文章编号: 1000-5641(2017)05-0101-16

跨领域推荐技术综述

陈雷慧[1], 匡 俊[1], 陈 辉[2], 曾 炜[2], 郑建兵[1], 高 明[1]

(1. 华东师范大学 数据科学与工程学院, 上海 200062;
2. 深圳腾讯计算机系统有限公司, 北京 100080)

摘要: 随着信息技术和互联网的飞速发展, 信息过载的问题日趋严重. 个性化推荐系统是解决这一问题的热门技术. 推荐系统的核心在于推荐算法, 在过去的十年里, 基于单领域的协同过滤推荐算法应用最为广泛. 但用户和项目数量的急剧增长使得传统的协同过滤推荐算法面临冷启动和数据稀疏问题的挑战. 跨领域推荐旨在整合来自不同领域的用户偏好特征, 针对每个用户自身特点进行智能化感知, 精准满足用户个性化需求, 从而提高目标领域推荐结果的准确性和多样性, 现已成为推荐系统研究领域中的热门话题. 本文首先对跨领域推荐技术进行系统地研究和分析, 概述跨领域推荐算法的相关概念、技术难点; 其次对现有的跨领域推荐技术进行分类, 总结出各自的优点及不足; 最后对跨领域推荐算法的性能分析方法进行详尽的介绍.

关键词: 信息过载; 个性化; 跨领域推荐算法
中图分类号: TP181 **文献标志码**: A **DOI:** 10.3969/j.issn.1000-5641.2017.05.010

Techniques for cross-domain recommendation: A survey

CHEN Lei-hui[1], KUANG Jun[1], CHEN Hui[2], ZENG Wei[2],
ZHENG Jian-bing[1], GAO Ming[1]

(1. *School of Data Science and Engineering, East China Normal University, Shanghai* 200062, *China*;
2. *Shenzhen Tencent Computer System Co. Ltd., Beijing* 100080, *China*)

Abstract: With the rapid development of information technology and Internet, the available information on the Internet has overwhelmed the human processing capabilities in some commercial applications. Personalized recommendation system is a popular technology to deal with the information overload and recommendation algorithms are the core of it. In the past decades, collaborative filtering recommendation algorithm based on single domain has been widely used in many applications. However, the problems of cold start and data sparsity usually result in overfitting and fail to give desirable performance. The

收稿日期: 2017-06-20
基金项目: 国家重点研发计划(2016YFB1000905); 国家自然科学基金广东省联合重点项目(U1401256); 国家自然科学基金(61402177, 61672234, 61402180, 61502236, 61363005, 61472321)
第一作者: 陈雷慧, 女, 硕士研究生, 研究方向为用户行为分析、点击率预测. E-mail: 15720622991@163.com.
通信作者: 郑建兵, 男, 高级工程师, 研究方向为信息处理技术. E-mail: zhengjb@js.chinamobile.com.

cross-domain recommendation techniques have been a hot topic in the field of recommender systems, which aim to utilize knowledge from related domains to perform or improve recommendation in the target domain. This paper carries out a systematic study and analysis of cross-domain recommendation techniques. First, we summarize the related concepts and the technical difficulties of cross-domain recommendation algorithms. Second, we present a general categorization of cross-domain recommendation techniques and sum up their respective advantages and disadvantages. Finally we introduce the method of performance analysis of cross-domain recommendation algorithm in detail.

Key words: information overload; personalization; cross-domain recommendation algorithms

0 引 言

随着互联网和 web 2.0 技术的飞速发展, 网络上信息资源迅猛增长, 进而导致“信息过载”的问题愈发严重. 用户从海量的文本、视频、图像和商品等资源中找到符合自己个性化需求的信息变得十分困难. 个性化推荐系统是解决上述问题的关键技术之一. 与搜索引擎相比, 推荐系统能够通过对用户的历史行为数据的研究, 统计分析出用户的兴趣偏好, 从而引导用户发现自己的信息需求, 实现个性化推荐. 因此, 这一技术已被广泛地应用于电子商务、社交网络和视频网站等方面.

传统的个性化推荐系统都是基于单一领域的, 即根据用户对某一领域的偏好特征, 为用户提供该领域的推荐服务. 例如, YouTube 网站依据用户观看视频的历史记录给用户推荐他可能感兴趣的视频; Last.fm 网站根据用户对音乐所打的标签给用户推荐符合他兴趣的音乐. 迄今为止, 应用最为广泛的单领域推荐技术是协同过滤, 其核心思想是给目标用户推荐与他兴趣偏好最为相似的用户喜欢的项目, 或者与他曾经喜欢过的项目最为相似的项目. 然而, 随着用户规模和项目数量的急剧增长, 传统的协同过滤推荐算法的缺陷逐渐暴露出来, 特别是新用户、新项目和新系统的冷启动以及用户行为数据稀疏的问题, 这些问题致使协同过滤推荐性能降低, 妨碍算法的进一步推广. 不难发现, web 2.0 模式下的用户不仅仅是互联网信息的使用者, 更是信息的生产者. 用户在不同的社交媒体和电子商务网站中直接或间接地表达出自己不同角度的兴趣偏好. 研究表明, 来自于不同平台(社交媒体和电子商务网站等)的用户兴趣偏好或项目特征(属性、类别等)之间存在很强的关联性和依赖性[1]. 例如, 通常情况下, 喜欢阅读推理小说的用户更倾向于观看悬疑类电影, 而观看电影之后也会购买一些与电影相关的周边, 如 CD、明星同款商品等. 基于这一现象, 学术界和业界提出了跨领域推荐技术: 从其它领域中获取有效的用户偏好或项目特征的信息来丰富目标领域中的数据, 精准地预测用户行为, 提供更加合理和个性化的推荐服务. 例如, 给购买学习参考书的用户, 推荐相关视频教程、在线练习题等; 根据出行游客的旅游目的景点, 给他们推荐酒店、特色美食等. 概括来说, 成熟领域积累了大量的用户行为数据, 通过领域间信息资源的共享和互补, 不仅可以有效地缓解用户访问量少的推荐系统所面临的数据稀疏和冷启动问题, 而且可以提高用户满意度、改善用户体验. 但是, 不同领域数据的异构性、知识的独立性使得传统的单领域推荐算法无法直接应用于提供推荐服务. 针对这一问题, 学术界和业界开展了大量的研究和实践工作, 提出了很多跨领域推荐算法的模型和框架. 本文主要研究了跨领域推荐技术, 对其做了系统的分类, 并结合各自的特点进行了分析和总结.

本文结构安排如下: 第 1 节概述跨领域推荐算法的相关概念、技术难点; 第 2 节对现有的跨领域推荐技术进行分类, 总结出各自的优点及不足; 第 3 节详尽地介绍跨领域推荐算法的性能分析方法; 最后 1 节对全文进行总结并对未来的研究热点做出展望.

1　跨领域推荐系统概述

跨领域推荐旨在整合来自不同领域的用户偏好特征, 针对每个用户自身特点进行智能化感知, 精准满足用户个性化需求, 从而提高目标领域推荐结果的准确性和多样性. 与传统的单领域推荐系统相似, 跨领域推荐系统也有 3 个重要的模块(如图 1 所示): 用户建模模块、推荐对象建模模块和推荐算法模块. 两者区别在于给用户和待推荐对象建模时, 跨领域推荐利用的是融合多个辅助领域信息的数据而不仅仅是目标领域提供的数据; 而在进行推荐的时候, 它也可以根据提高准确性或多样性需求的不同, 来灵活地选定用户群体或待推荐对象[2].

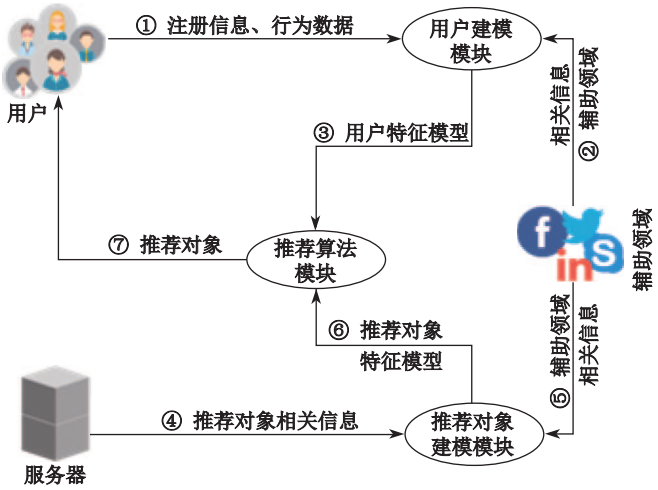

图 1　跨领域推荐系统流程

Fig. 1　The process for cross-domain recommendation

1.1　“域”的定义

学术界提出了多种关于“域”的定义. 例如, 文献 [3] 认为同一综合型网站上的图书和电影属于不同的领域; 文献 [4] 则将来自不同电影视频网站 (MoveLens, MoviePilot, Netflix) 的用户观看历史记录视为源自不同领域的用户行为数据. 据我们所知, 学术界和业界至今没有给出一个关于“域”的统一定义. 通过调研大量的相关研究工作[2,5-7], 本文将“域”分为三类: “系统域”、“概念域”和“时间域”.

