



## (12)发明专利申请

(10)申请公布号 CN 109508407 A

(43)申请公布日 2019.03.22

(21)申请号 201910032026.7

(22)申请日 2019.01.14

(71)申请人 上海电机学院

地址 200240 上海市闵行区江川路690号

(72)发明人 孔元元 白智远 张飒 吕品

(74)专利代理机构 上海伯瑞杰知识产权代理有限公司 31227

代理人 李庆

(51)Int.Cl.

G06F 16/78(2019.01)

G06F 16/735(2019.01)

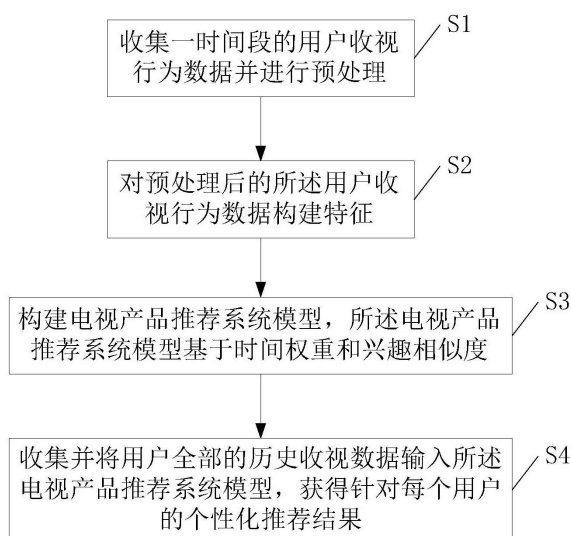
权利要求书2页 说明书6页 附图1页

### (54)发明名称

融合时间与兴趣相似度的电视产品推荐方法

### (57)摘要

本发明提供一种融合时间与兴趣相似度的电视产品推荐方法,包括步骤:S1:收集一时间段的用户收视行为数据并进行预处理,所述用户收视行为数据包括用户收视数据、用户订阅数据和用户点播数据;S2:对预处理后的所述用户收视行为数据构建特征;S3:构建电视产品推荐系统模型,所述电视产品推荐系统模型基于时间权重和兴趣相似度;S4:收集并将用户全部的历史收视数据输入所述电视产品推荐系统模型,获得针对每个用户的个性化推荐结果。本发明的一种融合时间与兴趣相似度的电视产品推荐方法,能够更为精确地对用户进行电视产品个性化推荐。



1. 一种融合时间与兴趣相似度的电视产品推荐方法,包括步骤:

S1: 收集一时间段的用户收视行为数据并进行预处理,所述用户收视行为数据包括用户收视数据、用户订阅数据和用户点播数据;

S2: 对预处理后的所述用户收视行为数据构建特征;

S3: 构建电视产品推荐系统模型,所述电视产品推荐系统模型基于时间权重和兴趣相似度;

S4: 收集并将用户全部的历史收视数据输入所述电视产品推荐系统模型,获得针对每个用户的个性化推荐结果。

2. 根据权利要求1所述的融合时间与兴趣相似度的电视产品推荐方法,其特征在于,所述预处理步骤包括:利用PHP对所述用户收视行为数据进行剔除异常值、去重和归一化。

3. 根据权利要求2所述的融合时间与兴趣相似度的电视产品推荐方法,其特征在于,所述去重步骤包括:删除一电视剧的分集信息,只保留所述电视剧的节目名称;所述归一化步骤包括建立一字典库,将节目内容相同但节目名称不同的数据统一化。

4. 根据权利要求3所述的融合时间与兴趣相似度的电视产品推荐方法,其特征在于,所述预处理步骤还包括:去除所述节目名称中的标签说明文字和去除收视时间小于一预设时间的用户收视数据。

5. 根据权利要求4所述的融合时间与兴趣相似度的电视产品推荐方法,其特征在于,所述S2步骤进一步包括步骤:

S21: 利用MariaDB对预处理后的所述用户收视数据进行合并去重,获得处理后用户收视行为数据;

S22: 利用所述处理后用户收视行为数据生成关联数据组;

