

融合社交网络的单类个性化协同排序算法

李 改^{1,2} 陈 强^{2,3} 李 磊² 潘进财¹

(顺德职业技术学院电子与信息工程学院 顺德 528333)¹

(中山大学数据科学与计算机学院 广州 510006)² (广东第二师范学院计算机科学系 广州 510303)³

摘 要 单类个性化协同排序算法的研究的核心思想是把单类协同过滤问题当成排序问题来看待。之前的研究仅仅使用了隐式反馈数据来对推荐对象进行排序,这限制了推荐的准确度。随着在线社交网络的出现,为了进一步提高单类个性化协同排序算法的准确度,提出了一种新的融合社交网络的单类个性化协同排序算法。在真实的包含社交网络的2个数据集上的实验验证了该算法在各个评价指标下的性能均优于几个经典的单类协同过滤算法。实验证明,社交网络信息对于提高单类个性化协同排序算法的性能具有重要作用。

关键词 推荐系统,协同排序,社交网络,单类协同过滤,隐式反馈数据

中图分类号 TP393 文献标识码 A DOI 10.11896/j.issn.1002-137X.2017.02.011

One-class Personalized Collaborative Ranking Algorithm Incorporating Social Network

LI Gai^{1,2} CHEN Qiang^{2,3} LI Lei² PAN Jin-cai¹

(Department of Electronic and Information Engineering, Shunde Polytechnic, Shunde 528333, China)¹

(School of Data and Computer Science, Sun Yat-Sen University, Guangzhou 510006, China)²

(Department of Computer Science, Guangdong University of Education, Guangzhou 510303, China)³

Abstract The research's key idea of one-class personalized collaborative Ranking Algorithm is to make use of partial order of items. In the early research of these problems, the training data are only implicit feedback dataset, this limits the sorting accuracy. With the advent of online social networks, in order to improve the performance of one-class personalized collaborative ranking algorithm, we proposed a new one-class personalized collaborative ranking algorithm incorporating social network. We conducted our experiment on two large real-world datasets with social information. The experiment results illustrate that our approach achieves a better performance than several traditional OCCF methods. Experiments also show that the social network information plays an important role in improving the performance of one-class personalized collaborative ranking algorithm.

Keywords Recommended systems, Collaborative ranking, Social network, One-class collaborative filtering, Implicit feedback dataset

1 引言

互联网的迅速发展以及各类电子商务的异军突起,使得互联网上的信息量呈几何级数增长,从而出现了“信息过载”问题——即尽管广大网民可以从互联网上轻易获取海量的信息,但这些信息中的绝大部分都是无用信息,如何从海量的互联网信息中获取自己所需的满足个性化需求的信息成了广大网民的急切需求。推荐系统正是在这种背景下产生的。基于协同过滤的推荐算法是推荐系统中运用最广泛的基础性核心算法^[1-3]。

协同过滤算法所处理的数据类型主要分为两类^[3-5]:1)显示评分(rating)数据,该类数据用1~5分/星表示用户对某个推荐对象的喜好程度,5分表示非常喜欢,1分代表厌恶;2)隐式反馈(implicit feedback)数据,不是在所有信息推荐系统中都能够收集到显式评分数据,在有些信息推荐系统中只能收集一些用户的行为信息(譬如用户是否阅读了某些新闻或购买了某些产品)。

在研究中一般假设用户的这些行为数据大致反映了用户的兴趣偏好,譬如用户购买了某个产品,可以认为用户是喜欢这个产品的,在这种假设下把这类行为数据称为基于隐式反

到稿日期:2015-11-11 返修日期:2016-03-12 本文受国家自然科学基金项目(61370186),广东省自然科学基金项目(2016A030310018),广东省科技计划项目(2014A010103040,2014B010116001),广州市科技计划项目(201604010049,201510010203),广东第二师范学院教授博士科研专项(2015ARF25),佛山市机电专业群工程技术开发中心2015年第二批开放课题(2015-KJZX139),广东省大学生科技创新培育专项(G2016Z08)资助。

李 改(1981—),男,博士,讲师,主要研究方向为数据挖掘、推荐系统、大数据处理,E-mail:ligai999@126.com;陈 强(1964—),男,教授,主要研究方向为知识发现、数据挖掘;李 磊(1951—),男,博士,教授,博士生导师,主要研究方向为人工智能、数据挖掘、数据库等;潘进财(1997—),男,主要研究方向为推荐系统。

馈的数据。单类协同过滤(OCCF)算法就是研究处理隐式反馈数据的协同过滤算法^[1-5]。单类协同过滤算法的任务是分析用户的这些隐式反馈信息,找出用户的偏好模型,进而按用户的喜欢程度对推荐对象进行排序。

