

(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 9 月 30 日 (30.09.2021)



(10) 国际公布号
WO 2021/189583 A1

(51) 国际专利分类号:
G06F 16/9535 (2019.01)

(21) 国际申请号: PCT/CN2020/086202

(22) 国际申请日: 2020 年 4 月 22 日 (22.04.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010216557.4 2020年3月25日 (25.03.2020) CN

(71) 申请人: 江苏科技大学(JIANGSU UNIVERSITY OF SCIENCE AND TECHNOLOGY) [CN/CN]; 中国江苏省镇江市梦溪路2号, Jiangsu 212003 (CN)。

(72) 发明人: 暴琳(BAO, Lin); 中国江苏省镇江市梦溪路2号, Jiangsu 212003 (CN)。 田雨波(TIAN, Yubo); 中国江苏省镇江市梦溪路2号, Jiangsu 212003 (CN)。 张贞凯(ZHANG, Zhenkai); 中国江苏省镇江市梦溪路2号, Jiangsu 212003 (CN)。 王敏(WANG, Min); 中国江苏省镇江市梦溪路2号, Jiangsu 212003 (CN)。 马国军(MA, Guojun);

中国江苏省镇江市梦溪路2号, Jiangsu 212003 (CN)。 邓小乔(DENG, Xiaoqiao); 中国江苏省镇江市梦溪路2号, Jiangsu 212003 (CN)。

(74) 代理人: 南京苏高专利商标事务所(普通合伙)(NANJING SUGAO PATENT AND TRADEMARK FIRM(ORDINARY PARTNERSHIP)); 中国江苏省南京市白下区中山东路198号龙台国际大厦1912室, Jiangsu 210005 (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。

(54) Title: RESTRICTED BOLTZMANN MACHINE-DRIVEN INTERACTIVE PERSONALIZED SEARCH METHOD

(54) 发明名称: 基于受限玻尔兹曼机驱动的交互式个性化搜索方法

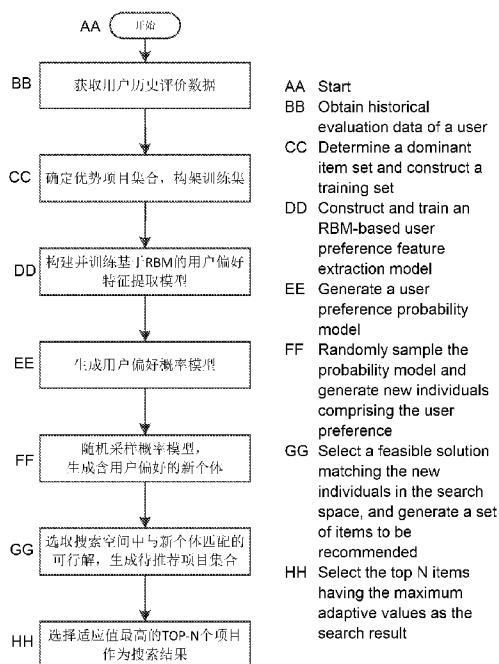


图 1

(57) Abstract: A restricted Boltzmann machine (RBM)-driven interactive personalized search method, comprising: 1. obtaining historical evaluation data of a user u, comprising all items that have been evaluated by the user u, scores and evaluation texts of the items, and performing vectorization on the evaluation texts; 2. constituting items having scores greater than a preset score threshold into a dominant item set D_u , and constructing a sample set; 3. constructing an RBM-based user preference feature extraction model; 4. constructing a user preference probability model; 5. generating N new items and setting category tags; 6. selecting N items having the highest category tag vector similarity with the N new items in the search space to constitute a set S_u of items to be recommended; 7. calculating adaptive values of the items in S_u ; and 8. selecting the top N items having the maximum adaptive values in S_u as the search result. The method can effectively guide the personalized search direction of a user and help the user search for the satisfactory solution as soon as possible, thereby improving the comprehensive performance of a personalized search algorithm.

(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: 一种基于受限玻尔兹曼机驱动的交互式个性化搜索方法, 包括: 1、获取用户 u 的历史评价数据, 包括用户 u 已评价的所有项目、对每个项目的评分和评价文本, 并将评价文本进行向量化; 2、将评分大于预设评分阈值的项目组成优势项目集合 Du , 并构建样本集; 3、构建基于RBM的用户偏好特征提取模型; 4、构建用户偏好概率模型; 5、生成 N 个新项目并设置类别标签; 6、在搜索空间中选取与 N 个新项目类别标签向量相似度最高的 N 个项目, 构成待推荐项目集合 Su ; 7、计算 Su 中每个项目的适应值; 8、选择 Su 中适应值最大的前 $TOPN$ 个项目作为搜索结果。该方法能够有效引导用户进行个性化搜索的方向, 尽快帮助用户搜索到用户满意解, 从而提高个性化搜索算法的综合性能。

基于受限玻尔兹曼机驱动的交互式个性化搜索方法

技术领域

本发明属于数据挖掘技术领域，具体涉及一种交互式个性化搜索方法。

背景技术

随着大数据、云计算等技术的迅猛发展，信息呈现爆炸式增长，各类互联网应用中聚集了大量用户生成内容，给用户带来新信息的同时，也增加了用户筛选有效信息并最终做出决策的难度。个性化搜索作为连接用户与信息的桥梁，能够引导用户进行快速搜索，尽可能帮助用户从海量信息中快速筛选出与用户兴趣相符的内容，有效缓解信息过载。

