

基于评分区间相似性的协同过滤推荐算法

项 武

(浙江省丽水市财政局(地方税务局)信息中心, 浙江 丽水 323000)

摘 要: 研究协同过滤推荐系统处理大规模稀疏评价数据的精度问题, 针对目前余弦相似性处理极稀疏的用户评价矩阵不能获取满意推荐质量, 且处理大规模推荐数据时存在性能缺陷的情况, 提出基于评分区间相似性的协同过滤推荐算法. 将用户评价的数值范围进行定性划分, 以项目在各评价区间上的得分建模项目质量, 以用户在各评价区间内的评分建模用户评分取向. 使用余弦相似性度量方法衡量用户和项目间的相似程度, 进而确定最近邻居. 此方法的实现过程降低了稀疏性对最近邻发现的影响, 且算法具有快速处理大规模评价数据的能力. 使用标准的 MovieLens 数据集, 测试提出的推荐算法, 实验结果表明, 本算法相比传统的最近邻推荐算法具有更好的推荐精度.

关键词: 推荐系统; 协同过滤; 相似性度量; 余弦相似性

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2010)07-0125-04

A Collaborative Filtering Algorithm Based on Similarity of Rating Interval

XIANG Wu

(Information Center of Lishui Treasure Department, Lishui 323000, China)

Abstract: This work focused on recommended accuracy of the collaborative filtering system to process large-scale sparse data. The cosine similarity measurement and correlation similarity measurement could not handle the sparse rating matrix very well. And these methods undermined the recommended system performance when dealing with large-scale data. This paper provided a new approach based on the similarity of rating interval. The user rating range was divided into several qualitative intervals. According to the rating ratios, user preferences and item qualities were represented by vectors. Finally, this work used the cosine similarity method to determine the nearest neighbors. This approach reduces the impact of the sparse matrix and has the ability to deal with massive data. On MovieLens data set, experiment result shows that the algorithm has reliable and accurate performance.

Key words: recommendation system; collaborative filtering; similarity measurement; cosine similarity

1 引言

随着电子商务的发展, 服务提供商需要在保证产品质量的前提下, 清楚地掌握用户的需求与偏好, 并对用户提供按需服务. 因此, 优秀的推荐系统在电子商务应用中显得尤为重要, 也成为研究者的关注热点之一. 目前, 主流的电子商务网站, 如 eBay、淘宝等, 均在一定程度上为用户提供服务推荐. 电子商务推荐系统根据其所采用的推荐技术大致可以分为下面几类^[1]: 基于协同过滤技术的推荐系统、基于

内容过滤的推荐系统、基于知识发现推荐系统与交互式推荐. 其他的一些非主流推荐系统还包括基于贝叶斯网络、神经网络等的统计推荐系统^[2]. 最近邻协同过滤推荐是当前最成功的推荐技术, 算法基于评分相似的最近邻居的评分数据向目标用户产生推荐. 由于最近邻居对项目的评分与目标用户非常相似, 因此目标用户对未评分项目的评分可以通过最近邻居对该项目进行加权预测. 然而, 随着实际应用的开展, 最近邻推荐也面临多种挑战, 其中最突出的是用户-项目评价矩阵的稀疏性问题, 文献[3]中

指出,用户评分的项目一般不会超过项目总数的 1%,面对如此稀疏的评价矩阵,传统相似性度量方法存在各自的弊端^[4].同时应该注意,网络应用的普及使得用户评价信息量急剧增加,这对推荐系统的大量信息处理能力提出了新要求.

针对上述情况,本文提出一种依据用户评分的区间比率建模用户偏好的方法.将用户对项目的评价结果划分为多个区间段,使用评价区间内的评分比率建模用户的评分取向,使用余弦相似性度量评分取向,寻找最近邻居.本研究结果降低了评价矩阵的稀疏性影响,并具备处理大规模评价数据的能力,同时获得了较传统余弦推荐算法更加优秀的推荐质量.