S23: 将数据导出至Python软件,获得导出数据。

6. 根据权利要求5所述的融合时间与兴趣相似度的电视产品推荐方法,其特征在于,所述S2中还包括步骤:所述关联数据组包括多个特征信息和多个标签信息,每一所述特征信息对应有一标签信息;所述特征信息包括收视数据特征信息、节目特征信息和用户行为特征信息;所述收视数据特征信息对应的所述标签信息包括用户收视时长、最早观看时间和用户总观看时长;所述节目特征信息所对应的所述标签信息包括节目频道所属分类标签和目标人群特征标签;所述用户行为特征信息所对应的所述标签信息包括点播金额、点播时间、点播次数和观看时间。

7. 根据权利要求6所述的融合时间与兴趣相似度的电视产品推荐方法,其特征在于,所述电视产品推荐系统模型包括一时间权重公式和一兴趣相似度公式;所述S3步骤进一步包括步骤:

S31: 构建一时间权重公式:

$$W_{time}(u, i) = (1 - a_1) + a_1 \cdot \frac{D_{ui}}{L_u} \quad (1);$$

其中, $W_{time}(u, i)$ 表示所述时间权重公式; $D_{ui}$ 表示第u用户观看第i节目的时间与第u用户最早观看节目的时间间隔;其中 $L_u$ 表示第u用户最早观看节目的时间与最近观看节目的时间间隔; $a_1 \in (0, 1)$ 表示时间权重增长指数;

S32: 构建一兴趣相似度公式:

$$W_{interest}(u, i) = (1 - a_2) + a_2 \cdot \frac{T_{ui}}{D_u} \quad (2);$$

其中,  $W_{interest}(u, i)$  表示所述兴趣相似度公式;  $T_{ui}$  表示第  $u$  用户观看第  $i$  节目的时长,  $D_u$  为第  $u$  用户观看的总时长, 且  $T_{ui} \leq D_u$ ;  $a_2 \in (0, 1)$  表示兴趣相似度权重增长指数;

S33: 对所述时间权重公式和所述兴趣相似度公式分别输入协同过滤推荐算法进行首次运算, 并对所述时间权重公式和所述兴趣相似度公式的输出结果按比例进行融合; 随机取部分所述导出数据输入当前电视产品推荐系统模型, 并调整融合比例进行训练, 获得多个推荐结果, 并使用剩余的数据计算当前所述电视产品推荐系统模型的召回率、准确率与推荐覆盖率, 作图, 得出最优融合比例, 获得最终的所述电视产品推荐系统模型。

8. 根据权利要求7所述的融合时间与兴趣相似度的电视产品推荐方法, 其特征在于, 所述随机取部分所述导出数据步骤进一步包括步骤:

S331: 将经过第一预处理后的所述用户收视数据进行随机排序, 切割成一训练集和一测试集;

S332: 去除所述训练集和所述测试集中的重复数据, 利用基于规则的方法对所述训练集和所述测试集去重后的数据进行归一化处理。

## 融合时间与兴趣相似度的电视产品推荐方法

### 技术领域

[0001] 本发明涉及智能信息处理和机器学习领域,尤其涉及一种融合时间与兴趣相似度的电视产品推荐方法。

### 背景技术

[0002] 现有的电视产品推荐方法中,较常见的推荐方法是传统协同过滤模型。其方法是:生成用户与节目、节目与节目或用户与用户的关系矩阵,寻找与每个用户最相似的节目作为推荐结果。现有的技术没有将用户的兴趣变化考虑在内,导致系统会基于用户早期观看过的节目记录来推荐相似的节目,使得推荐结果不符合用户现在的兴趣特征。

### 发明内容

[0003] 针对上述现有技术中的不足,本发明提供一种融合时间与兴趣相似度的电视产品推荐方法,能够更为精确地对用户进行电视产品个性化推荐。

[0004] 为了实现上述目的,本发明提供一种融合时间与兴趣相似度的电视产品推荐方法,包括步骤:

[0005] S1:收集一时间段的用户收视行为数据并进行预处理,所述用户收视行为数据包括用户收视数据、用户订阅数据和用户点播数据;

[0006] S2:对预处理后的所述用户收视行为数据构建特征;

[0007] S3:构建电视产品推荐系统模型,所述电视产品推荐系统模型基于时间权重和兴趣相似度;

