

评分可信度条件下的协同过滤模型

杨兴耀, 于 炯, 吐尔根·依布拉音, 英昌甜, 闫 歌

(新疆大学 信息科学与工程学院, 新疆 乌鲁木齐 830046)

摘 要: 通过对用户信任度进行量化, 再从项目的角度进行调整, 最终获得了较为准确的评分可信度量度. 在此基础上, 建立了评分可信度矩阵, 并对提出的四种可信度相似性模型进行了优化. 实验比较结果表明, 基于不同的数据集, 新提出的相似性模型在合理的时间开销下, 相对于传统模型在项目预测准确性方面拥有出色的表现.

关键词: 推荐系统; 协同过滤; 可信度; 相似性模型

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2013)10-0025-04

Collaborative Filtering Models with the Credibility of Ratings

YANG Xing-yao, YU Jiong, TURGUN Ibrahim, YING Chang-tian, YAN Ge

(College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China)

Abstract: This paper quantifies the degree of user credit and adjusts it by specific items, and then gains accurate quantification on the credibility of ratings. In the case, the paper establishes the credibility matrix of ratings, and proposes four optimized similarity models based on credibility. The comparative experimental results, tested on different data sets, demonstrate that the newly proposed models have better performance in prediction accuracy of items at a reasonable cost in time, compared to traditional models.

Key words: recommender system; collaborative filtering; credibility; similarity model

1 引言

互联网技术与应用的快速发展, 在带来信息红利的时候, 也产生了信息过载问题, 如何从海量信息中准确快速过滤出用户所需要的信息, 推荐系统应运而生^[1].

推荐系统中不论采用何种模型, 其目标通常是最大程度地减小预测偏差, 提高推荐质量, 为了进一步改善推荐系统的性能, 有必要对用户的可信任程度进行评估, 简称可信度, 以确保模型中用到的数据的真实性、可靠性. 目前大多数的可信度研究工作主要从用户的角度展开, 例如文献[2]中从用户评价的热情度、公正度、准确度等方面对用户可信度进行了系统研究. 文献[3]中对系统中的用户进行分类处理, 将用户分为专家型用户、可信用户与兴趣相似用

户三个维度, 基于此利用岭回归分析这三类用户的主观倾向, 获得目标用户近邻集合, 并且最终产生推荐, 对比实验证明了推荐结果的质量得到显著提高. 文献[4]中对此在模型中引入“用户权威性”概念, 即通过用户评价过的项目数量、用户身份背景和项目的角度三个方面来综合计算用户的权威值和相似性, 最终使得推荐模型拥有较小的评分预测偏差. 同样的还有基于用户特征和项目属性的协同过滤推荐算法^[5].

本文在上述研究成果的基础上, 综合用户、项目两方面从评分的角度出发, 提出了评分可信度条件下的各种可信度协同过滤相似性模型, 介绍了推荐系统中的预测过程, 整个验证实验基于不同的数据集进行.

收稿日期: 2012-11-30; **修回日期:** 2013-01-15

基金项目: 国家自然科学基金项目(61063042, 61262088, 61063026); 新疆大学博士创新项目基金(XJUBSCX-2011007); 新疆自治区自然科学基金项目(2011211A011)

2 问题描述

2.1 评分矩阵

在推荐系统中,用户对所有项目的评价数据可以构成一个 $m \times n$ 阶的二维评分矩阵 R ,其中 m 行分别表示 m 个用户, n 列分别表示 n 个项目,矩阵中的元素 $r_{u,i}$ 表示用户 u 对项目 i 的评价值,该值体现了用户对项目的某种兴趣和偏好,通常值越大,偏好程度越高,若用户没有评价过某项目,则对应的评价值为空值, null, 见表 1.

