



(12)发明专利

(10)授权公告号 CN 105630957 B

(45)授权公告日 2019.05.21

(21)申请号 201510983204.6

G06Q 10/04(2012.01)

(22)申请日 2015.12.24

(56)对比文件

(65)同一申请的已公布的文献号

申请公布号 CN 105630957 A

CN 102880501 A, 2013.01.16,

CN 102567477 A, 2012.07.11,

CN 101448025 A, 2009.06.03,

CN 101261604 A, 2008.09.10,

CN 101377739 A, 2009.03.04,

(43)申请公布日 2016.06.01

(73)专利权人 北京大学

地址 100871 北京市海淀区颐和园路5号

审查员 戴诚

(72)发明人 刘譞哲 黄罡 梅宏 李豁然

陆璇

(74)专利代理机构 北京君尚知识产权代理事务

所(普通合伙) 11200

代理人 邱晓锋

(51)Int.Cl.

G06F 16/958(2019.01)

G06F 17/18(2006.01)

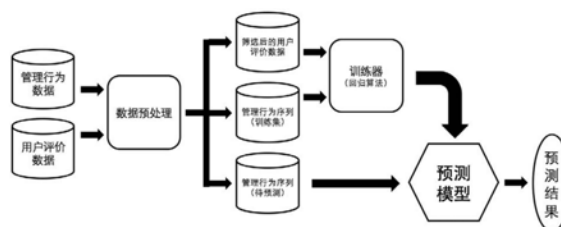
权利要求书2页 说明书5页 附图1页

(54)发明名称

一种基于用户管理应用行为的应用质量判别方法及系统

(57)摘要

本发明涉及一种基于用户管理应用行为的应用质量判别方法及系统。该方法包括:1)对用户的应用管理行为进行数据预处理,得到管理行为序列,作为训练集;2)利用得到的管理行为序列,对用户的应用管理行为进行特征抽取;3)利用抽取的行为特征建立多维自变量空间,每一个训练集中的应用被视作该空间当中的一个点,根据此空间进行模型训练,调试出最优的模型参数,得到一个确定的回归模型;4)将每一个待预测的应用的管理行为数据按照步骤1)和步骤2)进行整理,然后输入该回归模型,模型的输出即为对应用质量的判别结果。本发明能够更为有效地去除无关特征和发挥有关特征的质量表达能力,能够获得更为准确的应用质量评价结果。



1. 一种基于用户管理应用行为的应用质量判别方法,其特征在于,包括如下步骤:

1) 对用户的应用管理行为进行数据预处理,将每个用户与每个应用之间的所有管理行为按照时间先后顺序进行排序,得到管理行为序列,作为训练集;

2) 利用得到的管理行为序列,对用户的应用管理行为进行特征抽取;

3) 利用抽取的行为特征建立多维自变量空间,每一个训练集中的应用被视作该空间中的一个点,每个自变量维度的值即为该维度对应的特征的值,因变量的值为应用的用户评分;根据此空间进行模型训练,调试出最优的模型参数,最终得到一个确定的回归模型;

4) 利用训练得到的回归模型,将每一个待预测的应用的管理行为数据按照步骤1)和步骤2)进行整理,然后输入该回归模型,模型的输出即为对应用质量的判别结果。

2. 如权利要求1所述的方法,其特征在于,步骤1)所述用户的应用管理行为包括三种:下载应用、卸载应用、更新应用;行为记录中包含三个域:进行该行为的用户标识、该行为管理的应用的标识、该行为发生的时间。

3. 如权利要求2所述的方法,其特征在于,步骤1)所述管理行为序列以“D”表示下载应用行为,以“U”表示卸载应用行为,以“P”表示更新应用行为,并在序列的头部添加标识“S”表示序列开始,在尾部添加标识“E”表示序列结束。

4. 如权利要求3所述的方法,其特征在于,步骤1)还对管理行为序列进行筛选,过滤掉用户评价不明确的应用,只保留用户展现出明显喜好倾向的应用。

5. 如权利要求4所述的方法,其特征在于,所述进行筛选的方法是:通过过滤器对所有的应用根据其喜欢、不喜欢的评价个数进行二项分布概率测试,只有能够以95%概率证明该应用的用户倾向较为明显时才保留该应用;然后对于被保留应用的管理行为序列,只保留除S、E外至少包含两个操作的序列。

