

融合信任相似度的高校图书馆 个性化推荐研究*

林晓霞^{1,2} 刘敏¹ 杨晓东^{1,2} 徐尧¹

(1. 山东科技大学计算机科学与工程学院, 青岛 266590; 2. 山东科技大学信息工程系, 泰安 271019)

摘要: 随着数字图书馆的发展, 高校图书馆的个性化推荐问题成为热点, 而传统基于评分的推荐算法不适合高校图书馆的推荐需求。针对以上情况, 本文在研究传统协同过滤算法的基础上, 首先利用借阅记录中的数据构建评分矩阵, 然后融合信任相似度对用户相似度进行改进, 同时引入对新生、新书推荐的解决方法, 最后通过对山东科技大学图书馆数据集的对比实验, 结果表明, 提出的方法有效解决了数据稀疏和冷启动的问题, 具有更高的推荐精度。

关键词: 高校图书馆; 个性化推荐; 协同过滤; 信任相似度

中图分类号: G252

DOI: 10.3772/j.issn.1673-2286.2018.08.003

随着图书馆馆藏数量的快速增长, 用户越来越难以快速、准确地获取其需求的全部文献资源。为更好地满足用户需求, 馆藏资源浏览及检索过程中的个性化推荐服务应运而生, 即个人图书馆服务, 如美国康奈尔大学的MyLibrary个性化服务。目前, 推荐算法主要包括基于内容的推荐^[1]、基于协同过滤的推荐^[2]、基于关联规则的推荐^[3]、基于网络结构的推荐^[4]及混合推荐^[5]等, 而在图书馆个性化推荐中基于协同过滤的推荐算法应用最广泛, 且对于提升和改进图书馆服务起到一定作用。但是, 基于协同过滤的推荐算法较依赖用户对图书的评价或阅读反馈, 通常用户对所借阅图书的评价不多, 实际阅读发生的图书数量也有限(即数据稀疏问题), 并且不断加入新生和新书(即冷启动问题), 对协同过滤推荐算法的准确性与可用性都带来了很大挑战。

为应对数据稀疏问题, 刘健^[6]结合关联语义链与传统协同过滤算法, 充分挖掘用户信息, 缓解数据稀疏的问题; Gong^[7]通过聚类平滑的方法填充评分矩阵, 以降低评分数据的稀疏; Xu等^[8]通过融合用户和项目的预

测评分, 填充评分矩阵。以上方法虽然在一定程度上解决了数据稀疏的问题, 但是推荐精度有待提高。为解决冷启动问题, 王成等^[9]通过引入项目热度对用户相似度进行改进, 来提高推荐精度; 郑孝遥等^[10]在传统协同过滤算法中引入信任度, 算法在推荐精度上有较大提高; Guo等^[11]将显式或隐式的评分和信任用户相结合, 对算法进行改进; 郭艳红^[12]运用平均值法, 作为新项目的预测值, 改善了冷启动问题; Ahn^[13]提出一种新的启发式相似度, 改善了冷启动下的推荐性能。

这些方法一定程度上解决了已有算法存在的一些问题, 但是算法比较单一, 实际应用效果并不理想。

本文在上述研究的基础上, 提出一种融合信任相似度的个性化推荐方法, 即利用借阅记录中的借阅时长、借阅方式等构建用户-图书评分矩阵, 引入信任相似度对传统的相似度进行改进, 根据邻居用户计算得到图书推荐度, 结合新书的推荐度, 实现图书推荐。实验结果表明, 该改进方法降低了评分稀疏, 提高了图书推荐的精度。

*本研究得到全国统计科学研究项目“基于云计算的高精度空间插值方法与应用”(编号: 2016LZ12)和泰安市科技计划项目“智慧泰安背景下的高性能空间插值平台与泰安市高精度DEM和资源环境数据构建”(编号: 201630576)资助。

