结合用户判断力和相似性的协同推荐算法

张 莉 薛羽青
(对外经济贸易大学信息学院 北京 100029)

摘 要 作为解决信息超载问题的有效方式,协同过滤技术已被成功地应用到推荐系统。为进一步提高协同过滤算法的性能,首先利用用户评分的历史信息,估计用户的判断力。接着结合用户间的相似性,提出一种改进的协同推荐算法。实验结果表明用户的判断力可与用户的推荐能力正相关,也验证了用户判断力深入抽取用户评分信息以及影响用户采纳某项推荐结果的因素,以更好地刻画用户之间的相似性,从而提高算法的推荐准确度。

关键词 协同过滤,用户判断力,相似性,推荐系统
中图法分类号 TP391 **文献标识码** A

Collaborative Recommendation Algorithm Combining User's Judging Power and Similarity

ZHANG Li XUE Yu-qing
(School of Information Technology & Management Engineering, University of International Business and Economics, Beijing 100029, China)

Abstract As an effective way to solve information overload, collaborative filtering(CF) technology has been successfully used in recommendation system. To improve the performance of CF algorithm, first, this paper evaluated user's judging power based on historical scoring. Then combining user's judging power and similarity, an improved collaborative recommendation algorithm was proposed. Experimental results show that judging power has positive correlation with recommendation abilities of users, which also verify that judging power extracts the depth information from historical scoring and factors influencing a user on adopting recommendation results. So it can characterize the similarity between users better and improve the accuracy of a recommendation algorithm.

Keywords Collaborative filtering, User's judging power, Similarity, Recommendation system

1 引言

随着 Web2.0 技术的成熟,个性化推荐技术迅猛发展。个性化推荐系统在给电子商务领域带来巨大商业利益的同时,也以其准确、高效、个性化的特点在社会生活各领域服务大众。在不同领域研究人员努力下,个性化推荐技术的研究和发展也从未停止,推荐系统也日趋成熟,而协同过滤是目前推荐系统中广泛采用的推荐技术。然而,随着网络用户和商品数量的增加,协同过滤面临严峻的数据稀疏和推荐实时性的挑战。但这丝毫不影响它的受欢迎度,国内外学者对协同过滤算法的改进不断推陈出新,如基于概率的协同过滤算法、基于矩阵分解的协同过滤、基于神经网络的协同过滤、基于聚类的协同过滤等[1,2]。各种模型诸如概率模型、极大熵模型、线性回归、Gibbs 抽象、Bayes 模型等也大放异彩[3]。

本文在这些研究的基础上,充分利用用户评分的历史信息,计算用户的判断力,并结合用户的相似性度量,提出了改进的协同推荐算法,并通过实验验证了算法的性能。

2 相关工作

协同过滤算法依据的是相似用户具有相似的兴趣爱好,发现谁可以协同目标用户是协同过滤算法的主要过程之一,基于用户的协同过滤算法采用的是寻找目标用户的 N 个相似最近邻来协同目标用户。因而为了提高协同过滤算法的推荐准确性,降低计算复杂性,国内外学者在用户相似性计算和最近邻用户的选取方面进行了大量的研究,相关的研究可以划分成 3 个分支:1)在传统的相似性计算方法中加入时间、项目、兴趣等因素,调整相似性计算公式。Nathan 等人于 2010 年在用户相似性计算时增加时间权重[4];高滢等于 2008 年按照用户所评价的项目数量,引入用户等级函数,将用户分为不同的等级,改进用户相似性度量[5];文献[6]从用户共同评价的项目数量和用户对共同评价项目的评分方差两方面衡量用户之间的相似程度,提出最近邻用户动态重排序相似度计算方法,上述算法都取得了比较好的推荐效果;Tieli Sun 等于 2009 年利用项目的相似性改进用户相似性计算方法[7]。2)利用社会网络信息改进传统 CF 的相似性计算方法,部分学者 Yuan, Q 等于 2009 年、De Meo, P 等于 2011 年、Konstas I 等于 2009 年将社会网络的朋友关系、会员关系或社会网络标签数据与用户评分数据融合,改进传统 CF 算法中用户相似性计算方法,解决 CF 中的冷启动问题[8-10];另外 Chen W, Fong S 于 2010 年将用户间信任度融入到相似性计算,认为

本文受国家社科基金项目:社会网络中意见领袖对个性化信息推荐服务质量的影响研究(13BTQ027)资助。
张 莉(1972—),女,博士,副教授,主要研究方向为智能信息技术、电子商务、社会网络分析等,E-mail:tasummer@sina.com;薛羽青(1991—),女,硕士生,主要研究方向为社会网络分析、数据挖掘。