目前,协同过滤算法的研究热点是引入上下文信息以进一步提高现有协同过滤算法的性能。这些上下文信息包括:用户和推荐对象的内容信息^[6-7]、用户和推荐对象的标签信息^[7]、社交网络信息^[8-12]。如 Ma, Zhu 和 Jamali 等人就显示评分数据和社交网络数据相结合提出的新的社会化推荐模型提高了推荐性能^[8-10]。由于引入了社交网络信息,因此这类算法能在一定程度上解决处理显式评分数据的信息推荐系统所面临的数据稀疏性问题。对于处理隐式反馈数据,Ding 等人提出了 CSTR 模型^[13],该模型在传统的矩阵分解模型上融入用户的社交网络信息和推荐对象的文本描述或内容信息,可进一步缓解隐式反馈数据的数据稀疏性和不平衡性对基于评分预测的 OCCF 算法的影响,从而进一步提高 OCCF 算法的性能。Purushotham 等人提出了 CTR with SMF 模型^[14],该模型把用户的社交网络信息和推荐对象的文本描述或内容信息融入传统的矩阵分解模型中,可进一步缓解隐式反馈数据的数据稀疏性和不平衡性对基于评分预测的 OCCF 算法的影响,从而进一步提高 OCCF 算法的性能。但到目前为止,以上引入了上下文信息的协同过滤算法的核心思想都是通过评分预测来对推荐对象进行个性化排序。对于利用上下文信息,目前还没有以排序学习的思想来解决单类协同过滤问题的研究。

为了进一步提高单类个性化协同排序算法的性能,本文提出了一种全新的融合社交网络的单类个性化协同排序算法(S-BPR-MF),即在 Rendle S 等提出的基于矩阵分解的单类个性化协同排序算法^[15]的基础上引入社交网络信息。为了验证本文所提算法的优越性,文中的实验部分在真实的包含社交网络信息的数据集中实现了本文所提出的新算法。多个实验证明,本文所提出的算法在 AUC 值和 MAP 值这两个评价指标下均优于当前最新的几个单类个性化协同排序算法。

本文第2节全面介绍最新相关工作;第3节介绍所需的基本定义以及传统的单类个性化协同排序算法;第4节详细介绍本文所提出的融合社交网络的单类个性化协同排序算法;第5节设计各种实验,以验证本文所提算法的优越性,并给出实验结果分析;最后是总结与展望。

2 相关工作

近年来广大科研工作者对单类协同过滤算法进行了深入研究,研究方向主要分为两大类:1)基于排序学习思想的单类协同排序算法^[1-3,15,18];2)基于评分预测的单类协同过滤算法^[4-5,8,13-14]。

之前单类协同过滤的研究主要是侧重于基于评分预测的思想来对推荐对象进行个性化排序。如 Zhang S 等提出运用奇异值分解(SVD)技术来解决单类协同过滤问题^[16]。Wang C 等人运用概率矩阵分解(PMF)技术来解决单类协同过滤问题^[5],在他们的模型中,观察到的点击数据作为正例数据,其余的混合数据都作为负例数据。Pan 等提出了 ALS 算法,此算法运用加权的低秩逼近算法(OCCF)来解决该类问题^[4]。

Hu 等人提出了一种引入了信任级的因子分解模型^[17],该模型的核心思想是通过给 OCCF 数据集中的正例数据和负例数据分别分配一个变化的信任级来弥补简单的区分正/负例数据所带来的偏差,该模型的核心思想与 ALS 算法近似。

随着信息检索领域排序学习算法的发展,也出现了一些基于排序学习思想的单类协同排序算法。如 Li 等人提出了 MERR_SVD++^[18],该模型融合显式评分和隐式反馈信息进一步提高了基于排序学习思想的单类协同排序算法的性能;Li 等人还提出了 PPMF 模型,该模型提出了一种新的损失函数来优化基于排序学习思想的单类协同排序算法。Rendle S 等提出一种基于 Bayesian 个性化协同排序的优化框架[BPR]^[15],该框架运用偏序排序学习的思想来解决单类协同过滤问题,可融入矩阵分解模型[MF]和最近邻模型[KNN]。近年来有学者还提出了一些 BPR 模型的改进模型,如 Li 等人提出了 SPR 和 FPR 模型在一定程度上解决了 BPR 模型的噪声敏感性问题。Pan 等人提出了一种新的 GBPR 模型^[18],该模型的核心思想是假定一个用户的偏好受其所在兴趣组的其他用户的影响,其进一步提高了基于排序学习思想的二维单类协同排序算法的性能。本文继续深入研究 BPR 模型,在 BPR 模型中融入社交网络信息,以期进一步提高基于排序学习思想的单类协同排序算法的性能。

