7月22日

跨域推荐论文笔记（CDR)

DTCDR（2019)

论文解读系列第一篇：CIKM-19 论文 DTCDR - 知乎 (zhihu.com)

1 核心思想

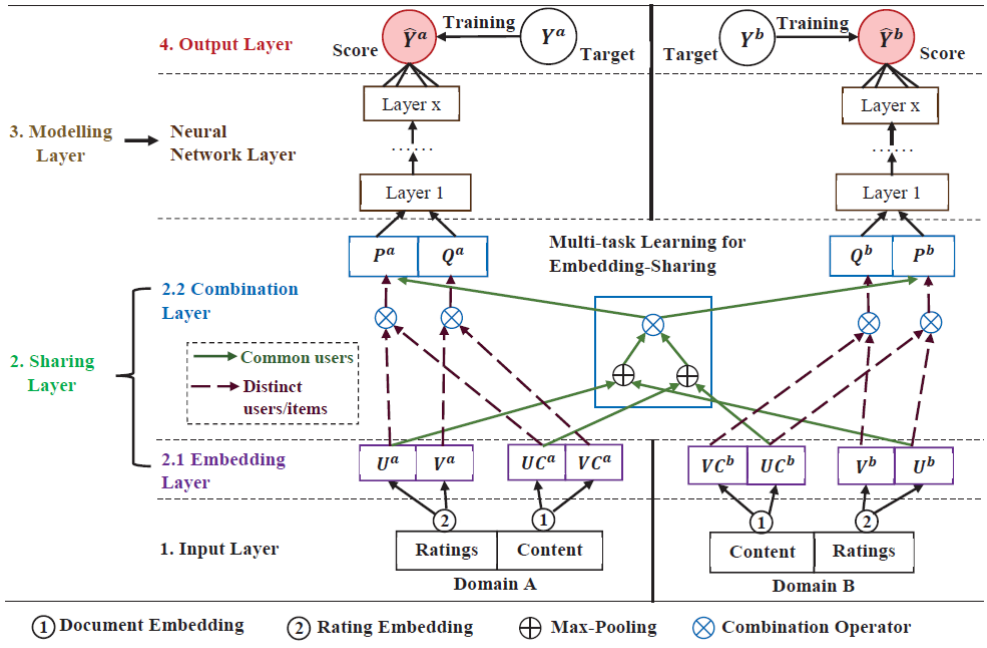


Figure 3: Our MTL-based solution for DTCDR framework.

- **传统的跨域推荐都是利用源领域（Source domain）的丰富数据来提升单个目标领域（Target domain）的推荐准确度**，即Source→Target。而源领域的推荐准确度没法直接通过现有的CDR方法来提升，因为没法直接改变现有的知识迁移方向，即没法从Source→Target变成Target→Source，否则会产生**“消极迁移”(Negative Transfer)的问题**。本论文中以两个域的共同用户或者共同商品作为桥梁，来实现双向的知识迁移，从而同时改善两个领域的推荐准确度。
- 嵌入层中，本论文除了常用的评分信息以外，还利用了其它多源的文本信息，如评论文本，标签，用户简介，商品详情。该论文使用**Doc2Vec模型来处理这些文本信息**，获取**用户和物品的文本表示**（document embedding）。此外，该论文**改进了NeuMF和DMF两个模型**，并分别获取**用户和物品的评分表示**（rating embedding）。
- 合并层中，使用**max-pooling策略分别合成共同用户分别来至领域A和B的文本以及评分信息**。接着使用**三种不同的合并策略**（Combination operators，具体的是指**average-pooling**，**max-pooling**，以及**Concatenation**）**来合成文本以及评分信息**。通过嵌入层和合并层，优化了领域A和B中共同用户的特征表示，并同两个领域分别的物品特征表示一起作为模型层的输入。在模型

层，使用了全连接的多层感知器（MLP）来学习用户和物品之间的非线性关系，并最终得到喜好预测（Score）给输出层。

DDTCDR（WSDM-2020）

[DDTCDR: Deep Dual Transfer Cross Domain Recommendation (aisoutu.com)]

