表 2　四个 OSD 节点的集群中 20 G 镜像文件的迁移时间对比

Tab.2　Comparison of migration time for 20 G image file cross clusters with four OSD nodes

次数	Ceph 块设备跨集群迁移算法迁移时间/s	传统迁移算法迁移时间/s
1	634	1567
2	636	1581
3	650	1537
4	646	1602
5	628	1514
平均时间	638.8	1560.2

从以上两组对比实验的结果可以看出，本文提出的 Ceph 块设备跨集群迁移算法在一定程度上加快了迁移速度，使迁移时间缩短至传统迁移算法的 40%左右.

本组对比测试中，两个集群 OSD 节点的数目均为 3 个，待迁移的镜像文件的大小为 10 G，两种算法分别迁移了 5 次，迁移时间结果如表 3 所示，同时测得使用 Ceph 块设备跨集群迁移算法时目的 OSD 节点的平均带宽利用率为 36.18%.

表 3　三个 OSD 节点的集群中 10 G 镜像文件的迁移时间对比

Tab.3　Comparison of migration time for 10 G image file cross clusters with three OSD nodes

次数	Ceph 块设备跨集群迁移算法迁移时间/s	传统迁移算法迁移时间/s
1	410	801
2	381	762
3	417	787
4	386	774
5	399	780
平均时间	398.6	780.8

本组对比测试中，待迁移的镜像文件的大小为 20 G，其余条件和第 3 组实验相同，两种算法分别迁移了 5 次，迁移时间结果如表 4 所示.

从以上两组对比实验的结果可以看出，减少一个 OSD 节点后，本文提出的 Ceph 块设备跨集群迁移算法虽然依旧加快了迁移速度，但是提升的效率相比 4 个 OSD 节点的集群更少，迁移时间缩短至传统迁移算法的 50%左右.

表 4　三个 OSD 节点的集群中 20 G 镜像文件的迁移时间对比

Tab.4　Comparison of migration time for 20 G image file cross clusters with three OSD nodes

次数	Ceph 块设备跨集群迁移算法迁移时间/s	传统迁移算法迁移时间/s
1	763	1632
2	760	1496
3	747	1526
4	775	1585
5	752	1513
平均时间	759.4	1550.4

综合上述 4 组对比实验的结果，本文提出的 Ceph 块设备跨集群迁移算法比起传统的迁移算法在迁移效率上有明显的优化提升，能节省相当多的迁移时间.在集群 OSD 节点数目不同的情况下，传统迁移算法的迁移时间没有很大的差别，说明镜像文件的存储方式对传统迁移算法几乎没有影响，这也印证了传统迁移算法中存储方式对用户透明的的特点.对于本文提出的 Ceph 块设备跨集群迁移算法，4 个 OSD 节点的集群对比 3 个 OSD 节点的集群能够提升更多的迁移效率，原因是在网络带宽足够的情况下，更多的 OSD 节点数目提供了更大的数据传输并行度，符合理论分析.由于受实验环境条件所限，无法增加更多的 OSD 节点，以测试传输效率的上限.推测当 OSD 节点的数目增加到一定程度之后，数据的读写速度或者节点的带宽成为瓶颈，使得传输效率无法继续提升，该想法有待今后的工作予以验证.

4　结论

我们考虑到分布式存储系统中存储节点的网络能力，针对 Ceph 块设备的跨集群迁移，对传统的具有普适性的迁移方式进行了修改，提出了存储节点并行向存储节点进行传输的算法，实验也表明了可以提升迁移效率.虽然该算法的使用有一些限制条件，只限定了 Ceph 块设备这一种分布式存储系统，并且存储节点需要有对集群外节点进行网络通信的能力，但是该算法大大加快了镜像文件的迁移速度，具有一定的应用前景.同时我们在实验中还发现，适当增加存储节点的数目，能提高该算法的并行度，进一步提升算法效率.

参考文献(References)

[1] CLARK C, FRASER K, HAND S, et al. Live migration of virtual machines[C]// Proceedings of the 2nd Conference on Symposium on Networked Systems Design & Implementation. Berkeley, USA: USENIX Association, 2005: 273-286.