3 基本定义和传统的单类个性化协同排序算法

3.1 基本定义

本文用大写斜体字母(如 X)表示矩阵,用小写字母(如 i, j)表示标量。给定矩阵 X, X^T 表示矩阵 X 的转置, X_{ij} 表示它的一个元素, $X_{:,j}$ 表示矩阵 X 的第 j 列, X_i 表示矩阵 X 的第 i 行。本文中,矩阵 X 表示隐式评分矩阵,该矩阵具有 n 个用户、 m 个对象; $\hat{X}$ 表示矩阵 X 的逼近/预测矩阵; U 表示所有用户的集合, I 表示所有推荐对象的集合;用户和推荐对象的特征矩阵分别用矩阵 W 和 H 表示; K 表示特征个数。

3.2 传统的单类个性化协同排序算法

有关单类个性化协同排序算法最经典的研究成果是由 Rendle S 等人提出的基于 Bayesian 函数的个性化协同排序框架(BPR)^[15]。BPR 框架的核心思想是运用推荐对象之间的偏序关系来构造推荐模型对推荐对象进行个性化排序。而非像传统的单类协同过滤算法那样通过基于回归的思想来预测评分矩阵中的空值,进而对推荐对象进行个性化排序。

定义基于用户的成对偏序数据集 $D_S: U \times I \times I$ 为训练数据集, $D_S = \{(u, i, j) | i \in I_u^+, j \in I_u^+, i \neq j\}$, 其中 $I_u^+ = \{i \in I | (u, i) \in X\}$ 表示用户 u 浏览过的推荐项目集合, I_u^- 表示 u 没有浏览过的推荐项目集合, 因此 $(u, i, j) \in D_S$ 表示用户 u 浏览过推荐项目 i 但没有浏览过推荐项目 j 。

BPR 框架的优化目标是找到任意一个模型去最大化如下所示的面向用户的成对偏序数据集的后验概率函数:

$$BPR-OPT = - \sum_{(u, i, j) \in D_S} \ln \sigma(\hat{X}_{ui} - \hat{X}_{uj}) + \lambda_\theta (\theta) \quad (1)$$

其中, θ 表示所选择的任意一个模型的参数, λ_θ 表示该模型的正则化参数, $\hat{X}_{ui} = \hat{X}_u - \hat{X}_j, \sigma(x) = 1/(1 + \exp(-x))$ 。

如果选择的任意模型是矩阵分解模型,则把这样的 BPR

框架叫作基于矩阵分解的 BPR 算法(BPR-MF),这时

$$\hat{X}_{ui} = \sum_{k=1}^K W_{uk} H_{ik} \quad (2)$$

为了估计用户 u 是否浏览过推荐项目 i 而没有浏览过推荐项目 j 。BPR-MF 算法需要优化下面的目标函数:

$$\Phi_1 = -\sum_{u=1}^n \sum_{i \in I_u^+} \sum_{j \in I_u^-} \ln \sigma(\hat{X}_{uij}) + \lambda(\|W\|^2 + \|H\|^2) \quad (3)$$

其中,参数 λ 是用于控制模型复杂度的正则化系数。BPR-MF 算法的具体求解过程见文献[15]。

4 融合社交网络的单类个性化协同排序算法

4.1 融合社交网络的单类个性化协同排序算法简介

为了进一步提高传统的单类个性化协同排序算法的性能,本节在传统的基于矩阵分解的 BPR 算法(BPR-MF)的基础上引入社会化规范。把本节所提出的算法称为融合社交网络的单类个性化协同排序算法(S-BPR-MF)。

在此假定,社交网络中互为朋友关系的两个用户之间的特征向量相似,但如果两个用户拥有不同的评分模式,那么这两个用户之间的特征向量的相似程度也不同。为了对社交网络中互为朋友关系的两个用户间的特征向量进行约束,在此引入如下社会化正则项^[8]:

$$\frac{\beta}{2} \sum_{f \in F(u)} \text{Sim}(u, f) \|W_u - W_f\|_F^2 \quad (4)$$

这里,添加社会化正则项的目的是使用户 u 的特征向量尽可能地与他的朋友们的特征向量相似。参数 β 用于控制社交网络信息的影响。 $F(u)$ 是用户 u 的朋友集。 $\|W_u - W_f\|_F^2$ 用于表示用户 u 和他的朋友 f 之间的偏好差异,该值越大,则 W_u 和 W_f 的差异越大,说明用户 u 和他的朋友 f 之间的偏好差异越大。其中 $\text{Sim}(u, f)$ 表示用户 u 和他的朋友 f 之间在评分模式上的相似度,可以用皮尔生相关系数(PCC)来计算,该相似度越小,则用户 u 的特征向量与他的朋友 f 的特征向量的差异越大。给定朋友间相似度矩阵 $S, S: U \times U$,如果用户 u 和用户 f 是朋友关系,则: $S_{uf} = \text{Sim}(u, f)$, 否则 $S_{uf} = 0$ 。因为 $\text{Sim}(u, f)$ 越小,则朋友 f 对用户 u 的特征向量的影响越小,故在这里采用 $\text{Sim}(u, f)$ 与 $\|W_u - W_f\|_F^2$ 相乘后的公式 $\text{Sim}(u, f) \|W_u - W_f\|_F^2$ 来表示用户 u 和他的朋友 f 通过相似度加权后的偏好差异。

