



(12)发明专利申请

(10)申请公布号 CN 105868298 A

(43)申请公布日 2016.08.17

(21)申请号 201610172881.4

(22)申请日 2016.03.23

(71)申请人 华南理工大学

地址 510640 广东省广州市天河区五山路
381号

(72)发明人 古万荣 董守斌 胡金龙 付佳兵
张铃启

(74)专利代理机构 广州市华学知识产权代理有
限公司 44245

代理人 陈宏升

(51)Int.Cl.

G06F 17/30(2006.01)

G06Q 30/02(2012.01)

G06K 9/62(2006.01)

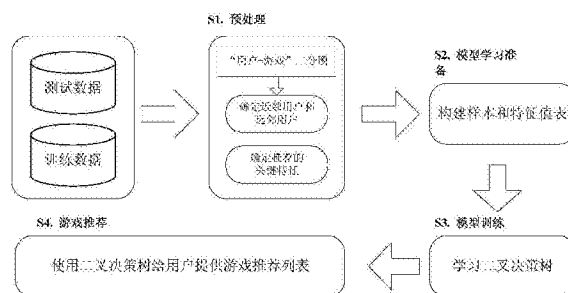
权利要求书2页 说明书6页 附图1页

(54)发明名称

一种基于二叉决策树的手机游戏推荐方法

(57)摘要

本发明公开了一种基于二叉决策树的手机游戏推荐方法,主要通过训练二叉决策树分类模型来确定用户是否对特定的待推荐手机游戏感兴趣,从而将特征权重对推荐的影响从主观赋值转变为学习模型自动预测。本发明的特色在于使用二叉决策树来确定游戏推荐场景中的特征选择和权重衡量问题,并给出相应的推荐模型。使用本发明可以较好地保证推荐结果和用户偏好的准确性,同时对推荐结果无需做额外二次挑选的前提下又可以保证推荐结果的多样性。



1. 一种基于二叉决策树的手机游戏推荐方法,其特征在于包括以下步骤:

S1、利用已有的用户玩游戏记录,进行预处理,确定近邻用户和远邻用户集合,获得用户、游戏、开发商、广告、地区、时段和分类;

S2、根据步骤S1中获得的信息特征构建样本和特征值表;

S3、训练游戏推荐的二叉决策树模型;

S4、根据训练模型,对用户进行游戏推荐,带推荐游戏可以是游戏的全体集合,也可以是经过预处理后的游戏子集。

2. 根据权利要求1所述的基于二叉决策树的手机游戏推荐方法,其特征在于:所述步骤S1具体为:先构建“用户-游戏”二分图,然后使用基于用户方法确定近邻和远邻用户,设置参数由应用数据的规模决定,待推荐的关键特征根据应用场景确定。

3. 根据权利要求1所述的基于二叉决策树的手机游戏推荐方法,其特征在于:在所述步骤S2的训练集中即近邻用户和远邻用户对应的内容,矩阵内容根据重合度计算,即当前用户和其列表中的用户在对应属性集中的重叠衡量,即考察集合的重叠问题,设某用户的特征值表为 D_{tr} ,公式如下:

$$D_{tr|v,a} = \begin{cases} 0, & \frac{|a_v| \cap |a_u|}{|a_v| \cup |a_u|} < Threshold \\ 1, & \frac{|a_v| \cap |a_u|}{|a_v| \cup |a_u|} \geq Threshold \end{cases}$$

其中,v表示表中对应的某个k近邻用户或k远邻用户,a表示特征,u表示当前考察的用户。

4. 根据权利要求1所述的基于二叉决策树的手机游戏推荐方法,其特征在于:所述步骤S2中候选游戏中的矩阵元素的赋值由以下公式确定:

$$D_{tr|n,a} = \begin{cases} 1, & a \text{ 同时出现于 } n \text{ 和 } u \text{ 玩过的游戏中} \\ 0, & \text{否则} \end{cases}$$

