基于Jaccard相似度和位置行为的协同过滤推荐算法

李 斌[1] 张 博[1] 刘学军[1] 章 玮[2]
(南京工业大学计算机科学与技术学院 南京 211816)[1] (中国人民解放军73677部队 南京 210016)[2]

摘 要 协同过滤是现今推荐系统中应用最为成功且最广泛的推荐方法之一,其中概率矩阵分解算法作为一类重要的协同过滤方式,能够通过学习低维的近似矩阵进行推荐。然而,传统的协同过滤推荐算法在推荐过程中只利用用户-项目评分信息,忽略了用户(项目)间的潜在影响力,影响了推荐精度。针对上述问题,首先利用Jaccard相似度对用户(项目)做预处理,而后通过用户(项目)间的位置信息挖掘出其间的潜在影响力,成功找到最近邻居集合;最后将该邻居集合融合到基于概率矩阵分解的协同过滤推荐算法中。实验证明该算法较传统的协同过滤推荐算法能够更有效地预测用户的实际评分,提高了推荐效果。

关键词 Jaccard相似度,位置行为,协同过滤,概率矩阵分解
中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.12.036

Collaborative Filtering Recommendation Algorithm Based on Jaccard Similarity and Locational Behaviors

LI Bin[1] ZHANG Bo[1] LIU Xue-jun[1] ZHANG Wei[2]
(College of Computer Science and Technology, Nanjing Tech University, Nanjing 211816, China)[1]
(The Chinese People's Liberation Army 73677 Troops, Nanjing 210016, China)[2]

Abstract Recently, collaborative filtering is one of the most widely used and successful recommendation technology in recommender system. And probabilistic matrix factorization is an important method of collaborative filtering and it can be recommended by learning the low dimensional approximation matrix. However, the traditional collaborative filtering recommendation algorithm has the disadvantages of using the ratings between users and items only, ignoring the potential impact of the users (items). At last, it affects the recommendation precision. In order to solve the problem, in this paper, we first used the Jaccard similarity to preprocess the users (items), and then dug out the potential impact through the users (items) location information, finding the set of nearest neighbors successfully. Furthermore, those nearest neighbors were successfully applied into the recommendation process based on probabilistic matrix factorization. Experimental results show that compared to traditional collaborative filtering recommendation algorithm, the proposed algorithm can achieve more accurate rating predictions and improve the quality of recommendation.

Keywords Jaccard similarity, Locational behaviors, Collaborative filtering, Probabilistic matrix factorization

1 引言

近年来,随着用户和网络规模的不断扩大,网络上记录的数据量急剧增长,无论是信息的消费者还是生产者都面临着严重的信息过载问题。推荐系统(Recommender System)作为一种有效的信息过滤技术,能帮助用户从规模庞大的信息中快速准确地找到其感兴趣的信息,为用户主动推荐满足其需求的资源,受到了研究者们的广泛关注。

目前主要的推荐方法有协同过滤(Collaborative Filtering,CF)推荐、基于内容的推荐、基于知识的推荐以及混合推荐等,其中协同过滤是应用最多的算法。其认为相似的用户具有相似的兴趣,通过寻找与目标用户相似的邻居用户,综合邻居用户对某一信息的评价,形成系统对该目标用户在产品喜好程度方面的预测,而后进行相应的推荐。这类方法已被广泛应用于推荐系统中,尤其是电子商务领域,并取得了显著的效果,带来了巨大的商业价值。

传统的协同过滤推荐算法在邻居选择时仅仅依靠用户(项目)间评分的相似度,没有考虑用户(项目)间关系,导致邻居选取的片面性,从而影响了推荐结果的准确度。为此,本文引入了Jaccard相似度和用户(项目)的位置行为,提出了基于Jaccard相似度和位置行为的协同过滤推荐算法。

本文第2节介绍协同过滤推荐算法的相关工作;第3节介绍研究问题和概率矩阵分解模型;第4节介绍CF-JSLB的具体实现;第5节给出实验和结论;最后对全文进行总结和展望。

2 相关工作

协同过滤推荐算法假设用户对某一物品的喜好同与其有着相似兴趣的用户相一致,通过对用户的历史评分数据进行分析,产生最后的推荐结果。目前,协同过滤推荐算法主要分

到稿日期:2015-10-17 返修日期:2016-03-22 本文受国家自然科学基金(61203072),江苏省重点研发计划(社会发展)(BE2015697)资助。
李 斌(1979-),男,硕士,讲师,主要研究方向为传感器网络、智能信息处理;张 博(1991-),女,硕士生,主要研究方向为数据挖掘、社会推荐,E-mail:zhangbo_hello@163.com;刘学军(1971-),男,博士,教授,CCF高级会员,主要研究方向为数据库、数据挖掘、传感器网络等;章 玮(1982-),女,硕士,工程师,主要研究方向为网络安全、社会计算。

为基于内存和基于模型两种。

基于内存的协同过滤推荐算法通过计算用户或物品在评分数据上的相似性，进而对用户可能的评分进行预测，但其需将数据全部加载到内存中，每次推荐都需要使用启发式的规则来重新计算相似性和预测评分。基于内存的协同过滤算法主要包括基于用户的和基于项目的两类。