2 评分区间划分方法

推荐系统的典型结构如图 1 所示,用户的评分记录作为系统的输入,从中提取用户对项目的评价偏好模型,并存储于知识库之中;对于用户的推荐需求,依据用户评价的偏好,选择项目库中最能够满足用户的项目,产生推荐结果.

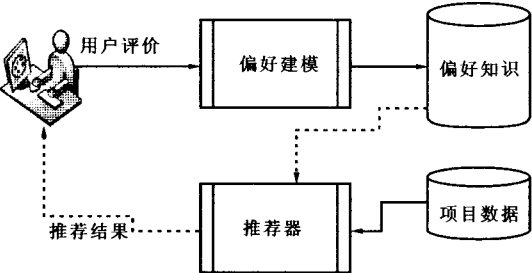


图 1 推荐系统结构图

推荐系统中,存在大量的用户-项目评分数据,数据结构中至少包含三个字段,分别表示项目编号($Item_i$)、用户编号($User_j$)和用户-项目评分(S_{ij}).表 1 给出了用户-项目评价矩阵的数据示意,其中存在大量的未评价信息,体现了用户-项目评价矩阵的稀疏性.

表 1 用户-项目评价矩阵数据示例

	User ₁	User ₂	...	User _k	...	User _M
Item ₁	S ₁₁	/	...	S _{1k}	...	/
Item ₂	/	S ₂₂	...	/	...	S _{2M}
...	...	...	...	...	...	...
Item _p	-	S _{p2}	...	...	...	...
...	...	...	...	...	...	...
Item _N	S _{N1}	...	...	...	...	/

考察用户-项目评价矩阵的特征,可以得到两点结论:首先,相同用户对不同项目的极其相似的评价,

其差异可以忽略,亦即,可将用户评价划分为若干等级,从而达到对评价结果范围的降维;其次,极少的用户-项目评价不能够说明用户的评价偏好或项目的本质优劣,需要依据规模性的统计结果来体现用户的评分取向或项目的实际质量.本文首先将用户评价划分为多个评价区间,然后依据用户评价结果的分布规律来建模用户评价取向和项目质量,进而获取用户-用户相似性与项目-项目相似性.

粗粒度的评价区间划分会降低用户或项目间的区分度,而过细力度的区间划分,会失去区间划分的意义,达不到对用户评价结果向量降维目的.本研究将用户评价结果范围定性划分 10 个区间,分别记为:极优 L_1 、优 L_2 、较优 L_3 、良好 L_4 、较良 L_5 、一般 L_6 、较差 L_7 、差 L_8 、很差 L_9 与极差 L_{10} .

表 2 给出了两种用户评价区间的划分规则.若用户对项目采用离散式评分方法,例如 5 分制评分(含 0.5 分),则可将 5 分定义为极优 L_1 ,而将 0 分或 0.5 分定义极差 L_{10} ;若用户的评分规则采用连续的百分制评定,可将 $[90, 100]$ 评价区间定义为最高等级的评价.实际应用中,可能存在用户采用离散式评分且不够 10 个等级的情况,此时可以降低评价区间的划分粒度^[5].

表 2 用户评价区间定性划分规则

	五分制	百分制	
		区间上限	区间下限
L_1	5	100	90
L_2	4.5	89	80
L_3	4	79	70
L_4	3.5	69	60
L_5	3	59	50
L_6	2.5	49	40
L_7	2	39	30
L_8	1.5	29	20
L_9	1	19	10
L_{10}	0.5/0	9	0

3 评分区间相似性度量方法

3.1 项目相似性度量

推荐系统中,同一项目会受到多个不同用户的评价,从统计学规律角度,多个用户的评价可反映项目本身的质量.因此,比较项目相似性时,可以忽略两个项目的评价用户集合的差异.

