Auxiliary Domain Selection in Cross-Domain Collaborative Filtering

Reporter--- YiChang

Content

研究背景

相关工作

研究方案

结果分析

研究总结

Background

Recommender System

- 主要思想：过去与目标用户具有相似偏好的用户，将来也可能有类似的偏好。因此，根据目标用户的最近邻居对目标商品的评分，可以计算出目标用户的预测评分。
- 主要方法：协同过滤、基于内容、基于网络结构、混合推荐、UCF/ICF/DF/Hits
- 遇到问题：cold start

Cross-domain Recommender System

- transfer learning
- 方法划分
  - 领域数据划分
  - 迁移方式

Related Work

按领域数据划分

1. 系统领域

跨电子商务系统用户行为的交叉推荐算法，利用了不同电商网站的共同用户行为，把交叉用户（同时在两个以上电商网站有浏览、购买等行为的用户）在不同电商网站之间的行为作为桥梁来连接不同的电商网站，从而实现不同电商网站间的交叉推荐[1]。

2. 数据类型领域(0/1, 1~5)

一些辅助域的数据类型（比如，链接行为数据0/1，或者浏览行为数据0/1）较目标域的数据类型（比如1~5的用户评分）更容易获得[2]。

3. 时间领域

通过将打分行为的时间跨度分割为更小的时间单位，可将数据领域划分成多个域。能够更好地捕获用户行为的动态变化[3]。

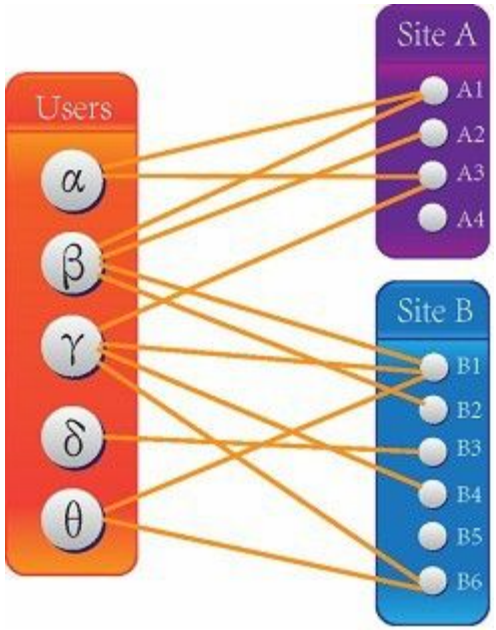

图1 用户商品二部分图

[1] 张亮，柏林森，周涛.基于跨电商行为的交叉推荐算法[J].电子科大学报.2012
[2] Pan W, Xiang E W, Liu N N, et al. Transfer Learning in Collaborative Filtering for Sparsity Reduction[C]. AAAI. 2010, 10: 230-235.
[3] B. Li, X. Zhu, R. Li, C. Zhang, X. Xue, and X. Wu, "Crossdomain collaborative filtering over time," in IJCAI, 2011, pp. 2293–2298.

Related Work

按迁移方式划分

1. 共享打分模式

Codebook Transfer（CBT）

主要是从辅助域系统中学习出一种用户打分模式codebook。这种用户打分模式是辅助域和目标域共同的“知识”。然后将该“知识”迁移到目标域中，帮助目标域完成推荐。

2. 基于主题学习模型

Cross-domain Topic Learning

首先从源域和辅助域中分别抽取主题模型。然后为源域和辅助域的主题分布，以及源域和辅助域之间主题的关系进行建模。再通过迁移源域和目标域中共同主题的分布“知识”，进而对目标域推荐[5]。