1 研究背景

协同过滤推荐最早被运用在电子商务领域^[14],包括基于用户的协同过滤推荐^[15]和基于项目的协同过滤推荐^[16]。

数据稀疏问题是协同过滤存在的最大问题,评分少将会导致评分数据稀疏,进而使用户相似度计算不准确,无法为用户找到最近邻居,影响推荐精度。在图书馆中,用户和图书的数量十分庞大,但是用户借阅记录中的评分极少,评价矩阵存在严重稀疏问题,无法准确为用户推荐图书。

冷启动问题是协同过滤存在的另一个问题,在高校中,每年都会有新用户,新用户没有借阅记录,在计算新用户相似度时,没有可依据的数据,因此新用户得不到良好的推荐;图书馆新购入的图书,没有被用户借阅过,也无法被推荐。

本文对传统协同过滤算法进行改进,通过借阅信息构建用户-图书评分矩阵来缓解数据稀疏和冷启动问题,融入信任相似度来提高用户的相似度,以解决存在的问题,提高个性化推荐精度。

2 融合信任相似度的个性化推荐模型构建

本文的图书馆借阅数据包括用户信息表(姓名、性别、单位)、图书信息表(书名、出版社、作者、索书号)、借阅信息表(借阅时长点、借阅次数)三个维度,利用图书馆借阅数据中这三个维度的信息,经过分析处理,建立个性化推荐模型,将符合用户需求的图书推荐给用户,为用户提供个性化阅读服务。融合信任相似度的个性化推荐模型如图1所示。

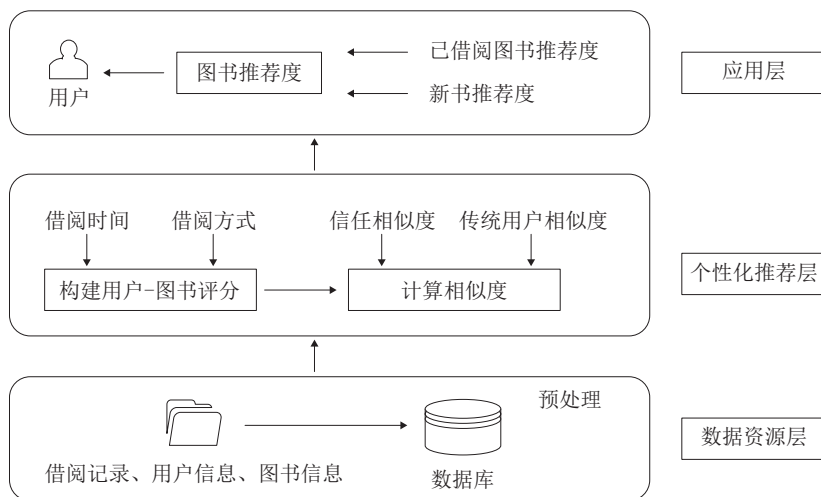


图1 融合信任相似度的个性化推荐模型

改进后的融合信任相似度的个性化推荐模型主要包括数据资源层、个性化推荐层和应用层。其中,数据资源层是将借阅记录、用户信息、图书信息进行预处理;个性化推荐层用于构建评分和计算相似度,寻找最近邻居;应用层是根据推荐度的高低将图书推荐给用户。

2.1 融合信任相似度的个性化推荐算法的步骤

传统的协同过滤个性化推荐算法由于评分数据稀疏,计算得出的用户相似度不准确,因而推荐精度

很低。根据上面的分析,本文将对传统的推荐算法改进,提出一种融合信任相似度的个性化推荐算法,可以将融合信任相似度的个性化推荐算法概括为以下步骤:①根据借阅记录中的借阅信息,构建用户-图书评分矩阵;②根据构建的评分矩阵,使用余弦相似度计算出用户相似度;③分别计算显性信任度、隐性信任度、新用户信任度,得出综合信任度;④将用户相似度与综合信任度相融合,计算出改进后的用户相似度;⑤根据改进后的用户相似度,找到用户的最近邻居;⑥根据最近邻居借阅的图书,得出图书的推荐度,与新书推荐度融合,得到最终的图书推荐度。其中,输入内容为用户借阅记录、用户信息、图书信息、

