结合全局与双重局部信息的社交推荐

钱付兰[1,2] 李启龙[1]

(安徽大学计算机科学与技术学院 合肥 230601)[1]

(安徽大学计算智能和信号处理教育部重点实验室 合肥 230601)[2]

摘 要 随着Web2.0的飞速发展,社交推荐逐渐成为推荐领域近几年的研究热点。如何更有效地利用用户的社交关系是社交推荐的关键,目前的社交推荐算法主要引入的是用户之间的直接联系(明确关系)。将社交关系进一步细分为明确关系和隐含关系,并结合历史评分得到的用户声誉信息刻画了由用户全局信息(声誉)与局部信息(明确关系和隐含关系)所构成的推荐系统框架。与现有的社交推荐算法相比,所提出的算法更全面地分析了用户的社交关系,且具有良好的可解释性。在Douban数据集和Epinions数据集上进行了实验,并将本算法与主流的推荐算法进行了比较,结果表明本算法具有更好的推荐精度。

关键词 社交推荐,矩阵分解,声誉,隐含关系

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.2.012

Social Recommendation Combining Global and Dual Local Information

QIAN Fu-lan[1,2] LI Qi-long[1]

(School of Computer Science and Technology, Anhui University, Hefei 230601, China)[1]

(Key Laboratory of Intelligent Computing & Signal Processing, Ministry of Education, Anhui University, Hefei 230601, China)[2]

Abstract With the rapid growth of Web2.0, social recommendation has become one of the hot research topics in the last few years. It is the key point to improve recommender systems using social contextual information in a more efficient way. The existing social recommendation approaches mainly take advantage of user's direct connection(explicit relation). This paper detailed social relation as explicit relation and implicit relation and obtained the user's reputation by using his/her historic records. Then we proposed a recommendation framework capturing user's global social relation (reputation)and local social relation(explicit relation and implicit relation). Using two real datasets, Douban and Epinions, we conducted a experimental study to investigate the performance of the proposed model GDLRec. We compared our approach with existing representative approaches. The results show that GDLRec outperforms other methods in terms of prediction accuracy.

Keywords Social recommendation, Matrix factorization, Reputation, Implicit relation

1 引言

在过去的十年间,由于Web2.0的飞速发展,社交网络的用户在网站上的互动信息越来越多。Facebook的月活跃用户数达到13亿,而国内的新浪微博每月也有1.76亿的活跃用户。在推荐研究领域,研究者们将社交关系引进传统推荐算法,形成了近几年的社交推荐研究热点。

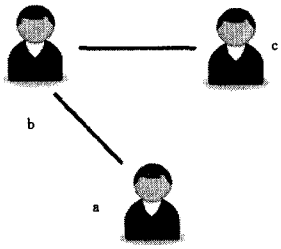
社会学和心理学有一些社会关系理论为社交推荐的可行性提供了理论支持。比如同质性(homophily)[1,2]:品味相似的人更有可能存在社交关系。而另一个理论,社会影响(social influence)[2,3]显示存在社交关系的人将很有可能有着相似的品味。根据这些理论,我们可以认为,在社交网络中,直接联系的两个人之间存在相似性,可以将这种联系称为明确关系(Explicit Relation);而那些没有直接联系却有共同朋友的人,也会对彼此产生一定的影响,这种联系定义为隐含关系(Implicit Relation)。现有的社交推荐算法主要考虑的是用户的明确关系,而对于隐含关系的研究仍然处于起步的阶段。

本文将社交关系进一步细分为明确关系和隐含关系,并结合历史评分得到的用户的声誉信息,刻画了由用户全局信息(声誉)与局部信息(明确关系和隐含关系)所构成的推荐系统框架。用户的明确关系与隐含关系如图1所示。

本文第2节介绍社交推荐的相关研究;第3节回顾概率矩阵分解模型,并结合社交关系提出结合用户全局与局部信息的社交推荐系统(combining user Global and Dual Local information for social Recommendation algorithm, GDLRec);第4节为实验结果与分析;最后作出总结并概析未来研究的方向。