1 核心思想

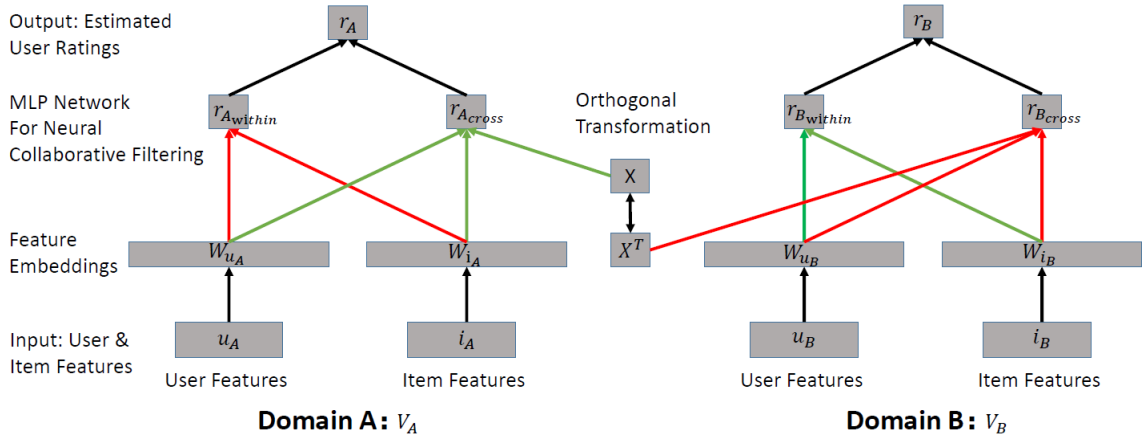


Figure 1: Model Framework: Red and blue lines represent the recommendation model for domain A and B respectively. We obtain the estimated ratings by taking the linear combination of within-domain and cross-domain user preferences and back-propogate the loss to update the two models and orthogonal mappings simultaneously.

- 以前提出的跨领域模型**没有考虑到用户和项目之间的双向潜在关系。**此外，它们**不明确地建模用户和商品特征的信息，而仅利用用户评级信息进行推荐。**本论文采用隐性嵌入方法，可以从数据记录中提取潜在的用户偏好，并有效地建模用户和项目特征。
- 作者建议使用两个组件来建模用户偏好：捕获用户交互并预测用户在目标域中的行为的域内偏好和利用源域用户动作的跨域偏好。**引入了传递率α作为超参数，它代表了在预测用户偏好时两种成分的相对重要性。**在域对(A, B)中估计用户评分如下：

$$r'_A = (1-\alpha)RS_A(W_{u_A}, W_{i_A}) + \alpha RS_B(X * W_{u_A}, W_{i_A}) \qquad (2)$$

$$r'_B = (1-\alpha)RS_B(W_{u_B}, W_{i_B}) + \alpha RS_A(X^T * W_{u_B}, W_{i_B}) \qquad (3)$$

- 利用深度双迁移学习机制实现用户偏好的双向迁移，该算法学习了两个域的潜在正交映射函数，既能保留用户偏好的相似性，又能有效地计算出反向映射函数。

Algorithm 1 Dual Neural Collaborative Filtering

Input: Domain V_A and V_B, autoencoder AE_A and AE_B, transfer rate α, learning rates γ_A and γ_B, initial recommendation models RS_A and RS_B, initial mapping function X
repeat
  Sample user-item records d_A and d_B from V_A and V_B respectively
  Unpack records d_A, d_B as user features u_A, u_B, item features i_A, i_B and ratings r_A, r_B
  Generate feature embeddings from autoencoder as $W_{u_A} = AE_A(u_A)$, $W_{u_B} = AE_B(u_B)$, $W_{i_A} = AE_A(i_A)$, $W_{i_B} = AE_B(i_B)$
  Estimate the ratings in domain A via $r'_A = (1-\alpha)RS_A(W_{u_A}, W_{i_A}) + \alpha RS_B(X * W_{u_A}, W_{i_A})$
  Estimate the ratings in domain B via $r'_B = (1-\alpha)RS_B(W_{u_B}, W_{i_B}) + \alpha RS_A(X^T * W_{u_B}, W_{i_B})$
  Compute MSE loss $\hat{r_A} = r_A - r'_A$, $\hat{r_B} = r_B - r'_B$
  Backpropogate $\hat{r_A}$, $\hat{r_B}$ and update RS_A, RS_B;
  Backpropogate orthogonal constraint on X; Orthogonalize X
until convergence