[0008] S4:收集并将用户全部的历史收视数据输入所述电视产品推荐系统模型,获得针对每个用户的个性化推荐结果。

[0009] 优选地,所述预处理步骤包括:利用PHP对所述用户收视行为数据进行剔除异常值、去重和归一化。

[0010] 优选地,所述去重步骤包括:删除一电视剧的分集信息,只保留所述电视剧的节目名称;所述归一化步骤包括建立一字典库,将节目内容相同但节目名称不同的数据统一化。

[0011] 优选地,所述预处理步骤还包括:去除所述节目名称中的标签说明文字和去除收视时间小于一预设时间的用户收视数据。

[0012] 优选地,所述S2步骤进一步包括步骤:

[0013] S21:利用MariaDB对预处理后的所述用户收视数据进行合并去重,获得处理后用户收视行为数据;

[0014] S22:利用所述处理后用户收视行为数据生成关联数据组;

[0015] S23:将数据导出至Python软件,获得导出数据。

[0016] 优选地,所述S2中还包括步骤:所述关联数据组包括多个特征信息和多个标签信息,每一所述特征信息对应有一标签信息;所述特征信息包括收视数据特征信息、节目特征信息和用户行为特征信息;所述收视数据特征信息对应的所述标签信息包括用户收视时

长、最早观看时间和用户总观看时长；所述节目特征信息所对应的所述标签信息包括节目频道所属分类标签和目标人群特征标签；所述用户行为特征信息所对应的所述标签信息包括点播金额、点播时间、点播次数和观看时间。

[0017] 优选地，所述电视产品推荐系统模型包括一时间权重公式和一兴趣相似度公式；所述S3步骤进一步包括步骤：

[0018] S31：构建一时间权重公式：

$$[0019] \quad W_{time}(u, i) = (1 - a_1) + a_1 \cdot \frac{D_{ui}}{L_u} \quad (1);$$

[0020] 其中， $W_{time}(u, i)$  表示所述时间权重公式； $D_{ui}$  表示第u用户观看第i节目的时间与第u用户最早观看节目的时间间隔；其中 $L_u$ 表示第u用户最早观看节目的时间与最近观看节目的时间间隔； $a_1 \in (0, 1)$  表示时间权重增长指数；

[0021] S32：构建一兴趣相似度公式：

$$[0022] \quad W_{interest}(u, i) = (1 - a_2) + a_2 \cdot \frac{T_{ui}}{D_u} \quad (2);$$

[0023] 其中， $W_{interest}(u, i)$  表示所述兴趣相似度公式； $T_{ui}$  表示第u用户观看第i节目的时长， $D_u$  为第u用户观看的总时长，且 $T_{ui} \leq D_u$ ； $a_2 \in (0, 1)$  表示兴趣相似度权重增长指数；

[0024] S33：对所述时间权重公式和所述兴趣相似度公式分别输入协同过滤推荐算法进行首次运算，并对所述时间权重公式和所述兴趣相似度公式的输出结果按比例进行融合；随机取部分所述导出数据输入当前电视产品推荐系统模型，并调整融合比例进行训练，获得多个推荐结果，并使用剩余的数据计算当前所述电视产品推荐系统模型的召回率、准确率与推荐覆盖率，作图，得出最优融合比例，获得最终的所述电视产品推荐系统模型。

[0025] 优选地，所述随机取部分所述导出数据步骤进一步包括步骤：

[0026] S331：将经过第一预处理后的所述用户收视数据进行随机排序，切割成一训练集和一测试集；

[0027] S332：去除所述训练集和所述测试集中的重复数据，利用基于规则的方法对所述训练集和所述测试集中去重后的数据进行归一化处理。

[0028] 本发明由于采用了以上技术方案，使其具有以下有益效果：

[0029] 本发明的电视产品推荐系统模型基于时间权重和兴趣相似度，其引入了“时间权重”与“兴趣相似度”作为额外特征，将用户的历史收视数据使用改进型的协同过滤模型进行计算，分别生成“时间权重”与“兴趣相似度”，继续将两者进行加权计算，得到针对某个用户的个性化推荐。由于结合了用户观看节目的时间与观看节目的次数，使得推荐结果很好地反馈了用户的兴趣变化，结果较传统协同过滤模型更精确。

