

(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
2021 年 2 月 18 日 (18.02.2021)



(10) 国际公布号
WO 2021/027149 A1

(51) 国际专利分类号:
G06F 16/9536 (2019.01)

(21) 国际申请号: PCT/CN2019/117794

(22) 国际申请日: 2019 年 11 月 13 日 (13.11.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910748591.3 2019 年 8 月 14 日 (14.08.2019) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518033 (CN)。

(72) 发明人: 刘利(LIU, Li); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518033 (CN)。

(74) 代理人: 北京鸿元知识产权代理有限公司(GRANDER IP LAW FIRM); 中国北京市朝阳区光华路 7 号汉威大厦东区 18A6 室, Beijing 100004 (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,

(54) Title: PORTRAIT SIMILARITY-BASED INFORMATION RETRIEVAL RECOMMENDATION METHOD AND DEVICE AND STORAGE MEDIUM

(54) 发明名称: 基于画像相似性的信息检索推荐方法、装置及存储介质

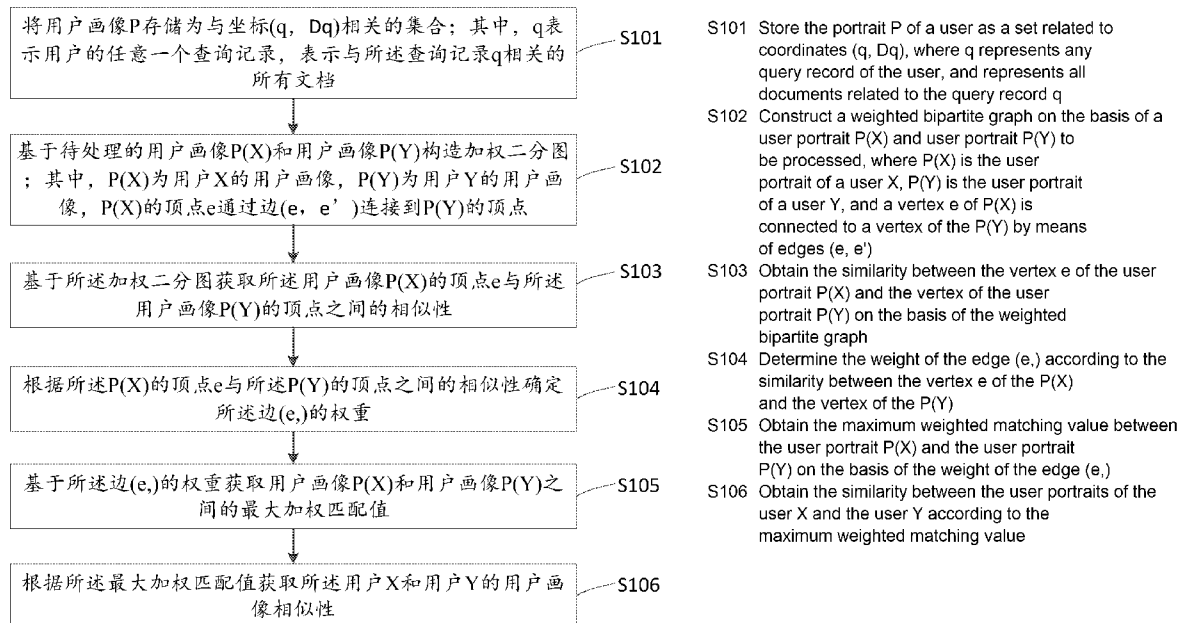


图 4

(57) Abstract: A portrait similarity-based information retrieval recommendation method, a device, a system and a storage medium, the method comprising: obtaining user portraits of different users, and determining the user portrait similarity between the user portraits (S11); creating a dynamic user community on the basis of the user portrait similarity, and enabling users having similar portraits to be grouped into the same dynamic user community (S12); and performing information retrieval recommendation on the user according to the dynamic user community and a query statement of the user (S13). By calculating the similarity between the user portraits, the similarity between different users may be obtained, and personalized information retrieval and recommendation may be achieved.

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, 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: 一种基于画像相似性的信息检索推荐方法、装置、系统及存储介质, 其中的方法包括:
获取不同用户的用户画像, 并确定各用户画像之间的用户画像相似性(S11); 基于所述用户画像
相似性创建用户动态社区, 使相似画像的用户分属到同一用户动态社区内(S12); 根据所述用户
动态社区及所述用户的查询语句对所述用户进行信息检索推荐(S13)。通过计算用户画像之间的
相似性, 获得不同用户之间的相似性, 可实现个性化信息检索和推荐。

基于画像相似性的信息检索推荐方法、装置及存储介质

5 本申请要求申请号为201910748591.3，申请日为2019年8月14日，发明创造名称为“基于画像相似性的信息检索推荐方法、装置及存储介质”的专利申请的优先权。

技术领域

10 本申请涉及数据分析技术领域，尤其涉及一种基于用户画像相似性的信息检索推荐方法、装置、系统及计算机可读存储介质。

背景技术

15 协同信息检索（Collaborative Information Retrieval, CIR）是一种基于社会关系的信息检索方法，该CIR协同信息检索系统能够对用户交互历史记录进行分析，以便更有效地响应后续的用户查询。但是，当两个用户同时向 CIR系统发送相同的查询时，由于两个用户的目标和行为特征可能不同，二者可能对两个不同的文档列表感兴趣，此时，CIR就面临个性化的查询推荐问题。

20 目前，信息检索是用户进行信息查询和获取的主要方式，是查找信息的方法和手段，信息的存储是实现信息检索的基础，这里要存储的信息包括原始文档数据、图片、视频和音频等，为实现信息检索首先要将这些原始信息进行计算机语言的转换，并将其存储在数据库中，否则无法进行机器识别。待用户根据意图输入查询请求后，检索系统根据用户的查询请求在数据库中搜索与查询相关的信息，通过一定的匹配机制计算出信息的相似度大小，并按从大到小的顺序将信息转换输出。

25 发明人意识到，现有的信息检索方法要么比较复杂，要么检索精度差，个性化不足，导致推荐效果差，用户体验不佳。

发明内容

30 本申请提供一种基于用户画像相似性的信息检索推荐方法、电子装置、系统及计算机可读存储介质，其主要目的在于通过加权二分图最大匹配获取

用户画像相似性的方式，获取不同用户之间的画像相似性，该方法能够在协同信息检索环境中动态构建用户社区，并将其应用于个性化信息检索，提高检索准确率，优化用户体验。

为实现上述目的，本申请提供一种基于用户画像相似性的信息检索推荐方法，应用于电子装置，方法包括：

获取不同用户的用户画像，并确定各用户画像之间的用户画像相似性；

基于用户画像相似性创建用户动态社区，使相似画像的用户分属到同一用户动态社区内；

根据用户动态社区及用户的查询语句对用户进行信息检索推荐。

为实现上述目的，本申请还提供一种电子装置，该电子装置包括：存储器及处理器，存储器中包括基于画像相似性的信息检索推荐程序，基于用户画像相似性的信息检索推荐程序被处理器执行时实现如下步骤：

获取不同用户的用户画像，并确定各用户画像之间的用户画像相似性；

基于用户画像相似性创建用户动态社区，使相似画像的用户分属到同一用户动态社区内；

根据用户动态社区及用户的查询语句对用户进行信息检索推荐。

为实现上述目的，本申请还提供一种基于画像相似性的信息检索推荐系统，包括：

用户画像相似性确定单元，用于获取不同用户的用户画像，并确定各用户画像之间的用户画像相似性；

动态社区创建单元，用户基于用户画像相似性创建用户动态社区，使相似画像的用户分属到同一用户动态社区内；

检索推荐单元，用于根据用户动态社区及用户的查询语句对用户进行信息检索推荐。

此外，为实现上述目的，本申请还提供一种计算机可读存储介质，计算机可读存储介质中包括基于用户画像相似性的信息检索推荐程序，基于用户画像相似性的信息检索推荐程序被处理器执行时，实现如上的基于用户画像相似性的信息检索推荐方法的步骤。