GA-DTCDR（IJCAI2020）

1 核心思想

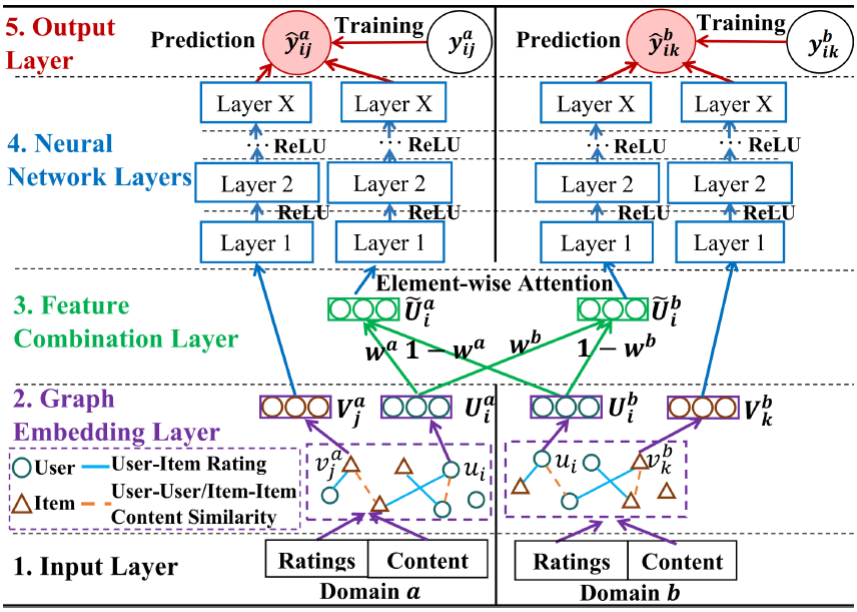


Figure 1: The overview of GA-DTCDR

- **此前的最好方法往往只考虑了用户-商品间的联系，而忽略了用户-用户以及商品-商品间的联系。**本论文通过构建异构图（heterogeneous graph）考虑了用户-商品间的联系（基于评分），还考虑了用户-用户以及商品-商品间的联系（基于内容相似性），然后通过 graph embedding technique（Node2vec）利用异构图生成相应的 user embedding matrix U 和 item embedding matrix V，这样能够更准确地捕捉到用户和物品的特征。
- **在融合共同用户/物品的embeddings时，此前的最好方法使用固定的合成策略，比如 average-pooling， max-pooling,，and concatenation**，本论文使用 element-wise attention networks 替代之前固定的合成策略，通过自注意力机制让机器自动学习到如何从本域的user以及common user中提取特征，也就是利用其他域的user特征来帮助生成本域的user特征。

2 小知识点

- [NDCG Normalized discounted cumulative gain 理解分析_Xiangyong58的专栏-CSDN博客]

