Journal of Computer Applications
计算机应用, 2022, 42(6): 1898 - 1913
ISSN 1001-9081
CODEN JYIIDU
2022-06-10
http://www.joca.cn

文章编号:1001-9081(2022)06-1898-16
DOI:10.11772/j.issn.1001-9081.2021040607

推荐系统综述

于　蒙,何文涛,周绪川*,崔梦天,吴克奇,周文杰
(计算机系统国家民委重点实验室(西南民族大学),成都 610041)
(*通信作者电子邮箱xczhou@swun.edu.cn)

摘　要:随着网络应用的不断发展,网络资源呈指数型增长,信息过载现象日益严重,如何高效获取符合需求的资源成为困扰人们的问题之一。推荐系统能对海量信息进行有效过滤,为用户推荐符合其需求的资源。对推荐系统的研究现状进行详细介绍,包括基于内容的推荐、协同过滤推荐和混合推荐这三种传统推荐方式,并重点分析了基于卷积神经网络(CNN)、深度神经网络(DNN)、循环神经网络(RNN)和图神经网络(GNN)这四种常见的深度学习推荐模型的研究进展;归纳整理了推荐领域常用的数据集,同时分析对比了传统推荐算法和基于深度学习的推荐算法的差异。最后,总结了实际应用中具有代表性的推荐模型,讨论了推荐系统面临的挑战和未来的研究方向。

关键词:推荐算法;协同过滤;深度学习;卷积神经网络;深度神经网络;循环神经网络;图神经网络
中图分类号:TP391　　**文献标志码:**A

Review of recommendation system

YU Meng, HE Wentao, ZHOU Xuchuan*, CUI Mengtian, WU Keqi, ZHOU Wenjie
(*The Key Laboratory for Computer Systems of State Ethnic Affairs Commission* (*Southwest Minzu University*), *Chengdu Sichuan* 610041, *China*)

Abstract: With the continuous development of network applications, network resources are growing exponentially and information overload is becoming increasingly serious, so how to efficiently obtain the resources that meet the user needs has become one of the problems that bothering people. Recommendation system can effectively filter mass information and recommend the resources that meet the users needs. The research status of the recommendation system was introduced in detail, including three traditional recommendation methods of content-based recommendation, collaborative filtering recommendation and hybrid recommendation, and the research progress of four common deep learning recommendation models based on Convolutional Neural Network (CNN), Deep Neural Network (DNN), Recurrent Neural Network (RNN) and Graph Neural Network (GNN) were analyzed in focus. The commonly used datasets in recommendation field were summarized, and the differences between the traditional recommendation algorithms and the deep learning-based recommendation algorithms were analyzed and compared. Finally, the representative recommendation models in practical applications were summarized, and the challenges and the future research directions of recommendation system were discussed.

Key words: recommendation algorithm; collaborative filtering; deep learning; Convolutional Neural Network (CNN); Deep Neural Network (DNN); Recurrent Neural Network (RNN); Graph Neural Network (GNN)

0　引言

近年来,网络应用尤其是移动应用的快速发展,使得人们能够方便地浏览大量的网络信息资源,如何为用户从海量的信息资源中推荐符合其需求的资源(如商品、电影、书籍等)成了目前研究者们关注的问题之一。推荐系统(Recommendation System, RS)[1]可以有效地对信息进行过滤和筛选,帮助用户以个性化的方式来检索符合其需求的信息资源,缓解信息过载(Information Overload)[2]的问题。推荐技术经过不断的发展和更新,已经在教育、音乐、电子商务、社交网络等领域广泛应用。协同过滤算法被提出后,推荐系统逐渐成为一个新的研究热点,同时也面临着数据稀疏问题(用户对推荐项目的评分数量太少)和冷启动问题(新的推荐项目和新用户无评分数据)。深度学习(Deep Learning, DL)是具备识别、分析、计算的机器学习算法,为缓解数据稀疏和冷启动问题带来了新的机遇,2015年以来,深度学习已经在

收稿日期:2021-04-19;**修回日期:**2021-07-14;**录用日期:**2021-07-20。　**基金项目:**国家自然科学基金资助项目(12050410248);四川省科技计划项目(2021YFH0120);西南民族大学研究生创新型科研项目(CX2020SZ04)。

作者简介:于蒙(1995—),女,宁夏固原人,硕士研究生,CCF会员,主要研究方向:推荐系统、信息过滤、数据挖掘;　何文涛(1996—),男,湖南永州人,硕士研究生,主要研究方向:深度学习、数据挖掘;　周绪川(1972—),男,重庆人,教授,博士,CCF会员,主要研究方向:数据挖掘、深度学习;　崔梦天(1972—),女,内蒙古乌兰浩特人,教授,博士,主要研究方向:智能信息处理;　吴克奇(1997—),男,湖北孝感人,硕士研究生,主要研究方向:推荐系统;　周文杰(1997—),男,四川广安人,硕士研究生,主要研究方向:数据挖掘。

语义挖掘、人脸识别、语音识别等领域广泛应用,深度学习模型的逐渐成熟也为推荐系统的发展带来了新的机遇。2016年的ACM推荐系统年会上,Song等[3]指出将深度学习和推荐系统融合作为推荐系统未来研究的重点,由此,国内外的学者和研究机构针对这一问题开展了大量的研究。2017年以来,机器学习方向的顶级会议(如:ICML、NIPS、COLT等)中有关深度学习的个性化推荐文章逐年增加。2019年,文献[4]的研究认为深度学习能够从数据中自动学习特征的不同层次表达和抽象,是解决传统推荐技术出现的冷启动、数据稀疏等问题的有效策略。

1　传统的推荐算法

推荐系统是数据挖掘、预测算法[5]、机器学习等多种学科结合而成的一个新的研究领域。文献[6]中最早对推荐系统给出定义,指出在日常生活中无论是了解的事件还是未知的事件,时刻需要人们做出决策,面对熟悉的事情,人们常常可以依赖过去的经验做出合理的决策,然而,在面对未知的事情时,人们则需要他人的口头建议、书评、影评、推荐等来进行判断,文献中认为推荐系统的意义是能够为推荐项目和用户建立适当的匹配关系。文献[7]中则认为推荐系统是为不同用户从大量的项目中匹配符合其兴趣偏好但是未被用户观察到的项目,它认为推荐系统正在成为一个具有重大经济影响的重要业务。

推荐系统从本质上来说是对人的某种行为的模拟,它通过推荐算法对特定的数据信息进行分析处理,然后将处理后的结果推荐给有相关需求的用户[8]。推荐算法是推荐系统的核心,它能根据用户的历史购买需求、行为记录或者相似偏好进行建模,从而发现符合用户偏好的需求,并将之推荐给用户。推荐系统的形式化定义[9-10]如下:

定义1　推荐系统。设P表示所有用户的集合,C为用户可推荐的对象的集合。实际问题中,P、C都是规模非常大的集合。函数f表示用户p对c的喜爱度,即$f: P \times C \rightarrow R$,其中$R$表示非负实数的有限序列,将让函数$f$取得最大值的推荐对象$c' \in C$推荐给用户。如式(1)所示:

$$\forall p \in P, c_p' = \arg\max_{c \in C} f(p, c) \tag{1}$$

c_p'表示最符合用户p偏好的推荐对象。因此,在为用户选择最感兴趣的对象之前,推荐系统必须利用已知的用户认可度去完成未知的推荐对象认可度的预测,这是推荐系统外推的过程。近年来,从不同的角度给推荐技术分类,不同的学者赋予推荐系统不同的内涵,目前传统的推荐系统分为三类[11]:基于内容过滤的推荐(Content-Based recommendation, CB)[12]、基于协同过滤的推荐(Collaborative Filtering recommendation, CF)[13]和混合推荐(Hybrid Recommendation)[14],如图1所示。

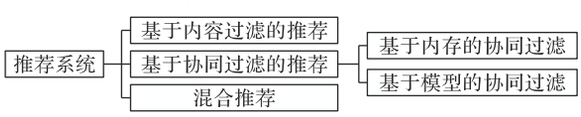

图1　传统推荐系统分类

Fig. 1　Classification of traditional recommendation systems

1.1　基于内容过滤的推荐技术

推荐系统最早被应用在电子商务网站,它通常根据用户的购买行为记录或购买评价来向用户推荐与其需求偏好相似的物品[15]。文献[16]中提出了一种基于上下文内容的方法来匹配和排序服务,认为上下文是用来描述一个给定文本的相关的语言术语集。该方法通过解析服务的底层文档提取作为文本术语的令牌,并使用字符串匹配函数来匹配这些令牌的本体。文献[17]中提出了一种匹配用户查询和服务描述以及相关上下文信息的服务发现方法。该方法将上下文提供者提供的上下文信息、服务提供者提供的服务描述和用户提供的服务请求三者用本体建模,然后将这三条信息逐个匹配。文献[18]中提出了一个Web服务上下文分类,然后使用本体来定义这个分类。上下文由一个两级机制建模,该机制涵盖了上下文规范和服务策略,提供了一个对等体系结构来完全匹配Web服务上下文策略,源服务的每个上下文都由候选服务的策略匹配。

总之,基于内容过滤的推荐(CB)技术的核心思想是:以用户历史的选择记录或偏好记录作为参考推荐,挖掘其他未知的记录中与参考推荐关联性高的项目作为系统推荐的内容。通过用户的显式反馈(如评价、认可度、喜欢\不喜欢)和隐式反馈(如浏览时间、点击次数、搜索次数、停留时间等)获取用户在某段时间内的交互记录,然后学习这些记录中用户的偏好并将其标记为特征;接着计算用户偏好与待测推荐对象在内容上的相似度(或匹配度);最后将待测推荐对象与用户偏好的相似度进行排序,从而为用户选择出符合其兴趣偏好的推荐对象。计算相似度是一个关键部分,会直接影响推荐的策略。计算相似度的方式有多种,常用式(2)计算相似度[19]:

$$u(p, c) = score(userprofile, content) \tag{2}$$

其中:p表示用户,c表示推荐内容,*userprofile*表示p偏好的内容,*content*表示系统为用户推荐的内容。*score*用来计算用户偏好和推荐内容的相似值,最终用效用函数$u()$来定义,根据u的值来排序,数值越大排序越靠前。

*score*有多种计算方式,通常使用向量夹角余弦的距离计算方式:

$$u(p, c) = cos(\boldsymbol{w}_p, \boldsymbol{w}_c) = \frac{\sum_{i=1}^{k} w_{i,p} w_{i,c}}{\sqrt{\sum_{i=1}^{k} w_{i,p}^2} \sqrt{\sum_{i=1}^{k} w_{i,c}^2}} \tag{3}$$

其中:$\boldsymbol{w}_p$表示*userprofile*的特征向量;$\boldsymbol{w}_c$表示*content*的关键词向量权重。

对计算得的u值进行排序,u值越大,说明推荐的对象越符合用户的喜好。例如为用户推荐电影时,系统会学习用户的历史观看记录并分析,然后找到这些电影的共性,预测出用户感兴趣的电影类型,然后从海量的电影清单中选择出与用户偏好相似的电影。用户偏好记录的特征标记和推荐内容是CB的关键,用户评价对基于内容的推荐系统影响较小[20]。

CB系统框架如图2所示,包含数据挖掘处理部分和自适

应推荐部分，对用户来说这两部分都是隐藏的。数据挖掘部分主要是通过建立向量空间模型对用户的偏好特征进行分析和提取；自适应推荐部分的主要作用就是将用户偏好的相似度排序，自动生成推荐列表，将推荐列表通过Web服务器推荐给用户。

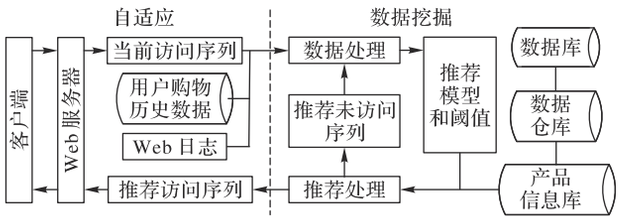

图2 CB系统框架

Fig. 2 Framework of CB

1.2 协同过滤推荐

协同过滤推荐(CF)算法的核心是通过分析评分矩阵(通常是用户对项目的评分)来得到用户、项目之间的依赖关系，并进一步预测新用户与项目之间的关联关系。CF算法是最早被研究和讨论的推荐技术之一，它有效地推动了个性化推荐的发展。1992年，文献[21]中利用传统的协同过滤技术解决了垃圾邮件分类问题；亚马逊(Amazon)是目前较大的网络购物平台之一，主要利用CF算法为用户推荐商品；Netflix在其主页上也使用CF算法为用户推荐喜爱的电视节目。

如今协同过滤技术被广泛应用在音乐推荐、电影推荐、电子商务等领域[22]，CF主要分为基于内存(Memory-Based)的推荐和基于模型(Model-Based)的推荐。