本申请提出的基于用户画像相似性的信息检索推荐方法、装置、系统及计算机可读存储介质，基于用户画像构造加权二分图，并采用加权二分图最

大匹配的方式获取用户画像之间的最大加权匹配值，能够在协同信息检索环境中基于用户画像相似性动态构建用户社区，并根据用户社区进行个性化信息检索推荐，能够提高用户检索准确率，优化用户体验，实现个性化推荐。

为了实现上述以及相关目的，本申请的一个或多个方面包括后面将详细说明的特征。下面的说明以及附图详细说明了本申请的某些示例性方面。然而，这些方面指示的仅仅是可使用本申请的原理的各种方式中的一些方式。此外，本申请旨在包括所有这些方面以及它们的等同物。

附图说明

图1为根据本申请基于用户画像相似性的信息检索推荐方法较佳实施例的应用环境示意图；

图2为根据本申请基于用户画像相似性的信息检索推荐系统较佳实施例的模块示意图；

图3为根据本申请基于用户画像相似性的信息检索推荐方法较佳实施例的流程图；

图4为基于图算法的用户画像相似性计算方法流程图；

图5为基于两个不同用户的用户画像构造的二分图。

本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

应当理解，此处所描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

本申请提供一种基于用户画像相似性的信息检索推荐方法，应用于一种电子装置1。参照图1所示，为本申请基于用户画像相似性的信息检索推荐方法较佳实施例的应用环境示意图。

在本实施例中，电子装置1可以是服务器、智能手机、平板电脑、便携计算机、桌上型计算机等具有运算功能的终端设备。

该电子装置1包括：处理器12、存储器11、网络接口14及通信总线15。

存储器11包括至少一种类型的可读存储介质。至少一种类型的可读存储

介质可为如闪存、硬盘、多媒体卡、卡型存储器 11 等的非易失性存储介质。在一些实施例中,所述可读存储介质可以是所述电子装置 1 的内部存储单元,例如该电子装置 1 的硬盘。在另一些实施例中,所述可读存储介质也可以是所述电子装置 1 的外部存储器 11,例如所述电子装置 1 上配备的插接式硬盘,智能存储卡 (Smart Media Card, SMC), 安全数字 (Secure Digital, SD) 卡, 5 闪存卡 (Flash Card) 等。

在本实施例中,所述存储器 11 的可读存储介质通常用于存储安装于所述电子装置 1 的基于用户画像相似性的信息检索推荐程序 10 等。所述存储器 11 还可以用于暂时地存储已经输出或者将要输出的数据。

10 处理器 12 在一些实施例中可以是一中央处理器 (Central Processing Unit, CPU), 微处理器或其他数据处理芯片, 用于运行存储器 11 中存储的程序代码或处理数据, 例如执行基于用户画像相似性的信息检索推荐程序 10 等。

网络接口 14 可选地可以包括标准的有线接口、无线接口(如 WI-FI 接口), 通常用于在该电子装置 1 与其他电子设备之间建立通信连接。

15 通信总线 15 用于实现这些组件之间的连接通信。

图 1 仅示出了具有组件 11-15 的电子装置 1, 但是应理解的是, 并不要求实施所有示出的组件, 可以替代的实施更多或者更少的组件。

可选地, 该电子装置 1 还可以包括用户接口, 用户接口可以包括输入单元比如键盘 (Keyboard)、语音输入装置比如麦克风 (microphone) 等具有语音识别功能的设备、语音输出装置比如音响、耳机等, 可选地用户接口还可以包括标准的有线接口、无线接口。

可选地, 该电子装置 1 还可以包括显示器, 显示器也可以称为显示屏或显示单元。在一些实施例中可以是 LED 显示器、液晶显示器、触控式液晶显示器以及有机发光二极管 (Organic Light-Emitting Diode, OLED) 触摸器等。25 显示器用于显示在电子装置 1 中处理的信息以及用于显示可视化的用户界面。

可选地, 该电子装置 1 还包括触摸传感器。所述触摸传感器所提供的供用户进行触摸操作的区域称为触控区域。此外, 这里所述的触摸传感器可以为电阻式触摸传感器、电容式触摸传感器等。而且, 所述触摸传感器不仅包括接触式的触摸传感器, 也可包括接近式的触摸传感器等。此外, 所述触摸30 传感器可以为单个传感器, 也可以为例如阵列布置的多个传感器。

此外,该电子装置 1 的显示器的面积可以与所述触摸传感器的面积相同,也可以不同。可选地,将显示器与所述触摸传感器层叠设置,以形成触摸显示屏。该装置基于触摸显示屏侦测用户触发的触控操作。

可选地,该电子装置 1 还可以包括射频(Radio Frequency, RF)电路,传感器、音频电路等等,在此不再赘述。

在图 1 所示的装置实施例中,作为一种计算机存储介质的存储器 11 中可以包括操作系统、以及基于用户画像相似性的信息检索推荐程序 10;处理器 12 执行存储器 11 中存储的基于用户画像相似性的信息检索推荐程序 10 时实现如下步骤:

10 获取不同用户的用户画像,并确定各用户画像之间的用户画像相似性;
基于用户画像相似性创建用户动态社区,使相似画像的用户分属到同一用户动态社区内;

根据用户动态社区及用户的查询语句对用户进行信息检索推荐。

在上述步骤中,获取不同用户的用户画像,并确定各用户画像之间的用户画像相似性可通过基于图算法的用户画像相似性计算方法来获取;

具体地,基于图算法的用户画像相似性计算方法包括以下步骤:

将用户画像 P 存储为与坐标(q, D_q)相关的集合;其中, q 表示用户的任意一个查询记录, D_q 表示与查询记录 q 相关的所有文档;

20 基于待处理的用户画像 $P(X)$ 和用户画像 $P(Y)$ 构造加权二分图;其中, $P(X)$ 为用户 X 的用户画像, $P(Y)$ 为用户 Y 的用户画像, $P(X)$ 的顶点 e 通过边($e, \acute{e}$)连接到 $P(Y)$ 的顶点 $\acute{e}$;

基于加权二分图获取用户画像 $P(X)$ 的顶点 e 与用户画像 $P(Y)$ 的顶点 $\acute{e}$ 之间的相似性;

根据 $P(X)$ 的顶点 e 与 $P(Y)$ 的顶点 $\acute{e}$ 之间的相似性确定边($e, \acute{e}$)的权重;

25 基于边($e, \acute{e}$)的权重获取用户画像 $P(X)$ 和用户画像 $P(Y)$ 之间的最大加权匹配值;

根据最大加权匹配值获取用户 X 和用户 Y 的用户画像相似性。

优选地,用户 X 的用户画像 $P(X)$ 存储为:

$$P(X) = \{(q_1^X, D_{q_1^X}^X), (q_2^X, D_{q_2^X}^X), \dots, (q_N^X, D_{q_N^X}^X)\}$$

用户 Y 的用户画像 $P(Y)$ 存储为:

$$P(Y) = \{(q_1^Y, D_{q_1^Y}^Y), (q_2^Y, D_{q_2^Y}^Y), \dots, (q_M^Y, D_{q_M^Y}^Y)\}$$

其中, q_i^X 表示用户 X 的第 i 个查询, $D_{q_i}^X$ 表示与查询 q_i^X 有关的所有文档;
 q_j^Y 表示用户 Y 的第 j 个查询, $D_{q_j}^Y$ 表示与查询 q_j^Y 有关的所有文档。

优选地, 用户画像 $P(X)$ 的顶点 e 包括对应的第一查询元素和第一文档元素, 用户画像 $P(Y)$ 的顶点 e 包括对应的第二查询元素和第二文档元素;

5 获取用户画像 $P(X)$ 的顶点 e 与用户画像 $P(Y)$ 的顶点 e 之间的相似性的过程包括:

获取第一查询元素和第二查询元素之间的第一相似性, 以及获取第一文档元素和第二文档元素之间的第二相似性;

基于第一相似性和第二相似性确定顶点 e 和顶点 e 之间的相似性。

10 优选地, 第一查询元素和第二查询元素的第一相似性通过编辑距离算法、杰卡德系数算法、TF 算法、TFIDF 算法或 Word2Vec 算法获取;

第一文档元素和第二文档元素之间的第二相似性通过 TFIDF 算法或基于空间向量的余弦算法获取。

15 优选地, 用户 X 的用户画像 $P(X)$ 包括元素 A、B、C、D、E, 其中元素 A、B、C、D、E 包含第一查询元素和第一文档元素;

