基于共享知识模型的跨领域推荐算法

李林峰,刘 真,魏港明,任 爽,葛梦凡
(北京交通大学计算机与信息技术学院,北京 100044)

摘 要: 互联网的普及使得大量信息不断累积,推荐系统作为解决信息过载的有效手段,能够帮助人们迅速准确地筛选出感兴趣的内容.但是由于用户项目评分数据过于稀疏,新用户或新商品存在"冷启动"问题,使得传统的推荐算法计算复杂性过高、准确性较低.考虑到用户会在互联网不同领域使用各类应用,在不同领域积累了大量行为数据和评价信息.而从用户群体的角度来说,在不同领域间存在着用户群体的偏好相似性,因此如果通过在不同领域中共享代表偏好的知识模型,将有助于提升在新领域推荐的准确性,解决冷启动问题.本文提出了基于共享知识模型的跨领域推荐算法 SKP(Sharing Knowledge Pattern),通过对各个领域中用户-项目的评分矩阵分解,得到用户的潜在特征矩阵和项目的潜在特征矩阵,对用户和项目的潜在特征分别聚类,得到了用户分组对项目分组的评分知识模型,最终利用目标领域的个性知识模型和各个领域的共性知识模型来得出推荐结果.本文对三个不同领域的数据集进行了分析和划分,并在物理集群环境下进行了实验.结果表明,通过利用数据稠密的辅助领域数据,本文提出的 SKP 算法与已有的单领域算法、跨领域算法相比,具有更高的准确率和更低的 RMSE 值.

关键词: 跨领域;推荐算法;冷启动;潜在因子;知识模型
中图分类号: TN95 **文献标识码:** A **文章编号:** 0372-2112 (2018)08-1947-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.08.020

Cross-Domain Recommendation Algorithm Based on Sharing Knowledge Pattern

LI Lin-feng, LIU Zhen, WEI Gang-ming, REN Shuang, GE Meng-fan
(*School of Computer and Information Technology, Beijing Jiaotong University, Beijing* 100044, *China*)

Abstract: With the popularity of the Internet and the accumulation of large amounts of data, recommendation system, as an effective means to solve the problem of information overload, can help people quickly select what they are interested in. Because of the sparse user-item rating data, and the cold start problem of new users or new items, traditional recommendation algorithm has the shortcoming of high complexity, low accuracy. Considering the accumulated users behavior or rating data across different domains can have the same preferences, we can share the knowledge pattern among different domains. Based on the matrix factorizationof user-itemrating data in different domains, we can obtain the latent feature matrix of users and items respectively. Considering the user group preference, the latent features of users and items are clustered separately as the domainknowledge pattern. Moreover, By clustering the cross-domain knowledge patterns, we can get shared common knowledge pattern. With the domain knowledge pattern and the shared common knowledge pattern, we can make the final-recommendation. Based on the above consideration, this paper proposes the SKP (Sharing Knowledge Pattern) algorithm. And the SKP is realized in a parallel manner. Experimentsare carried out in the physical cluster environment. By exploiting three different datasets, the results show that the SKP algorithm has better recommendation accuracy and lower RMSE values compared with the existing single-domain algorithm and other cross-domainalgorithms.

Key words: cross-domain; recommendation algorithm; cold start; latent factor; knowledge pattern

收稿日期:2017-07-26;修回日期:2017-12-17;责任编辑:蓝红杰
基金项目:科技部国家重点研发计划(No. 2016YFB1200100);国家自然科学基金(No. 61202429, No. 61763031);中央高校基本科研业务费专项(No. 2017JBM024)

1 引言

随着信息技术和互联网的快速发展,大众对社会化网络的参与和关注程度越来越高,每个领域都积累了大量的用户(User)对项目(Item)的使用、购买或评价信息.电子商务网站和在线社交媒体的迅速发展已经可以满足用户提供兴趣反馈并且在多个系统中维护个人信息,这些信息反映了用户的各种各样的品味和兴趣.有研究表明,利用用户在其它领域的信息和偏好,解决目标领域的"新用户"、"冷启动"[1]问题,能够取得较好的推荐效果.即通过跨领域(Cross-domain)推荐算法把其它领域的有效信息迁移到目标领域中,设计更精确的模型,提高目标领域的推荐准确性.然而,由于隶属于各个领域项目的多样性、用户对不同领域项目的不同偏好以及不同领域知识的相对独立性使得信息资源在领域间不能有效地得到利用.如何利用这些数据建立跨领域推荐系统,已成为当前工业界的研究热点.

目前,跨领域推荐的研究已成为推荐领域的一个新的研究方向,该问题的研究既具有良好的研究意义,也可产生明显的经济效益.工业界、学术界对此都开展了深入研究.工业界的交叉销售[2]、配套推荐,都是跨领域推荐的典型应用.人工智能、数据挖掘、信息检索、推荐系统等领域的重要国际学术会议,如 AAAI、SIGKDD、SIGIR、RecSys 等均有较多跨领域推荐相关的研究成果[3~5].

