软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(11):3382−3396 [doi: 10.13328/j.cnki.jos.005542]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种潜在特征同步学习和偏好引导的推荐方法*

李 琳[1], 朱 阁[1], 解 庆[1], 苏 畅[1], 杨征路[2]

[1](武汉理工大学 计算机科学与技术学院,湖北 武汉 430070)

[2](南开大学 计算机与控制工程学院,天津 300071)

通讯作者: 李琳, E-mail: cathylilin@whut.edu.cn

摘 要: 根据用户的历史评分数据为用户提供推荐的商品列表,是目前推荐系统研究的主流.研究者发现,随着用户参与度的不断提高,将反映用户偏好的评论文本与评分数据结合,可以进一步提高推荐的质量.提出了基于潜在特征同步学习和偏好引导的商品推荐方法,将评论文本的主题与用户的"打分偏好"进行关联,同步学习用户评论文本的潜在主题、评分矩阵的用户潜在因子和商品潜在因子,并将潜在主题作为用户个人偏好引导来约束推荐方法对商品的预测打分.该方法对推荐质量的优化主要体现在两个方面:一是在评论文本的潜在主题和评分数据的两种潜在因子之间建立映射关系,同步求解主题模型和矩阵分解模型;二是将从评论文本中学习得到的潜在主题作为用户对商品的个性偏好引入到矩阵分解中,进一步优化推荐方法.在来自 Amazon 网站的 28 组真实数据集上进行实验,以均方误差为评价指标,与已有的模型进行了对比分析.实验结果表明,该方法有效减少了推荐误差,与已有的 TopicMF 方法相比,均方误差在数据子集上最大减少了 3.32%,平均减少了 0.92%.

关键词: 评论文本;评分数据;推荐系统;潜在主题;潜在因子

中图法分类号: TP311

中文引用格式: 李琳,朱阁,解庆,苏畅,杨征路.一种潜在特征同步学习和偏好引导的推荐方法.软件学报,2019,30(11):3382−3396. http://www.jos.org.cn/1000-9825/5542.htm

英文引用格式: Li L, Zhu G, Xie Q, Su C, Yang ZL. Recommendation approach by simultaneous learning latent features and preferences guidance. Ruan Jian Xue Bao/Journal of Software, 2019,30(11):3382−3396 (in Chinese). http://www.jos.org.cn/1000-9825/5542.htm

Recommendation Approach by Simultaneous Learning Latent Features and Preferences Guidance

LI Lin[1], ZHU Ge[1], XIE Qing[1], SU Chang[1], YANG Zheng-Lu[2]

[1](School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430070, China)

[2](College of Computer and Control Engineering, Nankai University, Tianjing 300071, China)

Abstract: It is a popular way that makes use of users' rating data to recommend products or items to users. Currently, more and more users have contributed their reviews to recommender system for better online shopping experiences. Researchers have become interested in using review texts as extra information to improve recommendation quality. It is argued that reviews written by a user implicitly represent his/her preferences. In this study, a preference guidance recommendation approach is proposed that simultaneously learns latent factors from rating data and latent topics from review texts. More specifically, the learned latent topics are assumed to be positively

* 基金项目: 国家社会科学基金(15BGL048); 国家自然科学基金(61602353, 11431006, U1636116); 湖北省科技支撑计划(2015BAA072); 中央高校基本科研业务费专项资金(WUT:2017II39GX); 武汉理工大学研究生优秀学位论文培育项目(2016-YS-068)

Foundation item: National Social Science Foundation of China (15BGL048); National Natural Science Foundation of China (61602353, 11431006, U1636116); Hubei Province Science and Technology Support Project (2015BAA072); Fundamental Research Funds for the Central Universities (WUT:2017II39GX); Supported by the Excellent Dissertation Cultivation Funds of Wuhan University of Technology (2016-YS-068)

收稿时间: 2017-08-10; 修改时间: 2017-11-01; 采用时间: 2017-12-29

correlated with both of the corresponding user factors and item factors, which can further improve the accuracy of recommendation prediction. The proposed approach has two advantages. One is that in order to capture such a dependent correlation, a transformation function is used for simultaneously learning latent features, i.e., latent factors and latent topics. The other is that the predicted ratings of items are influenced by the implicit tastes of users, i.e., the latent topics from review texts. Experiments are conducted on the data from Amazon consisting of 28 categories. Experimental results show that the proposed approach obtains 3.32% improvement than the recent TopicMF approach in some category dataset and the average improvement is 0.92% in terms of mean square error.

Key words: review text; rating data; recommender system; latent topics; latent factors

1992 年,Goldberg 等人提出了协同过滤的方法,构建和实现了个性化的邮件推荐系统[1].1994 年,Resnick 等人在文献中阐述了 GroupLens 研究组研究的推荐系统[2].迄今为止,推荐系统已经在各大领域得到了广泛的应用和推广,包括 Netflix 的视频推荐系统、Amazon 的商品推荐系统、淘宝网、京东商城等知名的中国电子商务平台以及具有推荐功能的音乐等休闲娱乐平台[3].

通过用户历史评分来推荐商品的协同过滤方法是主流的推荐方法之一,其他还有基于内容的推荐、基于知识的推荐以及混合推荐方法等.为了进一步提高推荐质量,研究者们不断挖掘和分析能够反映用户特性或者用户偏好的信息,例如用户的性别、用户的评论、用户的关系等.相对于历史评分数据,评论文本是用户能够更具体表达自己喜好的一种方式.研究者通过总结用户评论文本中的重要观点[4],或者是挖掘具有相同观点的用户来提高推荐质量[5],还有将评论文本的语义分析和观点挖掘融入到推荐方法中[6].用户对商品的评论文本隐含了一定量的用户信息,因此可以与传统基于用户评分的推荐方法结合,进一步提高推荐质量.

图 1 中,用户对电脑产品的评分和评论数据来自于某电子商务平台.相对于评分数据(图 1 中的“star5”),评论文本表达了用户对商品不同方面的关注和偏好.用户 1 从价格、外观、材质这 3 个方面对某电脑产品进行了评论,总体比较满意,给了 5 分的评价.用户 2 从价格、硬件配置和系统软件这 3 个方面对进行了评论,给了 4 分的评价.评论文本可以反映用户的“打分偏好”(图 1 中价格、外观、材质、硬件配置和系统软件 5 个方面的评价),从一定程度上解释了用户评分的依据和原因.因此,从图 1 中的例子可以观察到,评分数据是用户对商品的整体评价,评论文本中蕴含着影响用户评分的偏好或者说是潜在因子.体现用户打分偏好的潜在因子表明评论数据和评论文本具有关联,因此可以将从评论文本主题发现模型学习得到的潜在主题作为个人打分偏好,融合到基于评分数据的隐因子矩阵分解模型中,从而提高面向用户的推荐质量.

Fig.1 Ratings and reviews of PC products on an e-commerce platform

图 1 某电子商务平台上对电脑商品的评分和评论

目前比较流行的是基于矩阵分解的协同过滤方法[7−11].这类方法在推荐准确率上表现出色.同时,有关融合评论文本的推荐方法也在不同方面上做了进一步的优化和改进,且以某个特定指标为评价方式(比如均方误差、绝对值误差、均方根误差等)最终提升了推荐质量[12−14].但是已有的方法未能充分地利用评论文本主题来

预测评分矩阵中对未知商品的打分,特别是评论文本与基于矩阵分解的协同过滤方法如何在模型求解过程中深度融合,同时考虑用户偏好和商品特性.本文的研究目标是探索融合评分矩阵分解模型与评论文本主题发现模型的推荐方法,考虑评分和评论数据都具有高层潜在特征,将矩阵分解中的潜在因子与评论文本的潜在主题建立映射关系后同步学习参数.通过在用户对商品的单条评论文本上进行主题发现学习,将潜在主题特征融入评分矩阵分解模型的求解中,从而提出了基于潜在特征同步学习和偏好引导的商品推荐方法(preference guided in matrix factorization,简称 PreferenceMF).该方法将评论文本的潜在主题与矩阵分解的潜在因子进行正向映射,同时又作为预测评分的引导项,并进一步将主题概率分布作为到正则约束的一部分.实验结果表明,PreferenceMF 降低了预测评分的误差,提高了推荐方法的质量.