1.2.1 基于内存的推荐

基于内存的协同过滤推荐通过用户-项(User-Item)的评价矩阵寻找相似用户和相似项目[23-24]之间的相似度，进而为新用户构建相似度矩阵，预测用户感兴趣的项目。通过寻找相似项目进行的推荐称为基于项目的推荐；通过寻找相似用户进行的推荐称为基于用户的推荐。

基于项目的协同过滤技术主要挖掘并分析的是不同推荐项目间隐藏的关系而不是用户之间的关系[25]，项目间的相似性计算是该技术的关键[26]，其推荐过程如图3。该过程可以理解为：若有2个不同用户A、B，且他们都对物品1、3表示出较高的喜爱，那么我们可以认为1、3物品存在某种相似。当系统中出现的新用户C并选择了物品1时，那么系统便会自动将与物品1相似度高的物品3推荐给他。

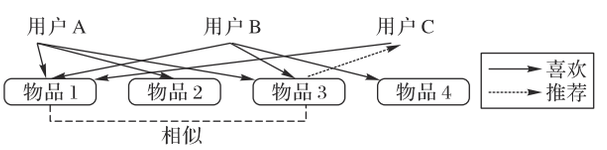

图3 基于项目的协同过滤推荐

Fig. 3 Item-based collaborative filtering recommendation

基于用户的推荐过程如图4所示，经过评价矩阵计算，认为用户A与B相似，在物品选择时，若用户A选择了物品1、2、3，用户B选择了物品1、3，那么在物品推荐时可以认为用户B的选择和用户A相似，因此推荐系统可以将物品2推荐给用户B。

文献[27]中通过分析用户矩阵来确定这些用户与用户以及不同用户与其感兴趣的项目之间的差异，从而根据差异有针对性地为用户推荐合适的项目。然而基于用户的推荐过程并不能依赖相似的用户都了解对方，于是，文献[28]中提出了一种基于匿名合作的协同过滤算法，专门用于解决为不同用户推荐新闻和电影的问题。基于用户的协同过滤算法虽然能够发现用户隐藏的兴趣点和偏好，但该技术存在严重的冷启动问题。在实际问题中，推荐系统中的用户种类不是一成不变的，当有新的用户类型出现时，系统中缺少该类用户的偏好记录，那么推荐系统就无法对该类用户提供符合其需求的推荐。为了解决协同过滤所面临的冷启动问题，文献[29]中将传统的协同过滤算法和神经网络算法相结合。神经网络算法是深度学习算法中的一种，能够分析并计算用户与项目之间的复杂的非线性关系，效率较高。文献[29]中的混合模型关注到了推荐对象的典型性和多样性，在韩国国民健康营养调查数据的应用中经过评估，结果表明它确实能提高推荐效果。

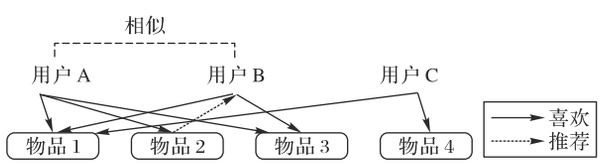

图4 基于用户的协同过滤推荐

Fig. 4 User-based collaborative filtering recommendation

1.2.2 基于模型的推荐

基于模型的推荐算法是通过训练数学模型来预测用户对未交互项目的评分情况，通常包括概率矩阵分解(Probabilistic Matrix Factorization, PMF)[30]和奇异值分解(Singular Value Decomposition, SVD)[31]。PMF和SVD的主要思路是先对用户与项目的历史交互数据记录建立适当的模型，然后产生符合用户需求的推荐列表，其中应用较为广泛的是基于矩阵分解的推荐。

PMF模型一般认为用户和推荐项目的交互行为仅仅由几个潜在的影响其兴趣偏好的因素决定，将高阶评分矩阵 $\boldsymbol{R}_{n\times m}$ 分解为两个低维的矩阵 $\boldsymbol{E}$、$\boldsymbol{Q}$，如式(4)所示：

$$\boldsymbol{R}\approx\boldsymbol{E}^{\mathrm{T}}\boldsymbol{Q} \tag{4}$$

其中：$\boldsymbol{E}=\{\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n\}$ 表示低维用户特征矩阵，$\boldsymbol{e}_i$ 表示用户 i 的 k 维特征向量；$\boldsymbol{Q}=\{\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_n\}$ 代表低维的推荐项目特征矩阵。

在实际推荐问题中，为了降低预测评分和实际评分之间的差值，得到更准确的推荐列表，一般将预测评分与实际评分之间误差的平方作为损失函数，如式(5)所示。

$$f(\boldsymbol{R},\boldsymbol{U},\boldsymbol{V})=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\boldsymbol{I}_{i,j}\left(\boldsymbol{R}_{ij}-\boldsymbol{E}_i^{\mathrm{T}}\boldsymbol{Q}_j\right)^2+\frac{\gamma_p}{2}\|\boldsymbol{E}\|_{\mathrm{F}}^2+\frac{\gamma_q}{2}\|\boldsymbol{Q}\|_{\mathrm{F}}^2 \tag{5}$$

其中：$\boldsymbol{I}_{ij}$ 表示一个示性函数，当 $\boldsymbol{I}_{ij}=1$ 时，代表用户 $\boldsymbol{u}_i$ 对推荐项目 $\boldsymbol{S}_i$ 已经评分了，否则就表示用户没有对推荐项目进行评分；γ_p、γ_q 代表着惩罚因子，是为了防止出现过拟合现象添加的正则化项，γ_p、γ_q 的值决定正则化程度，其值越大表示正则

化的程度越大;$\|\boldsymbol{E}\|_F$和$\|\boldsymbol{Q}\|_F$代表着矩阵范数,一般利用随机梯度下降法对目标函数进行优化处理,对原高阶评分矩阵$\boldsymbol{R}_{n\times m}$的缺失值进行预测。

2016年,文献[32]中利用用户对推荐项目的评分差异建模,建立了成对概率矩阵分解(Pairwise Probabilistic Matrix Factorization, PPMF)模型。该模型能够自动学习用户对交互过的项目的偏好程度,有效地降低了倒序排名的平均值,而不是降低协同过滤推荐的预测评分和实际评分之间的差值,解决了传统推荐问题中数据稀疏的问题。2020年,文献[33]中在PPMF模型的基础上,通过基于图的方式计算了概率矩阵的先验分布,提高了PPMF模型的推荐准确率。

由于社会环境的复杂化,系统为用户生成的推荐常常不能满足用户变化的兴趣爱好,大部分用户经常会接受来自身边朋友的推荐,两个用户之间的相似性常常受到商品的流行程度、用户偏好、社会关系等诸多因素的影响,因此文献[34]中提出了一种结合用户社交网络的矩阵分解模型,如图5所示。该模型首先为用户设计了一种基于粒子群优化(Particle Swarm Optimization, PSO)算法的K均值聚类算法(K-Harmonic Means, KHM)对用户进行聚类,然后将用户社交网络中用户的社交关系引入到相似计算模型中,利用矩阵分解技术计算出用户偏好的项目。这种办法有效地缓解了推荐系统中常遇到的冷启动和数据稀疏的问题;然而,当训练数据过大时,训练的复杂度也随之提高。

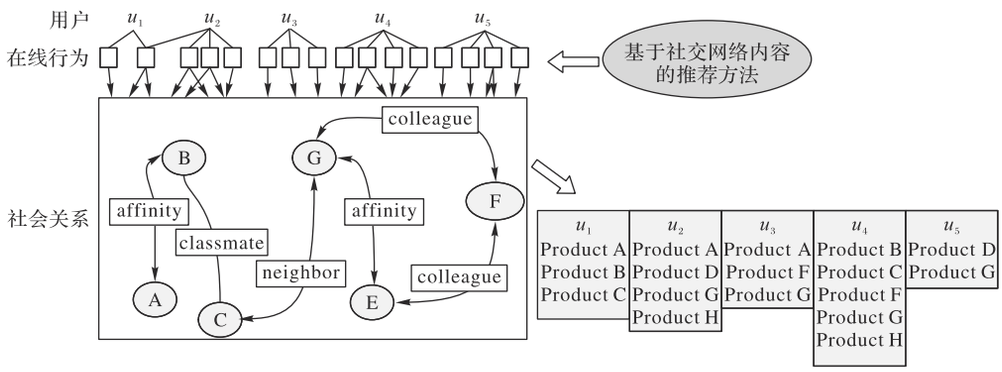

图5　基于社交网络的推荐系统框架

Fig. 5　Recommendation system framework based on social network

传统的矩阵分解算法中SVD的应用也较为广泛,与PMF不同的是,SVD是将用户-项目的评分矩阵通过降维、分解、计算成3个低阶矩阵乘积,对这3个低阶矩阵进行训练最后还原回初始的矩阵。2006年,文献[35]在优化SVD模型的基础上提出了FunkSVD模型,如图6所示。该模型首先把评分矩阵$\boldsymbol{R}_{m\times n}$分解成两个低阶的用户矩阵$\boldsymbol{U}$和推荐项目矩阵$\boldsymbol{V}$,用户和推荐项目都映射到一个$K$维空间,这$K$维空间对应着$K$个隐因子,用户对推荐项目的评分受这$K$个隐因子的影响。其优化函数如式(6)所示:

$$\min_{U^*,V^*}\sum\left(\boldsymbol{R}_{ui}-\boldsymbol{U}_u^{\mathrm{T}}\boldsymbol{V}_i\right)^2+\lambda\left(\boldsymbol{E}\boldsymbol{V}_i\boldsymbol{E}^2+\boldsymbol{E}\boldsymbol{U}_u\boldsymbol{E}^2\right)\tag{6}$$

其中:目标用户u的特征向量为$\boldsymbol{U}_u$;第i个推荐项目的特征向量用$\boldsymbol{V}_i$表示;矩阵中隐变量的数量用K表示;项目的预测评分和真实评分之间的误差用式(6)中的第一项来计算;式(6)中的第二项是为了避免过拟合而设置的正则项。

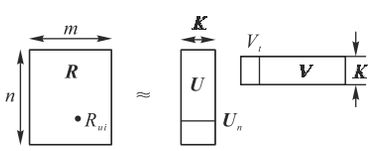

图6　FunkSVD模型

Fig. 6　FunkSVD model

为了解决用户本身特质对用户-项目评分矩阵的影响,文献[36]的研究以SVD模型为基础,提出了一种带偏置项的奇异值分解(Bias Singular Value Decomposition, BiasSVD)模型。2010年,文献[37-38]对BiasSVD模型进行了改进,提出了融合用户-项目的隐式反馈信息的SVD++模型,该模型解决了因只有显式反馈信息而缺少隐式反馈信息的推荐冷启动问题。之后,文献[37-38]的作者又将时间因子作为辅助信息融合到了SVD++中,提出了timeSVD++模型。该模型提高了用户近期隐示反馈行为的权重,而对用户的早期的反馈信息的权重进行了衰减,近似地实现了动态的推荐目的。

1.3　混合推荐

基于内容的推荐技术在处理规模较大的信息内容时,常常因为耗时久而造成信息时效性降低;协同过滤技术在面对新项目时容易遇到冷启动问题;而混合推荐技术是保留不同推荐技术优点而避免其缺点的一种推荐方式,将不同的算法融入到推荐系统中即混合推荐[39-40]。目前的混合推荐主要分为前融合、后融合、中融合。

1)前融合:指将多个推荐算法融合到一个模型中,如在商品推荐过程中,根据用户历史购买记录将其感兴趣的商品特征提取出来作为推荐模型的输入,由混合模型中的推荐算法通过自适应学习产生推荐结果。该混合推荐技术从本质上来说是数据库中所有不同用户特征的融合。如文献[41]中将层次聚类算法和集成相似度算法结合,构建了一种准确度和多样性相结合的混合推荐模型,在对推荐效果影响较小的情况下,通过调整混合模型的权重因子,可以达到推荐多样性且准确的目的。

2)中融合:该混合推荐技术一般先以某种推荐算法为参照,再将推荐效果与混合其他推荐算法的技术对比。如以基于内容的推荐为主框架,然后在该款框架中混合协同过滤推荐能够有效解决冷启动问题。从混合本质上来说,该融合是对不同模型的融合。如文献[42]的研究以深度学习算法作

为框架，将深度学习与改进的机器学习模型相结合，从多个角度学习项目和用户之间的交互，提出了一种称为深度度量因子分解学习（Deep Metric Factorization Learning，DMFL）的混合推荐模型。该混合推荐模型的泛化能力较好，能全面地反映用户的偏好。文献[43]中提出了一种基于潜在因子模型（Latent Factor Model，LFM）和基于图的个人排名（Personal Rank，PR）算法相结合的混合推荐算法，与单独使用PR算法相比，该混合模型的准确率和正确率表现更优。