用户生成内容中包含来自诸多用户提供的信息，包括用户行为数据、用户信息、项目信息等，这些多源异构数据从不同角度反映了用户的兴趣偏好，充分利用这些信息，可有效提高个性化搜索的质量。由于用户生成内容具有的数据稀疏性、不完整性、动态性等复杂特性，因此，处理融合多源异构用户生成内容数据的个性化搜索问题得难度更大。交互式进化计算有效利用用户对优化问题的主观评价和决策，将人类智能评价信息与传统进化优化算法相结合，是解决融合多源异构数据个性化搜索问题的有效途径。

申请号为 201410202346 .X 的中国专利公开了一种用于心理学图书个性化快速搜索的交互式进化优化方法，指导用户快速寻找满意图书，该方法利用用户的搜索记录和群体共性搜索信息来建立用户偏好模型，没有充分挖掘用户历史数据中隐含的偏好特征。申请号为 201910151051 .7 的中国专利公开了一种个性化商品推荐方法，该方法根据对用户的分类来进行商品推荐，只能体现出不同类用户之间的个性差别，而无法体现具体用户的个性化。申请号为 201910431441 .X 的中国专利公开了一种基于交互数据聚类的个性化推荐方法及系统，其中的个性化推荐方法通过对用户聚类，得到相似用户的兴趣点，从而对用户进行个性化推荐，同样体现的是同类用户的偏好。

发明内容

发明目的：针对现有技术中存在的问题，本发明公开了一种基于受限玻尔兹曼机驱动的交互式个性化搜索方法，该方法能够引导当前用户进行个性化搜索。

技术方案：本发明采用如下技术方案：

基于受限玻尔兹曼机驱动的交互式个性化搜索方法，包括：

(1) 获取用户 u 的历史评价数据，所述历史评价数据包括用户 u 已评价的所有项目、对每个项目的评分和评价文本，并将评价文本进行向量化；

(2) 将评分大于预设评分阈值的项目组成优势项目集合 D_u ，构成样本集 $S_{RBM} = \{(u, \mathbf{x}_i, \mathbf{C}_i^u, \mathbf{T}_i^u)\}$ ，其中 $\mathbf{x}_i \in D_u$ ， $\mathbf{C}_i^u$ 为项目 $\mathbf{x}_i$ 的类别标签向量，长度为类别总数 n_1 ； $\mathbf{T}_i^u$ 为用户 u 对项目 $\mathbf{x}_i$ 的评价文本向量化表示，长度为 n_2 ； $i = 1, 2, \dots, |D_u|$ ， $|D_u|$ 表示 D_u 中的项目数目；

(3) 构建基于 RBM 的用户偏好特征提取模型，所述用户偏好特征提取模型用于根据优势项目集合 D_u 提取用户的偏好特征；

(4) 构建用户偏好概率模型，所述用户偏好概率模型用于表示用户对每个类别项目的偏好概率；

(5) 设定种群大小 N ，采用分布估计算法生成 N 个新项目并设置每个项目的类别标签；

(6) 在搜索空间中选取分别与 N 个新项目类别标签向量 $\mathbf{C}_n^{new}$ 相似度最高的 N 个项目，构成待推荐项目集合 S^u ；

(7) 计算待推荐项目集合 S^u 中每个项目的适应值：

(7.1) 估计用户 u 对 S^u 中每个项目的个性化偏好评分 $\hat{R}_{ux}$ ：

$$\hat{R}_{ux} = \frac{\max(E_\theta) - E_\theta(\mathbf{x} | u^*)}{\max(E_\theta) - \min(E_\theta)}, \quad \mathbf{x} \in S^u$$

其中， $E_\theta(\mathbf{x} | u^*)$ 为与当前用户 u 相似度大于阈值的用户 u^* 对项目 $\mathbf{x}$ 的评价能量函数； $\max(E_\theta)$ 和 $\min(E_\theta)$ 分别为待推荐项目集合 S^u 的所有项目中获得的能量函数的最大和最小值；

(7.2) 计算 S^u 中每个项目的适应值：

$$\hat{f}_u(\mathbf{x}) = \sigma(\hat{R}_{ux})$$

其中 $\sigma(x) = 1/(1 + \exp(-x))$ 是归一化函数；

(8) 选择 S'' 中适应值最大的前 TOP_N 个项目作为搜索结果。

有益效果：本发明公开的交互式个性化搜索方法充分利用用户生成内容的多源异构数据，包括用户评分、文本标签、文本评论等信息，构建反映用户偏好特征的用户偏好特征提取模型，基于此模型，构建用户偏好概率模型，通过构建的模型，生成包含用户偏好的新项目，选择其中最能匹配用户偏好的项目作为搜索结果。该方法能够更好地解决多源异构复杂数据环境下的用户个性化搜索问题，能够有效引导用户进行个性化搜索的方向，尽快帮助用户搜索到用户满意解，从而提高个性化搜索算法的综合性能。

附图说明

图1为本发明公开的交互式个性化搜索方法的流程图；

图2为基于受限玻尔兹曼机的用户偏好模型的结构图。

具体实施方式

下面结合附图和具体实施方式，进一步阐明本发明。

如图1所示，本发明公开了一种基于受限玻尔兹曼机(Restricted Boltzmann Machine, RBM)驱动的交互式个性化搜索方法，包括：

步骤1、获取用户 u 的历史评价数据，所述历史评价数据包括用户 u 已评价的所有项目、对每个项目的评分和评价文本，并将评价文本进行向量化；

本实施例中评价文本进行向量化的步骤为：首先去除文本中的标点符号、停用词，其次采用文献：Le Q, Mikolov T. Distributed representations of sentences and documents[C]//International conference on machine learning. 2014: 1188-1196 中的 doc2vec 文本向量化表示模型，将用户的评价文本进行向量化。