与基于内存的协同过滤推荐算法相比，基于模型的协同过滤推荐算法需要事先使用机器学习的方法训练出一个有效的推荐模型，并利用此模型预测未知的数据。目前，基于模型的算法主要包括聚类模型、潜在特征模型、aspect 模型、贝叶斯层次模型等。其中，概率矩阵分解方法(Probabilistic Matrix Factorization，PMF)因其能使用概率的方法从已知评分数据中推导出表示用户和物品的潜在特征向量，且满足大数据时代高扩展性和准确性的要求，受到了研究者们的广泛青睐。

以上传统的协同过滤推荐算法虽然取得了较好的效果，但其往往仅采用历史评分信息，单一的用户-项目评分数据并不能全面准确地识别出用户的兴趣。为此，越来越多的研究者在已有算法的基础上不断加入新的元素以得到更准确的推荐结果。

Ma 等人[1]提出了利用用户间的社会关系来进一步提高传统推荐方法的精度，通过使用社会关系矩阵对目标函数进行约束，并给出了融合社会关系信息的概率矩阵分解框架。Jamali 等人[2]认为用户间的信任关系具有传播性，提出了一种基于信任传播的推荐方法。郭磊等人[3]认为要想取得好的推荐效果不能只从用户间社会关系的角度出发，假设推荐对象之间是相互独立的，即在已有社会化推荐算法的基础上提出了一种结合推荐对象间关联关系的推荐方法。Jiang 等人[4]提出了一种基于社会上下文信息的推荐算法，使用推荐对象间的内容相似性对目标函数进行约束，并对不同的社会上下文信息进行建模。

除了借助用户间的社会关系信息，如何利用丰富的上下文信息进行推荐也逐渐成为研究的热点。Liu 等人[5]为更好地符合用户需求，从上下文环境出发，提出了一种基于环境感知的推荐方法，提高了推荐精度。Dong 等人[6]提出了一种基于异构社会化网络的随机漫步排序模型和不同类型的边权值成对学习算法，利用 Web 应用环境下用户对各种资源的标签化命名关系构建一个异构的社会化网络结构，为用户提供信息推荐服务。Wang 等人[7]通过对两个基于位置的社会化网络用户签到位置的分布情况进行分析，提出了基于位置的社会网络位置推荐。Rendle 等人[8]将用户的情绪信息、陪同观看电影的用户信息与用户信息、电影信息编码在一起，实现对用户情绪和陪同人员感知的电影评分预测。

矩阵分解方法(Matrix Factorization，MF)也被广泛应用于推荐系统中。Chen 等人[9]针对 Twitter 数据的特点，提出了一种基于特征的矩阵分解模型，该模型重点考虑 tweet 主题水平因素 Ti 和用户社会关系因素 $dU(i)$ 及其他显著特征，实现对用户的 tweet 个性化协同推荐。Chen 等人[10]将用户社会特征、参与的互动、社会网络关系及项目的分类信息融合在一起，建立了递增树林模型，以获取用户的活动及其连续模式。Hong 等人[11]利用因子分解机模型，模拟用户兴趣潜在特征，实现对 Twitter 用户的个体决策分析与预测。

基于以上研究，本文提出了基于 Jaccard 相似度和位置行为的协同过滤推荐算法(CF-JSLB)，首先借助 Jaccard 相似度对用户(项目)进行预处理，而后借助位置信息挖掘出用户(项目)间的潜在影响力以确定最近邻居集合，最后将该邻居融合到基于概率矩阵分解的协同过滤推荐算法中，生成最后的推荐结果。

3 问题定义和概率矩阵分解

概率矩阵分解方法[12]通过对用户-项目评分矩阵进行分解，推导出两个分别表示用户和项目的低维潜在特征矩阵，且这些特征是刻画用户和项目的关键因素。同时，由于 PMF 方法在学习过程中使用的特征向量维度较低，具有较低的计算复杂度，使得其在社会网络这样的大规模数据集上具有较高的可扩展性和准确性。对于协同过滤算法而言，可利用概率矩阵分解模型学习用户(项目)的特征向量，然后基于此特征向量预测未知的评分。

假设在推荐系统中存在 m 个用户和 n 个项目，用户-项目评分矩阵 $R=\lfloor R_{i,j}\rfloor_{m\times n}$，其中 $R_{i,j}(1\leqslant i\leqslant m, 1\leqslant j\leqslant n)$ 表示用户 u_i 对项目 v_j 的评分。$U\in R^{l\times m}$ 和 $V\in R^{l\times n}$ 分别表示分解得到的与用户和项目相关的 l 维特征向量，其列向量 U_i 和 V_j 则分别表示相对应的潜在特征向量。根据以上定义，评分矩阵 R 的条件概率分布定义如下：

$$p(R|U,V,\sigma_R^2)=\prod_{i=1}^{m}\prod_{j=1}^{n}[N(R_{i,j}|g(U_i^TV_j),\sigma_R^2)]^{I_{ij}^R} \tag{1}$$