信任度高的用户，其相似性越大[11]。3）基于社会网络的结构图改进相似性计算，R. Zheng 等于 2007 年基于社会网络关系图计算两个用户间的距离，用距离修正用户相似度，认为两个用户距离越近，其相似性越大[12]，但是随着网络用户量的增加，用户间距离计算的复杂度也随之增加。部分学者将物质扩散与热传导理论引入到个性化推荐算法中，提出了一批性能良好的改进算法。周涛等于 2007 年提出了基于用户-产品二部分图资源分配模型，协同推荐中可以将用户的资源理解为推荐能力，据此设定用户的权重进行推荐[13]。杨兴耀、于炯等于 2013 年融合用户评分的奇异值以及文献[5]的资源分配模型改进了用户相似性计算方法[14]，这不仅从某种程度上提高了推荐算法的性能，缓解了冷启动问题，更重要的是为相关研究人员打开了全新的研究视角。

上述研究在一定程度上提高了协同过滤算法的性能，但没有考虑用户的专业判断能力对推荐结果的影响。由于相似度阈值的限制或者最近邻个数的限制，可能会导致判断能力较好的用户近邻未能进入用于推荐的最近邻集合。文献[15]实证研究了口碑传播过程中信息来源者所拥有的专业能力与判断能力已经成为影响信息接受者采纳的重要因素。因而在上述研究的基础上，受文献[15]的启发，借助文献[16]计算用户判断力的方法，本文提出了改进的协同过滤算法，它综合考虑用户间的相似性和用户的判断能力，更加真实地反映了用户接受推荐的影响因素。

3 传统的基于用户的协同过滤算法(SCF)

传统的基于用户的协同过滤算法首先计算目标用户与其它用户的相似性，选取相似性最大的 N 个用户组成目标用户的最近邻集合；然后对目标用户的最近邻评价的项目集合中的每个项目，预测目标用户的评分，并进行降序排列，把前 M 个项目推荐给目标用户。

3.1 用户相似性度量

Pearson 系数是目前常采用的相似性度量方法，Pearson 系数在用户共同的评分项目的基础上度量用户间相似度。设 $I(u)$、$I(v)$ 分别表示用户 u、v 的评分集合，$I(u)\cap I(v)$ 表示 u、v 共同评分的项目集合，则用户相似性($\mathrm{sim}(u,v)$)计算如式(1)所示。

$$\mathrm{sim}(u,v)=\frac{\sum_{i\in I(u)\cap I(v)}(R_{u,i}-\overline{R_u})((R_{v,i}-\overline{R_v})}{\sqrt{\sum_{i\in I(u)\cap I(v)}(R_{u,i}-\overline{R_u})^2}\sqrt{\sum_{i\in I(u)\cap I(v)}(R_{v,i}-\overline{R_v})^2}} \tag{1}$$

其中，$R_{u,i}$ 和 $R_{v,i}$ 分别表示用户 u 和用户 v 对项目 i 的评分值($i\in I(u)\cap I(v)$)，$\overline{R_u}$、$\overline{R_v}$ 分别表示用户 u 和用户 v 的评分均值，由于不同用户的评分标准可能不同，在本文涉及的算法中对评分数据进行了标准化处理。

3.2 用户评分预测方法

本文采用平均加权法预测评分并产生推荐结果，设用户 u 对未评分项目 $t\in I(N_u)$ 的预测评分为 $p_{u,t}$，其计算方法如式(2)所示。

$$p_{u,t}=\overline{R_u}+\frac{\sum_{i\in N(u)}\mathrm{sim}(u,i)*(R_{i,t}-\overline{R_1})}{\sum_{i\in N(u)}\mathrm{sim}(u,i)} \tag{2}$$

其中，$\overline{R_u}$ 是目标用户 u 对已评价项目的平均评分，$\overline{R_1}$ 是用户 i 的平均评分。

4 改进的基于用户判断力的协同过滤算法

在信息传播过程中，信息发送者的专业能力和判断能力影响信息接受者的接受意愿[16]，据此我们假设：用户的判断力与影响用户的推荐能力正相关，从而正向影响推荐算法的准确度。协同过滤算法中核心过程是发现谁可以协同目标用户，传统的基于用户的协同过滤算法是寻找目标用户的 N 个相似最近邻，通过最近邻实现推荐。因而最近邻集合的推荐能力直接影响算法的推荐性能，但是传统的算法在选取最近邻集合时没有区分用户的推荐能力。由于相似度阈值或最近邻个数的限制，可能会导致推荐能力较好的用户近邻未能进入用于推荐的最近邻集合。因而要得到推荐质量较好的最近邻集合，需要综合考虑了用户之间的相似性和用户的推荐能力。在本文中即综合考虑用户之间的相似性和用户自身的判断力。