到稿日期:2015-03-13 返修日期:2015-05-07 本文受安徽大学2014年本科生创新创业项目(201410357036),安徽大学"211工程"三期第三批杰出青年科学研究培育基金(KJQN1116)资助。

钱付兰(1978-),女,博士生,讲师,主要研究方向为社交网络、个性化推荐等,E-mail:qianfulan@hotmail.com;**李启龙**(1992-),男,主要研究方向为社交推荐,E-mail:tealee1992@163.com。



用户 a 和 b 是相邻用户，他们的社交关系属于明确关系；a 和 c 之间不存在直接联系，但存在共同的朋友，具有一定的相似性，他们之间存在隐含关系。

图 1　用户的明确关系与隐含关系

2　相关工作

上文提到过，为了提高传统推荐算法的推荐效果，许多研究者已经开始了社交推荐的相关研究。下面将对其中一部分算法进行简单的回顾。

文献[2]从全局与局部的角度分析了社交关系。用户的全局社交信息(global social context)指的是用户在其所在的整个社交网中所处的地位或所具有的影响力，即用户的声誉度；局部社交信息(local social context)是指用户与其相邻用户之间的相关性。

通过用户声誉系数对系统中一些不严谨用户在推荐过程中的作用进行限制，即过滤掉自然噪声，可以提升推荐精度[4]。文献[2]在计算用户声誉值时，先用 PageRank 算法计算出用户的排名，再通过函数将排名转化为取值范围为[0,1]的声誉值。但 PageRank 基于有向网络，不适用于对应无向网络的数据集。文献[4]在比较了几种声誉评估方法后，采用周涛等人提出的声誉相关系数[5]大小来衡量用户声誉，声誉值是由用户历史评分得到，这种方法具有更广的适用面和更好的可解释性。本文使用的声誉值即是通过这种方法计算得到的。

本文将在社交推荐中引入隐含关系，进一步补充用户的局部社交信息。关于明确关系和隐含关系，已经有了一些相关研究[6-8]。文献[6,7]将用户社交关系分为直接关系(Direct Relation)和非直接关系(Indirect Relation)，通过六度分割理论计算用户关系的权重。文献[8]通过结合 implicit interaction 与 explicit interaction，提出了新的推荐模型 RoRec。还有些研究对 implicit 和 explicit 赋予了不一样的含义，文献[9-11]将 explicit information 理解为社交网站直接提供的社交信息，而 implicit information 则需要通过用户互动的行为历史计算用户相似性得到。

社交推荐方法根据所用的传统推荐算法可以分为两类：基于记忆的方法和基于模型的方法[10]。其中基于矩阵分解模型的方法最为流行，比如 SoRec[12]、SocialMF[13] 以及 SoReg[14]。现有的社交推荐算法有从全局、局部的角度分析社交关系的，也有涉及明确关系与隐含关系的方法，但是还没有研究将这些综合进一个模型内。本文在这些研究的基础上提出了新的算法，主要有以下贡献：

(1)采用用户的历史评分计算得到用户的声誉值，以此作为衡量用户评分的权重指标。

(2)将局部社交关系细分为明确关系与隐含关系。

(3)结合(1)(2)提出基于声誉的局部关系细分推荐算法。将算法在 Douban、Epinions 数据集上实验，并与现有的社交推荐算法比较，结果表明所提算法在推荐精度上表现更好。