通过在传统的基于矩阵分解的 BPR 算法(BPR-MF)中引入社会化正则项,得到了本文所提出的融合社交网络的单类个性化协同排序算法(S-BPR-MF)。

$$\Phi_2 = -\sum_{u=1}^n \sum_{i \in I_u^+} \sum_{j \in I_u^-} \ln \sigma(\hat{X}_{uij}) + \lambda(\|W\|^2 + \|H\|^2) + \frac{\beta}{2} \sum_{u=1}^n \sum_{f \in F(u)} S_{uf} \|W_u - W_f\|_F^2 \quad (5)$$

在式(5)上,为了最小化损失函数,通过对 W_u, H_i 和 H_j 实施梯度下降法,得到:

$$\frac{\partial \Phi_2}{\partial W_u} = -\frac{\exp(-\hat{X}_{uij})}{1 + \exp(\hat{X}_{uij})} (H_i - H_j) + \lambda W_u + \beta \sum_{f \in F(u)} S_{uf} (W_u - W_f) \quad (6)$$

$$\frac{\partial \Phi_2}{\partial H_i} = -\frac{\exp(-\hat{X}_{uij})}{1 + \exp(\hat{X}_{uij})} W_u + \lambda H_i \quad (7)$$

$$\frac{\partial \Phi_2}{\partial H_j} = -\frac{\exp(-\hat{X}_{uij})}{1 + \exp(\hat{X}_{uij})} W_u + \lambda H_j \quad (8)$$

总结上述分析过程,S-BPR-MF 算法的求解算法如下所示。

算法 1 融合社交网络的单类个性化协同排序算法(S-BPR-MF)

输入:隐式评分矩阵 X ,特征矩阵的秩 K ,参数 λ 和 β ,学习率 α

输出:逼近/预测矩阵 $\hat{X}$ 。

计算朋友间的相似度矩阵 S

随机初始化特征矩阵: W 和 H

repeat

从数据集 D_s 中随机抽取一个偏序评分对 (u, i, j) 。

$$\hat{X}_{uij} \leftarrow \hat{X}_{ui} - \hat{X}_{uj}.$$

$$W_u \leftarrow W_u + \alpha \frac{\partial \Phi_2}{\partial W_u}.$$

$$H_i \leftarrow H_i + \alpha \frac{\partial \Phi_2}{\partial H_i}.$$

$$H_j \leftarrow H_j + \alpha \frac{\partial \Phi_2}{\partial H_j}.$$

until 收敛。

$$\hat{X} = WH^T.$$

return $\hat{X}$ 。

算法 1 中的特征矩阵 W 和 H 的特征值初始化时满足 $N(0, 0.01)$ 。

4.2 计算复杂度分析

本文所提出的 S-BPR-MF 算法的计算复杂度主要与目标函数 Φ_2 及其几个偏导项 $\frac{\partial \Phi_2}{\partial W_u}, \frac{\partial \Phi_2}{\partial H_i}, \frac{\partial \Phi_2}{\partial H_j}$ 的计算复杂度相关。分析式(5),得到 Φ_2 的计算复杂度为 $O(n\bar{r}(m-\bar{r})K)$, m 表示推荐对象数, K 表示特征矩阵 W 和 H 中的特征数, $\bar{r}$ 表示用户的平均评分点个数, n 表示用户数。分析式(6)~式(8),得到偏导项 $\frac{\partial \Phi_2}{\partial W_u}, \frac{\partial \Phi_2}{\partial H_i}$ 和 $\frac{\partial \Phi_2}{\partial H_j}$ 的计算复杂度分别为 $O(F_u + K), O(K)$ 和 $O(K)$ 。 F_u 表示用户的朋友数的均值。故算法迭代一轮的时间复杂度为 $O(n\bar{r}(m-\bar{r})K + F_u + K)$, 假定算法在迭代 d 轮后收敛,则整个算法的复杂度为 $O(d(n\bar{r}(m-\bar{r})K + F_u + K))$ 。从上述分析可知,算法复杂度与数据集中评分点的总数 $n\bar{r}$ 成正比,而在实际的应用中,数据集往往非常稀疏,也就是 $n\bar{r}$ 很小,因此该算法非常适合处理大规模数据集。

5 实验结果及分析

本节首先介绍实验所采用的数据集及评价标准;然后给出本文所提出的 S-BPR-MF 算法中的规范化参数 λ 和 β 对实验结果的影响;最后给出 S-BPR-MF 算法和几个传统的单类协同过滤算法的实验比较。