该公式表示,当游戏n对应的属性列a如果要赋值为1,则必须要其对应的属性a的值同时出现在游戏n和用户u玩过的游戏中。

5. 根据权利要求1所述的基于二叉决策树的手机游戏推荐方法,其特征在于:所述步骤S3中构建二叉决策树关键在于其节点及其树枝的构建,节点层次越低,说明该节点特征越重要,使用信息增益比来表示特征节点的重要性,即信息增益比越大,则其节点层次越低。

6. 根据权利要求1所述的基于二叉决策树的手机游戏推荐方法,其特征在于:所述步骤S3包括以下分步骤:

S3.1基本定义:信息增益比即属性的信息增益与训练数据集的经验熵的比值:

$$g_R(D, a) = \frac{g(D, a)}{H(D)}$$

其中,H(D)表示训练数据集的经验熵,D即训练数据集,g(D,a)表示特征属性a在训练数据集D中的信息增益;

数据集D的经验熵的计算公式如下:

$$H(D) = -\sum_{k=1}^2 \frac{|c_k|}{|D|} \log_2 \frac{|c_k|}{|D|},$$

其中, c_k 表示的是类别, 类别只有两种;

获得训练数据集整体的经验熵后, 需要计算指定特征属性 a 在此训练数据集中的经验条件熵 $H(D|a)$:

$$H(D|a) = \sum_{i=1}^n \frac{|D_i|}{|D|} H(D_i) = - \sum_{i=1}^n \frac{|D_i|}{|D|} \sum_{k=1}^2 \frac{|D_{ik}|}{|D_i|} \log_2 \frac{|D_{ik}|}{|D_i|} ,$$

信息增益 $g(D, a)$ 则是整体经验熵与经验条件熵的差:

$$g(D, a) = H(D) - H(D|a),$$

其中信息增益比是个相对概念, 即当考察的集合对象变化时, 需要重新计算, 如构建节点的过程中, 考察的集合已经在产生动态变化;

步骤S3.2、模型训练: 给定训练集 D , 即示例表中的 k 近邻用户和 k 远用户所对应的子表, 和特征集 A , 其中 $A = \{\text{手游开发商, 广告, 地区, 时段, 分类}\}$, 用户 u 的决策树 Tr 生成过程如下:

1) 若训练集 D 中所有记录都属于同一个类(0或1), 则 Tr 为单节点二叉树, 并将该类设置为该节点值, 返回 Tr ;

2) 如果特征集 A 为空集, 则将训练集 D 中具有最多记录的类别作为该节点的类别, 返回 Tr ;

3) 否则, 计算剩余特征集各特征的信息增益比 $g_R(D, a)$, 并选择出最大信息增益比对应的特征 a ;

4) 对选出来的最大特征 a 的每一个可能值(0或1)将 D 切割为两个非空子集 D_0 和 D_1 , 将子集 D_0 和 D_1 中记录数最多的类作为标记, 构建子节点, 由该节点及其子节点, 返回供递归使用的 Tr ;

5) 以非空子集 D_0 和 D_1 为训练集, 以 $A \setminus \{a\}$ 为新的特征集, 递归调用步骤1~步骤4, 得到子树 Tr_0 和 Tr_1 , 并返回相应子树。

一种基于二叉决策树的手机游戏推荐方法

技术领域

[0001] 本发明涉及数据挖掘、用户行为预测领域,特别涉及一种基于二叉决策树的手机游戏推荐方法。

背景技术

[0002] 在中国经济、互联网和娱乐业飞速发展的今天,越来越多的人注重业余休闲的质量。智能手机的发展,也带来了移动应用软件、移动广告和手机游戏业的飞速发展。由于游戏种类和适应层次的多样性,海量的手机游戏给用户也会带来信息过载问题,因此,对手机用户进行游戏推荐的方法也应运而生。

[0003] 游戏推荐是推荐技术研究的拓展应用形式之一。海量的游戏催生了推荐的需求,推荐的预测结果同时也可以用于指导游戏开发商专注于有效的游戏开发,节省开发和广告成本。套用经典推荐技术核心原理,如基于内容的推荐和基于协同过滤的推荐会导致一些问题:1)基于内容的推荐容易让用户获得的游戏过于集中,不利于新游戏的推广。2)基于协同过滤的推荐不能及时更新用户偏好度模型。3)传统推荐方法没有细化考虑特定应用,如游戏推荐中的特征及其权重衡量问题。4)现有方法没有考虑到游戏推荐扩展问题。即推荐模型的普适性不强,不能适应推荐场景中推荐条件的变化。