6. 如权利要求1所述的方法,其特征在于:步骤2)中,每个应用的行为特征包括:该应用拥有的管理行为序列数、管理行为序列的平均长度、管理行为序列中至多连续若干步考虑时间间隔长短的连续管理行为。

7. 如权利要求6所述的方法,其特征在于:步骤2)对于每一种行为特征,抽取得到该应用中该种特征的个数,最后得到平均每一个管理行为序列的该种特征个数,即为此种特征的值。

8. 如权利要求1所述的方法,其特征在于:步骤3)所述多维自变量空间中,因变量的值为应用的用户评分线性规约至 $[0,1]$ 区间后的值。

9. 如权利要求1所述的方法,其特征在于:步骤3)采用多折训练的方法对模型进行训练。

10. 一种基于用户管理应用行为的应用质量判别系统,其特征在于,包括:

数据预处理器,负责对用户的应用管理行为进行数据预处理,将每个用户与每个应用之间的所有管理行为按照时间先后顺序进行排序,得到管理行为序列,作为训练集;

行为特征抽取器,负责利用得到的管理行为序列,对用户的应用管理行为进行特征抽取;

训练器,负责利用抽取的行为特征建立多维自变量空间,每一个训练集中的应用被视作该空间其中的一个点,每个自变量维度的值即为该维度对应的特征的值,因变量的值为应用的用户评分;根据此空间进行模型训练,调试出最优的模型参数,最终得到一个确定的

回归模型；

预测器，负责通过数据预处理器和行为特征抽取器抽取其特征，然后将得到的特征输入训练好的回归模型，模型的输出即为对应用质量的判别结果。

一种基于用户管理应用行为的应用质量判别方法及系统

技术领域

[0001] 本发明是一种基于移动设备用户的应用管理行为抽取行为特征,然后基于该行为特征利用回归算法对缺乏用户评价的应用的质量进行预测的方法及系统。

背景技术

[0002] 随着移动设备(智能手机、平板电脑、智能手表等)的发展,移动设备应用程序(即app,简称应用)的种类、数量都呈现了爆发式增长的趋势。如何能够准确的衡量一个应用的质量好坏成为了一个非常重要的问题。这对于用户在下载应用前快速了解应用从而做出更合适的选择、对于开发者更快速地了解自己和竞争对手的应用在市场中的反馈、以及对于应用市场管理人员提供更准确的推荐都有着非常重要的意义。最为直观且准确的评价标准是应用市场上用户对于应用的评价,包括自然语言评价与打分,其中又以打分最为直观有效。但是,目前的移动应用市场上普遍存在着用户不习惯于评价应用、导致应用的用户评价很少的问题。对于这部分应用,其质量很难通过用户评价进行判别。

[0003] 除用户评价这种相对直观、准确的评价方式外,主流应用市场基本以应用的下载总次数或者以下载过该应用的人数来衡量应用的好坏,其基本出发点是通过用户对于应用的选择行为来判断应用质量的好坏。对于上述所提及的评价较少的应用,这成为了唯一的评价指标。但是,仅仅依靠这种简单的指标可能并不能真正准确的反映出应用的质量高低。例如,一些应用下载总量与下载总人数均较高,但是这可能是由于应用在上线期间所做的营销工作成功而导致的,应用本身质量并不高,用户在简单尝试使用之后很大部分会直接将其卸载。同时,也可能存在一类应用,它们的下载量与下载人数虽然不是非常多,但是其用户会对其保持使用,并跟随版本的迭代随时更新,这类应用往往是质量较高的。因此,单纯的通过下载量、下载人数这样简单的指标可能并不能得到准确的评价。

[0004] 该指标基于用户行为的出发点是合理的,但是其存在的一个问题在于浅尝辄止,没有充分挖掘出用户行为中的信息量。例如,如果一个应用的多数用户在下载了该应用后不久就进行了卸载,那么很可能说明该应用质量不高因此大量用户在使用之后选择了放弃;另一方面,如果一个应用存在部分用户在已经卸载了该应用之后又选择了重新安装,那么很有可能说明这个应用质量较好因此用户的确需要它因此重新安装。受此启示,用户对应用的管理行为序列中可能存在很多有价值的特征。如果能够挖掘出有效特征,就可能利用这些特征得出更准确的应用质量评价。