5.1 实验数据集

Delicious^[14] 和 Epinions^[19] 数据集是两个被广泛运用的

公用数据集,均包含社交网络数据。Delicious 数据集中包括用户的社交网络数据和隐式反馈数据,前者以用户所信任的朋友的关系存在。Delicious 数据集中包含 69226 个不同的推荐对象和 1867 个用户。总的隐式反馈数据点的个数是 104799 个。在用户社交网络中,共有 15328 个信任关系。

Epinions 数据集中的用户的社交网络信息也以用户所信任的朋友的关系存在。Epinions 数据集中包含 139738 个不同的推荐对象和 49290 个用户。总的评分点是 664824 个。在用户社交网络中,共有 511799 个信任关系。在此数据集中,用户对推荐对象的评分是 1~5 分的整数。为了将数据集变为隐式数据集,所有有评分数据的数据点均重新赋值为 1,其他没有评分数据的数据点赋值为 0。这样就使得 Epinions 数据集变成了适用于单类协同过滤算法的数据集。

5.2 实验的评价标准

这里采用的评价标准是 AUC 值^[15]和 MAP 值^[2],两个评价指标的详细说明见相应参考文献,下面给出计算公式:

$$AUC = \frac{1}{|U|} \sum_u \frac{1}{|E(u)|} \sum_{(i,j) \in E(u)} \delta(x_{ui} > x_{uj}) \quad (9)$$

在这里, δ 是一个指标函数:

$$\delta(b) = \begin{cases} 1, & \text{if } b \text{ is true} \\ 0, & \text{else} \end{cases}$$

每个用户的评估对象对如下:

$$E(u) = \{(i, j) | (u, i) \in S_{\text{test}} \wedge (u, j) \notin (S_{\text{test}} \vee S_{\text{train}})\}$$

AUC 值越高,表示该算法的性能越好。由随机模型产生的 AUC 值为 0.5,理想模型的 AUC 值为 1。

$$MAP = \frac{\sum_{u=1}^n AP_u}{n}$$

$$AP_u = \frac{\sum_{k=1}^m \text{prec}(k) \times \text{pref}(k)}{P_u}$$

$$\text{prec}(k) = \frac{\sum_{i=1}^k \text{pref}(i)}{k}$$

其中, m 是测试集中推荐项目的个数。 $\text{pref}(i) \in \{0, 1\}$, 测试集经过算法排序后的第 i 项是正例项时 $\text{pref}(i) = 1$, 否则 $\text{pref}(i) = 0$ 。 P_u 是用户 u 点击过的推荐项目的个数。 MAP 值越高,表示该算法的性能越好。

反复运行每个实验 10 次,构成新的训练集和测试集,最终结果取 10 次运算结果的平均值。

5.3 实验结果

各个算法的程序代码用 Java 开发工具 Eclipse 编写。所有算法均在装有 Linux 系统环境的计算机下运行,计算机配置环境是内存 7.6GB,4 核 CPU,每个核主频是 3.2GHz。反复运行各个算法 10 次后取平均结果作为该算法的最终运行结果。

5.3.1 社会化参数 β 对 S-BPR-MF 实验结果的影响

实验中可以通过交叉确认的方法学习得到正则化参数 λ ,在本文中设置该 λ 为 0.0125。算法中参数 β 用于控制社交网络信息对 S-BPR-MF 算法性能的影响。在极端情况下参数 $\beta=0$,这表示在信息推荐系统中不存在社交网络,或者信息推荐算法在推荐过程中用户之间的社交网络信息对推荐结果没有产生任何积极作用,这时 S-BPR-MF 算法退化为传统的单类个性化协同排序算法。图 1 显示了在 Epinions 数据集

上参数 β 对 S-BPR-MF 算法的性能的影响,此时特征数设为 30,纵轴表示算法 S-BPR-MF 的 AUC 值,横轴表示参数 β 值的变化。从图 1 不难看出:随着参数 β 的增大, S-BPR-MF 算法的 AUC 值首先快速上升,随后又逐渐降低,参数 β 的最优值是 0.01。这个实验结果正好符合现实世界中朋友之间相互推荐的实际效果,如果把自身的偏好和朋友的推荐结合起来考虑,则能更准确地找到符合自己喜好的推荐对象。如果仅仅只是考虑朋友的推荐而忽略自身的偏好,或者仅仅只是考虑自身偏好而忽略朋友的推荐,则都会走向极端,不能得到最好的推荐效果。实验结果显示,在 Delicious 数据集上参数 β 取值也为 0.01 时算法的性能最优。