发明内容

[0004] 本发明的目的在于克服现有技术的缺点与不足,提供一种基于二叉决策树的手机游戏推荐方法。

[0005] 本发明的目的通过如下技术方案实现:一种基于二叉决策树的手机游戏推荐方法,具体包括以下步骤:

[0006] S1、利用已有的用户玩游戏记录,进行预处理,确定近邻用户和远邻用户集合,获得用户、游戏、开发商、广告、地区、时段和分类;

[0007] S2、根据步骤S1中获得的信息特征构建样本和特征值表;

[0008] S3、训练游戏推荐的二叉决策树模型;

[0009] S4、根据训练模型,对用户进行游戏推荐,带推荐游戏可以是游戏的全体集合,也可以是经过预处理后的游戏子集。

[0010] 所述步骤S1具体为:先构建“用户-游戏”二分图,然后使用基于用户方法确定近邻和远邻用户,设置参数由应用数据的规模决定,待推荐的关键特征根据应用场景确定。

[0011] 在所述步骤S2的训练集中即近邻用户和远邻用户对应的内容,矩阵内容根据重合度计算,即当前用户和其列表中的用户在对属性集中的重叠衡量,即考察集合的重叠问题,设某用户的特征值表为 D_{tr} ,公式如下:

$$[0012] \quad D_{tr|v,a} = \begin{cases} 0, & \frac{|a_v| \cap |a_u|}{|a_v| \cup |a_u|} < Threshold \\ 1, & \frac{|a_v| \cap |a_u|}{|a_v| \cup |a_u|} \geq Threshold \end{cases}$$

[0013] 其中,v表示表中对应的某个k近邻用户或k远邻用户,a表示特征,u表示当前考察的用户。

[0014] 所述步骤S2中候选游戏中的矩阵元素的赋值由以下公式确定:

$$[0015] \quad D_{tr|n,a} = \begin{cases} 1, & \text{a 同时出现于 n 和 u 玩过的游戏中} \\ 0, & \text{否则} \end{cases}$$

[0016] 该公式表示,当游戏n对应的属性列a如果要赋值为1,则必须要其对应的属性a的值同时出现在游戏n和用户u玩过的游戏中。

[0017] 所述步骤S3中构建二叉决策树关键在于其节点及其树枝的构建,节点层次越低,说明该节点特征越重要,使用信息增益比来表示特征节点的重要性,即信息增益比越大,则其节点层次越低。

[0018] 所述步骤S3包括以下分步骤:

[0019] S3.1基本定义:信息增益比即属性的信息增益与训练数据集的经验熵的比值:

$$[0020] \quad g_R(D, a) = \frac{g(D, a)}{H(D)}$$

[0021] 其中,H(D)表示训练数据集的经验熵,D即训练数据集,g(D,a)表示特征属性a在训练数据集D中的信息增益;

[0022] 数据集D的经验熵的计算公式如下:

$$[0023] \quad H(D) = - \sum_{k=1}^2 \frac{|c_k|}{|D|} \log_2 \frac{|c_k|}{|D|},$$

[0024] 其中,c_k表示的是类别,类别只有两种;

[0025] 获得训练数据集整体的经验熵后,需要计算指定特征属性a在此训练数据集中的经验条件熵H(D|a):

$$[0026] \quad H(D|a) = \sum_{i=1}^n \frac{|D_i|}{|D|} H(D_i) = - \sum_{i=1}^n \frac{|D_i|}{|D|} \sum_{k=1}^2 \frac{|D_{ik}|}{|D_i|} \log_2 \frac{|D_{ik}|}{|D_i|},$$

[0027] 信息增益g(D,a)则是整体经验熵与经验条件熵的差:

$$[0028] \quad g(D, a) = H(D) - H(D|a),$$

[0029] 其中信息增益比是个相对概念,即当考察的集合对象变化时,需要重新计算,如构建节点的过程中,考察的集合已经在产生动态变化;