$$N(n) = Z_n \sum_{j=1}^{n} (2^{r(j)} - 1) / \log(1 + j)$$

Normalization　Cumulating　Gain　Position discount

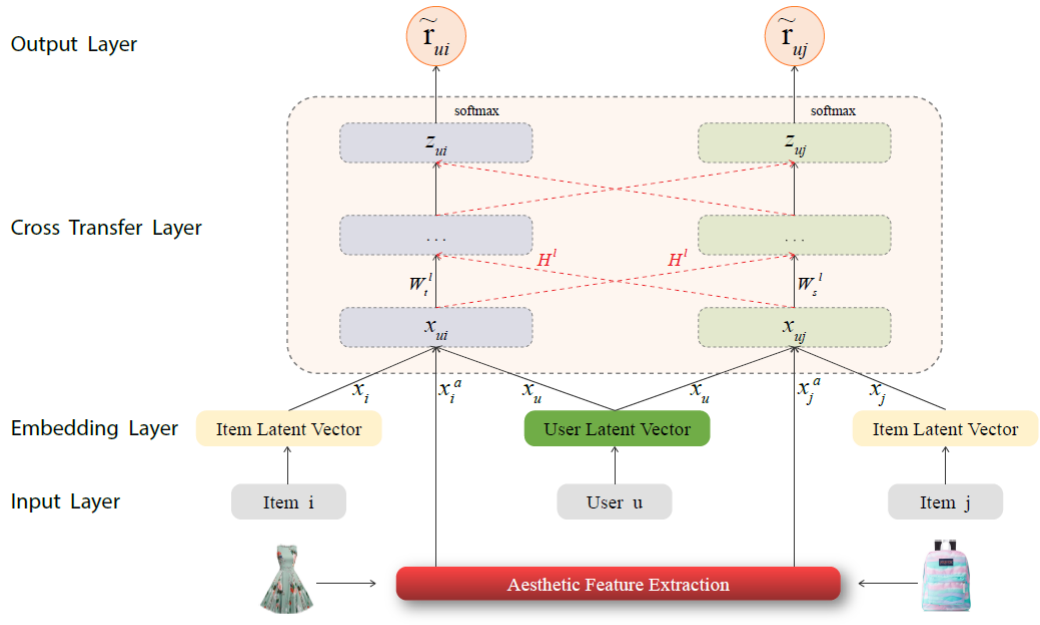
首先，计算NDCG，需要计算Gain，这个gain即是每条结果的质量的定义，NDCG把所有结果相加最终相加保证，整体质量越高的列表NDCG值越大。同时，Discounted的设计使得越靠前的结果权重越大，这保证了第一条，更相关的排在靠前的结果会有更大的NDCG值。从这两点看，**以NDCG为优化目标，保证了搜索引擎在返回结果总体质量好的情况下，把更高质量结果排在更前面。**

- [(92 封私信 / 80 条消息) 什么是 ablation study？ - 知乎 (zhihu.com)]

ablation study 就是你在同时提出多个思路提升某个模型的时候，为了验证这几个思路分别都是有效的，做的控制变量实验的工作。

ACDN（WWW-2020）

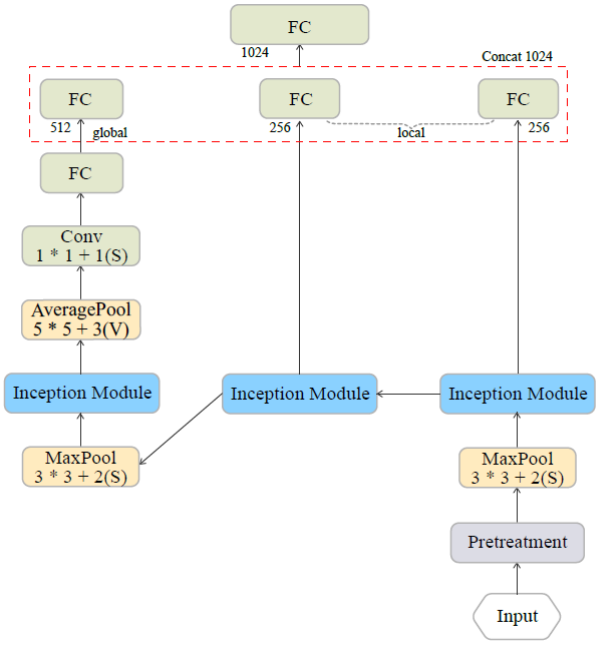
1 核心思想



(a) The proposed deep aesthetic preference cross-domain network architecture.

- 对于衣服鞋子这种仅商品图片就包含了大量信息的商品来说，商品图片的特征是十分重要的。而用户的审美偏好往往是独立于领域的，比如一个喜欢hiphop风格的顾客，在购买衣服，鞋子，项链等商品时往往也会选择更偏向hiphop风格的产品。因此，**视觉信息可以在提高外观优先推荐产品的性能方面发挥重要作用**，本论文通过捕获用户独立于领域的审美偏好，利用新颖的审美特征进行跨领域推荐。

- 使用ILGNet来提取产品图片的**局部以及宏观审美特征**：



(b) The aesthetic network (ILGNet) architecture.

