基于信任关系和项目流行度的矩阵分解推荐算法

李卫疆 郑雅民
(昆明理工大学信息工程与自动化学院 云南 昆明 650000)

摘 要 针对现有推荐系统推荐覆盖范围不高的问题,提出一种融合项目流行度和用户信任关系的矩阵分解推荐算法。合并用户-项目评分矩阵和用户-用户信任关系矩阵,通过矩阵分解的方式同时传递信任和推荐项目,极大提高了推荐算法的覆盖率,但损失了现有方法 8% 左右的精度。将项目流行度作为权重因子,引入到高稀疏性的用户-项目评分矩阵中,根据项目流行度对用户评分项目和未评分项目分别进行加权处理,提高了推荐算法的准确率。通过在 Epinions 数据集上的对比实验结果表明,该算法在大幅度改善推荐覆盖率的同时,保证了推荐的准确率,能够给于用户更好的推荐效果。

关键词 推荐 信任关系 项目流行度 矩阵分解

中图分类号 TP3 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.09.044

MATRIX FACTORIZATION RECOMMENDATION ALGORITHM BASED ON TRUST RELATIONSHIP AND ITEM POPULARITY

Li Weijiang Zheng Yamin
(*Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650000, Yunnan, China*)

Abstract To solve the problem of low coverage of existing recommendation systems, we proposed a matrix factorization recommendation algorithm that combined item popularity and user trust relationship. We merged the user-item scoring matrix and the user-user trust relationship matrix to transfer trust and recommendation items simultaneously through matrix factorization. It greatly improved the coverage of the recommended algorithm, but lost accuracy up to 8%. We introduced the item popularity as a weighting factor into the high sparse user-item rating matrix. According to the item popularity, we weighted user rating items and non-rating items respectively, which improved the accuracy of the recommendation algorithm. The experiment results on the Epinions dataset show that our algorithm can greatly improve the recommended coverage rate without reducing the accuracy of the recommendation, so it can give users better recommendation results.

Keywords Recommendation Trust relationship Item popularity Matrix factorization

0 引 言

随着互联网的飞速发展,推荐系统得到人们越来越多的关注,提供有效的用户个性化推荐是目前研究的热点问题,通过对用户历史行为的分析,预测用户偏好。近年来,结合用户社交网络的推荐算法目前已经得到广泛的应用,TidalTrust 通过在信任网络中找到从用户到每个其他用户的最短路径,然后将每个用户的信任值聚合到与 A 直接相邻的用户进行推荐[1]。Li 等基于奇异值分解的方法,根据人际关系中的六度分隔理论填充用户信任矩阵,再通过用户之间的信任度进行相关推荐[2]。用户的社交网络不仅可以有效缓解推荐中冷启动的问题,还可以通过好友推荐增加推荐的信任度。

但在社交网络中,好友关系不是基于共同兴趣产

收稿日期:2019-02-02。国家自然科学基金项目(61363045)。**李卫疆**,副教授,主研领域:自然语言处理,信息检索。**郑雅民**,硕士生。

生的,用户好友的兴趣往往和用户的兴趣不一致。同时,信任网络通常是一组不相交的子图,信任无法从一个子图传播到另一个子图上,如果一个项目没有被子图中的任何用户评分,那么它就不会被推荐给子图中有可能感兴趣的用户。因此,仅仅基于用户之间的信任关系进行推荐会存在项目的覆盖问题。同时,推荐算法的覆盖率和准确率存在内在的折中,在提高推荐覆盖率的同时会降低推荐的准确率,反之亦然。

为此,我们提出了一种融合项目流行度和用户信任关系的 PopTruMF 算法。PopTruMF 算法利用矩阵分解的传递性,将用户的信任关系与项目评分视为同一层级,在矩阵分解的过程中,项目和信任关系发生混合,使得信任传递和评分预测同时发生;同时 PopTruMF 算法根据项目的流行度,对用户评分项目和未评分项目分别进行加权处理,在保证了准确率的基础上,有效改善了以上方法的覆盖率。本文的主要创新点在于:

(1) 在基于信任网络的推荐算法中,提出在不相交的信任网络中进行推荐。

(2) 在基于项目的流行度的推荐算法中,提出项目流行度对用户评分项目和未评分项目有不同程度的影响。

(3) 本文在 Epinions 数据集上进行了大量的实验,实验结果表明 PopTruMF 算法在大幅度改善推荐覆盖率的同时,保证了推荐的准确率,能够给于用户更好的推荐结果。