用户 Y 的用户画像 $P(Y)$ 包含元素 1、2、3、4、5, 其中元素 1、2、3、4、5 包含第二查询元素和第二文档元素;

第一步: 通过以下公式获取加权二分图的所有的加权匹配值;

$$M_1 = w(A, 1) + w(B, 3) + w(C, 2) + w(D, 4) + w(E, 5)$$

$$M_2 = w(A, 1) + w(B, 3) + w(C, 5) + w(D, 4) + w(E, 2)$$

$$M_2 = w(A, 1) + w(B, 4) + w(C, 2) + w(D, 3) + w(E, 5)$$

$$M_2 = w(A, 1) + w(B, 4) + w(C, 5) + w(D, 3) + w(E, 2)$$

其中, $w(i, j)$ 表示元素 i 和元素 j 之间的相似性或者边 ij 的权重;

20 第二步: 从所有的加权匹配值中确定最大加权匹配值。

待获取不同用户的用户画像相似性之后, 可以基于用户 $P(X)$ 和用户 $P(Y)$ 之间的用户画像相似性创建用户社区, 并根据所创建的用户社区对用户查询结果进行排序推荐。

25 作为具体示例, 假设用户 U 需要查询的语句为 q, 基于用户 $P(X)$ 和用户 $P(Y)$ 之间的用户画像相似性进行查询的步骤包括:

步骤一: 寻找与查询 q 相似的历史查询记录 A。

设 $A = \{(U_1, q_1, D_{q_1}), (U_2, q_2, D_{q_2}), \dots, (U_m, q_m, D_{q_m})\}$

$$s(q, q_i) > \theta \text{ 且 } s(P(U), P(U_i)) > \omega \quad 1 \leq i \leq m$$

其中, U_m 表示用户, q_m 为用户 U_m 的查询, D_{qm} 为与查询 q_m 相关的所有文档, $P(U)$ 为用户 U 的用户画像, $P(U_i)$ 为用户 i 的用户画像, $s(P(U), P(U_i))$ 为用户 U 和用户 I 之间的用户画像相似性; $s(q, q_1)$ 为语句 q 与语句 q_i 之间的相似性, 上述相似性均可通过基于图算法的用户画像相似性计算方法获得。

5 步骤二: 计算所有与查询 q 相关的文档集合。

$$D_q = D_{q1} \cup D_{q2} \cup \dots \cup D_{qm}$$

其次, 对于语料库中每个文档 d 符合 $d \in D_q$, 计算如下得分:

$$R(U, d, q) = \sum_{d \in D_{q_i}} s(q, q_i) \times s(P(U), P(U_i))$$

而对于每个 d 不属于 D_q , 则默认 $R(U, d, q) = 0$;

步骤三: 对于语料库中的每个文档 d , 计算 d 和 q 之间的相似性, 获取相似性 $r(d, q)$;

10 步骤四: 计算文集中每个文档的最终排名:

$$R_{\text{final}}(U, d, q) = a * r(d, q) + b * R(U, d, q)$$

其中, a 和 b 为设定系数。

步骤五: 按文档的最终排名可对文档进行排序, 以构造输出列表, 根据输出列表即可对用户 U 需要查询的语句 q 进行查询输出。

上述实施例提出的电子装置 1, 通过加权二分图最大匹配的方式获取用户
15 画像之间的相似性, 能够在协同信息检索环境中基于用户画像相似性动态构建用户社区, 并根据用户社区进行个性化信息检索推荐, 能够提高用户检索准确率, 优化用户体验, 实现个性化推荐。

与上述电子装置相对应, 本申请还提供一种基于用户画像相似性的信息检索推荐系统。参照图 2 所示, 为本申请实施例中基于用户画像相似性的信息检索推荐系统较佳实施例的程序模块图。基于用户画像相似性的信息检索
20 推荐系统可以被分割为:

用户画像相似性确定单元 110, 用于获取不同用户的用户画像, 并确定各用户画像之间的用户画像相似性;

动态社区创建单元 120, 用户基于用户画像相似性创建用户动态社区, 使
25 相似画像的用户分属到同一用户动态社区内;

检索推荐单元 130, 用于根据用户动态社区及用户的查询语句对用户进行

信息检索推荐。

用户画像相似性确定单元 110 进一步包括：

用户画像存储模块 111，用于将用户画像 P 存储为与坐标(q, D_q)相关的集合；其中，q 表示用户的任意一个查询记录，D_q表示与查询记录 q 相关的所有文档；

加权二分图构造模块 112，用于基于待处理的用户画像 P(X)和用户画像 P(Y)构造加权二分图；其中，P(X)为用户 X 的用户画像，P(Y)为用户 Y 的用户画像，P(X)的顶点 e 通过边(e, é)连接到 P(Y)的顶点é；

相似性获取模块 113，用于基于加权二分图获取用户画像 P(X)的顶点 e 与用户画像 P(Y)的顶点é之间的相似性；

权重确定模块 114，用于根据 P(X)的顶点 e 与 P(Y)的顶点é之间的相似性确定边(e, é)的权重；

最大加权匹配值获取模块 115，用于基于边(e,é)的权重获取用户画像 P(X)和用户画像 P(Y)之间的最大加权匹配值；

用户画像相似性确定模块 116，用于根据最大加权匹配值获取用户 X 和用户 Y 的用户画像相似性。

具体地，用户 X 的用户画像 P(X)存储为：

$$P(X) = \{(q_1^X, D_{q_1^X}^X), (q_2^X, D_{q_2^X}^X), \dots, (q_N^X, D_{q_N^X}^X)\}$$

用户 Y 的用户画像 P(Y)存储为：

$$P(Y) = \{(q_1^Y, D_{q_1^Y}^Y), (q_2^Y, D_{q_2^Y}^Y), \dots, (q_M^Y, D_{q_M^Y}^Y)\}$$

其中， q_i^X 表示用户 X 的第 i 个查询， $D_{q_i^X}^X$ 表示与查询 q_i^X 有关的所有文档； q_j^Y 表示用户 Y 的第 j 个查询， $D_{q_j^Y}^Y$ 表示与查询 q_j^Y 有关的所有文档。

其中，用户画像 P(X)的顶点 e 包括对应的第一查询元素和第一文档元素，用户画像 P(Y)的顶点é包括对应的第二查询元素和第二文档元素；

相似性获取模块 113 包括：

查询元素和文档元素相似性获取模块 1131，用于获取第一查询元素和第二查询元素之间的第一相似性，以及获取第一文档元素和第二文档元素之间的第二相似性；

顶点间相似性确定模块 1132，用于基于第一相似性和第二相似性确定顶点 e 和顶点é之间的相似性。

查询元素和文档元素相似性获取模块 1131 包括：

第一相似性获取模块，用于通过编辑距离算法、杰卡德系数算法、TF 算法、TFIDF 算法或 Word2Vec 算法获取第一查询元素和第二查询元素的第一相似性；

第二相似性获取模块，用于通过 TFIDF 算法或基于空间向量的余弦算法获取第一文档元素和第二文档元素之间的第二相似性。

用户 X 的用户画像 P(X) 包括元素 A、B、C、D、E，其中元素 A、B、C、D、E 包含第一查询元素和第一文档元素；

用户 Y 的用户画像 P(Y) 包含元素 1、2、3、4、5，其中元素 1、2、3、4、5 包含第二查询元素和第二文档元素；

10 第一步：通过以下公式获取加权二分图的所有的加权匹配值；

$$M_1 = w(A, 1) + w(B, 3) + w(C, 2) + w(D, 4) + w(E, 5)$$

$$M_2 = w(A, 1) + w(B, 3) + w(C, 5) + w(D, 4) + w(E, 2)$$

$$M_2 = w(A, 1) + w(B, 4) + w(C, 2) + w(D, 3) + w(E, 5)$$

$$M_2 = w(A, 1) + w(B, 4) + w(C, 5) + w(D, 3) + w(E, 2)$$

其中， $w(i, j)$ 表示元素 i 和元素 j 之间的相似性或者边 ij 的权重；

第二步：从所有的加权匹配值中确定最大加权匹配值。

此外，本申请还提供一种基于用户画像相似性的信息检索推荐方法。参照图 3 所示，为本申请基于用户画像相似性的信息检索推荐方法较佳实施例的流程图。该方法可以由一个装置执行，该装置可以由软件和/或硬件实现。