3　结合用户全局与局部信息的社交推荐系统

3.1　符号说明

首先介绍本文将会使用到的符号，其中大部分符号会出现在式(2)里。用英文字母 i、j 表示某个用户，希腊字母 α 表示某个项目，用户 i 对项目 α 的评分为 $r_{i\alpha}$，预测评分为 $r^p_{i\alpha}$；R 表示评分矩阵，U、V 分别表示用户和项目的隐语义矩阵，U_i、V_α 表示用户和项目的隐语义向量。$I^R_{i\alpha}$ 表示指标函数，若用户 i 有对商品 α 评分，则 $I^R_{i\alpha}$ 为 1，否则为 0。I^S_{ij} 与 S_{ij} 有相同的取值，表示用户 i 与用户 j 之间是否存在社交关系，存在则值为 1，否则为 0。S_{ij} 描述的是用户 i、j 之间的关系强度。C_i 表示用户的声誉值。其余的符号会在后文详细介绍。

3.2　概率矩阵分解

概率矩阵分解(Probabilistic Matrix Factorization，PMF)假设用户和项目的隐语义矩阵都符合高斯分布。基于此，推导得到惩罚函数[15]：

$$E=\frac{1}{2}\sum_{i=1}^{m}\sum_{\alpha=1}^{n}I^R_{i\alpha}(r_{i\alpha}-U_i^TV_\alpha)^2+\frac{\lambda_U}{2}\sum_{i=1}^{m}\|U\|_F^2+\frac{\lambda_V}{2}\sum_{j=1}^{n}\|V\|_F^2 \tag{1}$$

其中，U、V 分别表示用户和项目的隐语义矩阵。而预测评分矩阵可近似分解为 U 和 V 矩阵相乘的形式，即 $r^p_{i\alpha}=U_i^TV_\alpha$，本文在此公式基础上加入了用户声誉值以及用户社交关系的影响，得到惩罚函数：

$$\min_{U,V}L(R,U,V)=\frac{1}{2}\sum_{i=1}^{m}\sum_{\alpha=1}^{n}C_iI^R_{i\alpha}(r_{i\alpha}-U_i^TV_\alpha)^2+\frac{\lambda_1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}I^S_{ij}I^R_{i\alpha}(S_{ij}-\hat{S}_{ij})^2+\frac{\lambda_2}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}C^U_{ij}I^R_{i\alpha}\|U_i-U_j\|_F^2+\frac{\lambda_3}{2}(\|U\|_F^2+\|V\|_F^2) \tag{2}$$

式中，在第一项加入用户声誉值，使声誉值高的用户对推荐结果具有更大的影响，进而提升推荐效果。第二、三项为正则项。第二项是通过用户明确关系描述用户的局部相似信息，其中 $\hat{S}_{ij}=U_i^TU_j$，当用户 i、j 存在明确关系时，S_{ij} 为 1，使得用户 i、j 的偏好向量(隐语义向量)U_i、U_j 更接近，即他们应该有更相似的喜好。第三项包含了用户之间的隐含关系，C^U_{ij} 表示用户 i 与用户 j 之间的相似度，两个用户有越多的共同朋友，即他们的隐含关系越强，则他们的相似度越高，也是在学习中迫使后面的偏好向量尽可能接近。C^U_{ij} 的计算公式如式(3)：

$$C^U_{ij}=\frac{\sum_{k=1}^{m}S_{ik}\cdot S_{jk}}{\sqrt{\sum_{k=1}^{m}S_{ik}^2}\cdot\sqrt{\sum_{k=1}^{m}S_{jk}^2}} \tag{3}$$

接下来只要对惩罚函数进行优化，得到最优的 U、V 即可。通常的优化方法分为两种：交叉最小二乘法(alternative least squares)和随机梯度下降法(stochastic gradient descent)。本文采用随机梯度下降法。先用式(2)分别对用户和项目的隐语义向量 U_i、V_α 求偏导，得到式(4)、式(5)：

$$\frac{\partial L}{\partial U_i}=-\sum_{\alpha=1}^{n}C_iI^R_{i\alpha}V_\alpha(R_{i\alpha}-U_i^TV_\alpha)-\lambda_1I^R_{i\alpha}\sum_{j=1}^{m}I^S_{ij}(S_{ij}-\hat{S}_{ij})U_j+\lambda_2I^R_{i\alpha}\sum_{j=1}^{m}C^U_{ij}\|U_i-U_j\|_F+\lambda_3\|U_i\|_F \tag{4}$$

$$\frac{\partial L}{\partial V_a}=-\sum_{i=1}^{m}C_i I_{ia}^R U_i(R_{ia}-U_i^T V_a)+\lambda_3 \| V_a \|_F \quad (5)$$