表 1 评分矩阵样例

项目用户	i_1	i_2	...	i_i	...	i_n
u_1	$r_{1,1}$	$r_{1,2}$	...	$r_{1,i}$	...	null
...	...	...	...	...	...	...
u_u	$r_{u,1}$	$r_{u,2}$	...	$r_{u,i}$	...	null
...	...	...	...	...	...	...
u_m	$r_{m,1}$	null	...	$r_{m,i}$	...	$r_{m,n}$

根据评分矩阵 R ,可以利用不同的相似性模型获得用户之间的相似性,为了更好地度量用户相似性,本文下面将从评分的角度出发,通过建立相应信任度度量标准来量化不同评分的可信度.

2.2 用户信任度

评分作为用户对项目的评价值,为了衡量评价值的可信度,首先需要度量不同用户的信任度,以便为评分的可信度度量提供依据^[6].一般情况下,用户的信任度直接体现了用户在推荐网络中向其他用户提供可靠信息的能力,它具体体现在用户关于项目的评分数量和评分质量上.通常可以认为,一个用户的评分数量越多,评分质量越高,其信任度就越高.因此对于用户信任度,我们可以从评价项目范围和用户评分偏差两个方面来衡量:

(1) 评价项目范围:定义 $Q_u = \{i \in I \mid r_{u,i} \neq \text{null}\}$ 为用户 u 评价过的项目集合,其中 I 为系统的项目集合, $||$ 表示某一集合中的元素个数,则对于用户 u ,其评价项目范围 scope_u 可以表示为

$$\text{scope}_u = |Q_u| / |I|, \text{且 } \text{scope}_u \in [0,1] \quad (1)$$

这种表示方法虽然易于理解,但由于推荐系统中项目数目巨大的特点以及评分数据的稀疏性问题,该方法在实际中存在较大的问题,导致的直接结果是 scope_u 趋于 0,失去了引入的意义.为此,本文采用某个用户评价项目的最大数目 $|Q|_{\max}$ 作为标准,来取代 scope_u 中的分母 $|I|$,这样便较好地避免

了上述问题,同时也可以鼓励用户多对项目进行评分,评分数目越多的用户,它对于其他用户来说就越可信,即所谓的见多识广.最终,用户 u 的评价项目范围 scope_u 表示为

$$\text{scope}_u = |Q_u| / |Q|_{\max}, \text{且 } \text{scope}_u \in [0,1] \quad (2)$$

(2) 用户评分偏差:首先需要确定一个参考标准来衡量每个用户评分的偏差幅度,从用户的角度来看,当一个项目获得足够多的评价时,评价均值便表明了该项目真正的品质或质量.因此可以利用项目获得的评价均值来衡量一个用户评价的可信程度,它同时也反映了一个用户的专业水准,例如专家用户往往能根据项目的本质特征给出一个较为客观的评价,即评价值较为接近评价均值.对于项目 i ,它获得的评分均值表示为

$$\bar{r}_i = \frac{1}{|Q_i|} \sum_{u \in Q_i} r_{u,i} \quad (3)$$

这里 $Q_i = \{u \in U \mid r_{u,i} \neq \text{null}\}$ 为对项目 i 进行过评价的用户集合, U 为系统的用户集合.

下面根据项目的评分均值来量化不同用户的评分偏差程度,对于用户 u ,表达式如下:

$$\text{var}_u = \frac{1}{|Q|_{\max}} \sum_{i \in Q_u} \left(1 - \frac{|r_{u,i} - \bar{r}_i|}{r_{\max,i} - r_{\min,i}} \right) \quad (4)$$

式中, $r_{\max,i}$ 和 $r_{\min,i}$ 分别为项目 i 获得用户评价的最大值和最小值, $\text{var}_u \in [0,1]$. $|Q|_{\max}$ 引入的目的主要是为了避免用户评价次数较少情况下的一些偶然情况.可以看出,用户评价与评分均值越接近,评价项目数目越多,则用户评价越可信性,这一点符合实际情况.