信任系数 (α_1 、 α_2 、 α_3)，用户相似度因子 (β)、邻居用户数，输出内容为图书的推荐度。

2.2 用户-图书评分矩阵构建

在实际应用中，高校图书馆虽然提供了评分功能，但是用户对图书的评分很少，甚至没有，评分矩阵比较稀疏，因此可以通过借阅信息来构建用户-图书评分矩阵。在图书馆的借阅系统中可以获取用户对图书的借阅时长及借阅方式等相关数据，本文以此为基础，构建用户-图书评分矩阵。

(1) 借阅时长。借阅时长能够反映用户对图书的喜爱程度，即可以体现用户对图书的评分。借阅时长有长有短，如果用户对一本书非常感兴趣，则借阅的时长会相对较长，对这本书的评分也会越高；如果用户对借阅的图书不感兴趣，则借阅时长较短，对图书的评分较低。本文的借阅时长以天为单位。同时，在高校中存在用户同质性低的情况，如教师、研究生、本科生对图书的借阅权限不同，因此，需要根据用户类型给予用户不同的借阅因子。若高校中有 n 类用户，借书时长依次为 $d_1, d_2, \dots, d_n$ ，最短借书时长为 $d_{\min} = \min(d_1, d_2, \dots, d_n)$ ，用 D_i 表示用户 i 的借阅因子，即 $D_i = \frac{\log_e d_{\min}}{\log_e d_i}$ 。

用 R_{ij} 表示用户 i 对图书 j 的借阅时长所得的评分， T_{ij} 表示用户 i 对图书 j 的借阅时长， $T_{\min}$ 表示该图书被借阅的最短时长， $T_{\max}$ 表示该图书被借阅的最长时长。

$$R_{ij} = \frac{T_{ij} - T_{\min}}{T_{\max} - T_{\min}} \times D_i \quad (1)$$

(2) 借阅方式。用户的借阅方式可以分为初借和续借。初次借阅时用户没有目的，只有所借阅的图书对其有意义，用户才会续借。因此，初借和续借两种借阅方式的比重不同，用 B_{ij} 表示用户 i 对书 j 的借阅方式所得的评分， B_{ij} 可以根据历史数据统计分析设置。

本文通过借阅时长、借阅方式两者之和构建评分矩阵，用 R'_{ij} 表示构建的评分矩阵，即 $R'_{ij} = R_{ij} + B_{ij}$ ，所构建的用户-图书评分矩阵如表1所示。

表1 用户-图书评分矩阵

	图书1	图书2	图书3	...	图书 n
用户1	0.4	0.7	0.6		0
用户2	1.0	0.7	0.9		0.9
用户3	0.3	0.5	0.5		0.5
...				R'_{ij}	
用户 m	0.7	0	0.5		0.1

2.3 用户相似度

个性化推荐算法中，相似度是核心。传统的协同过滤算法根据用户相似度对未评分图书进行预测。实际上，除用户间的相似关系外，用户间隐含的信任关系也会影响图书的推荐。在用户之间，也会存在一种信任与被信任的关系，因此本文引入信任相似度，将信任相似度融入用户相似度中。

(1) 显性信任度。显性信任指用户之间存在很明显的信任关系，本文即共同借阅过相同书籍的用户间存在显性信任关系。两个用户共同借阅的图书数量越多，显性信任度越大。

用 B_u 表示用户 u 的借阅书目集， B_v 表示用户 v 的借阅书目集， B_m 表示与用户 v 具有最多共同借阅书目的用户借阅书目集， $B_u \cap B_v$ 表示用户 u 和用户 v 的共同借阅书籍数量。

令 $T_{u,v}^D$ 表示用户 u 和用户 v 之间的直接信任关系，即用户 u 信任用户 v 的程度，直接信任度表示为公式 (2)。

$$T_{u,v}^D = \frac{B_u \cap B_v}{B_u \cap B_m} \quad (2)$$

其中， $B_u \cap B_m$ 表示用户 u 与用户 v 具有最多共同借阅书目的用户的借阅书目集的交集。

信任度的范围 $[0, 1]$ ，当 $B_u = B_m$ 时，与用户具有最多共同借阅书目的用户信任度等于1； $B_u = B_v$ 时，信任度等于0；其他情况下直接信任关系的信任度小于1。