步骤2、将评分大于预设评分阈值的项目组成优势项目集合 D_u ；

构建样本集 $S_{RBM} = \{(u, \mathbf{x}_i, \mathbf{C}_i'', \mathbf{T}_i'')\}$ ，其中 $\mathbf{x}_i \in D_u$ ， $\mathbf{C}_i''$ 为项目 $\mathbf{x}_i$ 的类别标签向量，长度为类别总数 n_1 ； $\mathbf{C}_i''$ 中每个元素 $c_{i,j}''$ 为二值变量； $c_{i,j}''=1$ 表示项目 $\mathbf{x}_i$ 具有第 j 类标签， $j=1,2,\dots,n_1$ ；值得注意的是，本发明中不同类别之间并不是互斥的，一个项目可以同时存在多个类别标签为1。比如，对于书籍搜索，即项目为书籍的情况，一本书可能有多个类别标签，如《张居正传》，可以同时有人物传记、中国历史这几个标签，那么对应的类别标签就为1；对于影视搜索，即项目

为影视作品的情况，一个影视作品也可能有多个类别标签，如《三国演义》可以同时有电视剧、历史这几个类别标签，同样地，对应的类别标签为 1。

$\mathbf{T}_i^u$ 为用户 u 对项目 $\mathbf{x}_i$ 的评价文本向量化表示，长度为 n_2 ； $i=1,2,\dots,|D_u|$ ， $|D_u|$ 表示 D_u 中的项目数目；

步骤 3、构建基于 RBM 的用户偏好特征提取模型，如图 2 所示，用户偏好特征提取模型包括第一可见层 $\mathbf{v}_1$ 、第二可见层 $\mathbf{v}_2$ 、隐藏层 $\mathbf{h}$ ；所述第一可见层有 n_1 个可见单元，每个可见单元为二值变量；第二可见层有 n_2 个可见单元，每个可见单元为实数；所述隐藏层有 m 个隐单元，每个隐单元为实数；所述用户偏好特征提取模型的参数为 $\theta = \{\mathbf{w}^1, \mathbf{a}^1, \mathbf{w}^2, \mathbf{a}^2, \mathbf{b}\}$ ，其中， $\mathbf{w}^1$ 、 $\mathbf{w}^2$ 分别表示第一可见层中可见单元与隐藏层中隐单元、第二可见层中可见单元与隐藏层中隐单元之间的连接权重； $\mathbf{a}^1$ 、 $\mathbf{a}^2$ 分别为第一可见层中可见单元、第二可见层中可见单元的偏置； $\mathbf{b}$ 为隐藏层中隐单元的偏置；隐藏层中隐单元个数 m 为类别总数 n_1 的 0.8-1.2 倍，本实施例中，设置 $m = \lceil 1.1 * n_1 \rceil$ ， $\lceil \cdot \rceil$ 为向上取整运算。

将样本集中每个样本的 $\mathbf{C}_i^u$ 和 $\mathbf{T}_i^u$ 输入到用户偏好特征提取模型的第一可见层和第二可见层，采用对比散度学习算法对所述模型进行训练，优化模型参数 θ ，得到训练好的用户偏好特征提取模型，其中隐单元的状态表征了当前用户 u 的偏好特征。

步骤 4、基于训练好的用户偏好特征提取模型，构建用户偏好概率模型：

$$P_u(\mathbf{x}) = [p(c_1=1), p(c_2=1), \dots, p(c_{n_1}=1)];$$

其中 $p(c_j=1)$ 表示用户 u 偏好的项目具有第 j 类标签项目的概率，计算步骤为：

(4.1) 估计用户偏好具有第 j 类标签项目的绝对概率：

$$\hat{p}(c_j=1) = \frac{\sum_{i=1}^{|D_u|} \delta_i(c_j^+) + \phi}{\sum_{i=1}^{|D_u|} \delta_i(c_j^+) + \sum_{i=1}^{|D_u|} \delta_i(c_j^-) + 2\phi}$$

其中 $\delta_i(c_j^+)$ 为项目 $\mathbf{x}_i$ 的第 j 个类别标签 $c_j=1$ 的边际分布, $\delta_i(c_j^-)$ 为项目 $\mathbf{x}_i$ 的第 j 个类别标签 $c_j=0$ 的边际分布, ϕ 为 S_{RBM} 中所有样本的平均边际分布;

$$\delta_i(c_j^+) = \begin{cases} \sum_{l=1}^m e^{-E_\theta(C_i^u, l)}, & c_{i,j}^u = 1 \\ 0, & c_{i,j}^u = 0 \end{cases},$$

$$\delta_i(c_j^-) = \begin{cases} \sum_{l=1}^m e^{-E_\theta(C_i^u, l)}, & c_{i,j}^u = 0 \\ 0, & c_{i,j}^u = 1 \end{cases},$$

$$\phi = \left(\sum_{i=1}^{|D_u|} \delta_i(c_j) \right) / |D_u|,$$

$$\delta_i(c_j) = (\delta_i(c_j^+) + \delta_i(c_j^-)) / 2,$$

其中 $c_{i,j}^u$ 为向量 $\mathbf{C}_i^u$ 的第 j 个元素, $E_\theta(C_i^u, l)$ 为第 i 个样本中项目的类别标签向量 $\mathbf{C}_i^u$ 在第 l 个隐单元的能量值, 其计算式为:

$$E_\theta(C_i^u, l) = -\sum_{j=1}^{n_l} a_j^l c_{i,j}^u - \ln \left(1 + e^{(b_l + \sum_{j=1}^{n_l} w_{jl}^l c_{i,j}^u)} \right)$$