• 系统域: 按照数据集所属的系统来划分. 例如, 豆瓣网站上的数据集和亚马逊网站上的数据集就分别属于不同的领域.

• 概念域: 将同一系统中的数据, 按照不同的概念层次进行划分. 例如, 题材层次(动作电影和喜剧电影为不同的领域)、对象层次(电影和图书为不同的领域, 即便在题材上有重复的地方).

• 时间域: 依据行为产生的时间对数据集进行域的划分. 例如 2017 年 1 月至 6 月的数据和 2017 年 7 月至 12 月的数据视为不同领域的数据.

总体说来, 前两种关于“域”的划分方式更为常见.

1.2 跨领域推荐的任务

本文用符号 $\mathcal{D}_T$ 表示目标领域、 $\mathcal{D}_S$ 表示源领域(或辅助领域). 源领域可由多个不同领域组成, 用于对目标领域 $\mathcal{D}_T$ 中的信息进行补充和丰富. 本文在结合I Fernández-Tobías 等人在文献 [2] 中所提出的 6 种跨领域推荐任务基础之上, 综合考虑实际应用需求, 将跨领域推荐的任务划分为以下 3 类:

• 缓解冷启动问题. 推荐系统需要根据用户的历史行为数据来预测用户对其他项目的偏好程度. 在面对新系统、新用户和新项目的时候, 会因为缺少用户行为数据而无法提供推荐服务. 利用从源领域中搜集到的用户偏好信息来预测用户的行为能够有效地弥补信息缺失的问题.

• 提高准确度. 个性化推荐系统中用户和项目数量都非常大, 但是大部分用户只会和一小部分的项目有交互, 这就导致用户项目评分矩阵十分稀疏, 降低推荐性能. 合理地应用源领域中的信息来增强目标领域评分矩阵的密集程度, 可以提高系统预测的精度.

• 增强多样性. 同一领域中的项目种类通常是单一的、相似的、冗余的, 并不能满足用户多样的兴趣需求. 将不同领域中的项目加入到待推荐对象中, 是提高推荐结果多样性的可靠方案.

1.3 跨领域推荐面临的挑战

跨领域推荐能够实施的一个关键性的假设是: 用户的兴趣偏好或项目特征在领域之间存在一致性或相关性. 这一假设也在一些研究工作[8-9]中得到佐证. 跨领域推荐利用的正是领域间的一致性或相关性, 如用户、项目的交集, 用户兴趣、项目特征的相似程度, 潜在因子的相互关系等进行知识迁移, 从而弥补目标领域所面临的信息不足的问题, 改善推荐性能. 同时跨领域推荐也是一个极具挑战性的研究领域, 其主要原因分析如下.

• 数据海量性: 海量数据是现今互联网应用的典型特征, 大多数推荐算法在海量数据场景下丧失优势, 因此简单、可扩展、可并行化等特点成为跨领域推荐算法的必备特征.

• 数据异构性: 不同领域具有不同的用户群体, 不同的推荐对象, 以及不同的用户行为数据结构, 譬如评分记录、购物列表和浏览日志等, 多源异构信息对象的融合是跨领域推荐所面临的最大挑战.

• 数据稀疏性: Power Law 是在社交网络普遍存在的一种现象. 简言之, 大部分用户只会和一小部分的项目有交互. 这就导致训练样本数据十分稀疏, 大大降低推荐模型的泛化能力. 而对大多数基于监督、半监督的学习模型而言, 它们往往是对训练数据集大小敏感的, 因此数据稀疏也就成为训练此类模型的一个特别棘手的问题.

• 数据相依性: 在实际生活中, 同一领域甚至不同领域中的用户的行为并不是互相独立的, 依据同质性原理(Homophily), 兴趣行为相似的用户偏向于喜好相似类型的项目, 如何挖掘和利用用户间隐藏的偏好关系成为一个难题.

• 数据低质性: 源领域中可获得的信息有用户注册信息、评分数据、浏览记录和点击情况等. 但是, 并不是所有的信息都有利于改善目标领域的推荐性能的. 不相关的信息如果被迁移进目标领域可能会成为“噪声”, 增加算法训练的复杂度, 降低推荐结果的准确性.

1.4 跨领域推荐的场景

在实际应用中, 不同领域间用户的重叠信息 (Overlapped Information) 对领域间信息资源或知识的共享和迁移起着至关重要的作用, 同时也是在设计跨域推荐方案时首先应

当考虑的问题. 按照用户重叠程度的不同可将跨域推荐的场景分为 3 类: 领域间用户完全重叠 (Fully-overlap)、领域间用户部分重叠 (Partially-overlap) 以及领域间用户完全不重叠 (Non-overlap). 之所以这么划分, 是因为领域间信息资源或知识共享和迁移的方式会随着有无用户交集而有所不同. 从图 2 可以看出, 当领域间用户完全重叠时, 可将两个领域合并, 从而轻易地将跨域推荐问题转换为单领域推荐; 当领域间用户部分重叠时, 这部分共享用户便成为领域间信息共享和迁移的桥梁; 当领域间用户完全不重叠时, 就需要通过挖掘领域间隐藏的共同用户或其他关系进行迁移学习. 当然, 领域间的项目也可能存在交集. 但用户和项目在推荐系统所担当的角色是对等的. 因此, 本文着重对领域间不同的用户重叠情况下的跨域推荐技术进行研究, 项目重叠情况下的推荐方案与其类似, 不做赘述.

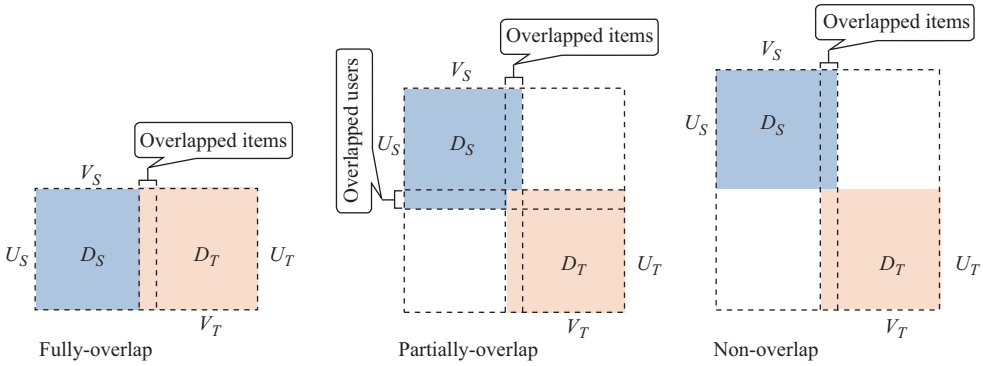

图 2　跨域推荐的3类场景

Fig. 2　Cross-domain recommendation scenarios

主流的跨域推荐算法有 3 类: 基于协同过滤关系的跨域推荐、基于语义关系的跨域推荐以及基于深度学习的跨域推荐. 其中, 协同过滤关系主要指用户或项目的近邻关系、隐语义模型等; 语义关系主要指项目属性、标签信息、语义网络关系和关联关系等[6]. 然而, 同一种方法在不同的跨域推荐场景下, 推荐性能不尽相同. 往往需要针对不同的推荐场景而采取不同的方案. 下面将依据不同的推荐场景来介绍跨域推荐技术.

2　领域间用户完全重叠的跨域推荐方法

现实生活中, 越来越多的网站呈现出向综合型的门户网站转变的趋势, 其所提供的推荐对象囊括了多个不同的领域. 例如, Amazon 除了提供图书购买外, 还有服饰、电子器件的销售; 著名的社区网站——豆瓣, 以书影音起家, 现在还提供线下同城活动, 小组话题交流等多种服务. 若从概念域角度来说, 书籍、电子器件和影音等便为不同领域中的项目, 而领域间的用户群体完全相同. 此时, 一种最直观的想法是将不同领域的用户行为数据整合为一个整体, 即一个更大的“单领域”, 从而将跨域推荐问题转化成单领域推荐问题.