获得用户 u 的评价项目范围 scope_u 和评分偏差 var_u 后,便可以定义用户信任度 Credit_u 了,如下:

$$\text{Credit}_u = \text{scope}_u \times \text{expert}_u =$$

$$\frac{|Q_u|}{|Q|_{\max}} \sum_{i \in Q_u} \left(1 - \frac{|r_{u,i} - \bar{r}_i|}{r_{\max,i} - r_{\min,i}} \right), \text{Credit}_u \in [0,1] \quad (5)$$

2.3 用户关于项目的信任度

用户信任度大致反映了一个用户评分的可信程度,但仅仅这样有些不够,因为一个用户的信任程度比较高,只是表明了这个用户值得信任,它的评分整体可信度较高而已,并不能代表用户每次的评分都是那么可信,那么客观.这一点反映在推荐系统中,就表现为用户关于单个项目评价的可信性问题上.

一般来说,用户针对项目的评价,从项目的角度同样可以感知用户评价的客观程度,比如某个项目

获得的评价一直比较高,而某个信任度较高的用户却给出了较低的评价,这只能说明用户在该项目的评价上有失偏颇.为此,我们可以沿用项目评分均值的概念,来尽量减小不公平评分对项目的客观评价所造成的影响,对于用户 u 关于项目 i 的评分信任度:

$$\text{Credit}_{u,i} = 1 - \frac{|r_{u,i} - \bar{r}_i|}{r_{\max,i} - r_{\min,i}} \quad (6)$$

上式中用户关于项目的评价越接近评价均值,则 $\text{Credit}_{u,i} \in [0,1]$ 值越大,它可以看成是用户信任度 Credit_u 在单个项目上的一个反映,这对于用户的信任度而言并没有大的影响,只是从项目的角度对其作了次再理解.

2.4 可信度矩阵

结合用户信任度以及用户关于具体项目的评分信任度,便可以评估矩阵 R 中评分 $r_{u,i}$ 的实际可信度值 $r_{u,i}'$,如下:

$$r_{u,i}' = \text{Credit}_u r_{u,i} \text{Credit}_{u,i} \quad (7)$$

可以看出 $r_{u,i}'$ 与 $r_{u,i}$ 对应,这样我们便得到了与评分矩阵维度相同的评分可信度矩阵 R' . 接下来,我们根据传统的相似性模型:PC,MSD,MAD 和 COS,提出基于可信度矩阵的可信度相似性度量模型:PC_{cre},MSD_{cre},MAD_{cre} 和 COS_{cre},以度量用户之间的相似性,表达式分别如下:

$$\text{PC}_{\text{cre}}(u,v) = \frac{\sum_{i \in I_{u,v}^*} (r_{u,i}' - \bar{r}_u') (r_{v,i}' - \bar{r}_v')}{\sqrt{\sum_{i \in I_{u,v}^*} (r_{u,i}' - \bar{r}_u')^2 \sum_{i \in I_{u,v}^*} (r_{v,i}' - \bar{r}_v')^2}} \quad (8)$$

$$\text{MSD}_{\text{cre}}(u,v) = 1 - \frac{1}{|I_{u,v}^*|} \sum_{i \in I_{u,v}^*} (r_{u,i}' - r_{v,i}')^2 \quad (9)$$

$$\text{MAD}_{\text{cre}}(u,v) = \frac{1}{|I_{u,v}^*|} \sum_{i \in I_{u,v}^*} (r_{u,i}' - r_{v,i}')^2 \quad (10)$$

$$\text{COS}_{\text{cre}}(u,v) = \frac{\sum_{i \in I_{u,v}^*} r_{u,i}' r_{v,i}'}{\sqrt{\sum_{i \in I_{u,v}^*} r_{u,i}'^2 \sum_{i \in I_{u,v}^*} r_{v,i}'^2}} \quad (11)$$

式中, $\bar{r}_u'$ 和 $\bar{r}_v'$ 分别为用户 u, v 的评分可信度均值, $I_{u,v}^* = \{i \in I \mid r_{u,i}' \neq \text{null} \wedge r_{v,i}' \neq \text{null}\}$ 为用户的可信度公共项目集合,它与通常的公共评分项目集合 $I_{u,v} = \{i \in I \mid r_{u,i} \neq \text{null} \wedge r_{v,i} \neq \text{null}\}$ 有所不同,此时显然有 $I_{u,v} \subseteq I_{u,v}^*$ 成立.