[2] NELSON M, LIM B H, HUTCHINS G. Fast transparent migration for virtual machines [C]// Proceedings of the USENIX Annual Technical Conference. Berkeley, USA: USENIX Association, 2005: 391-394.

[3] HANSEN J G, JUL E. Self-migration of operating systems[C]// Proceedings of the 11th Conference on ACM SIGOPS European Workshop. Leuven, Belgium: ACM Press, 2004: No.23(1-5) .

[4] ZHANG B, LUO Y, WANG X, et al. Whole-system live migration mechanism for virtual machines [J]. Acta Electronica Sinica, 2009,37(4): 894-899.

[5] HIROFUCHI T, NAKADA H, OGAWA H, et al. A live storage migration mechanism over wan and its performance evaluation[C]// Proceedings of the 3rd International Workshop on Virtualization Technologies in Distributed Computing. New York: ACM Press, 2009: 67-74.

[6] BRADFORD R, KOTSOVINOS E, FELDMANN A, et al. Live wide-area migration of virtual machines including local persistent state[C]// Proceedings of the 3rd International Conference on Virtual Execution Environments. San Diego, USA: ACM Press, 2007: 169-179.

[7] WEIL S A, BRANDT S A, MILLER E L, et al. Ceph: A scalable, high-performance distributed file system[C]// Proceedings of the 7th Symposium on Operating Systems Design and Implementation. Seattle, USA: USENIX Association, 2006: 307-320.

[8] WEIL S A. Ceph: Reliable, scalable, and high-performance distributed storage [D]. Santa Cruz: University of California, 2007.

[9] LIANG X Y, GUAN Z G. Ceph CRUSH data distribution algorithms [J]. Applied Mechanics & Materials, 2014 596: 196-199.

[10] 王远洋,周渊平,郭焕丽. Linux 下基于 socket 多线程并发通信的实现[J]. 微计算机信息, 2009, 25(3-5): 70-72.
WANG Yuanyang, ZHOU Yuanping, GUO Huanli. The accomplishment of multi-*p*thread communication based on socket model in the Linux [J]. Microcomputer Information, 2009, 25(3-5): 70-72.

(上接第 717 页)

[11] ROBERTS R S, BROWN W A, LOOMIS H H. Computationally efficient algorithms for cyclic spectral analysis[J]. IEEE Signal Processing Magazine, 1991, 8(2): 38-49.

[12] ANTONI J, XIN G, HAMZAOUI N. Fast computation of the spectral correlation[J]. Mechanical Systems and Signal Processing, 2017, 92: 248-277.

[13] NAPOLITANO A, PERNA I. Cyclic spectral analysis of the GPS signal[J]. Digital Signal Processing, 2014, 33: 13-33.

[14] GARDNER W A. Measurement of spectral correlation [J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1986, 34(5): 1111-1123.

[15] RAMKUMAR B. Automatic modulation classification for cognitive radios using cyclic feature detection[J]. IEEE Circuits and Systems Magazine, 2009, 9(2): 27-45.

[16] 赵宇峰, 曹玉健, 纪勇, 等. 基于循环频率特征的单信道混合通信信号的调制识别[J]. 电子与信息学报, 2014, 36(5): 1202-1208.
ZHAO Yufeng, CAO Yujian, JI Yong, et al. Modulation identification for single-channel mixed communication signals based on cyclic frequency features [J]. Journal of Electronics & Information Technology, 2014, 36(5): 1202-1208.

[17] HU Y, SONG M, and MENG B. GPS signal availability augmentation utilizing the navigation signal retransmission via the GEO Comsat [J]. Wireless Personal Communications, 2015, 82(4): 2655-2671.

[18] KAPLAN E D, HEGARTY C J. Understanding GPS: Principles and Applications[M]. 2ed, Boston, USA: Artech House Publishers, 2006: 153-200.

