(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 108876536 A
(43)申请公布日 2018.11.23

(21)申请号 201810621062.2

(22)申请日 2018.06.15

(71)申请人 天津大学
地址 300072 天津市南开区卫津路92号

(72)发明人 韩玥 王颖 张子洋 金志刚

(74)专利代理机构 天津市北洋有限责任专利代理事务所 12201
代理人 程毓英

(51)Int.Cl.
G06Q 30/06(2012.01)
G06F 17/30(2006.01)

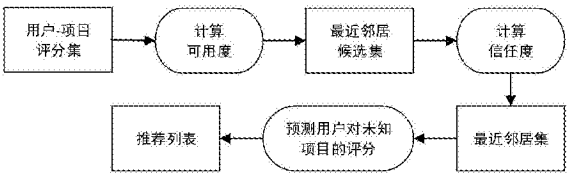
权利要求书1页 说明书4页 附图1页

(54)发明名称

基于最近邻信息的协同过滤推荐方法

(57)摘要

本发明涉及一种基于最近邻信息的协同过滤推荐方法，包括下列步骤：步骤1：输入用户-项目矩阵中的评分数据集，确定目标用户u；步骤2：通过评分数据和皮尔森相似度，对目标用户与其他用户之间的可用度进行计算，并选出可用度较高的k`个用户，生成目标用户u的最近邻居候选集；步骤3：计算同一时间窗口 θ $N_u$`中的每个用户v与目标用户u之间的信任度；步骤4：过滤掉可用度较高但是信任度较低的用户，选取信任度较高的K个用户，作为最近邻居用户，并生成最近邻居集；步骤5：利用N_u的评分信息，计算出目标用户u对所有未评分项目的预测评分；步骤6：选取预测评分高的项目生成推荐项目列表。



CN 108876536 A

1.一种基于最近邻信息的协同过滤推荐方法,包括下列步骤:

步骤1:输入用户-项目矩阵中的评分数据集,确定目标用户u;

步骤2:通过评分数据和皮尔森相似度,对目标用户与其他用户之间的可用度进行计算,并选出可用度较高的k`个用户,生成目标用户u的最近邻居候选集$N_u$`;

步骤3:计算同一时间窗口θN_u`中的每个用户v与目标用户u之间的信任度tru(u,v):

a)根据同一时间窗口θ中用户v对某个项目i的偏好是否接近或远离目标用户u的偏好,认为在窗口θ中用户u对用户v产生了一致偏好和不一致偏好,分别记为$g_i^{\theta}(u,v)$和$b_i^{\theta}(u,v)$;

b)由于用户v可能会在窗口θ中为目标用户u评价多个项目,窗口θ中一致或不一致偏好的值将被计算为所有项目上该用户一致或不一致偏好的总和,分别记为$g^{\theta}(u,v)$和$b^{\theta}(u,v)$

c)给落在同一时间窗口θ中的所有评分分配相同的遗忘因子f,逐渐遗忘用户以前的偏好:对于时间窗口θ,分别赋予一致偏好遗忘因子f_g和不一致偏好遗忘因子f_b,一致和不一致偏好的总值分别记为$G^{\theta}(u,v)$和$B^{\theta}(u,v)$;

d)最后,根据用户v对用户u一致和不一致偏好总值的高低,计算用户v与目标用户u之间的信任度,一致性偏好总值越高或不一致偏好总值越低,信任度越高;

步骤4:过滤掉可用度较高但是信任度较低的用户,选取信任度较高的K个用户,作为最近邻居用户,并生成最近邻居集N_u;

步骤5:利用N_u的评分信息,计算出目标用户u对所有未评分项目的预测评分;

步骤6:选取预测评分高的项目生成推荐项目列表。

基于最近邻信息的协同过滤推荐方法

技术领域

[0001] 本发明属于基于协同过滤的推荐技术领域，具体涉及一种基于最近邻信息的协同过滤推荐方法。

背景技术

[0002] 在互联网快速发展的今天，网络早已融入到人类的日常生活。与此同时，网络信息也变得纷繁复杂，并且数量迅速增长，使得互联网在信息共享时面临着“信息过载”等问题。而在电商领域中，所面临的“信息过载”现象相当严重。这此背景下，个性化推荐技术逐步发展起来。

[0003] 与传统的搜索方法相比较，个性化推荐系统为用户提供独特的服务，它能够通过收集用户的历史行为数据，分析用户的兴趣和潜在兴趣，为用户提供其感兴趣的商品，减少用户查找商品的时间和精力。另一方面，个性化推荐系统为电商网站吸引更多消费者，把商品推荐给消费者，提高网站销量，获得更多利润。达到用户和供应商的双赢。在个性化推荐系统中的核心技术是个性化推荐技术，个性化推荐系统中往往存在评分系统，而利用系统历史评分记录，可以对用户的偏好进行预估。除了评分系统，部分平台还仅有评价系统，利用现有技术如情感分析等可以将用户的文字、表情等评论信息转换为数值评分，进而可以转换为评分系统。