本文的工作主要包括以下 3 个方面.

(1) 提出 PreferenceMF 方法,以用户对商品的单条评论文本为潜在主题学习的处理单元,通过考虑主题与评分矩阵分解后的用户潜在因子和商品潜在因子的映射关系,使这些潜在特征相互关联,并依据此关联设计相应的求解算法,达到提升推荐质量的目的.

(2) 提出主题偏好引导,即在矩阵分解模型中引导潜在因子矩阵的计算.用户潜在因子矩阵解释为用户在某一种潜在特征上的偏好,评论文本中又能够反映用户的个人偏好,因此,将评论中发现的个人偏好融合进潜在因子矩阵,从而能够更好地反映用户偏好.具体来说,在 PreferenceMF 方法中添加基于主题偏好的引导,单条评论文文档的主题发现作为引导项,并将主题概率分布添加到正则约束中提高推荐质量.

(3) 在 Amazon28 组数据子集上测试了 PreferenceMF 方法的推荐质量,分析了数据的稀疏问题,并与已有的方法进行了对比,采用均方误差作为评价指标,与最近相关的 TopicMF 方法相比,在数据子集上均方误差最大减少了 3.32%,平均减少了 0.91%.

1 相关工作

1.1 基于评分数据的推荐方法

亚马逊购物网站是较早借助推荐方法进行商品销售的系统[15].在基于用户评分数据的推荐方法中,通常将用户、商品及评分关系表示成矩阵的数学形式.研究者提出了基于用户的协同过滤方法、基于商品的协同过滤方法和基于模型的推荐方法等.传统的基于用户或商品的协同过滤方法是通过计算用户之间或者商品之间的相似度之后,用最相似的若干个用户或者商品的历史评分数据来预测对未知商品的打分.相似度的计算大多采用欧氏距离、皮尔逊相关系数、余弦距离等,并且研究表明,相似度计算的改进可以提高推荐性能[7,8,16,17].

在基于模型的推荐方法中,目前比较有影响且被广泛研究的是 Koren 等人提出的潜在因子分解模型(latent factor models,简称 LFM)[9,18,19].传统的分解模型一般是从奇异值分解(SVD)模型开始的,需要首先将评分矩阵补全,再使用补全得到的稠密矩阵完成分解.这不仅在存储上带来了很大的限制,而且在计算复杂度上也显著地升高.2006 年 Netflix 竞赛后,Simon Funk 提出了一种矩阵分解的改进算法,称为 Funk-SVD,后来被 Netflix 竞赛的冠军 KorenY 进一步优化为 LFM,能够对不完整的评分数据实现矩阵分解,从此逐步有了其他的矩阵分解模型 NMF、PMF 等[3,13,20−22].

基于评分数据的推荐方法得到了广泛的研究和应用,然而也存在一定的问题.由于用户的数量和商品的数量非常庞大,用户一般不可能对所有的物品进行反馈评分,导致评分矩阵中存在很大的空缺.这种数据的稀疏性使得模型不能得到足够的训练,最终影响了推荐系统的性能.此外,新的消费者或者是新上架的商品由于没有历史评分,导致无法准确地进行相似度计算,从而引起推荐的冷启动问题.根据推荐系统的不同应用领域,研究者近年来通过挖掘商品的标签和用户的参与信息,例如用户的关系、用户的评论、用户的行为等来提高推荐系统的性能[3].本文通过挖掘评论文本的潜在主题,将其融合到基于评分的矩阵分解模型中.接下来主要介绍近年来这方面的研究现状.

1.2 加入评论文本的推荐方法

较早已有研究表明,评论文本有助于提高推荐系统的性能.Basilico 等人提出的推荐方法将回归分析应用于文本内容[10];Ganu 等人以句子为单元对评论文本分类和进行情感极性分析,并将这些信息融合到基于 KNN 的协同过滤推荐方法中[11].较早的这些研究侧重于发现评论文本中用户多方面的偏好,通过标注这些重要偏好来提高评分预测的精度.他们所采用的文本处理方法是以单词为单元的词袋模型,没有考虑到文字背后的语义关联.2003 年,由 Jordan 等人提出的潜在狄利克雷概率模型(latent Dirichlet allocation,简称 LDA)在概率语义分析之上加入了贝叶斯框架,是目前为主最重要的语义主题挖掘模型之一[23].此外,Lee 等人在文献中提出了非负矩阵分解模型(non-negative matrix factorization,简称 NMF),也可以用于文本的潜在主题分析[24].

近年来,对评分与评论文本的融合主要关注潜在特征(潜在因子和潜在主题)的分析.Mcauley 等人提出的 HFT(user)方法将用户评论集的 LDA 主题分布映射到矩阵分解的用户潜在因子向量,而 HFT(item)方法则将商品评论集的 LDA 主题分布映射到商品潜在因子向量[12].TopicMF 方法认为,应该同时考虑反映用户偏好的主题和商品特性的主题[13],并使用了 NMF 方法发现文本的主题,得到的每一条评论的主题分布是用户偏好和商品特性的综合,并将该主题分布映射到矩阵分解后的用户潜在因子向量和商品潜在因子向量.文献[14]首先利用主题模型挖掘评论文本中隐含的主题分布,然后用主题分布刻画用户偏好和商品画像,最后在逻辑回归模型上训练主题与打分的关系.文献[25]提出一种无监督的推荐方法,在评论文本中发现情感主题,主要考虑情感主题与评分的关系.Wang 等人提出将 PMF(probabilistic matrix factorization)与主题分布结合,用于推荐科学文章[22].文献[26]提出用混合高斯模型取代矩阵分解模型,再与 LDA 主题融合建模.黄璐等人提出将用户的主题模型和商品的主题模型与矩阵分解模型相结合,并将商品的标签信息和用户行为数据同时加以考虑,以提升推荐结果的多样性[27].Chen 等人提出矩阵分解模型结合文本信息的策略,直接通过用户评论集合和商品评论集的主题分布进行引导[28].Li 等人采用 Generative Models 来学习评论文本和评分数据在方面级(aspect)的关系[29].本文提出的方法是基于矩阵分解和主题模型来同步学习潜在主题和隐因子的映射关系,并使用潜在主题正则约束进行引导预测评分.此外,在跨领域的推荐问题上,融入评论文本的方法也显著提高了推荐准确率[30,31].文献[32]直接将主题模型用于评分数据,以提高单类协同过滤推荐算法(one-class CF)的质量.

与本文研究比较相近的有 HFT 方法和 TopicMF 方法.这两者之间有 3 点不同:一是主题模型,前者采用 LDA 模型,而后者使用了 NMF;二是文本粒度,前者是由某个用户的所有商品评论组成,而后者是某个用户对某个商品的单条评论;三是前者将潜在主题与用户潜在因子矩阵或者商品潜在因子矩阵之一进行融合建模,而后者是与两个潜在因子矩阵同时融合建模.通过以上文献调研,我们发现,基于评分矩阵和评论文本融合建模的推荐方法可以进一步深度融合,并且以用户对商品的单条评论文本为潜在主题学习的文档单元更有利于用户偏好的分析.本文提出的 PreferenceMF 推荐方法将单条评论文本的潜在主题与评分矩阵分解的用户潜在因子和商品潜在因子建立映射关系,同时,以此潜在主题作为用户偏好引入基本的矩阵分解模型中,并将主题概率分布作为正则约束,从而建立深度融合的推荐模型以及参数同步求解算法,达到减少预测评分的误差、提高推荐质量的目的.