针对当前跨推荐算法研究中[6~9],数据稀疏度高算法准确率低,以及新用户冷启动这两个问题,本文提出了基于共享知识模型的跨领域推荐算法 SKP(Sharing Knowledge Pattern),通过矩阵分解方法在不同领域的用户-项目的评分数据中抽取用户和项目的潜在特征,并考虑用户群体的偏好相似性,对用户和项目的潜在特征分别聚类,得到用户分组对项目分组的评分知识模型,最终利用目标领域中的个性知识模型和辅助领域的共享共性知识模型得出目标领域的推荐结果,提升了目标领域中非重合用户和重合用户的推荐效果.

2 相关工作

跨领域推荐的早期研究主要是对不同领域的用户-项目历史数据建模并拼接为一个整体的用户-项目评分矩阵,再进一步通过在单领域中使用经典的协同过滤方法.这类方法虽然利用了跨领域的相关信息,但算法本身并未从模型上得到优化,其效果难以优于单领域推荐.

文献[10]中,Singh 和 Gordan 提出了联合矩阵分解模型(Collective Matrix Factorization,CMF)的方法来学习用户与项目之间的关系,通过矩阵间不同权重的设定来构造损失函数,这种通过多个领域的潜在因子学习只能对领域间的重叠用户/重叠项目表现出较好的学习结果,而对领域间的非重叠用户/非重叠项目的学习结果较差,并且这种联合矩阵有大量的参数,在梯度下降求解最小化损失函数时消耗大量的时间和资源,并未得到广泛的应用.

在 CMF 的基础上,Jamali 等人[11]提出了异构矩阵分解模型(HeteroMF)的方法对联合矩阵分解(CMF)算法进行了改进.异构矩阵利用了领域间的上下文信息来进行矩阵分解,他把不同领域对目标领域的相关性参数添加到联合矩阵的潜在模型训练中去,即在 CMF 的基础上设定了各个领域的权重参数.引入这些参数之后,对用户/项目的潜在特征进行训练的准确率明显得到了提升.但是,该模型参数量大导致改进模型复杂程度很高,模型构建时间较长,不利于大规模数据集应用.

近年来,跨领域推荐与机器学习技术,尤其是迁移学习[12~14]技术相结合.迁移学习是在若干领域中提取出知识,并把该知识迁移到不同领域中的方法.在跨领域推荐中运用迁移学习的方法,可以通过迁移相关领域数据来解决目标领域的数据稀疏性问题,Li B 等人[15]通过把辅助领域稠密的用户-项目评分模式迁移到稀疏的目标领域中去,前提是这些领域必须是相关的.

文献[16]在基于迁移技术的跨领域推荐中特别地引入了"密码书(Codebook,CBT)"的概念:从用户-项目的评分矩阵来分析,用户/项目都有其自身特征,若把这些用户和项目分别进行聚类,能够得到不同类型的用户/项目的类,其中每一类内部的用户/项目差异很小,而类与类之间的用户/项目差异较大.CBT 就是通过用户/项目聚类后得到的能代表共性特征的一个"浓缩矩阵".CBT 矩阵可以在有差异的领域间进行迁移,而不用考虑用户/项目的重合.然而,这种强行迁移并没有考虑各自领域的特有特征,因此在不同的领域中表现出较大的差异,即在相关性强的领域这种方法迁移效果较好,在相关性较弱的领域甚至会导致"负迁移"[17,18]的现象.Gao 等人[19]提出了一种基于潜在因子模型的跨领域信息推荐算法(CLFM)不仅考虑了领域间的共性特征也考虑了单领域的个性特征,针对非重合用户提高了推荐的准确性,但是并没有在差异较大的领域间做尝试也没有进一步考虑重合用户群体的跨领域偏好性.

Berkovsky 和 Shi 等人[20]也提出了数据分布的概率相关性迁移方法,首先利用贝叶斯概率模型分析每个领域中用户-项目的评分数据分布的相关性,进而把这种相关性关系迁移到目标领域进行推荐.该方法并未对用户行为和兴趣进行建模,由于运算量较大,比较难于处理大规模数据.

3　基于共享知识模型的跨领域推荐算法

3.1　评分数据表示

假设 n 为用户数量,m 为项目的数量,用户-项目的 $n*m$ 矩阵 $\boldsymbol{R}$,用于表示所有用户对项目的评分. 推荐算法的目标就是通过评分矩阵已知的评分项去预测未知项的评分值. 评分矩阵 $\boldsymbol{R}$ 如图 1 所示:

$$\boldsymbol{R}=\begin{pmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nm} \end{pmatrix}$$

图 1　用户-项目评分矩阵

在本文提出的跨领域推荐算法中,将源领域和目标领域的评分数据分别用矩阵 $\boldsymbol{R}_S$ 和 $\boldsymbol{R}_T$ 表示.