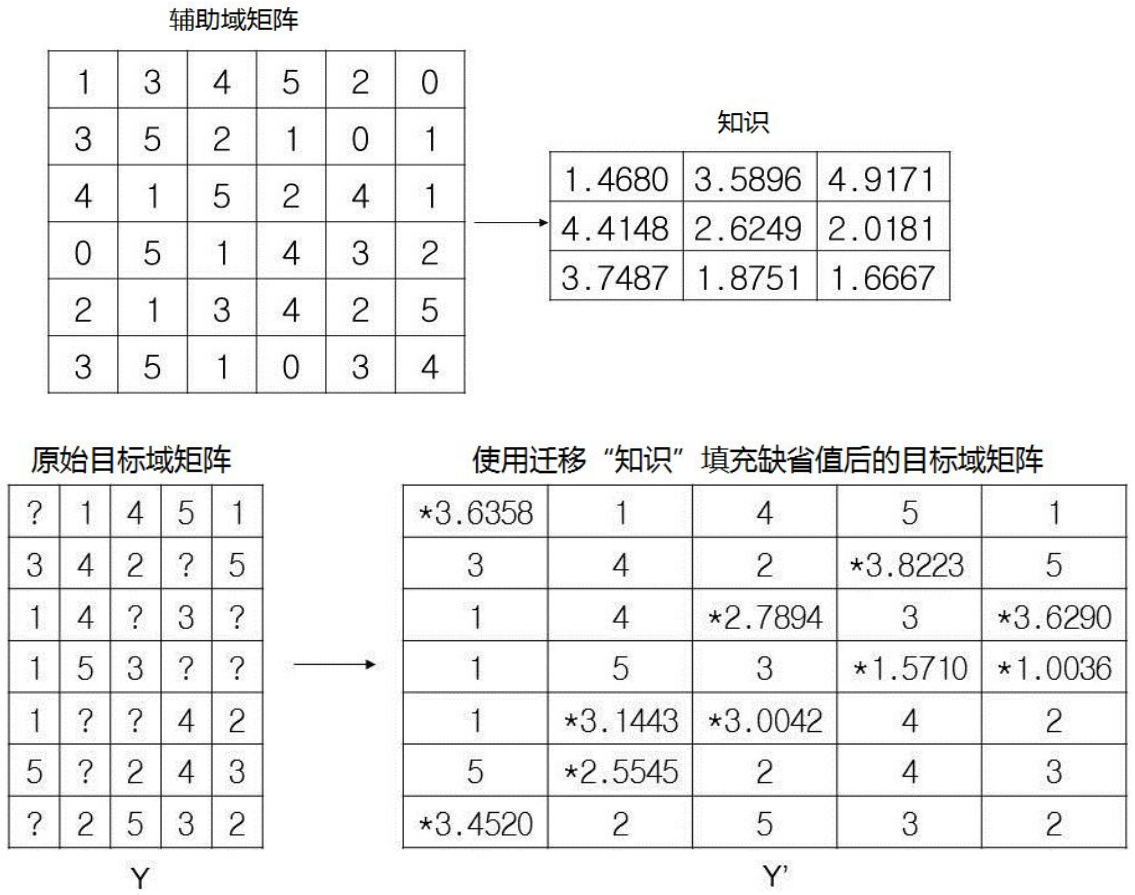

图2 获取辅助域的“知识”并迁移“知识”填充目标域数据

[4] Bin Li，Qiang Yang，Xiangyang Xue. Can Movies and Books Collaborate? Cross-Domain Collaborative Filtering for Sparsity Reduction[C]. Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence，2009.

Related Work

◆迁移方式

1. 共享打分模式

   Codebook Transfer（CBT）

2. 基于主题学习模型

   Cross-domain Topic Learning

首先从源域和辅助域中分别抽取主题模型。然后为源域和辅助域的主题分布，以及源域和辅助域之间主题的关系进行建模。再通过迁移源域和目标域中共同主题的分布“知识”，进而对目标域推荐[5]。

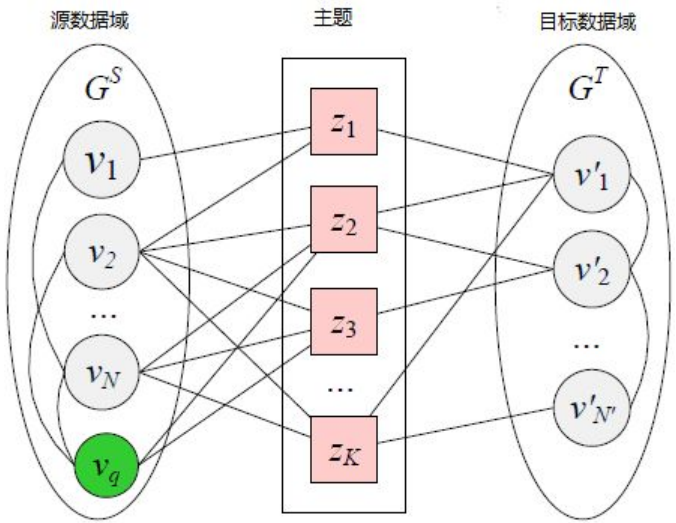

图3 跨域主学习模型

[5]Jie Tang，Sen Wu，Jimeng Sun，et al. Cross-domain Collaboration Recommendation[C]. KDD’12, August 12–16 , Beijing, China, 2012