2 PreferenceMF 商品推荐方法

本文主要研究在评分矩阵存在数据稀疏问题的情况下,如何将评论文本数据融合到基于评分矩阵的推荐模型中.本节分别从潜在特征同步学习和偏好引导两个方面入手给出 PreferenceMF 推荐方法的数学模型,并讨论推荐模型的建立和参数求解算法.

2.1 PreferenceMF中的潜在特征同步学习

2.1.1 基本思想

本节先考虑 PreferenceMF 中潜在特征同步学习部分.这部分将单条评论所隐藏的用户打分偏好融入到传统的隐因子模型(LFM)中.某个用户对某个商品的打分会受到用户的个人偏好与商品特性的影响,即用户的打分背后是综合了个人与商品两者的因素,在本文中统称为“打分偏好”.该打分偏好在这里可以理解为影响打分

的潜在隐性特征,而用户的评论文本能够反映出这种隐性特征.换言之,就是评论文本的潜在主题反映的是用户对商品不同方面的“打分偏好”[3],如图 1 所示的例子.

本文通过将用户的每一条评论文本的主题同步映射至打分矩阵的用户隐因子(个人偏好)和商品隐因子(商品特性),从而约束该用户对未打分商品的评分预测.这就是潜在特征同步学习的基本思想,也是本文提出的 PreferenceMF 推荐方法重点考虑的方面之一.与本文最为相近的两个方法是来自 McAuley 等人提出的 HFT (hidden factors and hidden topics)方法[12]和 Yang 等人提出的 TopicMF 方法[13].在第 1.2 节中,我们讨论了它们之间的不同点,这里直接给出 PreferenceMF 中潜在特征同步学习部分.

2.1.2 潜在特征同步学习

首先定义文档集,并引入主题模型来解决评论文本的主题发现问题.用户 i 对商品 j 的评论定义为 $d_{i,j}$,评论集为$\{d_{i,j}\}$,实例见表 1.

Table 1 Examples of users' review documents used for learning latent topics

表 1 用于学习潜在主题的用户评论文档集示例

Docs	Reviews	UsersID	ItemsID
1	Larger extra nooks putting extra accessories foot pedal manuals extra books material …	76	21
2	Received foot yesterday tiny size feels singer zipper cording footit compact version bulky …	76	48
3	Looked art supply stores crazy cutting scissors grandson fun metal plastic blades job …	24	58

根据表 1,这里将每一条评论文本作为一个文档(文本粒度),对应到用户 i 对商品 j 的打分上.接下来将用户 i 对商品 j 的打分(含个人偏好和商品特性)映射到评论文档集的潜在主题(打分偏好),构建如图 2 所示的参数逻辑关系.

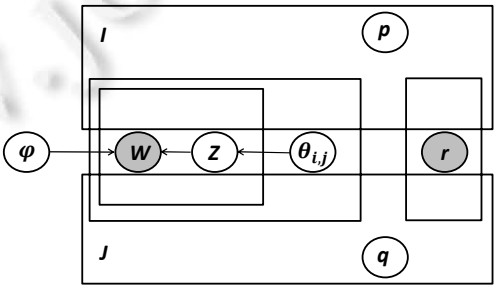


Fig.2 Logic relation diagram of parameters in learning latent features

图 2 潜在特征同步学习的参数逻辑关系图

图 2 中,I 表示用户集合,将产生用户和潜在特征关系的矩阵和用户评论集合;J 表示商品集合,将产生商品和潜在特征关系的矩阵和商品评论集合;r 表示用户 i 对商品 j 的打分;p 和 q 分别表示 r 在用户 i 上的潜在因子向量和在商品 j 上的潜在因子向量;$\theta_{i,j}$ 就是评论集$\{d_{i,j}\}$中评论 $d_{i,j}$ 的主题分布;W 是评论中词的集合;Z 是 W 的主题;φ表示主题词分布.从图 2 中可以看到,评论文本的潜在主题(打分偏好)分布与评分矩阵的潜在因子向量 p(个人偏好)和 q(商品特性)可以建立映射关系,即在上文中所描述的打分偏好可以映射到用户的个人偏好与商品特性上.本文采用 TopicMF[13]中推荐效果较好的映射关系,如公式(1)所示.

$$\theta_{i,j,k}=\frac{\exp(\kappa\mid p_{i,k}\cdot q_{j,k}\mid)}{\sum_{k'=1}^{K}\exp(\mid \kappa\cdot p_{i,k'}\cdot q_{j,k'}\mid)} \tag{1}$$

其中,$\theta_{i,j,k}$ 表示用户 i 对商品 j 在主题 k 上的概率,$p_{i,k}$ 和 $q_{j,k}$ 分别表示用户 i 和商品 j 对应于主题 k 的潜在因子.$\theta_{i,j,k}$ 与 $p_{i,k}$ 和 $q_{j,k}$ 成正相关关系.参数 κ 用于控制映射关系的权重,κ 越大,意味着用户倾向于最重要的主题;κ 越小,用户对所有主题倾向于同等对待.此外,$\theta_{i,j,k}$ 即对应于图 2 中的 $\theta_{i,j}$,$p_{i,k}$ 和 $q_{j,k}$ 对应于图 2 中的 p 和 q.建立如公式(1)所示的映射关系后,不需要同时对 $\theta_{i,j}$,$p_{i,k}$ 和 $q_{j,k}$ 这 3 个参数进行拟合求解.在 PreferenceMF 方法中,若先只考虑潜在特征同步学习,则需要最小化优化的目标函数参见公式(2),其中,Θ和Φ分别是评分矩阵分解和主题发现模型中

的参数集合,Z 是每一个词的主题,κ是映射关系公式(1)中的参数.

$$O(\tau \mid \Theta,\Phi,\kappa,z)=\sum_{r_{i,j}\in\tau}(r_{i,j}-\tilde{r}_{i,j})^2-\lambda\log L(\tau \mid \theta,\phi,z) \tag{2}$$

在公式(2)中,

- $r_{i,j}$ 为评分矩阵分解模型的预测打分,如公式(3)所示.

$$\tilde{r}_{i,j}=\sum_k p_{i,k}\cdot q_{j,k} \tag{3}$$

- $L(\tau|\theta,\phi,z)$即为 LDA 主题发现模型的似然函数,一个文档集合主题模型的似然函数参见公式(4).

$$p(\tau \mid \theta,\phi,z)=\prod_{m\in\tau}\prod_{j=1}^{N_m}\theta_{z_{m,j}}\phi_{z_{m,j},\omega_{m,j}} \tag{4}$$

公式(4)中,τ是语料中的所有文档,ω是文档中的词,N_m 表示 m 文档中的词数目,$\theta_{z_{m,j}}$ 表示从文档-主题分布 θ_m 中采样的主题 $z_{m,j}$,$\phi_{z_{m,j},\omega_{m,j}}$ 表示从主题-词分布 $\phi_{z_{m,j}}$ 中采样得到词$\theta_{m,j}$,相关细节和参数求解可参见文献[23].

可以看出,本文拟最小化的目标公式(2)本质上是用主题概率分布取代了传统评分矩阵分解中的正则化项,再通过公式(1)建立评论文本的潜在主题和评分矩阵的潜在因子之间的正向映射关系,从而达到潜在特征同步学习的目标.拟合求解公式(2)中参数的基本流程参见算法 1 所示.

算法 1. PreferenceMF 中潜在特征同步学习的参数求解基本流程.

输入:评分评论集合.

输出:公式(2)中的参数Θ,Φ,κ,Z.