3.2　矩阵分解和潜在因子提取

源领域和目标领域的评分矩阵 $\boldsymbol{R}_S$ 和 $\boldsymbol{R}_T$ 都非常稀疏,首先在各个领域先通过矩阵分解的方法将评分矩阵 $\boldsymbol{R}$ 分解为两个低维的用户特征矩阵 $\boldsymbol{P}$ 和项目特征矩阵 $\boldsymbol{Q}$ 的乘积,如图 2 所示:

$$\boldsymbol{P}=\begin{pmatrix} r_{11} & \cdots & r_{1k} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nk} \end{pmatrix}$$

$$\boldsymbol{Q}=\begin{pmatrix} r_{11} & \cdots & r_{1k} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mk} \end{pmatrix}$$

$$\boldsymbol{R}=\boldsymbol{P}\times\boldsymbol{Q}^{\mathrm{T}}=\begin{pmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nm} \end{pmatrix}$$

图 2　矩阵降维分解图示

其中 k 为用户和项目的隐含特征个数,按照隐语义模型(Latent Fator Model,LFM),用户 u 对项目 i 的评分的预测值 $\hat{R}(u,i)=\hat{r}_{ui}$ 通过公式(1)计算:

$$\hat{r}_{ui}=\sum_{k} p_{uk}q_{ik} \tag{1}$$

通过评价指标 RMSE 最小化,调整参数值,学习 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$. 构造的损失函数如公式(2):

$$f(\boldsymbol{p},\boldsymbol{q})=2\sum_{(u,i)\in Train}\left(r_{ui}-\sum_{f=1}^{F}p_{uf}q_{if}\right)^2+2\lambda\left(\|p_u\|^2+\|q_i\|^2\right) \tag{2}$$

其中,r_{ui} 为真实评分. 为了防止过拟合问题,加入正则化项 $\lambda(\|p_u\|^2+\|q_i\|^2)$,$\lambda$ 为正则化参数.

上述最优化问题,可以利用随机梯度下降法求解得到最终的 $\boldsymbol{P}$、$\boldsymbol{Q}$ 矩阵. 在迭代时,我们设定学习速率 α 的初始值为 0.1,为保证迭代收敛,每次的更新之后 α 值降为原来的 0.9 倍.

3.3　潜在因子聚类

在 3.2 节中求出了能代表用户和项目特征的矩阵 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$,进而可以根据某用户的特征向量与某项目的特征向量的内积求出该用户对项目的评分. 同理,可以求得用户类别对项目类别的评分. 例如在电影领域中,用户类别可以是{非常活跃群体,活跃群体,僵尸用户,…},电影类别可以是{喜剧,恐怖,古装,…},音乐类别可以是{回忆,古风,电视节目插曲,…}. 在不同的领域中,用户类别和项目类别可能存在着重叠情况,比如古装和古风可视为重叠类别,同样用户类别也会有重叠。

考虑到不同领域间用户群体偏好更具备相似性,我们在领域内根据所提取的用户特征矩阵和项目特征矩阵分别对用户和项目进行聚类,得到 k 个用户类别 $\{\boldsymbol{C}_u^1,\boldsymbol{C}_u^2,\cdots,\boldsymbol{C}_u^k\}$ 和 t 个项目类别 $\{\boldsymbol{C}_v^1,\boldsymbol{C}_v^2,\cdots,\boldsymbol{C}_v^t\}$.

具体的,我们采用 K-means 算法,对用户特征矩阵 $\boldsymbol{P}$ 中的用户特征向量 $u_1,u_2,\cdots,u_n\in P^n$ 进行聚类. 首先随机选取 k 个聚类中心点(cluster centroids)分别为 $C_u^1,\cdots,C_u^k\in P^n$. 对于每一个用户 u_i,通过公式(3)计算其与 k 个中心点中每一个中心点的距离,选择距离最近的那个用户类别作为该用户所属类别;通过公式(4)重新计算其中心点 C_u^j(对该用户分组中所有的用户坐标求平均). 重复迭代这两个步骤直到 k 个类的中心点不变或者变化小于某个值,则算法收敛. 同样的方法对项目特征矩阵 $\boldsymbol{Q}$ 中的项目特征向量 $\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n\in\boldsymbol{Q}^m$ 进行聚类.

$$c_i=\arg\min_j\|\boldsymbol{u}_i-\boldsymbol{c}_u^j\|^2 \tag{3}$$

$$c_u^j=\frac{\sum_{i=1}^{n}1\{\boldsymbol{c}_i=j\}\boldsymbol{u}_i}{\sum_{i=1}^{n}1\{\boldsymbol{c}_i=j\}\boldsymbol{u}_i} \tag{4}$$

上式中 c_i 代表用户 u_i 与 k 个类中距离最近的那个用户类别,c_i 的值是 1 到 k 中的一个. $\boldsymbol{c}_u^j$ 代表一个用户类别的用户中心点.

