软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(9):2869−2885 [doi: 10.13328/j.cnki.jos.005540]

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

融合多种数据信息的餐馆推荐模型*

戴 琳[1,2], 孟祥武[1,2], 张玉洁[1,2], 纪威宇[1,2]

[1](智能通信软件与多媒体北京市重点实验室(北京邮电大学),北京 100876)

[2](北京邮电大学 计算机学院,北京 100876)

通讯作者: 孟祥武, E-mail: mengxw@bupt.edu.cn

摘 要: 餐馆推荐可以利用用户的签到信息、时间上下文、地理上下文、餐馆属性信息以及用户的人口统计信息等挖掘用户的饮食偏好,为用户生成餐馆推荐列表.为了更加有效地融合这些数据信息,提出一种融合了多种数据信息的餐馆推荐模型,该模型首先利用签到信息和时间上下文构建“用户-餐馆-时间片”的三维张量,同时利用其他数据信息挖掘若干用户相似关系矩阵和餐馆相似关系矩阵;然后,在概率张量分解的基础上同时对这些关系矩阵进行分解,并利用 BPR 优化准则和梯度下降算法进行模型求解;最后得到预测张量,从而为目标用户在不同时间片生成相应的餐馆推荐列表.通过在两个真实数据集上的实验结果表明:相比于目前存在的餐馆推荐模型,所提出的模型有着更好的推荐效果和可接受的运行时间,并且缓解了数据稀疏性对推荐效果的影响.

关键词: 餐馆推荐;概率张量分解;相似关系;BPR 优化;梯度下降

中图法分类号: TP311

中文引用格式: 戴琳,孟祥武,张玉洁,纪威宇.融合多种数据信息的餐馆推荐模型.软件学报,2019,30(9):2869−2885. http://www.jos.org.cn/1000-9825/5540.htm

英文引用格式: Dai L, Meng XW, Zhang YJ, Ji WY. Restaurant recommendation model with multiple information fusion. Ruan Jian Xue Bao/Journal of Software, 2019,30(9):2869−2885 (in Chinese). http://www.jos.org.cn/1000-9825/5540.htm

Restaurant Recommendation Model with Multiple Information Fusion

DAI Lin[1,2], MENG Xiang-Wu[1,2], ZHANG Yu-Jie[1,2], JI Wei-Yu[1,2]

[1](Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia (Beijing University of Posts and Telecommunications), Beijing 100876, China)

[2](School of the Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: Restaurant recommendation can leverage check-ins, time, location, restaurant attributes, and user demographics to dig user's dining preference, and recommend a list of restaurants for each user. In order to fuse these data information more effectively, this study proposes a restaurant recommendation model with multiple information fusion. Firstly, this model constructs a three-dimensional tensor by using check-ins and time context, and digs some users' similar relation matrices and restaurants' similar relation matrices from additional data information. Secondly, these relation matrices and tensor are decomposed simultaneously. Then, Bayesian personalized ranking optimization criterion method (BPR Opt) and gradient descent algorithm are adopted to solve the model parameters. Finally, the proposed model generates a corresponding restaurant candidate list for target user at different time by calculating predicted tensor. A comprehensive experimental study is conducted on two real-world datasets. The experimental results not only validate the efficacy of the proposed model, which outperforms the current restaurant recommendation model and effectively alleviates influence of the data sparsity on recommendation performance, but also evaluate the efficiency of the proposed model, which has acceptable running time.

Key words: restaurant recommendation; probabilistic tensor factorization; similar relation; BPR-Opt; gradient descent

* 基金项目: 北京市教育委员会共建项目
Foundation item: The Mutual Project of Beijing Municipal Education Commission, China
收稿时间: 2017-07-19; 修改时间: 2017-11-01; 采用时间: 2017-12-29; jos 在线出版时间: 2019-01-21
CNKI 网络优先出版: 2019-01-22 13:48:59, http://kns.cnki.net/kcms/detail/11.2560.TP.20190122.1348.009.html

随着 GPS 在手持设备的广泛应用,基于位置社交网络[1−5]的服务应运而生,并且已经达到了一种前所未有的水平.在基于位置社交网络的服务上,用户可以签到自己的位置,并分享自己的评价.推荐系统通过用户的这些隐式反馈(签到信息)和显式反馈(用户的评分和评论)挖掘用户的偏好,就可以为用户生成兴趣点推荐列表.这种基于位置社交网络的兴趣点推荐系统有很多,典型的包括美国最大的点评网站 Yelp(https://www.yelp.com/sf)、基于用户地理位置的手机服务网站 Foursquare(https://foursquare.com/)以及国内的大众点评(http://www.dianping.com/)等,这些网站通过分析用户签到及评论信息等,挖掘用户个性化偏好,为用户提供合适的选择,从而为用户带来便利,同时为商家带来可观的利益.本文所研究的餐馆推荐是兴趣点推荐中一个典型的应用.

餐馆推荐可以利用多种数据信息挖掘用户的饮食偏好,为用户生成餐馆推荐列表.从文献[6,7]可知:除了用户的显式反馈(评分、评论等)和隐式反馈(签到)直接反映了用户的偏好,还有以下因素也影响用户的餐馆选择:(1) 时间上下文,在不同时间段用户的饮食偏好是不同的;(2) 地理上下文,用户通常会选择活动区域附近的餐馆;(3) 用户的人口统计信息,例如,不同年龄或不同性别的用户对于餐馆的需求是不同的,有的人看中服务,有的人看中环境等;(4) 餐馆的属性信息,例如,用户对于餐馆的选择通常集中在某几类风格.因此,如果能够同时考虑这些数据信息,并有效地进行整合,挖掘这些信息的最大价值,就能够提高推荐精度.

文献[8]考虑了用户行为的空间聚类现象,提出一种结合地理位置信息的矩阵分解模型,该模型在加权矩阵分解的基础上引入了地理上下文,将用户的活动区域向量融合到用户隐式空间中,将兴趣点的区域影响向量融合到兴趣点隐式空间中,通过这种融合,有效地解决了矩阵稀疏性问题,从而提高了推荐准确度;但是该模型没有考虑时间上下文信息,也没有考虑其他元数据信息.

文献[6]基于用户的隐式反馈提出了一种隐式偏好模型,该模型同时考虑了时间上下文、地理上下文以及餐馆属性信息,分别利用概率张量分解和逻辑回归获得用户的隐式偏好.该方法能够很好地提高推荐准确度,但是该方法是分别对各数据信息相对独立地进行分析,缺乏对所有数据信息进行有效地融合,且时间复杂度高.

文献[9]考虑到用户存在一些依赖关系,而用户的这些关系会相互影响用户的偏好,因此,作者提出了一种概率关系矩阵分解模型.该模型的创新点在于不再仅仅只考虑用户的社交关系,而通过学习用户的依赖提高了推荐准确度,但是该方法缺乏对时间上下文和地理上下文的研究.

文献[10]为了解决下一个兴趣点的推荐问题,提出了一种两步策略的方法:首先,基于用户签到的兴趣点种类构建三维张量,并提出一种新的优化准则(LBPR)对张量优化学习,进而得到预测的兴趣点种类;然后,根据预测的兴趣点种类获得位置列表.该方法的推荐效果要优于目前存在的方法,但是该方法没有考虑时间上下文的影响,也缺乏对其他元数据信息的研究,例如用户信息或者兴趣点信息等.

文献[11]认为,用户的兴趣点会随着时间和当前位置的变化而变化,因此,作者提出了两步策略的方法进行兴趣点的推荐:首先,根据用户签到的兴趣点种类和时间上下文构建四维张量,预测用户偏好的下一个兴趣点种类;然后,基于预测的兴趣点种类进一步获得位置列表.该模型的创新点在于既考虑了时间和位置,还降低了数据稀疏性的影响,但是缺乏对其他元数据信息的研究.

为了同时考虑多种数据信息,并进行有效的融合,本文提出了一种融合多种数据信息的餐馆推荐模型(a restaurant recommendation model with multiple information fusion,简称 RRMIF).该模型首先利用签到信息和时间上下文构建“用户-餐馆-时间片”的三维张量,同时,利用其他数据信息挖掘若干用户相似关系矩阵和餐馆相似关系矩阵;然后,在概率张量分解模型[12]的基础上同时对这些关系矩阵进行分解,并保证张量和矩阵分解后有共同的低维隐式因子矩阵;最后,利用 BPR[13,14]优化准则和梯度下降算法进行模型求解.可以看出,该模型将多种数据信息通过用户、餐馆以及时间片的隐式因子矩阵进行了有效地融合,这也就是本文提出的模型与目前存在的模型最大的区别.值得注意的是,为了降低模型的复杂度,本文提出的模型没有考虑用户的显式反馈.本文的贡献主要包括以下几点.

