基于个性化情境和项目的协同推荐研究

高　旻　吴中福

(重庆大学计算机学院,重庆 400044)

摘要:为提高基于项目的协同过滤推荐 SlopeOne算法的预测结果,在算法的项目相异性计算和评分预测过程中引入个性化情境因素.首先对基于项目的协同推荐方法进行综述,然后针对不同情境下的评分记录进行项目间相异性计算,根据此结果计算检验集中的项目在不同情境下的预测评分,并以预测结果为依据为每个用户得到个性化情境,进而为用户对新的资源项目进行评分预测.最后在标准的 Movielens数据集上进行实验,其中,U2的训练集和测试集用来训练个性化情境,其他数据集用来检验算法的预测结果.通过对改进的推荐算法与经典的基于项目的协同过滤算法 SlopeOne进行比较,实验数据表明改进后算法的推荐结果有较大提高.

关键字:协同过滤;推荐算法;项目相异性;情境;个性化

中图分类号:TP311　　**文献标识码**:A　　**文章编号**:1001-0505(2009)增刊(Ⅰ)-0027-05

Personalized context and item based collaborative filtering recommendation

Gao Min　Wu Zhongfu

(College of Computer Science, Chongqing University, Chongqing 400044, China)

Abstract: In order to improve the result of item-based collaborative filtering (CF) recommendation approach, this paper incorporates personalized context into the computation of item differences and rating prediction. First, the prior research and the problem of item-based CF approaches are reviewed. Then, item differences according to the ratings in different contexts are calculated. Based on the item differences, ratings are extrapolated for the items in examining dataset. The personalized context for every user is further identified and new items in training dataset are predicted according to the result of the prediction. Finally, an experiment is given to evaluate the proposed approach and it is compared with a typical item-based SlopeOne CF using Movielens dataset, in which the training and test datasets of U2 are used to obtain personalized context, the other datasets are applied to check the final prediction results. The experimental results show that the proposed approach provides better quality than SlopeOne.

Key words: collaborative filtering; recommendation algorithm; item difference; context personalization

随着互联网上资源的迅速增长,个性化推荐系统已经逐渐成为研究者和用户关注的重要研究内容.研究者提出多种推荐方法:基于内容的推荐、协同过滤推荐和混合推荐等,并结合先进的技术,如聚类、关联规则、贝叶斯网、神经网络和图模型等实现这些方法[1-2].

协同过滤(CF)是目前最成功的推荐技术,传统的协同过滤被称为基于用户的协同推荐(user-based CF),其基本思想是:目标用户对未评分项的评分可以通过相似用户对这些项目的评分进行预测.这种推荐方法存在扩展性差的问题,推荐效率随用户数目、项目数目的增多而明显降低.为解决这一问题,有研究者使用聚类缩小相似用户的搜索范围[3],或通过奇异值分解减少项目空间的维数[4-5].基于项目的协同过滤(item-based CF)更是从根本上解决了相似用户实时计算引起的时间复杂度随用户增长而指数上升的问题[6-7].基于项目的协同过滤的基本思想是:用户对项目的评分可以通过项目之间的相似性以及用户已

收稿日期:2009-05-12　作者简介:高旻(1980—),女,博士生;吴中福(联系人),男,教授,博士生导师,wzf@cqu.edu.cn.
基金项目:国家社科基金重大资助项目(ACA07004-08),中国博士后科学基金资助项目(20080440699),重庆市自然科学基金资助项目(2008BB2183),重庆市教育委员会科学技术研究资助项目(KJ071601),重庆市教育科学"十一五"规划资助项目(2008-ZJ-064).

有的评分进行预测.项目之间的相似性相对稳定,可以在较长时间间隔更新一次,明显改善了系统的扩展性问题.但基于项目的协同过滤的预测结果并没有得到明显的提高.