在本实施例中，基于用户画像相似性的信息检索推荐方法包括以下步骤：

步骤 S11：获取不同用户的用户画像，并确定各用户画像之间的用户画像相似性。

步骤 S12：基于用户画像相似性创建用户动态社区，使相似画像的用户分属到同一用户动态社区内。

步骤 S13：根据用户动态社区及用户的查询语句对用户进行信息检索推荐。

如图 4 基于图算法的用户画像相似性计算方法流程所示，上述步骤 S11 进一步包括以下步骤：

步骤 S101：将用户画像 P 存储为与坐标 (q, D_q) 相关的集合；其中，q 表示用户的任意一个查询记录， D_q 表示与查询记录 q 相关的所有文档。

其中，用户画像又称用户角色，作为一种勾画目标用户、联系用户诉求与设计方向的有效工具，用户画像在各领域得到了广泛的应用。我们在实际

操作的过程中往往会以最为浅显和贴近生活的话语将用户的属性、行为与期待联结起来。作为实际用户的虚拟代表，用户画像所形成的用户角色并不是脱离产品 and 市场之外所构建出来的，形成的用户角色需要有代表性能代表产品的主要受众和目标群体。

5 在本申请中，用户 X 的用户画像 P(X) 可存储为：

$$P(X) = \{(q_1^X, D_{q_1^X}^X), (q_2^X, D_{q_2^X}^X), \dots, (q_N^X, D_{q_N^X}^X)\}$$

用户 Y 的用户画像 P(Y) 可存储为：

$$P(Y) = \{(q_1^Y, D_{q_1^Y}^Y), (q_2^Y, D_{q_2^Y}^Y), \dots, (q_M^Y, D_{q_M^Y}^Y)\}$$

其中， q_i^X 表示用户 X 的第 i 个查询， $D_{q_i^X}^X$ 表示与查询 q_i^X 有关的所有文档； q_j^Y 表示用户 Y 的第 j 个查询， $D_{q_j^Y}^Y$ 表示与查询 q_j^Y 有关的所有文档。

因此，用户 X 和用户 Y 之间的用户画像相似性 (User Profile Similarity, UPS)

10 即为计算以上两组集合 P(x) 与 P(y) 之间的相似性。

步骤 S102: 基于待处理的用户画像 P(X) 和用户画像 P(Y) 构造加权二分图；其中，P(X) 为用户 X 的用户画像，P(Y) 为用户 Y 的用户画像，P(X) 的顶点 e 通过边 (e, é) 连接到 P(Y) 的顶点 é。

其中，二分图又称作二部图，是图论中的一种特殊模型。设 $G=(V,E)$ 是一个无向图，如果顶点 V 可分割为两个互不相交的子集 (A,B)，并且图中的每条边 (i, j) 所关联的两个顶点 i 和 j 分别属于这两个不同的顶点集 ($i \in A, j \in B$)，则称图 G 为一个二分图。

其中，基于上述用户画像 P(x) 和用户画像 P(Y) 构造一个加权二分图 $G=(V=(P(X), P(Y)), E)$ 。用户画像 P(X) 的元素构成图 G 的一部分，P(Y) 的元素构成图的另一部分。P(X) 的每个顶点 e 通过边 (e, é) 连接到 P(Y) 的每个顶点 é。而边 (e, é) 的权重则等于顶点 (或元素) e 和 é 之间的相似性。其中，边 (e, é) 的权重和元素类型有关，元素类型包括查询或者文档。

优选地，用户画像 P(X) 的顶点 e 包括对应的第一查询元素和第一文档元素，用户画像 P(Y) 的顶点 é 包括对应的第二查询元素和第二文档元素；

25 获取用户画像 P(X) 的顶点 e 与用户画像 P(Y) 的顶点 é 之间的相似性的过程包括：

首先，获取第一查询元素和第二查询元素之间的第一相似性，以及获取第一文档元素和第二文档元素之间的第二相似性；然后，基于第一相似性和第二相似性确定顶点 e 和顶点 é 之间的相似性。

步骤 S103: 基于加权二分图获取用户画像 $P(X)$ 的顶点 e 与用户画像 $P(Y)$ 的顶点 $\acute{e}$ 之间的相似性。

其中, 用户画像 $P(X)$ 的各顶点 e 包括相应的查询元素和文档元素, 用户画像 $P(Y)$ 的各顶点也包括相应的查询元素和文档元素, 在获取顶点 e 与顶点 $\acute{e}$ 之间的相似性的过程中, 首先获取用户画像 $P(X)$ 和用户画像 $P(Y)$ 的各查询元素之间的相似性, 以及用户画像 $P(X)$ 和用户画像 $P(Y)$ 之间的各文档元素之间的相似性, 基于用户画像 $P(x)$ 和用户画像 $P(Y)$ 的查询顶点之间的相似性及各文档顶点之间的相似性, 就可以确定所有顶点 e 与顶点 $\acute{e}$ 之间的相似性。

具体地, 用户画像 $P(X)$ 和用户画像 $P(Y)$ 的各查询之间的相似性的计算, 实际为各查询语句之间的相似性计算, 目前的查询语句相似性获取方法主要包括: 编辑距离算法、杰卡德系数算法、TF 算法、TFIDF 算法以及 Word2Vec 算法等。

进一步地, 编辑距离算法, 英文叫做 Edit Distance, 又称 Levenshtein 距离算法, 是指两个字串之间, 由一个转成另一个所需的最少编辑操作次数, 如果它们的距离越大, 说明它们越是不同。许可的编辑操作包括将一个字符替换成另一个字符、插入一个字符、删除一个字符等。

而杰卡德系数, 英文叫做 Jaccard index, 又称为 Jaccard 相似系数, 用于比较有限样本集之间的相似性与差异性。Jaccard 系数值越大, 样本相似度越高。实际上。杰卡德系数的计算方式非常简单, 就是两个样本的交集除以并集得到的数值, 当两个样本完全一致时, 结果为 1, 当两个样本完全不同时, 结果为 0。

另外, 用户画像 $P(X)$ 和用户画像 $P(Y)$ 的各文档之间的相似性计算方法主要包括 TFIDF 算法和基于空间向量的余弦算法等。

换言之, 第一查询元素和第二查询元素的第一相似性通过编辑距离算法、杰卡德系数算法、TF 算法、TFIDF 算法或 Word2Vec 算法等算法获取; 第一文档元素和第二文档元素之间的第二相似性通过 TFIDF 算法或基于空间向量的余弦算法等方法获取。

步骤 S104: 根据 $P(X)$ 的顶点 e 与 $P(Y)$ 的顶点 $\acute{e}$ 之间的相似性确定边 $(e, \acute{e})$ 的权重。

具体地, 可将边 $(e, \acute{e})$ 的权重设置为等于 $P(X)$ 的顶点 e 与 $P(Y)$ 的顶点 $\acute{e}$ 之

间的相似性。

步骤 S105: 基于边(e,é)的权重获取用户画像 P(X)和用户画像 P(Y)之间的最大加权匹配值。

其中, 二分图的最大匹配主要指: 给定一个二分图 G, 在二分图 G 的一个子图 M 中, M 的边集中的任意两条边都不依附于同一个顶点, 则称 M 是一个匹配。选择这样的边数最大的子集称为图的最大匹配问题 (maximal matching problem), 如果一个匹配中, 图中的每个顶点都和图中某条边相关联, 则称此匹配为完全匹配, 也称作完备匹配。

例如, 用户 X 的用户画像 P(X)包括元素 A、B、C、D、E, 其中 A、B、C、D、E 中包含第一查询元素和第一文档元素, 用户 Y 的用户画像 P(Y)包含元素 1、2、3、4、5, 其中 1、2、3、4、5 中包含第二查询元素和第二文档元素, 用户画像 P(X)和用户画像 P(Y)构造的二分图如图 4 所示。