对各个领域分别聚类,得出各个领域用户类别对项目类别的评分矩阵,称其为个性知识模型. 虽然该知识模型在各个领域间有可能存在重叠,但也必然存在着非重叠部分. 如果在各个领域共享该知识模型,会将不重叠的知识模型传递到目标领域,势必导致负迁移现象,所以需要提取出各个领域间的用户类别进行聚类,得到共性知识模型,通过共享该共性知识模型来提升目标领域的参考知识,进而解决冷启动、数据稀疏性问题.

3.4　源领域和目标领域潜在因子融合模型

融合源领域和目标领域基于潜在因子提取的知识模型求解过程如图 3 所示.

首先根据用户和项目类别和构建出各自领域的个性知识模型和各个领域的共性知识模型. 最后,融合各自领域的个性知识模型和各个领域的共性知识模型应用于目标领域,并通过调整不同的权重来求出用户对项目的最终评分,通过排序后进行项目的推荐.

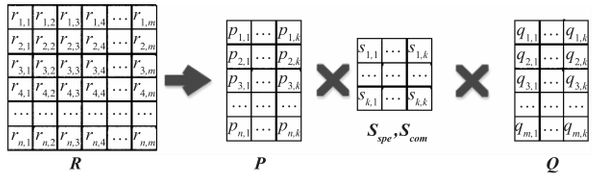

图3 知识模型融合过程

首先通过上一节中对潜在因子模型的聚类可得出各个领域中用户分组 c_u^k 对应项目分组 c_v^{T} 的评分知识模型 $\boldsymbol{S}_{spe}=\boldsymbol{C}_u\boldsymbol{C}_v$. Z 个领域的各自知识模型代表各个领域的各自用户分组对相应项目分组的评分知识模型 $\boldsymbol{S}_{spe}$. 以某个源领域为例,图 4 表示了该领域个性知识模型的提取过程. 其中,X、Y、Z 表示用户分组,A、B、C 表示项目分组.

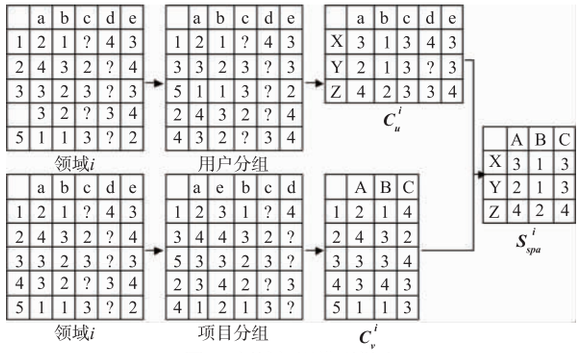

图4 领域知识模型提取

在得到各个领域的用户分组 $\boldsymbol{C}_u$/项目分组 $\boldsymbol{C}_v$ 后,再对所有 Z 个领域的所有用户分组 $\boldsymbol{C}_u$ 进行聚类,并对所有 Z 个领域的所有项目分组 $\boldsymbol{C}_v$ 进行聚类,分别得到的聚类结果中包含 Z 个领域的用户分组/项目分组的类别作为 Z 个领域的用户重叠分组 $\boldsymbol{C}_{u-com}$ 和项目重叠分组 $\boldsymbol{C}_{u-com}$,并基于 Z 个领域的用户重叠分组和项目重叠分组得出各领域间的共同评分知识模型 $\boldsymbol{S}_{com}=\boldsymbol{C}_{u-com}\boldsymbol{C}_{u-com}$,如图 5 所示:

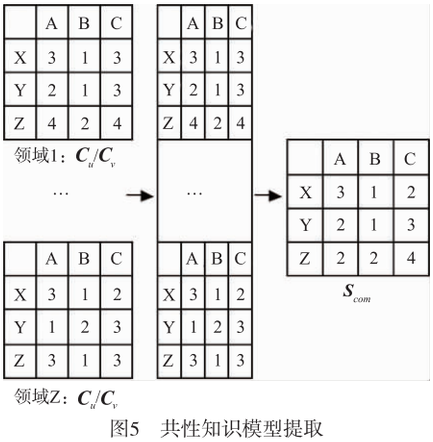

图5 共性知识模型提取

其中共性评分知识模型 $\boldsymbol{S}_{com}$ 代表各领域间共同用户分组对相应的共同项目分组的评分知识模型. 利用 $\boldsymbol{S}_{com}$ 解决目标领域的数据稀疏性问题,利用 $\boldsymbol{S}_{spe}$ 解决目标领域推荐的准确性问题. 综合考虑领域间共性知识模型 $\boldsymbol{S}_{com}$ 和各自领域的个性知识模型 $\boldsymbol{S}_{spe}$,并通过调整不同的权重来求出用户对项目的最终评分 $\hat{\boldsymbol{R}}$,通过 $\hat{\boldsymbol{R}}$ 排序后进行项目推荐. 如公式(5)所示,w 为衡量迁移共性知识模型的比重.