本文将个性化情境与基于项目的协同过滤结合,形成基于个性化情境和项目的协同过滤,其基本思想是通过某个性化情境中项目之间的相异性和用户的已知评分对未知项进行评分预测.这里的项目相异性不是通过整个评分集实现,而是基于某一情境,更适合于这一情境下的用户使用.推荐时只使用某一情境下的数据集,使得推荐的效率进一步提高,实验结果表明新算法能有效提高推荐质量.

1 相关工作

大多数协同过滤推荐算法都是基于用户-项目评分矩阵 R进行的. R是一个 $m\times n$阶矩阵,其中行代表用户,列代表项目,矩阵中的值 $r_{u,i}$是用户 u对项目 i的评分.协同过滤主要包括基于用户和基于项目的协同过滤.

1.1 基于用户的协同过滤

基于用户的协同过滤首先度量用户之间的相似性,相似性高的被作为目标用户的最近邻,然后根据最近邻对项目的评分预测目标用户对项目的评分.度量用户间相似性的传统方法,目前主要有修正的余弦相似性和皮尔森相关系数法等[8],计算方法如式(1)、式(2)所示.其中 u和 v表示用户, $\mathrm{sim}(u,v)$表示 u和 v的相似性, $I(u)$表示 u的已知评分项, $I(u)\cap I(v)$表示 u和 v评分项的交集.权重计算主要包括平均值、加权平均值和规范后加权平均值法(如式(3))等.式(3)中 $\bar{r}_u$是用户 u所有评分的平均值, $\mathrm{sim}(u)$表示 u的相似用户集合, $p_{u,i}$表示目标用户 u对项目 i的评分预测.

$$\mathrm{sim}(u,v)=\frac{\sum_{i\in I(u)\cap I(v)}(r_{u,i}-\bar{r}_u)(r_{v,i}-\bar{r}_v)}{\sqrt{\sum_{i\in I(u)\cap I(v)}(r_{u,i}-\bar{r}_u)^2}\sqrt{\sum_{i\in I(u)\cap I(v)}(r_{v,i}-\bar{r}_v)^2}} \tag{1}$$

$$\mathrm{sim}(u,v)=\frac{\sum_{i\in I(u)\cap I(v)}(r_{u,i}-\bar{r}_u)(r_{v,i}-\bar{r}_v)}{\sqrt{\sum_{i\in I(u)\cap I(v)}(r_{u,i}-\bar{r}_u)^2}\sqrt{\sum_{i\in I(u)\cap I(v)}(r_{v,i}-\bar{r}_v)^2}} \tag{2}$$

$$p_{u,i}=\bar{r}_u+\frac{\sum_{v\in \mathrm{sim}(u)}\mathrm{sim}(u,v)(r_{v,i}-\bar{r}_v)}{\sum_{v\in \mathrm{sim}(u)}\mathrm{sim}(u,v)} \tag{3}$$

1.2 基于项目的协同过滤

基于用户的协同过滤存在两大问题:稀疏性和扩展性问题.评分矩阵数据稀疏导致难以找到相似用户,影响预测效果.扩展性问题是指随着用户和项目的增加,计算量急剧增加,严重影响推荐的实时性.基于项目的协同过滤可以解决这些问题[7].

在大多数个性化推荐系统中,相对于用户的更新,项目信息的更新较慢,项目之间的关系相对稳定.所以可以先离线计算项目之间的相似性,对相似性进行评分预测.评分矩阵稀疏对项目间的相似性影响比对用户之间相似性的影响小得多.另外,评分预测过程只是相似项目的查找过程,在线计算速度快.