1 相关工作

1.1 矩阵分解

矩阵分解(Matrix Factorization, MF)是表示用户和项目潜在特征向量最有效的方式,一直都活跃在评分预测的推荐方式中[3]。根据已有的用户项目评分,通过矩阵分解的方式,可以有效挖掘用户和项目的潜在因子,进而估计矩阵中大量的缺失值。本文用 $\boldsymbol{R}(\boldsymbol{R}\in\mathbf{R}^{M\times N})$ 表示用户项目评分矩阵,其中,M 和 N 分别表示用户和项目的数目,表示用户 u 对项目 i 的评分,$\boldsymbol{P}_u$ 表示用户 u 的潜在特征向量,$\boldsymbol{Q}_i$ 表示项目 i 的潜在特征向量,同时,$\boldsymbol{P}\in\mathbf{R}^{M\times K}$ 和 $\boldsymbol{Q}\in\mathbf{R}^{N\times K}$ 分别表示用户和项目的潜在特征矩阵。矩阵分解将用户项目评分矩阵映射到 K 维的潜在特征空间,表示成用户特征矩阵和项目特征矩阵的乘积,即 $\boldsymbol{R}\approx\boldsymbol{P}^{\mathrm{T}}\boldsymbol{Q}$,$r_{ui}=p_u\cdot q_i$,从而将项目推荐的问题转换为了预测评分 $\hat{r}$ 的问题。通过最小化损失函数 E_{ui} 来获得 $\boldsymbol{P}_u$ 和 $\boldsymbol{Q}_i$ 的最优解,即:

$$E_{ui}^2=\sum_{(u,i)\in\boldsymbol{R}}w_0(r_{ui}-\hat{r}_{ui})^2+\frac{\beta}{2}\sum_{k=1}^{K}(p_{ik}^2+q_{ku}^2)\quad(1)$$

式中:$\boldsymbol{w}_0$ 代表缺失数据赋予的统一权重。

1.2 信任网络

美国著名的尼尔森调查表明 83% 的用户更愿意相信朋友对他们的推荐,因此,基于信任关系的推荐可以更好地模拟现实社会。信任网络是基于信任传递构建的有向图,通过使用用户之间的信任关系和用户的历史行为数据来预测评分,不仅可以提高推荐系统的推荐质量,还可以增加用户对系统的信任度[2,4-6]。

本文构建了一个简单的信任网络实例,如图 1 所示,其中,节点代表用户,边代表用户之间的信任关系。信任传递是信任网络的一个重要特征,用户可以通过具有直接信任关系的用户作为纽带传递给具有间接信任关系的用户。如图 1 所示,用户 u_1 对项目 i_2、i_3 有历史评分,用户 u_2 对项目 i_1、i_2、i_3、有历史评分,用户 u_3 对项目 i_2、i_3、i_4 有历史评分。因为用户 u_1 和用户 u_3 在相同的信任网络中,通过信任传递我们可以将用户 u_3 偏好的项目 i_4 推荐给用户 u_1。但由于左侧信任网络子图中没有用户节点对项目 i_4 进行评分,则不能通过信任关系将项目 i_4 推荐给用户 u_1。

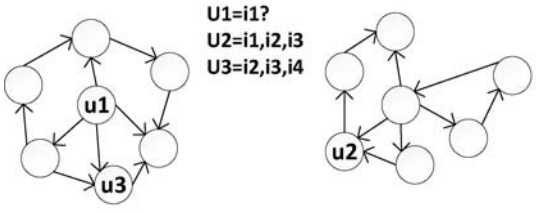


图 1 不相交的用户信任网络实例

1.3 项目流行度及数据稀疏问题

类似于头条新闻、微博热榜等根据 PV、UV、日均 PV 或分享率等数据,按某种热度排序推荐给用户,很多应用也以热门排行榜、最热卖排行榜、平均用户评分等形式展示当下流行的项目。项目被用户评分的次数越多,代表项目流行度越高;反之,则越低。作为推荐系统的新用户,或在人很难做出选择的时候,项目的历史评分确实是很有用的参考信息,且人们通常愿意接受流行项目的推荐[6-7]。基于项目流行度的推荐算法,可以深度挖掘用户偏好,给予不同用户个性化的推荐[9]。

郝立燕等基于 TopN 推荐预测提出,流行度越高的项目,体现用户兴趣的信息越少,相反,流行度越低的项目,体现用户兴趣的信息越可靠,提高了冷门项目的影响力[8]。但是用户-项目评分矩阵中评分项目一