## 附图说明

[0030] 图1为本发明实施例的融合时间与兴趣相似度的电视产品推荐方法的流程图。

## 具体实施方式

[0031] 下面根据附图1，给出本发明的较佳实施例，并予以详细描述，使能更好地理解本发明的功能、特点。

[0032] 请参阅图1,本发明实施例的一种融合时间与兴趣相似度的电视产品推荐方法,包括步骤:

[0033] S1:收集一时间段的用户收视行为数据并进行预处理,用户收视行为数据包括用户收视数据、用户订阅数据和用户点播数据。

[0034] 其中,预处理步骤包括:利用PHP对用户收视行为数据进行剔除异常值、去重和归一化。

[0035] 去重步骤包括:删除一电视剧的分集信息,只保留电视剧的节目名称;归一化步骤包括建立一字典库,将节目内容相同但节目名称不同的数据统一化。

[0036] 预处理步骤还包括:去除节目名称中的标签说明文字和去除收视时间小于一预设时间的用户收视数据。

[0037] S2:对预处理后的用户收视行为数据构建特征。

[0038] 其中,S2步骤进一步包括步骤:

[0039] S21:利用MariaDB对预处理后的用户收视数据进行合并去重,获得处理后用户收视行为数据;

[0040] S22:利用处理后用户收视行为数据生成关联数据组;

[0041] S23:将数据导出至Python软件,获得导出数据。

[0042] 本实施例中,S2中还包括步骤:关联数据组包括多个特征信息和多个标签信息,每一特征信息对应有一标签信息;特征信息包括收视数据特征信息、节目特征信息和用户行为特征信息;收视数据特征信息对应的标签信息包括用户收视时长、最早观看时间和用户总观看时长;节目特征信息所对应的标签信息包括节目频道所属分类标签和目标人群特征标签;用户行为特征信息所对应的标签信息包括点播金额、点播时间、点播次数和观看时间。

[0043] S3:构建电视产品推荐系统模型,电视产品推荐系统模型基于时间权重和兴趣相似度。

[0044] 电视产品推荐系统模型包括一时间权重公式和一兴趣相似度公式;S3步骤进一步包括步骤:

[0045] S31:构建一时间权重公式:

$$[0046] \quad W_{time}(u, i) = (1 - a_1) + a_1 \cdot \frac{D_{ui}}{L_u} \quad (1);$$

[0047] 其中, $W_{time}(u, i)$ 表示时间权重公式; $D_{ui}$ 表示第u用户观看第i节目的时间与第u用户最早观看节目的时间间隔;其中 $L_u$ 表示第u用户最早观看节目的时间与最近观看节目的时间间隔; $a_1 \in (0, 1)$ 表示时间权重增长指数;

[0048] S32:构建一兴趣相似度公式:

$$[0049] \quad W_{interest}(u, i) = (1 - a_2) + a_2 \cdot \frac{T_{ui}}{D_u} \quad (2);$$

[0050] 其中, $W_{interest}(u, i)$ 表示兴趣相似度公式; $T_{ui}$ 表示第u用户观看第i节目的时长, $D_u$ 为第u用户观看的总时长,且 $T_{ui} \leq D_u$ ;  $a_2 \in (0, 1)$ 表示兴趣相似度权重增长指数;

[0051] S33:对时间权重公式和兴趣相似度公式分别输入协同过滤推荐算法进行首次运算,并对时间权重公式和兴趣相似度公式的输出结果按比例进行融合;随机取部分导出数

据输入当前电视产品推荐系统模型,并调整融合比例进行训练,获得多个推荐结果,并使用剩余的数据计算当前电视产品推荐系统模型的召回率、准确率与推荐覆盖率,作图,得出最优融合比例,获得最终的电视产品推荐系统模型。

[0052] 本实施例中,随机取部分导出数据步骤进一步包括步骤:

[0053] S331:将经过第一预处理后的用户收视数据进行随机排序,切割成一训练集和一测试集;