例如, 项目有 2 个类别标签, 即 $n_1=2$, 有 3 个样本, $i=1,2,3$, 类别标签向量分别为: $C_1=(1,1)$ 、 $C_2=(1,0)$ 、 $C_3=(0,1)$;

则计算 $\hat{p}(c_1=1)$ 时, 需要计算 $\delta_i(c_1^+)$ 和 $\delta_i(c_1^-)$, 计算 $\delta_i(c_1^+)$ 时, 只考虑 $c_1=1$ 的样本, 即第一个、第二个样本, $\delta_3(c_1^+)=0$; 计算 $\delta_i(c_1^-)$ 时, 只考虑 $c_1=0$ 的样本, 即第三个样本, $\delta_1(c_1^-)=\delta_2(c_1^-)=0$ 。

(4.2) 对用户偏好具有第 j 类标签项目的绝对概率进行归一化:

$$p(c_j=1) = \frac{\hat{p}(c_j=1)}{\sum_{j=1}^{n_l} \hat{p}(c_j=1)}$$

得到的 $P_u(\mathbf{x})$ 即为用户 u 的用户偏好概率模型; $p(c_j=1)$ 值越大, 则用户越偏好具有第 j 类标签的项目;

步骤 5、设定种群大小 N , 采用分布估计算法 (Estimation of Distribution

Algorithms, EDA)生成 N 个新项目, 新项目的类别标签向量 $\mathbf{C}_n^{new}$ ($n=1, 2, \dots, N$) 的设置步骤如下:

(5.1) 令 $n=1$;

(5.2) 生成 $[0,1]$ 之间的随机数 z ; 如果 $z \leq p(c_j=1)$, 则第 n 个新项目的类别标签向量 $\mathbf{C}_n^{new}$ 的第 j 个元素为 1, 否则为 0;

(5.3) 令 n 加一, 重复步骤 (5.2), 直至 $n > N$;

需要说明的是, 此时这 N 个新项目只确定了类别标签向量, 具体项目内容还未确定;

步骤 6、在搜索空间中选取分别与 N 个新项目类别标签向量 $\mathbf{C}_n^{new}$ 相似度最高的 N 个项目, 构成待推荐项目集合 $\mathbf{S}''$; 本实施例中, 采用欧氏距离作为相似度的计算, 即两向量之间的欧氏距离越小, 二者相似度越高;

此时, N 个新项目的内容确定了, 下面步骤对其做进一步的筛选;

步骤 7、计算待推荐项目集合 $\mathbf{S}''$ 中每个项目的适应值:

(7.1) 估计用户 u 对 $\mathbf{S}''$ 中每个项目的个性化偏好评分 $\hat{R}_{ux}$:

$$\hat{R}_{ux} = \frac{\max(E_\theta) - E_\theta(\mathbf{x} | u^*)}{\max(E_\theta) - \min(E_\theta)}$$

其中, $E_\theta(\mathbf{x} | u^*)$ 为与当前用户 u 相似度大于阈值的用户 u^* 对项目 $\mathbf{x}$ 的评价能量函数, 其计算式为:

$$E_\theta(\mathbf{x} | u^*) = -\sum_{j=1}^{n_1} \alpha_j^1 c_j - \sum_{k=1}^{n_2} \alpha_k^2 t_k - \sum_{l=1}^m \ln \left(1 + e^{\left(b_l + \sum_{j=1}^{n_1} w_{jl}^1 c_j + \sum_{k=1}^{n_2} w_{kl}^2 t_k \right)} \right)$$

其中 c_j 为 $\mathbf{x}$ 类别标签向量中第 j 个元素; t_k 为用户 u^* 对 $\mathbf{x}$ 的评价文本向量化的第 k 个元素; α_j^1 为用户偏好特征提取模型中第一可见层中可见单元的偏置 $\mathbf{a}^1$ 的第 j 个元素; α_k^2 为用户偏好特征提取模型中第二可见层中可见单元的偏置 $\mathbf{a}^2$ 的第 k 个元素; b_l 为用户偏好特征提取模型中隐藏层中隐单元的偏置 $\mathbf{b}$ 的第 l 个元素; w_{jl}^1 为第一可见层中第 j 个可见单元与隐藏层中第 l 个隐单元之间的连接权重, w_{kl}^2 为第二可见层中第 k 个可见单元与隐藏层中第 l 个隐单元之间的连接权重;

从 $E_\theta(\mathbf{x}|u^*)$ 的计算式可以看出，项目 $\mathbf{x}$ 能量函数的大小与用户对其的评价文本有关，因此本步骤需要选择与当前用户 u 相似的用户，当前用户 u 和所有用户中任一用户 u' ($u' \neq u$) 的相似度 $Sim(u, u')$ 计算式为：

$$Sim(u, u') = \frac{\sum_{\mathbf{x}' \in I_{u, u'}} (R_{u\mathbf{x}'} - \bar{R}_u)(R_{u'\mathbf{x}'} - \bar{R}_{u'})}{\sqrt{\sum_{\mathbf{x}' \in I_{u, u'}} (R_{u\mathbf{x}'} - \bar{R}_u)^2 \times \sum_{\mathbf{x}' \in I_{u, u'}} (R_{u'\mathbf{x}'} - \bar{R}_{u'})^2}}$$