般不超过项目总数的 1%,更多的是缺失的数据,因此,解决数据稀疏问题通常是提高推荐质量的关键。文献[9]将这些缺失的数据统一假定为无关用户偏好进行建模,极大地减少了建模的工作量,然而这种方法是不现实的,会降低模型的预测性。在此假设的基础上,文献[10]将缺失的数据赋予统一的权重,然后在缺失数据上设置显式先验,当新的用户或项目评分进入系统时,可以及时更新因子,给与符合最新用户偏好的推荐,但这样的假设限制了推荐算法在实际应用中的可扩展性。针对以上问题,参考现有的推荐方法,本文提出了 PopTruMF 模型。

2　PopTruMF 模型

本文首先合并用户-项目评分矩阵和用户-用户信任关系矩阵来构建模型;然后,根据项目流行度的加权策略构建目标函数;最后,通过矩阵分解的方式同时传递信任和推荐项目。

2.1　TruMF 算法

2.1.1　基于项目节点传递信任

在信任网络中,由于好友关系不是基于共同兴趣产生的,用户好友的兴趣往往和用户的兴趣不一致,而不存在信任关系的用户却具有相似的用户偏好。如图 2 所示,用户 u_2 和用户 u_3 在不同的信任网络,但对项目 i_2、i_3 都有显示反馈,本文认为用户 u_2 和用户 u_3 具有相似的用户偏好,只不过他们不认识对方。

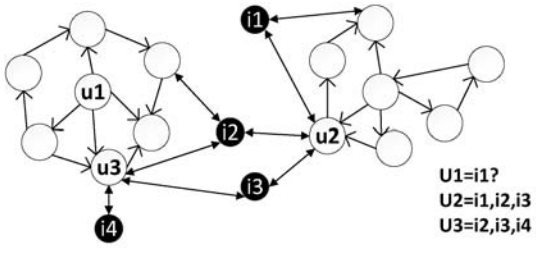
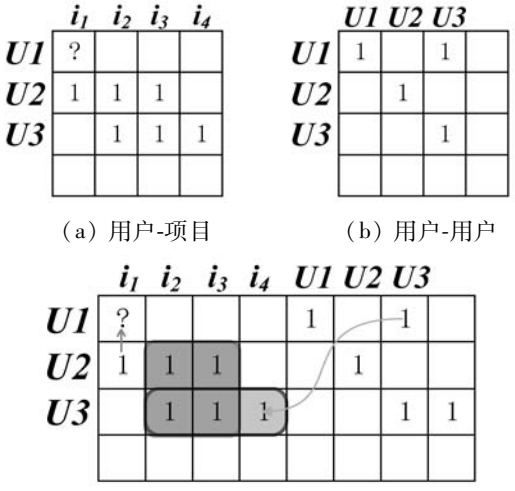


图 2　相交的用户信任网络实例

本文将项目作为节点引入信任网络,相似的用户偏好作为不同信任网络子图之间的纽带,用户可以通过项目节点连接不同的信任网络子图。此时,信任网络将不再是原始的信任网络,一个用户节点具有连接用户信任关系和用户相似偏好两种不同的边。用户 u_2 和用户 u_3 通过相似的用户偏好建立连接,项目 i_1 通过项目节点从用户 u_2 传递给用户 u_3,再通过信任传递从用户 u_3 最终推荐给用户 u_1。

2.1.2　基于矩阵分解传递信任

为了解决信任网络脱节的问题,首先,通过用户-项目评分和用户-用户信任关系分别构建两个矩阵,如图 3(a)、(b)所示。然后,通过合并这两个矩阵来构建模型,如图 3(c)所示。在模型中,项目评分和用户之间的信任关系视为同一层级,通过矩阵分解的方式可以同时传递信任和推荐项目。也就是说,给定 K 个潜在特征向量,每个用户通过这 K 个特征向量对每个项目进行评分,同时,每个用户通过相同的 K 个特征向量来信任其他的用户。



(a) 用户-项目　　(b) 用户-用户

(c) 用户-项目-用户

图 3　合并评分矩阵和信任关系矩阵

如图 3(c)所示,用户 u_2 和用户 u_3 之间不存在信任关系,但对项目 i_2 和项目 i_3 都具有显示反馈,本文认为用户 u_2 和用户 u_3 之间具有相同的特征向量。通过相同的特征向量,用户 u_2 可以将感兴趣的项目 i_1 推荐给用户 u_3,再通过用户之间的信任关系,最终用户 u_3 可以将项目 i_1 推荐给用户 u_1,从而解决了个性化推荐中广泛存在的信任网络脱节的问题。