3）后融合：这种方法对推荐结果十分看重，主要通过比较不同推荐算法的推荐效果从而得到可靠性较高的推荐对象序列，最后将这个序列推荐给用户。

在实际问题中，将不同的推荐算法相互结合从而得到效果更好的推荐是混合推荐技术的优势之一。目前Amazon、Google[44]、微软[45]等公司通过使用混合推荐技术在商品、广告、新闻等个性化推荐方面取得了巨大的成功。

以上推荐技术都属于传统的推荐技术，近年来，用户历史偏好记录的生成内容（如特征标签、位置、交友记录、评论记录）越来越多样化，传统的推荐技术已经无法满足用户的多样需求，因此产生了大量新的推荐算法，如：用户在社交网络中分享或者获取各种资源时，只希望将自己的兴趣或喜好公开给相似的用户，并不希望将个人信息等隐私信息公开。保护用户隐私的推荐逐渐成为学者们关注的问题[46-47]。文献[48]中提出了一种基于用户行为来保护用户好友隐私的算法，将该算法用于集中管理和分布管理相结合的混合社交网络中，能够让用户在实现兴趣偏好共享的同时又不暴露用户的隐私信息。

表1对上述三种不同的传统推荐技术的优缺点进行了总结和对比。

表1 传统推荐技术优缺点对比

Tab. 1 Comparison of advantages and disadvantages of traditional recommendation techniques

推荐技术	优点	缺点
CB	1. 解决冷启动问题 2. 可解释性强 3. 易实现	1. 缺少特征提取的方法 2. 易忽略推荐对象的典型性 3. 安全性差
CF	1. 适合小规模推荐 2. 简单易操作 3. 易建模	1. 存在冷启动问题 2. 无法处理运算复杂的推荐 3. 缺乏可解释性
混合推荐	1. 克服了数据稀疏 2. 弥补不同技术缺点 3. 适合用户多的推荐	1. 缺少高效的混合模式 2. 难以建立数学模型 3. 推荐过程较复杂

2 基于深度学习的推荐技术

深度学习算法强大之处在于能够像人类一样学习并处理复杂问题，面对规模复杂的数据能从多种维度来分析并计算线性或者非线性的特征序列，能从海量的数据中自动地学习符合用户需求的特征，已经成功地应用在图像识别、语音识别、自然语言处理等领域并取得了良好的效果，因此越来越多的研究者也尝试将深度学习应用在推荐系统中，如何把深度学习技术与推荐技术有效结合并深入研究已经成为了一个新的研究方向。深度学习技术除了能够发现用户行为记录隐藏的潜在特征表示，还能捕获用户与用户、用户与项目、项目与项目之间的非线性关系的交互特征，为系统的性能（如召回率、精度等）提高带来了更多机会，能够克服传统推荐技术中遇到的一些障碍，从而实现更精确的推荐。

2.1 基于深度神经网络的推荐

深度神经网络（Deep Neural Network，DNN）是深度学习模型中的一种[49-51]，也可以叫作多层神经网络或多层感知机（Multi-Layer Perceptron，MLP）。目前，在个性化推荐问题中引入深度神经网络技术的趋势越来越明显[52-57]。

文献[49]中首次将深度神经网络模型融入到视频推荐领域，并在YouTube视频网站进行了仿真实验，推荐流程如图7所示。

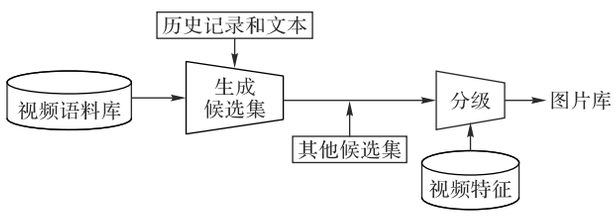

图7 YouTube视频推荐过程

Fig. 7 YouTube video recommendation process

YouTube视频网站的特点是注册用户多、视频更新速度快、视频时长不一、数量多，传统的推荐算法很难为用户推荐符合其偏好的视频内容。图7的推荐过程分为候选集生成和视频排序两个阶段。候选集生成阶段可以视为一个视频筛选的过程，即根据用户的观看记录从已有的视频中选择和用户观看历史记录相似的视频集合作为下次推荐的候选视频。候选集生成阶段将视频推荐问题视为一个多分类问题，利用深度神经网络对用户和视频建模，通过预测函数P来计算在C情况下，用户U在t时刻观看视频类型i的概率，i是所有视频集合V中的某一类。分类预测公式如下所示：

$$P\left(\boldsymbol{w}_t = i \mid U, C\right) = \mathrm{e}^{v_i u} \Big/ \sum_{j \in V} \mathrm{e}^{v_j u} \tag{7}$$

排序阶段则是从不同特征维度对视频进行分析，通过加权的逻辑回归输出层获得用户点击某类视频的概率预测。预测值与用户感兴趣的视频类型越相似，其得分就越高，最终选取得分最高的几十个视频作为推荐结果。仿真结果显示，文献[49]提出的推荐模型的召回速率和效率较高，能对百万级规模的视频数据集进行训练。

但该模型仍存在以下不足：1）面对海量的视频数据，该模型只对数据进行了简单的清洗，在后续研究中可以尝试引入注意力机制，从而对视频的权重进行分配，对用户关注较多的视频赋予更高权重，对用户关注较少的视频赋予较低权重；2）视频网站往往存在恶意视频（如广告等），在后续视频推荐研究中，可以尝试建立一种安全机制先对恶意视频进行拦截，从而更精确地捕获到用户的潜在偏好，不仅能提高用户的使用率，还能提高推荐的效果。

文献[50]中提出了一个深广（Wide & Deep）模型来解决大规模的在线推荐问题，该模型是由单层的Wide部分和多层的Deep部分相结合的一个模型，如图8所示。

Wide 部分是式(8)的广义线性模型,y 是模型的预测值,$\boldsymbol{x}$ 为特征向量,w 是模型的参数,b 为预测的偏差值。这部分的作用是让推荐模型具有较强的记忆能力。

$$y = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b \tag{8}$$

Deep 部分是深度神经网络,该部分模型对嵌入的向量进行抽象和初始化。接着,将抽象好的特征向量传递到隐藏层,每个隐藏层执行以下计算:

$$\boldsymbol{a}^{(l+1)} = f(\boldsymbol{w}^{(l)}\boldsymbol{a}^{(l)} + \boldsymbol{b}^{(l)}) \tag{9}$$

其中:l 是训练的层数;f 是激活函数,通常使用 ReLU 函数;$\boldsymbol{a}^{(l)}$、$\boldsymbol{b}^{(l)}$、$\boldsymbol{w}^{(l)}$ 分别是第 l 层的激活、偏置和模型权重矩阵。

Deep 部分的作用是让模型有更好的泛化能力,Wide 和 Deep 结合使得该模型不仅能够快速学习并处理大量的特征属性,还具有强大的表达能力。在 Google play(一个拥有超过 10 亿活跃用户和超过 100 万个应用程序的移动应用商店)上的实验结果表明,该模型明显增加了 APP 的下载量,达到了更精确的推荐目的。

Wide & Deep 模型主要利用 Wide 部分学习目标用户的特征,利用 Deep 部分来泛化相似的推荐项目,能对 5 千亿个样本进行训练,有效缓解数据稀疏问题,而且还可用于分类、回归、查找等问题;它的不足是需要人为的特征工程。

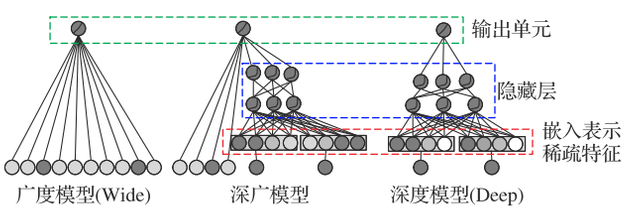

图 8　Wide & Deep 模型

Fig. 8　Wide & Deep model

文献[51]中提出了一种融合 DNN 和矩阵分解的推荐模型,能快速地为用户项建立其生成交互函数时所需要的非线性模型。相比单一的矩阵分解算法,该模型进一步提高了评分预测准确性结果,提高了推荐性能;但该模型并没有从多种维度提取用户的偏好,泛化能力较差。为了解决这个问题,文献[52]中提出了一种基于 DNN 的深度混合推荐模型。该模型将用户和项目信息输入到了改进的机器学习模型中进行训练,从多维度更深入地学习用户和推荐项目的交互关系。该模型对用户和项目的特征学习部分由 2 个并行的 DNN 组成,目的是为了提取静态项目的潜在特征和动态用户的潜在特征。这种将多个深度学习模型和机器学习模型相互融合的推荐模型,能较为准确地预测用户的偏好情况,提高推荐的性能。未来的研究中,可以尝试在推荐模型中融合多个深度学习模型和机器学习模型,提高推荐的泛化能力。

2.2　基于循环神经网络的推荐

循环神经网络(Recurrent Neural Network, RNN)[53]包括双向循环神经网络和长短期记忆(Long Short Term Memory, LSTM)网络。在深度神经网络中,模型训练好之后在输入层给定一个 $\boldsymbol{x}$,在输出层就能得到特定的 $\boldsymbol{y}$,但只适合于前后输入完全没有关系的序列。在推荐方面,通常使用 LSTM 和门控循环单元(Gated Recurrent Unit, GRU)处理推荐问题中的长序列信息。LSTM 和 GRU 属于 RNN 的改进版本,它们的关键是可以捕捉到序列比较长的 n 元信息序列,最大优势是能够为前后有关联的序列信息建模,已经在新闻推荐[54]、文字翻译[55]、语音识别[56]等领域得到了广泛的应用。

LSTM 模型最早由文献[57]提出,它可以学习较长序列信息之间的交互关系。从此,该模型不断地被研究者们改进和优化。文献[54]中认为传统的协同过滤技术无法为用户提供动态的个性化推荐,因此,将用户的行为记录抽象成有关联的数据序列,使用降噪自编码器(Auto-Encoder, AE)构建的深层网络来学习新闻文本的特征;RNN 用来训练输入序列(用户特征和浏览记录)。后来,日本雅虎(Yahoo)团队尝试将该文献中提到的推荐模型应用到手机端新闻主页中,整个推荐流程大致分为 5 步:1)将用户的历史浏览记录作为 RNN 的训练数据生成用户的偏好模型;2)利用一定的相似度计算规则计算出和用户偏好相符的新闻集合作为候选集;3)利用模型中的排序算法对新闻候选集排序;4)对重复的新闻内容进行去重;5)在适当时插入广告(如果需要)。经过实验和评估发现,GRU 模型需要设置的参数少而且能够为用户推荐更准确的新闻信息。文献[58]中则提出了一个多元递归神经网络(Multi-view Recurrent Neural Network, MV-RNN)模型,该模型能够将视频、文本、图片等信息整合,将不同的多视图特征进行组合作为输入项,然后在模型的隐藏层用一个单独且统一的结构来处理输入信息,动态有序地捕获用户的兴趣。

2.2.1　基于知识图谱和 RNN 的推荐

近年来互联网在多个领域快速发展,使得知识图谱从提升搜索引擎的质量逐渐发展到了推荐领域。2012 年,Google 公司为了提升用户使用搜索引擎时的搜索体验,提出了知识图谱的概念,知识图谱是用结构化网络对客观世界实体之间关系的一种描述,能够用形式化的方法表示现实生活中事物间的相互关系。文献[59]中结合知识图谱和 RNN 模型建立了一种能实时捕捉到用户兴趣点变化的序列化推荐模型。该模型将在线音乐平台的异构数据分为图形数据、文本数据和视觉数据三大类,用知识图谱将这三类异构数据的关系嵌入到实体中,再将结果作为输入嵌入到模型中;在解码阶段,RNN 和前馈层被用来获取序列中的信息,分析计算每个候选项的分数,最后预测推荐。该模型尝试将多源异构数据同时输入到模型中,提升了推荐的效率;但该模型只尝试了在音乐推荐方面的应用,因为它对异构数据具有较好的融合能力,未来可以尝试将其应用到视频、文本、社交网络推荐中,增强模型的可扩展性。

因为文献[59]中提出的模型无法记忆时间过长的序列,所以文献[60]提出了一个基于记忆的网络结构来长时间地保存用户个人信息和偏好。该结构可分为 Key 和 Value 两个模块,其中:Key 模块用来存储推荐项目的信息,从本质上来说这部分其实是知识图谱通过翻译嵌入(Translating Embedding, TransE)来获取实体和关系的表征信息;Value 模块存储用户的特征和偏好情况,在 RNN 对时间节点的信息迭代时,该模块能对 TransE 进行实时的记忆和更新,这一步充分利用了知识图谱中的信息。该网络结构有效地提升了