RMGM Algorithm

- 一种基于迁移学习的RMGM算法

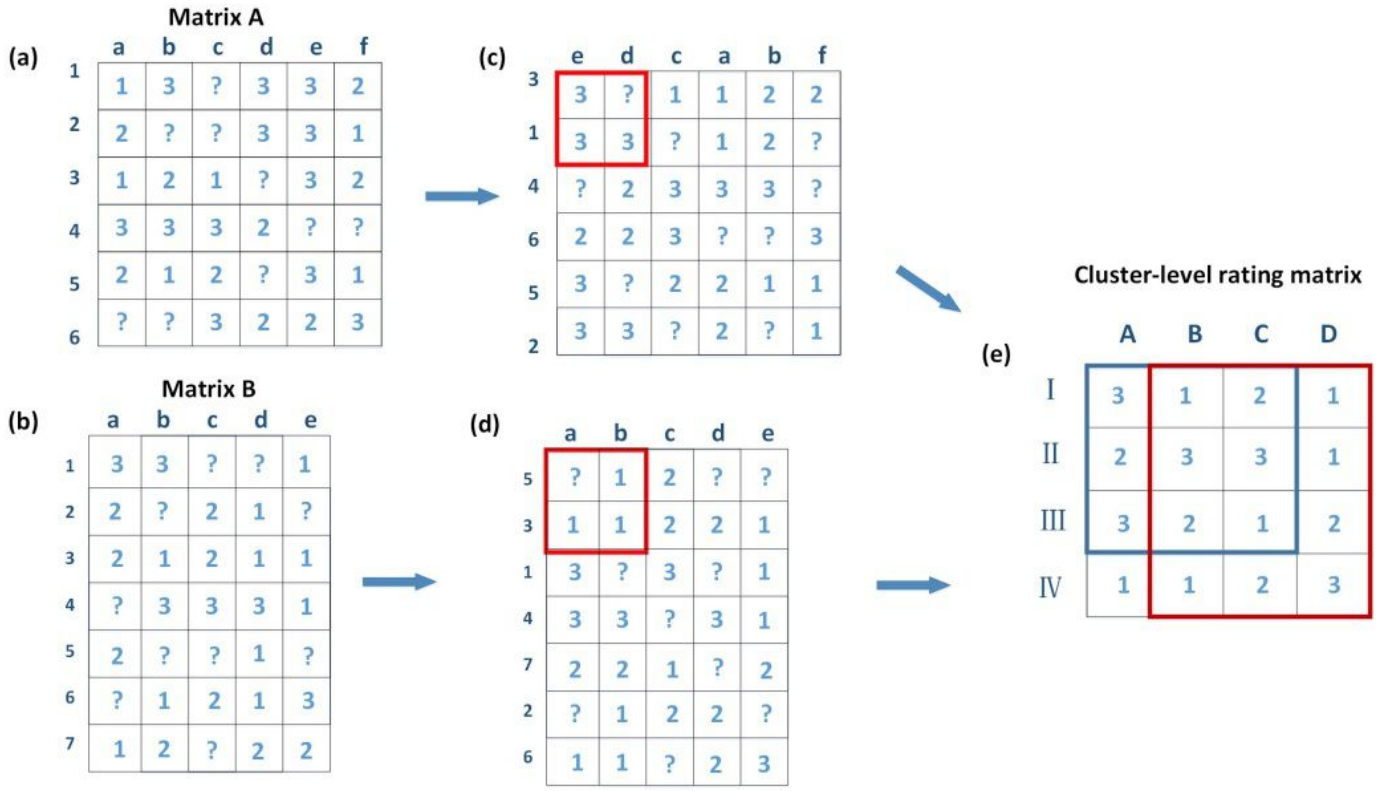

图4 RMGM算法类评分矩阵的成过程图

RMGM Learning

- RMGM learning

$P(C_u^{(k)})$ 表示用户组C(k)u的发生概率、$P(C_v^{(l)})$ 表示项目组 $C_v^{(l)}$ 的发生概率。

则用户u对项目v的预测评分为：

$$
\begin{aligned}
f_R(u,v) &= \sum_r rP(r \mid u,v) \\
&= \sum_r r \sum_{k,l} P(r \mid C_u^{(k)}, C_v^{(l)}) P(C_u^{(k)}, C_v^{(l)} \mid u,v) \\
&= \sum_r r \sum_{k,l} P(r \mid C_u^{(k)}, C_v^{(l)}) P(C_u^{(k)} \mid u) P(C_v^{(l)} \mid v)
\end{aligned}
\tag{1}
$$

RMGM Learning

E-STEP:

在E步过程中，首先计算联合概率：

$$P(C_{u_i}^{(k)}, C_{v_i}^{(l)} \mid u_i, v_i, r_i) = \frac{P(C_{u_i}^{(k)})P(\mathrm{u}_i \mid C_{u_i}^{(k)})P(C_{v_i}^{(l)})P(\mathrm{v}_i \mid C_{v_i}^{(l)})P(r_i \mid C_{u_i}^{(k)}, C_{v_i}^{(l)})}{\sum_{i=1:R} P(C_u^{(k)})P(\mathrm{u} \mid C_u^{(k)})P(C_v^{(l)})P(\mathrm{v} \mid C_v^{(l)})P(r \mid C_u^{(k)}, C_v^{(l)})} \tag{3}$$

M-STEP:

在M步过程中，对下面概率进行更新:

$$P(\mathrm{u}_i \mid C_{u_i}^{(k)}) = \frac{\sum_{l=1:L}\sum_{i=1:r} P(C_{u_i}^{(k)}, C_{v_i}^{(l)} \mid u_i, v_i, r_i)}{P(C_{u_i}^{(k)})R} \tag{4}$$

$$P(v_i \mid C_{v_i}^{(l)}) = \frac{\sum_{k=1:\mathrm{K}}\sum_{i=1:r} P(C_{u_i}^{(k)}, C_{v_i}^{(l)} \mid u_i, v_i, r_i)}{P(C_{v_i}^{(l)})R} \tag{5}$$

$$P(r \mid C_u^{(k)}, C_v^{(l)}) = \frac{\sum_{i \in r_i = r} P(C_{u_i}^{(k)}, C_{v_i}^{(l)} \mid u_i, v_i, r_i)}{\sum_{i=1:\mathrm{R}} P(C_{u_i}^{(k)}, C_{v_i}^{(l)} \mid u_i, v_i, r_i)} \tag{6}$$

计算每个用户ui和项目vi的相关概率，经过上述过程反复迭代计算，直致收敛。

Related Problem

◆ 交叉推荐存在的问题

我们发现使用同一种交叉推荐算法，不同辅助域的辅助推荐效果有着显著差异。有些甚至还会出现低于不使用辅助域的“负”迁移效果。然而以往的交叉推荐方法并没有研究辅助域的差异性对推荐效果的影响。所以，如何选取适当的辅助域成为了交叉推荐领域较为重要的目标。

◆ 辅助域特征分析

我们对辅助域的领域特征进行刻画。试图寻找和推荐效果相关的领域特征量。由于迁移学习推荐算法以及经典的协同过滤推荐算法都是基于相似用户的打分特征进行预测分析。

DATA

本实验数据采用Movielens-1M数据集。该数据集是由美国明尼苏达大学的grouplens小组提供（http://www.grouplens.org）。其包含了6040个匿名用户对3592部影片的100,0000多条打分记录。