其中 $I_{u, u'}$ 表示用户 u 和 u' 均已评分的项目集合； $R_{u\mathbf{x}'}$ 为用户 u 对 $I_{u, u'}$ 中的项目 $\mathbf{x}'$ 的评分， $R_{u'\mathbf{x}'}$ 为用户 u' 对 $\mathbf{x}'$ 的评分； $\bar{R}_u$ 为用户 u 对已被 u 评价的所有项目的平均评分； $\bar{R}_{u'}$ 为用户 u' 对已被 u' 评价的所有项目的平均评分。

$\max(E_\theta)$ 和 $\min(E_\theta)$ 分别为待推荐项目集合 S'' 的所有项目中获得的能量函数的最大和最小值；

(7.2) 计算 S'' 中每个项目的适应值：

为了充分利用用户的个性和用户所在的群体，本实施例中适应值包括用户 u 的个性化评价，以及与用户 u 相似的用户的群体评价，具体计算步骤为：

(7.2.1) 在所有用户中选择与用户 u 相似度最高的前 K 个用户，构成用户 u 的邻居用户集 S_{nb} ，计算 S_{nb} 对 S'' 中项目 $\mathbf{x}$ 的权重平均评分：

$$social(\mathbf{x}) = \sum_{u' \in S_{nb}} Sim(u, u') \times R_{u'\mathbf{x}}$$

其中 $R_{u'\mathbf{x}}$ 为 S_{nb} 中的用户 u' 对 S'' 中项目 $\mathbf{x}$ 的评分；如果用户 u' 对项目 $\mathbf{x}$ 没有评分，则 $R_{u'\mathbf{x}}=0$ ；

(7.2.2) S'' 中项目 $\mathbf{x}$ 的适应值为：

$$\hat{f}_u(\mathbf{x}) = \sigma(\alpha \times \hat{R}_{u\mathbf{x}} + (1 - \alpha) \times social(\mathbf{x}))$$

其中 α 是用来调整社交知识对于个体预测评分贡献度的系数，如果不考虑群体评价，将 α 设为 1 即可。 $\sigma(x) = 1/(1 + \exp(-x))$ 是归一化函数；

步骤 8、选择 S'' 中适应值最大的前 TOP_N 个项目作为搜索结果。

由于用户兴趣偏好的多样性和动态演化特性,在交互式个性化搜索过程的早期阶段,优势项目集合中所含当前用户的偏好信息不够充分,利用训练的基于RBM的用户偏好特征提取模型所提取的用户的偏好特征较粗略。同时,用户对项目的认知和偏好具有主观性、模糊性、渐进性和不确定性,在用户进行个性化搜索过程中随着环境变化、信息量增加等因素的影响,用户需求和兴趣偏好逐渐清晰甚至有可能发生变化。因此,为了获得准确且动态变化的用户偏好特征,随着用户交互式搜索过程的推进和用户行为动态演变,根据当前用户最近的评价数据,再次训练融合了多源异构数据的基于受限玻尔兹曼机的用户偏好特征提取模型,动态更新提取的用户偏好特征,同时,更新用户偏好概率模型,及时跟踪当前用户的偏好,有效引导交互式个性化搜索过程的前进方向,保证用户顺利完成个性化搜索。

权利要求书

1、基于受限玻尔兹曼机驱动的交互式个性化搜索方法，其特征在于，包括：

(1) 获取用户 u 的历史评价数据，所述历史评价数据包括用户 u 已评价的所有项目、对每个项目的评分和评价文本，并将评价文本进行向量化；

(2) 将评分大于预设评分阈值的项目组成优势项目集合 D_u ，构成样本集 $S_{RBM} = \{(u, \mathbf{x}_i, \mathbf{C}_i^u, \mathbf{T}_i^u)\}$ ，其中 $\mathbf{x}_i \in D_u$ ， $\mathbf{C}_i^u$ 为项目 $\mathbf{x}_i$ 的类别标签向量，长度为类别总数 n_1 ； $\mathbf{T}_i^u$ 为用户 u 对项目 $\mathbf{x}_i$ 的评价文本向量化表示，长度为 n_2 ； $i = 1, 2, \dots, |D_u|$ ， $|D_u|$ 表示 D_u 中的项目数目；

(3) 构建基于 RBM 的用户偏好特征提取模型，所述用户偏好特征提取模型用于根据优势项目集合 D_u 提取用户的偏好特征；

(4) 构建用户偏好概率模型，所述用户偏好概率模型用于表示用户对每个类别项目的偏好概率；

(5) 设定种群大小 N ，采用分布估计算法生成 N 个新项目并设置每个项目的类别标签；

(6) 在搜索空间中选取分别与 N 个新项目类别标签向量 $\mathbf{C}_n^{new}$ 相似度最高的 N 个项目，构成待推荐项目集合 S^u ；

(7) 计算待推荐项目集合 S^u 中每个项目的适应值：

(7.1) 估计用户 u 对 S^u 中每个项目的个性化偏好评分 $\hat{R}_{ux}$ ：

$$\hat{R}_{ux} = \frac{\max(E_\theta) - E_\theta(\mathbf{x} | u^*)}{\max(E_\theta) - \min(E_\theta)}, \quad \mathbf{x} \in S^u$$

其中， $E_\theta(\mathbf{x} | u^*)$ 为与当前用户 u 相似度大于阈值的用户 u^* 对项目 $\mathbf{x}$ 的评价能量函数； $\max(E_\theta)$ 和 $\min(E_\theta)$ 分别为待推荐项目集合 S^u 的所有项目中获得的能量函数的最大和最小值；

