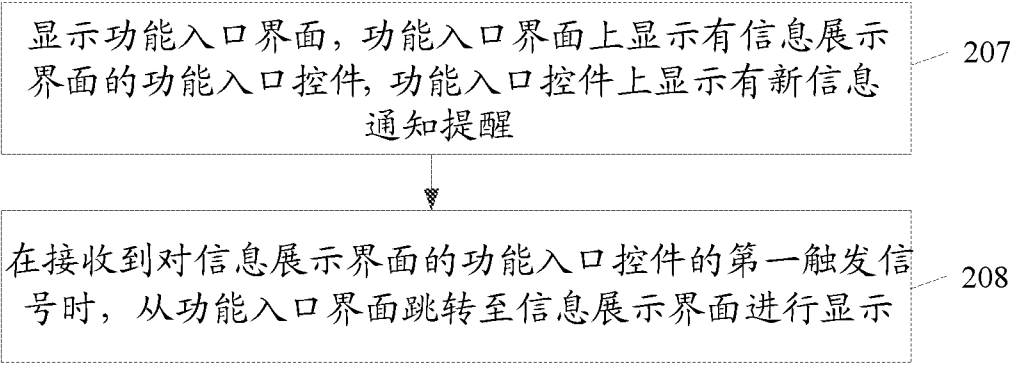


图 15

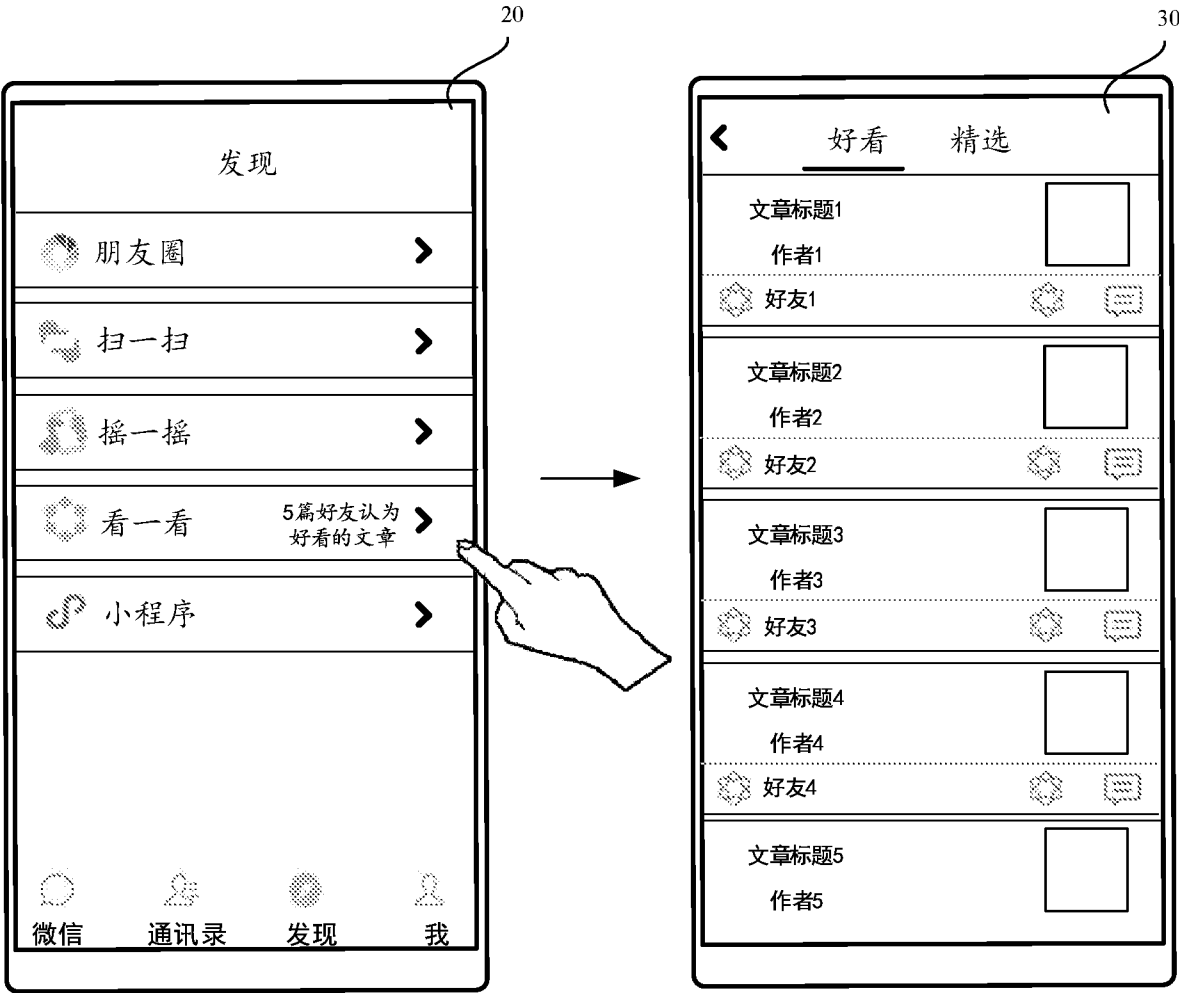


图 16

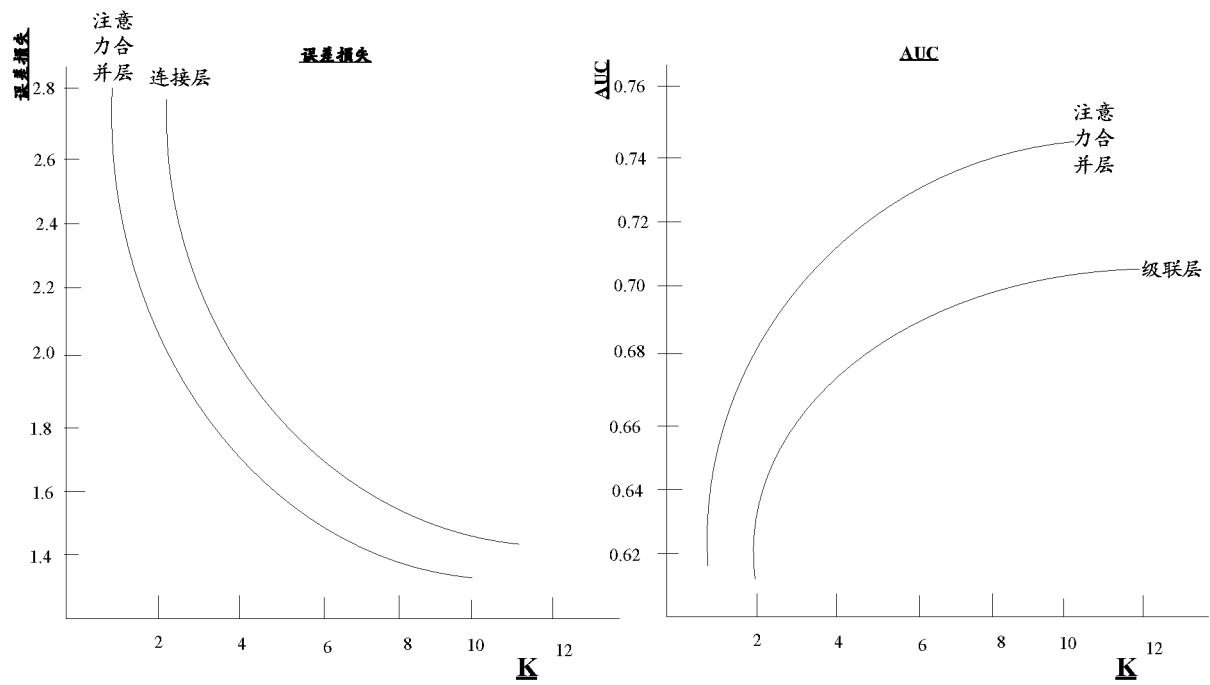


图 17