2.1　基于协同过滤关系的跨域推荐

文献[10-11]均提出一种集中式的协同过滤模型. 如图 3 所示, 模型将来自不同领域的评分矩阵 (R_s, R_t) 合并为一个评分矩阵 R, 并采用单领域协同过滤模型进行个性化推荐, 譬如基于项目的协同过滤推荐、基于用户的协同过滤推荐. 这种方式优点在于简单, 便于单领域

推荐算法的直接应用. 然而, 实施评分矩阵合并的前提是不同领域遵循相同的评分机制.

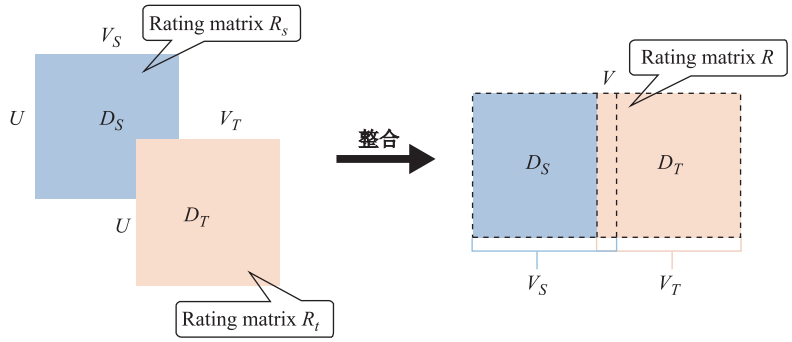

图 3 集中式的协同过滤模型

Fig. 3 Centralized collaborative filtering model

基于矩阵合并的跨域推荐方案的缺陷在于忽视了领域间的差异. 在某些情况下该方案并不能提高目标领域的推荐性能, 反而有可能引入“噪声”数据降低目标领域的预测精度. 为了克服这一缺点, 文献[12]提出了一种基于联合矩阵分解的跨领域推荐算法. 与传统的基于矩阵分解的单领域推荐算法相似, 均是通过最小化损失函数来获得两个特征矩阵: 用户特征向量矩阵 U 和项目特征向量矩阵 V, 最后再通过计算 UV^{T} 还原评分矩阵. 不同的是, 联合矩阵分解损失函数的构造是按照不同的权重系数将各领域矩阵分解的损失函数相加:

$$L(U,V^{(s)},V^{(t)})=\alpha||R_s-UV^{(s)\mathrm{T}}||_F^2+(1-\alpha)||R_t-UV^{(t)\mathrm{T}}||_F^2, \tag{1}$$

其中, U、$V^{(s)}$、$V^{(t)}$ 为模型的参数, 分别表示用户特征向量矩阵、源领域项目特征向量矩阵以及目标领域项目特征向量矩阵. 权重系数 α 控制两个领域中的损失函数在模型训练过程中所占的比重, 由反复试验来确定. 除了对两个评分矩阵进行联合训练外, 香港科技大学潘微科等人提出一种对二元信息矩阵(喜欢与不喜欢, 购买与不购买等)和评分矩阵进行联合分解的方案[13], 也有效地弥补了目标领域数据稀疏的问题, 但是该模型要求两个矩阵中的用户、项目必须严格一致. 文献 [14] 则同时对两个领域中的二元信息矩阵和用户评论信息进行联合建模, 得到用户特征向量; 并通过训练出两个非线性的映射函数, 一个用于将源领域中的用户偏好信息映射到目标领域中, 另一个则用于将源领域中的用户兴趣转换为目标领域中用户的兴趣. 相对来说, 这个模型更能够保留住领域间的独立性和差异性.

张量分解是近几年推荐系统的研究热点, 主要是通过在二维评分矩阵中加入一维或者多维信息, 如标签[15-16]、领域[17]等信息来获得更为全面的用户的偏好特征. 在评分矩阵中, 加入领域信息, 构造一个用户–项目–领域(user-item-domain)的三阶向量(如图 4 左图所示), 是一种较为新颖的用于解决领域间用户为共同维度的跨领域推荐问题的方法. 若将该三阶张量按正面切片的形式表示可发现每个切片正好是每个领域的评分矩阵 $R_d\in\mathbb{R}^{m\times n_d}(d=1,2,\cdots,n)$. 然而, 不同领域中项目数量的不同, 导致该三阶张量不是一个规则的“方块”. 相应的, 传统的基于“方块”的张量分解模型: CP 模型[18]、PARAFAC 模型[19]便不能直接应用.

文献 [17] 将基于PARAFAC2[20]的张量分解算法应用到跨域推荐中, 成功解决了上述问题. 该跨域推荐模型首先引入 n 个领域独立的矩阵 $P_d\in\mathbb{R}^{n_d\times n}$, 通过一个可逆的变换 $Y_d=R_dP_d$, $Y_d\in\mathbb{R}^{m\times n}$, 将不同领域的评分矩阵转变为具有相同维度 $m\times n$ 的信息矩阵, 从

而使得不规则的用户–项目–领域三阶张量成功地转换为规则的张量(如图 4 右图所示). 该模型张量分解的目标函数为:

$$L(U,V,C,P_k)=\frac{1}{2}\sum_{d=1}^{N}||w_d(R_dP_d-U\Sigma_dV^{\mathrm{T}})||_F^2+\frac{\lambda_U}{2}||U||_F^2+\frac{\lambda_V}{2}||V||_F^2+\frac{\lambda_C}{2}||C||_F^2, \quad (2)$$

其中, $\Sigma_d=\mathrm{diag}(C_d,.)$, w_d 为权重因子, 为了调整每个切片的损失值所占总体的比重. 最小化目标函数训练出模型参数 U,V,C,P_d 后, 通过计算 $U\Sigma_dV^{\mathrm{T}}$ 还原每个切片 Y_d, 最终由一个逆变换得到原来的评分矩阵. 权重因子的设置是求解和优化该模型最大的瓶颈, 文中采用神经网络模型自动找出最优的领域权重因子, 有效减少了人工设置权重参数的代价, 但一定程度上也加大了模型训练过程的复杂度. Song等人[21]认为相比较于评分信息, 用户的评论信息不仅能表达出用户对项目的喜好, 还能涵盖用户其他方面的兴趣偏好. 因此, 他们提出了一种基于评论信息的联合张量分解模型来进行跨域推荐. 该模型利用文献 [22] 所提出的 AIRS 方法评分信息进行训练, 从多个不同角度分析用户的评论, 得到用户在每一角度上的评分和关心程度. 以此作为输入构建用户–项目–角度 (user-item-aspect) 三阶张量, 通过源领域和目标领域共享特征向量实现知识迁移. 该模型可以较好地解决冷启动问题.

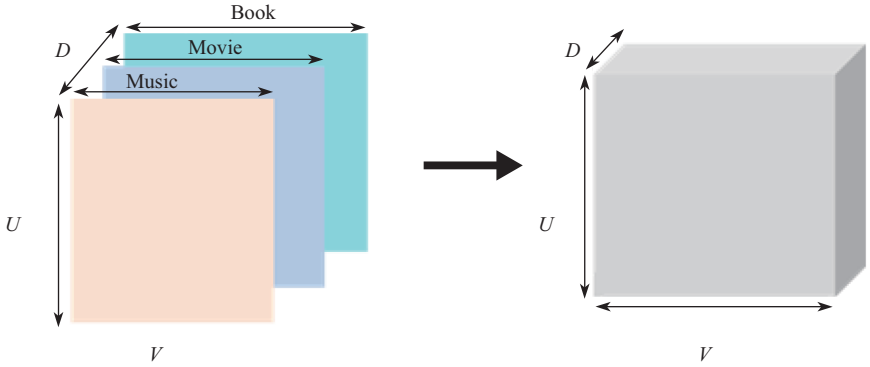

图 4　不规则的用户–项目–领域三阶张量转换为规则的张量

Fig. 4　Slices of rating matrices for each domain are transformed into a cubical tensor

2.2　基于语义关系的跨域推荐

在这一推荐场景下, 基于语义关系进行跨域推荐的研究相对较少, 主要为基于图模型的跨域推荐算法. 语义关系主要指项目属性、标签信息、语义网络关系和关联关系等, 图模型中会将上述的相关信息转换为边和权重. 2015 年, Jiang 等人提出跨域推荐模型[23]: 将不同的领域通过社交网络相互连接起来, 构成一个以社交网络为中心的星型结构的混合图 (star-structured hybrid graph). 对构建好的网络图采用 HRW(Hybrid Random Walk) 算法来预测用户与项目之间的关系. 特别地, 如图 5 所示, 除了考虑用户与各领域中项目的交互关系外(虚线表示), 每个领域中项目的语义关系(实线表示)也被用于知识的迁移. 这是解决领域间数据异构问题的一个行之有效的方法. 文献 [24] 提出利用标签体系解决异构问题, 成功实现了依据微博上的博文给用户推荐电影的跨域推荐服务. 其核心在于, 以用户博文上的标签和电影标签之间的语义关系为桥梁(如图 6 所示, 虚线表示用户–博文标签、电影–电影标签之间的关系, 实线表示语义关系), 将用户和电影关联起来, 组成一个多部图, 再基于图模型进行用户偏好预测. 这类模型在解决数据稀疏、冷启动以及领域间数据异构方面很有

优势.