1) 与现有研究不同,本文没有简单地将时间上下文按照小时制划分为 24 个时间片,而通过 *K*-means 聚类算法[15]将一天分为 4 个用餐时间段,这样不仅对用户的用餐行为进行了聚类,还降低了模型的复杂度;

2) 本文基于地理上下文、餐馆属性信息构建了两种餐馆相似关系矩阵,基于用户人口统计信息构建了用户相似关系矩阵;进而提出一种融合多种数据信息的餐馆推荐模型,该模型是在概率张量分解模型的基础上同时对用户相似关系和餐馆相似关系进行分解,它以用户和餐馆的隐式因子作为桥梁,更好地融入了多种数据信息;最后,采用 BPR 优化准则和梯度下降算法进行模型求解;

3) 实验在两种真实的数据集上进行,主要包括以下 4 个部分:1) 比较本文提出的模型和现有模型的推荐效果;2) 研究多种数据信息对于推荐效果的影响,包括时间上下文、地理上下文、餐馆属性信息以及用户人口统计信息;3) 研究本文提出的模型在不同稀疏度数据集的表现,并与现有模型作比较;4) 研究本文提出的模型的运行时间,并与现有模型作比较.实验结果表明:相比于目前存在的餐馆推荐模型,本文提出的餐馆推荐模型有着更好的推荐效果和可接受的运行时间,并且缓解了数据稀疏性对推荐效果的影响.

本文第 1 节介绍餐馆推荐的相关工作,包括兴趣点推荐和餐馆推荐.第 2 节介绍多种数据信息的研究,提出一种融合多种数据信息的餐馆推荐模型.第 3 节介绍实验的相关设置及实验结果分析.第 4 节是全文的总结.

1 相关工作

1.1 兴趣点推荐

兴趣点推荐[16]作为位置社交网络的一个重要应用,已经成为学术界和工业界的一个热门课题.它给用户和商家都带来了前所未有的便利和好处:对于用户而言,兴趣点推荐系统能够将其从海量的兴趣点搜索中解放出来,根据他们的历史数据挖掘其个性化偏好,为他们推荐合适的兴趣点;另一方面,对于商家而言,可以吸引大量感兴趣的用户,持续提高经济效益.

用户隐式反馈的研究是当前兴趣点推荐的一个热点,矩阵分解模型(MF)[17]就常常被用于隐式反馈数据的处理.该模型基于用户的历史签到数据构建“用户-兴趣点”的签到矩阵,再通过对签到矩阵的分解,得到用户隐式因子矩阵和兴趣点隐式因子矩阵,再利用这些隐式因子矩阵预测用户对于兴趣点的评分,进而为用户生成推荐列表.虽然该模型有着较高的准确率,且模型简单,但是隐式反馈中只有正反馈,当用户在某兴趣点没有签到时,并不能反映用户就不喜欢该兴趣点,传统的矩阵分解没有很好地解决在这一问题.为了从隐式反馈中获得额外的信息,Hu 等人[18]提出了加权矩阵分解模型,该模型在概率矩阵分解的基础上,通过分析用户的隐式反馈,得到正反馈和负反馈的置信水平,从而使得矩阵分解模型更好地应用于隐式反馈的研究.但是由于“用户-兴趣点”矩阵的稀疏性,该方法依旧面临着很大的挑战.为了解决稀疏性问题,Lian 等人[8]提出了结合地理上下文的矩阵分解模型(GeoMF).该模型在加权矩阵分解的基础上将地理上下文引入模型,让用户的活动区域向量融合到用户隐式空间中,将兴趣点的区域影响向量融合到兴趣点隐式空间中.通过这种融合,不仅考虑了用户行为的空间聚类现象,还有效地解决了矩阵稀疏性问题.但是该模型仅仅只考虑了用户的签到信息和地理上下文,而没有考虑用户的评论信息以及时间上下文信息.为了融合用户的评论信息,Li 等人[19]提出了一种多方面考虑的兴趣点推荐系统.该系统从用户对于兴趣点的评论中学习用户的偏好以及商家的质量标签,再结合用户评分矩阵分解得到的用户特征矩阵和商家特征矩阵,预测目标用户对于其他兴趣点的效用评分,最后生成推荐列表.该系统很好地结合了用户的评论信息和评分信息,提高了推荐的精度.但是该系统没有考虑到时间上下文信息.考虑到时间信息对于兴趣点推荐重要性,Luan 等人[20]提出了基于张量分解的协同过滤模型.该模型利用用户的签到行为构建一个“用户-兴趣点-时间片”的三维张量,同时从不同角度提取 3 个特征矩阵,例如“用户-兴趣点类别”或者“时间片-用户”等特征矩阵,然后利用特征矩阵协同地对张量进行分解,并使得目标函数最小,最后得到预测张量,通过预测张量就可以为目标用户在某一时间对生成兴趣点推荐列表.虽然该模型考虑了时间因素,提高了推荐准确度,但是该模型没有考虑地理上下文.

1.2 餐馆推荐

目前,国内对于餐馆推荐研究得不多,而国外则相对较多.例如,Fu 等人[21]提出了一种考虑多种评分数据,并

融合地理上下文以及用户和餐馆特征信息的餐馆推荐模型.该模型同时对多种评分矩阵进行分解,并利用用户和餐馆的位置信息构建位置相关性,利用用户和餐馆特征信息构建特征相关性,最后利用分解得到的隐式因子矩阵和这两种相关性得到用户对餐馆的预测评分.由于融合了多种评分数据,该模型提高了推荐准确度,但是该模型没有考虑时间上下文信息.为了融入时间上下文,Zhang 等人[6]提出了一种隐式反馈模型.该模型分别从用户签到信息、时间上下文以及其他数据信息获得两种偏好:(1) 构建“用户-餐馆-时间片”的三维签到张量,利用概率张量分解从签到张量中获得隐式空间的偏好;(2) 利用逻辑回归的方法分析用户依赖于其他数据信息的偏好.最后,结合这两种偏好得到用户在某一时间片对餐馆的预测评分,进而为目标用户在某一时间片生成餐馆推荐列表.该模型虽然同时考虑了多种数据信息,但是该模型是针对各种数据信息分别建模,没有完全挖掘这些数据信息的价值.除了考虑以上数据信息,Sun 等人[7]还考虑了用户的社交网络信息,提出了一种考虑多源信息的餐馆推荐模型.该模型的基本思想是:用户的隐式偏好不仅由用户的评分信息决定,还会受用户社交网络以及行为模式相似人群的影响.因此,作者在矩阵分解的框架中融入了评分信息、社交网络信息以及行为模式信息.该模型虽然有着较高的推荐精度,但是没有考虑餐馆的属性信息.

通过对餐馆推荐相关工作的分析,本文借鉴前人的经验,提出了一种融合多种数据信息的餐馆推荐模型.该模型综合考虑了用户的签到信息、时间上下文、地理上下文、餐馆属性信息以及用户人口统计信息,并有效地将这些信息进行融合,从而进一步提高了推荐精度.

2 融合多种数据信息的餐馆推荐模型

本节首先基于大众点评数据集(相关介绍见第 3.1 节)介绍了各种数据信息的研究,通过 *K*-means 算法将一天 24 小时划分为 4 个用餐时间段,另外,基于地理上下文、基于餐馆属性信息构建两种餐馆相似关系,基于用户人口统计信息构建用户相似关系;然后提出了一种融合多种数据信息的餐馆推荐模型;最后,利用 BPR 优化准则和梯度下降算法进行模型求解.

2.1 多种数据信息的研究

2.1.1 时间上下文

文献[6,22]指出:用户在不同时间片对于餐馆的选择是不同的,通过引入时间上下文,可以为目标用户在不同时间片生成相对应的餐馆推荐列表.因此,该文献按照小时制将一天划分成 24 个时间片,但是这种方法存在一定的缺点,即需要在每个时间片生成推荐列表,这会导致模型复杂度增大,运行时间变长.实际上,用户的用餐时间是存在聚类现象的,从图 1 可以看出:用户的聚餐时间存在明显的高峰和低谷,且高峰段大概有 4 段,因此,推荐系统只需要在每个用餐高峰之前为用户生成推荐列表.这样不仅可以降低模型的复杂度,还能保证模型具有较好的推荐效果.

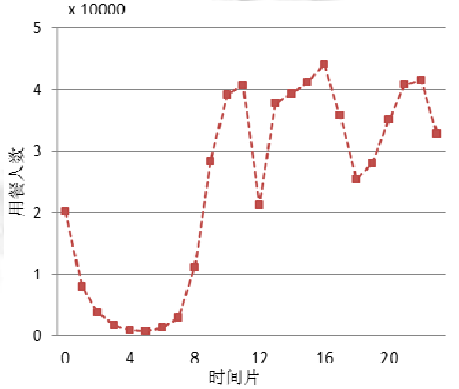


Fig.1 Number of diners at different time slots

图 1 用户在不同时间片的用餐人数

为了更精确地划分用户高峰,本文利用 *K*-means 算法对数据集的用餐时间进行聚类分析,将用户的用餐时间段分成 4 段,聚类结果见表 1.

Table 1 Clustering of dining time

表 1 用餐时间的聚类

用餐时间段	均值	最小值	最大值
时间段 1	0.4	0	5
时间段 2	10.5	6	13
时间段 3	15.8	14	18
时间段 4	21.1	19	23

从表 1 可以看出,通过聚类将一天划分为了 4 个用餐时间段,每一段分别对应[0,5],[6,13],[14,18],[19,23].