其中，$N(x|\mu,\sigma^2)$ 表示 x 服从均值为 μ、方差为 σ^2 的正态分布。I_{ij}^R 是指示函数，若用户 U_i 对项目 V_j 有评分，则其值为 1，否则为 0。函数 $g(x)=\frac{1}{(1+e^{-x})}$，使用它的目的是将预测值 $(U_i^TV_j)$ 的值映射到[0,1]区间内。

为了防止过拟合，假设 U 和 V 均服从均值为 0 的高斯先验分布：

$$\begin{aligned}p(U|\sigma_U^2)&=\prod_{i=1}^{m}N(U_i|0,\sigma_U^2I)\\ p(V|\sigma_V^2)&=\prod_{j=1}^{n}N(V_j|0,\sigma_V^2I)\end{aligned} \tag{2}$$

经过贝叶斯推断，特征向量 U 和 V 的后验概率如下：

$$\begin{aligned}p(U,V|R,\sigma_R^2,\sigma_U^2,\sigma_V^2)&\propto p(R|U,V,\sigma_R^2)p(U|\sigma_U^2)p(V\sigma_V^2)\\&=\prod_{i=1}^{m}\prod_{j=1}^{n}[N(R_{i,j}|g(U_i^TV_j),\sigma_R^2)]^{I_{ij}^R}\times\prod_{i=1}^{m}N(U_i|0,\sigma_U^2I)\times\\&\quad\prod_{j=1}^{n}N(V_j|0,\sigma_V^2I)\end{aligned} \tag{3}$$

由式(3)可知，在 PMF 方法中只需用户-项目评分矩阵即可学习出相应的特征向量，其模型图如图 1 所示。

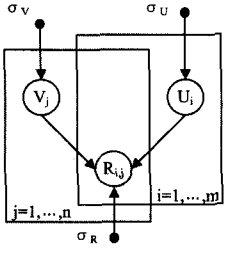

图 1 概率矩阵分解模型图

4 基于 Jaccard 相似度和位置行为的协同过滤推荐算法(CF-JSLB)

传统的协同过滤推荐算法在推荐过程中仅仅依靠用户-项目评分信息确定邻居，导致邻居选取的片面性，进而影响了

结果的准确性。为此，提出了基于 Jaccard 相似度和位置行为的协同过滤推荐算法。首先利用 Jaccard 相似度对用户(项目)进行预处理，然后利用用户消费的位置信息挖掘出用户(项目)间的潜在影响力，从而确定对用户(项目)影响最大的邻居集合，最后将该邻居集合融入到基于概率矩阵分解的协同过滤推荐算法中以生成最终的推荐结果。

4.1 基于 Jaccard 相似度的用户(项目)预处理

在协同过滤推荐算法中，邻居选择会直接影响到推荐结果的准确性，是推荐过程中至关重要的一步。为更好地确定影响最大的邻居集合，首先利用 Jaccard 相似度对用户进行预处理，认为与目标用户具有相同评价项目的用户才能成为候选邻居，项目同理。

对于用户而言，认为评价过相同项目的用户在一定程度上具有相似的兴趣，就目标用户而言，其更能接受来自兴趣爱好相似用户的推荐。假设用户 U_i 评价过的项目集合为 I_i，用户 U_j 评价过的项目集合为 I_j，则用户 U_i 和 U_j 的 Jaccard 相似度计算公式如下：

$$Jaccard(U_i,U_j)=\frac{|I_i\cap I_j|}{|I_i\cup I_j|} \tag{4}$$

其中，$|I_i\cap I_j|$ 表示用户 U_i 和 U_j 共同评价过的项目的个数，$|I_i\cup I_j|$ 表示用户 U_i 和 U_j 分别评价过的项目总数。使用 Jaccard 系数保证相似度在 0 到 1 之间，包含 0 或者 1，相似度值越大表示相似度越高，兴趣越相似。取与目标用户 U_i 的 Jaccard 相似度大于 0 的用户作为候选邻居，定义为 $C(U_i)$。

同理，对于项目而言，认为同一用户评价过的项目之间或许存在一定的联系。假设评价过项目 I_a 的用户集合为 U_a，评价过项目 I_b 的用户集合为 U_b，则项目 I_a 和 I_b 的 Jaccard 相似度计算公式如下：

$$Jaccard(I_a,I_b)=\frac{|U_a\cap U_b|}{|U_a\cup U_b|} \tag{5}$$

其中，$|U_a\cap U_b|$ 表示共同评价过项目 I_a 和 I_b 的用户个数，$|U_a\cup U_b|$ 表示分别评价过项目 I_a 和 I_b 的用户总数。取与目标项目 I_a 的 Jaccard 相似度大于 0 的项目作为候选邻居，定义为 $C(I_a)$。

4.2 基于位置行为的邻居选择 SNL

传统的协同过滤推荐算法在邻居选择时忽略了用户(项目)间关系，而用户(项目)间的关系对用户(项目)推荐具有重要影响。例如，用户 A 和 B 多次在同一位置消费，则 A 和 B 很可能存在相似的兴趣爱好或潜在影响。为此，本文在邻居选择阶段挖掘用户(项目)消费位置信息，该位置信息很可能会隐藏着一部分规律，利用这些规律可以从一定程度上分析出用户之间或项目之间的关系，从而产生更加准确的推荐。为了更好地描述这种关系，引入了用户(项目)消费位置图。