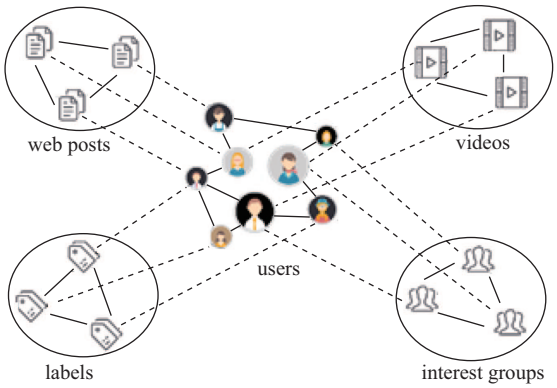

图 5 星型结构混合图

Fig. 5 A Star-structured hybrid graph

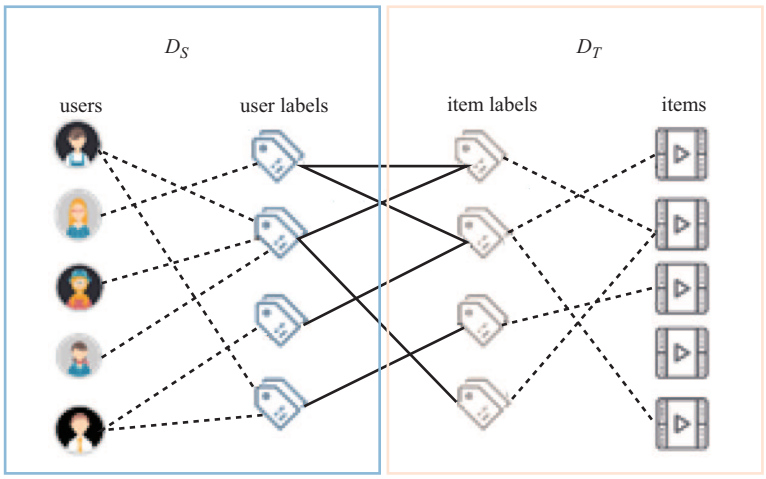

图 6 跨领域多部图

Fig. 6 A multi-partite graph across two domains

3 领域间用户不重叠的跨域推荐方法

隐私保护和商业竞争等原因使得跨域推荐算法的设计者难以获得不同领域中用户群体的重叠情况, 相应地就无法利用重叠的用户集作为领域间信息资源共享、迁移的桥梁. 解决这一场景下的跨域推荐问题有两个途径: ①采用用户匹配算法[25-27]挖掘出隐藏的重叠用户集, 将其转换为领域间用户有重叠的跨域推荐问题; ②基于协同过滤或语义关系进行知识迁移. 本节着重介绍途径二的相关技术, 途径一的相关算法可参照第 2 节 和第 4 节的内容.

3.1 基于协同过滤关系的跨域推荐

隐语义模型是隐含语义分析技术的一种, 也是推荐系统领域一个热门的研究话题. 其核心思想是通过聚类或矩阵分解等方法将稀疏高维的用户–项目矩阵映射到一个低维的隐空间(Latent Space)中, 找出潜在的主题或类别来表示用户的偏好和项目的特征从而能够以紧凑、简略的特征向量来表征用户、项目, 即用户、项目的隐语义模型. 那么, 在跨领域推荐

情境中, 自然可以想到将源领域中用户、项目的隐语义模型作为迁移学习对象, 来对目标领域中的用户、项目特征向量进行补充和增强. 然而, 用户、项目特征向量可以在领域间共享或融合的前提是用户、项目必须严格一致或存在很强的相似性. 因此, 如何有效地挖掘出领域间潜在的一致性关系或用户间的相似程度, 成为设计这类算法的核心问题.

文献 [28] 提出一种融合标签的协同过滤的跨领域推荐算法. 模型首先利用标签系统中丰富的、用户给项目所标注的标签信息, 计算出用户–用户的相似度矩阵 S^U 和项目–项目的相似度矩阵 S^V. 并将这一信息作为平滑项对概率矩阵分解模型 PMF[29]进行改进, 使得训练出的用户、项目特征向量在尽可能降低预测评分与实际评分误差的基础之上, 还能满足用户之间、项目之间的相似度关系. 模型的目标函数为:

$$\begin{aligned} L(U^{(S)},V^{(S)},U^{(T)},V^{(T)}) =& \frac{1}{2}\sum\nolimits_{d\in\{s,t\}}\sum\nolimits_{i=1}^{M_d}\sum\nolimits_{j=1}^{N_d} I_{ij}^{R^d}(R_{ij}^{(d)}-U_i^{(d)\mathrm{T}}V_j^{(d)})^2 \\ &+\frac{\alpha}{2}\sum\nolimits_{j=1}^{N_1}\sum\nolimits_{q=1}^{N_2} I_{jq}^{S^{(V)}}(S_{jq}^{(V)}-V_j^{(s)\mathrm{T}}V_q^{(t)})^2 \\ &+\frac{\beta}{2}\sum\nolimits_{i=1}^{M_1}\sum\nolimits_{p=1}^{M_2} I_{ip}^{S^{(U)}}(S_{ip}^{(U)}-U_i^{(s)\mathrm{T}}U_p^{(t)})^2 \\ &+\frac{\lambda}{2}\sum\nolimits_{d\in\{s,t\}}(||U^{(d)}||^2+||V^{(d)}||^2). \end{aligned} \tag{3}$$

其中, M_d、N_d、$R^{(d)}$、$U^{(d)}$、$V^{(d)}$ 分别代表领域 $d\in\{s,t\}$ 中的用户数量、项目数量、评分矩阵、用户特征向量矩阵、项目特征向量矩阵; I 为示性矩阵, 当对应的评分或相似度不为 0, 其值为 1, 否则为 0; α、β、λ 为模型训练参数. 该模型框架灵活, 对性能改善的效果明显. 但是领域间知识迁移完全依赖于用户、项目相似, 对用户、项目之间相似度的计算敏感.

除了利用用户打标签的行为挖掘领域间隐藏的关系外, Li 等人在 2010 年提出一种密码本迁移模型 (Codebook Transfer Model, CBT)[1]. 该模型从用户和项目两个角度对评分矩阵进行联合聚类, 发现来自不同领域的评分矩阵之间存在一个完全一致的用户–项目的聚级评分矩阵, 并将其形象地称之为“密码本”, 用于知识迁移. 具体做法如下: 首先通过正交非负三因式 (Orthogonal nonnegative matrix tri-factorization, ONMTF)[30]模型对源领域评分矩阵 R_S 进行分解得到两个特征向量矩阵 U_s、V_s, 然后利用公式

$$B=[U_s^{\mathrm{T}}R_sV_s]\oslash[U_s^{\mathrm{T}}\mathbf{1}\mathbf{1}^{\mathrm{T}}V_s], \tag{4}$$

求出“密码本”即矩阵 B, 其中符号 $\oslash$ 表示矩阵按元素相除. 最终通过最小化目标函数:

$$L=||[R_t-U_t^{\mathrm{T}}BV_t]\circ I||_F^2. \tag{5}$$

训练出目标领域中的用户、项目特征向量矩阵 U_t、V_t. 矩阵 I 是二值的示性矩阵. 在这个方法的基础之上, Li 等人又提出一个更为通用的模型——评分矩阵生成模型(Rating-Matrix Generative Model, RMGM)[8]. 该模型不再仅仅依赖于单个数据丰富的源领域, 而是将多个评分矩阵都合并到一起并同时对用户、项目两个维度进行共同聚类, 提取出聚级评分矩阵. 此外, 还学习出每个用户隶属于不同的用户聚类的概率分布, 每个项目隶属于不同项目组的概率分布, 以及每个聚类上的评分的概率分布, 至此, 评分矩阵生成模型就得到了. 当预测一个用户对项目的评分情况时, 首先按用户、项目隶属于用户组、项目组的概率分布情况找到用户项目聚类, 然后根据该聚类上的评分概率分布情况确定评分, 即为用户对项目的评分. 受这两种方法启发, 文献 [31] 在用户–项目二维评分矩阵中, 加入一维标签信息, 通过对从源

领域中用户、项目、标签同时聚类, 得到一个信息量更为丰富的簇级张量来缓解目标领域数据稀疏的问题.