2.1.2 地理上下文

在目前的兴趣点推荐和餐馆推荐的研究中,地理位置基本都被看做一种影响用户选择的重要因素.对于餐馆而言,用户通常会选择活动区域内的餐馆,那么活动区域内的餐馆之间理应具有相似性.文献[23,24]认为,距离用户 100km 以内的区域可以看做是用户的活动区域,那么就可以近似认为餐馆 *R* 与距离该餐馆 100km 以内的其他餐馆相似,它们之间的相似表现为:用户如果选择了餐馆 *R*,那么该用户也有可能选择距离餐馆 *R* 100km 以内的其他餐馆.因此,基于地理位置的餐馆相似关系 *A* 可以定义为公式(1):

$$A_{ij}=\begin{cases}1, & \text{if } |loc(i)-loc(j)|\leqslant 100\text{km}\\ 0, & \text{if } |loc(i)-loc(j)|>100\text{km}\end{cases} \tag{1}$$

其中,|*loc*(*i*)–*loc*(*j*)|表示餐馆 *i* 和餐馆 *j* 的距离.

2.1.3 餐馆属性信息

餐馆的属性信息包括餐馆的名字、餐馆的风格以及餐馆的平均消费等,为了研究餐馆的风格与用户偏好的关系,我们随机抽取了 3 名用户,计算了每个用户选择各种餐馆风格的次数,如图 2 所示.需要注意的是,这里的餐馆风格只是所有风格的一部分.

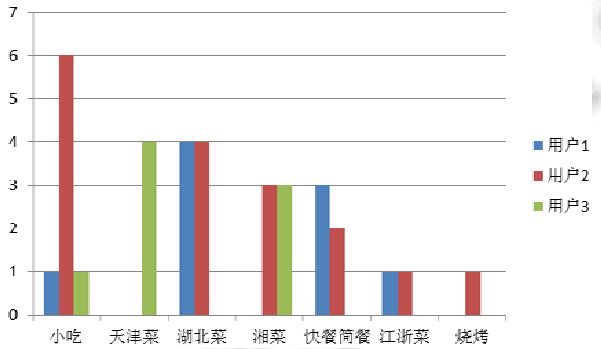


Fig.2 Number of selected times for different restaurant styles

图 2 用户选择不同餐馆风格的次数

从图 2 可以看出,用户通常会多次选择自己喜好的某类餐馆用餐.因此,餐馆风格相同的餐馆理应具有相似性.这种相似性表现为:当用户选择了餐馆 *R*,那么该用户也可能喜欢与餐馆 *R* 风格相同的其他餐馆.因此,基于餐馆风格的餐馆相似关系 *B* 可以定义为公式(2):

$$B_{ij}=\begin{cases}1, & \text{if } style(i)=style(j)\\ 0, & \text{if } style(i)!=style(j)\end{cases} \tag{2}$$

其中,*style*(*i*)表示餐馆 *i* 的风格.

除了餐馆的风格,餐馆的属性还包括餐馆的平均消费,但是由于很多餐馆的签到次数少,且用户通常很少标

注消费金额,导致很多餐馆的平均消费为 0.经统计,数据集中平均消费为 0 的餐馆占 45.1%,因此很难利用餐馆的平均消费进行研究.

2.1.4 用户人口统计信息

用户的人口统计信息一般包括用户的性别、年龄以及居住地,但是由于用户的隐私保护,有的数据很难获取.通过分析用户历史签到的餐馆的所在城市,可以将用户签到次数最多的城市看作是用户的居住地.文献[6]也分析了同一居住地的用户通常会有相似的偏好,例如北京人更偏爱北京菜,湖南人更爱湖南菜.因此,可以得到基于居住地的用户相似关系 E,如公式(3)所示:

$$E_{ij}=\begin{cases}1, & \text{if } residence(i)=residence(j)\\ 0, & \text{if } residence(i)!=residence(j)\end{cases} \tag{3}$$

其中,$residence(i)$表示用户的居住地.

2.2 模型描述

为了更好地融合用户签到信息、时间上下文、地理上下文、餐馆属性信息以及用户人口统计信息,提高推荐精度,本文提出了一种融合多种数据信息的餐馆推荐模型,如图 3 所示,其中,

- I,J,S 分别表示用户、餐馆以及时间片的个数;
- A,B,C,E 是可观察量,其中:A 表示基于地理位置的餐馆相似关系,B 表示基于餐馆风格的餐馆相似关系,C 表示用户的签到张量,E 表示基于居住地的用户相似关系;
- E_{iv} 表示用户 i 与用户 v 的居住地的相似性,A_{jq},B_{jp} 类似.

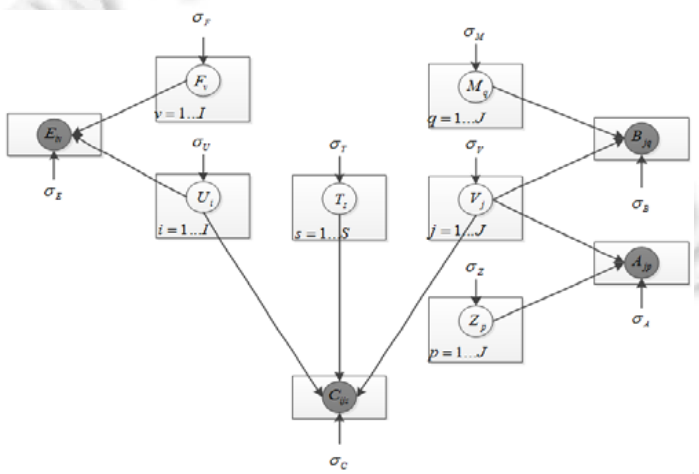


Fig.3 Graphical representation for a restaurant recommendation model with multiple information fusion

图 3 一种融合多种数据信息的餐馆推荐模型的图表示

关于用户签到张量 C 的构建及其含义,在下文会有介绍.另外,表 2 介绍了模型参数的符号及其定义.

Table 2 Symbols and definition of model parameters

表 2 模型参数的符号及定义

符号	意义
U,V,T,F,M,Z	低维隐式因子矩阵
$\sigma_U,\sigma_V,\sigma_T,\sigma_F,\sigma_M,\sigma_Z$	协方差矩阵参数
$\sigma_A,\sigma_B,\sigma_C,\sigma_E$	方差参数
$\alpha_C,\alpha_E,\alpha_A,\alpha_B$	误差权重
$\lambda_U,\lambda_V,\lambda_T$	正则化参数
u_C,u_E,u_A,u_B	学习速率参数
D	隐式因子维度

2.2.1 融合用户签到信息和时间上下文

根据用户的签到信息和时间上下文,可以构建“用户-餐馆-时间片”的三维签到张量 $C\in\{0,1\}^{I\times J\times S}$,其中, I,J,S 分别表示用户、餐馆以及时间片的个数,且通过第 2.1.1 节的分析,时间片的个数设定为 4.如果在时间片 s,用户 i 在餐馆 j 用餐,则 C_{ijs}=1,这代表了用户的正反馈;如果在时间片 s,用户 i 在餐馆 j 没有用餐,则 C_{ijs}=0,这并不一定代表用户的负反馈,因为用户 i 可能并不知道餐馆 j.从图 3 可以看出,通过对三维签到张量 C 的分解,可以得到用户、餐馆和时间片的隐式因子矩阵,分别用 $U_{I\times D},V_{J\times D},T_{S\times D}$ 表示,即:每一个张量元素 C_{ijs} 都可以分解为用户 i、餐馆 j 以及时间片 s 的隐式特征向量,分别表示为 U_i,V_j,T_s.

2.2.2 融合其他数据信息

文献[24]认为用户的社交关系影响了用户行为,因此同时分解社交网络矩阵和评分矩阵,且分解后有一个共同的用户隐式因子矩阵.基于此思想,我们认为:用户 i 的相似用户影响了用户 i 的特征向量,餐馆 j 的相似餐馆也影响了餐馆 j 的特征向量.因此,本文同时分解基于居住地的用户相似关系 E、基于地理位置的餐馆相似关系 A、基于餐馆风格的餐馆相似关系 B 和用户签到张量 C.从图 3 中可以看出,矩阵 E 分解为 $U_{I\times D}$ 和 $F_{I\times D}$,其中,$U_{I\times D}$ 是矩阵 E 和张量 C 分解后的共同隐式因子矩阵;矩阵 A 分解为 $V_{J\times D}$ 和 $Z_{J\times D}$,矩阵 B 分解为 $V_{J\times D}$ 和 $M_{J\times D}$,其中,$V_{J\times D}$ 是矩阵 A,B 以及张量 C 分解后的共同隐式因子矩阵.

2.2.3 模型的基本原理

在概率张量分解模型[12]的研究中,隐式特征向量是由多元高斯分布生成的,而张量的每个元素是由高斯分布生成的,且该分布的均值是由相应的隐式因子决定的.由于本文提出的模型是在概率张量分解模型基础上的扩展,因此该模型的基本原理可以描述如下.