模型记忆过长序列的效率和推荐效果。

综上所述，在推荐方面充分合理地利用知识图谱能提升推荐性能，尤其对于缓解数据稀疏和冷启动问题具有明显的效果；但这也仅适用于数据积累较为成熟的系统，当面对数据积累较少的新系统时，往往会出现推荐准确率低、推荐效果差的问题。因此，如何利用知识图谱对新系统产生较好的推荐，将是未来研究的一个重点。

2.2.2 基于注意力机制的RNN推荐方法

注意力机制能够根据用户的偏好差异为推荐项目的潜在特征划分区域，赋予大部分用户都关注的区域较高的权重，不关注无关部分，其原理类似人脑的注意力机制，从本质上来说其工作原理是利用注意力的概率分布，捕捉对输出有关键影响的输入。

近年来，文献[61-63]等的研究将注意力机制和深度学习模型融合，推动了推荐系统的发展。文献[61]中将动态的图注意力机制模型和RNN模型结合混合应用于社区推荐，该研究认为用户的偏好受社交平台朋友的偏好影响，图注意力机制模型能够动态地捕获用户朋友长短期偏好变化对用户产生的影响，其模型如图9所示。该模型的推荐过程如下：

1）为用户的偏好情况建立模型。这一步主要是由RNN模型来完成，RNN模型为用户的历史浏览行为记录建模，动态地捕获到用户 $\boldsymbol{u}_n$ 的偏好 $\boldsymbol{h}_n$。

2）对用户社交网络中朋友的偏好情况进行表示。RNN模型不仅可以对用户的历史浏览行为记录进行建模，也能对用户社交网络中朋友们的历史浏览行为进行建模，朋友 k 的短期偏好情况用输出向量 $\boldsymbol{s}_k^s$ 表示，长期偏好情况用输出向量 $\boldsymbol{s}_k^l$ 表示，最后将两种输出向量 $\boldsymbol{s}_k^s$ 和 $\boldsymbol{s}_k^l$ 连接得到 $\boldsymbol{S}_k$，$\boldsymbol{S}_k$ 就是朋友 k 的整体偏好。

3）动态的图注意力机制建模。首先，为用户建立一个图网络，网络中的每个节点代表着用户与社交网络中朋友的图网络，如式(10)所示：

$$\boldsymbol{a}_{uk}^{(l)} = \frac{\exp\left(f\left(\boldsymbol{h}_u^{(l)}, \boldsymbol{h}_k^{(l)}\right)\right)}{\sum_{j \in N(u) \cup \{u\}} \exp\left(f\left(\boldsymbol{h}_u^{(l)}, \boldsymbol{h}_k^{(l)}\right)\right)} \tag{10}$$

其中：$\boldsymbol{h}_u^{(l)}$ 代表用户的长期偏好；$\boldsymbol{h}_k^{(l)}$ 代表朋友的长期偏好；$\boldsymbol{a}_{uk}^{(l)}$ 代表用户和朋友之间相差的注意力分数。

4）为用户产生推荐序列。将用户的偏好 $\boldsymbol{h}_n$ 和合并后的朋友偏好 $\boldsymbol{h}_u^l$ 连接就可以得到融合了用户朋友偏好的用户偏好表示 $\boldsymbol{h}_n$，如式(11)所示：

$$\boldsymbol{h}_n = \boldsymbol{W}_2\left[\boldsymbol{h}_n; \boldsymbol{h}_u^{(l)}\right] \tag{11}$$

其中：$\boldsymbol{W}_2$ 代表的是一个线性变换。之后softmax函数计算出项目 y 被用户喜欢的概率，如式(12)所示：

$$p\left(y \mid \boldsymbol{i}_{T+1,1}^u, \cdots, \boldsymbol{i}_{T+1,n}^u\right)\left\{\boldsymbol{S}_T^k, k \in N(u)\right\} = \frac{\exp\left(\hat{\boldsymbol{h}}_n^{\mathrm{T}} \boldsymbol{z}_y\right)}{\sum_{j=1}^{I} \exp\left(\hat{\boldsymbol{h}}_n^{\mathrm{T}} \boldsymbol{z}_j\right)} \tag{12}$$

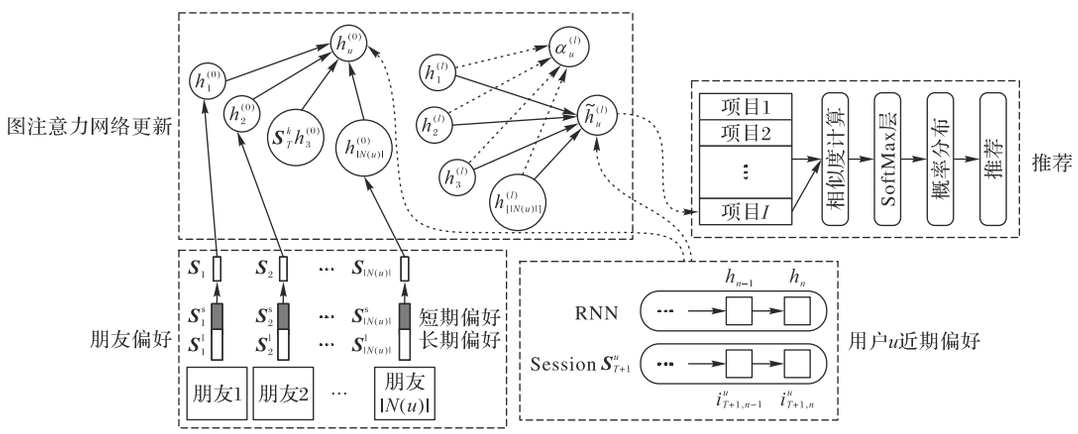

图9 动态图注意力网络社会推荐模型

Fig. 9 Social recommendation model via dynamic graph attention network

所有推荐项目的嵌入用 $\boldsymbol{z}_y$ 表示，推荐项目的总体数量用 I 表示，$\boldsymbol{S}_T^k$ 表示用户社交网络的第 k 个朋友在历史时刻为 T 的会话，$\boldsymbol{i}_{T,N_{k,T}}^u$ 表示用户社交网络中第 k 个用户在会话 T 中消费的第 $\boldsymbol{N}_{k,T}$ 个项目。

文献[63]中提出的模型充分利用了用户的社交关系，捕捉了用户的朋友的偏好；然而，该模型对用户和朋友的特征提取不够精准，没有考虑到用户和推荐项目之间长期的依赖关系。因此，在后续的研究中可以尝试将用户和推荐项目之间长期的依赖关系融合到模型中。

为了解决微博话题标签的时序数据问题，文献[64]中构建了一种基于主题注意力机制的LSTM模型，该模型考虑到了时间因子，将时序特征融入到了模型中，有效地提升了推荐的性能；但是该模型并没有考虑用户信息和微博标签文本长度问题等对推荐结果的影响。针对这一问题，文献[65]中提出了一种基于注意力机制的语句时态增强模型，该模型对微博特征从词级和语句级两方面进行分析和刻画，把时间信息融合在语句集注意力层，充分降低了微博标签数据中噪声数据对分类器的影响。因此，该模型除了解决微博话题标签推荐问题，还能用于解决文本识别、语言翻译和动态推荐等问题。然而，LSTM模型只能处理单一的欧几里得空间数据，无法处理较为复杂的非欧空间数据。

文献[66]中提出了双重注意力网络学习双重社会效应的推荐模型。该模型的双重注意力机制包括根据用户自己分配的注意力权重建模和通过上下文感知动态的注意力建

模两个方面,通过双重建模有效地把用户的社会效应传递到了推荐项目领域,缓解了传统推荐系统常常遇到的数据稀疏性问题。该模型对社会影响的有效表示能从多个维度学习,但是模型的复杂度也增加了。

2.3　卷积神经网络

卷积神经网络(Convolutional Neural Network, CNN)[67]的最大特点是具有表征学习能力,是包含深度卷积计算的前馈神经网络,它的核心是隐含层和卷积层的相互连接,常见的三种性能较好的CNN模型有VGGNet[68]、GoogLeNet和ResNet[69]。2014年提出的VGGNet模型取得了ILSVRC(ImageNet Large Scale Visual Recognition Challenge)分类组的亚军和图像识别组的冠军;2017年提出的ResNet模型在ILSVRC比赛中获得了冠军。ResNet模型有152层网络结构,将残差网络作为CNN的基本结构,这样做可以减少因网络结构过深造成的梯度爆炸问题。与其他的深度学习模型相比,CNN模型能够自动捕捉用户的潜在特征,发现潜在的规律,目前已广泛应用在图像识别、自然语言处理、目标分类等不同领域的推荐系统中[69-71]。

在音乐推荐方面,文献[70]中探讨了如何有效地缓解音乐推荐中新音乐冷启动问题,并提出了一种融合深度卷积神经网络的推荐模型,通过收集用户的历史收听记录和浏览过的音频数据,将这些数据投影到一个共有的隐空间中,从而学习用户和音频的隐表示。对于新的音乐,该研究利用深度卷积神经网络对新音频中的隐表示进行提取,从而在这个共有的空间中计算新音频和用户的相似度。经过实验和分析,该方法缓解了新音乐冷启动问题,提高了推荐的准确性。

在图像推荐方面,文献[71]中利用CNN模型学习用户和图像统一的特征表示,将异构用户图像网络转换为同质的低维数据,这样的转换有助于系统通过相似性向用户推荐图像。该模型能处理大型、稀疏和多样化的视觉图像。

针对文本推荐方面,文献[72]的研究认为用户对项目的评级矩阵如果过于稀疏,则会影响推荐质量,为此提出了一种混合推荐模型。该推荐模型基于上下文感知和卷积矩阵因式分解,将CNN集成在概率矩阵分解中,能有效捕获上下文信息,从而填补稀疏的用户评级矩阵,提高推荐的准确率。

文献[70-72]的研究针对的都是某一个特定的目标用户,但实际问题中的推荐场景往往更复杂,有时需要为特定的群组产生推荐列表,因此,未来的研究可尝试将CNN和社交关系、时间、文本等辅助信息相结合来进行群组推荐。

2016年,文献[73]中提出了一种基于注意力机制的CNN的新浪微博话题推荐模型,该模型设置了两个注意力通道(全局和局部),提高了推荐的准确率;但是该模型使用的数据都是文本类型,忽略了图像等其他形式的话题类型。为了解决这个问题,2017年,文献[74]中提出了协同注意力机制模型,充分考虑了文本、图像等与微博话题标签依赖关系,因此推荐性能优于仅考虑文本的推荐。

文献[75]的研究认为传统的推荐在提取评论文本信息方面有所欠缺,于是提出了一种基于注意力机制的深度协作神经网络(deep Cooperative Neural Network based on Attention, ACoNN)模型,其中注意力机制的作用是为文本矩阵的权重重新赋值,并行的CNN模型则充分挖掘用户和文本的信息以获取潜在的隐含特征。ACoNN模型的推荐流程如图10所示。

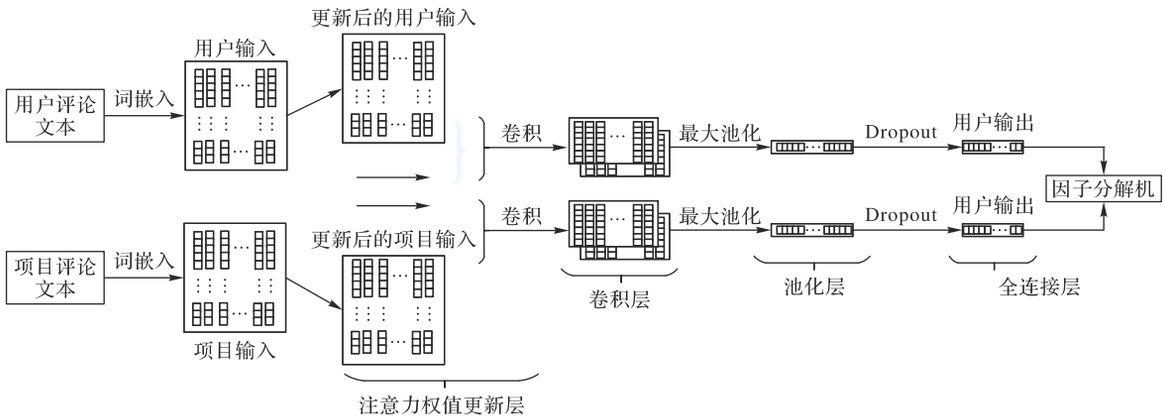

图10　ACoNN模型的推荐流程

Fig. 10　Recommendation flow of ACoNN model

具体描述如下:

1)将用户矩阵$\boldsymbol{M}_u$和文本信息矩阵$\boldsymbol{M}_i$利用词嵌入模型输入到输入层。sim_k代表户矩阵$\boldsymbol{M}_u$中的词向量$\boldsymbol{w}_k^u$与全部用户评论文本向量$\boldsymbol{w}_{1:n}^u$之间的相似度,如式(13)所示:

$$sim_k = \sum_{k=1}^{d} \cos\left(\left(\boldsymbol{w}_k^u\right), \left(\boldsymbol{w}_{1:n}^u\right)\right) \tag{13}$$

2)归一化处理式(13)得到的相似度系数,计算注意力机制层的每个词向量注意力权重,如式(14)所示:

$$a_k = \mathrm{e}^{sim_k} \Big/ \sum_{k=1}^{n} \mathrm{e}^{sim_k} \tag{14}$$

3)利用CNN模型对词向量矩阵进行卷积、池化和全连接,从而得到用户和项目的输出向量$\boldsymbol{output}_u$与$\boldsymbol{output}_i$。

4)连接$\boldsymbol{output}_u$与$\boldsymbol{output}_i$为用户-项矩阵构造特征向量$\boldsymbol{z}$;接着给向量$\boldsymbol{z}$融合因子分解机,用最小化损失函数训练向量$\boldsymbol{z}$,如式(15)所示:

$$J(\boldsymbol{w}_i) = y_{\text{real}} - \left(\boldsymbol{w}_0 + \sum_{i=1}^{|z|} \boldsymbol{w}_i z_i + \sum_{i=1}^{|z|} \sum_{j=i+1}^{|z|} w_{ij} z_i z_j\right) \tag{15}$$

其中:y_{real}代表用户对推荐项目的真实评分;$\boldsymbol{w}_0$代表全局的偏置量;$\boldsymbol{w}_i$代表的是第i个分量的权重值,z_i和z_j分别代表向量$\boldsymbol{z}$

的第i和第j个分量；$\boldsymbol{w}_{i,j}$代表的是z_i和z_j的交互值。

相比别的深度学习模型，该模型的优点是数据在训练阶段设置的参数比较少、模型的复杂度较低，充分利用了注意力机制能够捕捉权重较大信息的特点，以及CNN模型对权值能够共享、对局部连接的优势，结合了注意力机制的CNN模型在提取特征时对重点特征的提取效率有了很大的提升，因此，推荐项目的准确率也有了较大的改善；但是该模型无法对用户动态的偏好进行实时推荐。

文献[74-75]的研究将注意力机制和CNN模型相融合，尽管提升了推荐的效果，但是当数据规模足够大时，数据稀疏性问题仍然会逐渐显露。跨领域推荐是解决数据稀疏问题的一个重要的方法，多个领域的辅助信息可以为目标领域的推荐服务，通过输入辅助信息，模型可以学习到目标用户的潜在隐含特征，从而提升推荐的效果，因此，在后续研究中可以考虑将CNN和注意力机制融合到跨领域推荐任务中。

2.4 基于图神经网络的推荐

图神经网络（Graph Neural Network，GNN）借鉴RNN和CNN的思想，是一种重新定义和设计的用于处理非欧氏空间数据的深度学习算法。在实际的生活中，电子商务、推荐系统、动作识别等领域的数据抽象出来都是节点之间链接不固定的图谱，这些图谱不具备规则的空间结构，而GNN模型可以对该类数据进行高效的建模，精确地捕获到数据之间潜在的联系。文献[76]中针对电子商务领域出现的问题，提出了一种分层二分图神经网络的模型。该模型首先将多个GNN模型进行叠加，并在多个交替模块上使用聚类算法，聚类算法能够有效捕获到分层模块中推荐项目和用户的信息，进而有效地捕捉到用户的潜在偏好，提高推荐的准确率；但该模型利用的是用户某段时间内静态的交互记录，这与用户变化的偏好情况相矛盾。因此，文献[77]中建立了一种融合时间关注机制的图卷积推荐模型，图卷积神经层对用户在整个实际场景中的角色进行抽象，能大致反映出用户的短期偏好特征，文献[78]提出了一种卷积LSTM模型（Convolutional LSTM Network，ConvLSTM）来增强模型的鲁棒性，为了捕获到用户动态的偏好变化情况，模型首先融合了侧重分层学习和神经元排序的神经网络结构，最终通过学习模型捕获到的局部用户偏好的时空信息产生推荐序列。

2.4.1 基于PMF的图神经网络推荐

传统的矩阵分解模型具有很好的灵活性和可扩展性，但是仍然无法解决冷启动和数据稀疏的问题，于是，文献[79]中提出了一种融合PMF和GNN的推荐模型。该模型首先将社交网络图和用户项目图这两个图内在联系起来，然后对图进行建模，捕获用户在社会空间中的潜在特征向量和项目空间上的潜在特征向量；接着，将捕获到的特征向量进行相互串联，充分地学习目标用户的特征向量，将捕获到的特征向量集成在PMF模型中，产生项目的评分和推荐列表。在真实的数据集Epinions和Ciao上的实验结果表示，该模型是有效的，其均方根误差和平均绝对误差均有降低。但该模型只是将社交网络图作为辅助信息融合到模型中，在实际生活中，用户和项目之间的交互信息还体现在其他方面，例如，推荐项目的丰富属性与用户偏好的依赖性。未来可以考虑多方面地融合辅助信息，提高推荐模型的准确率和新颖性。

文献[80]提出了一种基于信任机制的社交网络推荐模型，该模型将神经网络集成到PMF模型中，用不同的神经网络的节点表示不同的用户，通过K最近邻（K-Nearest Neighbor，KNN）算法将用户特征和神经网络联系在一起形成图结构。但是该模型只考虑了单一的KNN联系用户特征和神经网络，未来的研究可以尝试多种方法联系用户特征和神经网络，尽可能从多方面考虑用户的特征。

文献[81]中通过GNN的节点来学习用户对特定推荐项目的置信度加权参数，该加权参数代表节点用户与推荐项目之间相交互的可能性。引入置信度加权参数是为了帮助用户模拟高阶信息，使得每个用户可以收集邻域节点间的高阶信息。对于比较稀疏的用户-项目矩阵，可以通过随机游走的方式对矩阵进行填充，缓解数据稀疏和冷启动问题。但该推荐方式仅考虑了用户项目之间历史交互的置信度参数，并没有考虑推荐系统所收集到的数据对加权参数的影响。

2.4.2 基于会话的图卷积神经网络推荐

近年来，匿名用户推荐问题逐渐成为推荐领域的一个重要研究方向，采用GNN模型解决该类问题已经取得了不错的进展；但是，GNN无法精确地捕获到用户会话间潜在的依赖信息。文献[82]中提出了一种基于会话的图卷积神经网络（Group-constrained Convolutional Recurrent Neural network，GCRNN）模型。该模型利用多层的图卷积模型能精确地捕获到用户会话图信息，利用递归神经网络层则能进一步捕获会话间的时序图来获得用户偏好的变化情况，而且递归神经网络层还能精确地捕获到会话之间的交互信息。因此，GCRNN模型能精确地捕获到会话间丰富的潜在隐含信息，从而提升推荐的准确性；然而，GCRNN模型并不能为用户产生动态的推荐列表，降低了模型的实效性，因此在今后的研究中可以考虑将用户的点击项作为辅助信息融合到模型中以产生更有效的推荐列表。

用户的兴趣是动态变化的，为了给用户产生实时推荐列表，文献[83]提出了一种基于会话的图卷积递归神经网络模型，模型的整体框架如图11所示。预测推荐项目的过程可分为三步：

1）对会话序列构建会话图，$\boldsymbol{I}=\{\boldsymbol{i}_1,\boldsymbol{i}_2,\boldsymbol{i}_3\cdots,\boldsymbol{i}_n\}$代表会话列表，$\boldsymbol{s}=[\boldsymbol{i}_{s,1},\boldsymbol{i}_{s,2},\boldsymbol{i}_{s,3},\cdots,\boldsymbol{i}_{s,t}]$代表按照时间戳进行排序的用户会话列表，$\boldsymbol{i}_{s,t}$代表用户在$t$时刻在会话$s$中的点击项，为用户的每个会话列表构建有向图$\boldsymbol{G}_S=(\boldsymbol{S}_S,\boldsymbol{E}_S)$，用户点击项$\boldsymbol{i}_{s,t}$作为会话图的节点，$\boldsymbol{i}_{s,t}$作为会话图的边，在用户会话列表$s$中，将节点向量作为RNN模型的输入，目的是节点向量能够被更新。接着，有向图$\boldsymbol{G}_S=(\boldsymbol{S}_S,\boldsymbol{E}_S)$被输入到嵌入层后，$\boldsymbol{i}_{s,t}$被映射到$\boldsymbol{G}$中，为了处理节点和会话图的收敛问题，文献[83]中对$\boldsymbol{G}_S=(\boldsymbol{S}_S,\boldsymbol{E}_S)$进行了卷积操作，如式(16)所示：

$$\boldsymbol{h}_\theta * \boldsymbol{g}=\boldsymbol{U}\boldsymbol{h}_\theta\boldsymbol{U}^{\mathrm{T}}\boldsymbol{g} \tag{16}$$

其中：$\boldsymbol{h}_\theta=\mathrm{diag}(\theta)$代表的是进行卷积操作时的滤波器；$\boldsymbol{g}$代表的是会话图；$\boldsymbol{U}$代表的是特征向量矩阵；$\boldsymbol{A}$代表的是邻接矩阵（若节点之间存在边，则$\boldsymbol{A}_{i,j}=1$，否则为0）。在建立图卷积模型时，获得会话图中的结构信息，利用多项式获取K阶

近似,K的阶数代表着有向图$\boldsymbol{G}_S=(\boldsymbol{S}_S,\boldsymbol{E}_S)$中每个节点在传播时的作用范围。

2)为了处理获取过程中遇到的梯度问题,选用GRU模型来获取节点向量,最终输出的$\boldsymbol{h}_i$的计算公式如(17)~(20)所示:

$$\boldsymbol{z}_i=\sigma\left(\boldsymbol{W}_z+\boldsymbol{U}_z\boldsymbol{h}_{i-1}\right) \tag{17}$$

$$\boldsymbol{r}_i=\sigma\left(\boldsymbol{W}_z+\boldsymbol{U}_r\boldsymbol{h}_{i-1}\right) \tag{18}$$

$$\tilde{\boldsymbol{h}}_i=\left(\frac{\pi}{2}-\theta\right)\tan h\left(\boldsymbol{W}_h+\boldsymbol{U}_h\left(\boldsymbol{r}\odot\boldsymbol{s}_{i-1}\right)\right) \tag{19}$$

$$\boldsymbol{h}_i=(1-\boldsymbol{z}_i)\odot\boldsymbol{s}_{i-1}+\boldsymbol{z}_i\odot\tilde{\boldsymbol{h}}_i \tag{20}$$

其中:$\boldsymbol{W}_Z$、$\boldsymbol{W}_h$和$\boldsymbol{U}_Z$、$\boldsymbol{U}_r$、$\boldsymbol{U}_h$为训练模型过程中得到的参数;$\sigma(\,)$代表sigmoid函数;$\odot$是代表乘法的运算符;$\boldsymbol{z}_i$和$\boldsymbol{r}_i$是GRU网络中的重置门与更新门,经过GRU编码,每个会话就被编码成一个个的嵌入序列$\boldsymbol{H}=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\boldsymbol{h}_3,\cdots,\boldsymbol{h}_n\}$,将嵌入向量经过线性变化变得到嵌入向量$\boldsymbol{h}_s$如式(21)所示。

$$\boldsymbol{h}_s=\boldsymbol{W}_s\left[\boldsymbol{h}_g;\ \boldsymbol{h}_l\right] \tag{21}$$

3)计算每个会话中点击项的得分$\bar{z}_i$,如式(22)所示:

$$\bar{\boldsymbol{z}}=\boldsymbol{h}_s^{\mathrm{T}}\boldsymbol{h}_i \tag{22}$$

其中:$\boldsymbol{h}_i$、$\boldsymbol{h}_s$分别代表点击项和会话的嵌入向量。接着计算会话被点击的概率,这一步是通过softmax层来完成的,如式(23)所示。

$$\bar{y}=\mathrm{softmax}(\bar{\boldsymbol{z}}) \tag{23}$$

预测出的y的值越大,则代表下一次被点击的可能性越大,那么通过对得到的y排序,将y值大的会话依次推荐给用户。

近年来基于会话的匿名推荐多关注的是用户的点击序列,但对于一个完整的推荐过程来说,其他信息(如推荐项目的种类和名称等)往往被忽略。为了解决上述问题,文献[84]中提出了一种基于会话的多粒度图神经网络推荐模型。该研究认为种类是推荐项目的一个重要特征属性,对推荐项目有聚合的作用,因此通过GNN获取推荐项目和用户的种类嵌入信息;接着,通过注意力机制捕获用户对项目分配的权重;最后使用RNN获得会话时序信息并对用户进行推荐,以提高推荐的泛化能力。但该模型并没有研究会话点击序列长度对推荐效果的影响,另外用户的长短期兴趣信息也可以尝试作为补充信息,从而进一步研究它对推荐的影响。