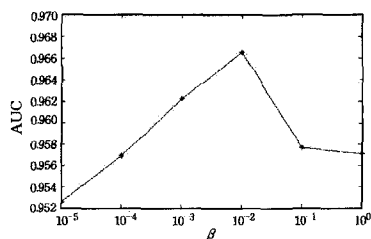


图 1 参数 β 对算法 S-BPR-MF 性能的影响

5.3.2 S-BPR-MF 算法和几个经典的单类协同过滤算法的性能比较

本节将在 2 个数据集上把本文所提出的 S-BPR-MF 算法与如下 5 个经典的单类协同过滤算法作比较。

S-OCCF: 基于评分预测的社会化单类协同过滤算法,也就是在 CTR with SMF 模型中去掉推荐对象的文本描述信息后的模型,其核心思想是在基于评分预测的单类协同过滤算法(ALS-OCCF)中引入社交网络信息。

GBPR: BPR 模型的扩展模型,其核心思想是放弃了 BPR 模型中有关用户之间相互独立的假设,认为一个用户的偏好受其所在的兴趣组的其他用户的影响。该算法是当前已知的最新的单类协同排序算法。

BPR-MF: 基于 Bayesian 个性化协同排序的优化框架的矩阵分解模型。该算法是最经典的单类协同排序算法。

ALS-OCCF: 运用加权的低秩逼近算法来解决(OCCF)问题的算法。

其中, S-OCCF 是本文在 CTR with SMF 模型的基础上去掉推荐对象的文本描述信息后抽象出来的基于评分预测的社会化单类协同过滤算法; GBPR 算法是当前已知的最新的单类协同排序算法; BPR-MF 算法是最经典的单类协同排序算法; ALS-OCCF 是当前已知的最新的基于评分预测的单类协同过滤算法。

实验结果如图 2—图 5 所示。

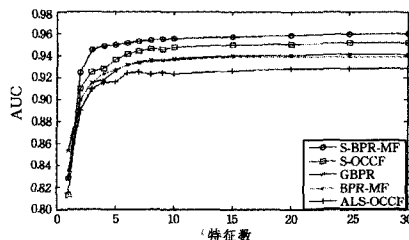


图 2 S-BPR-MF 算法与其他经典单类协同过滤算法的比较 (Delicious/AUC)

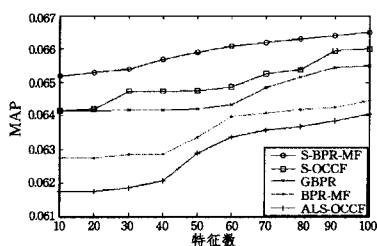


图3 S-BPR-MF算法与其他经典单类协同过滤算法的比较 (Delicious/MAP)

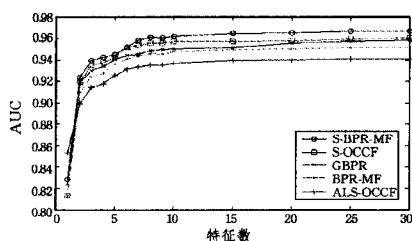


图4 S-BPR-MF算法与其他经典单类协同过滤算法的比较 (Epinions/AUC)

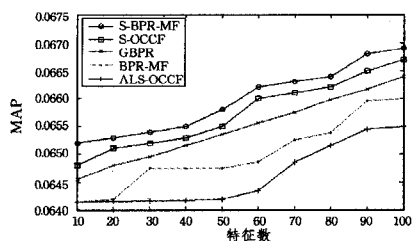


图5 S-BPR-MF算法与其他经典单类协同过滤算法的比较 (Epinions/MAP)

图2和图3表示在 Delicious 数据集上本文所提出的 S-BPR-MF 算法与传统的 S-OCCF 算法、GBPR 算法、BPR-MF 算法和 ALS-OCCF 算法分别在 AUC 值和 MAP 值这两个评价指标下的性能对比。图4和图5表示在 Epinions 数据集上本文所提出的 S-BPR-MF 算法与传统的 S-OCCF 算法、GBPR 算法、BPR-MF 算法和 ALS-OCCF 算法分别在 AUC 值和 MAP 值这两个评价指标下的性能对比。图2和图4中的纵轴表示各个算法的 AUC 值,横轴表示各个算法中用户/推荐对象的特征矩阵中的特征个数,特征个数从 1 逐渐增大到 30。图3和图5中的纵轴表示各个算法的 MAP 值,横轴表示各个算法中用户/推荐对象的特征矩阵中的特征个数,但此时特征个数从 1 逐渐增大到 100。通过实验验证, S-BPR-MF 算法在 Delicious 和 Epinions 数据集上取得最优 AUC 值和 MAP 值的正则化参数 λ 均为 0.0125。如图2—图5所示,在各个特征数下,本文所提出的 S-BPR-MF 算法均优于传统的 S-OCCF 算法、GBPR 算法、BPR-MF 算法和 ALS-OCCF 算法,并且这种优势随着特征数的增加越发明显。本文所提算法性能的显著提高归功于在基于排序学习思想的 BPR 算法中引入了社交网络信息。其中, S-BPR-MF 算法优于传统的 S-OCCF 算法表明:在基于排序学习思想的 BPR 算法中引入社交网络信息比在基于评分预测的单类协同过滤算法(ALS-OCCF)中引入社交网络信息更能有效地提高 OCCF 算法的性能; S-OCCF 算法的性能优于 GBPR 算法和 BPR 算法表