[0030] 步骤S3.2、模型训练:给定训练集D,即示例表中的k近邻用户和k远用户所对应的子表,和特征集A,其中A={手游开发商,广告,地区,时段,分类},用户u的决策树Tr生成过程如下:

[0031] 1)若训练集D中所有记录都属于同一个类(0或1),则Tr为单节点二叉树,并将该类设置为该节点值,返回Tr;

[0032] 2)如果特征集A为空集,则将训练集D中具有最多记录的类别作为该节点的类别,返回Tr;

[0033] 3)否则,计算剩余特征集各特征的信息增益比 $gr(D,a)$,并选择出最大信息增益比对应的特征 a ;

[0034] 4)对选出来的最大特征 a 的每一个可能值(0或1)将 D 切割为两个非空子集 D_0 和 D_1 ,将子集 D_0 和 D_1 中记录数最多的类作为标记,构建子节点,由该节点及其子节点,返回供递归使用的 Tr ;

[0035] 5)以非空子集 D_0 和 D_1 为训练集,以 $A \setminus \{a\}$ 为新的特征集,递归调用步骤1~步骤4,得到子树 Tr_0 和 Tr_1 ,并返回相应子树。

[0036] 本发明与现有技术相比,具有如下优点和有益效果:

[0037] 1、本发明可以有效地融合基于内容和基于协同过滤的基本思想,使推荐趋于融合推荐模型,有利于提高推荐准确度。

[0038] 2、本发明对游戏推荐中的各种特征进行重要度学习,而无需依赖于经验权重,有利于推荐结果的客观公正。

[0039] 3、本发明区别于以往的推荐方法,不容易产生基于内容的推荐游戏聚合,即较大限度地确保了推荐游戏的多样化。

[0040] 4、本发明的方法容易扩展,如特征可以根据不同的游戏应用需求进行增减,稍加修改还可以扩展到其他领域,如音乐推荐、广告推荐、应用推荐等。

附图说明

[0041] 图1为本发明所述的基于二叉决策树的手机游戏推荐方法的处理步骤流程图。

具体实施方式

[0042] 下面结合实施例及附图对本发明作进一步详细的描述,但本发明的实施方式不限于此。

[0043] 本发明具体做法步骤如下,如图1所示:

[0044] 步骤S1、预处理。

[0045] 该步骤首先要确定近邻用户和远邻用户集合。在本方法中,先构建“用户-游戏”二分图,然后使用基于用户(User-based)方法确定若干近邻和远邻用户,这里可以设置参数为 m ,该参数由应用数据的规模决定,一般将其设置为总用户数的10%。待推荐的关键特征根据应用场景确定,如应用场景中的游戏数据有“开发商”,则将其作为特征之一,以此类推。

[0046] 这些预处理产生的结果将为模型学习提供基本元素。

[0047] 步骤S2、模型学习准备。

[0048] 该步骤主要是构建用户的样本和特征值表,如表1所示,这里设某用户的特征值表为 D_{tr} 。

[0049] 表1 D_{tr} 表结构示例

[0050] (备注:表中行表示近邻用户 u_i 、远邻用户 u_i' 以及待推荐游戏 n_i ;列表示关联特征以及最后一列的是否接受的标记。表中的元素表示与当前用户是否有重合之处,如有重合,则元素为1,否则为0。)

[0051]

	开发 商	广 告	地 区	时 段	分 类	接 受
u_1	1	1	0	1	0	1
u_2	0	0	0	1	1	1
...	...	...	...	...	...	...
u_k	1	1	0	1	0	1
u_1'	0	1	1	0	0	0
u_2'	0	0	1	0	1	0
...	...	...	...	...	...	...
u_k'	0	0	1	0	0	0
n_1	0	1	0	0	1	?
n_2	1	0	0	1	1	?
...	...	...	...	...	...	...
n_m	0	1	1	0	0	?