根据图 5 基于用户画像构造的二分图所示, 最大匹配情况的加权匹配值计算通过以下公式计算:

$$M_1 = w(A, 1) + w(B, 3) + w(C, 2) + w(D, 4) + w(E, 5)$$

$$M_2 = w(A, 1) + w(B, 3) + w(C, 5) + w(D, 4) + w(E, 2)$$

$$M_2 = w(A, 1) + w(B, 4) + w(C, 2) + w(D, 3) + w(E, 5)$$

$$M_2 = w(A, 1) + w(B, 4) + w(C, 5) + w(D, 3) + w(E, 2)$$

其中, $w(i, j)$ 表示元素 i 和元素 j 之间的相似性或者边 ij 的权重; 例如, $w(A, 1)$ 表示为元素 A 和元素 1 之间的相似性, 其也表示边 A1 的权重, $w(B, 3)$ 、 $w(C, 2) \cdots w(E, 5)$ 等类似。

进而, 从所有的加权匹配值中确定最大加权匹配值, 在该具体实施例中, 最大加权匹配值为 3.5。

步骤 S106: 根据最大加权匹配值获取用户 X 和用户 Y 的用户画像相似性。

其中, 待获取不同用户的用户画像相似性之后, 可以基于用户 P(X)和用户 P(Y)之间的用户画像相似性创建用户社区, 并根据所创建的用户社区对用户查询结果进行排序推荐。

作为具体示例, 假设用户 U 需要查询的语句为 q, 基于用户 P(X)和用户 P(Y)之间的用户画像相似性进行查询的步骤包括:

步骤一: 寻找与查询 q 相似的历史查询记录 A。

设 $A = \{(U_1, q_1, D_{q1}), (U_2, q_2, D_{q2}), \dots (U_m, q_m, D_{qm})\}$

$$s(q, q_i) > \theta \text{ 且 } s(P(U), P(U_i)) > \omega \quad 1 \leq i \leq m$$

其中, U_m 表示用户, q_m 为用户 U_m 的查询, D_{qm} 为与查询 q_m 相关的所有文档, $P(U)$ 为用户 U 的用户画像, $P(U_i)$ 为用户 i 的用户画像, $s(P(U), P(U_i))$ 为用户 U 和用户 I 之间的用户画像相似性; $s(q, q_1)$ 为语句 q 与语句 q_i 之间的相似性, 上述相似性均可通过基于图算法的用户画像相似性计算方法获得。

5 步骤二: 计算所有与查询 q 相关的文档集合。

$$D_q = D_{q1} \cup D_{q2} \cup \dots \cup D_{qm}$$

其次, 对于语料库中每个文档 d 符合 $d \in D_q$, 计算如下得分:

$$R(U, d, q) = \sum_{d \in D_{q_i}} s(q, q_i) \times s(P(U), P(U_i))$$

而对于每个 d 不属于 D_q , 则默认 $R(U, d, q) = 0$;

步骤三: 对于语料库中的每个文档 d , 计算 d 和 q 之间的相似性, 获取相似性 $r(d, q)$;

10 步骤四: 计算文集中每个文档的最终排名:

$$R_{\text{final}}(U, d, q) = a * r(d, q) + b * R(U, d, q)$$

其中, a 和 b 为设定系数。

步骤五: 按文档的最终排名可对文档进行排序, 以构造输出列表, 根据输出列表即可对用户 U 需要查询的语句 q 进行查询输出。

15 利用上述基于用户选项相似性的信息检索推荐方法, 采用加权二分图最大匹配的方式获取用户画像之间的相似性, 能够在协同信息检索环境中基于用户画像相似性动态构建用户社区, 并根据用户社区进行个性化信息检索推荐, 能够提高用户检索准确率, 优化用户体验, 实现个性化推荐。

20 此外, 本申请实施例还提出一种计算机可读存储介质, 计算机可读存储介质中包括基于用户画像相似性的信息检索推荐程序, 基于用户画像相似性的信息检索推荐程序被处理器执行时实现如下操作:

获取不同用户的用户画像, 并确定各用户画像之间的用户画像相似性;

基于用户画像相似性创建用户动态社区, 使相似画像的用户分属到同一用户动态社区内;

根据用户动态社区及用户的查询语句对用户进行信息检索推荐。

25 本申请之计算机可读存储介质的具体实施方式与上述基于用户画像相似性的信息检索推荐方法、电子装置、系统的具体实施方式大致相同, 在此不

再赘述。

需要说明的是，在本文中，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、装置、物品或者方法不仅包括那些要素，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程、装置、物品或者方法所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、装置、物品或者方法中还存在另外的相同要素。

上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在如上所述的一个存储介质(如 ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台终端设备(可以是手机，计算机，服务器，或者网络设备等)执行本申请各个实施例所述的方法。

以上仅为本申请的优选实施例，并非因此限制本申请的专利范围，凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本申请的专利保护范围内。

权 利 要 求 书

1、一种基于画像相似性的信息检索推荐方法，应用于电子装置，其特征在于，所述方法包括：

- 5 获取不同用户的用户画像，并确定各用户画像之间的用户画像相似性；
 基于所述用户画像相似性创建用户动态社区，使相似画像的用户分属到同一用户动态社区内；

 根据所述用户动态社区及所述用户的查询语句对所述用户进行信息检索推荐。

- 10 2、根据权利要求 1 所述的基于画像相似性的信息检索推荐方法，其特征在于，所述获取不同用户的用户画像，并确定各用户画像之间的用户画像相似性的步骤包括：

 将用户画像 P 存储为与坐标 (q, D_q) 相关的集合；其中， q 表示用户的任意一个查询记录， D_q 表示与所述查询记录 q 相关的所有文档；

- 15 基于待处理的用户画像 $P(X)$ 和用户画像 $P(Y)$ 构造加权二分图；其中， $P(X)$ 为用户 X 的用户画像， $P(Y)$ 为用户 Y 的用户画像， $P(X)$ 的顶点 e 通过边 $(e, \acute{e})$ 连接到 $P(Y)$ 的顶点 $\acute{e}$ ；

 基于所述加权二分图获取所述用户画像 $P(X)$ 的顶点 e 与所述用户画像 $P(Y)$ 的顶点 $\acute{e}$ 之间的相似性；

- 20 根据所述 $P(X)$ 的顶点 e 与所述 $P(Y)$ 的顶点 $\acute{e}$ 之间的相似性确定所述边 $(e, \acute{e})$ 的权重；

 基于所述边 $(e, \acute{e})$ 的权重获取用户画像 $P(X)$ 和用户画像 $P(Y)$ 之间的最大加权匹配值；

 根据所述最大加权匹配值获取所述用户 X 和用户 Y 的用户画像相似性。

- 25 3、根据权利要求 2 所述的基于画像相似性的信息检索推荐方法，其特征在于，

 所述用户 X 的用户画像 $P(X)$ 存储为：

$$P(X) = \{(q_1^X, D_{q_1^X}^X), (q_2^X, D_{q_2^X}^X), \dots, (q_N^X, D_{q_N^X}^X)\}$$

 所述用户 Y 的用户画像 $P(Y)$ 存储为：

$$P(Y) = \{(q_1^Y, D_{q_1^Y}^Y), (q_2^Y, D_{q_2^Y}^Y), \dots, (q_M^Y, D_{q_M^Y}^Y)\}$$

其中, q_i^X 表示用户 X 的第 i 个查询, $D_{q_i^X}^X$ 表示与查询 q_i^X 有关的所有文档;
 q_j^Y 表示用户 Y 的第 j 个查询, $D_{q_j^Y}^Y$ 表示与查询 q_j^Y 有关的所有文档。

4、根据权利要求 2 所述的基于画像相似性的信息检索推荐方法, 其特征在于,

5 所述用户画像 $P(X)$ 的顶点 e 包括对应的第一查询元素和第一文档元素,
 所述用户画像 $P(Y)$ 的顶点 $\hat{e}$ 包括对应的第二查询元素和第二文档元素;

所述获取所述用户画像 $P(X)$ 的顶点 e 与所述用户画像 $P(Y)$ 的顶点 $\hat{e}$ 之间的相似性的过程包括:

获取所述第一查询元素和第二查询元素之间的第一相似性, 以及获取所述
 10 第一文档元素和所述第二文档元素之间的第二相似性;