1. 评分评论集合的预处理,例如评论文本分词等
2. 基于用户的评分构建评分矩阵模型 LFM(公式(3))
3. 定义每一条评论为一篇文档,构建 LDA 主题发现模型(公式(4))
4. 融合步骤 2、步骤 3 构建模型,生成误差目标函数(公式(2))
5. Gibbs 采样方法求解潜在主题参数(公式(4))
6. 拟牛顿法求解目标函数中的参数(公式(2))
7. 重复步骤 5、步骤 6,直到达到迭代次数

2.1.3 参数的拟合求解

在公式(2)中需要优化参数Θ,Φ,τ,Z,其中,$\Theta=\{\mu,b_i,b_j,p_i,q_j\}$,$\Phi=\{\theta,\phi\}$分别是评分矩阵模型中的参数和主题发现模型中的参数,然而,因为 p 和 q 与θ具有映射关系的约束,因此不能独立进行拟合求解,通过采用两步迭代方法进行求解,如公式(5)和公式(6)所示.

$$\arg\min_{\Theta^{(t)},\Phi^{(t)},\kappa^{(t)}}\sum_{r_{i,j}\in\tau}(r_{i,j}-\tilde{r}_{i,j})^2-\lambda\log\prod_{d\in D}\prod_{h=1}^{N_d}\theta_{d,z_{d,h}^{(t-1)}}\phi_{z_{d,h}^{(t-1)},w_{d,h}} \tag{5}$$

$$sample\rightarrow z_{d,h}^{(t)},\ p(z_{d,h}^{(t)}=k)=\theta_{d,k}\phi_{k,w_{d,h}}^{(t)} \tag{6}$$

主题发现似然函数是通过单条评论作为文档进行输入,并建立隐性特征之间的映射关系.在公式(5)中,通过对参数求偏导数,使用拟牛顿法进行求解.公式(5)经过变换,见公式(7).

$$\arg\min_{\Theta^{(t)},\Phi^{(t)},\kappa^{(t)}}\sum_{r_{i,j}\in\tau}(r_{i,j}-\tilde{r}_{i,j})^2-\lambda\sum_{d=1}^{D}\sum_{k=1}^{K}n_{d,k}\left(\kappa\cdot\mid p_{i,k}\mid\cdot\mid q_{j,k}\mid-\log\sum_{k'=1}^{K}\exp(\kappa\cdot\mid p_{i,k'}\mid\cdot\mid q_{j,k'}\mid)\right)-\lambda\sum_{d=1}^{D}\sum_{h=1}^{N_d}\log\phi_{z_{d,h},w_{d,h}} \tag{7}$$

其中,d 表示一个文档,D 表示文档数,$n_{d,k}$ 表示文档 d 中出现主题 k 的个数,其他参数可以参见前文描述的意义.为了能够使用拟牛顿法求解各个参数,需要求解未知参数的偏导数.μ,b_i,b_j可以在 $\sum_{r_{i,j}\in\tau}(r_{i,j}-\tilde{r}_{i,j})^2$ 中求得,ϕ可以在 $\sum_{d=1}^{D}\sum_{h=1}^{N}\log O_{z_{d,h},w_{d,h}}$ 求得.接下来需要从公式(8)中取得参数 p,q 以及控制参数κ的偏导数.

$$\lambda\sum_{d=1}^{D}\sum_{k=1}^{K}n_{d,k}\left(\kappa\cdot|p_{i,k}|\cdot|q_{j,k}|-\log\sum_{k'=1}^{K}\exp(\kappa\cdot|p_{i,k'}|\cdot|q_{j,k'}|)\right) \tag{8}$$

参数 p,q 和κ的部分重要偏导式如公式(9)~公式(11)所示.

$$\frac{\partial O}{\partial p_i}=\frac{\partial O}{\partial p_i}-\lambda\sum_{j=1}^{J}\sum_{d=1}^{D}\sum_{k=1}^{K}\kappa\cdot\left(n_{d,k}\cdot|q_{j,k}|-n_d\cdot|q_{j,k}|\cdot\exp(\kappa\cdot|p_{i,k}|\cdot|q_{j,k}|)\Big/\sum_{k'=1}^{K}\exp(\kappa\cdot|p_{i,k'}|\cdot|q_{j,k'}|)\right) \tag{9}$$

$$\frac{\partial O}{\partial q_i}=\frac{\partial O}{\partial q_i}-\lambda\sum_{i=1}^{I}\sum_{d=1}^{D}\sum_{k=1}^{K}\kappa\cdot\left(n_{d,k}\cdot|p_{i,k}|-n_d\cdot|p_{i,k}|\cdot\exp(\kappa\cdot|p_{i,k}|\cdot|q_{j,k}|)\Big/\sum_{k'=1}^{K}\exp(\kappa\cdot|p_{i,k'}|\cdot|q_{j,k'}|)\right) \tag{10}$$

$$\frac{\partial O}{\partial \kappa}=\frac{\partial O}{\partial \kappa}-\lambda\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{d=1}^{D}\sum_{k=1}^{K}|p_{i,k}|\cdot|q_{j,k}|\cdot\left(n_{d,k}-n_d\cdot\exp(\kappa\cdot|p_{i,k}|\cdot|q_{j,k}|)\Big/\sum_{k'=1}^{K}\exp(\kappa\cdot|p_{i,k'}|\cdot|q_{j,k'}|)\right) \tag{11}$$

从前面的阐述中可知,主题分布θ并非是从潜在狄利克雷分布中获取,而是从前一步中的$\Theta^{(t)}$得到.整个潜在特征同步学习算法的伪代码如算法 2 所示.

算法 2. 潜在特征同步学习的参数求解.

输入:*Ds* 评分评论集合.

输出:目标函数最小化而求得的局部最优参数解.

1. 分词等数据预处理
2. 得到评论文本 corpus
3. 定义一条评论为一篇文档
4. 统计主题词等用于偏导计算
5. 当迭代次数小于主题模型参数求解的最大迭代次数时(公式(4)和公式(5))
6. 当迭代次数小于梯度法迭代次数时
7. 构建误差目标函数 *f*(公式(7)),*f*←*lsq*(*W*[*NW*]),*W*[*NW*]为优化参数数组
8. 拟牛顿法求解 *Lib-BGFS-Operation*(*f*)
9. 重复步骤 7 和步骤 8,直到迭代次数不满足循环条件
10. 重复步骤 6~步骤 9,直到迭代次数不满足循环条件
11. 返回求解得到的参数数组 *W*[*NW*]

2.2 PreferenceMF中的用户个性偏好引导

基于邻域影响的矩阵分解模型(SVD++)是在传统的矩阵分解模型中添加了邻域的影响,邻域可以看作是一种用户偏好引导项.此外,目前矩阵分解模型结合评论文本主要有两种策略:一种是从正则项来约束,另一种是通过文本信息来引导.融合了上述两种策略的方法得到了较好的推荐效果.本文提出的方法也是融合两种策略的方法,然而不同于目前已有的研究工作在于:引导项不是直接通过用户评论集合和商品评论集的主题分布进行引导,而是用第 2.1 节中潜在特征同步学习中的映射关系进行引导,正则项使用公式(2)的主题正则约束.

潜在特征同步学习中,将单条评论文本的主题同步映射至打分矩阵的用户隐因子(个人偏好)和商品隐因子(商品特性),目标函数再通过与主题正则项结合,从而预测该用户对未打分商品的评分.本节在潜在特征同步学习的基础上考虑用户个性偏好引导,即在矩阵分解模型中引导潜在因子矩阵的计算.例如,用户潜在因子矩阵解释为用户在某一种潜在因素上的偏好,评论文本中又能反映用户的个人打分偏好,因此将评论中发现的个人打分偏好融合进潜在因子矩阵更能够反映用户的真实偏好.本节基于前面的潜在特征同步学习,再进一步添加基于主题引导的融合方法,得到完整的 PreferenceMF 方法.它以每一条评论文文档进行主题发现,将其映射到引导项,并将主题概率分布添加到正则约束中,PreferenceMF 的参数逻辑关系如图 3 所示,参数的含义同图 2.