1. 对于用户,$U_i\sim N(\mathbf{0},\sigma_U^2\boldsymbol{I}),F_v\sim N(\mathbf{0},\sigma_F^2\boldsymbol{I})$,其中,$U_i\sim N(\mathbf{0},\sigma_U^2\boldsymbol{I})$ 表示用户的隐式特征向量是由均值向量为 $\mathbf{0}$、协方差矩阵为 $\sigma_U^2\boldsymbol{I}$ 的多元高斯分布生成的,σ_U^2 表示方差参数,$\boldsymbol{I}$ 表示单位矩阵;
2. 对于餐馆,$V_j\sim N(\mathbf{0},\sigma_V^2\boldsymbol{I}),Z_p\sim N(\mathbf{0},\sigma_Z^2\boldsymbol{I}),M_q\sim N(\mathbf{0},\sigma_M^2\boldsymbol{I})$;
3. 对于时间片,$T_s\sim N(\mathbf{0},\sigma_T^2\boldsymbol{I})$;
4. 对于基于居住地的用户相似关系 E 中的每个元素 $E_{iv},E_{iv}\sim N(U_i\cdot F_v,\sigma_E^2)$,其中,$U_i\cdot F_v=\sum_{d=1}^{D}U_{id}F_{vd}$,$E_{iv}\sim N(U_i\cdot F_v,\sigma_E^2)$ 表示 E_{iv} 是由均值为 $U_i\cdot F_v$、方差为 σ_E^2 的高斯分布生成的;
5. 对于基于地理位置的餐馆相似关系 A 中的每个元素 $A_{jp},A_{jp}\sim N(V_j\cdot Z_p,\sigma_A^2)$,其中,$V_j\cdot Z_p=\sum_{d=1}^{D}V_{jd}Z_{pd}$;
6. 对于基于餐馆风格的餐馆相似关系 B 中的每个元素 $B_{jq},B_{jq}\sim N(V_j\cdot M_q,\sigma_B^2)$,其中,$V_j\cdot M_q=\sum_{d=1}^{D}V_{jd}M_{qd}$;
7. 对于三维签到张量 C 中的每个元素 $C_{ijs},C_{ijs}\sim N(U_i\cdot V_j\cdot T_s,\sigma_C^2)$,其中,$U_i\cdot V_j\cdot T_s=\sum_{d=1}^{D}U_{id}V_{jd}T_{sd}$.

2.3 模型参数求解

无论是张量分解还是矩阵分解,都是通过分解得到的低维隐式因子矩阵来求得预测张量或者预测矩阵,并使得预测张量和原始张量的误差最小,使得预测矩阵和原始矩阵的误差也最小.因此,该模型的目标函数可以定义为公式(4):

$$\underset{(U,V,T,F,Z,M)}{\arg\min}\ \alpha_C L(C,U\cdot V\cdot T)+\alpha_E L(E,UF^T)+\alpha_A L(A,VZ^T)+\alpha_B L(B,VM^T)+Reg(U,V,T,F,Z,M) \tag{4}$$

其中:L 是误差函数;Reg 是防止过拟合的正则项;$\alpha_C,\alpha_E,\alpha_A,\alpha_B$ 分别是各个误差项的比重,且$\alpha_C+\alpha_E+\alpha_A+\alpha_B$=1;$U\cdot V\cdot T$ 表示预测张量 $\hat{C}$,其中,预测张量的每一项 $\hat{C}_{ijs}=U_i\cdot V_j\cdot T_s=\sum_{d=1}^{D}U_{id}V_{jd}T_{sd}$.

由于用户的相似关系矩阵 E 是对称矩阵,则 $U=F$;同理,$V=Z=M$.那么,该模型的目标函数可以简化为公式(5):

$$\arg\min_{(U,V,T)} \alpha_C L(C,U\cdot V\cdot T)+\alpha_E L(E,UU^T)+\alpha_A L(A,VV^T)+\alpha_B L(B,VV^T)+Reg(U,V,T) \tag{5}$$

文献[13,14]认为:推荐列表的生成实际上是一种排名问题,通过优化项目的排名,可以优化目标函数,使得模型更快地收敛到最优解.因此,该文献通过 BPR 优化准则和梯度下降算法进行求解.实验表明,直接优化排名的方法要明显优于其他方法.受此启发,本文使用 BPR 优化准则来优化误差函数.

BPR 优化准则的基本思想是:当 C_{ijs}=1,$C_{ij's}$=0 时,用户 i 在时间片 s 对于餐馆 j 的排名要高于餐馆 j'.对于签到张量 C,定义一个集合 P_C 来表示 C 中的排名对,如公式(6)所示:

$$P_C:=\{(i,s,j,j')|C_{ijs}=1\wedge C_{ij's}=0\} \tag{6}$$

对于 $L(c,U\cdot V\cdot T)$的优化,可以转化为公式(7):

$$\arg\min_{(U,V,T)} L(C,U\cdot V\cdot T)=\arg\max_{(U,V,T)} BPR-Opt(C,U\cdot V\cdot T)=\arg\max \sum_{(i,s,j,j')\in P_C} \ln\sigma(\hat{x}^C_{ijj's}) \tag{7}$$

其中,$\hat{x}^C_{ijj's}=\hat{C}_{ijs}-\hat{C}_{ij's}$,$\hat{C}_{ijs}=U_i\cdot V_j\cdot T_s=\sum_{d=1}^{D}U_{id}V_{jd}T_{sd}$,$\sigma(x)=\dfrac{1}{1+e^{-x}}$.需要注意的是,最小化误差函数等价于最大化 BPR.

对于其他误差函数的优化,同样采用 BRP 优化准则,分别如公式(8)~公式(10)所示:

$$\arg\min_{(U)} L(E,UU^T)=\arg\max_{(U)} BPR-Opt(E,UU^T)=\arg\max \sum_{(i,j,j')\in P_E} \ln\sigma(\hat{x}^E_{ijj'}) \tag{8}$$

$$\arg\min_{(V)} L(A,VV^T)=\arg\max_{(V)} BPR-Opt(A,VV^T)=\arg\max \sum_{(i,j,j')\in P_A} \ln\sigma(\hat{x}^A_{ijj'}) \tag{9}$$

$$\arg\min_{(V)} L(B,VV^T)=\arg\max_{(V)} BPR-Opt(B,VV^T)=\arg\max \sum_{(i,j,j')\in P_B} \ln\sigma(\hat{x}^B_{ijj'}) \tag{10}$$

其中,

- 在公式(8)中,$P_E:=\{(i,j,j')\,|\,E_{ij}=1\wedge E_{ij'}=0\}$,$\hat{x}^E_{ijj'}=\hat{E}_{ij}-\hat{E}_{ij'}$,$\hat{E}_{ij}=U_i\cdot U_j=\sum_{d=1}^{D}U_{id}U_{jd}$;
- 在公式(9)中,$P_A:=\{(i,j,j')\,|\,A_{ij}=1\wedge A_{ij'}=0\}$,$\hat{x}^A_{ijj'}=\hat{A}_{ij}-\hat{A}_{ij'}$,$\hat{A}_{ij}=V_i\cdot V_j=\sum_{d=1}^{D}V_{id}V_{jd}$;
- 在公式(10)中,$P_B:=\{(i,j,j')\,|\,B_{ij}=1\wedge B_{ij'}=0\}$,$\hat{x}^B_{ijj'}=\hat{B}_{ij}-\hat{B}_{ij'}$,$\hat{B}_{ij}=V_i\cdot V_j=\sum_{d=1}^{D}V_{id}V_{jd}$.

对于正则项 $Reg(U,V,T)$,为了方便使用梯度下降算法[25]进行求解,本文采用 L_2-regularization[26],如公式(11)所示:

$$\mathrm{Re}\,g(U,V,T)=\lambda_U\|U\|_2^2+\lambda_V\|V\|_2^2+\lambda_T\|T\|_2^2 \tag{11}$$

其中,λ_U,λ_V和λ_T是正则化参数,$\|U\|_2^2$,$\|V\|_2^2$和$\|T\|_2^2$都是 L_2 范数的平方.

综上,该模型的目标函数可以转化为公式(12):

$$\begin{aligned}\arg\max_{(U,V,T)} \alpha_C \sum_{(i,s,j,j')\in P_C} \ln\sigma(\hat{x}^C_{ijj's})+\alpha_E \sum_{(i,j,j')\in P_E} \ln\sigma(\hat{x}^E_{ijj'})+\alpha_A \sum_{(i,j,j')\in P_A} \ln\sigma(\hat{x}^A_{ijj'})+\\ \alpha_B \sum_{(i,j,j')\in P_B} \ln\sigma(\hat{x}^B_{ijj'})+\lambda_U\|U\|_2^2+\lambda_V\|V\|_2^2+\lambda_T\|T\|_2^2\end{aligned} \tag{12}$$

利用梯度下降算法最大化这个目标函数,就可以得到模型参数 U,V,T.需要注意的是,由于是最大化目标函数,因此必须沿着梯度方向迭代,而不是负梯度方向.对于模型参数的完整求解过程见算法 1.

算法 1. 求解模型参数.

输入:用户的签到张量 C,基于居住地的用户相似关系 E,基于地理位置的餐馆相似关系 A,基于餐馆风格和消费的餐馆相似关系 B,方差参数$\sigma_U,\sigma_V,\sigma_T$,误差权重$\alpha_C,\alpha_E,\alpha_A,\alpha_B$,正则化参数$\lambda_U,\lambda_V,\lambda_T$,迭代次数 $Iter$,学习速率参数 u_C,u_E,u_A,u_B,隐式因子个数 D;

输出:模型参数 U,V,T.