基于所述第一相似性和所述第二相似性确定所述顶点 e 和顶点 $\hat{e}$ 之间的相似性。

5、根据权利要求 4 所述的基于画像相似性的信息检索推荐方法, 其特征在于,

15 所述第一查询元素和第二查询元素的第一相似性通过编辑距离算法、杰卡德系数算法、TF 算法、TFIDF 算法或 Word2Vec 算法获取;

所述第一文档元素和所述第二文档元素之间的第二相似性通过 TFIDF 算法或基于空间向量的余弦算法获取。

6、根据权利要求 2 所述的基于画像相似性的信息检索推荐方法, 其特征
 20 在于,

所述用户 X 的用户画像 $P(X)$ 包括元素 A、B、C、D、E, 其中元素 A、B、C、D、E 包含第一查询元素和第一文档元素;

用户 Y 的用户画像 $P(Y)$ 包含元素 1、2、3、4、5, 其中元素 1、2、3、4、5 包含第二查询元素和第二文档元素;

25 第一步: 通过以下公式获取所述加权二分图的所有的加权匹配值;

$$M_1 = w(A, 1) + w(B, 3) + w(C, 2) + w(D, 4) + w(E, 5)$$

$$M_2 = w(A, 1) + w(B, 3) + w(C, 5) + w(D, 4) + w(E, 2)$$

$$M_2 = w(A, 1) + w(B, 4) + w(C, 2) + w(D, 3) + w(E, 5)$$

$$M_2 = w(A, 1) + w(B, 4) + w(C, 5) + w(D, 3) + w(E, 2)$$

其中, $w(i, j)$ 表示元素 i 和元素 j 之间的相似性或者边 ij 的权重;

第二步: 从所有的加权匹配值中确定最大加权匹配值。

7、根据权利要求1所述的基于画像相似性的信息检索推荐方法，其特征在于，所述相似画像的用户分属到同一用户动态社区内的步骤包括：

将所述用户画像相似性符合相似性预设范围的所有用户分属到同一用户动态社区。

- 5 8、根据权利要求1所述的基于画像相似性的信息检索推荐方法，其特征在于，所述根据所述用户动态社区及所述用户的查询语句对所述用户进行信息检索推荐的步骤包括：

获取与所述查询语句相关的所有历史查询记录；

获取与所述历史查询记录相关的文档集合；

- 10 获取所述文档集合中的各文档与所述查询语句之间的相似性；

基于所述相似性对所述文档集合中的各文档进行排序，以根据排序构造输出列表；

基于所述输出列表输出与所述查询语句对应的查询结果。

- 15 9、根据权利要求8所述的基于画像相似性的信息检索推荐方法，其特征在于，所述所有历史查询记录的获取公式如下：

$$A = \{(U_1, q_1, D_{q1}), (U_2, q_2, D_{q2}), \dots (U_m, q_m, D_{qm})\}$$

$$s(q, q_i) > \theta \text{ 且 } s(P(U), P(U_i)) > \omega \quad 1 \leq i \leq m$$

其中，A表示历史查询记录， U_m 表示用户， q_m 为用户 U_m 的查询， D_{qm} 为与查询 q_m 相关的所有文档， $P(U)$ 为用户U的用户画像， $P(U_i)$ 为用户i的用户画像， $s(P(U), P(U_i))$ 为用户U和用户I之间的用户画像相似性； $s(q, q_1)$ 为语句q与语句 q_1 之间的相似性。

- 20 10、根据权利要求8所述的基于画像相似性的信息检索推荐方法，其特征在于，所述获取与所述历史查询记录相关的文档集合的公式如下：

$$D_q = D_{q1} \cup D_{q2} \cup \dots D_{qm}$$

其次， D_q 为文档集合， D_{qm} 为与查询 q_m 相关的所有文档，对于所述文档集合中的各文档d符合 $d \in D_q$ 。

- 25 11、一种电子装置，其特征在于，该电子装置包括：存储器及处理器，所述存储器中包括基于画像相似性的信息检索推荐程序，所述基于画像相似性的信息检索推荐程序被所述处理器执行时实现如下步骤：

获取不同用户的用户画像，并确定各用户画像之间的用户画像相似性；

基于所述用户画像相似性创建用户动态社区，使相似画像的用户分属到同一用户动态社区内；

根据所述用户动态社区及所述用户的查询语句对所述用户进行信息检索推荐。

- 5 12、根据权利要求 11 所述的电子装置，其特征在于，所述获取不同用户的用户画像，并确定各用户画像之间的用户画像相似性的步骤包括：

将用户画像 P 存储为与坐标 (q, D_q) 相关的集合；其中， q 表示用户的任意一个查询记录， D_q 表示与所述查询记录 q 相关的所有文档；

- 10 基于待处理的用户画像 $P(X)$ 和用户画像 $P(Y)$ 构造加权二分图；其中， $P(X)$ 为用户 X 的用户画像， $P(Y)$ 为用户 Y 的用户画像， $P(X)$ 的顶点 e 通过边 $(e, \acute{e})$ 连接到 $P(Y)$ 的顶点 $\acute{e}$ ；

基于所述加权二分图获取所述用户画像 $P(X)$ 的顶点 e 与所述用户画像 $P(Y)$ 的顶点 $\acute{e}$ 之间的相似性；

- 15 根据所述 $P(X)$ 的顶点 e 与所述 $P(Y)$ 的顶点 $\acute{e}$ 之间的相似性确定所述边 $(e, \acute{e})$ 的权重；

基于所述边 $(e, \acute{e})$ 的权重获取用户画像 $P(X)$ 和用户画像 $P(Y)$ 之间的最大加权匹配值；

根据所述最大加权匹配值获取所述用户 X 和用户 Y 的用户画像相似性。

- 20 13、根据权利要求 12 所述的电子装置，其特征在于，所述用户 X 的用户画像 $P(X)$ 存储为：

$$P(X) = \{(q_1^X, D_{q_1^X}), (q_2^X, D_{q_2^X}), \dots, (q_N^X, D_{q_N^X})\}$$

所述用户 Y 的用户画像 $P(Y)$ 存储为：

$$P(Y) = \{(q_1^Y, D_{q_1^Y}), (q_2^Y, D_{q_2^Y}), \dots, (q_M^Y, D_{q_M^Y})\}$$

其中， q_i^X 表示用户 X 的第 i 个查询， $D_{q_i^X}$ 表示与查询 q_i^X 有关的所有文档； q_j^Y 表示用户 Y 的第 j 个查询， $D_{q_j^Y}$ 表示与查询 q_j^Y 有关的所有文档。

14、一种基于画像相似性的信息检索推荐系统，其特征在于，包括：

- 25 用户画像相似性确定单元，用于获取不同用户的用户画像，并确定各用户画像之间的用户画像相似性；

动态社区创建单元，用户基于所述用户画像相似性创建用户动态社区，使相似画像的用户分属到同一用户动态社区内；

检索推荐单元，用于根据所述用户动态社区及所述用户的查询语句对所述用户进行信息检索推荐。

15、根据权利要求 14 所述的基于画像相似性的信息检索推荐系统，其特征在于，所述用户画像相似性确定单元包括：

5 用户画像存储模块，用于将用户画像 P 存储为与坐标 (q, D_q) 相关的集合；其中， q 表示用户的任意一个查询记录， D_q 表示与所述查询记录 q 相关的所有文档；

加权二分图构造模块，用于基于待处理的用户画像 $P(X)$ 和用户画像 $P(Y)$ 构造加权二分图；其中， $P(X)$ 为用户 X 的用户画像， $P(Y)$ 为用户 Y 的用户画像， $P(X)$ 的顶点 e 通过边 $(e, \acute{e})$ 连接到 $P(Y)$ 的顶点 $\acute{e}$ ；

相似性获取模块，用于基于所述加权二分图获取所述用户画像 $P(X)$ 的顶点 e 与所述用户画像 $P(Y)$ 的顶点 $\acute{e}$ 之间的相似性；

权重确定模块，用于根据所述 $P(X)$ 的顶点 e 与所述 $P(Y)$ 的顶点 $\acute{e}$ 之间的相似性确定所述边 $(e, \acute{e})$ 的权重；

15 最大加权匹配值获取模块，用于基于所述边 $(e, \acute{e})$ 的权重获取用户画像 $P(X)$ 和用户画像 $P(Y)$ 之间的最大加权匹配值；