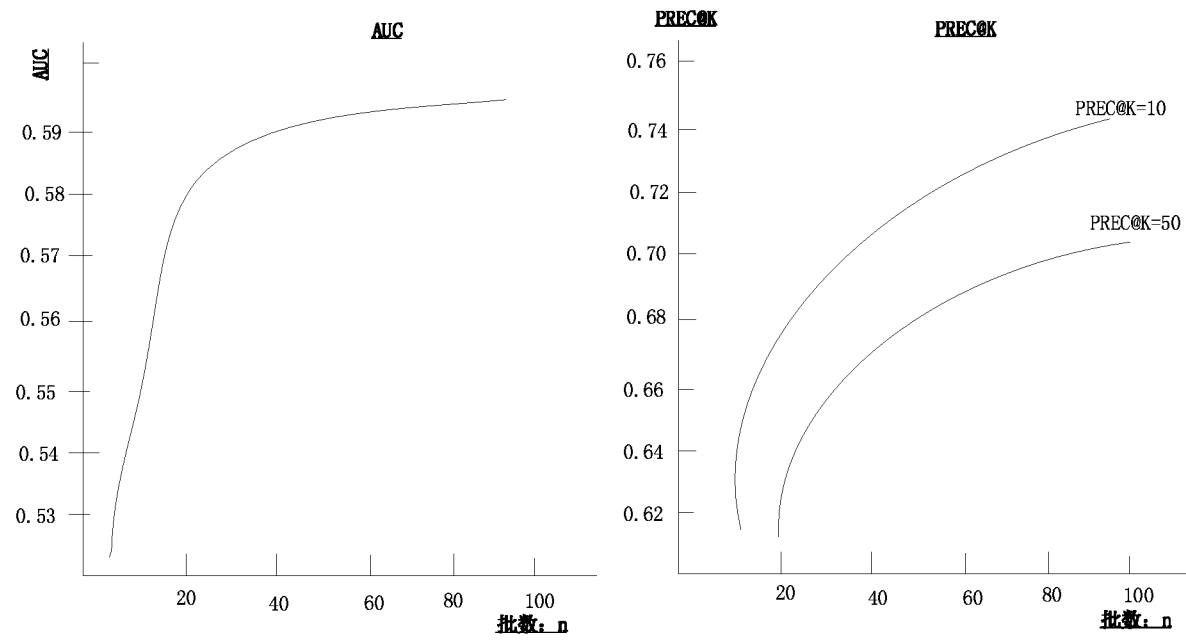


图 18

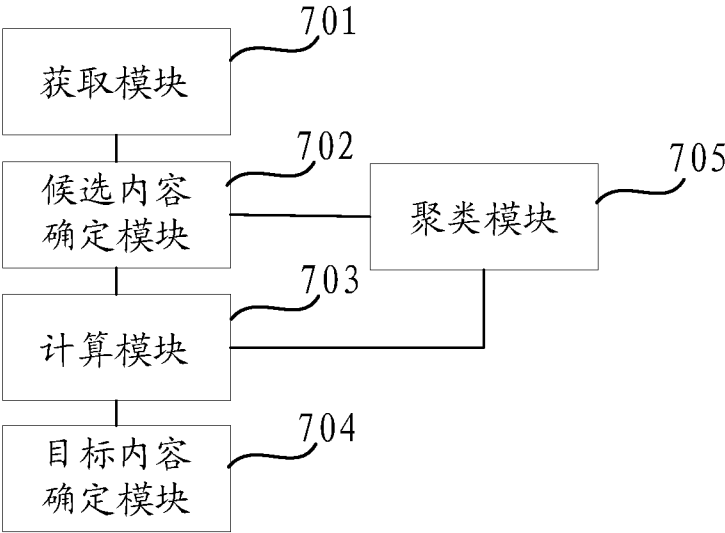


图 19

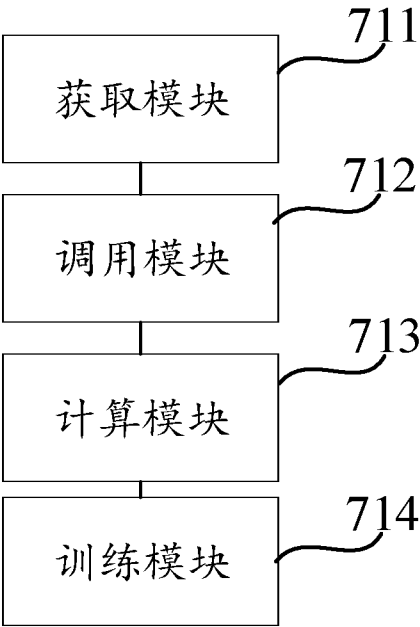


图 20

—15/16—

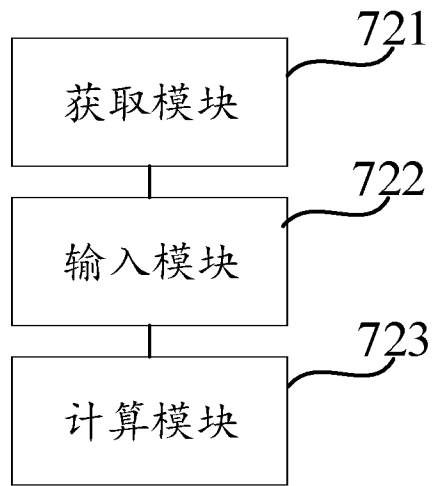