[0005] 现有技术中,授权公告号为CN101404650B的专利“一种细分网络应用服务质量的方法和系统”,根据用户当前使用的应用程序情况,提供一个合适的网络质量标记,实现细分网络应用服务质量。此外,“一种应用程序的质量确定方法和系统”(申请号:201110247330.7)通过获取的应用程序的质量指标的衡量值基于大量移动终端上报的平均值的收敛值,体现了应用程序的通用质量状态,能够反映应用程序的质量。但这些技术都没有有效利用用户对应用的管理行为中的有价值的特征,从而不能得出更准确的应用质量评价。

[0006] 综上所述,需要一种从应用管理行为中抽取行为特征,并利用抽取出的行为特征对用户评价较少的应用的质量进行判断的方法。

发明内容

[0007] 本发明的目的是通过用户管理行为中的特征,利用回归算法来判断应用质量的好坏。利用含有较多用户评价的应用,针对这部分应用的管理行为进行回归模型训练,用训练得到的模型对用户评论较少的应用进行质量预测。

[0008] 本发明中,用户的应用管理行为数据当中包括三种行为:下载应用、卸载应用、更新应用。行为记录中包含三个域:进行该行为的用户标识、该行为管理的应用的标识(即应用包名)、以及该行为发生的时间。

[0009] 本发明的基于用户管理应用行为的应用质量判别方法,包含四个主要步骤:(1)数据预处理;(2)行为特征抽取;(3)模型训练;(4)利用特征进行应用质量判别。

[0010] (1)数据预处理:

[0011] 首先根据阈值、统计分析等方法筛选出可靠的应用进入训练集。然后,对与训练集中应用相关的用户应用管理行为进行如下数据预处理:以“D”表示下载应用行为、“U”表示卸载应用行为、“P”表示更新应用行为。将每个用户与每个应用之间的所有管理行为排序,按时间先后顺序得到一个管理行为序列。例如,序列:

[0012] DPPU

[0013] 该序列表示该用户先对应用进行了一次卸载,然后更新了两次,最后卸载了该应用。此外,在上述原始序列的头部添加标识“S”表示序列开始,在尾部添加标识“E”表示序列结束,得到最终的序列:

[0014] SDPPUE

[0015] 为保证方法尽可能准确,需要进一步对序列数据进行筛选。最终根据筛选之后的管理行为序列构成模型的训练集。本方法的行为特征抽取工作将基于该训练集进行。

[0016] (2)行为特征抽取:

[0017] 每个应用的特征包括以下两项:该应用拥有的管理行为序列数、管理行为序列的平均长度。除此两项之外,应用的特征还可以包括:管理行为序列中至多连续若干步(包括S和E在内)考虑时间间隔长短的连续管理行为。

[0018] 对于每一种特征,抽取得到该应用中该种特征的个数,最后得到平均每一个管理行为序列的该种特征个数,即为此种特征的值(例如:某应用共有100个管理序列,其一个特征的。出现个数为200,则最终该特征的值 $200/100=2$)。特征总数为N。

[0019] (3)模型训练:

[0020] 行为特征抽取完成后,利用N个特征建立一个N+1维自变量空间。每一个训练集中的应用被视作空间其中的一个点,每个自变量维度的值即为该维度对应的特征的值,因变量的值为该应用的用户评分线性规约至[0,1]区间后的值(0为最低、1为最高)。根据此空间,用多折训练的方法对模型进行训练,调试出最优的模型参数,最终得到一个确定的回归模型。

[0021] (4)利用特征进行应用质量判别:

[0022] 利用训练得到的回归模型,将每一个待预测的应用的管理行为数据按照上述步骤

(1)、(2)的方式进行整理,然后输入模型,模型的输出即为对应用质量的判别结果。

[0023] 本发明的基于用户管理应用行为的应用质量判别系统,其包括:

[0024] 数据预处理器,负责对用户的应用管理行为进行数据预处理,将每个用户与每个应用之间的所有管理行为按照时间先后顺序进行排序,得到管理行为序列,作为训练集;

[0025] 行为特征抽取器,负责利用得到的管理行为序列,对用户的应用管理行为进行特征抽取;

[0026] 训练器,负责利用抽取的行为特征建立多维自变量空间,每一个训练集中的应用被视作该空间当中的一个点,每个自变量维度的值即为该维度对应的特征的值;根据此空间进行模型训练,调试出最优的模型参数,最终得到一个确定的回归模型;

[0027] 预测器,负责通过数据预处理器和行为特征抽取器抽取其特征,然后将得到的特征输入训练好的回归模型,模型的输出即为对应用质量的判别结果。

[0028] 本发明从应用管理行为中抽取行为特征,并利用抽取出的行为特征对用户评价较少的应用的质量进行判断,得出更准确的应用质量评价结果。该方法对于用户在下载应用前快速了解应用从而做出更合适的选择、对于开发者更快速地了解自己和竞争对手的应用在市場中的反馈、以及对于应用市场管理人员提供更准确的推荐等具有重要的意义。

[0029] 与现有技术相比,本发明是首个采用多个连续用户管理行为、及管理行为间的时间间隔来判别应用质量好坏的方法。同时,本发明也是首个采用回归算法,利用用户管理行为对应用质量进行判别的方法。以上特点使得本发明在判别应用质量的过程中可以更为有效地去除无关特征,并且更加有效地发挥有关特征的质量表达能力。因此,总结而言,本发明相较于现有技术可以获得更为准确的评价结果。

附图说明

[0030] 图1为本发明的技术架构图。

具体实施方式

[0031] 下面给出一个基于原始的评价数据进行应用质量判别的具体部署,包括数据描述、硬件部署、具体软件实现。图1为具体的技术架构图。

[0032] 1、数据描述

[0033] 本部署使用的数据来自于一个大型的Android移动应用市场,均为文本文件,包括以下两部分:

[0034] (1) 应用的管理行为数据。每个行为用一行数据来记录,包括前文描述的全部域,即进行该行为的用户标识、该行为管理的应用的标识(即应用包名)、以及该行为发生的时间。

[0035] (2) 用户评价数据。用户对应用的评价简单分为喜欢和不喜欢两种,因此应用的用户评分按照“喜欢的评价数/总评价数”的方式计算,取值范围为[0,1]。每个应用用一行数据来记录。

[0036] 2、硬件部署

[0037] 所需硬件环境为一台装有Ubuntu 14.04 LTS的服务器。所有源数据与软件均部署在该服务器上。

[0038] 3、软件实现

[0039] 软件部署由以下部分组成：

[0040] (1) 数据预处理器

[0041] 数据预处理器由Python脚本、Ubuntu指令及一个统一控制的脚本文件构成。数据预处理器以原始的文本数据为输入，对数据进行排序，然后生成上节中所描述的管理行为序列。

[0042] 然后，数据预处理器对数据进行筛选。应用筛选的目的是希望过滤掉用户评价不明确的应用，只保留用户展现出了明显的喜好倾向的应用。因此，过滤器对所有的应用根据其喜欢、不喜欢的评价个数进行二项分布概率测试，只有能够以95%概率证明该应用的用户倾向较为明显时才保留该应用。然后，对于被保留应用的管理行为序列，支保留除S、E外至少包含两个操作的序列，得到最后的训练集。

[0043] (2) 行为特征抽取器

[0044] 行为特征抽取器由Python脚本实现，以数据预处理器生成的训练集数据为输入，从统计管理行为序列中的特征个数。

[0045] 本部署中，采用的特征为至多包含3个连续管理行为（包含S、E）、同时考虑行为之间的时间间隔的管理行为特征。对于两个操作之间的时间间隔，分为不足24小时、24~48小时、48~72小时、72小时以上共四种情况。每24小时用一个T表示，超过72小时用*表示。以下为若干个符合上述定义的特征示例：