文章编号:0253-2778(2018)09-0755-07

一种基于Skill-LFM的知识点推荐方法

方建生[1],许言午[1],蔡瑞初[2],秦 艳[3]

(1.广州视源电子科技有限公司中央研究院,广东广州 510000; 2.广东工业大学数据挖掘与信息检索实验室,广东广州 510000;
3.中国电信股份有限公司广东分公司,广东广州 510000)

摘要:目前,知识库的用户主要是通过检索获取所需知识点,这种依赖搜索引擎解决信息过载的方法,对实时在线服务而言效率低下,对离线知识学习来说不具有完整性和连续性,为此提出由知识库系统根据用户技能水平主动推荐知识点给用户,提高决策效率,并有助于用户建立完备的知识学习体系.基于用户对知识点的历史行为以及用户对知识的学习能力,提出一种融合技能的隐语义模型的协同过滤推荐方法,将知识点难易程度作为潜在因子,同时考虑用户的能力水平预测用户对知识点的偏好水平.在呼叫中心知识库的数据集上进行测试,其均方根误差优于基础隐语义模型.综合知识点推荐的应用领域和知识学习行为数据的特点,对于知识点推荐方法,可从融合用户和知识点上下文信息的推荐技术上深入研究.

关键词:协同过滤;隐语义模型;知识库;决策支持;推荐系统;上下文感知

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2018.09.010

引用格式:方建生,许言午,蔡瑞初,等. 一种基于Skill-LFM的知识点推荐方法[J]. 中国科学技术大学学报,2018,48(9):755-761.

FANG Jiansheng, XU Yanwu, CAI Ruichu, et al. A method of knowledge item recommendation based on Skill-LFM[J]. Journal of University of Science and Technology of China, 2018,48(9):755-761.

A method of knowledge item recommendation based on Skill-LFM

FANG Jiansheng[1], XU Yanwu[1], CAI Ruichu[2], QIN Yan[3]

(1. *Research, CVTE, Guangzhou* 510000; 2. *DMIR, GDUT, Guangzhou* 510000;
3. *Guangdong branch, China Telecom, Guangzhou* 510000)

Abstract: At present, the users of knowledge base mainly get the required knowledge items through search, which relies on the search engine to solve the information overload problem. It is inefficient for real-time online services, and has no integrity and continuity of offline knowledge learning. Therefore, it is proposed that knowledge items should be actively recommended to users by the knowledge base system according to their level of skills, to improve the efficiency of decision making, and also to help users establish a complete knowledge learning system. A collaborative filtering recommendation method is proposed to predict every user's preference on knowledge items, based on the historical behavior of a user on the knowledge items, and the knowledge learning ability of this user. This method combines latent factor model with skill, named Skill-LFM, where the difficulties of knowledge items are taken as potential factors, and users' ability level is considered to give personalized recommendations. Tested on the data from a call center knowledge base, the proposed Skill-LFM outperforms the baseline latent factor model in

收稿日期:2018-05-28; **修回日期**:2018-09-18

作者简介:方建生(通讯作者),男,1982年生,硕士. 研究方向:人工智能及数据挖掘. E-mail: fangjiansheng@cvte.com

terms of lower RMSE. Considering the characteristics of the application domain and the historical behavior data of the knowledge base, this paper demonstrates the possibility of further improving knowledge item recommendation through integrating user and knowledge item context information.

Key words: collaborative filtering; latent factor model; knowledge base; decision support; recommender system; context-aware

0 引言

知识库[1](knowledge base,KB)是知识工程中有组织的知识片集合,是针对某一领域问题求解的事实、规则、概念的组合.用户掌握知识库中的知识点用于决策支持,如呼叫中心的客服代表(customer service representative,CSR),一方面离线学习产品介绍或服务标准等知识点,另一方面在线通话时依托知识库解答客户问题.