[0054] S332:去除训练集和测试集中的重复数据,利用基于规则的方法对训练集和测试集中去重后的数据进行归一化处理。

[0055] S4:收集并将用户全部的历史收视数据输入电视产品推荐系统模型,获得针对每个用户的个性化推荐结果。

[0056] 例如:针对某广电网络运营公司,自2017年7月1日至2017年9月30日的用户收视行为数据。

[0057] 步骤一,对用户收视数据进行预处理。

[0058] 用户收视行为数据包含“节目信息”数据、“用户观看时间”数据和“用户点播行为”数据。由于直接使用原始数据训练模型不仅会产生误差,还会耗费大量的计算资源。因此,对原始数据集中存在的异常值进行剔除、去重、归一化等处理。一方面,由于数据中同一电视剧的分集也单独算作一个电视节目,当推荐系统向用户推荐电视剧中某一分集时,会使得用户很难完整地理解剧情,这对用户的观看体验会产生负面的影响,因此在预处理时会将同一电视剧的分集作为重复数据剔除,仅保留电视剧名;另一方面,平台出于管理目的,会在电视节目的名称中添加标签说明文字,这些信息对于用户来说是无用的,因此也需要在预处理时剔除;另外,数据集中还存在某些用户在短时间内观看了大量节目,但每部节目的收视时间都较短,这种情况产生的原因可能是用户在短时间内进行无目的地换台寻找自己感兴趣的节目导致的,因此这类收视时间较短的记录也需要被剔除。

[0059] 步骤二、对经过预处理的数据构建特征。

[0060] 为提高模型预测的准确性,本发明采集每个电视节目与频道所属分类与目标人群类型作为额外数据,使其成为节目的标签特征。由于用户所偏好的节目类型通常是固定的,因此结合用户的标签特征进行推荐会提高推荐系统的命中率。最终,模型训练与预测使用的特征信息与标签信息如表1所示。

[0061] 表1模型训练与预测使用的特征信息和标签信息表

[0062]

特征信息	标签信息
收视数据	用户收视时长、最早观看时间、用户总观看时长
节目	节目与频道所属分类与目标人群特征标签
用户行为	用户对节目的点播金额、点播时间、点播次数、观看时间

[0063] 步骤三、基于时间权重与兴趣相似度构建协同过滤推荐模型,即电视产品推荐系统模型。

[0064] 为获得精确度高的电视节目推荐模型,本发明实施例采用了两个阶段的训练方法。第一次阶段的训练中,使用电视产品推荐系统模型将用户观看的节目特征信息的标签信息数据作为输入数据进行训练,得出用户的偏好特征。第二次阶段的训练中,使用本实施

例的电视产品推荐系统模型将用户完整的收视数据作为输入数据进行训练,得出用户的电视节目推荐结果。两个阶段的训练生成了两个推荐模型,即标签推荐模型和节目推荐模型,其参数如表2所示。

[0065] 表2标签推荐模型和节目推荐模型的参数对照表

[0066]

推荐模型参数特征	标签推荐模型	节目推荐模型
时间权重增长指数 $a_1$	0.5	0.6
兴趣相似度权重增长指数 $a_2$	3	5
时间权重与兴趣相似度融合比例	0.5	0.5
推荐数量	5	10

[0067] 本发明实施例调整了权重增长系数、两种权重融合比例与推荐数量等参数。其中时间权重增长指数与兴趣相似度权重增长指数的取值范围落在0~1之间,值越小,权重对推荐结果的影响越小。本实施例使用多个参数,分别进行实验,得出表中的增长指数使得推荐模型的准确率最高。

[0068] 步骤四、对已经训练好的标签推荐模型和节目推荐模型进行多模型加权融合,得到每个用户的个性化推荐。将用户历史全部的收视数据作为两个模型的输入数据,分别得出每个用户的标签推荐指数与节目推荐指数。再将每个用户的节目推荐指数与节目所属标签对用户的推荐指数按比例融合,得出最终的节目推荐指数,从大大小排序,即可得出每个用户的个性化推荐结果。

[0069] 通过优化算法参数,采用测试集样本对推荐结果进行评估,算法的运行结果和精度测试如表3所示。

[0070] 表3历史均值与提升树融合模型精度测试结果表

[0071]

融合比例	准确率	召回率	推荐覆盖率
0.3	0.4192%	0.5663%	60.4626%
0.4	0.2616%	0.3641%	59.8673%
0.5	0.4281%	0.5663%	61.6833%
0.6	0.2655%	0.3641%	58.0716%
0.7	0.2564%	0.3641%	55.4757%
0.8	0.3625%	0.4854%	56.2431%