进而根据随机梯度下降法求得 U_i、V_a 的更新公式：

$$U_i=U_i-\gamma\frac{\partial L}{\partial U_i} \quad (6)$$

$$V_a=V_a-\gamma\frac{\partial L}{\partial V_a} \quad (7)$$

式中，γ 是学习速率，即迭代步长。在学习完用户和项目的隐语义矩阵 U、V 后，就可以得到预测评分矩阵 $R^p=U^TV$。具体算法见算法 1。

算法 1 结合用户全局与局部信息的社交推荐算法

输入：用户-项目评分矩阵 R，用户声誉值 rt，用户关系文件 trust，语义维数 F，随机梯度下降法的学习速率 γ(迭代步长)

输出：隐语义矩阵 U、V

Step1：采用随机数填充的方法初始化矩阵 U、V

Step2：计算 $\hat{S}_{ij}$ 以及用户相似度 C_{ij}^U

Step3：计算 $\frac{\partial L}{\partial U_i}$ 和 $\frac{\partial L}{\partial V_a}$

Step4：$U \leftarrow U-\gamma\frac{\partial L}{\partial U_i}$，$V \leftarrow V-\gamma\frac{\partial L}{\partial V_a}$

Step5：若 U、V 不收敛，转至 Step3。

4 实验结果及分析

4.1 数据集

实验采用了 Douban 数据和 Epinions 数据，这两个数据集的部分统计如表 1 所列。实验所用的数据包括用户的电影评分数据(评分范围为 1～5)、社交关系集(trust)和用户声誉值(rt)，评分数据按照训练集(train)：测试集(test)为 60%：40%、70%：30%、80%：20%的比例进行随机划分。

表 1 数据集统计

Datasets	Douban	Epinions
User Number	7700	22168
Item Number	6497	41369
Ratings	851220	583968
Relations	5874	342037

4.2 评价指标

推荐的准确度是评价推荐算法最基本的指标，它衡量的是推荐算法在多大程度上能够准确预测用户对推荐商品的喜欢程度。本文使用的评分准确度的指标为平均绝对误差(Mean Absolute Error，MAE)以及均方根误差(Root Mean Squared Error，RMSE)：

$$MAE=\frac{1}{|E^p|}\sum_{i,a\in E^p}(r_{ia}-r_{ia}^p)^2 \quad (8)$$

$$RMSE=\sqrt{\frac{1}{|E^p|}\sum_{i,a\in E^p}(r_{ia}-r_{ia}^p)^2} \quad (9)$$

式中，E^p 为 test 集中评分数据的数目。指标的值越小，说明算法的推荐精度越好。

4.3 推荐精度对比

为了评价算法的推荐精度，将本文算法与以下算法进行对比。

PMF[15]：概率矩阵分解仅用用户历史评分数据的传统推荐模型，本文提出的算法即是基于此模型。

SoRec[12]：该方法基于概率矩阵分解，通过将用户社交信息和用户评分数据相结合，在数据稀疏性问题的解决上有很好的表现。

SocialMF[13]：基于矩阵分解的社交推荐模型，引入了信任传播机制，可以减轻冷启动问题的影响。

SoReg[14]：该方法也是基于矩阵分解，通过引入社交正则项(Social Regularization)将用户的社交信息融入推荐系统中。

除了 PMF 没有考虑社交关系外，其他算法虽然都在传统的矩阵分解模型上引入了用户的社交信息，但没有综合考虑到本文提出的双重局部信息，即明确社交关系和隐含社交关系。