根据日常生活经验,用户间的相似性并不能通过一两个的公共评价项目来衡量,因为即使两次的用户评分完全相同,此时的相似性值非常大,也有可

能出于偶然.因此,有必要利用公共评价项目数目来对相似性值作一定程度的改进.通常的处理方法是取一个正整数常数阈值 μ ,让它与用户的公共评价项目数目 $|I_{u,v}|$ 进行比较,以 $\text{PC}_{\text{cre}}(u,v)$ 为例:

$$\text{PC}_{\text{cre}}(u,v) = \frac{\min(\mu, |I_{u,v}|)}{\mu} \times \text{PC}_{\text{cre}}(u,v) \quad (12)$$

可以看出,当公共评价项目数目超过 μ 时,相似性值保持不变,但当小于 μ 时,相似性值会随公共评价项目数目的减小而减小,从而达到优化用户相似性值的效果.这个阈值 μ 比较关键,后面实验部分会对其在特定环境下的具体取值作专门讨论.

3 预测过程与时间复杂度分析

3.1 预测过程

为目标用户提供若干个推荐项目,首先需要根据用户相似性值为目标用户构造用户近邻集合,然后收集集合中用户的评分来对目标用户未评分的项目进行评分预测.下面,本节来详细介绍整个预测过程:

输入:评分矩阵 R ,评分可信度矩阵 R' .

输出:目标用户 u 的 M 个推荐项目.

Step1. 利用矩阵 R' 和可信度相似性模型计算用户间的相似性.

Step2. 根据相似性大小选取相似性较大的用户来为目标用户 u 构造近邻用户集合 K_u ,然后通过 K_u 中用户对项目 i 的评分,利用预测函数计算获得用户 u 关于未评分项目 i 的预测评分 $P_{u,i}$,通常选择的预测函数为改进型的权重函数,如下:

$$P_{u,i} = \bar{r}_u + \frac{1}{\sum_{x \in G_{u,i}} \text{sim}(u,x)} \sum_{x \in G_{u,i}} \text{sim}(u,x) (r_{x,i} - \bar{r}_x) \quad (13)$$

式中, $\text{sim}(u,x)$ 为用户 u, x 通过可信度相似性模型获得的相似性值如 $\text{PC}_{\text{cre}}(u,v)$, $\bar{r}_u$ 和 $\bar{r}_x$ 为用户的评分均值, $G_{u,i} = \{x \in K_u \mid r_{x,i} \neq \text{null}\}$ 为 u 的近邻集合 K_u 中对项目 i 评分过的用户集合.此时, $P_{u,i}$ 存在的条件是 $G_{u,i} \neq \emptyset$,否则 $P_{u,i}$ 为空值,因为当 $G_{u,i} = \emptyset$ 时, K_u 中用户均未对 i 进行过评价,所以无法进行预测.

Step 3. 重复执行 Step 2,直到用户 u 的所有未评分项目预测完毕,然后进入 Step 4.

Step 4. 从用户 u 未评分但可以获得评分预测的项目集合 $E_u = \{i \in I \mid r_{u,i} = \text{null} \wedge P_{u,i} \neq \text{null}\}$ 中,选取预测评分值最大的 M 个项目向 u 进行推荐.

特殊情况下,当 E_u 中的项目个数不足 M 的时候,全部选取.