本文定义符号 y_i 表示用户的个人偏好向量,即从评论文本中学习得到的潜在主题分布,用于引导潜在因子向量,得到预测评分公式,参见公式(12).

$$\tilde{r}_{i,j}=\mu+b_i+b_j+(p_i+y_i)\cdot q_j \tag{12}$$

其中,μ表示所有评分记录的全局平均值,作为整体模型的偏置;b_i 和 b_j 作为用户偏置和商品偏置.考虑到在实际

的评分情况中存在一些固有的特性是和用户或物品没有直接关系的,例如某些商品由于装饰和布局等等因素,用户的评分往往是高分.相反,有些商户的商品通常是低分.另外,用户自身也存在着差异,存在着一些用户习惯于给高分,另一些用户习惯于给低分.同理,在物品上,有些物品的质量较高,样式较好,通常都是高分;然而另一些则通常是低分.因此,为了能够体现这些特性,在模型中考虑添加了偏置项μ,b_i和b_j.

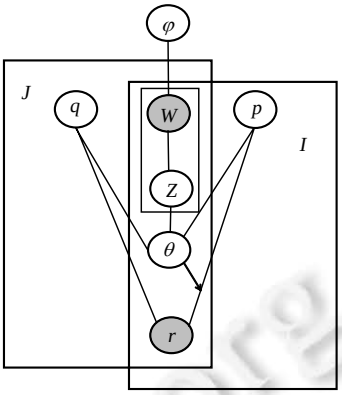


Fig.3 Logic relation diagram of parameters in PreferenceMF

图 3 PreferenceMF 的参数逻辑关系图

公式(12)中的p_i和q_j分别表示用户和商品潜在因子向量,用户的个人偏好y_i与潜在因子向量p_i和q_j的映射关系与主题分布映射θ_i一样,参见公式(1).再继续加入主题偏好作为正则项,得到拟优化的目标函数,整合后参见公式(13)所示.

$$\min_{P,Q}\sum_{\tilde{r}_{i,j}\in\tau}(r_{i,j}-(\mu+b_i+b_j+(p_i+y_i)\cdot q_j))^2-\lambda\log L(\tau\mid\theta,\phi,z) \tag{13}$$

其中,θ,ϕ,z表示 LDA 主题发现模型中的参数,与p和q的映射关系同公式(1).求解公式(13)可以通过求解各参数的偏导数,最后通过梯度下降法或拟牛顿法求解,流程和参数求解与算法 1 和算法 2 类似.

3 实验结果与分析

本节首先介绍实验数据和评价指标,然后通过对比本文的 PreferenceMF 方法与已有方法在推荐质量上的差异,对实验结果进行分析和讨论.

3.1 数据集

本文使用的数据集取自于 Amazon.com 电商网站,该数据集的主要内容是用户对该网站商品的打分、评论及评论的用户有用性反馈.数据的评分与评论的时间跨度是 1994 年 6 月~2013 年 3 月约为 18 年,大小约为 3.3GB,共约 3 500 万条评论数据,统计数量见表 2.

Table 2 Dataset statistics

表 2 数据集统计信息

评论数	用户数	商品数	平均词数	时间跨度
34 686 770	6 643 669	2 441 053	82	1995.6~2013.3

在整个数据集中按照该电商网站的商品分类,共抓取了 28 种类别.每一条样本中包含有如下字段:商品标识、商品名称、商品价格、用户标识、署名、有用性反馈、评分、时间、评论标题文本、评论正文文本.实验中,为了更好地提升推荐质量,对评论文本进行了预处理,包括分词(本文使用 IKAnalyzer)、去停用词、去噪等.将数据集按 4:1 随机划分为训练集和测试集,进行了 20 次的交叉验证,将平均值作为最终结果.由于实验在单机下完成,因此对 28 组子集中大于 1GB 的数据子集进行随机抽取.硬件配置为 CentOS6.5,4 核 Intel CORE i5 CPU,

8GB 内存,使用 C/C++语言实现.

3.2 评价指标

当预测输出为实数值时,通常用于评估推荐模型的指标有均方根误差(RMSE)、均方误差(MSE)和平均绝对值误差(MAE)等,都是计算预测值和真实值之间的误差.为了与对比算法[12,13]在同一评估方法下进行对比,本文采用 MSE 作为推荐模型的评价指标,定义如公式(14)所示.

$$MSE=\frac{\sum_{u,i\in TestSet}(r_{i,j}-\tilde{r}_{i,j})^2}{|TestSet|} \tag{14}$$

其中,*TestSet* 表示测试样本集,|*TestSet*|表示样本集数目.MSE 值越小,表示系统的推荐质量越好.同时,RMSE 是 MSE 开根号后的值,而 MSE 为 MAE 的平方,因此它们对推荐质量的评价与 MSE 的保持一致.

3.3 评分数据的稀疏性分析

如果将电商平台中用户对商品的评分构成一个矩阵,则其规模巨大,且用户数量和商品数量都在不断地增加.然而每一个用户购买的商品总数相对来说只占很少一部分,矩阵中大部分的元素都为 0,只有较少的用户评分和评论数据能够用于模型计算,这个现象被称为数据的稀疏性.本文对 Amazon 全部子集中评分数据的统计结果表明其确实具有稀疏性,如图 4 所示,其中,横坐标代表不同类别的数据集,纵坐标代表评分个数占商品总数的比例.

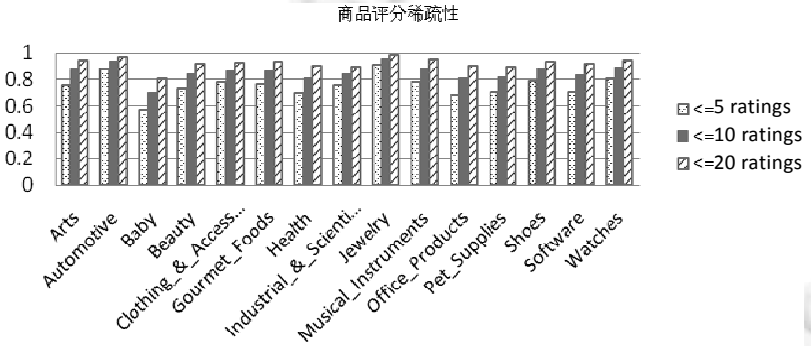


Fig.4 Data sparsity of item ratings in Amazon datasets

图 4 Amazon 数据集中商品评分稀疏性

例如,在图 4 中,Arts 数据集中具有 4 211 件商品,其中 75.3%的商品评分少于 5 个,87.1%少于 10 个,94.3%少于 20 个;Automotive 数据集中具有 47 571 件商品,其中 87.7%的商品少于 5 个,93.8%少于 10 个,97.1%少于 20 个等等.总体来看,平均 75.3%的商品评分少于 5 个,平均 85.3%的商品评分数量少于 10 个,平均 92.1%的商品评分数量少于 20 个.上述统计结果表明,在评分评论数据集中,评分的数据具有稀疏性,将影响矩阵分解模型的预测评分准确性.以往的研究使用数据初始化填充,基于内容的推荐等来缓解该问题.本文考虑融合评论文本的数据,对商品评论文本中单词的统计量结果如图 5 所示,横坐标代表不同类别的数据集,纵坐标代表评论文本单词个数占单词总数的比例.

图 5 反映了每种商品的评论单词的数量,例如,在 Arts 数据集中,67.1%的商品评论中词数大于 20,42.5%的商品评论中词数大于 50;Automotive 数据集中,60.5%的商品评论中词数大于 20,31.1%的商品评论中词数大于 50;Baby 数据集中,84.5%的商品评论词数大于 20,66.1%的商品评论中词数大于 50.总体来看,平均统计后得到 71.99%的商品评论中词数大于 20,47.33%的商品评论中词数大于 50.统计结果表明,文字性的评论文本能够从不同方面反映用户和商品特性,有助于缓解数据的稀疏性问题.因此,可以对评论文本内容加以利用,并融合到传统的评分数据中.