[0072] 实验中使用准确率、召回率与推荐覆盖率作为模型的评价指标。准确率precision的计算公式是:  $precision = \frac{Hits}{N}$ , 其中Hits表示算法产生的正确推荐数,N表示算法生成的

推荐总数;召回率recall的计算公式是:  $recall = \frac{Hits}{A}$ , Hits表示算法产生的正确推荐数,A

表示所有的推荐总数;推荐覆盖率coverage的计算公式:  $coverage = \frac{R(u)}{I}$ , R(u)表示推荐系统给用户推荐的电视产品列表,I表示系统中所有电视产品信息的总数。coverage的值越大,则推荐列表中的电视产品信息越丰富。推荐覆盖率能够从一定程度上反映用户潜在兴

趣的程度,但是推荐结果的准确性却依赖于准确度。观察表3可知,当融合比例为0.6时模型在准确率、召回率、推荐覆盖率上性能较优。这是由于它结合了两种加权方法优点,不仅能突出近期用户观看电视产品信息的重要性,又避免了早期所观看的电视产品信息被忽略,从而更准确地反映了用户的兴趣变化趋势。

[0073] 本发明实施例的一种融合时间与兴趣相似度的电视产品推荐方法,引入了“时间权重”与“兴趣相似度”作为额外特征,将用户的历史收视数据使用改进型的协同过滤模型进行计算,分别生成“时间权重”与“兴趣相似度”,继续将两者进行加权计算,得到针对某个用户的个性化推荐。由于结合了用户观看节目的时间与观看节目的次数,使得推荐结果很好地反馈了用户的兴趣变化,结果较传统协同过滤模型更精确。

[0074] 以上结合附图实施例对本发明进行了详细说明,本领域中普通技术人员可根据上述说明对本发明做出种种变化例。因而,实施例中的某些细节不应构成对本发明的限定,本发明将以所附权利要求书界定的范围作为本发明的保护范围。

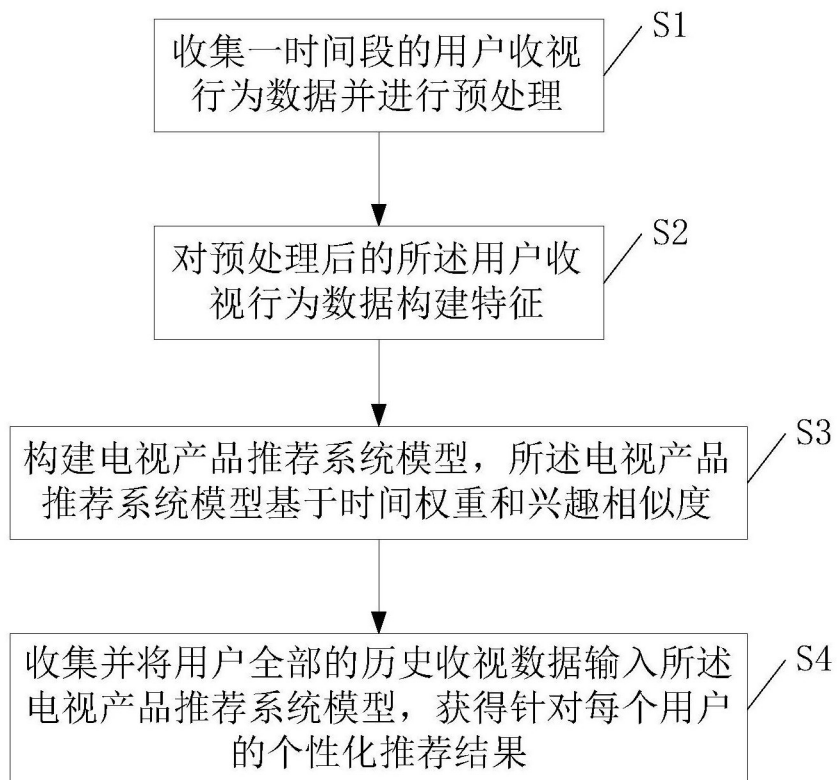


图1