为支持用户使用知识库,知识库系统[2]提供模式匹配、检索推理等功能反馈用户所需的知识点.用户从知识库中获取知识通过检索方式存在两个问题:

(Ⅰ)检索操作耗时,在线服务时要求及时反馈客户的问题,用户从关键词搜索到选择合适的知识点阅读和理解,对即时通讯(instant messaging,IM)在线服务来说,这个操作时长既影响客户感知,又付出高昂的人力时间成本;

(Ⅱ)知识学习不够系统,用户离线学习按需而为,对某一类知识点的学习持续性和完整性不足,尤其在庞大的知识点下,用户不可能记住过去一段时间学习过某个知识点,而当下是否需重新理解或阅读同类其他知识点.

基于知识库系统检索功能的不足,知识库系统开启了集成推荐功能[3-4]的探索,相关推荐技术的研究也逐渐受到关注.本文立足于在知识库系统中集成推荐功能,为用户提供个性化知识点支持其在线服务和离线学习,提出基于矩阵分解技术[5-6](matrix factorization,MF)并融合技能的隐语义模型(skill-based latent factor models,Skill-LFM),该推荐模型适用于知识点推荐.本文的主要贡献如下:

(Ⅰ)基于用户学习知识点的二元关系,考虑到不同用户对难易不同的知识点有不同的理解、记忆、学习能力,将用户的学习技能水平集成到隐语义模型(latent factor models,LFM)中,预测用户对知识点的偏好水平;

(Ⅱ)将知识点的难易程度作为隐因子,并融合技能的隐语义模型,直观上理解为,不同技能水平的用户面对不同的知识点有不同的学习需求,同时本文提出还可进一步加入用户和知识点上下文来提升推荐准确度;

(Ⅲ)在呼叫中心知识库系统的数据集测试中融合技能的隐语义模型在观测值和预测值误差上的表现优于隐语义模型.

1 相关工作

随着基于物品的协同过滤[7-8](item-based collaborative filtering,ItemCF)方法在商业应用上的重大成功,如 Amazon、Spotify、Facebook 等,推荐方法的研究已成个性化服务领域的重要分支.推荐具有发现用户感兴趣内容和长尾内容的能力.

数据的膨胀必然给信息选择带来困扰,解决信息过载的问题,搜索引擎是有效的工具,但个性化不足[9].随着机器学习和数据挖掘方法的应用深入,推荐系统[10-11](recommender system,RS)成为个性化信息推荐的重要研究方向.

推荐系统从大量数据中帮助用户发现可能感兴趣的项目,目前在电子商务、移动应用、互联网广告等众多领域有成熟应用,尤以 Netflix 电影推荐比赛[12]输出了一批技术成果.推荐策略[13]一般可分为:

(Ⅰ)基于内容的推荐(content-based recommendation),基于用户个人的历史行为来推荐,通过对用户和物品的特征构建有监督机器学习模型进行分类以识别与用户兴趣匹配的物品.

(Ⅱ)协同过滤推荐[14](collaborative filtering recommendation),基于相似用户的历史行为协同推荐,体现邻域和集体思想,可分为基于记忆和基于模型两种类型,前者是基于用户或物品相似度,后者基于用户对物品的评分构建模型用于预测评分.

(Ⅲ)混合推荐[15](hybrid recommendation),将不同推荐算法组合进行推荐,混合策略有合并、并行、串联等.

基于内容的推荐具有强可解释性和稳定性,但

推荐多样性不足，而且采集用户和物品的信息在一些场景存在困难；协同过滤推荐则可进行新颖性推荐，且仅依赖用户历史行为，如交易或评分，但存在冷启动、稀疏性以及推荐结果不可解释性等问题，文献[13]更是指出协同过滤技术更易于被攻击，即伪造历史行为数据误导推荐结果，如恶意评分.

基于矩阵分解技术的隐语义模型是协同过滤推荐方法中基于模型一类，相比于基于近邻技术的基于记忆一类在 Netflix 比赛中表现更为优越，而且可以融合除评分显式反馈信息外的隐式反馈信息[16]，如时间[17]、信任度[18]等.

本文基于矩阵分解的隐语义模型，提出融合用户技能的推荐方法，该方法可集成到知识库系统中，为用户推荐个性化知识点，用于在线客服问答或离线强化学习.