表-1 按影片类别划分数据域的基本信息表

ID	Genre	N_{user}	N_{movie}	$\langle k_{user} \rangle$	$\langle k_{movie} \rangle$	$\langle r \rangle$	σ_r
1	Documentary	2243	110	3.53	15.98	3.93	1.0672
2	Western	4100	67	5.04	73.61	3.64	1.2096
3	Film-Noir	4150	44	4.40	173.91	4.08	0.8698
4	Musical	4754	113	8.74	166.13	3.67	1.2123
5	Animation	4808	105	9	37.23	3.68	1.1709
6	Fantasy	4850	68	7.48	180.6	3.45	1.2841
7	Mystery	5133	104	7.83	365.25	3.67	1.1810
8	Children's	5283	250	13.66	48.35	3.42	1.3476
9	Horror	5300	339	14.41	1123.32	3.22	1.5019
10	Crime	5662	201	14.05	1807.75	3.71	1.1615
11	War	5769	141	11.88	202.14	3.89	1.1348
12	Adventure	5894	281	22.73	1185.42	3.48	1.2757
13	Sci-Fi	5911	274	26.61	1512.44	3.47	1.3392
14	Romance	5961	459	24.75	321.4	3.61	1.1380
15	Thriller	5989	485	31.67	692.26	3.57	1.2247
16	Action	6012	495	42.82	530.84	3.49	1.2848
17	Crime	6031	1163	59.12	2528.94	3.52	1.2560
18	Drama	6037	1493	58.73	5291.48	3.77	1.0937

Analysis & Results

- 不同辅助域的辅助差异

$$MAE = \sum_{r_{ua} \in r^P} \frac{|r_{ua} - \widehat{r_{ua}}|}{|r^P|}$$

这里u、a分别指代测试集中的用户u和商品a。

$\widehat{r}_{ua}$ r_{ua} 分别表示预测的评分和测试集中实际的评分。

$|r^P|$ 为测试集中的评分数量。所以，较低的MAE值代表较高的预测准确性。

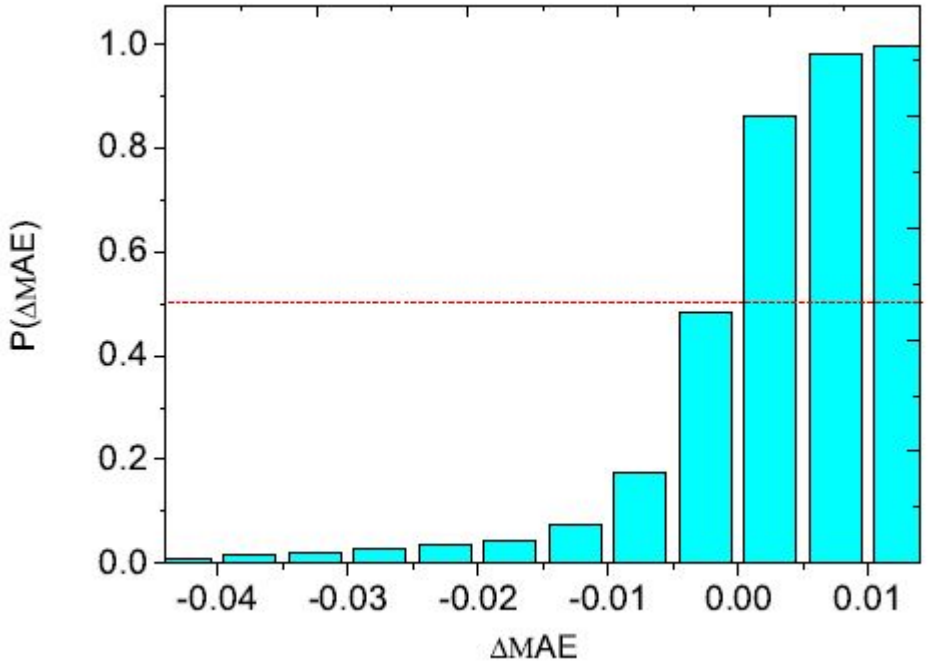

图-5 RMGM算法的MAE差值累积分布图

Analysis & Results

◆ 辅助域特征分析

1. 用户置信度

$$\mathrm{UCC}(S,T)=\frac{|U_S \cap U_T|}{|U_T|} \tag{7}$$

US表示为辅助域电影打分的用户集合，|US |表示该集合的数量。明显的，较高的UCC(S,T)值表示UT有较多的用户被US覆盖。

2. KL散度

$$\mathrm{KL}(P,Q)=\sum_i P(i)\ln\frac{P(i)}{Q(i)} \tag{8}$$

Q与P的KL散度表示Q相对于P的信息损失。当且仅当Q与P的概率分布完全一致时，KL(P,Q)=0。因为打分的范围为1~5，所以将Q和P设置为分别代表辅助域和目标域的打分概率5维向量。