设推荐系统 RS 中,共存在 M 个用户(User), N 个项目(Item).且,项目 I_p ($0 < p \leq N$) 接受了 k ($0 < k \leq M$) 个用户的评价,评分结果记为 $SI_p =$

$\{S_{p1}, S_{p2}, \dots, S_{pk}\}$, 则项目 I_p 的质量 Q_p 可表示为

$$Q_p = \{IS_1, IS_2, \dots, IS_{10}\} \quad (1)$$

式中, $IS_i (1 \leq i \leq 10)$ 称为项目 I_p 在评价区间 i 上的得分率, 有

$$IS_i = \frac{\text{COUNT}(S_{pi} \in L_i)}{k} \quad 1 \leq i \leq 10, 1 \leq j \leq k \quad (2)$$

经过变换, 每个项目的用户评价结果均可映射为项目质量向量 Q , 显然 Q 为统一形式的 10 维向量, 且有 $\sum_{i=1}^{10} IS_i = 1$. 对项目 x 与项目 y , 其项目质量为 Q_x 与 Q_y , 依据余弦夹角公式, 项目相似性定义为

$$\text{sim}(x, y) = \frac{Q_x \cdot Q_y}{\|Q_x\| \|Q_y\|} \quad (3)$$

3.2 用户相似性度量

与项目相似性不同, 度量用户对项目评价结果时, 不同用户应该在相同的项目集合上进行比较, 否则, 由于项目质量的不同而导致的用户评价差异, 并不能说明用户评价取向的本质不同.

相关相似性度量方法中, 评价用户 i 与用户 j 的相似性之前, 首先获取两者的评价项目集合的交集 I_{ij} , 以 I_{ij} 为基础, 能够客观地反映用户 i 与用户 j 的评价取向差异. 然而, 由于用户 - 项目评价矩阵的极端稀疏性, I_{ij} 中不会拥有足够多的元素, 因此失去了统计评价的基础. 文献[4]中给出了度量用户相似性的新思路, 将 I_{ij} 定义为用户评价项目的并集, 而非交集. 这一思路可以保证用户评价项目的丰富性, 本研究将借鉴此方法.

设推荐系统 RS 中, 共存在 M 个用户 (User), N 个项目 (Item). 且, 用户 $U_i (0 < i \leq M)$ 共对 $k (0 < k \leq N)$ 个项目进行了评价, 用户 $U_j (0 < j \leq M, j \neq i)$ 共对 $w (0 < w \leq N)$ 个项目进行了评价. 评分结果记为 $SU_i = \{S_{1i}, S_{2i}, \dots, S_{ki}\}$, $SU_j = \{S_{1j}, S_{2j}, \dots, S_{wj}\}$, 取 $I_{ij} = SU_i \cup SU_j$, I_{ij} 的维度为 L , 则用户 i 相对用户 j 的评分取向记为 T_{ij} :

$$T_{ij} = \{US_1, US_2, \dots, US_{10}\} \quad (4)$$

式中, $US_a (1 \leq a \leq 10)$ 表示用户 U_i 相对用户 U_j 在评价区间 a 上的评分率, 有

$$US_a = \frac{\text{COUNT}(S_{ri} \in La)}{L} \quad 1 \leq a \leq 10, 1 \leq w \leq L \quad (5)$$

获得用户 i 相对用户 j 的评分取向 T_{ij} 与用户 j 相对用户 i 的评分取向 T_{ji} 后, 可仿照式(3), 定义两用户相似性为

$$\text{sim}(i, j) = \frac{T_{ij} \cdot T_{ji}}{\|T_{ij}\| \|T_{ji}\|} \quad (6)$$

4 仿真过程及结果分析

4.1 算法实现

基于项目评价区间余弦相似性的协同过滤推荐算法的实现可分为两个主要部分: 首先, 通过推荐系统已有的用户 - 项目评价记录, 进行机器学习, 判定用户的最近邻居; 其次, 依据第一步的学习结果, 对未评价项目给出评分预测. 具体而言, 本文的推荐算法可以划分为 5 个关键步骤:

输入: 系统已有的用户 - 项目集合 UI , UI 包含 M 个用户对 N 个项目的 s 项评分, 且任意 $n \in N$ 或 $m \in M$, UI 中至少包含一项相关评价记录; 待推荐的用户 - 项目实例 R_{mn} , 求解用户 m 对项目 n 的评分;

输出: 用户 m 对项目 n 的预测评分 S_{mn} ;

核心步骤:

(1) 依据式(1)、式(2), 获取项目 $p (1 \leq p \leq N)$ 的评价质量向量 Q_p , 而后进行迭代, 获取所有 N 个项目的项目质量;

(2) 依据式(3), 度量所有项目的相似性, 并记录结果为 IS , IS 为 N 阶方阵;

(3) 从 UI 中依次提取用户 i 与用户 $j (1 \leq i \leq M, 1 \leq j \leq M, i \neq j)$, 求评价项目交集 I_{ij} , 对于用户 i 未评价的项目 $p (p \in I_j)$, 依据项目相似性结果 IS 估计用户 i 对项目 p 的评分 S_{pi} [4]. 依据式(4)、式(5)获取用户 i 相对用户 j 的评分取向 T_{ij} . 而后迭代此步骤直至获取所有用户间的评分取向;

$$S_{pi} = \frac{\sum_{l \in I_i} \text{Sim}(p, l) \times S_{li}}{\sum_{l \in I_i} \text{Sim}(p, l)} \quad (7)$$

(4) 使用式(6)计算所有用户间的相似性, 并记录为 US , US 为 M 阶方阵;

(5) 依据输入, 在已经对项目 n 进行过评价的用户中寻找与用户 m 最为接近的 k 个邻居, 依据 k 邻居的评分, 计算用户 m 对项目 n 的评分估值 S_{mn} :

$$S_{mn} = \frac{\sum_{j \in \text{Top-K}} \text{Sim}(m, j) \times S_{pj}}{\sum_{j \in \text{Top-K}} \text{Sim}(m, j)} \quad (8)$$

算法中, 步骤(1) ~ (4) 完成机器学习阶段的任务, 相比传统的余弦相似性度量方式, 本算法在评价区间的基础上首先将项目评分和用户评分均化简为统一形式的 10 维向量, 表示用户评分取向和项目质量, 而后进行余弦相似性度量. 项目质量向量的映射

过程完全不受评价矩阵稀疏性影响,而用户评分取向的映射过程,由于使用了两个用户的评价项目并集作为度量基础,同样最大限度地消除了稀疏性带来的负面影响.同时,在面对大量用户-项目评价数据时,本算法可以有效降低相似性度量的系统开销,无论用户-项目间存在多大规模的评分数据,最终均将被转换为一个 10 维向量,以此向量计算相似性显然比在原始数据基础上计算相似性更加高效;算法的步骤(5)可产生推荐结果,属于评价推测过程,其中涉及的式(8)属于平局加权的推荐产生策略,在同类研究中得到广泛的应用^[6-7].

4.2 实验与分析

推荐系统的推荐质量评价标准主要有统计精度度量和决策支持精度度量方法两类^[3].其中,平均绝对偏差 MAE(Mean Absolute Error)是一种常用度量方法.

MAE 通过计算预测的用户评分与实际的用户评分之间的偏差度量预测的准确性,越小的 MAE 值意味着越高的推荐质量.对 N 个待评价项目,设预测的用户评分集合为 $\{s_1, s_2, \dots, s_N\}$,而实际评分集合为 $\{r_1, r_2, \dots, r_N\}$,则 MAE 为

$$MAE = \frac{\sum_{i=1}^N |s_i - r_i|}{N} \quad (9)$$

参考本领域的主要研究文献^[4,8],通常使用 MovieLens 站点提供的测试数据集作为算法推荐准确性的验证基准.为验证本文算法的推荐质量,本文同样选择 MovieLens 数据库为验证平台.