对于用户而言，以候选邻居集 C 中的用户为节点、用户间的消费位置联系 E 为边构建用户消费位置图 $G=\{C,E\}$。图中，如果两个用户曾经在同一个位置消费过，则用边连接这两个节点，且权重为消费次数，定义为 T，次数越多，权重越大。假设用户 U_i 和 U_j 在同一个位置消费过 3 次，则 U_i 和 U_j 间边的权重 T_{ij} 为 3。

设 W_{ij} 表示用户 U_i 对用户 U_j 的潜在影响力，计算公式如下：

$$W_{ij}=\frac{T_{ij}}{A_{ij}} \tag{6}$$

其中，T_{ij} 表示用户 U_i 和 U_j 间的权重大小，A_{ij} 表示用户 U_i 和 U_j 分别消费的次数的总和。

同理，对于项目而言，以候选邻居集 C 中的项目为节点、项目间的消费位置联系 E 为边构建项目消费位置图 $G=\{C,E\}$。如果两个项目曾在同一个位置被用户所消费，则用边连接这两个节点，且权重为消费次数，定义为 P，次数越多，权重越大。假设项目 I_a 和 I_b 在同一位置只被用户消费了一次，则 I_a 和 I_b 间边的权重 P_{ij} 为 1。

设 S_{ab} 表示项目 I_a 对项目 I_b 的潜在影响力，计算公式如下：

$$S_{ab}=\frac{P_{ab}}{B_{ab}} \tag{7}$$

其中，P_{ab} 表示项目 I_a 和 I_b 间的权重大小，B_{ab} 表示项目 I_a 和 I_b 分别被用户消费的次数的总和。

4.3 推荐算法 CF-JSLB

在协同过滤推荐算法中，寻找到最近邻居后则进入推荐阶段。文中借助用户(项目)间的潜在影响力找到最近邻居集合后，将其应用到概率矩阵分解模型中生成推荐结果。

本文虽然引入了用户(项目)间的潜在影响力，但并没有改变用户-项目评分矩阵 R 的条件分布，用户(项目)间的潜在影响力只对用户(向量)的特征向量产生影响。R 的条件概率分布依然定义为：

$$p(R|U,V,\sigma_R^2)=\prod_{i=1}^{m}\prod_{j=1}^{n}[N(R_{i,j}|g(U_i^TV_j),\sigma_R^2)]^{I_{ij}^R} \tag{8}$$

其中，I_{ij}^R 是指示函数，如果用户 U_i 对项目 V_j 有评分，则其值为 1，否则为 0。

由于用户(项目)的特征向量会受到其最近邻居集合的影响，即相似的用户(项目)具有相似的特征向量，则：

$$\hat{U}_i=\sum_{t\in N_i}W_{it}U_t,\hat{V}_j=\sum_{k\in N_j}S_{kj}V_k \tag{9}$$

其中，$\hat{U}$和$\hat{V}$表示近似的特征向量，N_i 和 N_j 分别表示用户 u_i 和 v_j 的邻居集合。

用户(项目)特征向量 $U(V)$ 主要表现为以下两方面：1)为了防止过拟合，均采用均值为 0 的高斯先验分布；2)同时综合考虑能反映出对用户(项目)有潜在影响力的特征向量。因此，可以得到：

$$p(U|W,\sigma_U^2,\sigma_W^2)\propto p(U|\sigma_U^2)\times p(U|W,\sigma_W^2)=\prod_{i=1}^{m}N(U_i|0,\sigma_U^2I)\times\prod_{i=1}^{m}N(U_i|\sum_{t\in N_i}W_{it}U_t,\sigma_W^2I) \tag{10}$$

$$p(V|S,\sigma_V^2,\sigma_S^2)\propto p(V|\sigma_V^2)\times p(V|S,\sigma_S^2)=\prod_{j=1}^{n}N(V_j|0,\sigma_V^2I)\times\prod_{j=1}^{n}N(V_j|\sum_{k\in N_j}S_{kj}V_k,\sigma_S^2I) \tag{11}$$

经过贝叶斯推断，可以进一步得出：

$$p(U,V|R,W,S,\sigma_R^2,\sigma_U^2,\sigma_V^2)\propto p(R|U,V,\sigma_R^2)p(U|W,\sigma_U^2,\sigma_W^2)\times p(V|S,\sigma_V^2,\sigma_S^2)=\prod_{i=1}^{m}\prod_{j=1}^{n}[N(R_{i,j}|g(U_i^TV_j),\sigma_R^2)]^{I_{ij}^R}\times\prod_{i=1}^{m}N(U_i|\sum_{t\in N_i}W_{it}U_t,\sigma_W^2I)\times\prod_{j=1}^{n}N(V_j|\sum_{k\in N_j}S_{kj}V_k,\sigma_S^2I)\times\prod_{i=1}^{m}N(U_i|0,\sigma_U^2I)\times\prod_{j=1}^{n}N(V_j|0,\sigma_V^2I) \tag{12}$$