2.2　项目流行度的加权策略

随着 Web 2.0 的发展,在所有其他因素相同的情况下,流行度较高的项目更有可能被用户认识。本文认为,项目流行度对用户评分项目和未评分项目分别有不同程度的影响。

2.2.1　项目流行度对未评分项目的影响

在推荐系统中,缺失数据(未评分项目)是负反馈和未知反馈的混合物,然而区分这两种情况是众所周知的困难。文献[12]提出,在缺失数据中,流行的项目更可能是用户的负反馈而非未知的反馈,因此,在模型训练的过程中应该给予更高的权重。本文引入文献[11]中的一个权重因子:

$$c_i = c_0 \times item\_pop(i) \tag{2}$$

$$item_{pop(i)} = \frac{f_i^{\partial}}{\sum_{j=1}^{N} f_j^{\partial}} \tag{3}$$

式中：c_i 表示缺失数据是负反馈的权重，f_i 代表项目 i 的流行度，通过项目 i 在数据中的频率来表示，c_0 代表缺失数据的总权重。指数 ∂ 控制项目流行度对缺失数据的影响。当 $\partial>1$ 时，会加强项目流行度对缺失数据权重的影响；$\partial\in(0,1)$ 时，会抑制项目流行度对缺失数据权重的影响；当 $\partial=0$ 时，$c_i=\frac{c_0}{N}=w_0$ 代表缺失数据赋予的统一权重 w_0，此时与文献[9]中的 w_0 相同。

2.2.2 项目流行度对评分项目的影响

随着互联网的发展，通过网络传播、媒体曝光、社区讨论等方式，流行度较高的项目，更有可能被用户认识，因此会获得更多偏离用户个人偏好的评分。对于矩阵中评分项目的数据，文献[8]认为项目的流行度越高，用户评分体现用户偏好的信息越少，相反，项目的流行度越低，用户评分体现用户偏好的信息越可靠。参考文献[8]，本文提出了一个与项目流行度相关的误差权重因子：

$$w_{ui}=\frac{\log(2)}{\log(1+f_i)} \tag{4}$$

式中：$r=w_{ui}r_{ui}$ 表示体现用户偏好的项目评分。f_i 最小为 1，此时 w_{ui} 为最大值 1。f_i 越低，w_{ui} 越大，用户项目评分 r_{ui} 越接近用户偏好评分 r；反之，w_{ui} 越小，用户项目评分 r_{ui} 越偏离用户偏好评分 r。

2.3 PopTruMF 算法

根据以上分析，本文设计了一个基于项目流行度的目标函数：

$$E_{ui}^2=\sum_{(u,i)\in\boldsymbol{R}}w_0(r-\hat{r}_{ui})^2+\sum_{k=1}^{K}c_i\,\hat{r}_{ui}^{\,2}+\frac{\beta}{2}\sum_{k=1}^{K}(p_{ik}^2+q_{ku}^2) \tag{5}$$

首先，本文对用户项目评分根据项目流行度进行预处理，得到更接近用户偏好的评分 r。式(5)中：第一项代表用户偏好评分预测的误差，在显示评分预测中广泛使用[12]；第二项代表缺失数据中负反馈对评分预测的影响[11]；第三项代表一个正则项来防止目标函数的过拟合。为了获得目标函数的最小值，本文通过在 p_{ik}、q_{ku} 上使用梯度下降的方法训练模型：

$$\frac{\partial}{\partial p_{ik}}E_{ui}^2=-2w_0\Big(r-\sum_{k=1}^{K}p_{ik}q_{ku}\Big)q_{ku}+2c_ip_{ik}q_{ku}^2+\beta p_{ik}=-2w_0e_{ui}q_{ku}+2c_ip_{ik}q_{ku}^2+\beta p_{ik} \tag{6}$$

$$\frac{\partial}{\partial q_{kj}}E_{ui}^2=-2w_0\Big(r-\sum_{k=1}^{K}p_{ik}q_{ku}\Big)q_{ik}+2c_iq_{ku}p_{ik}^2+\beta q_{ku}=-2w_0e_{ui}p_{ik}+2c_iq_{ku}p_{ik}^2+\beta q_{ku} \tag{7}$$

$$p'_{ik}=p_{ik}+\alpha(w_0e_{ui}q_{ku}-2c_ip_{ik}q_{ku}^2-\beta p_{ik}) \tag{8}$$

$$q'_{ku}=q_{ku}+\alpha(w_{ui}e_{ui}q_{ik}-2c_ip_{ik}q_{ku}^2-\beta q_{ku}) \tag{9}$$