MovieLens 站点是一个基于 Web 的研究型推荐系统(<http://movielens.umn.edu>),用于统计用户对电影的评分并提供电影推荐. MovieLens 对研究人员提供各种规模的标准推荐测试数据集,目前,最完整的数据集涵盖了 71 567 名用户对 10 681 部电影的 10 000 000 条评分记录.相关文献中,在评测算法推荐准确性时,均未对样本的选取方式进行详细说明,原始数据和测试数据的划分方式亦比较模糊.不同的数据基础使得本文无法直接利用其实验结果.现选用 MovieLens 官方提供的标准 u1 数据文件模拟推荐系统的已有数据,其中包含 80 000 条训练数据和 20 000 条测试数据.

图 2 给出了文献^[4]中实验结果,该研究表明,基于项目评分预测的推荐方法要优于一般的基于余弦相似性和相关相似性的协同过滤推荐算法.本文是在传统余弦相似性度量的基础上,提出将评价结

果依据区间进行划分,完成用户评分(项目评分)向用户评价取向(项目质量)向量的映射,本质上是对基于余弦相似性的协同过滤推荐的改进,同时又参考了文献^[4]中的用户评价项目并集策略,故将对比目标选定为余弦相似性协同推荐、相关相似性协同推荐和文献^[4]的算法.

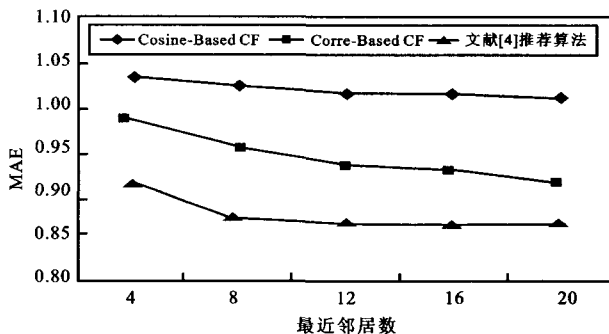


图2 文献^[4]的仿真结果

在上述数据基础上实现四种推荐算法,实验结果如图3所示.最近邻居个数的选择从4开始,递增至20,步长为4,整个过程中,本文提出的基于评价区间余弦相似性的推荐算法获得了更低的平均绝对偏差,算法相比传统的余弦相似性与相关相似性方法的推荐精度优势明显,较文献^[4]中算法的推荐质量高.

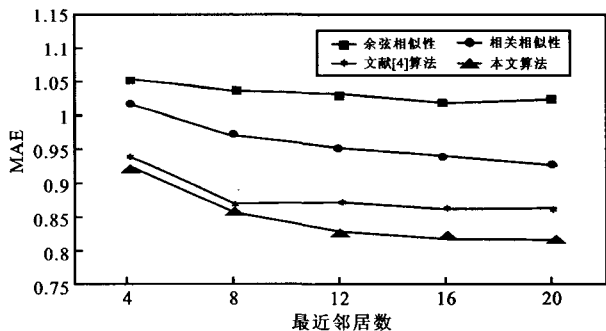


图3 推荐算法推荐精度比较

实验结果可以证明,将用户评价进行按区间划分,进而获取用户的相似性的推荐思路可以获得更好的推荐准确度.同时应该注意,本文提出的算法在有效地规避用户-项目评价矩阵的稀疏性问题的同时,还能够较好地适应大规模的评价数据,保证推荐系统的推荐效率.