基于项目的协同过滤推荐算法由 Sarwar于 2001年提出[7]. Sarwar介绍了基于项目之间相似性进行计算的思想,计算方法与基于用户的协同计算方法相似.后来研究者提出建立回归模型,用线性或非线性的回归函数对未评分项进行评分计算.目前经典算法的是 Lemire提出的 Slope One算法[9].算法先计算项目之间的相异性 $\mathrm{dev}_{i,j}=\sum_{u\in U(i)\cap U(j)}\frac{(r_{u,i}-r_{u,j})}{|U(i)\cap U(j)|}$,然后进行评分预测 $p_{u,i}=\bar{r}_u+\sum_{i\in R_u}\mathrm{dev}_{i,j}/|R_u|$.公式中 $|U(i)\cap U(j)|$是同时对 i和 j进行评分的用户个数, R_u是用户 u除对 i之外的所有已知评分.算法只用了线性回归函数 $f(x)=x+b$的简单形式进行评分的预测,简单实用、效率高,预测结果不比其他算法差,得到了研究者的广泛关注.

虽然基于项目的协同推荐有很多优点,但其预测准确率还有待于进一步提高.本文分析用户个性化情境,结合个性化情境和基于项目的协同过滤算法进行评分预测,有效地提高推荐质量.

2 基于个性化情境和项目的协同推荐

情境信息是指人的行为或事件发生影响的上下文信息或者场景信息,如时间、地点等[10],情境信息在

很大程度上影响用户最终的选择.在一个系统中,情境包括多个因素,个性化情境是指最影响目标用户做出选择决策的因素.

2.1　情境信息的重要性

传统的推荐都很少考虑情境因素.而情境因素在一些个性化系统中可能很重要,例如,当用户看电影的时候,不同的时间、地点(家里或电影院)、和不同的人看等情境都会影响其选择[11].文献[2,11]提出了集成情境的多维信息推荐系统,将情境和基于用户的协同过滤方法相结合进行评分预测.虽然他们考虑了情境,但都是针对整个用户群体得到整体较优情境,然后再与经典的协同过滤算法相结合进行推荐.这些方法中的情境是面向全部用户的[12],没有为每个用户的个性化情境进行分析.

个性化情境分析是指在特定的推荐系统中,为每个用户找出哪个情境对他的评分最有影响.这种情境信息是因人而异的,因此被称为是个性化的情境.个性化情境是非常重要的,例如,新闻系统中时间和地点是很重要的情境.但有的人对时间和地点敏感,而有的人却只对时间或地点敏感,甚至时间和地点对他都没有影响.分析出对最有影响的情境因素,能使预测结果更加理想.

2.2　基于情境项目相异性矩阵的生成

步骤1　将评分矩阵D按照用户信息中的不同情境C_k分成n个子集$D_1, D_2, \cdots, D_n$.子集D_k中包括k_n个用户的评分集合.由于一个用户可能同时处于几个情境中,因此任意2个D_x和D_y之间可能存在交集.

步骤2　对2个不同的项目i和j,从评分矩阵中提取出在某一情境C下所有同时对i和j进行评分的用户组U和评分值$r_{u,i}$和$r_{u,j}$,组成U(i)和U(j).

步骤3　计算项目i和j的评分差异值$d_{i,j}=\dfrac{\sum_{u\in U(i)\cap U(j)}(r_{u,i}-r_{u,j})}{|U(i)\cap U(j)|}$,$|U(i)\cap U(j)|$是指同时对i和j进行评分的用户的个数.由$d_{i,j}$构成情境$C_k$项目相异性矩阵$D_{A_k}$.

2.3　个性化情境分析

分析哪些情境信息会对目标用户的选择产生明显的影响,对用户行为不产生影响的情境将不会在推荐中用到.

步骤1　将用户评分矩阵的80%作为训练集D,20%作为测试集T,将D按照不同情境C_k分成n个子集$d_1, d_2, \cdots, d_n$.