算法比较的结果如表 2 和表 3 所列，表中 80%、70%、60%表示训练集所占比例。结果表明，本文提出的算法在推荐精度方面要比其他算法更加出色。

表 2 Douban 数据集上算法精度的对比

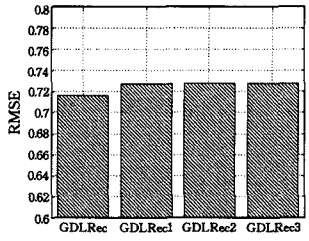
算法	80%		70%		60%	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
PMF	0.5714	0.7301	0.58	0.7336	0.5922	0.7383
SocialMF	0.5691	0.7253	0.5761	0.7310	0.5904	0.7370
SoRec	0.5705	0.7287	0.5783	0.7321	0.5916	0.7377
SoReg	0.5683	0.7278	0.5764	0.7319	0.5893	0.7362
GDLRec	0.5608	0.7170	0.5609	0.7172	0.5683	0.7280

表 3 Epinions 数据集上算法精度的对比

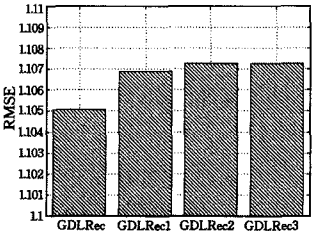
算法	80%		70%		60%	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
PMF	0.8641	1.1503	0.9123	1.1713	0.9572	1.1832
SocialMF	0.8531	1.1342	0.8514	1.1425	0.8729	1.1596
SoRec	0.8502	1.1437	0.8677	1.1583	0.8964	1.1649
SoReg	0.8496	1.1312	0.8519	1.1396	0.8691	1.1528
GDLRec	0.8351	1.0805	0.8513	1.1050	0.8773	1.1474

4.4 明确关系与隐含关系对算法的影响

上面的实验结果显示本文提出的算法在推荐精度方面优于现有的算法；下面通过实验研究明确关系与隐含关系对该算法推荐效果的影响。分别通过去除惩罚函数中第二项(明确关系)、第三项(隐含关系)以及两项均去除，得到 GDLRec 的 3 个变种算法：GDLRec1、GDLRec2、GDLRec3。分别选取 Douban、Epinions 按训练集为 70%划分的评分数据进行实验，比较结果如图 2 所示。



(a) Douban 数据集



(b) Epinions 数据集

图 2 明确关系与隐含关系对算法的影响

从图 2 中可以看出，综合考虑了明确关系与隐含关系的

(下转第 94 页)

mation[J]. Pattern Recognition,2010,43(7):2367-2379

[31] Wang Guo-quan, Zhou Xiao-hong, Yu Li-lei. Image Segmentation Based on Watershed Algorithm[J]. Computer Simulation, 2006,26(5):255-258(in Cinese)
王国权,周小红,蔚立磊. 基于分水岭算法的图像分割方法研究[J]. 计算机仿真,2006,26(5):255-258

[32] Vincent L, Soille P. Watersheds in digital spaces: an efficient algorithm based on immersion simulations[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, 13(6): 583-598

[33] Meyer F. Topographic distance and watershed lines[J]. Signal Process, 1994, 38(1): 113-125

[34] Stoev S. Rafsi—a fast watershed algorithm based on rainfalling simulation[C]// Proceedings of the Eighth International Conference on Computer Graphics, Visualization, and Interactive Digital Media. 2000: 100-107

[35] Osma-Ruiz V, Godino-Llorente J I, Sáenz-Lechón N, et al. An improved watershed algorithm based on efficient computation of shortest paths[J]. Pattern Recognition, 2007, 40(3): 1078-1090

[36] Camps-Valls G, Gomez-Chova L, Muñoz-Marí J, et al. Composite Kernels for Hyperspectral Image Classification[J]. IEEE Geoscience and Remote Sensing Letters, 2006, 3(1): 93-97

(上接第 59 页)