1: 初始化 U,V,T

2: **FOR** 迭代=1 到 *Iter* **DO**

3: 采样 P_C

4: **FOR** (i,s,j,j') from P_C

5:
$$U \leftarrow U + u_C\left(\alpha_C \frac{e^{-\hat{x}^C_{ijj's}}}{1+e^{-\hat{x}^C_{ijj's}}} \cdot \frac{\partial}{\partial U}\hat{x}^C_{ijj's} + \lambda_U \cdot U\right)$$

$$V \leftarrow V + u_C\left(\alpha_C \frac{e^{-\hat{x}^C_{ijj's}}}{1+e^{-\hat{x}^C_{ijj's}}} \cdot \frac{\partial}{\partial V}\hat{x}^C_{ijj's} + \lambda_V \cdot V\right)$$

$$T \leftarrow T + u_C\left(\alpha_C \frac{e^{-\hat{x}^C_{ijj's}}}{1+e^{-\hat{x}^C_{ijj's}}} \cdot \frac{\partial}{\partial T}\hat{x}^C_{ijj's} + \lambda_T \cdot T\right)$$

6: **END FOR**

7: 采样 P_E

8: **FOR** (i,j,j') from P_E

9:
$$U \leftarrow U + u_E\left(\alpha_E \frac{e^{-\hat{x}^E_{ijj'}}}{1+e^{-\hat{x}^E_{ijj'}}} \cdot \frac{\partial}{\partial U}\hat{x}^E_{ijj'} + \lambda_U \cdot U\right)$$

10: **END FOR**

11: 采样 P_A

12: **FOR** (i,j,j') from P_A

13:
$$V \leftarrow V + u_A\left(\alpha_A \frac{e^{-\hat{x}^A_{ijj'}}}{1+e^{-\hat{x}^A_{ijj'}}} \cdot \frac{\partial}{\partial V}\hat{x}^A_{ijj'} + \lambda_V \cdot V\right)$$

14: **END FOR**

15: 采样 P_B

16: **FOR** (i,j,j') from P_B

17:
$$V \leftarrow V + u_B\left(\alpha_B \frac{e^{-\hat{x}^B_{ijj'}}}{1+e^{-\hat{x}^B_{ijj'}}} \cdot \frac{\partial}{\partial V}\hat{x}^B_{ijj'} + \lambda_V \cdot V\right)$$

18: **END FOR**

19: **END FOR**

20: **RETURN** 模型参数 U,V,T

从算法 1 可以看出,模型求解的时间复杂度为 $O(Iter\times(|P_C|+|P_E|+|P_A|+|P_B|)\times D)$,其中:*Iter* 为迭代次数;$|P_C|$,$|P_E|$,$|P_A|$和$|P_B|$分别表示集合 P_C,P_E,P_A 和 P_B 中元素的个数,即排名对的个数;D 表示隐性因子的个数.通过学习到的模型参数 U,V,T,就可以得到预测签到张量 $\hat{C}$,从而为目标用户在某一时间片生成餐馆推荐列表.

3 实验及结果分析

3.1 数据集及实验环境

本文使用大众点评数据集(http://yongfeng.me/)和 Yelp 数据集(https://www.yelp.com/dataset)进行餐馆推荐研究.这两种数据集都包括用户的评论数据集和商家的属性数据集,由于用户隐式的保护,都没有公开用户的人口统计信息.用户的评论数据集中都包括用户的评分、评论以及消费信息,不同的是:Yelp 数据集给出的评论时间只具体到哪一天,而大众点评数据集具体到了一天中的某个时刻.另外,本文将用户评论一次可以看做签到一次.商家的属性数据集都包括商家的 ID、名字、位置(城市、经纬度等)、餐馆的种类以及标签等信息.

在这两种数据集中,商家不仅包括餐馆,还包括超市、酒吧以及茶馆等.由于除了餐馆的其他商家与研究的主题无关,因此首先必须从数据集中剔除与餐馆无关的信息.对处理后的两种数据集,各项统计信息见表 3.

Table 3 Statistics of dataset

表 3 数据集的各项统计信息

统计信息	Dianping	Yelp
用户数量	224 021	60 137
餐馆数量	29 147	1 979
签到数量	590 599	98 411
平均每个用户的签到数量	2.6	1.6
平均每个餐馆的签到数量	20.3	49.7
用户的最大签到数量	218	108
用户的最小签到数量	1	1
餐馆的最大签到数量	213	3 517
餐馆的最小签到数量	1	1

从表 3 中可以看出,Yelp 数据集中,用户的数量是 Dianping 数据集中用户数量的将近 3 倍;而 Dianping 数据集中,餐馆的数量是 Yelp 数据集中餐馆数量的近 15 倍.这可能导致模型在 Yelp 数据集中的推荐效果要优于 Dianping 数据集.

为了更好地验证推荐模型的效果,分别将两种数据集划分为训练集(80%)和测试集(20%),训练集主要是用来学习推荐模型中的参数,测试集主要是用来验证模型的推荐效果.

本文的实验环境为:Windows7 操作系统,4GB 内存,Intel(R) Core(TM)2 Duo CPU 2.93GHz,实验程序使用 java1.6 语言开发.

3.2 评价指标

对于基于隐式反馈的推荐而言,MAE 不是一个很好的评价指标,因为与评分不同,隐式反馈只有 1 或者 0,而且由于数据的稀疏性,1 的数目会很少,这样,计算 MAE 是没有意义的.因此,为了验证推荐的效果,本文采用 *recall@K* 作为评价指标,对于 *recall@K* 的定义如公式(13):

$$recall@K=\frac{hit}{recall} \tag{13}$$

其中,*recall@K* 表示在 top-*K* 列表中的召回率;*hit* 表示测试集中的命中次数,所谓命中是指如果测试集中的签到餐馆出现在了 top-*K* 的推荐列表中,那么就表示命中一次;*recall* 表示测试集签到总次数.

3.3 实验结果与分析

在这一节,主要在两种数据集中分别进行以下实验:(1) 比较本文提出的模型(RRMIF 模型)和现有模型的推荐效果;(2) 研究多种数据信息对于推荐效果的影响,包括时间上下文、地理上下文、餐馆风格以及用户居住地;(3) 研究 RRMIF 模型在不同稀疏度数据子集的表现,并与现有模型作比较;(4) 研究 RRMIF 模型的运行时间,并与现有模型作比较.

通过网格搜索法对 RRMIF 模型的参数进行调优,得到实验效果最好的模型参数如下:方差参数σ_U=σ_V=σ_I=0.1;正则化参数λ_U=λ_V=λ_I=0.004;误差参数α_C=0.45,α_E=0.2,α_A=0.3,α_B=0.05;迭代次数 *Iter*=40;学习速率参数u_C=0.2,u_E=0.06,u_A=0.1,u_B=0.04 以及隐性因子个数 *D*=10.

值得注意的是:由于 Yelp 数据集中没有提供用户在一天中具体的签到时间,因此时间片的大小为 1;而对于 Dianping 数据集,按照上文的聚类结果,将时间片的大小设置为 4.

3.3.1 与其他对比模型的比较

为了验证文中提出的 RRMIF 模型的推荐效果,本文选取下面几种利用用户隐式反馈来进行兴趣点推荐的相关模型作为对比.

(1) 概率矩阵分解模型(PMF)[17]:该模型主要是利用了用户的签到信息,将“用户-兴趣点”的签到矩阵分解

为用户隐式因子矩阵和兴趣点隐式因子矩阵,再利用这些隐式因子矩阵预测用户对于兴趣点的评分,进而为用户生成推荐列表;

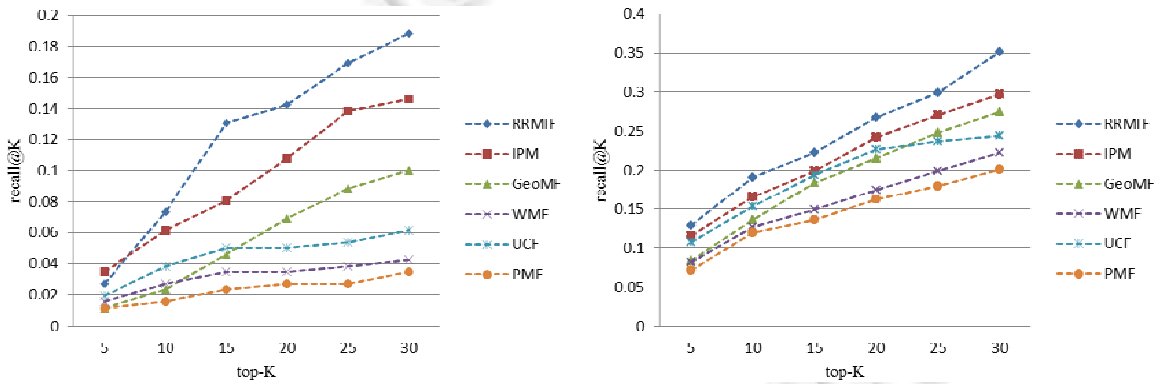
(2) 加权矩阵分解模型(WMF)[18]:该模型在概率矩阵分解的基础上,通过分析用户的隐式反馈,得到正反馈和负反馈的置信水平,从而使得矩阵分解模型更好地应用于隐式反馈的研究;

(3) 基于用户的协同过滤模型(UCF)[27]:该模型主要利用相似度计算目标用户的相似用户集合,然后根据相似用户对于目标项目的评分预测目标用户对目标项目的评分,最后给出推荐列表;

(4) 结合地理位置信息的矩阵分解模型(GeoMF)[8]:该模型在加权矩阵分解的基础上引入了地理上下文,将用户的活动区域向量融合到用户隐式空间中,将兴趣点的区域影响向量融合到兴趣点隐式空间中.通过这种融合,不仅考虑了用户行为的空间聚类现象,还有效地解决了矩阵稀疏性问题;

(5) 隐式反馈模型(IPM)[6]:该模型首先构建“用户-餐馆-时间片”的三维签到张量,利用概率张量分解从签到张量中获得隐式空间的偏好;其次,利用逻辑回归的方法分析用户依赖于其他数据信息(餐馆属性、用户人口统计信息)的偏好;最后,结合这两种偏好得到用户在某一时间片对餐馆的预测评分,进而为目标用户在某一时间片生成餐馆推荐列表.