4.1 用户判断力的计算

假设 N 个用户对 M 个产品进行了评分，每个用户判断能力不同，每个商品都有不同的内在的质量。用户的判断能力和商品内在的质量隐含在已有的评分信息中，人们经常用 N 个用户评分均值表示商品的内在质量，这种计算方法受评分的噪声数据影响很大，并且认为每个用户具有相同的判断力，这与前面的假设不符。为此本文根据文献[16]中的用户判断力估算方法，将用户的判断力用其评分与项目质量的方差来衡量，项目质量由所有用户的评分来决定，其计算步骤如下：

(1)利用式(3)计算数据集中项目 j 的质量 q_j

$$q_j=\sum_{u=1}^{N}A_{uj}\times w_u\times r_{uj} \tag{3}$$

其中，每个用户的初始权重均为 $w_u=1/N$，A_{uj} 表示用户 u 是否评价过项目 j，若评价过则 $A_{uj}=0$，反之为 1，r_{uj} 表示用户 u 对项目 j 的评分。

(2)利用式(4)计算用户的评价方差

$$v_u=\frac{1}{\sum_{i=1}^{M}A_{u,i}}\sum_{j=1}^{M}A_{u,j}(r_{u,j}-q_j)^2 \tag{4}$$

(3)利用式(5)计算用户的判断力

$$w_u=\frac{v_u^{-1}}{\sum_{v=1}^{N}v_v^{-1}} \tag{5}$$

(4)重复(1)－(3)，直至前后两轮的项目质量的差达到设定的阈值，得到最终所有用户的判断力 w_u。

4.2 改进的协同过滤算法

本文提出的算法主要是改进目标用户最近邻集合的选取，为了查看用户判断力对协同推荐算法性能的影响，采用了两种方法构造目标用户的最近邻用户集合，分别用 UPCF1、UPCF2 标识。

UPCF1：将用户判断力作为权重融入到用户相似性计算中，如式(6)所示，并用其计算目标用户 u 与其他用户的相似性，选取 K 个最相似用户作为推荐的最近邻集合 N_u。

$$\mathrm{sim}(u,v)=w_v\frac{\sum_{i\in I(u)\cap I(v)}(R_{u,i}-\overline{R_u})(R_{v,i}-\overline{R_v})}{\sqrt{\sum_{i\in I(u)\cap I(v)}(R_{u,i}-\overline{R_u})^2}\sqrt{\sum_{i\in I(u)\cap I(v)}(R_{v,i}-\overline{R_v})^2}} \tag{6}$$

由于用户的判断力不同，因此通过式(6)计算得到的用户间相似性矩阵是不对称的。

UPCF2：采用式(1)计算目标用户 u 与其它用户的相似性，选取 $N1$ 个用户组成候选最近邻集合，然后根据用户的判断力将最近邻集合中的用户重新排序，选取判断力最大的前 $K(K<N1)$ 个用户组成用于推荐的最近邻集合 N_u。

5 实验

实验数据采用美国 GroupLens 实验室提供的 MovieLens 数据集(100KB)。数据集中包含了 943 个用户对 1682 部电影的 100000 个评分，评分范围为 1 至 5，在实验时将数据归一化处理在[0,1]范围，实验结果也进行了标准化处理。

5.1 评价标准

为了检验算法的性能，本文采用平均绝对偏差 MAE 对算法进行衡量，MAE 易于理解，可以直观地对推荐质量进行度量，是最常用的一种推荐质量度量方法。MAE 仅对测试集中用户 i 已经评过分的项目进行度量，这些项目数 $m_i \leqslant m$(项目总数)，预测评分为 $p_{i,t}(t=1,2,\cdots,m_i)$，用户实际评分是 $r_{it}(t=1,2,\cdots,m_i)$，那么对用户推荐结果的评价如式(6)所示。

$$MAE=\frac{\sum_{t=1}^{m_i}|p_{i,t}-r_{it}|}{m_i} \tag{7}$$

MAE 是通过计算目标用户的预测评分与实际评分之间的偏差来度量预测的准确性，因而 MAE 指标的值越小，推荐质量越高。

5.2 实验结果

为了验证用户判断力对推荐算法性能的影响，我们采用前面的 SCF、UPCF1、UPCF2 分别进行了 3 个实验，3 个算法的 MAE 的对比实验结果见图 1。其中，实验用于推荐的用户最近邻数 $k=10,20,30,40,50,60,70,80$，最终取推荐项目数为 20。而 UPCF2 先根据相似性计算(式(1))选取 $N1$ 个最相似用户作为候选近邻集合，$N1=3/2k$。由图 1 可以看出，将用户的判断力作为一种权重加入传统的协同过滤算法中，对算法的性能改进不是很显现，但将用户判断力用于改变最近邻用户顺序(UPCF2)，可以有效改善推荐的准确度，特别当选取的近邻个数越大时，用户判断力对算法性能的影响越明显，这说明用户的判断力会影响其推荐能力。