基于项目流行度的加权策略，本文模型训练如算法 1 所示：

算法 1 Matrix Factorization

输入：合并矩阵 $\boldsymbol{R}$，潜在特征向量 $\boldsymbol{K}$，预测评分的误差权重 w_{ui}，缺失数据的权重 c_0

输出：用户潜在特征矩阵 $\boldsymbol{P}$，项目潜在特征矩阵 $\boldsymbol{Q}$

```
Randomly initialize P and Q;
for (u,i) ∈ R do r_ui = p_u q_i;
while Stopping criteria is not
  //Update user、item factors
  for i←1 to len (R)
    for u←1 to len (R_1)
      for k←1 to K
        w_ui, c_i←E_q.(2,3,4);
p_ik = p_ik + α(2w_ui e_iu q_ku − 2c_i p_ik q_ku^2 − β p_ik);
q_ku = q_ku + α(2w_ui e_iu p_ik − 2c_i p_ik q_ku^2 − β q_ku);
  for i ← 1 to len (R)
    for u ← 1 to len (R_i)
      E = e + w_0 × (r − Σ_{k=1}^{K} P_ik Q_ku)^2;
      for k ← 1 to K
        E = e + c_i (P_ik Q_ku)^2 + (β/2)‖P_ik‖^2 + Q_ku^2;
      If e < 0.001
      break;
Return P, Q · T;
```

预测评分由模型训练得到的用户特征向量 $\boldsymbol{p}_u$ 和项目特征向量 $\boldsymbol{q}_i$ 的内积产生。在预测评分构成的新矩阵中，由于信任矩阵部分通常被认为是算法的次要结果，本文对此先进行过滤，再生成最终的推荐列表。本文推荐算法的基本框架如图 4 所示，主要输入是用户-项目评分矩阵、用户-用户信任矩阵和项目的流行权重。该算法的主要步骤是合并矩阵、引入项目流行度的加权策略、模型训练、预测评分、过滤信任矩阵、生成推荐列表。

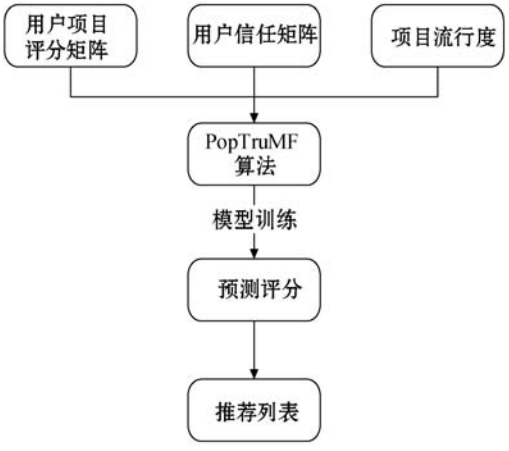


图 4 PopTruMF 模型基本框架

3　实验结果与分析

3.1　实验设置

数据集 本文采用公开的 Epinions 数据集,其中包括40 163 个用户对 139 738 个项目的664 824 条评分,评分数值在 1 ~5 之间,用户之间的信任信息包括487 181条,数据的稀疏度为 0.011 845 84%。由于原始数据集的高稀疏性,使得推荐算法在模型训练的过程中存在一定的难度。例如,一半以上的用户只有一个评论。对此,本文按照文献[13]的方法先进行数据过滤,使得本实验中数据的稀疏度为 99.89%。

(1) *度量方法*　我们采用均方根误差(RMSE)、准确率(precison)、覆盖率(coverage)、F 值四个性能指标评估推荐算法的推荐质量。

$$RMSE = \sqrt{\frac{\sum_{(u,i)\in D}(r_{ui} - \hat{r}_{ui})^2}{|D|}} \tag{10}$$

$$coverage = \frac{|R(u)|}{|I|} \tag{11}$$

式中:I 表示所有物品的集合,*coverage* 表示推荐列表中的物品占总物品数的比例。其中,RMSE 的值越小,代表推荐效果越好。由于实际评分和预测评分之间最大差值是 5 - 1 =4,本文算法的最差精度为:

$$precision = 1 - \frac{RMSE}{4} \tag{12}$$

为了更好地评估算法的推荐质量,本文引入 F 值指标,将准确率和覆盖率结合在一个数字度量中。其中,F 值越大,代表推荐算法的总体效果越好。

$$F = \frac{2 \times precision \times coverage}{precision \times coverage} \tag{13}$$

(2) *对比方法*　将本文算法与下面的算法进行比较,评估各算法的性能。

SocialMF 算法[4]　基于矩阵分解的方法,结合用户之间的信任关系,学习用户和项目的潜在特征向量,提出每个用户的特征向量依赖于社交网络中他直接邻居特征向量的加权平均。