(2) 隐性信任度。很多用户之间没有共同借阅过相同书籍，不存在显性信任关系，无法直接计算信任度。本文利用信任的传递性来计算两个用户之间的隐性信任度，可以根据用户之间的显性信任关系推导出用户之间隐含的信任关系，通过共同朋友的借阅信息，来实现个性化推荐。

用 $T_{u,v}^F$ 表示用户 u 与用户 v 的隐性信任， i 表示用户之间的共同朋友， m 表示两个用户之间共同的朋友数量。

$$T_{u,v}^F = \frac{\sum_{i=m}^1 T_{u,i}^D \times T_{i,v}^D}{m} \quad (3)$$

(3) 新用户信任度。新生刚入校，借阅图书时没有目标，通常情况下更愿意信任同专业的同学，且新生在填写注册信息时会有具体的学院专业信息。因此，本文提出新用户信任，可以在一定程度上对新生进行个性化推荐。新用户信任用 $T_{u,v}^S$ 表示，通过历史统计数据进行分析，相同专业的新用户信任为0.6，不同专业的新用户信

任为0.4。

(4) 综合信任度。将上述3种信任进行加权融合, 得到用户的综合信任度。用 $T_{u,v}^T$ 来表示综合信任度。

$$T_{u,v}^T = \alpha_1 T_{u,v}^D + \alpha_2 T_{u,v}^F + \alpha_3 T_{u,v}^S \quad (4)$$

其中, α_1 、 α_2 、 α_3 分别表示显性信任、隐性信任、新用户信任的权值, 并且 $\alpha_1 + \alpha_2 + \alpha_3 = 1$ 。

(5) 相似度计算。将信任相似度与用户相似度进行融合, 得到最终的用户相似度。 $sim_{new}(u, v)$ 表示改进后的用户相似度, $sim(u, v)$ 表示用户相似度, $T_{u,v}^T$ 表示信任相似度。

$$sim_{new}(u, v) = \beta sim(u, v) + (1 - \beta) T_{u,v}^T \quad (5)$$

其中, $0 \leq \beta \leq 1$ 。

2.4 图书推荐度

图书推荐度指图书的可推荐程度, 计算出每本图书推荐度后, 将推荐度高的图书推荐给用户。本文将图

书分为已借阅图书和新书两类。

(1) 已借阅图书的推荐度。已借阅图书的推荐度指已经被借阅过的图书的可推荐度。用 $P_{u,i}$ 表示被借阅图书的推荐度, 若用户已借阅多本图书, $\bar{R}_u$ 表示用户 u 对图书的平均评分, $\bar{R}_v$ 表示用户 v 对图书的平均评分。

$$P_{u,i} = \bar{R}_u + \frac{\sum_{v \in N} sim_{new}(u, v) \times (R_{v,i} - \bar{R}_v)}{\sum_{v \in N} |sim_{new}(u, v)|} \quad (6)$$

(2) 新书推荐度。图书馆每年都会引进一批新书, 这些新书没有用户的评分及借阅记录, 无法推荐给用户。同一院系中, 虽然专业不同, 但基础课程差别不大, 用户的借阅需求基本相同。本文利用中图分类号, 进行新书推荐。中图分类法将相同学科、相同主题的图书归为一类。在图书信息表中, 通过索书号中的中图分类号可以了解用户借阅图书类目, 根据用户以往的借阅记录分析各类图书所占的比重。以山东科技大学泰安校区的3个院系为例, 根据用户的以往借阅记录, 分析得出不同类型的图书在各个院系所占的比重, 如表2所示。