该实验设置 top-K=5,10,15,20,25,30,且当上述 5 种模型的参数设置为最优参数时,比较各模型在 Dianping 数据集和 Yelp 数据集中的召回率(*recall@K*),结果如图 4 所示.



(a) 大众点评数据集 (b) Yelp 数据集

Fig.4 *recall@K* comparison of different model

图 4 不同模型的召回率比较

观察图 4 可以得到以下结果.

(1) 无论在哪种数据集中,PMF 模型的推荐效果都是最差的,这主要是因为用户的签到数据是稀疏的;另外,PMF 模型没有考虑时间上下文,也没有融合其他数据信息;

(2) 在两种数据集中,WMF 模型的推荐效果都要优于 PMF 模型.这是因为 WMF 模型在 PMF 模型的基础上还分析了正负反馈的置信水平,从而提升了推荐效果;

(3) 在两种数据集中,UCF 模型的推荐效果都要明显优于 WMF 模型和 PMF 模型.但是 UCF 模型存在以下缺点:面对大数据集,模型的复杂度会急剧增大;很难融入其他数据信息,不易扩展.而对于 WMF 模型和 PMF 模型,则更易扩展,且面对大数据集仍然表现不错.因此,WMF 模型和 PMF 模型在适用性上要强于 UCF 模型;

(4) 与 WMF 模型相比,在两种数据集中,GeoMF 模型的推荐效果都要明显优于 WMF 模型.这主要是因为 GeoMF 模型考虑了餐馆的地理位置信息,从而缓解了数据的稀疏性问题,也提高了推荐效果;

(5) 在两种数据集中,IPM 模型的推荐效果都要明显优于 GeoMF 模型、WMF 模型、UCF 模型和 PMF 模型.这是因为 IPM 模型考虑了餐馆的地理位置信息,并融入了时间上下文和其他数据信息;

(6) 在两种数据集中,RRMIF 模型的推荐效果都是最好的.这充分地说明了,虽然 IPM 模型和 RRMIF 模型都考虑了多种数据信息,但是 RRMIF 模型相比于 IPM 模型更加有效地将多种数据信息进行融合,充分挖掘了多种数据信息的价值,从而提升了推荐效果;

(7) Yelp 数据集中,各模型的召回率趋势相比于 Dianping 数据集更加紧凑,而且各模型在 Yelp 数据集中的推荐效果要明显优于 Dianping 数据集.这主要是因为 Yelp 数据集中餐馆的数量要明显少于 Dianping 数据集,导致各模型的差距相对缩小.

实验结果表明:RRMIF 模型相比于现有模型,更加有效地融合了多种数据信息,提升了推荐效果.

3.3.2 多种数据信息对推荐效果的影响

为了研究多种数据信息对于推荐效果的影响,该实验对下面 9 种推荐模型进行对比.

(1) 概率矩阵分解模型(PMF):该模型仅仅考虑用户的签到信息,而没有考虑时间上下文、地理上下文以及其他数据信息;

(2) 概率张量分解模型(PTF)[12]:该模型仅仅考虑用户的签到信息和时间上下文,对用户签到张量进行分解,而不考虑地理上下文和其他数据信息;

(3) PTF+E:该模型不仅考虑用户的签到信息和时间上下文,还考虑了基于居住地的用户相似关系 E,同时对签到张量和关系矩阵 E 进行分解;

(4) PTF+A:该模型不仅考虑用户的签到信息和时间上下文,还考虑了基于地理位置的餐馆相似关系 A,同时对签到张量和关系矩阵 A 进行分解;

(5) PTF+B:该模型不仅考虑用户的签到信息和时间上下文,还考虑了基于餐馆风格的餐馆相似关系 B,同时对签到张量和关系矩阵 B 进行分解;

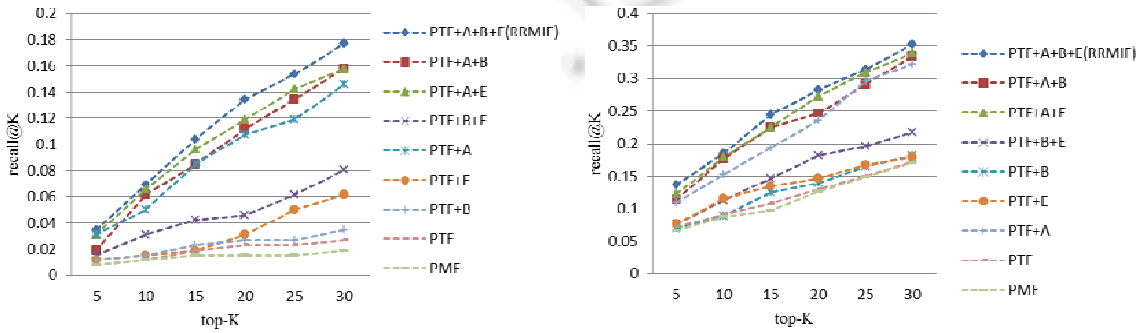
(6) PTF+E+A:该模型不仅考虑用户的签到信息和时间上下文,还考虑了基于居住地的用户相似关系 E 和基于地理位置的餐馆相似关系 A,同时对签到张量、关系矩阵 E 和关系矩阵 A 进行分解;

(7) PTF+E+B:该模型不仅考虑用户的签到信息和时间上下文,还考虑了基于居住地的用户相似关系 E 和基于餐馆风格的餐馆相似关系 B,同时对签到张量、关系矩阵 E 和关系矩阵 B 进行分解;

(8) PTF+A+B:该模型不仅考虑用户的签到信息和时间上下文,还考虑了基于地理位置的餐馆相似关系 A 和基于餐馆风格的餐馆相似关系 B,同时对签到张量、关系矩阵 A 和关系矩阵 B 进行分解;

(9) PTF+A+B+E(RRMIF):该模型就是本文提出的 RRMIF 模型,它不仅考虑用户的签到信息和时间上下文,还考虑了基于居住地的用户相似关系 E、基于地理位置的餐馆相似关系 A 以及基于餐馆风格的餐馆相似关系 B,同时对签到张量、关系矩阵 E、关系矩阵 A 和关系矩阵 B 进行分解.

该实验设置 top-K=5,10,15,20,30,且当上述 9 种模型的参数设置为最优参数时,比较各种模型在两种数据集中的召回率(*recall@K*),如图 5 所示.



(a) 大众点评数据集 (b) Yelp 数据集

Fig.5 Research of multiple data information

图 5 多种数据信息的研究

从图 5 中可以看出:

(1) 在 Dianping 数据集中,PMF 模型的推荐效果要低于 PTF 模型,这主要是因为 PTF 模型考虑了时间上下文信息;而在 Yelp 数据集中,PMF 模型的推荐效果几乎与 PTF 模型一样,这主要是因为 Yelp 数据集中没有用户签到的具体时间,因此时间片设置为 1,因此,PTF 模型的效果基本和 PMF 模型相同;

(2) 在两种数据集中,PTF+A>PTF+E>PTF+B>PTF.这验证多种数据信息对于用户餐馆选择的影响,并且餐馆的地理位置影响最大,其次是用户的居住地,最后是餐馆的风格;

(3) 在两种数据集中,PTF+A+E>PTF+A+B>PTF+B+E.这说明如果只考虑两种数据信息,同时考虑餐馆地理位置和用户居住地的效果是最好的;

(4) 在两种数据集中,PTF+A>PTF+B+E.这说明考虑餐馆的地理位置对于推荐效果的提升比同时考虑用户的居住地和餐馆风格还要大;

(5) 无论在哪种数据集中,PTF+A+B+E 的推荐效果都是最好的,这个模型也就是本文提出 RRMIF 模型.这说明当同时考虑了餐馆的地理位置、餐馆的风格以及用户的居住地时,模型的效果是最好的;

(6) Yelp 数据集中,各模型的召回率要相比于 Dianping 数据集更紧凑.这是由于数据集本身特性的影响,Yelp 数据集中餐馆的数量明显少于 Dianping 数据集.

实验结果表明,多种数据信息能够有效地提高推荐的效果,按照作用大小排序如下:餐馆地理位置>用户居住地>餐馆风格;另外,RRMIF 模型通过融合多种数据信息,使得推荐效果明显要优于只融合一种或者两种数据信息的模型.