步骤2　设评价推荐质量的评价指标为$P_A(U,X)$,A为推荐算法,U为用户集合,X为训练集.例如$P_A(u_1,d_1)$是算法A在基于情境C_1的训练集d_1上进行训练,为用户u_1进行推荐的推荐质量的评价.最后形成基于情境的推荐评价矩阵P_C,

$$P_C=\begin{bmatrix} P_{A(1,d_1)} & P_{A(1,d_2)} & \cdots & P_{A(1,d_n)} \\ P_{A(2,d_1)} & P_{A(2,d_2)} & \cdots & P_{A(2,d_n)} \\ \vdots & \vdots & & \vdots \\ P_{A(p,d_1)} & P_{A(p,d_2)} & \cdots & P_{A(p,d_n)} \end{bmatrix} \tag{4}$$

步骤3　根据P_C为每个用户u找到相应行中具有最优值的$P_{C_{x,j}}$,就找到用户u的个性化情境C_x和其对应的训练集D_x.

2.4　基于个性化情境的评分预测算法

把传统项目之间相异度矩阵的计算推广到基于情境相异度矩阵的计算,推荐时根据用户的个性化情境,选择相异度矩阵,然后根据用户已评分项和待推荐项之间的差异估计待推荐项的评分.这里的情境可以是用户的性别、年龄、职业、住址、时间等所有可能对评分产生影响的因素.

输入:用户u,待推荐项j,用户表U,评分表D.

输出:目标用户对待推荐项目的评分$p_{u,j}$.

步骤1　针对用户表U和数据表D,根据2.1中步骤1得到各情境C_k对应的数据集合D_k.

步骤2　针对D_k,根据2.1中步骤3得到各情境对应的项目相异度矩阵D_{A_k}.

步骤3　根据2.2中的计算为用户u得到个性和情境C_x,得到对应的D_{A_x},找到用户u所有已知评分R_u,根据这些评分和这些项与j之间的相异度d预测对j的评分,使用均值法(式(5))进行计算.

$$p_{u,j}=\frac{\sum_{i\in R_u}(r_{u,i}-d_{i,j})}{|R_u|}=\bar{r}_u+\frac{\sum_{i\in R_u}d_{i,j}}{|R_u|} \tag{5}$$

循环步骤 3可以预测用户对所有待推荐项目的评分.算法中步骤 1和 2学习阶段相对稳定,可以离线计算.步骤 3是为用户推荐的计算阶段,可以在线性时间内完成 $T(n)=O(n)$, n是用户已知的评分数.

3 实验仿真与分析

3.1 实验数据集与评价标准

采用美国明尼苏达大学开发并公布的 Movielens十万数据集作为实验数据集.该数据集每个用户至少对 20部电影进行评分.用户信息结构为(性别,年龄,职业和住址).数据集分 5次从这 10^6条记录中选取了 80%作为训练集,20%作为测试集,形成了 U1 base,…, U5 base和 U1 test,…, U5 test.不失一般性,我们从第 2个数据集 U2 base和 U2 test进行个性化的情境选择.并使用 U1, U3, U4, U5进行方法的评分预测,最后根据 MAE(mean absolute error)评价标准对我们的算法和 SlopeOne进行比较.

数据集中并不含有时间、地点等信息,只包含用户信息和项目信息,因此在算法中,我们采用用户类别信息作为项目的情境集合.

实验采用统计精度度量方法中广泛使用的平均绝对误差 MAE作为评价标准.MAE通过计算预测的评分与实际评分之间的偏差度量预测的准确性,MAE越小,推荐质量越高.设所有预测评分集合 $\{p_1, p_2, \cdots, p_N\}$,对应的所有的实际评分为 $\{r_1, r_2, \cdots, r_N\}$,则预测结果的误差为

$$MAE=\frac{\sum_{i=1}^{N}|p_i-r_i|}{N}$$

3.2 实验过程及结果分析

我们对训练集 U2 base的 80000个评分进行分析得到基于情境的项目相异性矩阵.并对测试集 U2 test中的 20000个评分进行预测,得到每个用户个性化情境.最后使用 U1, U3, U4, U5进行方法的评分预测,根据预测结果和测试集中的实际评分计算平均绝对偏差 MAE.