上述 3 种方法的不足在于抹平了领域间的差异. 针对 CBT 模型, 文献 [6] 提出一种既考虑领域之间相同因素也考虑差异信息的跨域推荐算法对其进行改进. 算法将源领域和目标领域潜在聚级评分矩阵划分为共有部分 B_0 和本领域个性化部分 B_s, B_t, 求解的目标函数为:

$$L = |||[R_s - U_{\mathrm{s}}[B_0, B_s]V_s^{\mathrm{T}}] \circ I_s||_F^2 + |||[R_t - U_{\mathrm{t}}[B_0, B_t]V_t^{\mathrm{T}}] \circ I_t||_F^2. \tag{6}$$

通过不断地迭代更新, 求出最后的模型参数 B_0, B_s, B_t. 与其类似, 文献 [32] 提出的 PCLF(Probabilistic Cluster-level Latent Factor)模型以及文献 [33] 提出的 CLFM(Cluster-level Based Latent Factor Model) 模型也是同时训练出领域间共享的用户–项目聚级评分矩阵和各领域的个性化的特征矩阵来提高跨域推荐性能的. 而对 RMGM 模型的改进有 TALMUD(Transfer Learning for Multiple Domains)模型[34], 该模型通过对每个源领域都训练出一个互相独立的聚级评分矩阵并以不同的权重比例对目标领域进行数据补充, 以保留领域间的差异性和独立性. 基于评分聚类模型的最大缺陷在于缺少理论支撑. 只有在领域间具有很强的相关性的情况下, 才能起到改善目标领域推荐准确度的作用.

3.2 基于语义关系的跨域推荐

Chuang等人[35]基于项目属性交集提出了一种用于提高推荐结果多样性的模型: 将那些在项目属性上和用户历史购买的项目有交集的项目, 推荐给用户. 但实际上不同领域中项目的高度异构性导致项目间共同属性很少甚至没有.

因此, 有一些工作借助于社交网络中的标签信息, 来挖掘领域间用户、项目隐藏的关系. 其中一种方案是以Wikipedia[36]、WordNet[37]和情绪[38]分类体系为中间载体, 基于语义相似度、关联规则将不同领域中的标签映射到上述分类体系中, 构建由分类体系中的类别而构成的用户偏好特征, 从而获得更为精准的用户相似度信息. 另外一种方案是利用 LDA 主题模型[39]对用户所打的标签信息进行建模[40]构建出一个不同领域共享的用户特征 (user profile) 主题分布空间, 再基于这一空间找出不同领域中偏好相近的用户, 实施跨域推荐.

另外, 还有一些工作利用外部知识库 (Wikipedia, DBpedia) 构造语义网络, 来解决领域间数据异构问题. 文献 [41] 通过分析用户登录日志获取用户信息 (User profile) 和待推荐对象的文本信息 (Recommender context), 并将这两部分信息与 Wikipedia 的页面建立对应关系. 再利用 Wikipedia 页面间的链接信息 (Wikipedia hyperlinks), 构建语义关系网络. 最终基于马尔科夫模型获得用户到达每个待推荐对象的概率, 产生推荐结果. 文献[42-43]通过类似的方法构建语义网络, 实现了音乐和名胜古迹的跨域推荐. 此外, Benjamin Heitmann 等人[44]利用由 DBpedia 构建的知识图谱来连接不同的领域, 设计出一种即使在目标领域没有用户行为数据也能提供推荐服务的跨域推荐算法 SemStim.

4 领域间用户部分重叠的跨域推荐方法

文献 [45] 中提到不同领域中的用户集合完全重叠和完全不重叠是两种比较极端的情况, 现实生活中领域间的用户集合更多的是存在部分重叠. 关于这一点, 其实不难理解. 因为现在很多网站都会提供其他账号登录的入口, 从这一角度出发, 就能够找到不同领域中的同一用户. 此外文中通过实验证明了重叠的这一小部分用户其实在每个领域中都和超过 80% 的项目都有过交互行为. 利用这部分信息作为领域间信息共享和迁移的桥梁是可靠且有效的.

4.1 基于协同过滤关系的跨域推荐

Berkovsky 等人[11]提出一种启发式跨领域推荐算法: 首先利用源领域中的用户评分矩阵计算出用户的 K 近邻列表. 再依据重叠的用户将近邻信息导入到目标领域中以丰富用户模型. 与其类似的, Shapira 等人[46]用 Facebook 社交网络中的好友关系来增强目标领域中的用户模型. Tiroshi 等人[47]则进一步采用随机游走算法从社交网络中挖掘出更多隐含的用户近邻信息.

Jiang 等人[45]提出一种半监督的基于联合矩阵分解的迁移学习方法, 该模型认为在源领域中兴趣偏好相似的用户在目标领域中的兴趣偏好也应当相似. 最终矩阵分解的最小化目标函数为:

$$\begin{aligned} L = &\sum\nolimits_{ij} W_{i,j}^{(s)}(R_{i,j}^{(s)} - U_i^{(s)\mathrm{T}}V_j^{(s)})^2 + \lambda\sum\nolimits_{i,j} W_{i,j}^{(t)}(R_{i,j}^{(t)} - U_i^{(t)\mathrm{T}}V_j^{(t)})^2 \\ &+ \sum\nolimits_{i,j} W_{i_1,j_1}^{(s,t)}W_{i_2,j_2}^{(s,t)}(A_{i_1,i_2}^{(s)} - A_{j_1,j_2}^{(t)})^2. \end{aligned} \tag{7}$$

其中, λ 为经验参数, 通过实验获得. $W^{(s)}$、$W^{(t)}$ 为源领域和目标领域评分矩阵的二值示性矩阵, $W^{(s,t)}$ 为源领域和目标领域用户是否为同一用户的二值示性矩阵, 若用户 i 和用户 j 为同一用户, 则 $W_{i,j}^{(s,t)}$ 为 1, 否则为 0. $A^{(s)}$、$A^{(t)}$ 为源领域和目标领域中用户相似度矩阵, 计算公式如下:

$$A_{i_1,i_2}^{(s)} = U_{i_1}^{(s)\mathrm{T}}U_{i_2}^{(s)}, \quad A_{j_1,j_2}^{(t)} = V_{j_1}^{(t)\mathrm{T}}V_{j_2}^{(t)}. \tag{8}$$

上述的几种方法对于领域间的用户集合的交集大小十分敏感. 交集越大, 对目标领域推荐性能的改善越大; 反之, 交集越小, 对目标领域推荐性能的提升越不明显. 然而, 实际应用中, 能够直接被观测到的领域间的用户交集是很小的, 大部分的用户关系被隐藏起来. 为了充分利用领域间潜藏的用户、项目关系, 一些工作[48-49]基于共同的用户将两个领域连接成一个连通的图, 采用随机游走算法挖掘和利用领域间潜藏的关系进行迁移学习, 并取得了很好的推荐效果.

4.2 基于深度学习的跨域推荐

迄今为止, 深度学习在跨域推荐系统中的应用不是很广泛. 通常是被用于模型训练的某一过程. 例如, 用神经网络模型自动的找出最优的领域权重因子[17], 减少人工设置权重参数的代价; 或者基于语言模型训练用户、项目特征向量[50]: 将用户和项目的交互历史记录视为语言模型中的一个句子, 项目为语言模型的单词. 利用 word2vec 工具训练出源领域和目标领域中用户特征向量, 并以领域间重叠的用户为桥梁, 通过训练一个知识转移矩阵, 将源领域中的用户特征信息迁移进目标领域中.

5 各种跨域推荐技术的总结和对比

前面介绍了各种跨域推荐技术, 针对不同的推荐场景需要采用不同的用户行为预测模型. 不同的跨域推荐模型各自的优缺点不尽相同, 具体的比较如表 1 所示. 概括来说, 基于协同过滤关系的跨域推荐算法在 3 种推荐场景下, 都能取得较高的推荐质量. 尤其是将用户、项目的隐语义特征向量作为共享和迁移对象的方案, 框架灵活, 效果显著. 基于语义关系的跨域推荐算法, 是解决领域间数据异构问题的上策. 而基于深度学习的跨域推荐模型相对较少, 现处于初步研究阶段, 还有很多值得探索的方向, 存在很大的进步空间.