3.3.3 稀疏性验证

为了验证 RRMIF 模型在不同稀疏度数据集的表现,该实验从 Dianping 数据集和 Yelp 数据集中分别抽取了 4 种稀疏度不同的数据子集,各数据子集的统计信息见表 4.其中,稀疏度定义为

$$\text{稀疏度}=1-\frac{\text{签到数量}}{\text{用户数量}\times\text{餐馆数量}}.$$

Table 4 Statistics of different sparse subset

表 4 不同稀疏度数据子集的统计信息

数据集	数据子集	用户数量	餐馆数量	签到数量	稀疏度
Dianping	Dianping-1	1 745	3 332	4 124	0.999 29
	Dianping-2	975	3 309	4 059	0.998 75
	Dianping-3	796	3 334	4 167	0.998 43
	Dianping-4	650	3 318	4 050	0.997 78
Yelp	Yelp-1	3 551	981	3 975	0.998 85
	Yelp-2	2 309	914	2 884	0.998 63
	Yelp-3	2 315	1 146	4 759	0.998 2
	Yelp-4	768	1 072	3 935	0.995 22

从表 4 中可以看出,按照稀疏度排序:

Dianping-1>Dianping-2>Dianping-3>Dianping-4,Yelp-1>Yelp-2>Yelp-3>Yelp-4.

RRMIF 模型与第 3.3.1 节中的 5 种对比模型在以上各数据子集的召回率(*recall*@30)比较如图 6 所示,从图 6 中可以看出:

(1) 随着数据稀疏度的降低,PMF 模型的推荐效果逐渐提升.例如,PMF 模型在 Dianping-3 的效果相比于 Dianping-1 提升了 1.5 倍,相比于 Dianping-2 提升了 35%.这是因为 PMF 模型的目的是通过矩阵中的“1”来填充整个矩阵的值,因此极易受数据的稀疏性影响.当数据相对稠密时,也就是说矩阵中“1”的比例相对较多时,模型的效果就会得到提升.另外,PMF 模型在 Dianping-4 的效果相比于 Dianping-3 降低了 2.3%.这说明 PMF 模型随着数据稀疏度的继续增加,PMF 模型的推荐效果会逐渐趋于平稳,并有降低的趋势;

(2) 与 PMF 模型类似,UCF 模型和 WMF 模型随着数据稀疏度的降低,推荐效果都有明显的提升,且最后

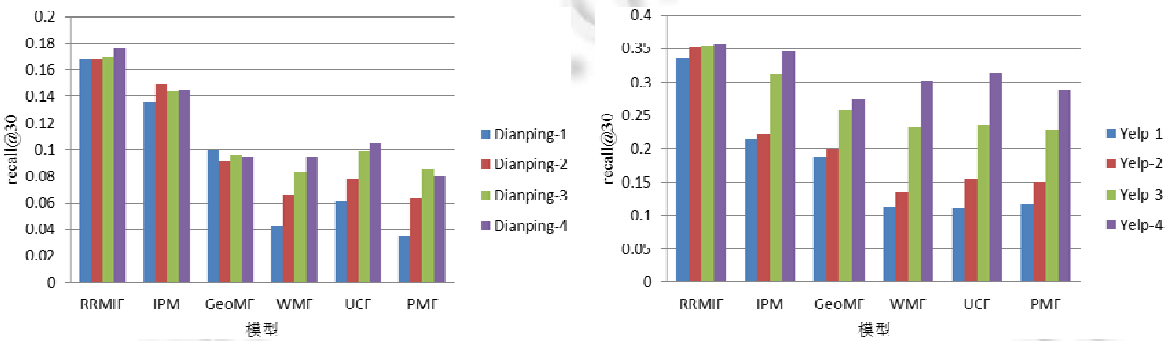
也会趋于平稳.WMF 模型通过引入正负反馈的置信水平来提升推荐效果,但是正负反馈的置信水平仍然受到数据稀疏性的影响.UCF 模型是通过用户的共同评分来计算用户相似度的,因此数据的稀疏性影响着 UCF 模型的推荐效果;

(3) 从 Dianping 的数据子集中可以看出,随着数据稀疏度的变化,GeoMF 模型推荐效果变化不大.这是因为该模型考虑了餐馆的地理位置信息,缓解了矩阵的稀疏性问题;

(4) 在 Dianping 的数据子集中,随着数据集稀疏度的变化,IPM 模型的推荐效果相对稳定;而在 Yelp 数据子集中,IPM 模型的推荐效果还是受到了稀疏性的影响;

(5) 不论在 Dianping 的数据子集还是在 Yelp 的数据子集,RRMIF 模型的推荐效果都是最好的,而且基本不受数据稀疏性的影响.这说明本文提出的模型不仅在有效性上优于其他对比模型,在稳定性上也要优于其他对比模型;

(6) 随着数据稀疏度的降低,其他对比模型的推荐效果有一定的提升,但仍然低于 RRMIF 模型.



(a) 大众点评数据子集 (b) Yelp 数据子集

Fig.6 Sparsity verification

图 6 稀疏性验证

实验结果表明:相比于现有其他模型,RRMIF 模型在数据极其稀疏的情况下仍然能够很好地进行推荐,这说明 RRMIF 模型有效地缓解了数据稀疏性对于推荐效果的影响.

3.3.4 效率评估

在实际应用中,推荐模型的运行时间往往是重要的评价指标之一,如果模型的运行时间超过了可接受范围,那么这个模型的实用性就会受到限制.该实验主要对 RRMIF 模型和第 3.3.1 节中的 5 种对比模型分别在 Dianping 数据集和 Yelp 数据集进行效率的评估.首先,从 Dianping 数据集分别抽取 4 种用户数量不同的数据子集,用户数量分别为 300、600、1200 和 2400,然后得到 RRMIF 模型和其他 5 种对比模型在 4 种数据子集的运行时间.为了全面地进行效率比较,在 Yelp 数据集上按照相同的操作进行.最后,分别得到各模型基于两种数据集的效率评估,如图 7 所示.横轴表示用户数量,纵轴表示为运行时间,单位为秒(s).

从图 7 中可以看出:

(1) 在两种数据集中,PMF 模型、WMF 模型、UCF 模型和 GeoMF 模型的运行时间大致排序为:GeoMF>UCF>WMF>PMF.这是因为 WMF 模型是基于 PMF 模型,考虑了正负反馈的置信水平,增加了一个权重矩阵;而 GeoMF 模型是基于 WMF 模型,考虑餐馆的地理位置信息,增加了特征向量的维数.对于 UCF 模型,需要计算每个用户的相似用户,受用户数量影响明显,当用户数量增加时,会出现运行时间陡增的情况;

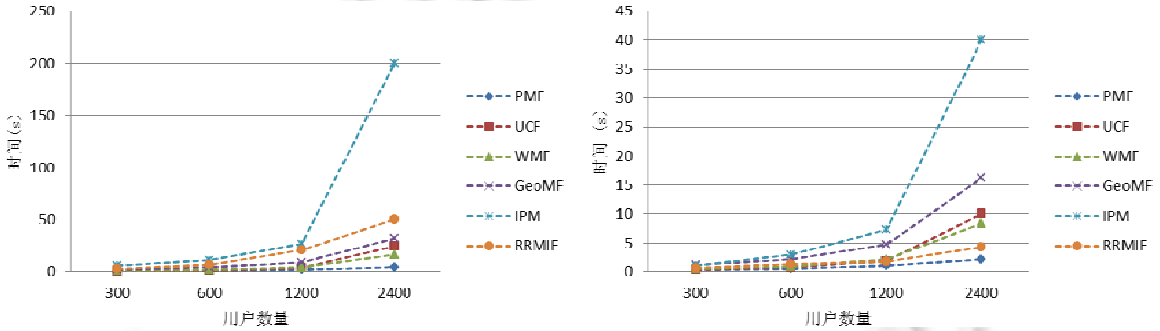
(2) Dianping 数据集中,模型的运行时间相比于 Yelp 数据集显著地增加,这主要有两个原因:第一,因为 Yelp 数据集中用户的签到时间没有具体到一天的某个时刻,因此时间片设置为 1,这样就导致 Yelp 数据集中的模型的运行时间显著地降低;第二,Dianping 数据集的稀疏度要低于 Yelp 数据集,这有可能

导致了运行时间较高;

(3) 无论在 Dianping 数据集还是在 Yelp 数据集,IPM 模型的运行时间都是最多的.IPM 模型的运行时间是最长的,而且随着数据集的增大,运行时间也随之增长,并且增长速率也逐渐增大,几乎呈现指数增长的趋势.这主要有两方面的原因:第一,该模型对多种数据信息预处理的复杂度相对较高;第二,该模型除了要通过概率张量分解进行求解,还需要通过逻辑回归求解;

(4) 基于 Dianping 数据集,RRMIF 模型的运行时间要高于 GeoMF 模型、WMF 模型、UCF 模型和 PMF 模型,这主要是因为 RRMIF 模型考虑了时间上下文信息,需要为用户在某一时间片生成推荐列表,因此计算的时间复杂度增大.而基于 Yelp 数据集,RRMIF 模型的运行时间虽然高于 PMF 模型,但是要低于 GeoMF 模型、WMF 模型和 UCF 模型.这是因为基于 Yelp 数据集的实验,时间片设置为 1,这样,RRMIF 模型的运行时间显著降低.这也说明了在同样不考虑时间上下文的情况下,RRMIF 模型的运行效率是相当可观的;

(5) 从这两种数据集可以看出:RRMIF 模型的运行时间虽然随着数据集的增大而增大,但是没有呈现出急剧增大的趋势.换句话说,该模型的运行时间随着数据集的增大而平稳增大.因此,本文提出的模型在大数据集下仍然是实用的.