(7.2) 计算 S^u 中每个项目的适应值：

$$\hat{f}_u(\mathbf{x}) = \sigma(\hat{R}_{ux})$$

其中 $\sigma(x) = 1/(1 + \exp(-x))$ 是归一化函数；

(8) 选择 S^u 中适应值最大的前 TOP_N 个项目作为搜索结果。

2、根据权利要求 1 所述的交互式个性化搜索方法，其特征在于，所述用户偏好特征提取模型包括第一可见层、第二可见层、隐藏层；所述第一可见层有 n_1 个可见单元，每个可见单元为二值变量；第二可见层有 n_2 个可见单元，每个可见单元为实数；所述隐藏层有 m 个隐单元，每个隐单元为实数；所述用户偏好特征提取模型的参数为 $\theta = \{\mathbf{w}^1, \mathbf{a}^1, \mathbf{w}^2, \mathbf{a}^2, \mathbf{b}\}$ ，其中， $\mathbf{w}^1$ 、 $\mathbf{w}^2$ 分别表示第一可见层中可见单元与隐藏层中隐单元、第二可见层中可见单元与隐藏层中隐单元之间的连接权重； $\mathbf{a}^1$ 、 $\mathbf{a}^2$ 分别为第一可见层中可见单元、第二可见层中可见单元的偏置； $\mathbf{b}$ 为隐藏层中隐单元的偏置；

将样本集中每个样本的 $\mathbf{C}_i''$ 和 $\mathbf{T}_i''$ 输入到用户偏好特征提取模型的第一可见层和第二可见层，采用对比散度学习算法对所述模型进行训练，优化模型参数 θ ，得到训练好的用户偏好特征提取模型。

3、根据权利要求 2 所述的交互式个性化搜索方法，其特征在于，所述用户偏好概率模型为：

$$P_u(\mathbf{x}) = [p(c_1=1), p(c_2=1), \dots, p(c_{n_1}=1)];$$

其中 $p(c_j=1)$ 表示用户 u 偏好的项目具有第 j 类标签项目的概率。

4、根据权利要求 3 所述的交互式个性化搜索方法，其特征在于， $p(c_j=1)$ 的计算步骤为：

(4.1) 估计用户偏好具有第 j 类标签项目的绝对概率：

$$\hat{p}(c_j=1) = \frac{\sum_{i=1}^{|D_u|} \delta_i(c_j^+) + \phi}{\sum_{i=1}^{|D_u|} \delta_i(c_j^+) + \sum_{i=1}^{|D_u|} \delta_i(c_j^-) + 2\phi}$$

其中 $\delta_i(c_j^+)$ 为项目 $\mathbf{x}_i$ 的第 j 个类别标签 $c_j=1$ 的边际分布， $\delta_i(c_j^-)$ 为项目 $\mathbf{x}_i$ 的第 j 个类别标签 $c_j=0$ 的边际分布， ϕ 为 S_{RBM} 中所有样本的平均边际分布；

$$\delta_i(c_j^+) = \begin{cases} \sum_{l=1}^m e^{-E_{\theta}(\mathbf{C}_i'', l)}, & c_{i,j}'' = 1 \\ 0, & c_{i,j}'' = 0 \end{cases},$$

$$\delta_i(c_j^-) = \begin{cases} \sum_{l=1}^m e^{-E_\theta(C_i^u, l)}, & c_{i,j}^u = 0 \\ 0, & c_{i,j}^u = 1 \end{cases},$$

$$\phi = \left(\sum_{i=1}^{|D_u|} \delta_i(c_j) \right) / |D_u|, \quad \delta_i(c_j) = (\delta_i(c_j^+) + \delta_i(c_j^-)) / 2$$

其中 $c_{i,j}^u$ 为向量 $\mathbf{C}_i^u$ 的第 j 个元素, $E_\theta(C_i^u, l)$ 为第 i 个样本中项目的类别标签

向量 $\mathbf{C}_i^u$ 在第 l 个隐单元的能量值, 其计算式为:

$$E_\theta(C_i^u, l) = -\sum_{j=1}^{n_1} a_j^1 c_{i,j}^u - \ln \left(1 + e^{\left(b_l + \sum_{j=1}^{n_1} w_{jl}^1 c_{i,j}^u \right)} \right)$$

(4.2) 对用户偏好具有第 j 类标签项目的绝对概率进行归一化:

$$p(c_j = 1) = \frac{\hat{p}(c_j = 1)}{\sum_{j=1}^{n_1} \hat{p}(c_j = 1)}$$

得到 $P_u(\mathbf{x})$ 即为用户 u 的用户偏好概率模型。

5、根据权利要求 2 所述的交互式个性化搜索方法, 其特征在于, 用户 u^* 对项目 $\mathbf{x}$ 的评价能量函数 $E_\theta(\mathbf{x} | u^*)$ 的计算式为:

$$E_\theta(\mathbf{x} | u^*) = -\sum_{j=1}^{n_1} a_j^1 c_j - \sum_{k=1}^{n_2} a_k^2 t_k - \sum_{l=1}^m \ln \left(1 + e^{\left(b_l + \sum_{j=1}^{n_1} w_{jl}^1 c_j + \sum_{k=1}^{n_2} w_{kl}^2 t_k \right)} \right)$$