2　基于技能的隐语义模型

奇异值分解（singular value decomposition，SVD）在信息检索中识别潜在语义的功能，使其在推荐系统的方法研究[19-20]在近年呈现上升趋势，文献[16]比较体系地分析了 SVD 及各种改良 SVD 的性能.

传统奇异值分解在推荐系统中的应用首先面临数据稀疏的问题，因此要对用户-物品的评分矩阵进行补全，比如用全局平均值或用户、物品平均值进行补全.补全后的矩阵存储需要很大空间，且分解成低维矩阵时的计算复杂度也很高.

鉴于 basic SVD 的数据稀疏以及计算成本问题，Funk 提出 Funk-SVD 方法，将评分矩阵分解为用户因子矩阵和物品因子矩阵，利用线性回归的思想，通过不断的迭代训练使得用户因子矩阵和物品因子矩阵乘积得到的评分残差尽可能的小.基于 Funk-SVD 方法，Netflix 电影推荐比赛的冠军 Koren 构建了隐语义模型（LFM），后续有众多的研究改进，如加入正则项的 RSVD、加入偏置项的 bias-SVD、增加用户隐式反馈的 SVD＋＋、非对称性 asymmetric-SVD 等.

基于 Funk-SVD 方法的隐语义模型（LFM），本文在面向知识点推荐应用上提出了融合技能的隐语义模型（Skill-LFM）.

对于知识点的学习能力，从阅读、理解到掌握是一个过程，不同技能的用户对不同难度的知识点会有不同的学习过程，通过历史学习次数可隐式观测.

Skill-LFM 在构建模型上以显式的学习次数为用户对知识点的偏好度，融合隐式观测到的技能水平，隐因子映射为知识点难易程度，预测用户对知识点的偏好度，作为用户在线服务和离线学习时推荐知识点的依据.

下面给出融合技能隐语义模型的描述：

（Ⅰ）给定 m 个 CSR、n 个知识点，以及 CSR 对知识点的评分矩阵 R，其中第 u 个 CSR 对第 i 个知识点的评分为 $r_{u,i}$.

$r_{u,i}$ 是第 u 个 CSR 对第 i 个知识点学习次数 $n_{u,i}$ 经过区间缩放后的值，区间缩放公式为

$$r_{u,i}=\frac{n_{u,i}-n_{\min}+1}{n_{\max}-n_{\min}+1},$$

其中，n_{min} 和 n_{max} 分别是最小评分次数和最大评分次数，$r_{u,i}$ 区间值在 (0,1] 范围内.

（Ⅱ）隐语义模型在 CSR 和知识点之间构造潜在因子 k，可以将 CSR 对知识点的评分矩阵 R 分解成两个稠密的因子矩阵 $P\in R^{k\times m}$ 和 $Q\in R^{k\times n}$，其中 $p_u\in R^k$ 表示矩阵 P 的第 u 列，$q_i\in R^k$ 表示矩阵 Q 的第 i 列，则第 u 个 CSR 对第 i 个知识点的预测评分表示为 $\hat{r}_{u,i}=q_i^{\mathrm{T}}p_u$.

（Ⅲ）隐语义模型利用线性回归思想，通过优化预测评分和真实评分之间的评分误差最小来拟合因子矩阵 P 和 Q.评分预测值和真实值的误差定义为

$$e_{u,i}=r_{u,i}-\hat{r}_{u,i}.$$

（Ⅳ）隐语义模型优化问题定义为

$$\min_{P,Q}\sum_{(u,i)\in R}(r_{u,i}-q_i^{\mathrm{T}}p_u)^2+\lambda(\|p_u\|_F^2+\|q_i\|_F^2),$$

其中 $\|\cdot\|_F$ 是 F-范数，$(u,i)\in R$ 是可用的评分，λ 是为防止过拟合而给出的正则化系数[21].

所构建的优化问题是关于矩阵 P 和 Q 的非凸函数，随机梯度下降[22]（stochastic gradient descent，SGD）可用于求解其最优解.