3.2.1 基于情境的项目间的相异度

如果不考虑情境,项目间的相异度矩阵是对整个训练集进行计算的.考虑情境时,则针对基于情境的评分子集进行计算.例如,情境为男性的相异度矩阵是从训练集中提取所有男性用户的评分,由 2.2中的计算方法确定相异度矩阵中的值.式(6)是基于男性的相异度矩阵.矩阵中的行和列是各个电影,例如 0.793是电影 1和电影 2的评分相异度.

$$D_{A_1}=\begin{bmatrix}0 & 0.793 & 0.9 & \cdots\\ -0.793 & 0 & 0.352 & \cdots\\ -0.9 & -0.352 & 0 & \cdots\\ \vdots & \vdots & \vdots & \vdots\end{bmatrix} \tag{6}$$

3.2.2 个性化情境矩阵的生成

根据评分子集和相异度矩阵,对相应情境下测试集中的用户进行评分预测.预测后的值与真实值进行 MAE计算,得到各情境下各用户评分的 MAE,见式(7).矩阵中的行是各个用户,列式对应的情境,每个值表示用户 u_i在情境 c_j下的 MAE值,MAE最小的那个情境即为该用户的个性化情境.

$$\begin{array}{c}\\ U_1\\ U_2\\ U_3\\ \vdots\end{array}\begin{array}{c}\begin{array}{cccc}\text{全局} & \text{性别} & \text{年龄} & \cdots\end{array}\\ \begin{bmatrix}0.869 & 0.884 & 0.856 & \cdots\\ 0.701 & 0.615 & 0.659 & \cdots\\ 1.179 & 1.236 & 0.983 & \cdots\\ \vdots & \vdots & \vdots & \vdots\end{bmatrix}\end{array} \tag{7}$$

3.2.3 预测结果的统计特征分析

把算法与经典算法 SlopeOne进行比较,具体的比较结果见图 1.图中横坐标是已知评分的项目数,例如 15~40表明项目已经有 15~40个用户进行了评分.纵坐标是对应项目在测试集中的预测评分与真实评分比较后的 MAE值.图中每组柱状线,左侧柱状线是 SlopeOne算法预测值的 MAE值,右侧柱状线是改进算法的 MAE值.MAE值越小,说明算法预测的准确率越高.可见,

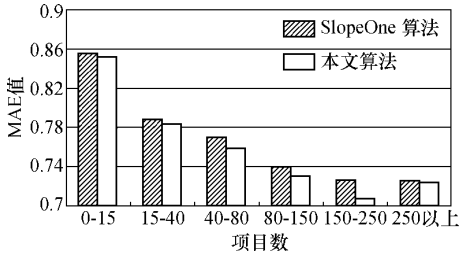

图 1 预测结果比较

本文算法的MAE值比SlopeOne算法好.同时,图1给出了项目已知的评分数与预测值之间的关系:

1)当已知评分项较少,在0~15项时,2个算法的结果相似,MAE值较大,预测质量不高.这是因为已知评分较少,导致项目之间的相异度未能反映项目之间的真实差距,导致预测准确性较差.

2)随着项目的已知评分项增多,MAE值不断下降,预测质量较高.当项目的已知评分在150~250时,本文算法达到最优值,接近0.7.

3)当项目的已知评分项在250项以上时,MAE有所反弹,略高于0.72.原因是当评分非常多时,其评分呈现多样性,反而降低了项目相异度的真实性.但无论在哪个取值段内,本文算法的预测质量都有明显提高.