用户画像相似性确定模块，用于根据所述最大加权匹配值获取所述用户 X 和用户 Y 的用户画像相似性。

20 16、根据权利要求 15 所述的基于画像相似性的信息检索推荐系统，其特征在于，

所述用户 X 的用户画像 $P(X)$ 存储为：

$$P(X) = \{(q_1^X, D_{q_1^X}), (q_2^X, D_{q_2^X}), \dots, (q_N^X, D_{q_N^X})\}$$

所述用户 Y 的用户画像 $P(Y)$ 存储为：

$$P(Y) = \{(q_1^Y, D_{q_1^Y}), (q_2^Y, D_{q_2^Y}), \dots, (q_M^Y, D_{q_M^Y})\}$$

其中， q_i^X 表示用户 X 的第 i 个查询， $D_{q_i^X}$ 表示与查询 q_i^X 有关的所有文档； q_j^Y 表示用户 Y 的第 j 个查询， $D_{q_j^Y}$ 表示与查询 q_j^Y 有关的所有文档。

25 17、根据权利要求 15 所述的基于画像相似性的信息检索推荐系统，其特征在于，

所述用户画像 $P(X)$ 的顶点 e 包括对应的第一查询元素和第一文档元素，所述用户画像 $P(Y)$ 的顶点 $\acute{e}$ 包括对应的第二查询元素和第二文档元素；

所述相似性获取模块包括：

查询元素和文档元素相似性获取模块，用于获取所述第一查询元素和第二查询元素之间的第一相似性，以及获取所述第一文档元素和所述第二文档元素之间的第二相似性；

- 5 顶点间相似性确定模块，用于基于所述第一相似性和所述第二相似性确定所述顶点 e 和顶点 e' 之间的相似性。

18、根据权利要求 17 所述的基于画像相似性的信息检索推荐系统，其特征在于，所述查询元素和文档元素相似性获取模块包括：

- 10 第一相似性获取模块，用于通过编辑距离算法、杰卡德系数算法、TF 算法、TFIDF 算法或 Word2Vec 算法获取所述第一查询元素和第二查询元素的第一相似性；

第二相似性获取模块，用于通过 TFIDF 算法或基于空间向量的余弦算法获取所述第一文档元素和所述第二文档元素之间的第二相似性。

- 15 19、根据权利要求 15 所述的基于画像相似性的信息检索推荐系统，其特征在于，

所述用户 X 的用户画像 $P(X)$ 包括元素 A 、 B 、 C 、 D 、 E ，其中元素 A 、 B 、 C 、 D 、 E 包含第一查询元素和第一文档元素；

用户 Y 的用户画像 $P(Y)$ 包含元素 1、2、3、4、5，其中元素 1、2、3、4、5 包含第二查询元素和第二文档元素；

- 20 第一步：通过以下公式获取所述加权二分图的所有的加权匹配值；

$$M_1 = w(A, 1) + w(B, 3) + w(C, 2) + w(D, 4) + w(E, 5)$$

$$M_2 = w(A, 1) + w(B, 3) + w(C, 5) + w(D, 4) + w(E, 2)$$

$$M_2 = w(A, 1) + w(B, 4) + w(C, 2) + w(D, 3) + w(E, 5)$$

$$M_2 = w(A, 1) + w(B, 4) + w(C, 5) + w(D, 3) + w(E, 2)$$

其中， $w(i, j)$ 表示元素 i 和元素 j 之间的相似性或者边 ij 的权重；

第二步：从所有的加权匹配值中确定最大加权匹配值。

- 25 20、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质中包括基于画像相似性的信息检索推荐程序，所述基于画像相似性的信息检索推荐程序被处理器执行时，实现如权利要求 1 至 10 中任一项所述的基于画像相似性的信息检索推荐方法的步骤。