（Ⅴ）第 u 个 CSR 当前的技能水平 n_u 表示为

$$n_u=\frac{|N(u)|+1}{\mathrm{avg}_u+1},$$

其中，$|N(u)|$ 表示第 u 个 CSR 学习过的知识点集合，avg_u 表示第 u 个 CSR 平均学习次数.

s_u 是 n_u 经过区间缩放的值，区间缩放公式为

$$s_u=\frac{n_u-n_{u,\min}+1}{n_{u,\max}-_{u,\min}+1},$$

其中，$n_{u,\min}$ 和 $n_{u,\max}$ 分别是最低技能水平和最高评分次数，s_u 区间值在 (0,1] 范围内.

如果是新 CSR，没学习过任何知识点，技能为

0.5;随着知识点的学习行为累积,技能得到反映,值越大说明技能越高.

CSR 的当前总体技能水平 s_u 是通过学习次数的隐式反馈的,如果 CSR 学习的知识点数量很多且平均学习次数较低,则 s_u 越高;反之,如果 CSR 学习的知识点数量较少,而平均学习次数较高,则 s_u 越低.

(Ⅵ)基于知识点学习次数所隐式反馈出的 CSR 技能水平,融合技能的隐语义模型将第 u 个 CSR 学习的第 i 个知识点的技能水平作为一个属性融合到模型中.

Skill-LFM 模型中第 u 个 CSR 对第 i 个知识点的预测评分表示为

$$\hat{r}_{u,i}=q_i^{\mathrm{T}}(p_u+s_{u,k}+s_{i,k}),$$

其中,$s_{u,k}\in R^k$ 表示第 u 个 CSR 当前技能水平 s_u 的 k 维向量化;$s_{i,k}\in R^k$ 表示第 u 个 CSR 技能水平在第 i 个知识点上的偏置的 k 维向量化.

$s_{i,k}$ 用来描述 CSR 技能水平面对不同难易程度的知识点时有不同的差异,随机初始值,在模型训练过程中不断迭代优化.

(Ⅶ)Skill-LFM 模型的优化问题是使预测偏好和真实偏好的误差最小,目标函数定义为

$$\min_{p^*,q^*,s^*}\sum_{(u,i)\in R}(r_{u,i}-\hat{r}_{u,i})^2+\lambda_1(\|p_u\|_F^2+\|q_i\|_F^2)+\lambda_2 s_{i,k}^2,$$

其中,λ_1 和 λ_2 是为防止过拟合而给出的正则化系数.

(Ⅷ)通过随机梯度下降(SGD)求解 Skill-LFM 的目标函数:

①随机选择一条 (u,i) 记录获得预测评分:

$$\hat{r}_{u,i}=q_i^{\mathrm{T}}(p_u+s_{u,k}+s_{i,k});$$

②计算评分误差:

$$e_{u,i}=r_{u,i}-q_i^{\mathrm{T}}(p_u+s_{u,k}+s_{i,k});$$

③求解 $(r_{u,i}-q_i^{\mathrm{T}}(p_u+s_{u,k}+s_{i,k}))^2+\lambda_1 p_u^{\mathrm{T}}p_u+\lambda_1 q_i^{\mathrm{T}}q_i+\lambda_2 s_{i,k}^2$ 对 p_u、q_i、$s_{i,k}$ 的偏导并更新其负梯度方向:

$$p_u\leftarrow p_u+\gamma(e_{u,i}q_i-\lambda_1 p_u)$$

$$q_i\leftarrow q_i+\gamma(e_{u,i}(p_u+s_{u,k}+s_{i,k})-\lambda_1 q_i)$$

$$s_{i,k}\leftarrow s_{i,k}+\gamma(e_{u,i}q_i-\lambda_2 s_{i,k}).$$

其中,γ 是学习速率.

④循环选择 (u,i) 记录,迭代更新前 3 个步骤,直至迭代次数到达或满足既定条件退出.

矩阵分解技术在推荐系统中的应用,除了比较流行的 SVD 外,还有概率矩阵分解[23](probabilistic matrix factorization, PMF)、非负矩阵分解[24-26](non-negative matrix factorization, NMF)、高阶奇异值分解[27-29](higher-order singular value decomposition, HOSVD)等的探索.