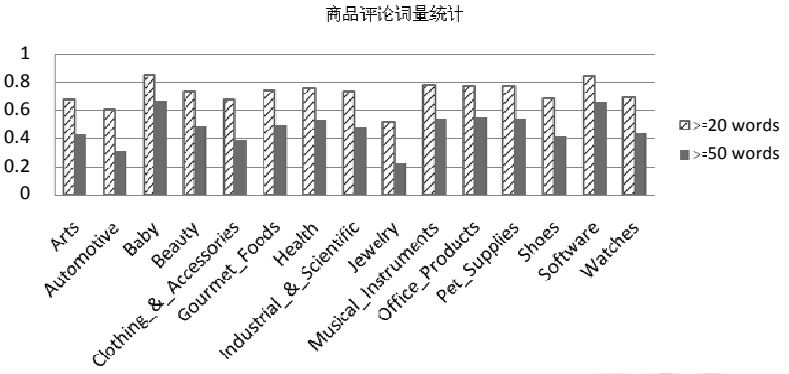


Fig.5 Word statistics in the item reviews of Amazon datasets

图 5 Amazon 数据集中商品评论词量统计

3.4 推荐方法结果对比

3.4.1 对比方法

本文提出的 PreferenceMF 方法与以下 4 种传统方法进行实验对比.

(1) Offsets 方法将直接采用全局偏置,即用户评分的平均值作为预测.这是最简单的预测方法,作为对比的基准方法.

(2) LFM(latent factor model),隐因子分解方法[9].通过矩阵分解方法进行未知商品的评分预测.此方法中没有应用到用户的评论文本信息.

(3) HFT(user)和 HFT(item)方法[10],融合评分与评论文本的推荐方法.将评论划分为某用户的全部评论集或某商品的评论集作为 LDA 主题模型的处理对象,得到的主题与 LFM 矩阵分解中的用户潜在因子或者商品潜在因子进行融合.

(4) TopicMF 方法[11],面向全部评论集,得到每一条评论的主题分布,将主题概率分布添加 LFM 矩阵分解的正则项中,其核心思想是,通过每一条评论的主题分布去正则化 LFM.

(5) PreferenceMF 方法,为本文提出的优化模型.以每一条评论文文档进行 LDA 主题发现,将主题偏好作为用户潜在因子的引导项,并将主题概率分布添加到正则约束中构建融合推荐模型.

3.4.2 对比结果

实验结果表明,潜在因子数 K 等于 5 或者 10 以上时,大部分方法的误差趋于稳定.此外,主题正则权重参数经过实验也选取了结果较优的参数值.因此,本文采用 K=5,正则参数λ=0.5.对上述方法进行 20 组交叉验证,得到平均 MSE 结果,如图 6 所示.

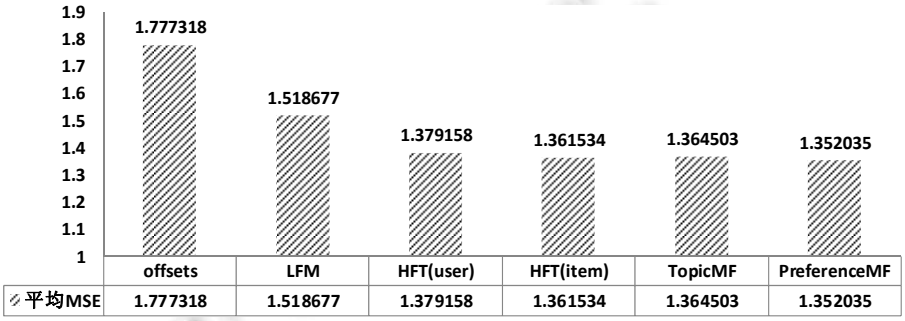


Fig.6 Results of six recommendation approaches in terms of mean square error (MSE)

图 6 6 种推荐方法平均均方误差(MSE)结果

从图 6 中 6 种推荐方法的比较结果可以发现,本文中提出的 PreferenceMF 方法 MSE 值最小,将没有考虑用户偏好引导的 TopicMF 方法的误差降低了 0.91%,并且与 offsets,LFM 以及 HFT(user)相比优势明显.

上述 6 种方法可以分为 3 类:第 1 类是非矩阵分解方法(offsets),第 2 类是潜在因子矩阵分解模型(LFM),第 3 类是融入了评论文本信息的潜在因子矩阵分解模型(HFT(user),HFT(item),TopicMF,PreferenceMF).

第 1 类方法没有采用基于潜在特征的矩阵分解,第 2 类方法是使用了矩阵分解方法.从结果上来看,矩阵分解方法推荐质量较高,减少了预测评分误差.从理论上看,矩阵分解得到的潜在因子为高层特征,反映了有用户历史信息的偏好,这对于预测未评分的用户有显著的积极作用.

第 3 类方法融入了评论文本主题,差别在于融合的方法以及融合程度,以及在评论文本主题挖掘上的不同.TopicMF 方法以单条评论为一篇文档进行主题发现,与用户和商品隐因子矩阵同时融合,比仅融合用户偏好的 HFT(user)方法更优.本文提出的 PreferenceMF 不仅将主题发现融合到潜在因子分解矩阵,而且用主题分布作为偏好引导项,其优点在于同时考虑了用户偏好和商品特性.

3.4.3 数据子集上的对比结果(融入评论文本的方法)

本文提出的 PreferenceMF 方法添加入用户偏好的引导,以 MSE 作为评估方法,在 28 组 Amazon 的数据子集上进行了实验,结果同 HFT(user),HFT(item)、TopicMF 这 3 种融入了评论文本的推荐方法进行了比较,并给出相对于 TopicMF 的预测评分误差减少的比例,见表 3.

Table 3 Comparison of PreferenceMF with other recommendation approaches using review text (MSE)

表 3 PreferenceMF 与其他融入评论文本的推荐方法比较(MSE)

Datasets	HFT(user)	HFT(item)	TopicMF	PreferenceMF	Imp (%)	Imp (%)	Imp (%)
	a	b	c	e	e vs. a	e vs.b	e vs. c
Amazon_Instant_Video	1.403 331	**1.237 879**	1.241 715	1.240 183	**11.63**	–0.18	**0.12**
Arts	1.403 269	1.398 137	1.405 139	**1.393 092**	**0.73**	**0.36**	**0.86**
Automotive	**1.423 605**	1.426 024	1.446 262	1.425 431	–0.13	**0.04**	**1.44**
Baby	1.465 442	**1.449 722**	1.485 505	1.470 345	–0.33	–1.42	**1.02**
Beauty	1.392 671	1.336 476	1.347 839	**1.335 463**	**4.11**	**0.08**	**0.92**
Books	1.268 35	1.250 076	1.260 907	**1.244 52**	**1.88**	**0.44**	**1.30**
Cell_Phones&Accessories	**2.107 897**	2.112 647	2.117 026	2.114 434	–0.31	–0.08	**0.12**
Clothing&Accessories	0.375 229	0.349 138	0.354 129	**0.342 365**	**8.76**	**1.94**	**3.32**
Electronics	1.816 131	1.819 961	1.811 834	**1.764 93**	**2.82**	**3.02**	**2.59**
Gourmet_Foods	**1.457 064**	1.460 673	1.486 422	1.469 823	–0.88	–0.63	**1.12**
Health	1.554 054	1.505 57	1.516 725	**1.503 543**	**3.25**	**0.13**	**0.87**
Home&Kitchen	1.664 797	1.530 157	1.535 965	**1.530 346**	**8.08**	–0.01	**0.37**
Industrial&Scientific	0.344 32	0.344 535	0.345 636	**0.342 029**	**0.67**	**0.73**	**1.04**
Jewelry	**1.197 654**	1.202 037	1.223 719	1.203 432	–0.48	–0.12	**1.66**
Kindle_Store	1.451 514	1.431 611	1.428 913	**1.414 756**	**2.53**	**1.18**	**0.99**
Movies&TV	1.382 538	1.376 415	1.349 808	**1.339 873**	**3.09**	**2.65**	**0.74**
Music	1.031 931	1.030 399	1.014 069	**1.001 421**	**2.96**	**2.81**	**1.25**
Musical_Instruments	1.387 048	1.389 323	1.370 014	**1.369 748**	**1.25**	**1.41**	**0.02**
Office_Products	**1.666 77**	1.676 403	1.677 756	1.669 872	–0.19	**0.39**	**0.47**
Patio	**1.708 102**	1.715 282	1.725 853	1.713 843	–0.34	**0.08**	**0.70**
Pet_Supplies	1.555 542	**1.549 682**	1.560 526	1.553 243	**0.15**	–0.23	**0.47**
Shoes	0.235 528	0.218 553	**0.217 364**	0.220 19	**6.51**	–0.75	–1.30
Software	2.245 577	2.270 249	2.249 234	**2.232 983**	**0.56**	**1.64**	**0.72**
Sports&Outdoors	1.177 441	**1.147 139**	1.151 593	1.147 937	**2.51**	–0.07	**0.32**
Tools&Home_Improvement	1.498 481	1.499 123	1.515 904	**1.493 213**	**0.35**	**0.39**	**1.50**
Toys&Games	1.390 398	1.378 072	1.361 513	**1.349 734**	**2.92**	**2.06**	**0.87**
Video_Games	1.515 444	1.515 565	1.490 55	**1.470 873**	**2.94**	**2.95**	**1.32**
Watches	**1.496 303**	1.502 108	1.514 18	1.499 345	–0.20	**0.18**	**0.98**
AVG	**1.379 158**	**1.361 534**	**1.364 504**	**1.352 035**	**2.32**	**0.68**	**0.92**