算法(GDLRec)要优于单独考虑任意一个的情况(GDLRec1、GDLRec2);而两种局部社交关系均不考虑时(GDLRec3)与只引入明确关系时(GDLRec2)的推荐效果非常接近。这说明:(1)明确关系与隐含关系的结合使算法的推荐精度有很大的提升;(2)在不考虑明确关系时,隐含关系对算法有一定的作用,但并不明显;(3)在不考虑隐含关系时,明确关系对算法几乎没有影响。

结束语 本文从全局、局部的角度研究如何将社交关系应用于推荐系统中,主要着眼于局部社交关系里的明确关系(explicit relation)和隐含关系(implicit relation);并提出结合用户全局与局部信息的社交推荐系统(GDLRec)。全局社交关系体现在用户声誉值系数的引入;局部社交关系的两类关系对应于在 PMF 里加入的正则项。通过在 Douban、Epinions 数据集上的实验,验证了本文基于声誉的局部关系细分推荐算法在推荐精度上有很好的表现。与现有的算法相比,本文提出的算法具有更好的可解释性和更高的推荐精度。

随着社交网络的普及,社交推荐的研究工作显得尤为重要。未来关于社交推荐的研究主要围绕以下几个方面展开:(1)继续探讨全局关系在社交推荐中的作用,尝试用不同方法得到声誉系数;(2)由于框架中的明确关系单独作用不是很明显,可以尝试用其他方法计算用户之间的明确关系的关系强度。

参 考 文 献

[1] Mcpherson M, Smith-Lovin L, Cook J M. Birds of a feather: Homophily in social networks[J]. Annual Review of Sociology, 2001, 15(4): 344-349

[2] Tang J, Hu X, Gao H, et al. Exploiting local and global social context for recommendation[C]// Proceedings of the Twenty-Third international joint conference on Artificial Intelligence. AAAI Press, 2013: 2712-2718

[3] Marsden P V, Friedkin N E. Network studies of social influence [J]. Sociological Methods & Research, 1993, 22(1): 127-151

[4] Zhang Yan-ping, Zhang Shun, Qian Fu-lan, et al. Robust Collaborative Recommendation Algorithm Based on User's Reputation[J]. Acta Automatica Sinica, 2015(5): 1004-1012(in Chinese)
张燕平,张顺,钱付兰,等. 基于用户声誉的鲁棒协同推荐算法[J]. 自动化学报,2015(5):1004-1012

[5] Zhou Y B, Lei T, Zhou T. A robust ranking algorithm to spamming[J]. EPL (Europhysics Letters), 2011, 94(4): 1034-1054

[6] Ha I, Oh K J, Hong M D, et al. Social filtering using social relationship for movie recommendation[M]// Computational Collective Intelligence: Technologies and Applications. Springer Berlin Heidelberg, 2012: 395-404

[7] Ha I, Oh K J, Jo G S. Personalized advertisement system using social relationship based user modeling[J]. Multimedia Tools and Applications, 2013, 74(20): 8801-8819

[8] Yao W, He J, Huang G, et al. Modeling dual role preferences for trust-aware recommendation[C]// Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM, 2014: 975-978

[9] Fazeli S, Loni B, Bellogin A, et al. Implicit vs. explicit trust in social matrix factorization[C]// Proceedings of the 8th ACM Conference on Recommender systems. ACM, 2014: 317-320

[10] Ma H. An experimental study on implicit social recommendation [C]// Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013: 73-82

[11] Zhao T, Hu J, He P, et al. Exploiting homophily-based implicit social network to improve recommendation performance[C]// 2014 International Joint Conference on Neural Networks (IJCNN). IEEE, 2014: 2539-2547

[12] Ma H, Yang H, Lyu M R, et al. Sorec: social recommendation using probabilistic matrix factorization[C]// Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008: 931-940

[13] 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

[14] Ma H, Zhou D, Liu C, et al. Recommender systems with social regularization[C]// Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. ACM, 2011: 287-296

[15] Salakhutdinov R, Mnih A. Probabilistic Matrix Factorization [C]// NIPS. 2012: 1257-1264