不同的技术有不同的适用场景,Skill-LFM 模型主要是基于 SVD++[16]思想,集成隐式反馈出的 CSR 技能水平,将 CSR 的技能属性融合到模型中训练,挖掘出 CSR 在不同知识点上学习差异,从而在学习次数偏好上刻画的更精准.

3 实验与结果

3.1 评价指标

本文主要通过 Skill-LFM 预测用户对未学习知识点的学习偏好,作为在线服务和离线学习知识点推荐的依据,不直接推荐 Top-n 结果,所以不采用 Top-n 推荐[30]的评价指标.

本文通过回归评价指标均方根误差[31-32](root mean square error, RMSE)来评价模型性能.

RMSE 是预测值与真实值的误差平方根的均值为

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{t=1}^{N}(r_t-\hat{r}_t)^2},$$

其中,N 是测试集记录数,r_t 是实际观测到的值,$\hat{r}_t$ 是预测的值.RMSE 的值越小,说明预测模型描述实验数据具有更好的精确度.

3.2 数据集

本文的实验数据来自某呼叫中心知识库系统中 CSR 学习知识点(knowledge item,KI)的行为记录.提取 2017 年 12 月 1 日-2017 年 12 月 31 日共 211 万条学习记录、1.2 万名 CSR、1.4 万个知识点,包括 CSR、KI、Time 三个字段,记录了 CSR 什么时间学习了那个知识点的行为.

基于这批记录,建立一个 CSR 对 KI 的评分矩阵 R,剔除部分异常数值(大于 60 的学习次数),共有 68 万个评分值,其他为零.矩阵值是对 CSR 学习 KI 的次数进行区间缩放,最高学习次数 59、最小学习次数 1,平均学习次数 2.88 次.

3.3 实验结果

LFM 和 Skill-LFM 模型的实现是在矩阵分解[33-35]基础代码上改进,采用 python 语言编写.从样本中抽取 10% 比例作为测试集,约 7 万条评分记录.

实验一　Skill-LFM 和 ItemCF、NMF

ItemCF 是传统协同过滤推荐方法，先建立知识点之间的相似度矩阵，然后从相似度最高的 k 个知识点加以推荐，如命中则依据相似度可以得出预测值和真实值的误差.对相似度最高的 k 个知识点进行试验，在 $k=10$ 时取得 RMSE＝0.175681 的最好成绩.

NMF 是在矩阵分解的基础上加上非负的约束条件，对基于平方距离的损失函数采用乘法规则求解.与 LFM 一样，其关键是选择合适的隐因子 k，实验结果显示 $k=3$ 时，RMSE＝0.160680 为最好成绩.

Skill-LFM 在 $k=3$ 时取得 RMSE＝0.115101 的成绩，表现比传统 ItemCF、NMF 方法更为优异.

实验二　Skill-LFM 和基础 LFM 及变异模型

实验选择隐语义模型(LFM)中的 Funk-SVD、Bias-SVD 两种方法和基于 SVD＋＋思想并融合用户技能水平的隐语义模型(Skill-LFM)在 RMSE 上进行比较，如表 1.

表 1　LFM 和 Skill-LFM 模型 RMSE 性能

Tab.1　RMSE on LFM and Skill-LFM

参数\\方法		LFM		Skill-LFM
		Funk-SVD	Bias-SVD	SVD＋＋
$k=3$	steps＝100	0.153 822	0.280 491	0.174 702
	steps＝200	0.126 224	0.187 339	**0.115 101**
$k=5$	steps＝100	0.187 236	0.252 827	0.185 359
	steps＝200	0.142 596	0.168 075	0.117 486
$k=8$	steps＝100	0.219 507	0.236 100	0.188 148
	steps＝200	0.160 834	0.152 611	0.120 649
$k=10$	steps＝100	0.237 658	0.232 116	0.187 363
	steps＝200	0.169 922	0.148 874	0.121 932

实验中设置正则化系数 $\lambda=\lambda_1=\lambda_2=0.002$、学习速率 $\gamma=0.0002$，其中 k 是隐因子数、steps 是迭代次数.通过不同隐因子数以及迭代次数的实验观察，表明融合技能属性的 Skill-LFM 相比于 LFM 在 RMSE 上表现更好.