参数 U,V 的对数联合后验概率分布可以表示为：

$$\ln p(U,V|R,W,S,\sigma_R^2,\sigma_W^2,\sigma_S^2,\sigma_U^2,\sigma_V^2)=-\frac{1}{2\sigma_R^2}\sum_{i=1}^{m}\sum_{j=1}^{n}I_{ij}^R(R_{i,j}-g(U_i^TV_j))^2-\frac{1}{2\sigma_W^2}\sum_{i=1}^{m}((U_i-\sum_{t\in N_i}W_{it}U_t)^T(U_i-\sum_{t\in N_i}W_{it}U_t))-\frac{1}{2\sigma_S^2}\sum_{j=1}^{n}((V_j-\sum_{k\in N_j}S_{kj}V_k)^T$$

$$(V_j-\sum_{k\in N_j}S_{kj}V_k))-\frac{1}{2\sigma_U^2}\sum_{i=1}^{m}U_i^{\mathrm{T}}U_i-\frac{1}{2\sigma_V^2}\sum_{j=1}^{n}V_j^{\mathrm{T}}V_j-(\sum_{i=1}^{m}\sum_{j=1}^{n}I_{ij}^{R})\ln\sigma_R^2-\frac{1}{2}(ml\ln\sigma_U^2+nl\ln\sigma_V^2+ml\ln\sigma_W^2+nl\ln\sigma_S^2)+C \tag{13}$$

其中，C 是与参数无关的常量。求参数固定时 U 和 V 的极大后验概率，相当于最小化以下带正则项的误差平方和函数：

$$L(R,W,S,U,V)=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}I_{ij}^{R}(R_{i,j}-g(U_i^{\mathrm{T}}V_j))^2+\frac{\lambda_W}{2}\sum_{i=1}^{m}((U_i-\sum_{t\in N_i}W_{it}U_t)^{\mathrm{T}}(U_i-\sum_{t\in N_i}W_{it}U_t))+\frac{\lambda_S}{2}\sum_{j=1}^{n}((V_j-\sum_{k\in N_j}S_{kj}V_k)^{\mathrm{T}}(V_j-\sum_{k\in N_j}S_{kj}V_k))+\frac{\lambda_U}{2}\|U\|_F^2+\frac{\lambda_V}{2}\|V\|_F^2 \tag{14}$$

其中，$\lambda_U=\frac{\sigma_R^2}{\sigma_U^2}$，$\lambda_V=\frac{\sigma_R^2}{\sigma_V^2}$，$\lambda_W=\frac{\sigma_R^2}{\sigma_W^2}$，$\lambda_S=\frac{\sigma_R^2}{\sigma_S^2}$，$\|\cdot\|_F^2$ 表示 Frobenius 范数。对于式(14)所示的目标函数，通过在 U 和 V 上使用梯度下降的方法进行求解，可以使目标函数达到局部极小值：

$$\frac{\partial L}{\partial U_i}=\sum_{j=1}^{n}I_{ij}^{R}V_jg'(U_i^{\mathrm{T}}V_j)(g(U_i^{\mathrm{T}}V_j)-R_{i,j})+\lambda_UU_i+\lambda_W(U_i-\sum_{t\in N_i}W_{it}U_t)-\lambda_W\sum_{\{t|i\in N_t\}}W_{it}(U_t-\sum_{x\in N_t}W_{xt}U_x) \tag{15}$$

$$\frac{\partial L}{\partial V_j}=\sum_{i=1}^{m}I_{ij}^{R}U_ig'(U_i^{\mathrm{T}}V_j)(g(U_i^{\mathrm{T}}V_j)-R_{i,j})+\lambda_VV_j+\lambda_S(V_j-\sum_{k\in N_j}S_{kj}V_k)-\lambda_S\sum_{\{k|j\in N_k\}}S_{jk}(V_k-\sum_{h\in N_k}S_{hk}V_h) \tag{16}$$