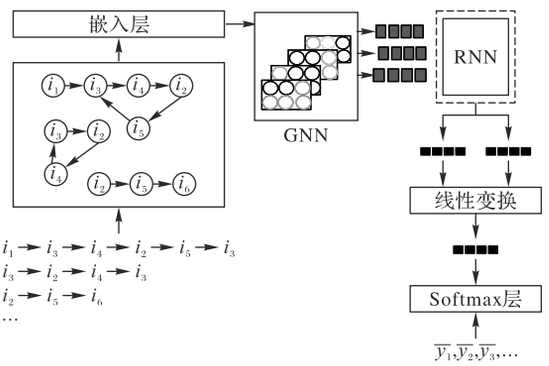

图11　基于会话的图卷积神经网络推荐框架

Fig. 11　Session-based graph convolutional neural network recommendation framework

GNN对非欧几里得数据有其强大的提取和表示能力,这是GNN模型的优势之一。在现有的研究中,基于GNN的推荐还存在以下问题:1)本节中介绍的模型的输入都是单一的数据类型,然而在现实生活中,数据的形式有文本、音频、图片等,如何对异构数据进行统一的输入是目前GNN推荐面临的问题;2)目前文献中所用到的GNN模型都是图的节点不为空的情况,然而在现实生活中,会存在节点对象没有存放任何数据的情况,目前相关算法难以处理该类情况。

随着GPU计算能力的提高,深度学习在大数据分析及个性化推荐中的应用越来越广泛。深度学习模型将用户的显式数据、隐式数据、用户画像等多源异构数据作为输入融入到推荐过程中,通过辅助数据分析用户和推荐项目隐藏的特征,从而建立预测模型,可有效地缓解数据稀疏和冷启动问题,达到提升推荐准确率和推荐质量的目的。表2归纳比较了深度学习模型在不同推荐系统中差异。

表2　不同的深度学习模型的文献总结及优点

Tab. 2　Literature summary and advantages of different deep learning models

模型	辅助数据类型	主要优点	主要难点	文献
DNN	视频、标签、用户和项目特征	1. 从多维度学习行为记录特征 2. 数据稀疏问题得到了有效解决 3. 缓解新用户面临的冷启动问题	1. 如何使推荐结果更新颖 2. 如何建立并且实现非线性特征的推荐模型	[49-52]
RNN	用户和项目特征	1. 动态地为用户推荐商品 2. 时效性强 3. 可解释性强	1. 多源异构数据特征如何有效表达 2. 如何为用户和项目的特征动态建模	[53-66]
CNN	图像、视频、音乐、文本	1. 有效地利用了辅助信息 2. 对用户的隐藏特征进行了挖掘 3. 提高了推荐的新颖性	1. 如何提高推荐模型的训练效率、响应时间以及可扩展性 2. 如何建立融入辅助信息的深度学习推荐模型	[67-75]
GNN	会话、文本;用户-项目特征	1. 能充分地挖掘节点信息之间的交互信息 2. 提高了图节点之间的敏感度 3. 可以在图领域对数据特征进行提取。	1. 如何有效地捕获到图节点的信息传递 2. 数据规模较大时,难以进行实时推荐	[76-83]

近年来,随着抖音、快手等短视频平台的快速发展,推荐系统成为了更加流行的研究热点。目前,深度学习模型因能与显、隐式反馈信息结合,并将多源异构数据融合到推荐系统中,从而有效缓解了传统推荐所面临的冷启动和数据稀疏

等问题，提高了推荐效果，其优点主要表现在：1）当遇到非结构化的数据（如图片、视频）时，数据隐含的特征信息仍然能通过深度学习其强大的表示学习能力被提取到。2）对原始数据的类型无要求，异构的数据均可以作为输入，从而进一步地获取目标用户的特征。然而，在不同的应用领域，融合深度学习模型的推荐算法仍然存在以下不足：1）深度学习模型虽然在 YouTube 视频、Google 地图等实际应用中取得了不错的效果，但由于视频、图片均属于非结构化的数据，且大量的非结构化数据训练起来复杂度极高且耗时。因此，未来对于在视频、图片等领域的推荐模型，应尽可能设计复杂度较低高效的模型。2）融合深度学习模型的推荐算法类似于一个黑盒，尤其在类似社交网络推荐的问题时，对于基于目标用户的社交网络推荐问题，深度学习模型往往都是个性化的推荐，很少有文献对此类推荐尝试群组推荐，未来对此方面的改进研究可以尝试建立群组推荐。表 3 列举了不同深度模型在不同的推荐领域所需要的数据类型及未来改进重点。

表 3 深度学习在不同推荐领域改进方向的比较

Tab. 3 Comparison of improvement directions of deep learning in different recommendation fields

应用方向	深度学习模型	数据类型	优点	未来改进方向
视频、图片推荐	AM、CNN、RNN、GNN 等	用户的隐、显式反馈信息，项目内容、用户生成内容、用户-项目的评分矩阵	1. 对大规模的非线性数据进行处理和计算 2. 不存在新项目或者新用户冷启动问题	1. 建立复杂度较低且高效的模型 2. 对异构数据能进行统一的处理（如可以同时输入视频、图片）
音乐推荐	CNN、RNN 等	用户-项目的评分矩阵、用户画像、社会化标注、项目数据、用户特征	1. 能动态地为用户进行有效推荐 2. 不存在新项目或者新用户冷启动问题	1. 将用户对音乐的情感表达作为特征属性融合到推荐模型中 2. 跨平台获取用户在不同情境下的音乐偏好
新闻推荐	MLP、RNN、CNN、GCN 等	目标用户的社会关系图、用户的隐显式反馈信息、知识图谱等	1. 新闻推荐的时效性高 2. 有效地解决数据稀疏的问题	1. 获取用户短期新闻偏好变化，从而动态地为用户推荐具有时效性的新闻 2. 建立机制对虚假、垃圾新闻进行有效屏蔽
社交网络推荐	RNN、RNN、CNN、GCN 等	目标用户的社会关系图、知识图谱、时间数据、位置数据等	1. 能对社交网络中社交信息的权重进行重新分配 2. 能跨平台捕获用户的社交网络	1. 用户隐私和安全的保护，需推荐系统建立相应的隐私保护机制 2. 建立对推荐新颖性、可靠性 、安全性评价指标的评估方法

3 常用数据集

推荐模型及其推荐效果要想获得公认、客观的评价，权威的数据集和统一的评价指标必不可少，本章主要介绍电影推荐、电子商务推荐、音乐推荐、新闻推荐领域一些公开的数据集以及近年来一些典型推荐模型的性能指标的对比。表 4 归纳整理了近年来有关推荐问题研究中所用到的公开数据集。

1）电影推荐[85]。MovieLens 数据集是由明尼苏达大学发布的一个包含多个用户对多部电影评级的数据集，包含了用户个人信息和有关电影的相关数据，因数据集的大小不同，目前包括 MovieLens 1M、MovieLens 10M、MovieLens 20M 三个版本。

2）电子商务[86]。Epinions 数据集包含了 139 738 个商品、49 290 个匿名用户，这些商品至少被评价过一次，共有 664 824 条评价记录，该数据集被广泛应用在商品推荐领域。Amazon 数据集由 Amazon 公司内部团队收集数据并创建，包含了商品的类别、数量、标价、用户的点击次数、浏览记录、购买情况等。

3）音乐推荐[87]。Last. fm 数据集是由马德里自治大学的研究小组创建并发布的，于 2011 年在第二届推荐系统信息异构与融合国际研讨会正式公开发布，音乐推荐算法的模型常常通过这个数据集进行仿真实验，也有研究学者将此数据集用于新闻推荐，以验证算法的通用性。

4）新闻推荐[88]。MIND 数据集是从微软新闻网站提取的匿名行为日志的新闻推荐数据集，有 MIND 和 MIND-small 两个版本。MIND 数据集包含了 1 000 000 名用户所浏览过的 161 031 篇新闻，包含了用户 24 155 470 条行为日志；MIND-small 数据集则包含了 50 000 名用户浏览过的 93 698 篇新闻以及 230 117 条用户的行为日志。Adressa 数据集是由挪威新闻出版社和挪威科技大学共同收集和发布的，不过因其新闻内容多为挪威语，因此应用常常受限。

5）文本推荐[89]。Yelp 数据集是美国最大的点评网站内部整理得到的数据集，常用于教育、研究和学术；Goodbooks-10k 数据集来自 goodreads 网站，包含用户的文本评论，图书的标签，被用户评论过的书的详细信息（作者、年份、书的类型等）。Yelp 和 Goodbooks-10k 数据集常用于基于用户评论的文本推荐、图书推荐等领域。

4 应用及比较

传统推荐算法中的 CF 是最早被提出且发展最好的推荐算法。近年来，以 CF 为主的改进算法不断涌现，如基于 PFM 的协同过滤、融合时间因素的协同过滤、基于知识图谱的协同过滤、基于信任因子的协同过滤等，这些算法都取得了令人满意的推荐效果。相比 CF，CB 更多地是作为辅助算法，CB 包括特征提取和产生推荐列表两个过程，很容易造成推荐性能低的问题。混合推荐算法是各种推荐算法的组合，能够让不同的推荐算法相互弥补不足，能有效地缓解数据稀疏

的问题。目前,基于深度学习的推荐的核心是将不同的深度学习模型与CF或CB组合,其推荐过程可分为两步:1)让深度学习模型学习用户或项目隐含的潜在特征,并和CF结合构建优化函数对参数进行训练;2)从完成训练的模型中获取最终的隐向量,接着完成向用户推荐的过程。

在面对复杂庞大的数据时,传统推荐算法常常无法快速建模且表示性较差,而深度学习可以对复杂问题分层处理,能快速发现每一层数据之间潜在的规律和联系。基于深度学习的推荐通过融入辅助信息能有效地缓解传统推荐技术的数据稀疏和冷启动等问题。现有研究大多根据具体的辅助信息而选取不同的深度学习模型,在以后的研究中可以尝试针对所有的辅助信息建立一个统一的混合推荐模型。

表5整理了不同推荐技术在电影、音乐、新闻、社交网络、视频和广告等六个典型领域中的应用,并列举了这六个领域中代表性的推荐模型、需要的数据信息以及模型特点。

表4　常用公开数据集归纳统计信息

Tab. 4　Summarization and statistics of commonly used open datasets

数据集的类型及名称		用户数量	项目数量	评论数量	稀疏度/%	获取链接
电影推荐	MovieLens 1M	6 040	3 883	1 000 209	4. 26	https://grouplens. org/datasets/movielens/
	MovieLens10M	71 567	9 164	10 000 054	1. 3	https://grouplens. org/datasets/movielens/
	MovieLens 20M	138 493	27 278	20 000 263	0. 52	https://grouplens. org/datasets/movielens/
电子商务	Epinions	49 290	139 738	664 824	0. 011	http://www. trustlet. org/wiki/Epinions_datasets
	Amazon	5 786	26 573	14 280 000	0. 002	http://jmcauley. ucsd. edu/data/amazon/
音乐推荐	Last. fm	1 892	17 632	92 834	0. 28	https://grouplens. org/datasets/hetrec-2 011/
新闻推荐	Mind-small	50 000	93 698	230 117	0. 056	https://msnews. github. io
	Mind	1 000 000	161 013	24 155 470	0. 012	https://msnews. github. io
文本推荐	Yelp	2 189 457	1 162 119	8 635 403	0. 043	https://www. yelp. com/dataset
	Goodbooks-10k	865 456	10 000	6 000 000	0. 12	https://github. com/zygmuntz/goodbooks-10k