表 1 跨域推荐各模型的优点和缺点

Tab. 1 Advantages and disadvantages of different methods in cross-domain recommendation

跨域推荐场景	方法		优点	缺点
领域间用户完全重叠	基于协同过滤关系的方法	基于矩阵合并的跨域推荐[10-11]	简单, 能直接应用传统的协同过滤推荐算法	不同领域需要一致的评分机制, 忽略领域差异性
		基于联合矩阵分解的跨域推荐[12-14]	框架灵活, 效果显著	对权重参数敏感, 导致难以完全保留领域间的差异性
		基于张量分解的跨域推荐[15-17,21]	保留领域独立性特征, 对解决冷启动问题效果显著	必须构建出规则的多阶张量
	基于语义关系的方法	基于图模型的跨域推荐[23-24]	能够解决数据异构、数据稀疏、冷启动问题	需要学习大量的参数, 训练时间长
	基于深度学习的方法	/	/	/
领域间用户完全不重叠	基于协同过滤关系的方法	基于联合矩阵分解的跨域推荐[28-29]	框架灵活, 效果显著	对用户、项目之间相似度的计算敏感
		基于评分聚类的跨域推荐[1,6,8,32]	模型简单, 易于训练	缺少理论支撑, 领域间必须存在强相关
	基于语义关系的方法	基于标签分类体系的跨域推荐[36-38]	利用外部知识库最大程度上找出领域间偏好相似的用户	对标签信息和外部知识库要求高
		基于语义关系图的跨域推荐[41-44]	很好的解决数据异构和数据稀疏问题	对外部知识库敏感, 训练复杂
		基于LDA主题模型的跨域推荐[40]	在目标领域没有用户行为数据时也能有较好的推荐性能	对标签信息稀疏程度敏感, 影响模型准确度
	基于深度学习的方法	/	/	/
领域间部分重叠	基于协同过滤关系的方法	基于用户近邻的跨域推荐[11,46-47]	模型简单, 能够提供推荐解释	对领域间用户重叠程度敏感
		基于联合矩阵分解的跨域推荐[45]	框架灵活, 效果显著	
		基于图模型的跨域推荐[48−49]	能够解决数据异构、数据稀疏、冷启动问题	需要学习大量的参数, 训练时间长
	基于语义关系的方法	/	/	/
	基于深度学习的方法	基于Embedding技术的跨域推荐[50]	基于已有word2vec工具, 训练过程相对简单、稳定	处于初步研究阶段, 还有很多需要进一步探索

Network Embedding 技术是数据挖掘和机器学习领域中一项很重要的工作. 其核心思想是将大规模的网络降维到低维空间表示, 即用低维空间中的向量来表示网络中每个节点的特征, 如与其它节点的相互关系、在网络中的重要程度等. 从而能够基于每一个节点的特征向量来更高效、更精确地完成诸如分类 (classification)、连接预测 (link prediction) 以及推荐 (recommendation) 等任务. 近年来, Network Embedding 领域中涌现出大量基于深度学习的模型, 并在解决上述 3 种任务上取得了很好的效果. 譬如, 基于随机游走和神经网络来学习网络非线性结构的 DeepWalk 模型[51]和 Node2vec 模型[52]; 譬如, 基于节点的 first-order proximity 和 second-order proximity 获取网络局部结构和全局结构的 SDNE 模型[53]和 LINE 模型[54], 甚至有基于节点的 k-step proximity 的 GraRep 模型[55]; 譬如, 融合标签[56]和领域专家知识给出的节点间的相似度[57], 对 DeepWalk 结果进行修正的. 其实, Network Embedding 技术与推荐系统的中的隐语义模型本质上是相同的, 都是以特征向量来表征实体(节点、用户和项目)特征. 3 种跨域推荐场景下, 都能够轻松地构造出一个连接两个领域的网络图. 因此, 我们认为如何有效

地将 Network Embedding 领域中基于深度学习的技术应用于跨域推荐是个值得研究方向.

6　跨领域推荐算法的评测与分析

本节介绍评价和分析跨领域推荐算法性能的方法. 主要从实验方法、评测指标、数据集以及影响因素分析 4 个方面来阐述.

6.1　性能评测指标与方法

与传统的单领域推荐相似, 评测跨领域推荐算法性能的指标有: 准确度、覆盖度、多样性、新颖度、惊喜度和用户满意度等[58]. 从表 2 中可以看出, 对于跨域推荐算法性能评价集中在准确度这一指标上, 没有相关工作从覆盖率、多样性以及与用户体验相关的指标来分析跨域推荐算法的性能. 获得上述指标的实验方法主要有 3 种[59]: 离线实验 (offline experiment)、在线实验 (online experiment) 和用户调查(user study).

表 2　跨领域推荐算法性能评测指标

Tab. 2　Summary of metrics used for the evaluation of cross-domain recommendation

类别	度量指标		相关文献	实验方法
准确度	预测的精度指标	MAE	[1][4][8][9][10][11][12][13][15] [17][12][23][28][32][33][34][48]	离线实验、在线实验
		RMSE	[4][13][12][23][27][45]	
	分类的精度指标	Precision	[23][24][31][40][43][44]	
		Recall	[17][23][24][31][40][44]	
	排序的精度指标	MAP	[12][23][45]	
		AUC	[14][17]	
		F1-Measure	[10][23][40]	
		nDCG	[24]	
覆盖率	信息熵、基尼系数		/	
多样性	推荐列表中项目两两之间的不相似性		/	
用户体验	新颖度、惊喜度、用户满意度		/	在线实验、用户调查

离线实验是将处理好的数据集按照一定规则划分为训练数据集和测试数据集. 并在训练数据集上训练用户兴趣模型, 在测试数据集上进行预测. 整个实验过程都是在预先准备的数据集上完成, 不需要真实用户参加, 能够快速测试大量不同的算法. 但离线实验无法获得很多商业上关注的指标, 如转化率、点击率, 且其指标和商业指标存在一定的差距. 因此, 离线实验通常被用来批量验证多个推荐模型的性能优劣. 对于离线实验来说, 最重要的就是模拟出真实的在线推荐场景. 但现有的公开数据集中, 没有适用于跨域推荐的数据集. 究其原因是无法获取不同公开数据集间用户重叠的情况. 为了模拟出真实的跨域推荐场景, 通常是将某一公共数据集根据需求划分成一个个子集. 当然, 对于模拟领域间用户不存在交集的推荐场景, 就不会有这一问题.

由于离线测试的指标和实际的商业指标存在差距, 所以如果要准确地评测一个算法, 最好的方法是直接上线进行测试. 但在对用户满意度没有把握的情况下, 直接上线测试有一定的风险性. 为了降低风险, 企业会在上线测试之前做用户调查. 即安排一些用户在测试系统上行完成一些任务或回答一些问题, 并据此分析推荐系统的性能. 这样就能在降低在线实验风险的同时发现体现用户感受的指标. 但招募被测试者代价高.

最具代表性的在线实验的 AB 测试, 通过一定的策略将用户随机分成几组, 并对不同组的用户采用不同的算法, 然后通过统计不同的测评指标比较不同的算法. 其优点是能够公平的获得不同算法包括商业指标在内的实际在线性能指标, 但周期长.

6.2　影响因素分析

跨领域推荐算法的性能主要受 3 方面的因素影响: 源领域的信息密集程度、目标领域的信

息密集程度以及领域间的相关性. 因此, 在分析跨域推荐算法的性能的影响因素时, 往往会从这 3 个方面着手.

源领域的信息密集性一定程度上影响了被共享或迁移进目标领域的用户偏好信息及项目特征信息的准确性. 若源领域本身所包含的信息不足以训练出准确的用户、项目模型, 那便会成为训练目标领域推荐模型的噪声信息, 起到适得其反的作用. 但是仅有少量的工作对这一因素进行详尽的分析. 文献 [8,13,32-33,60] 通过改变的源领域评分数据集的大小, 观测这一因素对模型性能的影响; 文献 [34] 通过改变组成源领域中领域的数量, 分析源领域数据信息量对推荐性能的影响.

然而, 也有一些工作对于目标领域信息密集性的进行了系统的分析. 例如, 设定不同的大小的评分数据矩阵[1,8,10-11,17,32-33], 设定不同大小用户项目标签数量[28,60]来分析目标领域信息密集程度对跨域推荐性能的影响.