Analysis & Results

◆ 用户置信度相关性分析

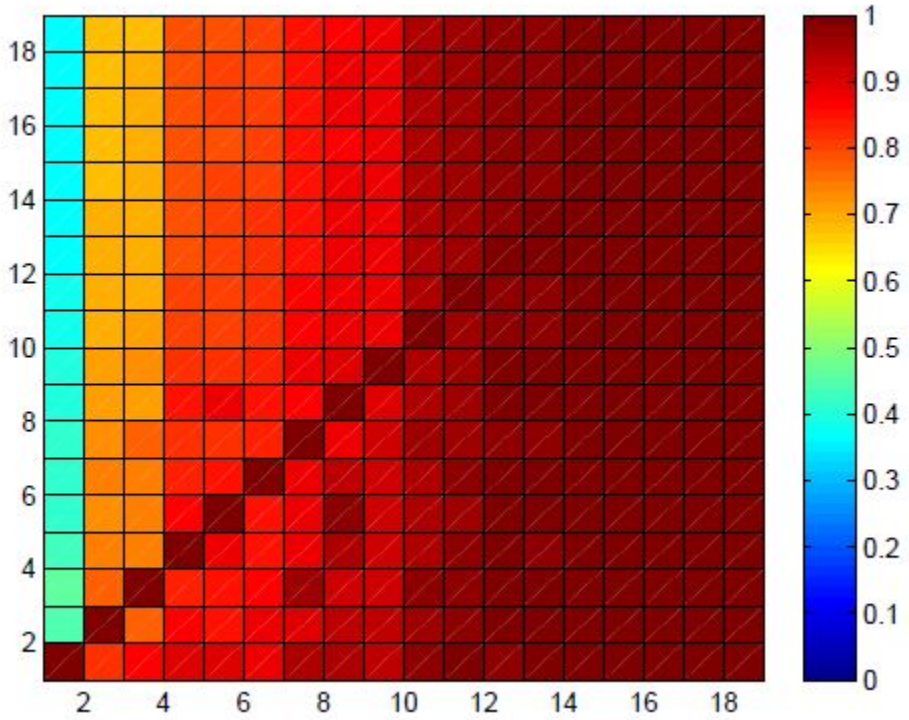

图-6 18组电影域的用户置信度

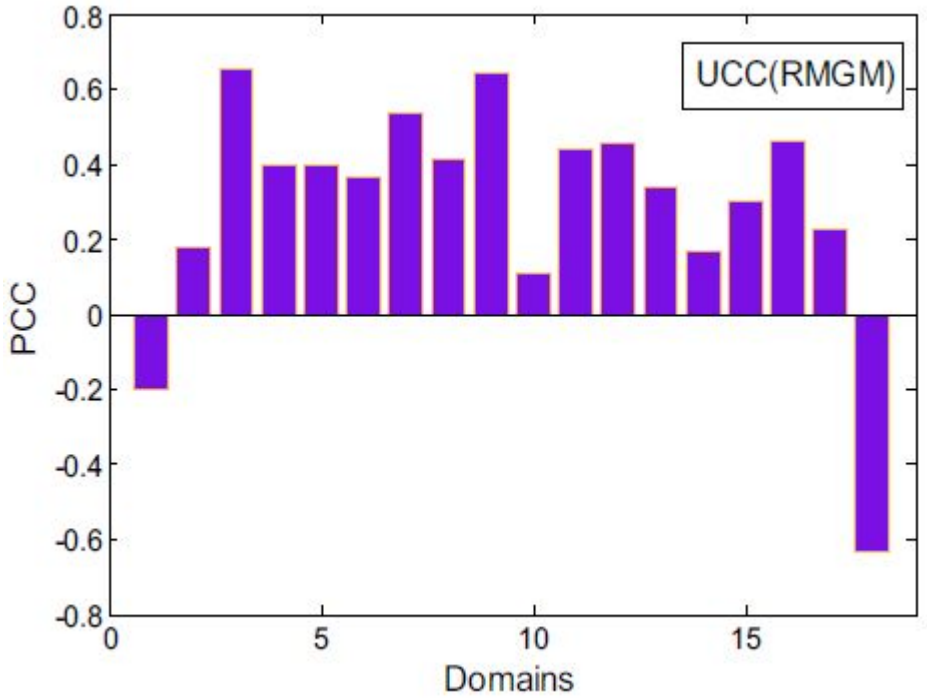

图-7 RMGM模型的MAE差异和用户置信度的相关性

Analysis & Results

- KL散度相关性分析