2 小知识点

- word2vec -- 负采样 -- skip-gram - 简书 (jianshu.com)

  negative sampling：不同于原本每个训练样本更新所有的权重，负采样每次让一个训练样本仅仅更新一部分的权重，这样就好降低梯度下降过程中的计算量。

- (1条消息) leave-one-out之个人理解西红是番茄-*CSDN*博客leave-one-out

  LOOCV - Leave-One-Out-Cross-Validation 留一交叉验证_很吵请安静-CSDN博客

  leave-one-out：留一法交叉验证是一种用来训练和测试分类器的方法，会用到图像数据集里所有的数据，假定数据集有N个样本（N1、N2、...Nn），将这个样本分为两份，第一份N-1个样本用来训练分类器，另一份1个样本用来测试，如此从N1到Nn迭代N次，所有的样本里所有对象都经历了测试和训练。

最新综述（IJCAI-2021）

论文解读系列第十六篇：IJCAI 2021--跨域推荐（Cross-Domain Recommendation）的最新综述 - 知乎 (zhihu.com)

1 背景介绍

- 基于**协同过滤技术（Collaborative Filtering，CF）**(3条消息) 个性化智能推荐(协同过滤算法)技术研究_zolalad的专栏-CSDN博客的推荐系统，都多多少少受到数据稀疏的影响
- 特别对于新用户或者新产品**（Cold-Start，冷启动问题）**来说，由于在系统内还没有产生任何的交互信息（评分、评论等等），推荐的精度会比较低

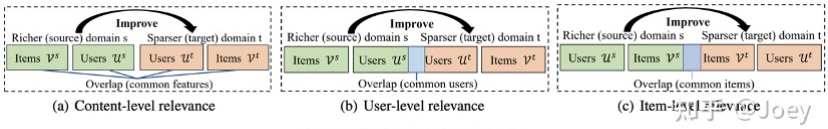
- 跨域推荐的提出就是为了解决这个数据稀疏问题。基本思路是利用**丰富领域（richer domain又称为source domain）**的较为丰富训练数据来提升**稀疏领域（sparser domain又称为target domain）**的推荐精度。
- 领域（domain）的定义：
  - **内容层级相关性（content-level relevance）**：两个或多个领域中，用户之间或产品之间存在共同的内容或者特征（例如关键字、标签）。但是这些领域不存在共同的用户或者产品。例如：亚马逊音乐（Amazon music，音乐相关）和奈飞（Netflix，电影相关）。
  - **用户层级相关性（user-level relevance）**：两个或多个领域中，存在共同用户但是产品层级不同。产品层级的不同又可以细分为属性层级（attribute-level）的不同（即类型相同（例如图书）但是属性不同，例如教科书、小说、自传等等）和类型层级（type-level）的不同（例如图书、电影、音乐、服装等等）。
  - **产品层级相关性（item-level relevance）**：两个或多个领域中，存在共同产品但是用户不同。例如：MovieLens和奈飞，都是电影相关的系统，存在大量相同的电影，但是用户不同，或者很难识别用户一致性。这种类型在一些文献里又被称为cross-system recommendation（跨系统推荐）。

2 综述动机

- 近些年来，在跨域推荐领域出现了一些新的或愈发明显的挑战，例如：
  - 特征映射问题
  - 嵌入（embedding）优化问题
  - 负面迁移问题（negative transfer）
- 随着应用场景的拓展，跨域推荐也出现了一些新的方向：
  - 双目标跨域推荐（dual-target CDR）
  - 多目标跨域推荐（multi-target CDR）

3 不同的跨域推荐场景以及挑战

- **场景1. 单目标跨域推荐（single-target CDR）：**传统的跨域推荐主要是针对的单目标跨域推荐（single-target CDR），即利用丰富领域（源领域，source domain）来提升稀疏领域（目标领域，target domain）的推荐精度。根据上面'领域'的不同定义，单目标跨域推荐可以细分为以下三个应用场景：



(a) Content-level relevance (b) User-level relevance (c) Item-level relevance

Figure 1: Single-target CDR scenarios

- 这个场景下，研究人员将面临如下三个挑战：
  - 构建基于内容的关系
  - 生成准确的用户/产品嵌入表达（embeddings）或者评价模式（rating patterns）
  - 学习准确的映射关系：领域之间embedding或rating pattern的迁移对应关系。
- **场景2. 多领域推荐（Multi-Domain Recommendation）：**在单目标跨域推荐中，有一个研究分支，即多领域推荐（Multi-Domain Recommendation）。这个分支主要针对来之多个领域的特定用户集中用户，推荐来之多个领域的特定产品，本质上这也是single-target CDR，因为推荐的目标是特定一个集合（用户集合）。因此，本综述认为这个研究方向还是single-target CDR的一个分支。