相对而言, 大部分跨领域推荐算法的研究工作, 集中在领域间相关性对推荐性能的影响上. 而领域间相关性可以从领域间用户交集、项目交集以及用户、项目的属性交集等方面来体现, 交集越大, 相关性越高. 文献 [3,45-46] 通过改变领域间用户交集的大小来观测性能变化; 文献 [3] 研究领域间项目的重合程度对于推荐性能的影响; 文献 [28] 从标签重合角度对这一因素进行分析. 此外, 还有一些工作[9,46-47]通过设置不同的源领域和目标领域, 来观测领域间的相关性对目标领域推荐性能的影响.

7 总结和展望

目前, 推荐的应用的场景越来越多, 如 Yahoo 的个性化广告显示, Google 的网页排名, OK Cupid 的在线约会等. 显而易见, 推荐系统已经成为计算广告、信息检索和社交网络分析等众多领域的核心技术之一. 而近 5 年, 国内的很多互联网公司先后成立了独立研发团队来研究跨领域推荐技术在工业上的运用. 如百度的“跨领域推荐”的搜索技术, 腾讯的基于腾讯云的搜索引擎等. 本文对跨领域推荐算法进行了系统地研究和分析, 概述了跨领域推荐算法的相关概念、技术难点; 对现有的跨领域推荐技术的进行了分类, 总结出各自的优点及不足; 最后对跨领域推荐算法的性能分析方法进行了详尽的介绍.

随着互联网、云计算、人工智能等领域的发展, 跨领域推荐算法也面临了一些新的研究问题, 这些问题也是未来的研究热点.

• 可扩展性: 现有的技术有各自特定的应用场景和算法的优势. 在不同的应用场景或数据集上往往表现出不同的结果. 因此, 设计具有可扩展性的推荐算法, 使其能够很好地应用于工业就显得尤为重要.

• 并行性: 单机已经不能满足对海量的用户行为数据和项目信息进行处理和分析的行业需求. 因此, 要考虑跨域推荐算法的并行化.

• 实时性: 推荐系统的精确度和实时性一直是一对矛盾. 因为数据量巨大, 所以大部分系统已经采用离线计算推荐的方式, 相应的推荐质量也会因此而打折扣. 因此就提高精确度的同时兼顾实时性是一个重要的研究问题.

• 可评测性: 惊喜度和新颖性这两个指标截至目前还没有什么标准的定义方式, 需要进一步研究; 此外, 仅仅从预测准确度来分析算法性能, 存在片面性, 还需要从覆盖率、多样性以及用户体验等角度来分析, 以获得更为全面的信息.

• 应用场景多元化: 将跨域推荐技术融入到可穿戴设备、智能家居的研究中, 以及与医疗、食品等领域相结合, 提供诸如健康生活建议、疾病预处理、个性化营养配餐等与人们生活休戚相关的服务也将会是未来研究的热点之一.

[参 考 文 献]

[1] LI B, YANG Q, XUE X. Can movies and books collaborate? Cross-domain collaborative filtering for sparsity reduction.[C]// Proceedings of the International Joint Conference on Artificial Intelligence.USA: DBLP, 2009: 2052-2057.

[2] CANTADOR I, FERNáNDEZ-TOBíAS I, BERKOVSKY S, et al. Cross-Domain Recommender Systems[M]// Recommender Systems Handbook. US: Springer, 2015: 919-959.

[3] ZHAO L, XIANG E W, XIANG E W, et al. Active transfer learning for cross-system recommendation[C]// Twenty-Seventh AAAI Conference on Artificial Intelligence. USA: AAAI Press, 2013: 1205-1211.

[4] PAN W, XIANG E W, YANG Q. Transfer learning in collaborative filtering with uncertain ratings[C]// Twenty-Sixth AAAI Conference on Artificial Intelligence. USA: AAAI Press, 2012: 662-668.

[5] LI B. Cross-domain collaborative filtering: A brief survey[C]// IEEE, International Conference on TOOLS with Artificial Intelligence. [S.l.]: IEEE Computer Society, 2011: 1085-1086.

[6] 罗浩. 基于跨域信息推荐的算法研究[D]. 北京: 北京邮电大学, 2014.

[7] FERNÁNDEZ-TOBÍAS I, CANTADOR I, KAMINSKAS M, et al. Cross-domain recommender systems: A survey of the State of the Art[C]//Proc 2nd Spanish Conf Inf Retrieval. [S.l.]: [S.n.], 2012: 187-198.

[8] LI B, YANG Q, XUE X. Transfer learning for collaborative filtering via a rating-matrix generative model[C]// International Conference on Machine Learning, ICML 2009. Canada: DBLP, 2009: 617-624.

[9] WINOTO P, TANG T. If you like the Devil Wears Prada the book, will you also enjoy the Dvil Wears Prada the movie? A study of cross-domain recommendations[J]. New Generation Computing, 2008, 26(3): 209-225.

[10] BERKOVSKY S, KUFLIK T, RICCI F. Mediation of user models for enhanced personalization in recommender systems[J]. User Modeling and User-Adapted Interaction, 2008, 18(3): 245-286.

[11] BERKOVSKY S, KUFLIK T, RICCI F. Cross-domain mediation in collaborative filtering[C]// User Modeling 2007, International Conference. Greece: DBLP, 2007: 355-359.

[12] SINGH, AJIT P, GORDON, et al. Relational learning via collective matrix factorization[J]. Relational Learning via Collective Matrix Factorization, 2008: 650-658.

[13] PAN W, YANG Q. Transfer learning in heterogeneous collaborative filtering domains[J]. Artificial Intelligence, 2013, 197(4): 39-55.

[14] XIN X, LIU Z, LIN C Y, et al. Cross-domain collaborative filtering with review text[C]// International Conference on Artificial Intelligence. USA: AAAI Press, 2015: 1827-1833.

[15] WEI C, HSU W, LEE M L. A unified framework for recommendations based on quaternary semantic analysis[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2011: 1023-1032.

[16] ARORA A, TANEJA V, PARASHAR S, et al. Cross-domain based event recommendation using tensor factorization [J]. Open Computer Science, 2016, 6(1): 32-37.

[17] HU L, CAO J, XU G, et al. Personalized recommendation via cross-domain triadic factorization[J]. Proc 22nd Int World Wide Web Conf, 2014: 595-606.

[18] ZHOU G, HE Z, ZHANG Y, et al. Canonical polyadic decomposition: From 3-way to N-way[C]// Eighth International Conference on Computational Intelligence and Security. [S.l.]: IEEE, 2012: 391-395.

[19] SCHMITZ S K, HASSELBACH P P, EBISCH B, et al. Application of parallel factor analysis (PARAFAC) to electrophysiological data.[J]. Front Neuroinform, 2014(8): 84.

[20] KIERS H A L. An alternating least squares algorithm for PARAFAC2 and three-way DEDICOM[J]. Computational Statistics & Data Analysis, 1993, 16(1): 103-118.

[21] SONG T, PENG Z, WANG S, et al. Review-based cross-domain recommendation through joint tensor factorization[C]// Database Systems for Advanced Applications. [S.l.]: DASFAA, 2017: 525-540.

[22] LI H, LIN R, HONG R, et al. Generative models for mining latent aspects and their ratings from short reviews[C]// 2015 IEEE International Conference on Data Mining. USA: IEEE, 2015: 241-250.

[23] JIANG M, CUI P, CHEN X, et al. Social recommendation with cross-domain transferable knowledge[J]. IEEE Transactions on Knowledge & Data Engineering, 2015, 27(11): 3084-3097.

[24] YANG D, HE J, QIN H, et al. A graph-based recommendation across heterogeneous domains[J]. 2016: 1075-1080.

[25] ZHANG J, YU P S. Multiple anonymized social networks alignment[C]// IEEE International Conference on Data Mining. [S.l.]: IEEE Computer Society, 2015: 599-608.

[26] KOUTRA D, TONG H, LUBENSKY D. BIG-ALIGN: Fast bipartite graph alignment[C]// IEEE International Conference on Data Mining. [S.l.]: IEEE, 2013: 389-398.

[27] LI C Y, LIN S D. Matching Users and Items Across Domains to Improve the Recommendation Quality [M]. New York: ACM, 2014: 801-810.

[28] SHI Y, LARSON M, HANJALIC A. Tags as bridges between domains: Improving recommendation with tag-induced cross-domain collaborative filtering[C]// User Modeling, Adaption and Personalization, International Conference. USA: DBLP, 2011: 305-316.

[29] SALAKHUTDINOV R, MNIH A. Probabilistic matrix factorization[C]// International Conference on Neural Information Processing Systems. USA: Curran Associates, 2007: 1257-1264.