表2 不同类型的图书在各个院系所占比重示例

	A	B	C	D	E	F	G	H	I	J	K	N	O	P	Q	R	S	T	U	V	X	Z
信息系	0.05	0.00	0.00	0.05	0.00	0.03	0.00	0.05	0.05	0.02	0.04	0.00	0.10	0.00	0.01	0.00	0.01	0.45	0.01	0.02	0.04	0.02
经管系	0.10	0.10	0.10	0.05	0.20	0.40	0.00	0.05	0.02	0.01	0.02	0.01	0.01	0.10	0.00	0.20	0.01	0.02	0.00	0.00	0.10	0.50
文法系	0.03	0.05	0.03	0.40	0.03	0.04	0.10	0.05	0.04	0.05	0.05	0.00	0.00	0.00	0.01	0.01	0.00	0.02	0.02	0.01	0.01	0.05

(3) 图书推荐度。融合新书的推荐度和被借阅图书的推荐度, 得到最终的图书推荐度。设 P_z 表示图书推荐度。

$$P_z = P_a + P_{u,i} \quad (7)$$

其中, P_a 表示新书的推荐度。

3 实验结果及其分析

3.1 实验数据集

为验证本文所提出推荐算法的可用性, 选取山东科技大学图书馆的后台借阅记录数据进行试验, 将数据范围定位为7 197名学生的借阅记录, 时间跨度为2012年9月—2017年7月, 共100 986条借阅记录, 涉及283 568本图书。将这些数据划分为训练集和测试集, 训练集为2012年9月—2016年7月的借阅数据, 测试集为2016年9月—2017年7月的借阅数据。以随机方式从信

息系、经管系、文法系中各抽取30名学生的借阅记录进行试验。

3.2 评价标准

推荐准确率是衡量算法推荐能力的标准之一, 本文使用推荐准确率(MAE)来验证算法的推荐质量。MAE越小, 则表示推荐越准确。

$$MAE = \frac{\sum_{u,i \in N} |R_{ui} - P_{ui}|}{|N|} \quad (8)$$

其中, N 为进行预测图书的数量, R_{ui} 为用户的实际评分, P_{ui} 为预测的评分。

3.3 实验结果及分析

本文主要分析设计参数 α_1 、 α_2 、 α_3 、 β 、不同邻居数对推荐准确率的影响, 以及所构建评分矩阵的有效性

验证,设计5组不同的实验,选择最优参数,最后与其他模型进行对比分析。

(1) 评分矩阵的有效性验证。用评分数据的稀疏来衡量用户-图书评分矩阵的有效性,计算得到改进前的评分数据稀疏为96.19%,通过构建用户-图书评分矩阵,得出评分数据稀疏仅为0.90%。

可以看出,本文提出的构建方法,极大地改善了评分数据的稀疏,解决了高校图书馆评分严重稀疏的问题。

(2) 信任参数 α_1 、 α_2 对实验结果的影响。由于对于所有老用户, $\alpha_3=0$;因此将信任参数分成10组分别进行验证,对应的MAE值如表3所示。

表3 参数组合对应的MAE值

参数组合	信任参数 α_1	信任参数 α_2	MAE
1	0.9	0.1	0.099 955 907
2	0.8	0.2	0.097 778 050
3	0.7	0.3	0.096 429 999
4	0.6	0.4	0.094 871 995
5	0.5	0.5	0.091 050 847
6	0.4	0.6	0.090 318 914
7	0.3	0.7	0.086 298 256
8	0.2	0.8	0.078 115 656
9	0.1	0.9	0.081 121 317
10	0	1.0	0.090 283 469

实验表明,在其他参数默认的情况下,信任参数组合8时MAE值最小,即 $\alpha_1=0.2$, $\alpha_2=0.8$ 时,MAE值最小,推荐效果最好。

(3) 相似度权重因子 β 的取值。此实验验证 β 取值的大小对推荐效果的影响,推荐效果用也MAE来衡量,MAE值越小,推荐质量也就越高。为找到MAE的最小值,实验将在训练集上以不同值进行训练,在测试集中进行测试,测试结果如图2所示。