$$\hat{\boldsymbol{R}}=w\boldsymbol{P}\boldsymbol{S}_{com}\boldsymbol{Q}+(1-w)\boldsymbol{P}\boldsymbol{S}_{spe}\boldsymbol{Q} \tag{5}$$

3.5 基于共享知识模型的跨领域推荐算法

基于上述工作,本文提出基于共享知识模型的跨领域推荐算法 SKP(Sharing Knowledge Pattern,SKP)的框架如图 6 所示.

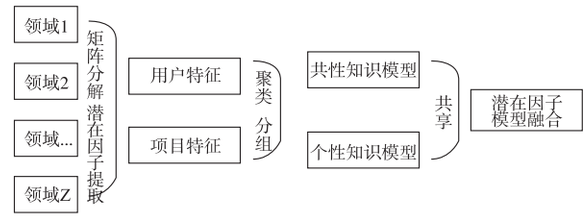

图6 SKP算法框架

SKP 算法设计描述如算法 1 所示.

算法 1 基于共享知识模型的跨领域推荐算 SKP

输入:各源领域的用户-项目评分数据
输出:目标领域每个用户的推荐结果
(1)$\boldsymbol{P},\boldsymbol{Q}\leftarrow\min(\boldsymbol{f}(\boldsymbol{p},\boldsymbol{q}))$;//根据公式(2),对各领域的用户-项目评分数据求出各领域的用户特征矩阵 $\boldsymbol{P}$ 和项目特征矩阵 $\boldsymbol{Q}$.
(2)for $i=1$ to n do
(3)$\boldsymbol{C}_u\leftarrow$Clustering($\boldsymbol{P}$);//根据公式(3),(4),求出用户分组.
(4)end for
(5)for $j=1$ to m do
(6)$\boldsymbol{C}_v\leftarrow$Clustering($\boldsymbol{Q}$);//根据公式(3),(4),求出项目分组.
(7)end for
(8)$\boldsymbol{S}_{spe}=\boldsymbol{C}_u*\boldsymbol{C}_v$
(9)$\boldsymbol{C}_{u\text{-}com}\leftarrow$Clustering($U_1^Z\boldsymbol{C}_u$);//对 $\boldsymbol{Z}$ 个领域的所有用户分组进行聚类.
(10)$\boldsymbol{C}_{v\text{-}com}\leftarrow$Clustering($U_1^Z\boldsymbol{C}_v$);//对 $\boldsymbol{Z}$ 个领域的所有项目分组进行聚类.
(11)$\boldsymbol{S}_{com}=\boldsymbol{C}_{u\text{-}com}*\boldsymbol{C}_{v\text{-}com}$
(12)$\hat{\boldsymbol{R}}=w\boldsymbol{P}\boldsymbol{S}_{com}\boldsymbol{Q}+(1-w)\boldsymbol{P}\boldsymbol{S}_{spe}\boldsymbol{Q}$//得到目标领域的用户预测评分矩阵

在步骤(1)中首先初始化各个源领域评分矩阵 $\boldsymbol{R}$,分别把各个领域的用户-项目-评分数据广播(broadcast)到所有节点上,继而通过公式(2)根据各领域的用户-项目-评分数据计算出各领域的用户特征矩阵 $\boldsymbol{P}$,项目特征矩阵 $\boldsymbol{Q}$. 在步骤(2)中根据内存中的用户特征 P_z 和项目特征 $\boldsymbol{Q}_z$,通过公式(3)、(4)计算各个分组之间的距离,求出

各个领域中的用户和项目所属的分组.

在步骤(3)中,通过代表各个领域的用户特征 $\boldsymbol{P}_z$ 和项目特征 $\boldsymbol{Q}_z$ 求取所有领域的个性知识模型和共性知识模型,综合考虑各自领域的个性知识模型和共享领域间的共性知识模型的跨领域推荐结果:将目标领域的个性知识模型和各个领域间的共性知识模型根根据公式(5)得出用户项目评分矩阵. 最后通过对预测结果的排序 sortBy 操作后给用户推荐出 top-N 的项目集合.