[0052] 表1中涵盖了决策树学习的训练集和待预测集,其中,训练集由k近邻用户和k远用户构成,待预测集即本节需要确认该用户是否接受待推荐游戏的构成。该表有一些特点:

[0053] 1)k近邻用户对应的“接受”字段全部为1。

[0054] 2)k远用户对应的“接受”字段全部为0。

[0055] 3)待预测集的“接受”字段未知,即算法需要填补的字段。

[0056] 4)在训练集中(k近邻用户和k远用户对应的内容),矩阵内容根据重合度计算,即当前用户和其列表中的用户在对应属性集中的重叠衡量,即考察集合的重叠问题,公式如下:

$$[0057] \quad \mathbf{D}_{tr|v,a} = \begin{cases} 0, & \frac{|a_v| \cap |a_u|}{|a_v| \cup |a_u|} < Threshold \\ 1, & \frac{|a_v| \cap |a_u|}{|a_v| \cup |a_u|} \geq Threshold \end{cases} \quad (1)$$

[0058] 其中,k表示表中对应的某个k近邻用户或k远用户,a表示特征(即开发商、广告、地区、时段或分类),u表示当前考察的用户。该公式表示,当用户u和用户v在某个属性中的集合重叠度大于等于阈值T时,对应矩阵元素赋值为1,否则赋值为0。

[0059] 5)候选游戏中的矩阵元素的赋值由以下公式确定:

$$[0060] \quad \mathbf{D}_{tr|n,a} = \begin{cases} 1, & a \text{ 同时出现于 } n \text{ 和 } u \text{ 玩过的游戏中} \\ 0, & \text{否则} \end{cases} \quad (2)$$

[0061] 该公式表示,当游戏n对应的属性列a如果要赋值为1,则必须要其对应的属性a的值同时出现在游戏n和用户u玩过的游戏中。例如,待推荐游戏n的开发商出现了“网易”,且

当前用户 u 玩过的游戏也有同样的开发商关联,则对应的矩阵元素 $D_{tr}|_{n,a}$ 赋值为1。

[0062] 步骤S3、二叉决策树学习

[0063] 构建二叉决策树关键在于其节点及其树枝的构建,节点层次越低,说明该节点特征越重要,本方法使用信息增益比来表示特征节点的重要性,即信息增益比越大,则其节点层次越低。

[0064] 步骤S3.1基本定义

[0065] 信息增益比即属性的信息增益与训练数据集的经验熵的比值:

$$[0066] \quad g_R(D, a) = \frac{g(D, a)}{H(D)} \quad (3)$$

[0067] 其中, $H(D)$ 表示训练数据集的经验熵, D 即训练数据集, $g(D, a)$ 表示特征属性 a 在训练数据集 D 中的信息增益。数据集 D 的经验熵的计算公式如下:

$$[0068] \quad H(D) = - \sum_{k=1}^2 \frac{|c_k|}{|D|} \log_2 \frac{|c_k|}{|D|} \quad (4)$$

[0069] 其中, c_k 表示的是类别,在小节提出的方法中,类别只有两种。获得训练数据集整体的经验熵后,需要计算指定特征属性 a 在此训练数据集中的经验条件熵 $H(D|a)$:

$$[0070] \quad H(D|a) = \sum_{i=1}^n \frac{|D_i|}{|D|} H(D_i) = - \sum_{i=1}^n \frac{|D_i|}{|D|} \sum_{k=1}^2 \frac{|D_{ik}|}{|D_i|} \log_2 \frac{|D_{ik}|}{|D_i|} \quad (5)$$

[0071] 信息增益 $g(D, a)$ 则是整体经验熵与经验条件熵的差:

$$[0072] \quad g(D, a) = H(D) - H(D|a) \quad (6)$$

[0073] 这里需要注意的是,信息增益比是个相对概念,即当考察的集合对象变化时,需要重新计算,如构建节点的过程中,考察的集合已经在产生动态变化。

[0074] 步骤S3.2、模型训练

[0075] 给定训练集 D (即示例表中的 k 近邻用户和 k 远用户所对应的子表)和特征集 A , ($A = \{\text{手游开发商, 广告, 地区, 时段, 分类}\}$), 用户 u 的决策树 Tr 生成过程如下:

[0076] 若训练集 D 中所有记录都属于同一个类(0或1), 则 Tr 为单节点二叉树, 并将该类设置为该节点值, 返回 Tr 。

[0077] 2) 如果特征集 A 为空集, 则将训练集 D 中具有最多记录的类别作为该节点的类别, 返回 Tr 。

[0078] 3) 否则, 计算剩余特征集各特征的信息增益比 $g_R(D, a)$, 并选择出最大信息增益比对应的特征 a 。

[0079] 4) 对选出来的最大特征 a 的每一个可能值(0或1)将 D 切割为两个非空子集 D_0 和 D_1 , 将子集 D_0 和 D_1 中记录数最多的类作为标记, 构建子节点, 由该节点及其子节点, 返回供递归使用的 Tr 。

[0080] 5) 以非空子集 D_0 和 D_1 为训练集, 以 $A \setminus \{a\}$ 为新的特征集, 递归调用步骤1~步骤4, 得到子树 Tr_0 和 Tr_1 , 并返回相应子树。

[0081] 步骤S4、游戏推荐。

[0082] 根据训练模型, 对用户进行游戏推荐, 带推荐游戏可以是游戏的全体集合, 也可以是经过预处理后的游戏子集。

[0083] (1)根据本发明的步骤S1,利用已有的用户玩游戏记录,进行预处理,获得用户、游戏、开发商、广告、地区、时段和分类。

[0084] (2)根据本发明的步骤S2,利用这些特征构建样本和特征值表。

[0085] (3)根据本发明的步骤S3,训练游戏推荐的二叉决策树模型。

[0086] (4)根据本发明的步骤S4,利用二叉决策树模型给用户推荐游戏。

[0087] 上述实施例为本发明较佳的实施方式,但本发明的实施方式并不受上述实施例的限制,其他的任何未背离本发明的精神实质与原理下所作的改变、修饰、替代、组合、简化,均应为等效的置换方式,都包含在本发明的保护范围之内。

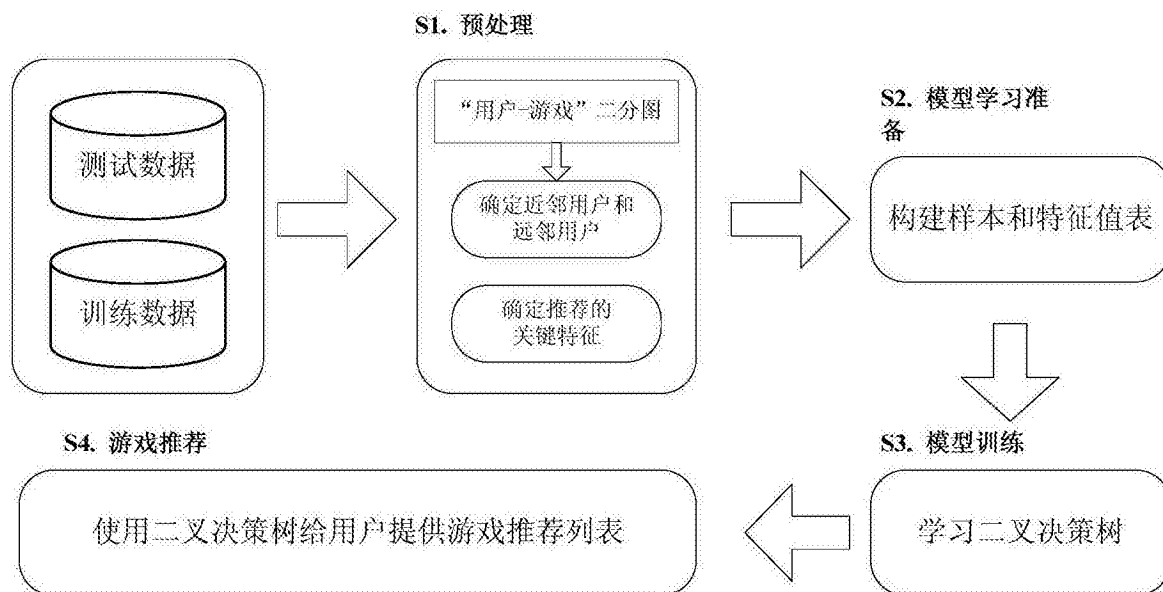


图1