实验中的 3 种方法，预测评分的表示分别如下：

①LFM 的 Funk-SVD 方法预测评分表示为 $\hat{r}_{u,i}=q_i^{\mathrm{T}}p_u$；

②LFM 的 Bias-SVD 方法预测评分表示为 $\hat{r}_{u,i}=q_i^{\mathrm{T}}p_u+b_{u,i}$，其中，$b_{u,i}$ 是 CSR 对知识点的偏好偏，其负梯度方向更新表示为 $b_{u,i}\leftarrow b_{u,i}+\gamma(e_{u,i}-\lambda b_{u,i})$；

③Skill-LFM 的 SVD＋＋方法预测评分表示为 $\hat{r}_{u,i}=q_i^{\mathrm{T}}(p_u+s_{u,k}+s_{i,k})$.

将 3 种种方法在 steps＝500 下的 RMSE 进行比较，如图 1 所示，可观察到 3 点现象.

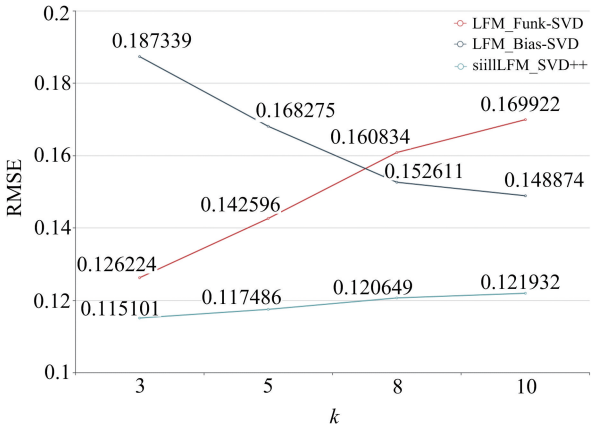

图 1　LFM 和 Skill-LFM 在 RMSE 上的比较

Fig.1　Comparison of LFM and Skill-LFM on RMSE

①LFM 模型中，bias-SVD 方法将当前 CSR 和知识点偏置 $b_{u,i}$ 作为模型的偏置项加入到模型中.从图 1 中可观察到，随着 k 值的增加，bias-SVD 方法的 RMSE 值越来越低，而 Funk-SVD 方法则相反，在 $k=8$ 时 bias-SVD 超越 Funk-SVD.

在 bias-SVD 方法中，$b_{u,i}$ 是在训练时最小化残差下所获取的.以 RMSE 作为因变量来看，相比于 Funk-SVD 方法来说，Bias-SVD 方法除了隐因子是自变量，还有偏置项 $b_{u,i}$ 是自变量，这说明 $b_{u,i}$ 在模型中发挥了重要作用，使 k 越大 RMSE 越低.

②Skill-LFM 模型中，基于 SVD＋＋思想并在用户因子矩阵 P 中加入 CSR 技能水平及其知识点偏置，随着 k 值增加，RMSE 值渐渐上升.在隐因子 $k=3$ 时 RMSE 最低，直观上，对应为知识点的难、中、易三个层次，体现了 CSR 技能水平面对不同难度知识点的学习偏好程度不一样.

③一个比较有趣的问题是，随着 k 的增加，从理论上说，当 k 越接近于知识点的数量 n 时，LFM 模型 bias-SVD 方法在 RMSE 值上会接近，甚至低于 Skill-LFM 模型 SVD＋＋方法.这是因为在 LFM 模型 bias-SVD 方法中，$b_{u,i}$ 是主要作用，k 越大、大到和知识点数量相同时，$b_{u,i}$ 就越接近为每个具体知识点的偏置，而不是一类知识点的偏置.Skill-LFM 模型中，$s_{i,k}$ 体现为一类知识点的偏置作用更大，k