SocialSVD 算法[2]　基于奇异值分解的方法,根据人际关系中的六度分隔理论计算用户之间信任度,填充用户信任矩阵,提高了预测的准确率。

TrustWalker 算法[5]　结合用户之间的信任关系,在信任网络上随机游走以预测项目评分。如果在特定的深度阈值内未找到该项目,则基于相似的项目预测项目评分。

3.2　实验结果及分析

本文采用上述的 Epinions 数据集,依次对SocialMF 算法、SocialSVD 算法、TrustWalker 算法和本文提出的合并矩阵 TruMF 算法以及完整的 PopTruMF 算法做对比实验。在实验中,我们设置潜在特征向量 $K = 25$、$\beta = 0.01$,且所有方法都是用 python 实现的,以便公平比较所有方法。实验结果如表 1 所示。

表 1　TruMF 算法与各算法实验结果

Method	RMSE	precison	Coverage/%	F
SocialMF	1.085	0.729	NA	NA
SocialSVD	1.059	0.735	82.8	0.779
TrustWalker	1.077	0.731	95.4	0.827
TruMF	1.172	0.707	98.7	0.824

如表 1 所示,本文提出的 TruMF 算法的 F 值最大,推荐效果最好。在现有的推荐算法中,TrustWalker 算法是推荐结果覆盖范围最广的方法之一,高达 95.4%,而本文的 TruMF 算法相比 TrustWalker 算法项目覆盖率略有提高,约 98.7%。TruMF 算法的覆盖率同样也高于其他两种算法,因为 SocialMF 算法和 Social SVD 算法都是基于信任网络中的信任传递进行推荐,因此存在信任网络脱节的问题。但在准确率方面,TruMF 算法分别低于 SocialMF 算法、SocialSVD 算法和 TrustWalker 算法。

本文结合项目流行度加权策略,在 TruMF 算法的基础上作出进一步优化,提出 PopTruMF 算法。其中 c_0 代表缺失数据的总权重,∂ 控制项目流行度对以上权重的影响。首先,本文将缺失数据赋予的统一权重,即 $\partial = 0$。如图 5 所示,当 c_0 取 128 左右时,RMSE 最小,达到预测准确性的峰值,而当 c_0 变小或设置的太大时,RMSE 的值显著升高,这说明在 TruMF 算法中适当考虑缺失数据的必要性。

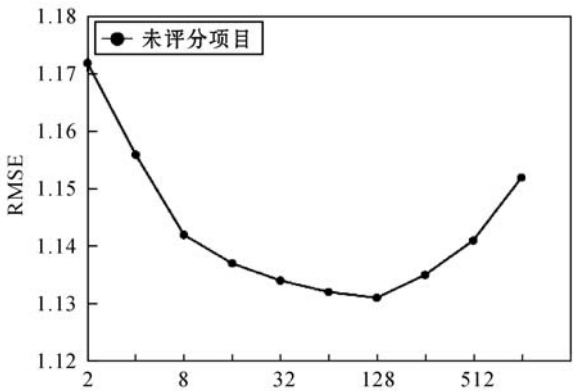


图 5　c_0($\partial = 0$)

然后,我们将 c_0 设置为 128,来观察项目流行度对

推荐效果的影响。如图 6 所示,随着 α 的增加,*RMSE* 的值会再次减小,到达一个最优值后又会大幅度升高,这说明在 TruMF 算法中适当考虑项目流行度的有效性。本文设置 $c_0 = 128$、$\partial = 0.4$,同时引入项目流行度的误差权重因子 w_{ui},适度增加 ∂,会达到更好的预测效果。如图 6 所示,根据项目流行度对用户评分项目和未评分项目分别进行加权处理,可以更有效提高推荐算法的准确率。

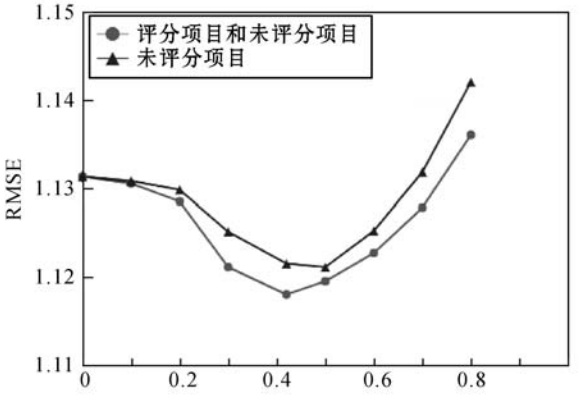


图 6 $\partial(c_0 = 128)$

PopTruMF 算法与各算法的实验结果如表 2 所示。可以看出,PopTruMF 算法的 F 值在 TruMF 算法的基础上得到进一步的改善,准确率和覆盖率得到了更好的折中。与前几种方法相比,PopTruMF 算法虽然损失了 5% 的 RMSE,但在推荐中覆盖项目的范围最广,能够给于用户更好的推荐效果。