4　实验结果及分析

4.1　实验数据

本实验是取自 3 个独立的领域,分别是 MovieLens 数据集①,Book-Crossing 数据集②和豆瓣的图书评分数据③. 数据集介绍如下:

MovieLens 数据集:实验中采用的是 135MB 的 ml-10M 数据,包括 35594 个用户对 10533 个电影的 5095567 条评分数据,评分范围在 1 – 5 分. 每一个用户至少对 20 个电影评过分.

Book-Crossing 数据集:实验中采用的是 120MB 的数据集,包括 278,858 个用户对 271,379 本图书的 1.1 百万条评分数据,评分范围在 0 – 10 分. 为了减少领域间重叠用户群体对项目群体的评分知识模型偏差,把该领域的评分数据平滑处理到 1 – 5 分.

豆瓣图书数据集:实验中采用的是 102MB 的图书评分数据集,包括 383033 个用户对 80008 本图书的 3.6 百万条评分数据,评分范围在 1 – 5. 选取豆瓣的图书评分信息,而不是音乐,是考虑和其中的另一个领域 Book-Crossing 可以有较多的重叠.

实验对各个领域数据集进行了分析和划分. 实验需要验证共享知识模型后对自身领域的影响程度,每个领域既是源领域也是目标领域,故而把每个领域的数据集划分为 80% 的训练集和 20% 的测试集. 考虑到各个领域的数据稀疏性对自身领域和其他领域的影响,又对各个领域的数据集根据稀疏程度进行了划分,根据用户对项目的平均评分个数把数据集划分为 4 种,其中 M3,B3,D3 非常稀疏,M4,B4,D4 与原始数据稀疏度相当,M5,B5,D5 比原始数据稠密,M6,B6,D6 更加稠密,如表 1、表 2、表 3 所示.

表 1　MovieLens 数据集划分

编号	用户数	项目数	平均评分个数	描述
M1	35594	10463	114	目标领域训练数据集
M2	35543	9859	28	目标领域测试数据集
M3	35594	1937	13	辅助领域数据集
M4	35594	10533	97	辅助领域数据集
M5	34583	10532	146	辅助领域数据集
M6	21791	10529	213	辅助领域数据集

表 2　Book-Crossing 数据集划分

编号	用户数	项目数	平均评分个数	描述
B1	92839	298511	9	目标领域训练数据集
B2	42033	118539	21	目标领域测试数据集
B3	105283	340556	10	辅助领域数据集
B4	12053	304791	79	辅助领域数据集
B5	7078	290350	124	辅助领域数据集
B6	3371	265499	227	辅助领域数据集

表 3　豆瓣数据集划分

编号	用户数	项目数	平均评分个数	描述
D1	378634	79358	7	目标领域训练数据集
D2	247836	69446	11	目标领域测试数据集
D3	383033	74510	4	辅助领域数据集
D4	37523	77972	46	辅助领域数据集
D5	9164	73031	94	辅助领域数据集
D6	2323	63432	172	辅助领域数据集

4.2　实验结果与分析

首先通过实验对比在不共享辅助领域数据与本文提出的共享不同稀疏程度的辅助领域数据下 SKP 算法对目标领域预测的 RMSE 值. 数据集如表 4 所示,选取 MovieLens 作为目标领域,其他组数据作为辅助领域. 实验结果如表 5 所示,共享了稀疏程度较高的数据集后对结果预测不如单领域的预测,而共享了相对稠密的数据集后对结果预测优于单领域预测.

表 4　数据集分组

组号	数据集编号
1	M1 + M2
2	M1 + M2 + (B3 + D3)
3	M1 + M2 + (B4 + D4)
4	M1 + M2 + (B5 + D5)
5	M1 + M2 + (B5 + D5)

表 5　SKP 算法共享不同稀疏度数据的影响

	1	2	3	4	5
RMSE	0.973	1.0938	0.9234	0.8977	0.8703
Precision	0.159	0.1497	0.1994	0.2224	0.2497
Recall	0.180	0.150	0.251	0.282	0.306
F1 值	0.16	0.1498	0.2222	0.2486	0.2745

基于上述分析,验证了本文提出的 SKP 算法在共

① http://grouplens.org/datasets/movielens/

② http://www2.informatik.uni-freiburg.de/~cziegler/BX/

③ http://www.datatang.com/data/shop-page.html? k = 豆瓣

享了知识稠密领域的数据后使目标领域的推荐效果更好. 为了进一步验证,本文又选取了其他 2 种算法与 SKP 进行对比,分别是基于非负矩阵分解(non-negative matrix factorization,NMF)[21]的算法,在单领域中利用评分矩阵的隐含特征进行推荐. 基于评分模式(Code Book Transfer,CBT)[15]的算法,在跨领域中通过矩阵分解的方式得到一个代表评分的码书(codebook),利用该码书对目标领域稀疏矩阵进行填充.