明:在基于评分预测的单类协同过滤算法(ALS-OCCF)中引入社交网络信息比纯粹使用隐式评分数据的单类协同排序算法(GBPR 算法和 BPR 算法)也能够更有效地提升 OCCF 算法的性能。GBPR 算法的性能优于 BPR 算法,这与文献[18]中的实验结果一致。BPR-MF 算法的性能优于 ALS-OCCF 算法,这也与文献[15]中的实验结果一致。

结束语 本文在传统的单类个性化协同排序算法(BPR 算法)的基础上引入了社交网络信息,提出了一种新的融合社交网络的单类个性化协同排序算法,并以 AUC/MAP 值为性能评价标准,在 2 个包含用户社交网络信息的数据集上将其与传统的 S-OCCF 算法、GBPR 算法、BPR-MF 算法和 ALS-OCCF 算法的性能进行了全面的比较,实验结果证明本文所提出的 S-BPR-MF 算法远优于其他几种经典的单类协同过滤推荐算法。本研究成果证明,社交网络信息能显著提高单类个性化协同排序算法的性能。在以后的工作中还将考虑 S-BPR-MF 算法在大数据下的并行化问题和实时推荐问题。

参考文献

- [1] LI G, OU W H. Pairwise Probabilistic Matrix Factorization for Implicit Feedback Collaborative Filtering [J]. Neurocomputing, 2016, 204: 17-25.
- [2] LI G, WANG L Y, OU W H. Robust Personalized Ranking from Implicit Feedback [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2016, 30(1): 1-28.
- [3] LI G, CHEN Q. Exploiting Explicit and Implicit Feedbacks for Personalized Ranking [J]. Mathematical Problems in Engineering, 2016, 2016: 1-11.
- [4] PAN R, ZHOU Y, CAO B, et al. One-class collaborative Filtering [C]//Proceedings of the IEEE International Conference on Data Mining. 2008: 502-511.
- [5] WANG C, BLEI D M. Collaborative topic modeling for Recommending scientific articles [C]//Proceedings of the 2011 Conference of the Knowledge Discovery and Data Mining. California, 2011: 448-45.
- [6] GU Q, ZHOU J, DING C. Collaborative filtering: Weighted Nonnegative Matrix Factorization Incorporating User and Item Graphs [C]//Proceedings of the 2010 SIAM Conference on Data Mining. 2010: 199-210.
- [7] ZHEN Y, LI W, YEUNG D. TagiCofi: tag informed collaborative filtering [C]//Proceedings of the Fifth ACM Conference on Recommender Systems. 2009: 69-76.
- [8] MA H, ZHOU D Y, LIU C, et al. Recommendation Systems with Social Regularization [C]//Proceedings of the 4th ACM International Conference on Web Search and Data Mining. Hongkong, China: ACM, 2011: 287-296.
- [9] ZHU J K, MA H, CHEN C, et al. Social Recommendation Using Low-Rank Semidefinite Program [C]//Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI, 2011: 158-163.
- [10] JAMALI M, ESTER M. A Matrix Factorization Technique with Trust Propagation for Recommendation in Social Networks [C]//Proceedings of the Twenty-third International Conference on

感知的数据融合取得了较高的 F 值,且都在 0.93 以上。

$$F\text{-measure} = \frac{2P \cdot R}{P + R} \quad (6)$$

表 3 时间感知的数据融合算法 TDF 的性能

Entity Type	P	R	F-Measure
Author	0.92	0.94	0.93
Paper	0.97	0.95	0.96

T-EI 的可伸缩性:不同数据量下的执行时间如图 3 所示。从图 3 中可以看出,在两种实体类型(作者(Author)、论文(Paper))上随着数据量的增大,执行时间缓慢增加,这说明 T-EI 具有较好的可伸缩性。

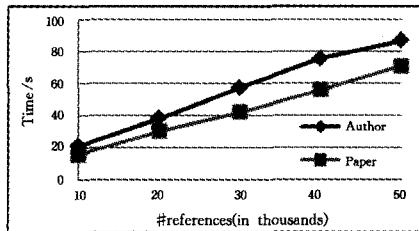


图 3 T-EI 的可伸缩性

结束语 针对异构信息空间中来自异构数据源的具有时间信息的异构实体,提出一个时间感知的实体集成框架 T-EI。在实体识别和数据融合阶段,提出时间感知的实体识别算法 TC 和时间感知的数据融合算法 TDF。在真实数据集上的实验结果表明了 T-EI 的可行性和有效性。