其中，$g'(x)=\frac{e^{-x}}{(1+e^{-x})^2}$ 是 $g(x)$ 的导数。

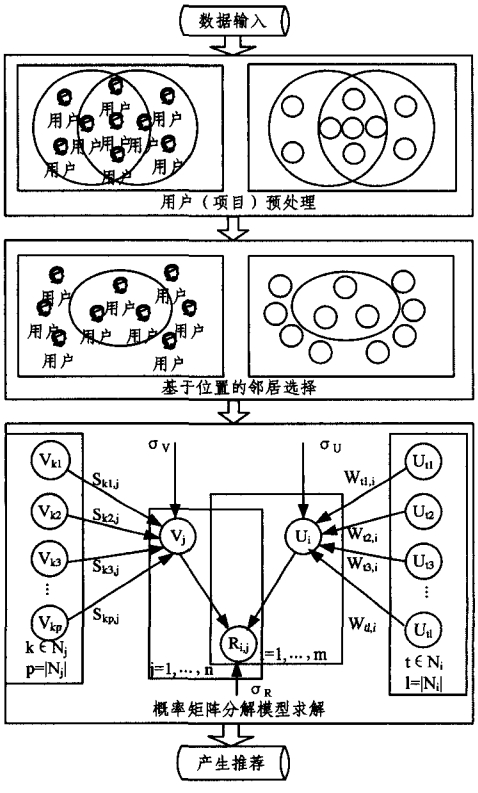

图 2　基于 Jaccard 相似度和位置行为的协同过滤推荐算法的框架

5　实验结果及分析

5.1　实验数据集

本文实验采用了两种数据集。

(1)Movie Lens 数据集。该数据集由美国明尼苏达州立大学 Group Lens 研究小组提供，包括 6040 个用户对 3952 部电影的约 100 万条电影评分信息，每个用户至少评价了 20 部电影，且评分范围为 1～5，“1”表示“poor”(不喜欢)，“5”表示“perfect”(非常喜欢)。本文从中选取了 943 个用户对 1682 部电影的大约 100000 次评分作为实验数据集。由于该数据集中不包含位置信息，而位置信息是本文算法中的一个重要组成部分，因此对其随机生成了地理位置信息，且一经生成就将其作为数据集的一部分，后面不予变动。

(2)Synthetic 数据集。该数据集是在美国明尼苏达州通过随机生成用户和物品位置，从而产生的一个包含空间物品位置评分的真实数据集，其中包括 2000 个用户对 1000 个物品的约 500000 个评分信息，且评分范围为 0～5。本文将整个数据集作为实验数据集。

在实验中，从每个数据集中随机抽取 80%作为训练集，其余的 20%作为测试集。

5.2　评价标准

本实验采用均方根误差(Root Mean Square Error，RMSE)作为度量标准。RMSE 通过计算预测用户评分与实际用户评分之间的偏差来度量预测的准确性，可以直观地对推荐质量进行评估，是一种常用的推荐质量度量方法。RMSE 越小，推荐质量越高。

假设预测的用户评分集合为 $\{p_1,p_2,\cdots,p_M\}$，对应的实际用户评分集合为 $\{q_1,q_2,\cdots,q_M\}$，M 表示预测的次数，则 RMSE 的计算公式如下：

$$RMSE=\sqrt{\frac{\sum_{i=1}^{M}(p_i-q_i)^2}{M}}$$

5.3　比较算法

为了展示 CF-JSLB 算法在推荐性能上的提升，本文与以下方法进行了比较。

(1)PMF[12]。该方法是一种基于概率的矩阵分解方法，在推荐过程中只考虑了用户对物品的评分信息生成推荐，而没有考虑用户(产品)间的关系。

(2)SPCF[13]。这是一种基于内存的传播式协同过滤推荐算法，通过相似度传播，寻找到更多、更可靠的邻居，在此基础上进行推荐。

(3)SocialMF[14]。该方法是一种对用户间的兴趣传播进行建模的算法，其将用户的社交网络关系融入推荐模型中，而忽略了隐藏在社交关系背后更深层次的用户特征。

(4)PMFUI[3]。该方法从推荐对象间关联关系的角度出发，假设具有关联关系的推荐对象更容易受到同一用户的关注，利用此关系提高推荐效果。

实验中的参数设置如下：$\lambda_U=\lambda_V=0.01$，并且假设 $\lambda_w=\lambda_s=\lambda$，算法的最大迭代次数设为 100，特征向量维度分别设置为 $K=5,10,15$。

5.4　最近邻选择方法比较

在协同过滤推荐算法中，只有选择较佳的最近邻才能更好地进行推荐。本实验对 4.2 节提出的基于位置行为的邻居选择 SNL 方法与 DSN 方法[15]、传统的 kNN 方法进行了对比。DSN 方法在考虑用户相似度的同时考虑了两者之间共同评价产品的个数来选取目标对象的推荐对象。以选择的近

邻个数为横坐标，RMSE 为纵坐标，近邻个数从 10 开始，逐次增加，直至 100，结果如图 3 和图 4 所示。

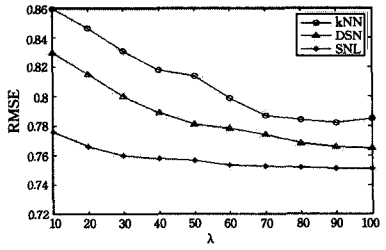

图 3　Movie Lens 数据集上最近邻选择方法比较

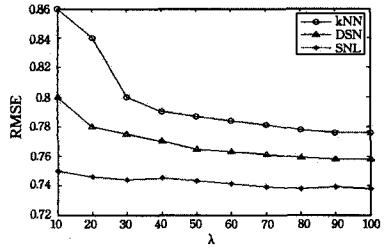

图 4　Synthetic 数据集上最近邻选择方法比较

从图中可以看出，随着近邻个数的增加，RMSE 值逐次减小，且更趋于稳定化。与此同时，可以发现 SNL 方法的 RMSE 值都明显小于 DSN 方法和 kNN 方法。随着训练用户数目的增多，目标用户的最近邻就越容易找到，RMSE 值也越小，这意味着通过更多用户的训练结果，得到的推荐结果也更好。由此认为基于位置行为的邻居选择 SNL 方法能够为用户进行有效的推荐。