[0004] 如上所述，个性化推荐存在于淘宝、亚马逊、京东、今日头条等各大平台。但作为用户不难发现其中依然存在一些问题，例如某用户在电商平台买了一台笔记本电脑，可想而知该用户在一定时间内不会再去购买笔记本电脑，但是系统会在近期多次推荐笔记本电脑给该用户。同时不难发现，对于一些只有评分或者评论的平台来说，之所以能够实现较高的个性化推荐，是因为在启动个性化推荐之前，拥有着大量的用户和数据，而对于某些中小型的公司来说，并没有大量的用户，即使拥有大量用户也没用较多的用户历史行为记录，导致现存数据相当稀疏，无法实现良好的个性化推荐。

[0005] 为了解决最近邻居选取精度低、数据稀疏情况下推荐准确度不高等问题，本发明在传统协同过滤算法的基础上，建立了最近邻居优化选取方法，为个性化推荐提供更好的技术支持。

发明内容

[0006] 为解决最近邻居选取精度低、数据稀疏情况下推荐准确度低的问题，本发明提供一种更加准确的个性化推荐方法，在传统协同过滤算法的基础上，建立了基于最近邻信息的协同过滤推荐方法。为实现上述目的，本发明采取以下技术方案：

[0007] 一种基于最近邻信息的协同过滤推荐方法，包括下列步骤：

[0008] 步骤1：输入用户-项目矩阵中的评分数据集，确定目标用户u；

[0009] 步骤2：通过评分数据和皮尔森相似度，对目标用户与其他用户之间的可用度进行计算，并选出可用度较高的k`个用户，生成目标用户u的最近邻居候选集$N_u$`；

[0010] 步骤3:计算同一时间窗口θN_u`中的每个用户v与目标用户u之间的信任度tru(u,v):

[0011] a)根据同一时间窗口θ中用户v对某个项目i的偏好是否接近或远离目标用户u的偏好,认为在窗口θ中用户u对用户v产生了一致偏好和不一致偏好,分别记为$g_i^{\theta}(u,v)$和$b_i^{\theta}(u,v)$;

[0012] b)由于用户v可能会在窗口θ中为目标用户u评价多个项目,窗口θ中一致或不一致偏好的值将被计算为所有项目上该用户一致或不一致偏好的总和,分别记为$g^{\theta}(u,v)$和$b^{\theta}(u,v)$

[0013] c)给落在同一时间窗口θ中的所有评分分配相同的遗忘因子f,逐渐遗忘用户以前的偏好:对于时间窗口θ,分别赋予一致偏好遗忘因子f_g和不一致偏好遗忘因子f_b,一致和不一致偏好的总值分别记为$G^{\theta}(u,v)$和$B^{\theta}(u,v)$;

[0014] d)最后,根据用户v对用户u一致和不一致偏好总值的高低,计算用户v与目标用户u之间的信任度,一致性偏好总值越高或不一致偏好总值越低,信任度越高;

[0015] 步骤4:过滤掉可用度较高但是信任度较低的用户,选取信任度较高的K个用户,作为最近邻居用户,并生成最近邻居集N_u;

[0016] 步骤5:利用N_u的评分信息,计算出目标用户u对所有未评分项目的预测评分;

[0017] 步骤6:选取预测评分高的项目生成推荐项目列表。

[0018] 本发明由于采取以上技术方案,其具有以下优点:

[0019] (1)大多数现有用户相似度计算机制,如Pearson相似度,将一对用户之间的相似度计算为对称值,而实际情况中两个用户互相推荐的能力并不相同。本发明基于传统相似度计算方法,考虑用户相似度的不对称性和推荐可用性,保证最近邻居的推荐能力;

[0020] (2)大多数当前的邻居选择方法不考虑用户对不同项目的偏好的一致性,忽略用户偏好随着时间的推移的动态变化。本发明考虑用户对不同项目的偏好一致性,增加时间窗口和遗忘因子,更进一步地保证最近邻居与目标用户在不同时间段的偏好持续一致。

附图说明

[0021] 图1为基于最近邻居优化选取方法的协同过滤推荐方法流程。

具体实施方式

[0022] 本发明在传统协同过滤算法的基础上,建立了基于最近邻信息的协同过滤推荐方法,其中包括可用度计算模型和动态信任度计算模型两个关键的模型。

[0023] 为实现上述目的,本发明采取以下技术方案:

[0024] (1)可用度计算模型:传统的相似度计算方法主要依赖于共同评分项集,因此当两个用户的共同评分项目数较少时,得到的相似度与实际情况的偏差较大。同时,传统的相似度计算方法认为用户间的推荐能力对称,而实际情况中,两个用户对彼此的推荐能力并非相同。针对以上问题,本发明考虑到两个用户间共同评分项个数占目标用户的评分项个数的比例,结合传统的相似度计算方法,提出用户的可用度模型。

[0025] (2)动态信任度计算模型:在过去传统推荐算法中,计算相似度的方法中主要依赖的数据是用户之间的共同评分项值,然而在当今众多推荐系统当中,推荐技术主要面临的

关键问题之一就是数据稀疏性问题。因此，在较少共同评分数据的情况下，计算得出的相似度并不能够准确代表用户之间的相似程度，负面影响到了后续的推荐结果。此外，推荐系统中存在许多虚假或者恶意评分的用户，才使协同过滤推荐系统很容易受到攻击，对商品的不真实的评价或者别有目的评价内容，使得其误导其他用户对该商品的认知。本发明提出了信任度模型，一方面可以改善数据稀疏的问题，另一方面通过计算用户之间的信任度，选出那些值得信任的用户，排除那些恶意用户，进一步提高了推荐的准确率。

[0026] 本发明的基于最近邻居优化选取方法的协同过滤推荐方法，在计算动态信任度模型的过程中，根据实际情况改进了传统的信任模型，提出时间窗口和遗忘因子，从目标用户的角度来计算用户的信任值，并考虑时间因素来随着时间的推移来捕获用户的偏好变化，选择具有较高信任度的邻居。本发明将大大提高系统向目标用户做出推荐的准确度，图1显示了本发明的流程。具体实施步骤如下：

[0027] (1)输入用户-项目矩阵中的评分数据集，确定目标用户u。

[0028] (2)通过评分数据对目标用户与其他用户之间的可用度进行计算，公式如下：

[0029] $$ava(u,v)=\begin{cases}0, & I_v\subset I_u \ or \ sim(u,v)<0\\ |I_u\cap I_v|/|I_u|\times sim(u,v), & other\end{cases}$$

[0030] 其中ava(u,v)表示用户v对用户u的可用度，I_v、I_u分别表示用户v、u的评分项，sim(u,v)表示用户v、u的相似度。可以看出，当$I_v\subset I_u$即用户v的所有评分项都被用户u评价过时，用户v对于用户u来说毫无推荐能力，因此，此时ava(u,v)＝0。当sim(u,v)＜0时，说明用户u和v的相似性过低，因此，此时同样有ava(u,v)＝0。$|I_u\cap I_v|$表示用户u和v共同评分项的数目，当用户v中存在用户u未评分的项目时，通过考虑共同评分项数目占用户u的评分项数目的比例并结合传统的相似度计算方法，抵消传统相似度计算方法对共同评分项的过分依赖，当sim(u,v)相同时，$|I_u\cap I_v|/|I_u|$越大，说明计算传统相似度时使用的评分数据越多，因此用户v对用户u的可用度越高。

[0031] (3)选出可用度较高的K′个用户，生成目标用户u的最近邻居候选集N′(u)。

[0032] K′＝[ε×K]

[0033] 其中ε为最近邻居选取系数，ε∈{ε∈R|ε≥1}，[x]为取整函数，即不超过实数x的最大整数。N′(u)集中的用户一般具有较高的可用度，即与目标用户具有较高的相似度且具有较高的推荐能力。

[0034] (4)计算N′(u)中的所有用户w与目标用户u之间的信任度tru(u,w)。

[0035] a)本发明通过引入时间窗口和遗忘因子来逐渐遗忘用户以前的偏好。给定在时间t_c提供的特定评分，该评分落入的窗口的索引被标记为θ，

[0036] $\theta=[(t_c-t_s)/t_w]+1$

[0037] 其中t_s和t_w分别表示训练开始时间和窗口长度。在本发明提出的方法中，落在同一时间窗口θ中的所有评分将被分配相同的遗忘因子进行处理。

[0038] b)根据v对某个项目的偏好是否接近或远离目标用户u的偏好，本发明认为用户v对用户u产生了“好行为”(即一致偏好)或“坏行为”(即不一致的偏好)，分别记为$g_i(u-v)$和$b_i(u-v)$，其中的每一个部分均可以被量化为在(0,1)的范围内的连续值。本发明可以对窗口θ中的每个用户v的评分行为进行量化(记为$g_i^{\theta}(u,v)$或$b_i^{\theta}(u,v)$)：