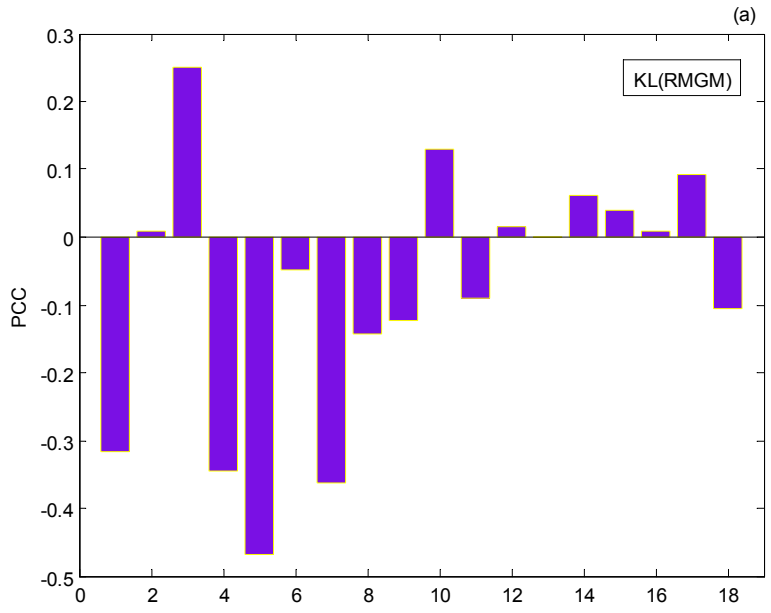

图-8 RMGM模型的MAE差值M(:,y)与KL(:,y)的相关性

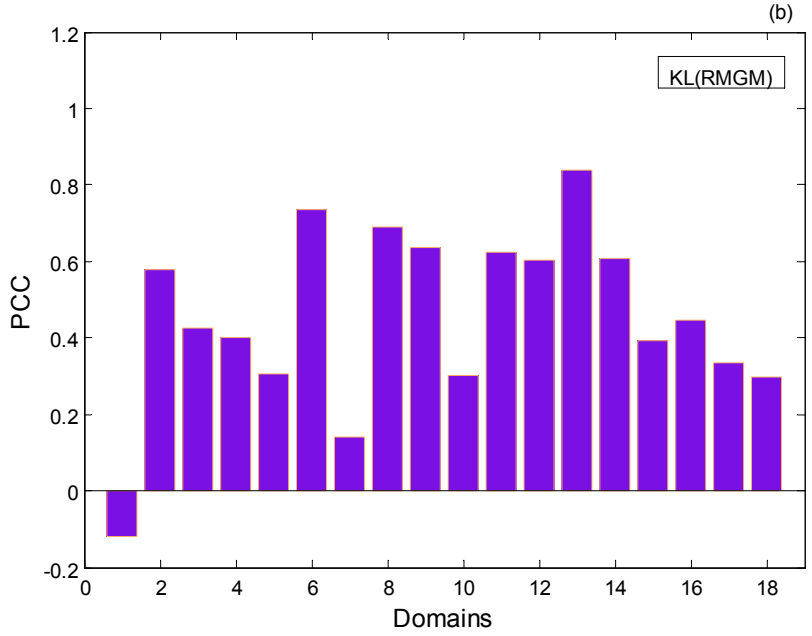

图-9 RMGM模型的MAE差值M(:,y)与KL(:,y)的相关性（去掉相同用户）

Conclusion

- 实现了一种基于迁移学习的交叉推荐算法。
- 提出了两种挑选更加匹配目标域的辅助域特征指标。


Thank you!