5 结束语

本文首先给出了余弦相似性、相关相似性和修正的余弦相似性度量方法,这些方法无法适应项目-用户评价矩阵的极端稀疏性的问题,且面对大规

(下转第 132 页)

以及报文长时间不到的情况进行了特殊处理,保证了 97% 以上的重组率。

6 结束语

本算法的低开销、高灵活度,为繁杂的应用层协议分析系统节省了宝贵的系统资源。但是,由于 Libpcap 在高速网络下的数据包捕获能力有限,导致 TCP 报文段的大量丢失,一定程度上影响了算法的效率与数据的还原率。因此,下一步将研究更为高效的抓包机制,提高系统的整体性能。目前,本算法实际运用于自主开发的网络行为管理系统中,运行稳定。

参考文献:

- [1] 韩东东,杨立国,赖杰贤. 校园网环境下的电子邮件监控与审计系统[J]. 中国教育信息化,2008(5):53-55.
- [2] 姜恩勇,岳猛. 一种可扩展的 webmail 监视系统的设计

与实现[J]. 计算机安全,2009(4):18-20.

- [3] 邓子宽,范明钰,王光卫,等. Snort 入侵检测系统中 TCP 流重组的研究[J]. 信息安全与通信保密,2007(2):65-67.
- [4] 熊兵,陈晓苏,陈宁. A real-time TCP stream reassembly mechanism in high-speed network[J]. 西南交通大学学报:英文版,2009,17(3):185-191.
- [5] Ethan Blanton, Mark Allman. On making TCP more robust to packet reordering[J]. ACM,2002,32(1):20-30.
- [6] Richard Stevens W. TCP/IP 详解 卷 1:协议 [M]. 北京:机械工业出版社,2000:174-269.
- [7] Charles M Kozierok. TCP/IP 指南,卷 1:底层核心协议 [M]. 北京:人民邮电出版社,2008:511-587.

作者简介:

赵启升 男,(1970-),高级工程师.研究方向为计算机网络.

李存华 男,(1963-),博士.研究方向为数据挖掘及图形图像处理.

(上接第 128 页)

模评价数据时性能堪忧.针对此情况,本文提出一种基于评价区间余弦相似性的协同过滤推荐算法.首先将用户评价的取值范围划分为 10 个定性评价区间,依据用户评价落入各区间的比率建模项目质量和用户评价取向,从而降低稀疏性和大规模数据对推荐系统造成的负面影响.实验结果表明,基于用户评价区间的协同过滤推荐算法可明显提高推荐系统的推荐质量。

参考文献:

- [1] 朱岩,林泽楠. 电子商务中的个性化推荐方法评述[J]. 中国软科学,2009(2):183-192.
- [2] 张海燕,丁峰,姜丽红. 基于模糊聚类的协同过滤推荐方法[J]. 计算机仿真,2005,22(8):144-147.
- [3] Sarwar B, Karypis G, Konstan J, et al. Item-Based collaborative filtering recommendation algorithms[C]// Proceedings of the 10th International World Wide Web Conference. Horgliorg, 2001: 285-295.
- [4] 邓爱林,朱扬勇,施伯乐. 基于项目评分预测的协同过

滤推荐算法[J]. 软件学报,2003,14(9):1621-1628.

- [5] Arwar B, Karypis G, Konstan J, et al. Analysis of recommendation algorithms for E-commerce[C]// Processing of 2nd ACM Conference on Electronic Commerce. New York, 2000:158-167.
- [6] Kim BM, Li Q, Park C S, et al. A new approach for combining content-based and collaborative filters[J]. Journal of Intelligent Information System, 2006,27(1):79-91.
- [7] Chedrawy Z, Abidi SSR. An adaptive personalized recommendation strategy featuring context sensitive content adaptation [C]//Proceedings of Adaptive Hypermedia and Adaptive Web-based Systems. Germany, 2006, 4018: 61-70.
- [8] 张锋,常会友. 使用 BP 神经网络缓解协同过滤推荐算法的稀疏性问题[J]. 计算机研究与发展,2006,43(4):667-672.

作者简介:

项武 男,(1968-),硕士,高级工程师.研究方向为自动控制、人工智能。