表5　推荐技术在不同领域的应用

Tab. 5　Applications of recommended techniques in different fields

应用领域	推荐技术	代表性模型	数据信息类型	模型特点
电影推荐	DNN,CF	TransHR[90]	用户信息、电影信息、观看时间、评分	超关系数据翻译嵌入模型(Translating embedding for Hyper-Relational data, TransHR)更关注电影之间的关系,它将电影之间的关系嵌入到关系空间中,电影之间的多种向量关系能得到保留,但也因此增加了模型的空间复杂度
音乐推荐	CNN,GRU,CB	PEIA[91];潜在因素模型[92]	音乐信息、歌手信息、听歌时长、点击数据	整合性格情绪专注的模型(Personality and Emotion Integrated Attentive model, PEIA)是将用户的人格和情感结合的模型,充分利用了用户的兴趣偏好变化和社交数据;潜在因素模型通过人的听觉效应特征学习用户潜在偏好的音频,处理的音频被输入到CNN模型中,缓解了冷启动问题
新闻推荐	LSTM,GRU	NPA[93];LSTUR[94]	新闻信息、时间、位置	个性化注意力的神经网络新闻推荐(Neural news recommendation with Personalized Attention, NPA)模型主要关注不同用户对同一篇新闻的感兴趣程度,从而利用模型中的注意力机制部分对用户的兴趣建模;长短期用户表示(Long and Short-Term User Representation, LSTUR)模型融合了用户偏好和时序兴趣,能动态为用户产生新闻推荐
社交网络	MLP,CF,MF	NSCR[95]	用户信息、时间、位置等环境信息	神经社会协作分级(Neural Social Collaborative Ranking, NSCR)模型是将MLP和CF结合的一种深度协同过滤推荐算法,其输入是用户特征信息,经过MLP预测用户潜在偏好
视频推荐	MLP	YouTube视频推荐过程[49]	播放量、点击次数、访问日志	文献[49]中将DNN融入到了视频推荐中,将推荐过程分为候选集生成阶段和视频排序阶段,传统的推荐方法很难为用户推荐符合其偏好的视频
广告推荐	CF,DL	AdROSA[96];FBARS[97]	网页信息、文本数据、面部信息	自适应个性化的网络广告推荐模型(Adaptive personalization of web advertising, AdROSA)通过用户对广告的响应时间和对广告的评论信息捕获用户的潜在偏好;基于深度学习的人脸广告推荐模型(Face Based Advertisement Recommendation System with deep learning, FBARS)的关键是实现对人脸的三维特征提取,然后将特征提取结果以三维数组的形式传入深度学习推荐模块

5　推荐系统面临的挑战和研究趋势

推荐系统旨在从海量的推荐对象中帮助用户发现符合其偏好的推荐项。本文分析了四类不同的推荐系统,包括基于内容的推荐技术、基于协同过滤的推荐技术、混合推荐技术以及基于深度学习的推荐系统,虽然这些推荐技术已经取得了令人满意的推荐效果,但仍面临以下挑战,未来可以尝试在以下这些方面进行研究:

1)通过动态信息为用户推荐项目。大部分文献中所提到的推荐技术都是通过静态信息(假定用户的行为记录不改变)对用户推荐商品。然而,在实际生活中,用户的喜好会随

着时间、空间以及内和外部环境的变化而变化，因此，未来对用户偏好建模时可以考虑动态的推荐算法。文献[98]中通过建立深度递归神经网络模型，使得用户每打开一个新的Web页面，都会刷新推荐结果，从而实现实时动态推荐服务。类似地，文献[99]中通过建立基于递归神经网络的协作序列模型，能够准确地捕获用户上下文状态隐藏的特征向量，为用户动态地推荐项目。然而，目前动态实时地为用户推荐方面的研究较少，如何根据用户的偏好变化动态地为用户推荐项目，仍是未来推荐系统研究的热点之一。

2)推荐系统安全性有待提高。大规模的在线网站吸引了海量用户的加入，尤其是社交网站的发展，精确地为用户推荐感兴趣的项目成了各个网站吸引用户手段之一，而只有对用户的多维度(特征)信息的挖掘才能更容易找到符合其偏好的推荐对象。事实上，用户在期望推荐系统推荐感兴趣的商品时，并不希望个人的其他隐私被公开，目前的研究都是通过数据扭曲和数据模糊的算法扰乱用户的信息。这种数据扰乱虽对用户的个人隐私做到了保护，但也会导致提取到的用户信息并不准确，大大降低了推荐的准确性，因此，接下来可以着重研究一种既能保护用户隐私又可以提高推荐准确性的方法。

3)缺少提取用户偏好特征的方法。目前的推荐系统推荐对象更多的是依赖用户对推荐项目的评分或者反馈信息，而忽略了用户和推荐对象本身的特征，目前研究缺少适当的建模方法对用户和推荐项目的特征、线性和非线性关系进行多维的提取。因此，接下来的研究中需引入更多样的方式来提取用户和推荐对象的特征。

4)评价推荐系统的性能指标单一。现有的研究在衡量推荐系统性能时，它们多关注的是推荐结果是否准确以及准确率是多少，它们认为准确率是衡量推荐系统好坏的最关键指标。推荐的准确率高，则认为这个推荐系统是好的；反之，则不是一个好的推荐。但是，用户在真正使用这些应用程序时，不仅希望系统可以精确地推荐感兴趣的项目，也期待出现更加多样且新颖的推荐。因此，在未来研究中推荐项目的新颖性、多样性都应该作为推荐系统的评价指标。

6 结语

随着深度学习、数据挖掘、预测算法等技术的不断成熟，提高推荐系统的准确率、安全性、隐私性将成为未来研究的热点。本文深入分析了传统推荐方法以及融入了不同深度学习模型的推荐方法，整理总结了不同推荐领域常用的数据集，对比了传统推荐模型和基于深度学习模型的区别，尝试对推荐系统现存问题进行了总结并对推荐系统的未来发展方向做了展望，希望能对推荐系统领域或深度学习领域感兴趣的研究人员提供有益的帮助。

参考文献(References)

[1] 周惠宏，柳益君，张尉青，等. 推荐技术在电子商务中的运用综述[J]. 计算机应用研究，2004，21(1)：8-12.(ZHOU H H，LIU Y J，ZHANG W Q，et al. A survey of recommender system applied in E-commerce[J]. Application Research of Computers，2004，21(1)：8-12.)

[2] 刘君良，李晓光. 个性化推荐系统技术进展[J]. 计算机科学，2020，47(7)：47-55.(LIU J L，LI X G. Techniques for recommendation system：a survey[J]. Computer Science，2020，47(7)：47-55.)

[3] SONG Y，ELKAHKY A M，HE X D. Multi-rate deep learning for temporal recommendation [C]// Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York：ACM，2016：909-912.

[4] ZHANG S，YAO L N，SUN A X，et al. Deep learning based recommender system：a survey and new perspectives [J]. ACM Computing Surveys，2020，52(1)：No. 5.

[5] GOODWIN P，KEITH ORD J，ÖLLER L-E，et al. Principles of Forecasting：A Handbook for Researchers and Practitioners[M]. Boston，MA：Kluwer Academic，2001：61-70.

[6] RESNICK P，VARIAN H R. Recommender systems [J]. Communications of the ACM，1997，40(3)：56-58.

[7] SUN M X，LEBANON G，KIDWELL P. Estimating probabilities in recommendation systems[C]// Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. New York：JMLR. org，2011：734-742.

[8] LIU H Y，HE J，WANG T T，et al. Combining user preferences and user opinions for accurate recommendation [J]. Electronic Commerce Research and Applications，2013，12(1)：14-23.

[9] ADOMAVICIUS G，TUZHILIN A. Toward the next generation of recommender systems：a survey of the state-of-the-art and possible extensions [J]. IEEE Transactions on Knowledge and Data Engineering，2005，17(6)：734-749.

[10] HA S H. Helping online customers decide though Web personalization [J]. IEEE Intelligent Systems，2002，17(6)：34-43.

[11] VERBERT K，MANOUSELIS N，OCHOA X，et al. Context-aware recommender systems for learning：a survey and future challenges [J]. IEEE Transactions on Learning Technologies，2012，5(4)：318-335.

[12] MOONEY R J，ROY L. Content-based book recommending using learning for text categorization[C]// Proceedings of the 5th ACM Conference on Digital Libraries. New York：ACM，2000：195-204.

[13] BREESE J S，HECKERMAN D，KADIE C. Empirical analysis of predictive algorithms for collaborative filtering[C]// Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence. San Francisco：Morgan Kaufmann Publishers Inc.，1998：43-52.

[14] BALABANOVIĆ M，SHOHAM Y. Fab：content-based，collaborative recommendation[J]. Communications of the ACM，1997，40(3)：66-72.

[15] LIU L W，LECUE F，MEHANDJIEV N. Semantic content-based recommendation of software services using context [J]. ACM Transactions on the Web，2013，7(3)：No. 17.

[16] SEGEV A，TOCH E. Context-based matching and ranking of Web services for composition [J]. IEEE Transactions on Services Computing，2009，2(3)：210-222.

[17] BROENS T，POKRAEV S，SINDEREN M van，et al. Context-aware，ontology-based service discovery[C]// Proceedings of the 2004 European Symposium on Ambient Intelligence，LNCS 3295. Berlin：Springer，2004：72-83.

[18] MEDJAHED B，ATIF Y. Context-based matching for Web service composition [J]. Distributed and Parallel Databases，2007，21(1)：5-37.

[19] 许海玲，吴潇，李晓东，等. 互联网推荐系统比较研究[J]. 软件

学报, 2009, 20(2):350-362. (XU H L, WU X, LI X D, et al. Comparative research on Internet recommendation systems [J]. Journal of Software, 2009, 20(2): 350-362.)

[20] 黄立威,江碧涛,吕守业,等. 基于深度学习的推荐系统研究综述[J]. 计算机学报, 2018, 41(7): 1619-1647. (HUANG L W, JIANG B T, LYU S Y, et al. Comparison study of Internet recommendation system[J]. Chinese Journal of Computers, 2018, 41(7): 1619-1647.)

[21] GOLDBERG D, NICHOLS D, OKI B M, et al. Using collaborative filtering to weave an information tapestry [J]. Communications of the ACM, 1992, 35(12):61-70.

[22] CAI Y, LEUNG H F, LI Q, et al. Typicality-based collaborative filtering recommendation [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(3): 766-779.

[23] DESHPANDE M, KARYPIS G. Item-based top-*N* recommendation algorithms [J]. ACM Transactions on Information Systems, 2004, 22(1): 143-177.

[24] KOREN Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model [C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2008: 426-434.

[25] MA W M, SHI J F, ZHAO R D. Normalizing item-based collaborative filter using context-aware scaled baseline predictor [J]. Mathematical Problems in Engineering, 2017, 2017: No. 6562371.

[26] 周万珍,曹迪,许云峰,等. 推荐系统研究综述[J]. 河北科技大学学报, 2020, 41(1): 76-87. (ZHOU W Z, CAO D, XU Y F, et al. A survey of recommender systems [J]. Journal of Hebei University of Science and Technology, 2020, 41(1): 76-87)

[27] SARWAR B, KARYPIS G, KONSTAN J, et al. Item-based collaborative filtering recommendation algorithms [C]// Proceedings of the 10th International Conference on World Wide Web. New York: ACM, 2001: 285-295.

[28] RESNICK P, IACOVOU N, SUCHAK M, et al. GroupLens: an open architecture for collaborative filtering of netnews [C]// Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work. New York: ACM, 1994: 175-186.

[29] YOO H, CHUNG K. Deep learning-based evolutionary recommendation model for heterogeneous big data integration [J]. KSII Transactions on Internet and Information Systems, 2020, 14(9): 3730-3744.

[30] SALAKHUTDINOV R, MNIH A. Probabilistic matrix factorization [C]// Proceedings of the 20th International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc. , 2007: 1257-1264.

[31] FUNK S. Funk-SVD [EB/OL]. (2006-12-11) [2020-11-01]. http://sifter. org/simon/journal/20061211. html.

[32] LI G, OU W H. Pairwise probabilistic matrix factorization for implicit feedback collaborative filtering [J]. Neurocomputing, 2016, 204: 17-25.

[33] STRAHL J, PELTONEN J, MAMITSUKA H, et al. Scalable probabilistic matrix factorization with graph-based priors [C]// Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2020: 5851-5858.

[34] XU C H. A novel recommendation method based on social network using matrix factorization technique [J]. Information Processing and Management, 2018, 54(3): 463-474.

[35] 王运,倪静,马刚. 基于FunkSVD矩阵分解和相似度矩阵的推荐算法[J]. 计算机应用与软件, 2019, 36(12): 245-250. (WANG Y, NI J, MA G. Recommendation algorithm based on FunkSVD matrix decomposition and similarity matrix [J]. Computer Applications and Software, 2019, 36(12):245-250.)

[36] DEEP K, THAKUR M. A new crossover operator for real coded genetic algorithms [J]. Applied Mathematics and Computation, 2007, 188(1): 895-911.

[37] KOREN Y. Collaborative filtering with temporal dynamics [C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2009: 447-456.

[38] KOREN Y. Factor in the neighbors: scalable and accurate collaborative filtering [J]. ACM Transactions on Knowledge Discovery from Data, 2010, 4(1): No. 1.

[39] DE CAMPOS L M, FERNÁNDEZ-LUNA J M, HUETE J F, et al. Combining content-based and collaborative recommendations: a hybrid approach based on Bayesian networks [J]. International Journal of Approximate Reasoning, 2010, 51(7): 785-799.

[40] PAZZANI M J. A framework for collaborative, content-based and demographic filtering[J]. Artificial Intelligence Review, 1999, 13(5/6): 393-408.