5.5　各算法实验结果的比较

为了验证用户间的潜在影响力在推荐过程中所起到的作用以及 CF-JSLB 算法的准确性，本实验将 CF-JSLB 方法与 PMF，SocialMF，PMFUI 方法以及基于内存的 SPCF 方法在两个数据集下就维度为 5，10，15 的情况分别做了对比实验。需要说明的是，在与 SocialMF 和 PMFUI 算法做对比实验时，由于 Movie Lens 数据集和 Synthetic 数据集均不包含用户间关系，因此实验中涉及到的用户间和推荐对象间关系是随机生成的，结果如图 5—图 7 所示。

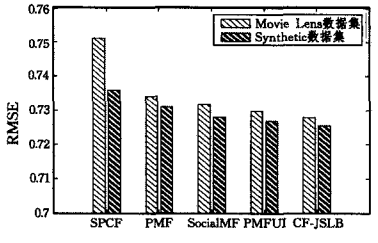

图 5　不同数据集下各算法结果比较($K=5$)

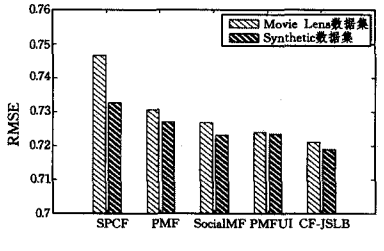

图 6　不同数据集下各算法结果比较($K=10$)

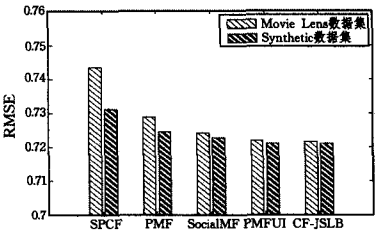

图 7　不同数据集下各算法结果比较($K=15$)

从图中可以看出，随着维度的上升，RMSE 逐渐降低，算法精度逐渐提高。与此同时，发现在 3 种不同维度下，CF-JSLB 算法的 RMSE 在几种算法中最低，说明经常被忽视的用户间的潜在影响力是推荐过程中的重要因素，能对推荐的最终结果产生重要影响。另外，从图中可以看出基于矩阵分解的算法与基于内存的 SPCF 算法相比在 RMSE 指标上有明显的优势，这也说明了矩阵分解方法的有效性。本文提出的 CF-JSLB 方法与 SocialMF 和 PMFUI 方法相比，在推荐精度上取得了更好的结果，说明简单地从用户的社交网络和推荐对象间关联关系的角度出发，并不能深层次挖掘用户和项目间的联系及产生更准确的推荐，进一步说明了本文提出的从用户(项目)间的潜在影响力出发的 CF-JSLB 算法的有效性。同时可以发现随着维度的上升，CF-JSLB 算法在两个数据集上的 RMSE 逐渐趋于相等，由此也可以说明随机生成的数据在真实的实验环境下也具有有效性。

5.6　参数 λ 对算法的影响

在 CF-JSLB 算法中，参数 λ 起到了很重要的作用，它可以控制用户(项目)间的潜在影响力在整个推荐过程中的比重，即推荐方法受信息的影响程度，λ 越大表明算法受用户(项目)间的潜在影响力作用越大。实验中为了降低复杂度，设定 $\lambda_w=\lambda_s=\lambda$，$\lambda$ 的设定值分别为 0.1，0.5，1，5，10，20。同时，分别在 $K=5$ 和 $K=10$ 的情况下对参数 λ 做对比实验，结果如表 1 和表 2 所列。

表 1　不同数据集下参数 λ 对 RMSE 的影响($K=5$)

λ	Movie Lens 数据集	Synthetic 数据集
0.1	0.7455	0.7446
0.5	0.7396	0.7392
1	0.7354	0.7326
5	0.7317	0.7302
10	0.7293	0.7288
20	0.7309	0.7299

表 2　不同数据集下参数 λ 对 RMSE 的影响($K=10$)

λ	Movie Lens 数据集	Synthetic 数据集
0.1	0.7449	0.7422
0.5	0.7383	0.7374
1	0.7347	0.7333
5	0.7310	0.7304
10	0.7286	0.7271
20	0.7291	0.7286

从表中可以发现，参数 λ 的取值对实验结果有较为显著的影响，随着 λ 的增加，RMSE 逐渐降低，算法精度不断提高；但当 λ 的值增大到阈值时(本文中是 20)，由于 λ 过大会导致算法过拟合，算法精度开始下降。推荐结果随着 λ 的取值不断变化，这也充分说明了本文增加的用户(项目)间潜在影响力对推荐结果产生的重要影响，验证了本文算法的有效性。

5.7 各算法运行时间比较

本实验对各算法的运行时间做了具体比较。该实验的运行环境为：Intel Core i5 CPU，2.67GHz 主频，Windows7 系统，4GB 内存，同时设定 $K=10$。实验结果如图 8 所示。

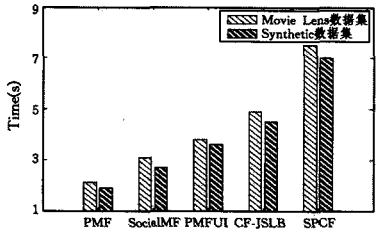

图 8 不同数据集下各算法运行时间比较($K=10$)