图 21

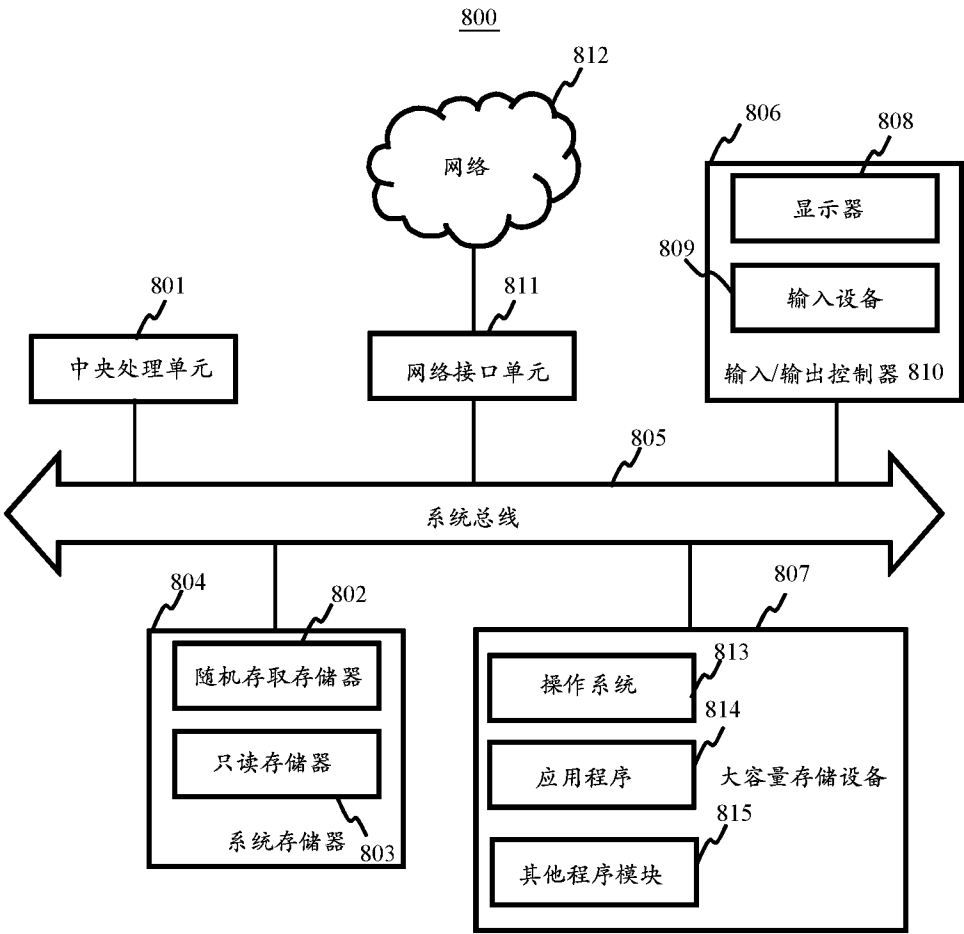


图 22

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/086946

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/9535(2019.01)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 16/-, G06F 17/-, G06N 3/-

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)