[0039] $$g_i^{\theta}(u,v)=1-\frac{\left|R_{vi}^{\theta}-R_{ui}^{\theta}\right|}{R_{\max}-R_{\min}}$$

[0040] $$b_i^{\theta}(u,v)=\frac{\left|R_{vi}^{\theta}-R_{ui}^{\theta}\right|}{R_{\max}-R_{\min}}$$

[0041] 其中R_{max}和R_{min}分别表示推荐系统中的最大和最小评分值，R_{vi}^{θ}表示时间窗口θ中用户v对项目i的评分值，R_{ui}^{θ}表示时间窗口θ中目标用户u对项目i的评分值。

[0042] c)由于用户v可能会在窗口θ中为目标用户u评价多个项目，窗口θ中的好或坏行为的总值将被计算为所有项目上该用户的好或坏行为值的总和。

[0043] $$g^{\theta}(u,v)=\sum_{i\in I_{uv}^{\theta}} g_i^{\theta}(u,v)$$

[0044] $$b^{\theta}(u,v)=\sum_{i\in I_{uv}^{\theta}} b_i^{\theta}(u,v)$$

[0045] d)为了让时间更近的行为在计算用户的可信度方面获得较高的权重，本发明使用遗忘因子来逐渐遗忘用户以前的行为。具体来说，对于时间窗口θ，“好行为”和“坏行为”的总值(记为$G^{\theta}(u,v)$和$B^{\theta}(u,v)$)可以计算为：

[0046] $G^{\theta}(u,v)=G^{\theta-1}(u,v)\times f_g+g^{\theta}(u,v)$

[0047] $B^{\theta}(u,v)=B^{\theta-1}(u,v)\times f_b+b^{\theta}(u,v)$

[0048] 其中f_g和f_b是表示“好行为”和“坏行为”的遗忘因子，取值范围为(0,1)。

[0049] e)最后，目标用户u对用户v的信任度tru(u,v)计算如下，可以看出用户v进行的行为越好，该用户将获得的信任值越高。

[0050] $$tru(u,v)=\frac{G^{\theta}(u,v)+1}{G^{\theta}(u,v)+B^{\theta}(u,v)+2}$$

[0051] (5)过滤掉可用度较高但是信任度较低的用户，选取信任度较高的K个用户，作为最近邻居用户，并生成最近邻居集N(u)，从而得到了可用度和信任度均较高的用户作为目标用户的最近邻居，这部分用户与用户的偏好比较相似，且相似持续性比较高。

[0052] (6)利用最近邻居用户集的评分信息，计算出目标用户u对所有未评分项目的预测评分(例如对目标项目i的预测评分P_{ui})。主要利用下面的计算方法。

[0053] $$P_{ui}=\overline{R_u}+\frac{\sum_{v\in N(u)}\left(R_{vi}-\overline{R_v}\right)\times ava(u,v)\times tru(u,v)}{\sum_{v\in N(u)}\left|ava(u,v)\times tru(u,v)\right|}$$

[0054] (7)预测评分列表中的评分数据越高，表示本发明认为目标用户对该评分对应的项目的喜爱程度越高。因此选取预测评分高的项目生成推荐项目列表。

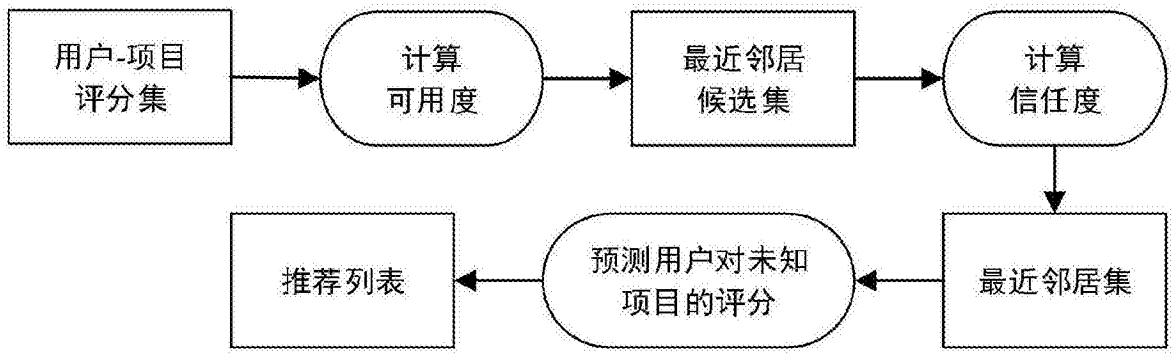


图1