实验结果如图 7 所示,其中 NMF 算法的各组实验集中未迁移其他领域的数据集,在辅助领域数据集最为密集的实验组 5 中,跨领域算法得到了最低的 RMSE 值,其中 CBT 算法的 RMSE 值为 0.8903,SKP 算法的 RMSE 值为 0.8703. 实验结果表明,跨领域算法 CBT、SKP 的 RMSE 值在数据稠密的实验分组中均小于 NMF,说明共享了其他数据稠密领域的知识模型后推荐算法预测效果较好,本文提出的 SKP 算法与另外两种算法相比 RMSE 值更低.

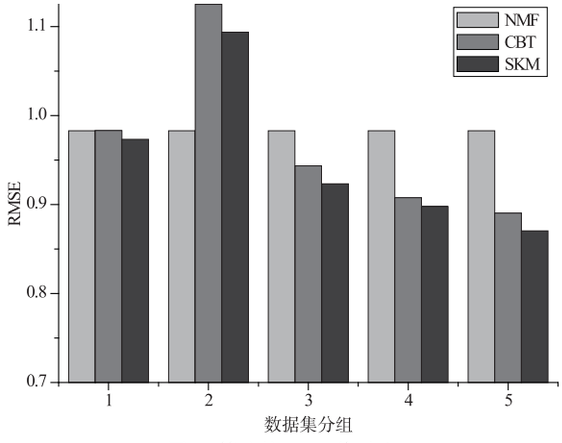

图7 算法的RMSE值对比

评价指标 Precision 和 Recall 的实验结果分别如图 8 和图 9 所示,在辅助领域数据集最为密集的实验组 5 中,SKP 算法和 CBT 算法的 Precision 和 Recall 比单领域的 NMF 算法提升近 10%,说明跨领域算法在中通过利用数据稠密的辅助领域数据后能达到更高的准确率和召回率. 其中 SKP 算法的 Precision 和 Recall 分别为 0.2497 和 0.306,CBT 算法的 Precision 和 Recall 分别为 0.2027 和 0.266. SKP 算法相比 CBT 算法提升 4%.

5 结束语

针对当前推荐领域中数据稀疏度高算法准确率低等问题,本文提出了基于共享知识模型的跨领域推荐算法 SKP(Sharing Knowledge Pattern),通过矩阵分解方法在不同领域的用户-项目的评分数据中抽取用户和项目的潜在特征,并考虑用户群体的偏好相似性,对用户和项目的潜在特征分别聚类,得到用户分组对项目分

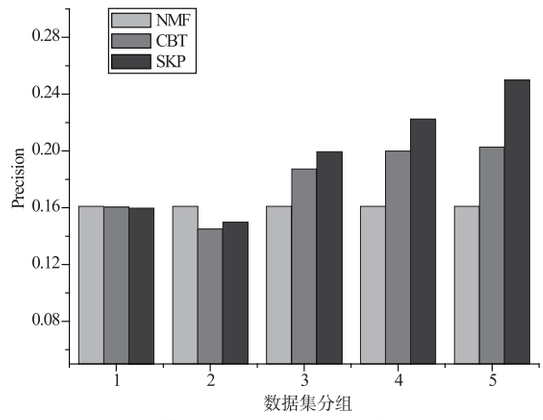

图8 算法的Precision对比

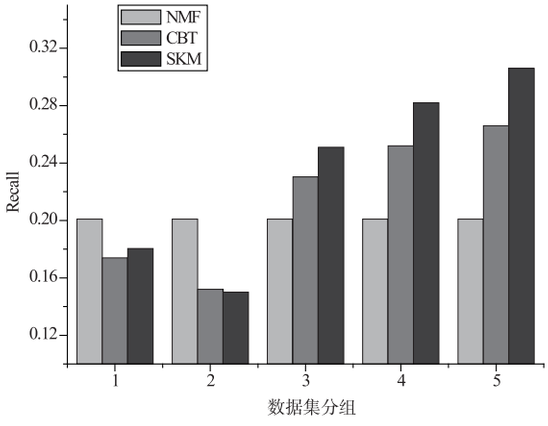

图9 算法的Recall对比

组的评分知识模型,还抽取了多个领域中重叠的用户分组和项目分组,使得共同的评分知识模型可以在多个领域中共享,最终利用目标领域中的个性知识模型和辅助领域的共享共性知识模型来得出目标领域最终的推荐结果.

进一步的跨领域推荐还可以通过构建语境(context)和跨领域推荐之间的联系,将不同的语境(位置,时间和心情)视为不同的领域. 可将上下文感知技术应用在跨领域推荐中,找出上下文与推荐领域之间的关系,并将这种关系视为连接不同领域的桥梁.

参考文献

[1] Mirbakhsh N, Ling C X. Improving top-n recommendation for cold-start users via cross-domain information[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2015, 9(4): http://dx.doi.org/10.1145/2724720.