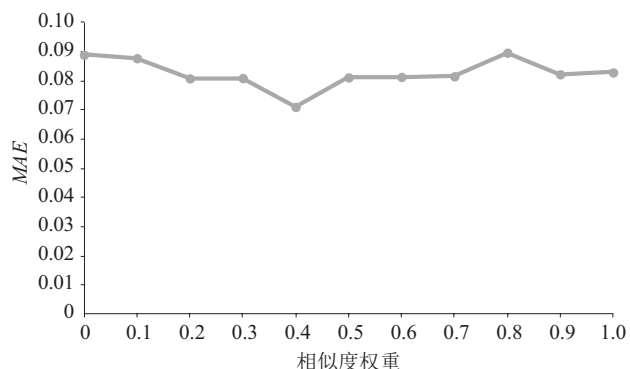


图2 不同权重下MAE的值

由图2可以看出,在 $\beta=0.4$,根据公式(5),即信任相似度为0.6时,MAE值最小,此时推荐最准确,因此本文将用户相似度权重 β 设为0.4。

(4) 设置邻居个数。选择不同的用户规模进行推荐,从用户信息表中随机抽取用户,最近邻居数分别取5、10、15、20、25、30,间隔为5。实验结果如图3所示。

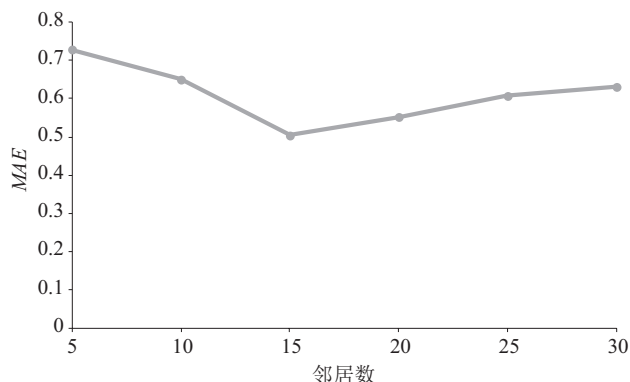


图3 不同邻居数下的MAE值

由图3可以看出,在邻居数15之前,MAE值逐渐减小,当邻居用户为15时,MAE值最小,此时对用户的推荐精度最高。

(5) 本文提出的改进算法与传统协同过滤算法的对比。为验证本文算法的准确性,基于相同的邻居数下,对比传统的协同过滤算法与改进后的协同过滤算法下MAE的大小,实验结果如图4所示。

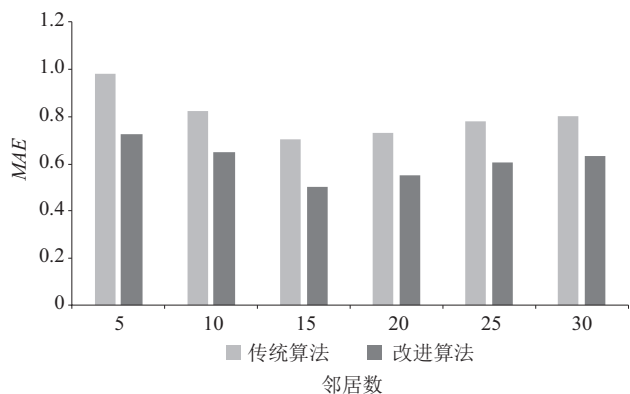


图4 两种不同算法的比较

从图4可以清楚地看出,本文提出的改进算法推荐精度明显高于传统协同过滤算法。

4 结语

本文针对传统协同过滤算法存在的问题,结合高校图书馆的实际特点,提出了一种融合信任相似度的

高校图书馆个性化推荐方法。以用户的借阅信息构建用户-图书矩阵,有效地解决了评分数据严重稀疏的问题,将信任相似度融入传统的用户相似度,极大地提高了推荐精度,同时,提出的新生信任度及新书推荐度改善了冷启动现象。实验结果表明,融合信任相似度的高校图书馆个性化推荐方法具有很好的推荐效果,但本文仍存在一定欠缺,后续将针对冷门图书进行推荐,提高资源利用率,并进一步完善算法,以求达到更精确的推荐效果,为高校用户提供更好的服务。