从表 3 中可以看到,带有偏好引导的 PreferenceMF 方法在 16 组数据子集上优于其他 3 种融合方法,最大提升了 3.32%.PreferenceMF 的平均 MSE 相对于 TopicMF 的平均 MSE 提高了 0.92%.同时,通过计算可知,相对于 HFT(user),PreferenceMF 提升了 2.32%;相对于 HFT(item),PreferenceMF 方法提升了 0.68%.综上可知,同时添加主题偏好引导和主题正则化项的 PreferenceMF 推荐质量最优.

3.4.4 数据子集上的对比结果(未融合评论文本的方法)

从图 4 中可以看到,融合评论文本的方法比没有融合评论文本的方法推荐质量要高.本节对没有融合评论文本 offsets 和 LFM 两种方法进一步对比.基于全局偏置的 Offsets 方法,是通过商品的平均值作为该商品的预测值,即用户 i 对商品 j 没有打分,使用商品 j 所有打分的平均分来预测.LFM 方法采用的是带有偏置的 SVD 模型,如公式(15)所示.

$$\min \sum_{(i,j)\in TrainSet}\left(r_{ij}-\mu-b_i-b_j-\sum_{k=1}^{K}p_{i,k}p_{j,k}\right)^2+\lambda(\|p_i\|^2+\|q_j\|^2) \tag{15}$$

公式(15)的参数含义与公式(12)一致,模型使用随机梯度下降法(SGD)求解在 Amazon 不同数据子集上 Offsets 和 LFM 方法的结果见表 4.

Table 4 Comparison experimental results of offsets and LFM without using review text (MSE)[3]

表 4 不考虑评论文本的 Offsets 和 LFM 的实验结果比较(MSE)[3]

Datasets	Offset	LFM	Imp. (%)
	a	b	b vs. a
Amazon_Instant_Video	1.816 718	**1.387 078**	23.65
Arts	1.717 373	**1.532 365**	10.77
Automotive	1.802 015	**1.594 104**	11.54
Baby	1.906 211	**1.770 408**	7.12
Beauty	1.755 679	**1.412 187**	19.56
Books	1.474 779	**1.362 384**	7.62
Cell_Phones_&_Accessories	2.318 377	**2.287 076**	1.35
Clothing_&_Accessories	1.597 384	**0.392 945**	75.40
Electronics	2.125 713	**2.006 641**	5.60
Gourmet_Foods	1.680 040	**1.650 013**	1.79
Health	1.852 469	**1.652 961**	10.77
Home_&_Kitchen	2.001 117	**1.759 401**	12.08
Industrial_&_Scientific	1.285 686	**0.414 979**	67.72
Jewelry	1.486 781	**1.260 129**	15.24
Kindle_Store	1.679 401	**1.633 415**	2.74
Movies_&_TV	1.688 753	**1.530 791**	9.35
Music	1.211 798	**1.099 868**	9.24
Musical_Instruments	1.596 279	**1.551 396**	2.81
Office_Products	2.069 281	**1.821 759**	11.96
Patio	2.095 429	**1.915 252**	8.60
Pet_Supplies	1.879 021	**1.795 096**	4.47
Shoes	1.374 472	**0.262 346**	80.91
Software	2.788 576	**2.528 138**	9.34
Sports_&_Outdoors	1.620 245	**1.261 090**	22.17
Tools_&_Home_Improvement	1.855 477	**1.681 265**	9.39
Toys_&_Games	1.670 877	**1.585 947**	5.08
Video_Games	1.836 921	**1.786 549**	2.74
Watches	1.578 050	1.587 378	–0.59
AVG.	**1.777 318**	**1.518 677**	**14.55**

表 4 中粗体表示 MSE 的值最低.从表中可以看出,基于矩阵分解的方法在几乎所有子集合上(除 watches 子集合外),都比基于全局偏置的 Offsets 方法更优,将 Offsets 方法的均方误差减少了 14.55%.因此可以得出结论,潜在因子矩阵分解模型是推荐质量较好的方法.

3.5 实验结果分析

从上面的实验结果可以看出,本文提出的 PreferenceMF 方法在整体上表现最优.我们对上述实验中的 6 种方法进行归纳总结如下.

- 没有融入评论文本的 Offsets 和 LFM 方法推荐质量不如融入评论文本的方法、基于矩阵分解的 LFM 优于 Offsets.
- HFT(user)方法是将用户评论文本的主题分布与用户评分矩阵的潜在因子向量构成映射关系,融入到训练模型中,它考虑到了用户偏好对用户评分的影响.
- HFT(item)方法是以商品特性分布与商品潜在因子向量构成映射关系,融入到训练模型中.文献[12]中

指出,用户在对商品打分时关注商品特性多于关注自身的偏好.实验结果表明,HFT(item)效果优于HFT(user).

- TopicMF 方法是更换了主题发现的评论集主体,采用单条评论作为主题模型的输入.而 HFT(user)和HFT(item)的对主题偏好的学习是通过整体评论集使用主题发现,定义了主题偏好分布与用户潜在因子向量和商品潜在因子向量的映射.实验结果表明,TopicMF 方法均优于 HFT(item)和 HFT(user),单条评论能够更精准地表达某用户对某商品的个性偏好.
- 本文提出的 PreferenceMF 是在 TopicMF 基础上加入了主题偏好的引导.经过实验验证,该引导项可以进一步提升推荐质量.后续有待进一步研究该方法在解决冷启动问题的能力.

4 总结和展望

本文提出了融合评分与评论文本的推荐方法,即加入了偏好引导的 PreferenceMF 方法.从不同角度进行分析后,给出了模型的数学表述形式和参数拟合求解的算法.最后,在 28 组 Amazon 数据子集上与多种相关的推荐方法进行实验对比,结果表明,本文提出的 PreferenceMF 方法的推荐质量相对于目前主流的方法有进一步提升.并分析了各个方法的实验结果以及存在的问题.今后可以对隐因子数量和映射关系做进一步研究.本文在融合方法中假设主题发现的主题数量与潜在因子数相等,并且具有相同的权重,可以考虑不同权重的情况.在解决数据稀疏性、模型冷启动和长尾现象等问题上,都有待进一步展开研究.此外,本文是从最大化似然的角度来使推荐模型适应数据,今后可以采用贝叶斯估计,从最大化后验分布的角度来对模型进行建模和求解.

References:

[1] Goldberg D, Nichols D, Oki B M, Terry D. Using collaborative filtering to weave an information tapestry. Communications of the ACM, 1992,35(12):61−70.

[2] Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J. GroupLens: An open architecture for collaborative filtering of netnews. In: Proc. of the 1994 ACM Conf. on Computer Supported Cooperative Work. New York: ACM Press, 1994. 175−186.