[2] 张亮,柏林森,周涛. 基于跨电商行为的交叉推荐算法[J]. 电子科技大学学报,2013(1):154-160.
Zhang Liang, Bai Linsen, Zhou Tao. Crossover recommendation algorithm based on cross-merchants behavior[J]. Journal of University of Electronic Science and Technology

of China,2013(1):154 – 160. (in Chinese)
[3] Elkahky A M, Song Y, He X. A multi-view deep learning approach for cross domain user modeling in recommendation systems[A]. Proceedings of the 24th International Conference on World Wide Web[C]. International World Wide Web Conferences Steering Committee,2015. 278 – 288.
[4] Fernández-Tobías I, Tomeo P, Cantador I, et al. Accuracy and diversity in cross-domain recommendations for cold-start users with positive-only feedback[A]. Proceedings of the 10th ACM Conference on Recommender Systems[C]. ACM,2016. 119 – 122.
[5] Chen L, Zheng J, Gao M, et al. TLRec: transfer learning for cross-domain recommendation [A]. Big Knowledge (ICBK), 2017 IEEE International Conference on [C]. IEEE,2017. 167 – 172.
[6] Taneja A, Arora A. Cross domain recommendation using multidimensional tensor factorization[J]. Expert Systems with Applications,2018,92: 304 – 316.
[7] Zheng L, Noroozi V, Yu P S. Joint deep modeling of users and items using reviews for recommendation[A]. Proceedings of the Tenth ACM International Conference on Web Search and Data Mining[C]. ACM,2017. 425 – 434.
[8] 王兴茂,张兴明,吴毅涛,潘俊池. 基于启发式聚类模型和类别相似度的协同过滤推荐算法[J]. 电子学报,2016,44(7):1708 – 1713.
Wang XingMao, Zhang Xingming, Wu Yitao, Pan Junchi. Collaborative filtering recommendation algorithm based on heuristic clustering model and class similarity[J]. Acta Electronica Sinica,2016,44(7):1708 – 1713. (in Chinese)
[9] Xu Z, Zhang F, Wang W, et al. Exploiting trust and usage context for cross-domain recommendation[J]. IEEE Access,2016,4: 2398 – 2407.
[10] Singh A P, Gordon G J. Relational learning via collective matrix factorization [A]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [C]. ACM,2008. 650 – 658.
[11] Jamali M, Lakshmanan L. HeteroMF: recommendation in heterogeneous information networks using context dependent factor models[A]. International Conference on World Wide Web[C]. ACM,2013. 643 – 654.
[12] Shi J, Long M, Liu Q, et al. Twin Bridge Transfer Learning for Sparse Collaborative Filtering[M]. Advances in Knowledge Discovery and Data Mining. 2013. 496 – 507.
[13] Yan Z, Wei L, Lu Y, et al. You are what apps you use: Transfer Learning for Personalized Content and Ad Recommendation[A]. Proceedings of the Eleventh ACM Conference on Recommender Systems[C]. ACM, 2017. 350 – 350.
[14] 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.
[15] Li B, Yang Q, Xue X. Can Movies and books collaborate cross-domain collaborative filtering for sparsity reduction [A]. IJCAI 2009, Proceedings of the International Joint Conference on Artificial Intelligence[C]. Pasadena, California, Usa: DBLP,2009. 2052 – 2057.
[16] Li B, Yang Q, Xue X. Transfer learning for collaborative filtering via a rating-matrix generative model[A]. International Conference on Machine Learning, ICML 2009[C]. Montreal, Quebec, Canada: DBLP,2009. 78.
[17] Fernández-Tobías I, Tomeo P, Cantador I, et al. Accuracy and diversity in cross-domain recommendations for cold-start users with positive-only feedback[A]. Proceedings of the 10th ACM Conference on Recommender Systems [C]. ACM,2016. 119 – 122.
[18] Pan W. A survey of transfer learning for collaborative recommendation with auxiliary data[J]. Neurocomputing, 2016,177: 447 – 453.
[19] Gao S, Luo H, Chen D, et al. Cross-domain recommendation via cluster-level latent factor model[A]. Joint European Conference on Machine Learning and Knowledge Discovery in Databases[C]. Berlin, Heidelberg: Springer, 2013. 161 – 176.
[20] 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.
[21] Hoyer P O. Non-negative matrix factorization with sparseness constraints [J]. Journal of Machine Learning Research,2004,5(1):1457 – 1469.

作者简介

李林峰　女,1993 年生于河南唐河. 北京交通大学计算机科学与技术学院研究生. 研究方向为大数据、推荐系统.

刘　真(通信作者)　女,1977 年生于江西南昌. 北京交通大学计算机与信息技术学院副教授. 研究方向为推荐系统、移动社会网络、云计算与虚拟化.
E-mail:zhliu@ bjtu. edu. cn