3.2 时间复杂度分析

时间复杂度分析从两个阶段来进行:

(1) 离线阶段即上述的 Step 1,以 PC_{cre} 模型为例:对于一个 $m \times n$ 维的可信度矩阵 R' 来说, PC_{cre} 模型需要进行 $4n$ 次加法操作和 $3n$ 次乘法操作来获得两个用户间的相似性. 这样由于需要计算系统中所有用户之间的相似性且该相似性具备对称性,所以总共需要计算 $(m-1) + (m-2) + \dots + 1 = m(m-1)/2$ 次,因此该阶段模型的时间复杂度为 $O(m^2n)$.

(2) 在线阶段即上述的 Step 2、3 和 4:获得全体用户间的相似性后,需要根据相似性值为目标用户选取一定数量 P 的相似用户作为用户近邻,然后通过预测函数经过常数次的加法乘法操作获得项目的评分预测,由于该阶段最多需要运行 n 次,然后根据预测结果为目标用户提供 M 个推荐项目,因此该阶段模型的时间复杂度为 $O(nP)$.

在实际中,在线阶段的时间花费即为用户获得项目推荐而需要等待的时间,这个时间对于系统性能的影响非常明显,对用户来说也比较敏感. 因为即使一个系统可以提供符合用户兴趣的项目,但是却需要等待较长的时间,大多数用户也会谨慎使用该系统,所以后面实验部分会对这段时间进行专门统计,以比较不同模型的运行效率.

4 结束语

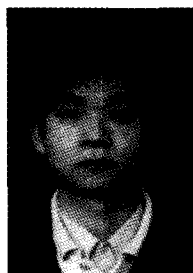
本文从评分的可信度量入手,结合用户信任度以及用户关于项目的评分信任度,获得了与用户一项目评分矩阵相对应的可信度矩阵,在此基础上,本文提出了各种不同的可信度相似性模型,并基于 MovieLens、NetFlix 数据集将其与传统的相似性模型进行对比. 实验结果表明了可信度相似性模型的性能优越性,下一步的研究方向之一是进一步研究影响评分信任度的关键因素,并建立合理的度量标准来对其进行更加精准的量化;另一个是信任度的更新问题,文中在计算评分信任度时,相当于仅仅考虑了一个周期内信任度没有发生变化的情形,而没有考虑在不同时间周期间信任度不断变化的特点,这与实际情况不是完全相符的,未来的工作应该考

虑这一因素.

参考文献:

- [1] Kwon H J, Hong K S. Personalized smart TV program recommender based on collaborative filtering and a novel similarity method[J]. IEEE Transactions on Consumer Electronics, 2011, 57(3): 1416-1423.
- [2] 陈春明,徐义峰. 协同过滤算法中一种改进的预测值计算方法[J]. 情报探索, 2009(8): 79-82.
- [3] 张尧,冯玉强. 协同过滤推荐中基于用户分类的邻居选择方法[J]. 计算机应用研究, 2012, 29(11): 4216-4219.
- [4] 吴杉. 引入用户权威性的分类协同过滤推荐算法的研究[J]. 焦作师范高等专科学校学报, 2008, 24(1): 78-79.
- [5] 陈志敏,李志强. 基于用户特征和项目属性的协同过滤推荐算法[J]. 计算机应用, 2011, 31(7): 1748-1750.
- [6] 景民昌,唐弟官,于迎辉. 基于专家信任优先的协同过滤推荐算法[J]. 图书情报工作, 2012, 56(11): 105-108.

作者简介:



杨兴耀 男, (1984-), 博士研究生. 研究方向为推荐系统、网络计算与云计算、可信计算.



于炯 男, (1964-), 教授, 博士生导师. 研究方向为网络安全、网络与分布式计算.

吐尔根·依布拉音 男, (1958-), 教授, 博士生导师. 研究方向为计算机应用、自然语言处理、软件工程.

英昌甜 女, (1989-), 博士研究生. 研究方向为网络调度算法、云计算.

闫歌 女, (1988-), 硕士研究生. 研究方向为网络安全、网络与云计算.