表 2 PopTruMF 算法与各算法实验结果

Method	RMSE	precison	Coverage/%	F
SocialMF	1.085	0.729	NA	NA
SocialSVD	1.059	0.735	82.8	0.779
TrustWalker	1.077	0.731	95.4	0.827
TruMF	1.172	0.707	98.7	0.824
PopTruMF	1.118	0.721	98.9	0.834

4 结 语

针对现有推荐算法在推荐中项目覆盖范围有限的问题(如用户信任网络不相容、项目评分矩阵的高稀疏性、用户个性化推荐存在偏差等),本文提出了一种基于合并用户信任关系和项目流行度的 PopTruMF 算法,并讨论了如何通过信任关系将项目推荐给不同信任网络的用户,以及在推荐算法中考虑缺失数据和项目流行度的有效性。通过在 Epinions 上的实验表明,PopTruMF 算法在大幅度改善推荐覆盖率的同时,保证了推荐的准确率,能够给于用户更好的推荐效果。

在下一步的研究中,我们将进一步探索项目和用户基于多个潜在特征对个性化推荐的影响,以及通过给与用户实时的推荐,进一步提升模型性能。

参 考 文 献

[1] Golbeck J. Computing and Applying Trust in Web-based Social Networks [D]. University of Maryland College Park, 2005.

[2] 李卫疆, 齐静, 余正涛, 等. 融合信任传播和奇异值分解的社会化推荐算法[J]. 计算机工程, 2017,43(8): 236 – 242.

[3] Zhang H, Shen F, Liu W, et al. Discrete Collaborative Filtering[C]//the 39th International ACM SIGIR conference. ACM, 2016.

[4] Jamali M, Ester M. A Matrix Factorization Technique with trust Propagation for Recommendation in Social Networks [C]//Proceedings of the fourth ACM conference on Recommender systems. ACM, 2010: 135 – 142.

[5] Jamali M, Ester M. TrustWalker: A Random Walk Model for Combining Trust-Based and Item-Based Recommendation [C]//Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.

[6] Ratkiewicz J, Fortunato S, Flammini A, et al. Characterizing and Modeling the Dynamics of Online Popularity[J]. Physical Review Letters, 2010, 105(15): 158701.

[7] He H, Garcia E A. Learning from Imbalanced Data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263 – 1284.

[8] 郝立燕, 王靖. 基于项目流行度的协同过滤 TopN 推荐算法[J]. 计算机工程与设计, 2013, 34(10): 3497 – 3501.

[9] Hu Y, Koren Y, Volinsky C. Collaborative Filtering for Implicit Feedback Datasets[C]//Proceedings of the IEEE International Conference on Data Mining. IEEE, 2008: 263 – 272.

[10] Devooght R, Kourtellis N, Mantrach A. Dynamic Matrix Factorization with Priors on Unknown Values[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 189 – 198.

[11] He X, Zhang H, Kan M Y, et al. Fast Matrix Factorization for Online Recommendation with Implicit Feedback[C]// Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. ACM,2017: 549 – 558.

[12] Koren Y, Bell R. Advances in Collaborative Filtering[M]// Recommender systems handbook. Springer,2011:145 – 186.

(下转第 275 页)

引导爬虫抓取主题相关的网页。此外,通过核心内容提取算法准确定位网页核心内容的起始和终止位置,对网页噪声信息进行过滤能够准确抓取网页核心内容,提高了主题相似度计算的准确度,进而引导爬虫抓取更多主题相关的网页。因此,相较于 SSVSMC 以及 VSMC,TDSFC 抓取的网页平均相似度较高。

表 3　部分平均相似度对比实验数据

N	BFC	VSMC	SSVSMC	TDSFC
1 000	0.782	0.914	0.922	0.939
2 000	0.781	0.901	0.909	0.922
3 000	0.784	0.932	0.901	0.929
4 000	0.781	0.931	0.906	0.948
5 000	0.777	0.928	0.901	0.944