从图中可以看出，PMF 的运行速度最快，SocialMF 和 PMFUI 次之，CF-JSLB 排第 4 名，SPCF 的运行速度最慢。由此发现，矩阵分解方法与基于内存的方法相比，效率要高得多，且考虑的关系越多，其运行速度就越慢，时间复杂度就越高。由于 CF-JSLB 算法在邻居选择阶段做了 Jaccard 相似度的预处理工作，因此速度稍微有所下降，但总的来说，CF-JSLB 算法的运行时间是可以接受的。

结束语 本文提出了一种基于 Jaccard 相似度和位置行为的协同过滤推荐算法。针对推荐过程中存在的只利用用户-项目评分信息的问题，首先利用 Jaccard 相似度对用户(项目)做预处理工作，而后通过用户(项目)间的位置信息挖掘出用户(项目)间的潜在影响力找到最近邻居集合，最后将其融入到概率矩阵分解模型中以生成最终的推荐结果。通过一系列的实验结果表明，该方法与传统的推荐算法相比进一步改善了推荐效果。

由于基于矩阵分解的协同过滤推荐算法存在数据稀疏的共性问题，目前也有许多相关的研究，因此在未来的工作中，如何缓解数据稀疏性问题是下一步工作的重点。另外还将考虑如何将本文提出的方法应用到其他的推荐算法上。此外，除了概率矩阵分解算法，还将利用更多的上下文信息到已有的推荐算法上以进一步改善推荐的效果。

参 考 文 献

[1] Ma H, Zhou D, Liu C, et al. Recommender systems with social regularization[C]// Proceedings of the 4th ACM International Conference on Web Search and Data Mining. Hong Kong, China, 2011: 287-296

[2] Jamali M, Ester M. A Matrix factorization technique with trust propagation for recommendation in social networks[C]// Proceedings of the 4th ACM Conference on Recommender Systems. Barcelona, Spain, 2010: 135-142

[3] Guo Lei, Ma Jun, Chen Zhu-min, et al. Incorporating Item Relations for Social Recommendation[J]. Chinese Journal of Computers, 2014, 37(1): 219-228(in Chinese)
郭磊，马军，陈竹敏，等. 一种结合推荐对象间关联关系的社会化推荐算法[J]. 计算机学报，2014，37(1)：219-228

[4] Jiang Meng, Cui Peng, Liu Rui, et al. Social contextual recommendation[C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management. Maui, USA, 2012: 45-54

[5] Liu Chen-guan, Lin Hui-ping, Xiong Yi-bing. A Web service recommendation approach based on situation awareness[C]// Proceedings of the International Conference on Services Computing (SCC). California, USA, 2013: 432-437

[6] Dong Yu-xiao, Tang Jie, Wu Sen, et al. Link prediction and recommendation across heterogeneous social networks[C]// Proceedings of the ICMD. Washington: IEEE Computer Society, 2012: 181-190

[7] Wang Hao, Manolis T, Nikos M. Location recommendation in location-based social networks using user check-in data[C]// Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information System. New York, USA, 2013: 374-383

[8] Steffen R, Gantner Z, Freudenthaler C. Fast context-aware recommendations with factorization machines[C]// Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2011: 635-644

[9] Chen Kai-long, Chen Tian-qi, Zheng Guo-qing. Collaborative personalized tweet recommendation[C]// Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2012: 661-670

[10] Chen Tian-qi, Tang Lin-peng, Liu Qin. Combining factorization model and additive forest for collaborative followee recommendation[C]// Proceedings of the KDD Cup Workshop. New York: ACM Press, 2012

[11] Hong Liang-jie, Doumith AzizS, Davison BrianD. Co-Factorization machines: Modeling user interests and predicting individual decisions in twitter[C]// Proceedings of the WSDM. New York: ACM Press, 2013: 557-566

[12] Salakhutdinov R, Mnih A. Probabilistic matrix[C]// Proc of the 21st Annual Conf on Neural Information Processing Systems. New York: Curran Associates Inc, 2008: 1257-1264

[13] Zhao Qin-qin, Lu Kai, Wang Bin. SPCF: A Memory Based Collaborative Filtering Algorithm via Propagation[J]. Chinese Journal of Computers, 2013, 36(3): 671-676(in Chinese)
赵琴琴，鲁凯，王斌. SPCF：一种基于内存的传播式协同过滤推荐算法[J]. 计算机学报，2013，36(3)：671-676

[14] Jamali M, Ester M. A matrix factorization technique with trust propagation for recommendation in social networks[C]// Proc of the 4th ACM Conf on Recommender Systems. New York, 2010: 135-142

[15] Huang Chuang-guang, Yin Jian, Wang Jing, et al. Uncertain neighbours' collaborative filtering recommendation algorithm[J]. Chinese Journal of Computers, 2010, 33(8): 1369-1377(in Chinese)
黄创光，印鉴，汪静，等. 不确定近邻的协同过滤推荐算法[J]. 计算机学报，2010，33(8)：1369-1377