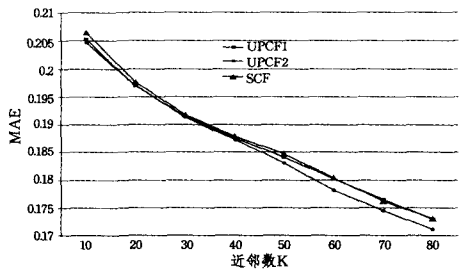

图 1 SCF、UPCF1、UPCF2 实验结果对比分析

从图 1 可以看出，当 $k\in[10,40]$ 时，3 个算法的推荐准确性没有明显的差异。但随着 k 的增加，UPCF2 表现出比较好的性能，UPCF1 越来越逼近 SCF 算法的性能。这显示用户对其他用户的影响力与相似性有相同的变化趋势，即用户与其他用户越相似，对其影响力也越大。

结束语 在信息传播过程中，信息传播者所拥有的判断能力已经成为影响信息接受者接受意愿的重要因素。本文将用户自身属性信息-判断力融入到基于用户的协同过滤算法，综合考虑用户之间的相似性和用户的判断能力，以更加真实地反映推荐性能的影响因素。实验结果显示了用户的判断力对推荐算法的性能有正向的影响作用。但本文所采用的全局环境下的用户判断力，对推荐算法性能的改观还不够明显。由于不同的用户在不同主题信息的影响力是不同的，因此下一步将会研究不同主题下的用户判断力估计方法及其此在协同过滤算法中的应用，期望推荐性能有更大的改进。

参 考 文 献

[1] Su Xiao-yuan, Taghi M K. A Survey of Collaborative Filtering Techniques[J]. Advances in Artificial Intelligence, 2009(1): 1-19

[2] Fidel C, V'ictor C, Diego F, et al. Comparison of Collaborative Filtering Algorithms: Limitations of Current Techniques and Proposals for Scalable, High-Performance Recommender Systems[J]. ACM Transactions on the Web, 2011, 5(1): 2-33

[3] 刘建国，周涛，汪秉宏. 个性化推荐系统的研究进展[J]. 自然科学进展，2009，19(1)：1-15

[4] Liu N N, Zhao Min, Xiang E, et al. Online evolutionary collaborative filtering[C]// Proceedings of the fourth ACM Conference on Recommender systems. New York, 2010: 95-102

[5] 高滢，齐红，刘杰，等. 结合似然关系模型和用户等级的协同过滤推荐算法[J]. 计算机研究与发展，2008，45(9)：1463-1469

[6] 张迎峰，陈超，俞能海. 基于最近邻用户动态重排序的协同过滤方法[J]. 小型微型计算机系统，2011，32(8)：1581-1586

[7] Sun Tie-li, Wang Li-jun, Guoin Qing-he. A Collaborative Filtering Recommendation Algorithm Based on Item Similarity of User Preference [C] // The 2th International Workshop on Knowledge Discovery and Data Mining. 2009, 1: 60-63

[8] Yuan Q, Zhao S, Chen L, et al. Augmenting collaborative recommender by fusing explicit social relationships[C]// Workshop on Recommender Systems and the Social Web. Recsys, 2009: 49-56

[9] De Meo P, Ferrara E, Fiumara G, et al. Improving recommendation quality by merging collaborative filtering and social relationships[C]// 2011 11th International Conference on Intelligent Systems Design and Applications(ISDA). 2011: 587-592

[10] Konstas I, Stathopoulos V, Jose J M. On social networks and collaborative recommendation[C]// Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2009: 195-202

[11] Chen W, Fong S. Social network collaborative filtering framework and online trust factors: a case study on Facebook[C]// 2010 Fifth International Conference on Digital Information Management (ICDIM). 2010: 266-273

[12] Zheng R, Provost F, Ghose A. Social Network Collaborative Filtering: Preliminary Results [C] // Proceedings of the Sixth Workshop on eBusiness(WEB2007). Montreal, 2007: 47-55

[13] Zhou T, Ren J, Medo M, et al. Bipartite network projection and personal recommendation[J]. Physical Review E, 2007, 76(4): 046115

[14] 杨兴耀，于炯，等. 融合奇异性和扩散过程的协同过滤模型[J]. 软件学报，2013，24(8)：1868-1884

[15] Bansal Harvir S, Voyer P A. Word-of-Mouth Processes within a Services Purchase Decision Context [J]. Journal of Service Research, 2000, 3(2): 166-177

[16] Yu Y K, Zhang Y C, Laureti P, et al. Decoding information from noisy, redundant, and intentionally distorted sources[J]. Physica A: Statistical Mechanics and its Applications, 2006, 371(2): 732-744