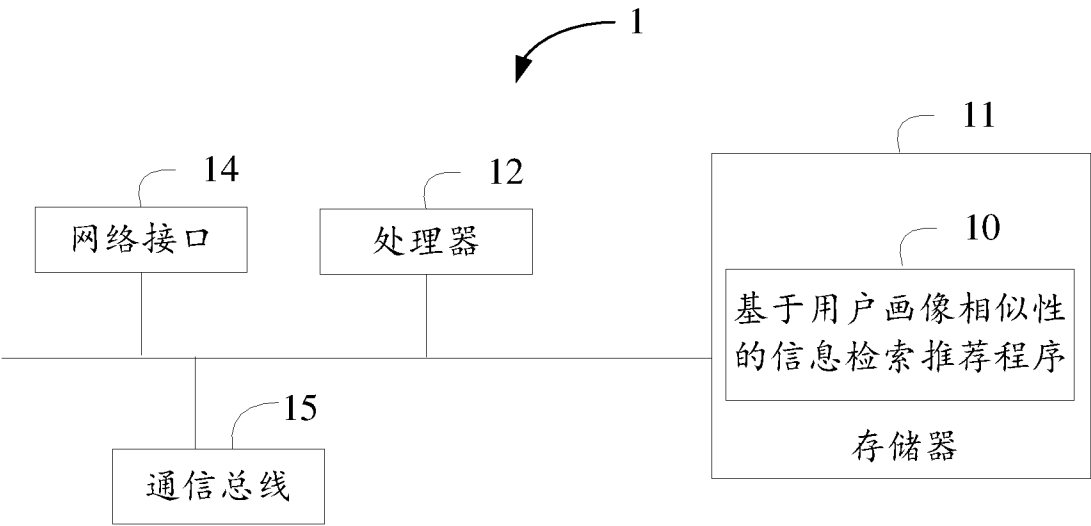


图 1



图 2

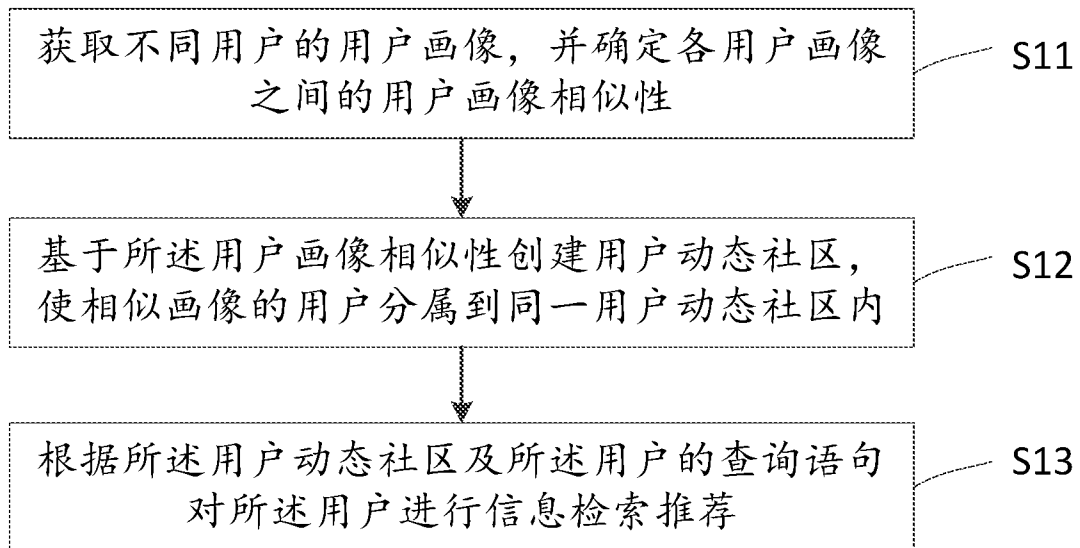


图 3

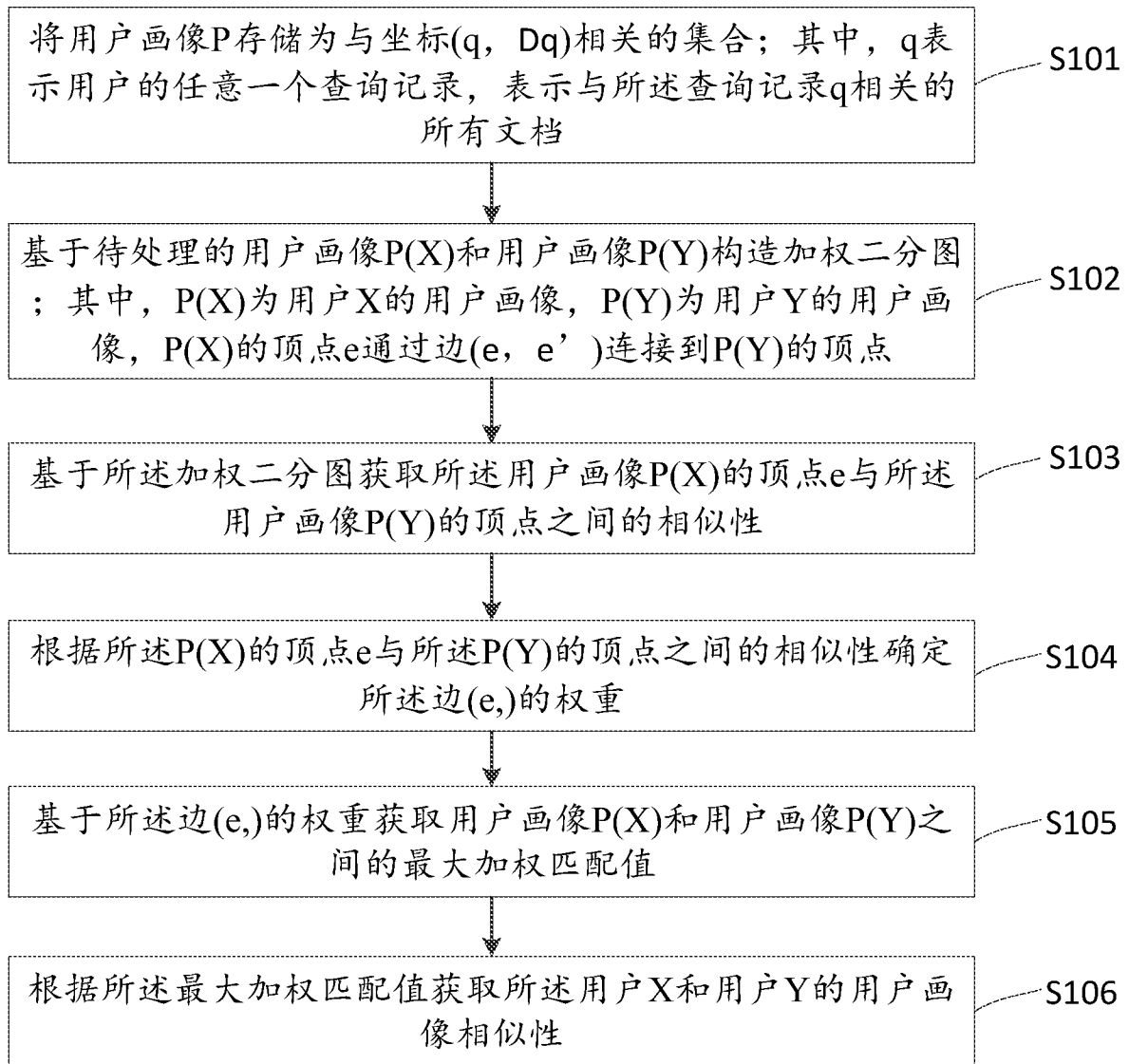
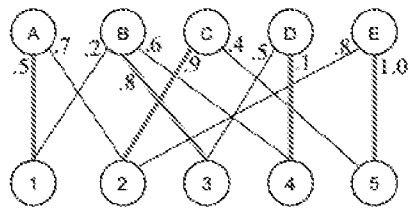
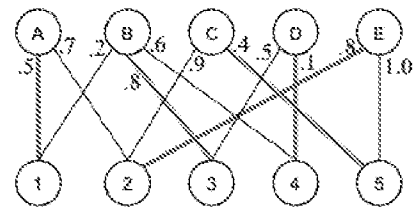


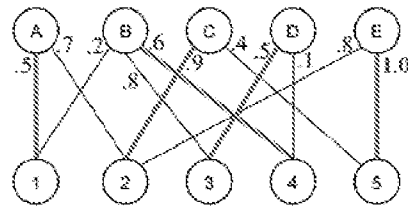
图 4



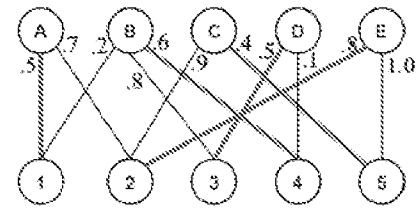
$$M = w(A,1) + w(B,3) + w(C,2) + w(D,4) + w(E,5) = 0.5 + 0.8 + 0.9 + 0.1 + 1.0 = 3.3$$



$$M = w(A,1) + w(B,3) + w(C,5) + w(D,4) + w(E,2) = 0.5 + 0.8 + 0.4 + 0.1 + 0.8 = 2.6$$



$$M = w(A,1) + w(B,4) + w(C,2) + w(D,3) + w(E,5) = 0.5 + 0.6 + 0.9 + 0.5 + 1.0 = 3.5$$



$$M = w(A,1) + w(B,4) + w(C,5) + w(D,3) + w(E,2) = 0.5 + 0.6 + 0.4 + 0.5 + 0.8 = 2.8$$

图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/117794

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/9536(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

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

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

CNPAT, CNKI, WPI, EPODOC, IEEE: 相似, 用户, 画像, 行为, 聚类, 社区, 群, 推荐, user?, similar+, behavior, cluster+, zone, group?, recommend+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 110111167 A (BEIJING JINGDONG SHANGKE INFORMATION TECHNOLOGY CO., LTD. et al.) 09 August 2019 (2019-08-09) claims 1-10	1, 7-11, 14, 20
Y	CN 110111167 A (BEIJING JINGDONG SHANGKE INFORMATION TECHNOLOGY CO., LTD. et al.) 09 August 2019 (2019-08-09) claims 1-10	2-6, 12-13, 15-19
Y	CN 101398854 A (TSINGHUA UNIVERSITY) 01 April 2009 (2009-04-01) description, page 5 paragraphs 6-11	2-6, 12-13, 15-19
A	KR 101752636 B1 (SKELTERS INC.) 03 July 2017 (2017-07-03) entire document	1-20
A	CN 102521659 A (BEIHANG UNIVERSITY) 27 June 2012 (2012-06-27) entire document	1-20



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

17 March 2020

Date of mailing of the international search report

01 April 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
Information on patent family members

International application No.

PCT/CN2019/117794

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	110111167	A	09 August 2019	None	
CN	101398854	A	01 April 2009	None	
KR	101752636	B1	03 July 2017	None	
CN	102521659	A	27 June 2012	None	

国际检索报告

国际申请号

PCT/CN2019/117794

A. 主题的分类

G06F 16/9536 (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

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))

CNPAT, CNKI, WPI, EPDOC, IEEE: 相似, 用户, 画像, 行为, 聚类, 社区, 群, 推荐, user?, similar+, behavior, cluster+, zone, group?, recommend+

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
X	CN 110111167 A (北京京东尚科信息技术有限公司 等) 2019年 8月 9日 (2019 - 08 - 09) 权利要求1-10	1, 7-11, 14, 20
Y	CN 110111167 A (北京京东尚科信息技术有限公司 等) 2019年 8月 9日 (2019 - 08 - 09) 权利要求1-10	2-6, 12-13, 15-19
Y	CN 101398854 A (清华大学) 2009年 4月 1日 (2009 - 04 - 01) 说明书第5页第6-11段	2-6, 12-13, 15-19
A	KR 101752636 B1 (SKELTERS INC.) 2017年 7月 3日 (2017 - 07 - 03) 全文	1-20
A	CN 102521659 A (北京航空航天大学) 2012年 6月 27日 (2012 - 06 - 27) 全文	1-20

☐ 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 3月 17日

国际检索报告邮寄日期

2020年 4月 1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

姚楠

电话号码 86-(10)-53961521

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/117794

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	110111167	A	2019年 8月 9日	无	
CN	101398854	A	2009年 4月 1日	无	
KR	101752636	B1	2017年 7月 3日	无	
CN	102521659	A	2012年 6月 27日	无	