(a) 大众点评数据集 (b) Yelp 数据集

Fig.7 Efficiency evaluation of different model

图 7 各模型的效率评估

实验结果表明,RRMIF 模型的运行时间要远低于 IPM 模型,略高于其他模型.但是随着数据量的增大,RRMIF 模型的运行时间几乎呈现稳定增长的趋势.这说明 RRMIF 模型的运行时间虽然不是最少的,但是它的增长趋势仍然是可接受的.

4 总 结

本文首先介绍了餐馆推荐的相关工作,包括兴趣点推荐和餐馆推荐,并分析了现有推荐模型的优缺点.然后,基于大众点评数据集研究了多种数据信息对于用户餐馆选择的影响,并通过 *K*-means 聚类算法将一天 24 小时分为 4 个用餐时间段.另外,基于地理上下文、餐馆属性信息构建了两种餐馆相似关系矩阵,基于用户人口统计信息构建了用户相似关系矩阵.在此基础上,本文提出了一种融合多种数据信息的餐馆推荐模型(RRMIF).该模型首先利用签到信息和时间上下文构建“用户-餐馆-时间片”的三维张量;然后,在概率张量分解模型的基础上同时分解用户相似关系矩阵和餐馆相似关系矩阵;最后,利用 BPR 优化准则和梯度下降算法进行模型求解.为了更加全面地验证 RRMIF 模型的有效性,本文基于大众点评数据集和 Yelp 数据集做了以下 4 个实验:(1) 与现有模型比较;(2) 研究多种数据信息对推荐效果的影响;(3) 稀疏性验证;(4) 效率评估.最后,实验结果表明,相比于目前存在的餐馆推荐模型,RRMIF 模型有着更好的推荐效果和可接受的运行时间,并且缓解了数据稀疏性对推荐效果的影响.

References:

[1] Cho E, Myers SA, Leskovec J. Friendship and mobility: User movement in location-based social networks. In: Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2011. 1082−1090. [doi: 10.1145/2020408.2020579]

[2] Noulas A, Scellato S, Lathia N, Mascolo C. Mining user mobility features for next place prediction in location-based services. In: Proc. of the 12th IEEE Int'l Conf. on Data Mining (ICDM). IEEE, 2012. 1038−1043. [doi: 10.1109/icdm.2012.113]

[3] Gao H, Tang J, Hu X, Liu H. Content-aware point of interest recommendation on location-based social networks. In: Proc. of AAAI. 2015. 1721−1727.

[4] Liu SD, Meng XW. Approach to network services recommendation based on mobile users' location. Ruan Jian Xue Bao/Journal of Software, 2014,25(11):2556−2574 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4561.htm [doi: 10.13328/j.cnki.jos.004561]

[5] Chen T, Zhu Q, Zhou MX, Wang S. Trust-based recommendation algorithm in social network. Ruan Jian Xue Bao/Journal of Software, 2017,28(3):721−731 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5159.htm [doi: 10.13328/j.cnki.jos.005159]

[6] Zhang F, Yuan NJ, Zheng K, Lian D, Xie X, Rui Y. Exploiting dining preference for restaurant recommendation. In: Proc. of the 25th Int'l Conf. on World Wide Web, Inte'l World Wide Web Conf. on Steering Committee. 2016. 725−735. [doi: 10.1145/2872427.2882995]

[7] Sun J, Xiong Y, Zhu Y, Liu J, Guan C, Xiong H. Multi-source information fusion for personalized restaurant recommendation. In: Proc. of the 38th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM Press, 2015. 983−986. [doi: 10.1145/2766462.2767818]

[8] Lian D, Zhao C, Xie X, Sun G, Chen E, Rui Y. GeoMF: Joint geographical modeling and matrix factorization for point-of-interest recommendation. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2014. 831−840. [doi: 10.1145/2623330.2623638]

[9] Liu Y, Zhao P, Liu X, Wu M, Duan L, Li X L. Learning user dependencies for recommendation. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence. 2017. 2379−2385. [doi: 10.24963/ijcai.2017/331]

[10] He J, Li X, Liao L, He J, Li X, Liao L. Category-aware next point-of-interest recommendation via listwise Bayesian personalized ranking. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence. 2017. 1837−1843. [doi: 10.24963/ijcai.2017/255]

[11] Li X, Jiang M, Hong H, Liao L. A time-aware personalized point-of-interest recommendation via high-order tensor factorization. ACM Trans. on Information Systems (TOIS), 2017,35(4):31. [doi: 10.1145/3057283]

[12] Shan H, Banerjee A, Natarajan R. Probabilistic Tensor Factorization for Tensor Completion. 2011.

[13] Krohn-Grimberghe A, Drumond L, Freudenthaler C, Schmidt-Thieme L. Multi-relational matrix factorization using Bayesian personalized ranking for social network data. In: Proc. of the 5th ACM Int'l Conf. on Web Search and Data Mining. ACM Press, 2012. 173−182. [doi: 10.1145/2124295.2124317]

[14] Rendle S, Freudenthaler C, Gantner Z, Schmidt-Thieme L. BPR: Bayesian personalized ranking from implicit feedback. In: Proc. of the 25th Conf. on Uncertainty in Artificial Intelligence. AUAI Press, 2009. 452−461.

[15] Wagstaff K, Cardie C, Rogers S, Schrödl S. Constrained *k*-means clustering with background knowledge. In: Proc. of the ICML. 2001. 577−584.

[16] Yao Z, Fu Y, Liu B, Liu Y, Xiong H. POI recommendation: A temporal matching between POI popularity and user regularity. In: Proc. of the 16th IEEE Int'l Conf. on Data Mining (ICDM). IEEE, 2016. 549−558. [doi: 10.1109/icdm.2016.0066]

[17] Salakhutdinov R, Mnih A. Probabilistic matrix factorization. Advances in Neural Information Processing Systems, 2008,20(1):1257−1264.

[18] 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). IEEE, 2008. 263−272. [doi: 10.1109/icdm.2008.22]

[19] Li X, Xu G, Chen E, Li L. Mars: A multi-aspect recommender system for point-of-interest. In: Proc. of the 31st IEEE Int'l Conf. on Data Engineering (ICDE). IEEE, 2015. 1436−1439. [doi: 10.1109/icde.2015.7113395]

[20] Luan W, Liu G, Jiang C. Collaborative tensor factorization and its application in POI recommendation. In: Proc. of the 13th IEEE Int'l Conf. on Networking, Sensing, and Control (ICNSC). IEEE, 2016. 1–6. [doi: 10.1109/icnsc.2016.7478984]

[21] Fu Y, Liu B, Ge Y, Yao Z, Xiong H. User preference learning with multiple information fusion for restaurant recommendation. In: Proc. of the 2014 SIAM Int'l Conf. on Data Mining. SIAM, 2014. 470–478. [doi: 10.1137/1.9781611973440.54]

[22] Yuan Q, Cong G, Ma Z, Sun A, Thalmann NM. Time-aware point-of-interest recommendation. In: Proc. of the 36th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM Press, 2013. 363–372. [doi: 10.1145/2484028.2484030]

[23] Ference G, Ye M, Lee WC. Location recommendation for out-of-town users in location-based social networks. In: Proc. of the 22nd ACM Int'l Conf. on Information & Knowledge Management. ACM Press, 2013. 721–726. [doi: 10.1145/2505515.2505637]

[24] Ma H, Yang H, Lyu MR, King I. Sorec: Social recommendation using probabilistic matrix factorization. In: Proc. of the 17th ACM Conf. on Information and Knowledge Management. ACM Press, 2008. 931–940. [doi: 10.1145/1458082.1458205]

[25] Zhang T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: Proc. of the 21st Int'l Conf. on Machine Learning. ACM Press, 2004. 116. [doi: 10.1145/1015330.1015332]

[26] Kakade SM, Shalev-Shwartz S, Tewari A. Regularization techniques for learning with matrices. Journal of Machine Learning Research, 2012,13(1):1865–1890.

[27] Ye M, Yin P, Lee WC, Lee DL. Exploiting geographical influence for collaborative point-of-interest recommendation. In: Proc. of the 34th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM Press, 2011. 325–334. [doi: 10.1145/2009916.2009962]

附中文参考文献:

[4] 刘树栋,孟祥武.一种基于移动用户位置的网络服务推荐方法.软件学报,2014,25(11):2556–2574. http://www.jos.org.cn/1000-9825/4561.htm [doi: 10.13328/j.cnki.jos.004561]

[5] 陈婷,朱青,周梦溪,王珊.社交网络环境下基于信任的推荐算法.软件学报,2017,28(3):721−731. http://www.jos.org.cn/1000-9825/5159.htm [doi: 10.13328/j.cnki.jos.005159]

戴琳(1994－),男,安徽安庆人,硕士生,主要研究领域为推荐系统,数据挖掘,机器学习.

张玉洁(1969－),女,副教授,主要研究领域为网络服务,用户需求,推荐服务.

孟祥武(1966－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为网络服务,用户需求,推荐服务.

纪威宇(1987－),男,博士生,CCF 学生会员,主要研究领域为数据挖掘,机器学习,推荐系统.