[30] DING C, LI T, PENG W, et al. Orthogonal nonnegative matrix t-factorizations for clustering[J]. Proc 12th ACM SIGKDD, 2006: 126-135.

[31] CHEN W, HSU W, LEE M L. Making recommendations from multiple domains[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. USA: ACM, 2013: 892-900.

[32] REN S, GAO S, LIAO J, et al. Improving cross-domain recommendation through probabilistic cluster-level latent factor model[C]// Twenty-Ninth AAAI Conference on Artificial Intelligence. USA: AAAI Press, 2015: 4200-4201.

[33] GAO S, LUO H, CHEN D, et al. Cross-domain recommendation via cluster-level latent factor model[C]// Proceedings, Part II, of the European Conference on Machine Learning and Knowledge Discovery in Databases. New York: Springer-Verlag, 2013: 161-176.

[34] MORENO O, SHAPIRA B, ROKACH L, et al. TALMUD:transfer learning for multiple domains[C]// ACM International Conference on Information and Knowledge Management. New York: ACM, 2012: 425-434.

[35] CHUNG R, SUNDARAM D, SRINIVASAN A. Integrated personal recommender systems[C]// International Conference on Electronic Commerce: the Wireless World of Electronic Commerce. USA: DBLP, 2007: 65-74.

[36] SZOMSZOR M, ALANI H, CANTADOR I, et al. Semantic Modelling of User Interests Based on Cross-Folksonomy Analysis[M]. Germany: Springer Berlin Heidelberg, 2008: 632-648.

[37] ABEL F, HERDER E, HOUBEN G J, et al. Cross-system user modeling and personalization on the social web[J]. User Modeling and User-Adapted Interaction, 2013, 23(2-3): 169-209.

[38] FERNÁNDEZ-TOBíAS I, CANTADOR I, PLAZA L. An emotion dimensional model based on social tags: Crossing folksonomies and enhancing recommendations[J]. Lecture Notes in Business Information Processing, 2013, 152: 88-100.

[39] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[M]. J Mach Learn Res, 2003(3): 993-1022.

[40] KUMAR A, KUMAR N, HUSSAIN M, et al. Semantic clustering-based cross-domain recommendation[C]// Computational Intelligence and Data Mining. [S.l.]: IEEE, 2014: 137-141.

[41] LOIZOU A. How to recommend music to film buffs: Enabling the provision of recommendations from multiple domains[J]. University of Southampton, 2009.

[42] KAMINSKAS M, RICCI F. A generic semantic-based framework for cross-domain recommendation[C]// International Workshop on Information Heterogeneity and Fusion in Recommender Systems. New York: ACM, 2011: 25-32.

[43] KAMINSKAS M, FERNÁNDEZ-TOBÍAS I, CANTADOR I, et al. Ontology-Based Identification of Music for Places[M]// Information and Communication Technologies in Tourism. Germany: Springer Berlin Heidelberg, 2013: 436-447.

[44] HEITMANN B, HAYES C. SemStim at the LOD-RecSys 2014 Challenge[M]// Semantic Web Evaluation Challenge. Germany: Springer International Publishing, 2014: 170-175.

[45] JIANG M, CUI P, YUAN N J, et al. Little is much: bridging cross-platform behaviors through overlapped crowds[C]// Thirtieth AAAI Conference on Artificial Intelligence. USA: AAAI Press, 2016: 13-19.

[46] SHAPIRA B, ROKACH L, FREILIKHMAN S. Facebook single and cross domain data for recommendation systems[J]. User Modeling and User-Adapted Interaction, 2013, 23(2/3): 211-247.

[47] TIROSHI A, KUFLIK T. Domain Ranking for Cross Domain Collaborative Filtering[M]// User Modeling, Adaptation, and Personalization. Germany: Springer Berlin Heidelberg, 2012: 328-333.

[48] NAKATSUJI M, FUJIWARA Y, TANAKA A, et al. Recommendations over domain specific user graphs[C]// European Conference on Artificial Intelligence. USA: DBLP, 2010: 607-612.

[49] TIROSHI A, BERKOVSKY S, KAAFAR M A, et al. Cross social networks interests predictions based ongraph features[C]// ACM Conference on Recommender Systems. New York: ACM, 2013: 319-322.

[50] KRISHNAMURTHY B, PURI N, GOEL R. Learning vector-space representations of items for recommendations using word embedding models [J]. Procedia Computer Science, 2016, 80: 2205-2210.

[51] PEROZZI B, ALRFOU R, SKIENA S. Deepwalk: Online learning of social representations[C]//The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: ACM, 2014: 701-710.

[52] GROVER A, LESKOVEC J. node2vec: Scalable feature learning for networks[C]// ACM SIGKDD International Conference. New York: ACM, 2016: 855-864.

[53] WANG D, CUI P, ZHU W. Structural deep network embedding[C]// ACM SIGKDD International Conference. New York: ACM, 2016: 1225-1234.

(下转第 137 页)

[18] ESTER M, KRIEGEL H P, XU X. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise[C]// International Conference on Knowledge Discovery and Data Mining. USA:AAAI Press, 1996: 226-231.

[19] NG R T, HAN J. Efficient and effective clustering methods for spatial data mining[C]//International Conference on Very Large Data Bases. San Francisco: Margan Kaufmann, 1994: 144-155.

[20] ZHANG T, RAMAKRISHNAN R, LIVNY M. BIRCH: An efficient data clustering method for very large databases[J]. ACM SIGMOD Record, 1999, 25(2): 103-114.

[21] SUN C F, SHI Y L, LI Q I, et al. A hybrid approach for detecting fraudulent medical insurance claims: (Extended abstract)[C]//Proceedings of the 2016 Interational Conference on Autonomous) Agents & Multiagent Systems. Singapore: IFAAMS, 2016: 1287-1288.

[22] MOYANO L G, APPEL A P, SANTANA V F D, et al. GraPhys: Understanding health care insurance data through graph analytics[C]// International Conference Companion on World Wide Web. [S.l.]: International World Wide Web Conferences Steering Committee, 2016: 227-230.

[23] BAUDER R A, KHOSHGOFTAAR T M. A novel method for fraudulent medicare claims detection from expected payment deviations (Application Paper)[C]// IEEE, International Conference on Information Reuse and Integration. [S.l.]: IEEE, 2016: 11-19.

[24] 关皓文. 基于离群点检测方法的医保异常发现 [D]. 济南: 山东大学, 2016.

[25] HE Z, XU X, DENG S. Squeezer: An efficient algorithm for clustering categorical data[J]. Journal of Computer Science and Technology, 2002, 17(5): 611-624.

(责任编辑: 张　晶)

(上接第 116 页)

[54] TANG J, QU M, WANG M, et al. LINE: Large-scale information network embedding[C]//Proceedings of the 24th International Conference on World Wide Web. [S.l.]: International World Wide Web Conference Committee, 2015: 1067–1077.

[55] CAO S, LU W, XU Q. GraRep:Learning graph representations with global structural information[C]// Proceedings of the 24th ACM International Conference on Information and Knowledge Management. New York: ACM, 2015: 891–900.

[56] LI C, WANG S, YANG D, et al. PPNE: Property Preserving Network Embedding[C]//Database Systems for Advanced Applications,22nd International Conference. [S.l.]: DASFAA, 2017: 163–179.

[57] LI C, LI Z, WANG S, et al. Semi-supervised network embedding[C]//Database Systems for Advanced Applications, 22nd International Conference. [S.l.]: DASFAA, 2017: 131–147.

[58] 项亮.推荐系统实践[M]. 北京: 人民邮电出版社, 2012.

[59] SHANI G, GUNAWARDANA A. Evaluating Recommendation Systems[M]// Recommender Systems Handbook, 2011: 257-297.

[60] SAHEBI S, BRUSILOVSKY P. Cross-Domain Collaborative Recommendation in a Cold-Start Context: The Impact of User Profile Size on the Quality of Recommendation[M]. Germany: Springer,2013: 289-295.

(责任编辑: 张　晶)

(上接第 124 页)

[24] HERBRICH R, GRAEPEL T, OBERMAYER K. Large margin rank boundaries for ordinal regression[J]. Advances in Neural Information Processing Systems, 2000, 10(3): 115-132.

[25] CALVANESE D, DE GIACOMO G, LENZERINI M. A framework for ontology integration[C]//Proceedings of the First International Conference on Semantic Web Working. 2001: 303-316.

[26] ICTCLAS. [EB/OL]. [2017-02-01]. http://ictclas.nlpir.org/.

[27] SVM^{rank}. [EB/OL]. [2016-09-01]. https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html.

(责任编辑: 林　磊)