[0046] D: 单次下载

[0047] DU: 下载之后卸载，时间间隔不超过24小时

[0048] SDTTP: 序列开始的第一个行为为下载，经过48~72小时范围内的一段时间后进行了一次更新

[0049] D*PE: 下载之后经过了超过72小时的一段时间进行了卸载，然后序列结束

[0050] 统计过程通过字符串匹配的方式进行，统计结束后按照上节中提出的方法进行求平均，得到提取后的特征统计结果。

[0051] (3) 训练器。训练器以Python脚本实现，内部使用了scikit-learn库的GradientBoostingRegressor类。scikit-learn是一个机遇Python语言开发的开源工具库，主要用于统计学习、机器学习领域；scikit-learn中的GradientBoostingRegressor类是scikit-learn对gradient boosted regression tree算法的实现，后者是一个广泛使用的基于梯度下降思想的回归算法。训练器会根据人工预先设定好的取值范围及粒度进行参数调试。调试具体过程为：针对每一组模型参数，训练器将数据随机等分为5份，然后进行5折交叉验证。每一次实验以4份数据为训练集，1份数据为测试集，得到当前参数下预测结果与实际结果之间的kendall's tau值。5组实验全部完成后，5组的kendall's tau值的平均值即为本组参数的最终结果。尝试所有参数后，选取结果最好的一组，作为训练的最终结果，确定模型。

[0052] (4) 预测器。对于每一个待预测的应用，预测器首先通过数据预处理器和行为特征抽取器抽取其特征，然后将得到的特征输入训练好的模型，产生一个取值范围为[0,1]的预测结果，即为对该应用质量的预测结果。

[0053] 本发明对于具体的实施部署要求较为灵活，除上述实施方式外，也可以采取多种

其它方式进行部署。硬件部署方面可采用其它硬件形式(如大规模集群、云服务器等),亦可以采用其它适用于硬件形式的系统软件(如Windows Server、RedHat等)。数据预处理器、行为特征抽取器、训练器和预测期等软件实现,可采用其它编程语言编写(如Java、C/C++、R等)。数据也可采用其它可行的方式进行组织(如SQL、MongoDB等)。

[0054] 以上实施例仅用以说明本发明的技术方案而非对其进行限制,本领域的普通技术人员可以对本发明的技术方案进行修改或者等同替换,而不脱离本发明的精神和范围,本发明的保护范围应以权利要求书所述为准。

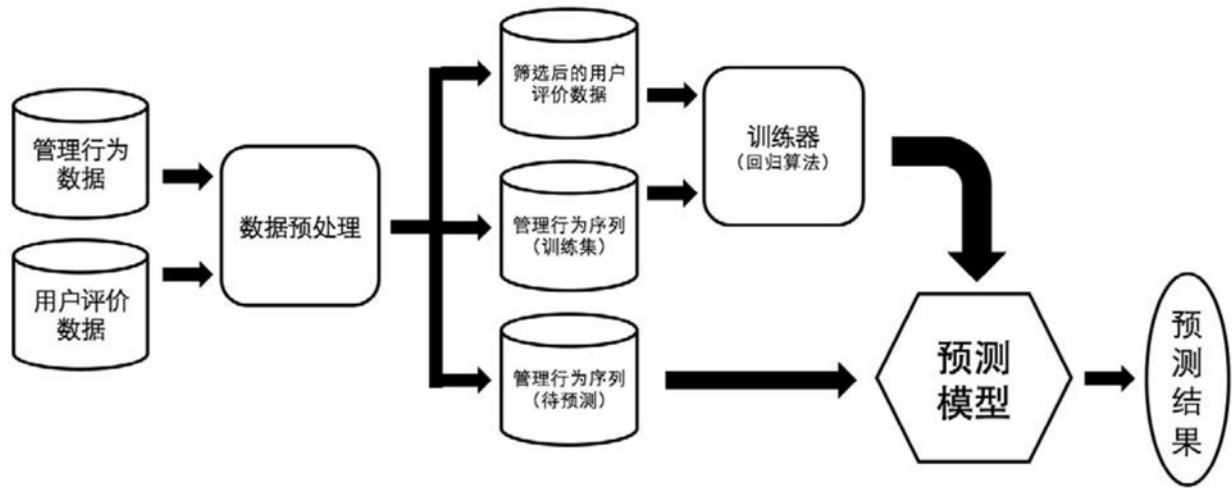


图1