其中 c_j 为 $\mathbf{x}$ 类别标签向量中第 j 个元素; t_k 为用户 u^* 对 $\mathbf{x}$ 的评价文本向量化的第 k 个元素; a_j^1 为用户偏好特征提取模型中第一可见层中可见单元的偏置 $\mathbf{a}^1$ 的第 j 个元素; a_k^2 为用户偏好特征提取模型中第二可见层中可见单元的偏置 $\mathbf{a}^2$ 的第 k 个元素; b_l 为用户偏好特征提取模型中隐藏层中隐单元的偏置 $\mathbf{b}$ 的第 l 个元素; w_{jl}^1 为第一可见层中第 j 个可见单元与隐藏层中第 l 个隐单元之间的连接权重, w_{kl}^2 为第二可见层中第 k 个可见单元与隐藏层中第 l 个隐单元之间的连接权重。

6、根据权利要求 1 所述的交互式个性化搜索方法, 其特征在于, 所述步骤

(5) 中新项目的类别标签向量 $\mathbf{C}_n^{new}$ 的设置步骤如下:

(5.1) 令 $n = 1$;

(5.2) 生成 $[0,1]$ 之间的随机数 z ；如果 $z \leq p(c_j=1)$ ，则第 n 个新项目的类别标签向量 $\mathbf{C}_n^{new}$ 的第 j 个元素为1，否则为0；

(5.3) 令 n 加一，重复步骤(5.2)，直至 $n > N$ 。

7、根据权利要求1所述的交互式个性化搜索方法，其特征在于，所述步骤(7.2)中每个项目的适应值还包括群组估计评分，具体计算步骤为：

(7.2.1) 在所有用户中选择与用户 u 相似度最高的前 K 个用户，构成用户 u 的邻居用户集 S_{nb} ，计算 S_{nb} 对 S^u 中项目 $\mathbf{x}$ 的权重平均评分：

$$social(\mathbf{x}) = \sum_{u' \in S_{nb}} Sim(u, u') \times R_{u'\mathbf{x}}$$

其中 $R_{u'\mathbf{x}}$ 为 S_{nb} 中的用户 u' 对 S^u 中项目 $\mathbf{x}$ 的评分；如果用户 u' 对项目 $\mathbf{x}$ 没有评分，则 $R_{u'\mathbf{x}}=0$ ；

(7.2.2) S^u 中项目 $\mathbf{x}$ 的适应值为：

$$\hat{f}_u(\mathbf{x}) = \sigma(\alpha \times \hat{R}_{u\mathbf{x}} + (1-\alpha) \times social(\mathbf{x}))$$

其中 α 是用来调整社交知识对于个体预测评分贡献度的系数。

8、根据权利要求2所述的交互式个性化搜索方法，其特征在于，所述隐藏层中隐单元个数 m 为类别总数 n_1 的0.8-1.2倍。

9、根据权利要求1所述的交互式个性化搜索方法，其特征在于，对于与用户 u 存在共同评分项目的用户 u' ， $u' \neq u$ ， u 和 u' 的相似度 $Sim(u, u')$ 为：

$$Sim(u, u') = \frac{\sum_{\mathbf{x}' \in I_{u, u'}} (R_{u\mathbf{x}'} - \bar{R}_u)(R_{u'\mathbf{x}'} - \bar{R}_{u'})}{\sqrt{\sum_{\mathbf{x}' \in I_{u, u'}} (R_{u\mathbf{x}'} - \bar{R}_u)^2 \times \sum_{\mathbf{x}' \in I_{u, u'}} (R_{u'\mathbf{x}'} - \bar{R}_{u'})^2}}$$

其中 $I_{u, u'}$ 表示用户 u 和 u' 均评分的项目集合； $R_{u\mathbf{x}'}$ 为用户 u 对 $I_{u, u'}$ 中的项目 $\mathbf{x}'$ 的评分， $R_{u'\mathbf{x}'}$ 为用户 u' 对 $\mathbf{x}'$ 的评分； $\bar{R}_u$ 为用户 u 对已评价的所有项目的平均评分； $\bar{R}_{u'}$ 为用户 u' 对已评价的所有项目的平均评分。