[41] ZHANG H, GE D C, ZHANG S Y. Hybrid recommendation system based on semantic interest community and trusted neighbors [J]. Multimedia Tools and Applications, 2018, 77(4): 4187-4202.

[42] HUANG Z H, YU C, NI J, et al. An efficient hybrid recommendation model with deep neural networks [J]. IEEE Access, 2019, 7: 137900-137912.

[43] HU J J, LIU L Z, ZHANG C Y, et al. Hybrid recommendation algorithm based on latent factor model and PersonalRank [J]. Journal of Internet Technology, 2018, 19(3): 919-926.

[44] CHEN S J, QIN Z, WILSON Z, et al. Improving recommendation quality in Google Drive [C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery. New York: ACM, 2020: 2900-2908.

[45] SHAN Y, HOENS T R, JIAO J, et al. Deep crossing: web-scale modeling without manually crafted combinatorial features [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 255-262.

[46] HOENS T R, BLANTON M, STEELE A, et al. Reliable medical recommendation systems with patient privacy [J]. ACM Transactions on Intelligent Systems and Technology, 2010, 4(4): No. 67.

[47] FANG L J, KIM H, LeFEVRE K, et al. A privacy recommendation wizard for users of social networking sites [C]// Proceedings of the 17th ACM Conference on Computer and Communications Security. New York: ACM, 2010: 630-632.

[48] 蒋伟. 推荐系统若干关键技术研究[D]. 成都:电子科技大学, 2018: 10-20. ((JIANG W. Research on some key technologies of recommender systems [D]. Chengdu: University of Electronic Science and Technology of China, 2018: 10-20.)

[49] COVINGTON P, ADAMS J, SARGIN E. Deep neural networks for YouTube recommendations [C]// Proceedings of the 10th ACM Conference on Recommender Systems. New York: ACM, 2016: 191-198.

[50] CHENG H T, KOC L, HARMSEN J, et al. Wide & Deep learning for recommender systems [C]// Proceedings of the 1st

Workshop on Deep Learning for Recommender Systems. New York: ACM, 2016: 7-10.

[51] XU X. Matrix factorization recommendation algorithm based on deep neural network [C]// Proceedings of the 2nd International Conference on Information Systems and Computer Aided Education. Piscataway: IEEE, 2019: 320-323.

[52] ZHANG L, LUO T, ZHANG F, et al. A recommendation model based on deep neural network [J]. IEEE Access, 2018, 6: 9454-9463.

[53] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors [J]. Nature, 1986, 323 (6088): 533-536.

[54] OKURA S, TAGAMI Y, ONO S, et al. Embedding-based news recommendation for millions of users [C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2017: 1933-1942.

[55] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks [C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 3104-3112.

[56] ZHANG Y, PEZESHKI M, BRAKEL P, et al. Towards end-to-end speech recognition with deep convolutional neural networks [C]// Proceedings of the InterSpeech 2016. [S. l.]: International Speech Communication Association, 2016: 410-414.

[57] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[58] CUI Q, WU S, LIU Q, et al. MV-RNN: a multi-view recurrent neural network for sequential recommendation [J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 32(2): 317-331.

[59] HUANG J, ZHAO W X, DOU H j, et al. Improving sequential recommendation with knowledge-enhanced memory networks [C]// Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2018: 505-514.

[60] LIN Q K, NIU Y Q, ZHU Y F, et al. Heterogeneous knowledge-based attentive neural networks for short-term music recommendations [J]. IEEE Access, 2018, 6: 58990-59000.

[61] SONG W P, XIAO Z P, WANG Y F, et al. Session-based social recommendation via dynamic graph attention networks [C]// Proceedings of the 12th ACM International Conference on Web Search and Data Mining. New York: ACM, 2019: 555-563.

[62] 张昕,刘思远,徐雁翎. 结合注意力机制的知识感知推荐算法[J/OL]. 计算机工程与应用. (2021-03-31) [2021-04-11]. https://kns. cnki. net/kcms/detail/11. 2127. TP. 20210330. 1708. 026. html. (ZHANG X, LIU S Y, XU Y L. Knowledge-aware recommendation algorithm combined with attention mechanism [J/OL]. Computer Engineering and Applications. (2021-03-31) [2021-04-11]. https://kns. cnki. net/kcms/detail/11. 2127. TP. 20210330. 1708. 026. html.)

[63] 任柯舟,彭甫镕,郭鑫,等. 动态融合社交信息的社会化推荐[J]. 计算机应用, 2021, 41(10): 2806-2812. (REN K Z, PENG F R, GUO X, et al. Social recommendation based on dynamic integration of social information [J]. Journal of Computer Applications, 2021, 41(10): 2806-2812.)

[64] LI Y, LIU T, JIANG J, et al. Hashtag recommendation with topical attention-based LSTM [C]// Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. [S. l.]: The COLING 2016 Organizing Committee, 2016: 3019-3029.

[65] DEY K, SHRIVASTAVA R, KAUSHIK S. Topical stance detection for Twitter: a two-phase LSTM model using attention [C]// Proceedings of the 2018 European Conference on Information Retrieval. Cham: Springer, 2018: 529-536.

[66] WU Q T, ZHANG H R, GAO X F, et al. Dual graph attention networks for deep latent representation of multifaceted social effects in recommender systems [C]// Proceedings of the 2019 World Wide Web Conference. New York: ACM, 2019: 2091-2102.

[67] LeCUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[68] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10) [2021-02-20]. https://arxiv. org/pdf/1409. 1556. pdf.

[69] HUANG G, LIU Z, MAATEN L van der, et al. Densely connected convolutional networks [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 4700-4708.

[70] OORD A van den, DIELEMAN S, SCHRAUWEN B. Deep content-based music recommendation [C]// Proceedings of the 26th International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2013: 2643-2651.

[71] GENG X, ZHANG H W, BIAN J W, et al. Learning image and user features for recommendation in social networks [C]// Proceedings of the 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015: 4274-4282.

[72] KIM D, PARK C, OH J, et al. Convolutional matrix factorization for document context-aware recommendation [C]// Proceedings of the 10th ACM Conference on Recommender Systems. New York: ACM, 2016: 233-240.

[73] GONG Y Y, ZHANG Q. Hashtag recommendation using attention-based convolutional neural network [C]// Proceedings of the 25th International Joint Conference on Artificial Intelligence. California: ijcai. org, 2016: 2782-2788.

[74] ZHANG Q, WANG J W, HUANG H R, et al. Hashtag recommendation for multimodal microblog using co-attention network [C]// Proceedings of the 26th International Joint Conference on Artificial Intelligence. California: ijcai. org, 2017: 3420-3426.

[75] 黄文明,卫万成,张健,等. 基于注意力机制与评论文本深度模型的推荐方法[J]. 计算机工程, 2019, 45(9): 176-182. (HUANG W M, WEI W C, ZHANG J, et al. Recommendation method based on attention mechanism and review text deep model [J]. Computer Engineering, 2019, 45(9): 176-182.)

[76] LI Z, SHEN X, JIAO Y H, et al. Hierarchical bipartite graph neural networks: towards large-scale e-commerce applications [C]// Proceedings of the IEEE 36th International Conference on Data Engineering. Piscataway: IEEE, 2020: 1677-1688.

[77] ZHANG M G, YANG Z Y. GACOforRec: session-based graph convolutional neural networks recommendation model [J]. IEEE Access, 2019, 7: 114077-114085.

[78] SHI X J, CHEN Z R, WANG H, et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting [C]// Proceedings of the 28th International Conference on Neural

Information Processing Systems. Cambridge: MIT Press, 2015: 802-810.

[79] 王英博,孙永荻. 基于GNN的矩阵分解推荐算法[J]. 计算机工程与应用, 2021, 57(19):129-134. (WANG Y B, SUN Y D. GNN-based matrix factorization recommendation algorithm [J]. Computer Engineering and Applications, 2021, 57 (19): 129-134.)

[80] DENG S G, HUANG L T, XU G D, et al. On deep learning for trust-aware recommendations in social networks [J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28 (5): 1164-1177.

[81] YANG J H, CHEN C M, WANG C J, et al. HOP-Rec: high-order proximity for implicit recommendation [C]// Proceedings of the 12th ACM Conference on Recommender Systems. New York: ACM, 2018: 140-144.

[82] LIN S D, RUNGER G C. GCRNN: group-constrained convolutional recurrent neural network [J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29 (10): 4709-4718.

[83] 曹万平,周刚,陈黎,等. 基于会话的图卷积递归神经网络推荐模型[J]. 四川大学学报(自然科学版), 2021, 58(2):66-72. (CAO W P, ZHOU G, CHEN L, et al. Session-based graph convolutional recurrent neural networks recommendation model [J]. Journal of Sichuan University (Natural Science Edition), 2021, 58(2):66-72.)

[84] 任俊伟,曾诚,肖丝雨,等. 基于会话的多粒度图神经网络推荐模型[J]. 计算机应用, 2021, 41(11): 3164-3170. (REN J W, ZENG C, XIAO S Y, et al. Session-based recommendation model of multi-granular graph neural network [J]. Journal of Computer Applications, 2021, 41(11): 3164-3170.)

[85] HARPER F M, KONSTAN J A. The MovieLens datasets: history and context [J]. ACM Transactions on Interactive Intelligent Systems, 2016, 5(4): No. 19.

[86] CHIA P H, PITSILIS G. Exploring the use of explicit trust links for filtering recommenders: a study on Epinions. com[J]. Journal of Information Processing, 2011, 19: 332-344.

[87] CANTADOR I, BRUSILOVSKY P, KUFLIK T. Second workshop on information heterogeneity and fusion in recommender systems (HetRec2011)[C]// Proceedings of the 5th ACM Conference on Recommender Systems. New York: ACM, 2011: 387-388.

[88] WU F Z, QIAO Y, CHEN J H, et al. MIND: a large-scale dataset for news recommendation [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2020: 3597-3606.

[89] KRONMUELLER M, CHANG D J, HU H Q, et al. A graph database of Yelp Dataset Challenge 2018 and using cypher for basic statistics and graph pattern exploration[C]// Proceedings of the 2018 IEEE International Symposium on Signal Processing and Information Technology. Piscataway: IEEE, 2018: 135-140.

[90] ZHOU M, ZHANG C H, HAN X, et al. Knowledge graph completion for hyper-relational data[C]// Proceedings of the 2016 International Conference on Big Data Computing and Communications, LNISA 9784. Cham: Springer, 2016: 236-246.

[91] SHEN T C, JIA J, LI Y, et al. PEIA: personality and emotion integrated attentive model for music recommendation on social media platforms[C]// Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2020: 206-213.

[92] LIU C L, CHEN Y C. Background music recommendation based on latent factors and moods[J]. Knowledge-Based Systems, 2018, 159: 158-170.)

[93] WU C H, WU F Z, AN M X, et al. NPA: neural news recommendation with personalized attention [C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2019: 2576-2584.

[94] AN M X, WU F Z, WU C H, et al. Neural news recommendation with long-and short-term user representations[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2019: 336-345.

[95] WANG X, HE X N, NIE L Q, et al. Item silk road: recommending items from information domains to social users[C]// Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2017: 185-194.

[96] KAZIENKO P, ADAMSKI M. AdROSA — adaptive personalization of web advertising [J]. Information Sciences, 2007, 177(11): 2269-2295.

[97] YAO X Z, CHEN Y Y, LIAO R , et al. Face based advertisement recommendation with deep learning: a case study [C]// Proceedings of the 2017 International Conference on Smart Computing and Communication, LNISA 10699. Cham: Springer, 2018: 96-102.

[98] WU S, REN W C, YU C C, et al. Personal recommendation using deep recurrent neural networks in NetEase[C]// Proceedings of the IEEE 32nd International Conference on Data Engineering. Piscataway: IEEE, 2016: 1218-1229.

[99] WU S, REN W C, YU C C, et al. Personal recommendation using deep recurrent neural networks in NetEase[C]// Proceedings of the IEEE 32nd International Conference on Data Engineering. Piscataway: IEEE, 2016: 1218-1229.

This work is partially supported by National Natural Science Foundation of China (12050410248), Science and Technology Program of Sichuan Province (2021YFH0120), Southwest Minzu University Graduate Innovative Research Project (CX2020SZ04).

YU Meng, born in 1995, M. S. candidate. Her research interests include recommendation system, information filtering, data mining.

HE Wentao, born in 1996, M. S. candidate. His research interests include deep learning, data mining.

ZHOU Xuchuan, born in 1972, Ph. D., professor. His research interests include data mining, deep learning.

CUI Mengtian, born in 1972, Ph. D., professor. Her research interests include intelligent information processing.

WU Keqi, born in 1997, M. S. candidate. His research interests include recommendation system.

ZHOU Wenjie, born in 1997, M. S. candidate. His research interests include data mining.