4 结语

提出了一种结合用户个性化情境和基于项目的协同过滤推荐方法.该方法首先计算各情境下的项目相异度矩阵,根据相异度矩阵和基于项目的协同过滤算法对每个用户进行预推荐,根据预推荐的结果确定用户的个性化情境.然后基于个性化情境和该情境下的相异度矩阵对用户进行推荐.由于该方法先判断对用户推荐影响最大的情境因素,更符合人类在不同情境下有不同需求的事实.最后在Movielens数据集上进行了实验,对算法的推荐效率进行分析并与传统的基于项目的协同推荐算法进行比较.实验数据表明本文提出的算法能有效提高预测准确度并提高推荐质量.下一步的研究工作将把基于用户的协同推荐结合与个性化情境结合,只在相同情境下选择相似性用户,在提高计算效率的同时进一步提高推荐质量.

参考文献(References)

[1] 李聪,梁昌勇,马丽.基于领域最近邻的协同过滤推荐算法[J].计算机研究与发展,2008,45(9):1532-1538.
Li Cong, Liang Changyong, Ma Li. A collaborative filtering recommendation algorithm based on domain nearest neighbor [J]. Journal of Computer Research and Development, 2008, 45(9): 1532-1538. (in Chinese)

[2] 张光卫,康建初,李鹤松,等.面向场景的协同过滤推荐算法[J].系统仿真学报,2006,18(增2):595-601.
Zhang Guangwei, Kang Jianchu, Li Hesong, et al. Context based collaborative filtering recommendation algorithm [J]. Journal of System Simulation, 2006, 18(z2): 595-601. (in Chinese)

[3] 邓爱林,左子叶,朱扬勇.基于项目聚类的协同过滤推荐算法[J].小型微型计算机系统,2004,25(9):1665-1670.
Deng Ailin, Zuo Ziye, Zhu Yangyong. Collaborative filtering recommendation algorithm based on item clustering [J]. Journal of Chinese Computer Systems, 2004, 25(9): 1665-1670. (in Chinese)

[4] 赵亮,胡乃静.个性化推荐算法设计[J].计算机研究与发展,2002,39(8):986-991.
Zhao Liang, Hu Naijing. Algorithm design for personalization recommendation systems [J]. Journal of Computer Research and Development, 2002, 39(8): 986-991. (in Chinese)

[5] Sarwar B, Karypis G, Konstan J, et al. Incremental singular value decomposition algorithms for highly scalable recommender systems [C]//Fifth International Conference on Computer and Information Technology. Shanghai, China, 2002: 399-404.

[6] 邓爱林,朱扬勇,施伯乐.基于项目评分预测的协同过滤推荐算法[J].软件学报,2003,14(9):1621-1628.
Deng Ailin, Zhu Yangyong, Shi Bole. A collaborative filtering recommendation algorithm based on item rating prediction [J]. Journal of Software, 2003, 14(9): 1621-1628. (in Chinese)

[7] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms [C]//Proceedings of the 10th International Conference on World Wide Web. Hong Kong, 2001: 285-295.

[8] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6): 734-749.

[9] Lemire D, Maclachlan A. Slope one predictors for online rating-based collaborative filtering [C]//Proceedings of the SIAM Data Mining Conference. Newport Beach, CA, USA: Society for Industrial Mathematics, 2005: 21-25.

[10] 姚忠,吴跃,常娜.集成项目类别与语境信息的协同过滤推荐算法[J].计算机集成制造系统,2008,14(7):1449-1456.
Yao Zhong, Wu Yue, Chang Na. Collaborative filtering recommender algorithm for integrating item category and contextual information [J]. Computer Integrated Manufacturing Systems, 2008, 14(7): 1449-1456. (in Chinese)

[11] Adomavicius G, Sankaranarayanan R, Sen S, et al. Incorporating contextual information in recommender systems using a multidimensional approach [J]. ACM Transactions on Information Systems (TOIS), 2005, 23(1): 103-145.

[12] Gao Min, Wu Zhongfu. Incorporating pragmatic information in personalized recommendation systems [C]//The 11th International Conference on Informatics and Semiotics in Organizations. Beijing, China, 2009: 156-164.