[3] Su C. Recommendation algorithms by exploiting rating matrix and review texts [MS. Thesis]. Wuhan: Wuhan University of Technology, 2016.

[4] Lerman K, Blair-Goldensohn S, McDonald RT. Sentiment summarization: Evaluating and learning user preferences. In: Proc. of the 12th Conf. of the European Chapter of the Association for Computational Linguistics (EACL). Stroudsburg: Association for Computational Linguistics, 2009. 514−522.

[5] Sharma A, Cosley D. Do social explanations work? Studying and modeling the effects of social explanations in recommender systems. In: Proc. of the 22nd Int'l World Wide Web Conf. (WWW). New York: ACM Press, 2013. 1133−1144.

[6] Chen L, Chen G, Wang F. Recommender systems based on user reviews: The state of the art. User Modeling and User-adapted Interaction, 2015,25(2):99−154.

[7] Wen JH, Shu S. Improved collaborative filtering recommendation algorithm of similarity measure. Computer Science, 2014,41(5): 68−71 (in Chinese with English abstract).

[8] Huang X, Qin Z, Chen H. A new user similarity measurement based on a local item space in collaborative filtering recommendation. Journal of Computational Information Systems, 2015,11(10):3501−3508.

[9] Hu Y, Koren Y, Volinsky C. Collaborative filtering for implicit feedback datasets. In: Proc. of the 8th IEEE Int'l Conf. on Data Mining (ICDM). Washington: IEEE, 2008. 263−272.

[10] Basilico J, Hofmann T. Unifying collaborative and content-based filtering. In: Proc. of the 21st Int'l Conf. on Machine Learning (ICML). New York: ACM Press, 2004. 65−72.

[11] Ganu G, Elhadad N, Marian A. Beyond the stars: Improving rating predictions using review text content. In: Proc. of the 12th Int'l Workshop on the Web and Databases. New York: ACM Press, 2009. 1−6.

[12] McAuley J, Leskovec J. Hidden factors and hidden topics: Understanding rating dimensions with review text. In: Proc. of the 7th ACM Conf. on Recommender Systems (RecSys). New York: ACM Press, 2013. 165−172.

[13] Bao Y, Fang H, Zhang J. Topicmf: Simultaneously exploiting ratings and reviews for recommendation. In: Proc. of the 28th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2014. 2−8.

[14] Gao YF, Yu WZ, Chao PF, Zheng ZL, Zhang R. Analyzing reviews for rating prediction and item recommendation. Journal of East China Normal University (Natural Science), 2015,(3):80−90 (in Chinese with English abstract).

[15] Linden G, Smith B, York J. Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 2003, 7(1):76−80.

[16] Choi K, Suh Y. A new similarity function for selecting neighbors for each target item in collaborative filtering. Knowledge-based Systems, 2013,37(1):146−153.

[17] Wu X, Cheng B, Chen JL. Collaborative filtering service recommendation based on a novel similarity computation method. IEEE Trans. on Services Computing, 2017,10(3):352−365.

[18] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. IEEE Computer, 2009,42(8):30 – 37.

[19] Koren Y. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In: Proc. of the 14th ACM Int'l Conf. on Knowledge Discovery and Data Mining (KDD). New York: ACM Press, 2008. 426−434.

[20] Salakhutdinov R, Mnih A. Probabilistic matrix factorization. In: Proc. of the Advances in Neural Information Processing Systems (NIPS). Red Hook: Curran Associates Inc., 2007. 1257−1264.

[21] Xiao R, Li Y, Chen H, *et al*. SRSP-PMF: A novel probabilistic matrix factorization recommendation algorithm using social reliable similarity propagation. In: Proc. of the Intelligent Computing Theories and Methodologies. Springer Int'l Publishing, 2015. 80−91.

[22] Wang C, Blei DM. Collaborative topic modeling for recommending scientific articles. In: Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD). New York: ACM Press, 2011. 448−456.

[23] Jordan MI, Blei DM, Ng AY. Latent dirichlet allocation. Journal of Machine Learning Research, 2003,3:993−1022.

[24] DDL, HS.S. Learning the parts of objects by non-negative matrix factorization. Nature, 1999,401(6755):788−791.

[25] Diao Q, Qiu M, Wu CY, Smola AJ, Jiang J. Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS). In: Proc. of the 20th ACM Sigkdd Int'l Conf. on Knowledge Discovery and Data Mining (KDD). New York: ACM Press, 2014. 193−202.

[26] Ling G, Lyu MR, King I. Ratings meet reviews, a combined approach to recommend. In: Proc. of the 8th ACM Conf. on Recommender Systems (RecSys). New York: ACM Press, 2014. 105−112.

[27] Huang L, Lin CJ, He J, Liu HY, Du XY. Diversified mobile app recommendation combining topic model and collaborative filtering. Ruan Jian Xue Bao/Journal of Software, 2017,28(3):708−720 (in Chinese with English abstract). http://www. jos.org.cn/1000-9825/5163.htm [doi: 10.13328/j.cnki.jos.005163]

[28] Chen X, Yao Y, Xu F, Lu J. Exploring review content for recommendation via latent factor model. In: Proc. of the13th Pacific Rim Int'l Conf. on Artificial Intelligence. Springer Int'l Publishing, 2014. 668−679.

[29] Li H, Lin R, Hong R, Ge Y. Generative models for mining latent aspects and their ratings from short reviews. In: Proc. of the 15th IEEE Int'l Conf. on Data Mining (ICDM). Washington: IEEE, 2015. 241−250.

[30] Song TH, Peng ZH, Wang SZ, Fu WJ, Hong XG, Yu PS. Review-based cross-domain recommendation through joint tensor factorization. In: Proc. of the 22nd Int'l Conf. on Database Systems for Advanced Applications (DASFAA). Springer Int'l Publishing, 2017. 525−540.

[31] Xin X, Liu Z, Lin C, Huang H, Wei X, Guo P. Cross-domain collaborative filtering with review text. In: Proc. of the 24th Int'l Joint Conf. on Artificial Intelligence (IJCAI). Palo Alto: AAAI Press, 2015. 1827−1833.

[32] Zhang HJ, Li ZJ, Chen Y, Zhang XM, Wang SZ. Exploit latent dirichlet allocation for one-class collaborative filtering. In: Proc. of the 23rd ACM Int'l Conf. on Information and Knowledge Management (CIKM). New York: ACM Press, 2014. 1991−1994.

附中文参考文献:

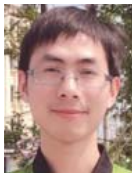
[3] 苏畅.融合评分矩阵和评论文本的推荐算法研究[硕士学位论文].武汉:武汉理工大学,2016.

[7] 文俊浩,舒珊.一种改进相似性度量的协同过滤推荐算法.计算机科学,2014,41(5):68−71.

[14] 高祎璠,余文喆,晁平复,郑芷凌,张蓉.基于评论分析的评分预测与推荐.华东师范大学学报(自然科学版),2015,(3):80−90.

[27] 黄璐,林川杰,何军,刘红岩,杜小勇.融合主题模型和协同过滤的多样化移动应用推荐.软件学报,2017,28(3):708−720. http://www.jos.org.cn/1000-9825/5163.htm [doi: 10.13328/j.cnki.jos.005163]

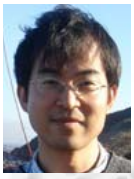
李琳(1977－),女,湖南衡阳人,博士,教授,CCF 专业会员,主要研究领域为数据挖掘,推荐系统,信息检索.



朱阁(1993－),男,硕士,主要研究领域为机器学习,数据挖掘.

解庆(1986－),男,博士,副教授,CCF 专业会员,主要研究领域为数据库,数据挖掘,推荐系统,知识服务.

苏畅(1992－),男,硕士,主要研究领域为机器学习,数据挖掘.



杨征路(1976－),男,博士,教授,博士生导师,主要研究领域为自然语言处理,人工智能,数据挖掘,信息检索.