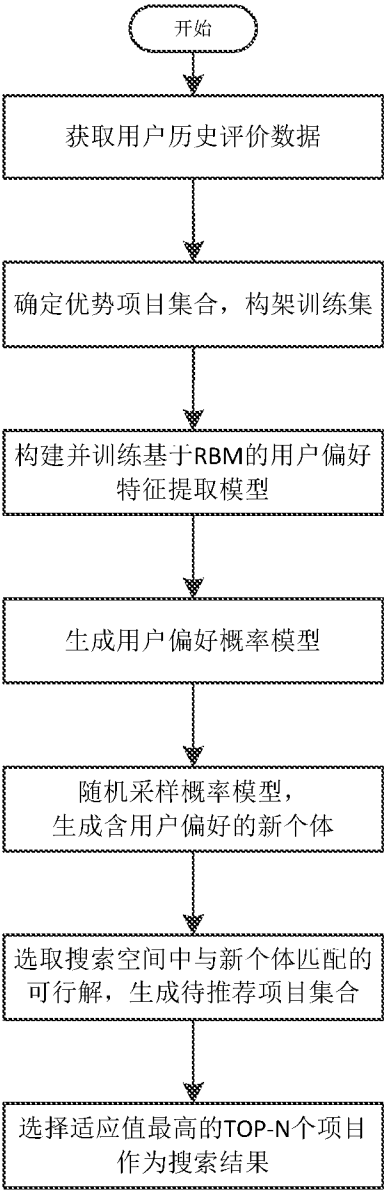


图 1

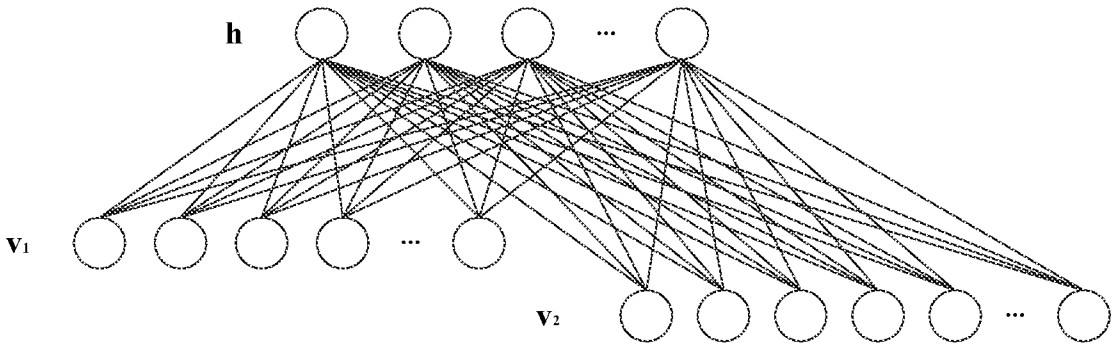


图 2

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/086202

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

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, WPI, EPODOC, CNKI, IEEE, GOOGLE: 受限玻尔兹曼机, 概率, 模型, 能量, 分布估计算法, 相似度, Restricted Boltzmann Machine, RBM, probability, model, energy, Estimation of Distribution Algorithms, EDA, similarity

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Lin Bao et al. "Restricted Boltzmann Machine-Assisted Estimation of Distribution Algorithm for Complex Problems" 《Complexity》, Vol. no. 2018, 01 November 2018 (2018-11-01), pp. 3-6	1-9
A	CN 106777359 A (ANHUI AGRICULTURAL UNIVERSITY) 31 May 2017 (2017-05-31) entire document	1-9
A	CN 110162706 A (NANJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 23 August 2019 (2019-08-23) entire document	1-9
A	US 2016188726 A1 (TCL RESEARCH AMERICA INC.) 30 June 2016 (2016-06-30) entire document	1-9



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

02 December 2020

Date of mailing of the international search report

21 December 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/CN2020/086202

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	106777359	A	31 May 2017	None			
CN	110162706	A	23 August 2019	None			
US	2016188726	A1	30 June 2016	CN	105677780	A	15 June 2016

国际检索报告

国际申请号

PCT/CN2020/086202

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 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, WPI, EPDOC, CNKI, IEEE, GOOGLE: 受限玻尔兹曼机, 概率, 模型, 能量, 分布估计算法, 相似度, Restricted Boltzmann Machine, RBM, probability, model, energy, Estimation of Distribution Algorithms, EDA, similarity																	
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>A</td> <td>Lin Bao 等. "Restricted Boltzmann Machine-Assisted Estimation of Distribution Algorithm for Complex Problems" 《Complexity》, 第2018卷, 2018年 11月 1日 (2018 - 11 - 01), 第3-6页</td> <td>1-9</td> </tr> <tr> <td>A</td> <td>CN 106777359 A (安徽农业大学) 2017年 5月 31日 (2017 - 05 - 31) 全文</td> <td>1-9</td> </tr> <tr> <td>A</td> <td>CN 110162706 A (南京邮电大学) 2019年 8月 23日 (2019 - 08 - 23) 全文</td> <td>1-9</td> </tr> <tr> <td>A</td> <td>US 2016188726 A1 (TCL RESEARCH AMERICA INC.) 2016年 6月 30日 (2016 - 06 - 30) 全文</td> <td>1-9</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	A	Lin Bao 等. "Restricted Boltzmann Machine-Assisted Estimation of Distribution Algorithm for Complex Problems" 《Complexity》, 第2018卷, 2018年 11月 1日 (2018 - 11 - 01), 第3-6页	1-9	A	CN 106777359 A (安徽农业大学) 2017年 5月 31日 (2017 - 05 - 31) 全文	1-9	A	CN 110162706 A (南京邮电大学) 2019年 8月 23日 (2019 - 08 - 23) 全文	1-9	A	US 2016188726 A1 (TCL RESEARCH AMERICA INC.) 2016年 6月 30日 (2016 - 06 - 30) 全文	1-9
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
A	Lin Bao 等. "Restricted Boltzmann Machine-Assisted Estimation of Distribution Algorithm for Complex Problems" 《Complexity》, 第2018卷, 2018年 11月 1日 (2018 - 11 - 01), 第3-6页	1-9															
A	CN 106777359 A (安徽农业大学) 2017年 5月 31日 (2017 - 05 - 31) 全文	1-9															
A	CN 110162706 A (南京邮电大学) 2019年 8月 23日 (2019 - 08 - 23) 全文	1-9															
A	US 2016188726 A1 (TCL RESEARCH AMERICA INC.) 2016年 6月 30日 (2016 - 06 - 30) 全文	1-9															
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																	
<table border="0"> <tr> <td style="vertical-align: top;"> * 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 </td> <td style="vertical-align: top;"> "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 </td> </tr> </table>			* 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													
* 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 12月 2日		国际检索报告邮寄日期 2020年 12月 21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		授权官员 孙韬敏 电话号码 86-(10)-53961420															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2020/086202

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	106777359	A	2017年 5月 31日	无	
CN	110162706	A	2019年 8月 23日	无	
US	2016188726	A1	2016年 6月 30日	CN 105677780 A	2016年 6月 15日