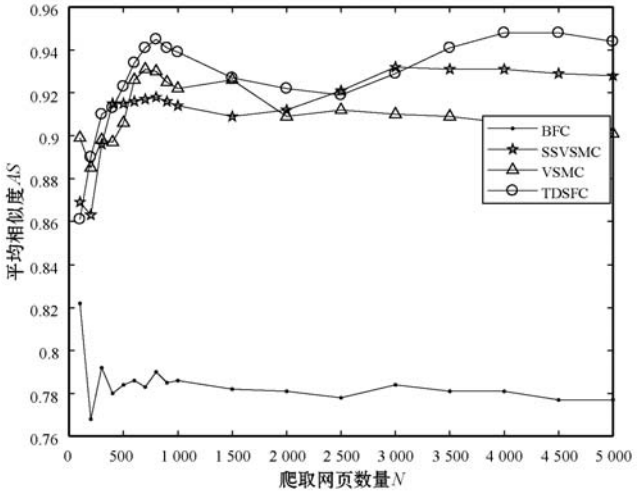


图 5　聚焦爬虫平均相似度对比图

由于爬取网页数量有限,TDSFC、SSVSMC 以及 VSMC 的平均相似度数值上差异并不明显,但是通过对表 3 数据以及图 5 分析可知,随着爬取网数量的增多,TDSFC 爬取准确度优势会更明显。

4　结　语

爬虫的爬取准确度和效率是衡量爬虫性能的重要指标。提出的结合文本密度的语义聚焦爬虫方法,通过使用网页标题结合 LCS 算法提取网页核心内容,提高了主题相似度计算的准确度。考虑词项语义信息引入主题相关度计算模型,结合主题重要度计算链接优先级,优化链接的提取。此外,为了提高全局搜索性能,结合关键词使用搜索引擎扩展链接集。结果显示,TDSFC 能够提高爬虫的爬取准确度和爬取效率。

TDSFC 虽能通过核心文本内容提取算法准确提取网页核心内容文本,但是整个爬虫仍然需要不断处理大量文本内容,爬取效率有待提高。下一步考虑提取网页文本内容特征用于主题相似度计算,进一步优化主题相似度计算模型。

参　考　文　献

[1] Khan M N A, Mahmood A. A distinctive approach to obtain higher page rank through search engine optimization[J]. Sādhanā, 2018, 43(3):43.

[2] 肖江, 季节. 基于 Heritrix 的主题爬虫在互联网舆情系统中应用[J]. 电子设计工程, 2015,23(6):30 - 32.

[3] 费晨杰,刘柏嵩. 基于 LDA 扩展主题词库的主题爬虫研究[J]. 计算机应用与软件,2018,35(4):49 - 54.

[4] Seyfi A. A Focused Crawler Combinatory Link and Content Model Based on T-Graph Principles[J]. Computer Standards & Interfaces, 2016, 43:1 - 11.

[5] Du Y, Liu W, Lv X, et al. An improved focused crawler based on Semantic Similarity Vector Space Model[J]. Applied Soft Computing, 2015, 36(C):392 - 407.

[6] Tarik B, Mahmoud D D, Zakaria E. Classifying Web Pages by Aimed Nation Using Machine Learning[J]. International Journal of Organizational and Collective Intelligence,2017,7(1):20 - 35.

[7] Gali N, Mariescu-Istodor R, Fränti P. Using linguistic features to automatically extract web page title[J]. Expert Systems with Applications, 2017, 79:296 - 312.

[8] 王飞,谭新. 一种基于 Word2vec 的训练效果优化策略研究[J]. 计算机应用与软件,2018,35(1):97 - 102,174.

[9] Wensen L, Zewen C, Jun W, et al. Short text classification based on Wikipedia and Word2vec[C]//IEEE International Conference on Computer & Communications. IEEE, 2017.

[10] Putri I, Kusumaningrum R. Latent Dirichlet Allocation(LDA) for Sentiment Analysis Toward Tourism Review in Indonesia[J]. Journal of Physics: Conference Series, 2017, 801: 012073.

[11] Yan W, Pan L. Designing focused crawler based on improved genetic algorithm[C]//Tenth International Conference on Advanced Computational Intelligence. IEEE, 2018.

(上接第 254 页)

[13] Rendle S, Freudenthaler C, Gantner Z, et al. BPR: Bayesian Personalized Ranking from Implicit Feedback[C]//Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2012: 452 - 461.

[14] Deqing L, Honghui M, Yi S, et al. ECharts: A Declarative Framework for Rapid Construction of Web-Based Visualization[J]. Visual Informatics, 2018,2(2): 136 - 146.