参考文献

- [1] NARDUCCI F, BASILE P, MUSTOC, et al. Concept-based item representations for a cross-lingual content-based recommendation process [J]. *Information Sciences An International Journal*, 2016, 374 (C): 15-31.
- [2] 刘青文. 基于协同过滤的推荐算法研究 [D]. 合肥: 中国科学技术大学, 2013.
- [3] 孔维梁. 协同过滤推荐系统关键问题研究 [D]. 武汉: 华中师范大学, 2013.
- [4] YU F, ZENG A, GILLARD S, et al. Network-based recommendation algorithms: a review [J]. *Physica A: Statistical Mechanics and Its Applications*, 2016, 452: 192-208.
- [5] 钟克吟. 基于混合推荐的学术资源推荐系统的服务模式与数据挖掘 [J]. *图书馆学研究*, 2013 (11): 58-61, 66.
- [6] 刘健. 数字图书馆资源聚合与服务推荐研究 [D]. 长春: 吉林大学, 2017.
- [7] GONG S J. A collaborative filtering recommendation algorithm based on user clustering and item clustering [J]. *Journal of Software*, 2010, 5 (7): 745-752.
- [8] XU R Z, WANG S Q, ZHENG X W, et al. Distributed collaborative filtering with singular ratings for large scale recommendation [J]. *Journal of Systems and Software*, 2014, 95 (9): 231-241.
- [9] 王成, 朱志刚, 张玉侠, 等. 基于用户的协同过滤算法的推荐效率和个性化改进 [J]. *小型微型计算机系统*, 2016, 37 (3): 428-432.
- [10] 郑孝遥, 鲍煜, 孙忠宝, 等. 一种基于信任的协同过滤推荐模型 [J]. *计算机工程与应用*, 2016, 52 (5): 50-54, 60.
- [11] GUO G, ZHANG J, YORKE-SMITH N. A novel recommendation model regularized with user trust and item ratings [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2016, 28 (7): 1607-1620.
- [12] 郭艳红. 推荐系统的协同过滤算法与应用研究 [D]. 大连: 大连理工大学, 2008.
- [13] AHN H J. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem [J]. *Information Sciences*, 2008, 178 (1): 37-51.
- [14] 冷亚军. 协同过滤技术及其在推荐系统中的应用研究 [D]. 合肥: 合肥工业大学, 2013.
- [15] LU L, MEDO M, CHI H Y, et al. Recommender systems [J]. *Physics Reports*, 2012, 519 (1): 1-49.
- [16] PIRASTEH P, JUNG J J, HWANG D. Item-based collaborative filtering with attribute correlation: A case study on movie recommendation [C] // CHIDS 2014: Proceedings of the 6th Asian Conference on Intelligent Information and Database Systems. Berlin: Springer, 2014: 245-252.

作者简介

林晓霞, 女, 1971年生, 硕士, 副教授, 研究方向: 大数据分析、商业智能。

刘敏, 女, 1992年生, 硕士研究生, 通信作者, 研究方向: 大数据分析、数据挖掘, E-mail: sdustlau@163.com。

杨晓东, 男, 1970年生, 硕士, 讲师, 研究方向: 数据挖掘、人工智能。

徐尧, 男, 1991年生, 硕士研究生, 研究方向: 数据存储。

Research on Personalized Recommendation of University Libraries Based on Trust Similarity

LIN XiaoXia^{1,2} LIU Min¹ YANG XiaoDong^{1,2} XU Yao¹

(1.School of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China;

2.Department of Information Engineering, Shandong University of Science and Technology, Tai'an 271019, China)

Abstract: With the development of the digital library, the personalized recommendation problem of university libraries has become a hot spot, and the traditional rating-based recommendation algorithm does not apply to the recommendation requirements of university libraries. In view of the above situation, based on the study of traditional collaborative filtering algorithm, firstly, use the borrowing recording data to build the scoring matrix, then introduces trust similarity to user similarity, at the same time, integrate the solution of recommended for new students and books. Experiment on the Library Dataset of Shandong University of Science and Technology, the results show, the improved algorithm effectively solves the problem of matrix sparse and cold start, and has a higher recommendation accuracy.

Keywords: Academic Library; Personalized Recommendation; Collaborative Filtering; Trust Similarity

(收稿日期: 2018-08-02)