DWPI, CNABS, SIPOABS, CNTXT, USTXT, CNKI: 内容, 推荐, 目标, 用户, 向量, 候选, 备选, 相似, 相似度, 注意力, 模型, 扩散, 实时, 同步, MLP, RALM, content, context, recommendation, push, target, user, vector, candidate, alternative, similarity, attention, model, diffusion, real time, look alike, attention mechanism, CTR, click through rate, click rate, synchronization

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110929206 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 27 March 2020 (2020-03-27) description, paragraphs 53-186	1-14
X	CN 108921221 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 30 November 2018 (2018-11-30) description, paragraphs 35-147	1, 12-14
A	CN 105447730 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 30 March 2016 (2016-03-30) entire document	1-14
A	CN 106250464 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 21 December 2016 (2016-12-21) entire document	1-14
A	CN 109389168 A (GUOXIN YOUER DATA CO., LTD.) 26 February 2019 (2019-02-26) entire document	1-14
A	CN 109492157 A (HUAQIAO UNIVERSITY) 19 March 2019 (2019-03-19) entire document	1-14



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

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

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

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

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

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

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

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

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

“&” document member of the same patent family

Date of the actual completion of the international search

26 July 2020

Date of mailing of the international search report

07 August 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/086946

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2019130257 A1 (SENTIENT TECHNOLOGIES BARBADOS LTD) 02 May 2019 (2019-05-02) entire document	1-14

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2020/086946

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)		Publication date (day/month/year)
CN	110929206	A	27 March 2020	None		
CN	108921221	A	30 November 2018	WO	2020007164	A1 09 January 2020
CN	105447730	A	30 March 2016	None		
CN	106250464	A	21 December 2016	CN	106250464	B 28 April 2020
CN	109389168	A	26 February 2019	None		
CN	109492157	A	19 March 2019	None		
US	2019130257	A1	02 May 2019	None		

国际检索报告

国际申请号

PCT/CN2020/086946

A. 主题的分类 G06F 16/9535 (2019.01) 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 16/-, G06F 17/-, G06N 3/- 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)) DWPI, CNABS, SIPOABS, CNTXT, USTXT, CNKI: 内容, 推荐, 目标, 用户, 向量, 候选, 备选, 相似, 相似度, 注意力, 模型, 扩散, 实时, 同步, MLP, RALM, content, context, recommendation, push, target, user, vector, candidate, alternative, similarity, attention, model, diffusion, real time, look alike, attention mechanism, CTR, click through rate, click rate, synchronization		
C. 相关文件		
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 110929206 A (腾讯科技深圳有限公司) 2020年 3月 27日 (2020 - 03 - 27) 说明书第53-186段	1-14
X	CN 108921221 A (腾讯科技深圳有限公司) 2018年 11月 30日 (2018 - 11 - 30) 说明书第35-147段	1, 12-14
A	CN 105447730 A (腾讯科技深圳有限公司) 2016年 3月 30日 (2016 - 03 - 30) 全文	1-14
A	CN 106250464 A (北京百度网讯科技有限公司) 2016年 12月 21日 (2016 - 12 - 21) 全文	1-14
A	CN 109389168 A (国信优易数据有限公司) 2019年 2月 26日 (2019 - 02 - 26) 全文	1-14
A	CN 109492157 A (华侨大学) 2019年 3月 19日 (2019 - 03 - 19) 全文	1-14
A	US 2019130257 A1 (SENTIENT TECHNOLOGIES BARBADOS LTD) 2019年 5月 2日 (2019 - 05 - 02) 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 2020年 7月 26日		国际检索报告邮寄日期 2020年 8月 7日
ISA/CN的名称和邮寄地址 中国国家知识产权局 (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		受权官员 田竞 电话号码 62411749

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2020/086946

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	110929206	A	2020年 3月 27日	无	
CN	108921221	A	2018年 11月 30日	WO 2020007164 A1	2020年 1月 9日
CN	105447730	A	2016年 3月 30日	无	
CN	106250464	A	2016年 12月 21日	CN 106250464 B	2020年 4月 28日
CN	109389168	A	2019年 2月 26日	无	
CN	109492157	A	2019年 3月 19日	无	
US	2019130257	A1	2019年 5月 2日	无	