参考文献

- [1] IOANNOU E, NIEDEREE C, WOLFGAN N. Probabilistic entity linkage for heterogeneous information spaces [C] // Proc of CAISE, 2008.
- [2] LORENZO, HACID, PAIK, et al. Data integration in mashups [J]. SIGMOD Record, 2009, 38(1): 59-66.
- [3] EENDRULLIS S, THOR A, RAHM E. WETSUIT: an efficient mashup tool for searching and fusing web entities[J]. PVLDB, 2012, 5(12): 1970-1973.
- [4] THOR A, RAHM E. CloudFuice: A flexible cloud-based data integration System [C] // Proc of 10th Intl. Conference on Web Engineering (ICWE), 2011.
- [5] HERNADEZ M, KOUTRIKA G, KRISHNAMURTHY R. HIL: a high-level scripting language for entity integration [C] // Proc of EDBT, 2013.
- [6] LI P, DONG X. Linking temporal records [C] // Proc. of VLDB, 2011.
- [7] CHIANG Y H, DOAN A H, NAUGHTON J F. Modeling Entity evolution for temporal record matching [C] // Proc. of SIGMOD, 2014: 1175-1186.
- [8] CHIANG Y H, DOAN A H, NAUGHTON J F. Tracking entities in the dynamic world: a fast algorithm for matching temporal records [J]. PVLDB, 2014, 7(6): 469-480.
- [9] LI F R, LI M, HSU W, et al. Linking temporal records for profiling entities [C] // Proc. of SIGMOD, 2015: 593-605.
- [10] BLEIHOLDER J, NAUMANN F. Data Fusion [J]. ACM Comput. Surv., 2008, 41(1): 137-153.
- [11] DONG X, NAUMANN F. Data fusion-resolving data conflicts for integration [C] // Proc. of VLDB, 2009.
- [12] LIU X, DONG X L, Ooi B C, et al. Online data fusion [J]. PVLDB, 2011, 4(12): 932-943.
- [13] FAN W F, GEERTS F, et al. Determining the currency of data [C] // Proc. of PODS, 2011.
- [14] FAN W F, GEERTS F, TANG N, et al. Inferring data currency and consistency for conflict resolution [C] // Proc. of ICDE, 2013: 470-481.
- [15] ITHOFF T, DONG X L, MURPHY K. TimeMachine: Timeline Generation for Knowledge-Base Entities [C] // Proc. of KDD, 2015.
- [16] ALEXE B, ROTH M, WANG C T. Preference-aware Integration of Temporal Data [J]. PVLDB, 2014, 8(4): 365-376.
- [17] LU W, LOANNIDIS S, BHAGAT S, et al. Optimal Recommendations under Attraction, Aversion, and Social Influence [C] // Proceedings of the 20th International Conference on Knowledge Discovery and Data Mining, New York, USA: ACM, 2014: 657-666.
- [18] YAO W L, HE J, HUANG G Y, et al. SoRank: Incorporating Social Information into Learning to Rank Models for Recommendation [C] // Proceedings of the 23th ACM International Conference on World Wide Web, Seoul, Korea: ACM, 2014: 409-410.
- [19] DING X T, JIN X M, LI Y J, et al. Celebrity Recommendation with Collaborative Social Topic Regression [C] // Proceedings of the Twenty-third International Conference on Artificial Intelligence, Beijing, China: ACM, 2013: 2612-2618.
- [20] PURUSHOTHAM S, LIU Y, KUO C. Collaborative topic regression with social matrix factorization for recommendation systems [C] // Proceedings of the 29th ACM International Conference on Machine Learning, Edinburgh, Scotland, UK: ACM, 2012: 1255-1265.
- [21] RENDLE S, FREUDENTHALER C, GANTNER Z, et al. BPR: Bayesian personalized ranking from implicit feedback [C] // Proceedings of the 22nd International Conference on Uncertainty in Artificial Intelligence, Montreal, Canada, 2009: 452-461.
- [22] PATEREK A. Improving regularized singular value decomposition for collaborative filtering [C] // Proceedings of KDD Cup and Workshop, ACM Press, 2007: 39-42.
- [23] HU Y, KOREN Y, VOLINSKY C. Collaborative filtering for implicit feedback datasets [C] // Proceedings of the IEEE International Conference on Data Mining, Pisa, Italy: IEEE, 2008: 263-272.
- [24] PAN W K, CHEN L. GBPR: Group Preference based Bayesian Personalized Ranking for One-Class Collaborative Filtering [C] // Proceedings of the Twenty-third International Conference on Artificial Intelligence, Beijing, China: ACM, 2013: 3007-3011.
- [25] YU L, PAN R, LI Z F. Adaptive social similarities for recommender systems [C] // Proceedings of the fifth ACM Conference on Recommender Systems (RecSys), 2011: 257-260.

(上接第 92 页)

Artificial Intelligence. Barcelona, Catalonia, Spain: ACM, 2011: 2644-2649.