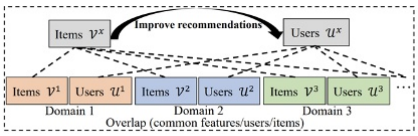


Figure 2: MDR scenario

知乎 @Joey

- **场景3. 双目标跨域推荐（dual-target CDR）：** 这是一个新的跨域推荐场景，即同时利用两个领域的数据来同时提升两个领域的推荐精度。这个场景面临以下两个新的挑战：
  - 构建一个可行的双目标跨域推荐的框架
  - 各领域之间的对应关系，优化用户/产品的嵌入表达（embeddings）
- **场景4. 多目标跨域推荐（Multi-target CDR）：** 这也是一个新的跨域推荐场景，即同时利用多个领域的数据来同时提升多个领域的推荐精度。这个场景面临以下挑战：
  - 避免负面迁移（negative transfer），这个问题随着越来越多领域的加入会愈加严重。

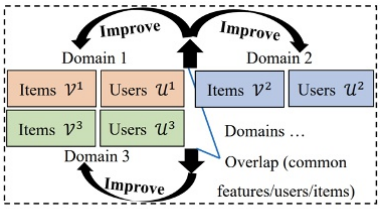


Figure 4: Multi-target CDR scenario

知乎 @Joey

4 相关的研究进展

- 现有的跨域推荐方法主要划分如下图：

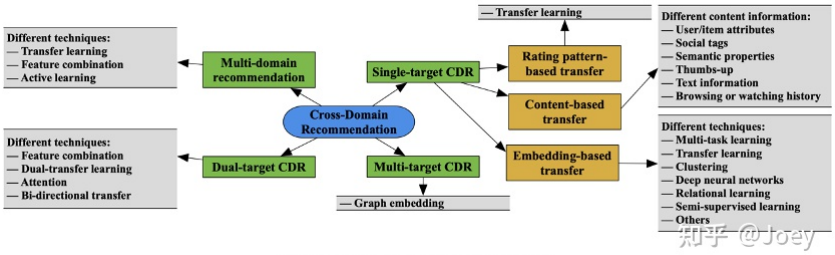


Figure 5: A categorization of CDR approaches

知乎 @Joey

5 可用数据集合

Table 3: Summary of datasets for CDR

Datasets	Domains	Data types	Scale	Website
Arnetminer [Tang *et al.*, 2012]	Research domains (user-level relevance — attribute-level)	Paper & author & conference name ...	1 million	https://www.aminer.org/collaboration
MovieLens + Netflix [Zhu *et al.*, 2018]	Movie (item-level relevance)	Rating & tag	25 million & 100 million	https://grouplens.org/datasets/movielens/ https://www.kaggle.com/netflix-inc/netflix-prize-data
Amazon [Fu *et al.*, 2019]	Book & music & movie ... (user-level relevance — type-level)	Rating & review & side information	100 million+	http://jmcauley.ucsd.edu/data/amazon/
Douban [Zhu *et al.*, 2018] [Zhu *et al.*, 2019; Zhu *et al.*, 2020]	Book & music & movie (user-level relevance — type-level)	Rating & review & side information	1 million+	https://github.com/FengZhu-Joey/GA-DTCDR/tree/main/Data

知乎 @Joey

6 未来研究方向

- 异质化跨域推荐：现有跨域推荐的假设前提是跨域的信息是同质的，但是实际应用场景中，存在跨域的异质信息。
- 序列化跨域推荐：跨域推荐系统也和传统推荐系统一样，面临如何序列化地给用户/产品建模。
- 隐私保护的跨域推荐：现有的跨域推荐方法忽略了信息孤岛的问题。而实际应用场景，用户敏感信息是无法直接跨域分享的。
- 另外，数据集稀疏程度、领域间的重叠规模（overlap scale）、以及领域间的关联度，如何分别影响跨域推荐的性能？这些问题同样也值得进一步研究。

7 结论

近些年，随着深度神经网络以及图学习的发展，跨域推荐越来越受工业界、学术界的关注。本篇综述分别从单目标跨域推荐、多领域推荐、双目标跨域推荐、以及多目标跨域推荐来系统性地分析、总结现有跨域推荐方法。综述的最后，给出了三个未来值得研究方向以供读者参考，并希望这些方向能得到进一步的研究，最终解决困恼推荐领域多年的问题—数据稀疏问题。