

一种改进的混合推荐算法

宋文君, 郭 强, 刘建国

(上海理工大学 管理学院 复杂科学研究中心, 上海 200093)

摘要: 基于用户的近期行为能够更好地反映其潜在的兴趣偏好的思想,提出了一种基于有限时间窗口的改进混合推荐算法.在标准数据集 Netflix 上的实验结果表明,只采用大约 31.11% 的用户近期历史记录,所得到的推荐结果准确性可以平均提高 4.22%,而推荐列表多样性可以提高 13.74%.另外还发现新提出的算法适用于不同活跃程度的用户,这可以极大地降低大规模数据所引发的计算复杂性问题.

关键词: 混合推荐算法; 时间窗口; 用户-产品二部分网络

中图分类号: TP 312 **文献标志码:** A

Improved Hybrid Recommendation Algorithm

SONG Wenjun, GUO Qiang, LIU Jianguo

(Research Center of Complex Systems Science, Business School, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Since the recent behaviors are more effective to capture the users' potential interests, an improved hybrid recommendation algorithm was proposed for making use of the partial recent information. The experimental results on the benchmark dataset Netflix indicate that by only adopting approximately 31.11% recent rating records, the accuracy can be improved by an average of 4.22%, and the diversity can be improved by 13.74%. Furthermore, it is found that the improved algorithm is suitable for the users with different level of activeness. The study is valuable in both theory and practice, and it could effectively handle the calculation complexity triggered by massive data.

Key words: *hybrid recommendation algorithm; time window; user-object bipartite network*

互联网技术的迅猛发展使得信息超载问题日益严重^[1]. 个性化推荐系统利用用户的历史行为来预测其潜在的需求, 能够有效地解决信息超载问

题^[2-3]. 现有的推荐算法包括协同过滤算法(collaborative filtering)^[4-5]、基于内容的推荐算法(content-based algorithm)^[6]、基于网络结构的推荐

收稿日期: 2014-03-01

基金项目: 国家自然科学基金资助项目(61374177, 71371125, 71171136); 国家教育部人文社科基金资助项目(13YJA630023, 14ZR1427800)

第一作者: 宋文君(1989-), 女, 硕士研究生. 研究方向: 个性化推荐. E-mail: swj8912@sina.com

通信作者: 刘建国(1979-), 男, 教授. 研究方向: 复杂网络、个性化推荐. E-mail: liujg004@ustc.edu.cn

算法(network-based algorithm)等^[7-13]. 其中, 基于网络结构的推荐算法就是将热传导(heat conduction)^[7-9]、物质扩散(mass diffusion)^[10-12]等原理应用到个性化推荐算法的研究中, 已经取得了很好的研究成果. Zhang 等^[7]利用用户对产品的打分信息, 实现了热传导的推荐算法. Zhou 等^[10]利用用户-产品二部分网络提出了一种基于物质扩散的推荐算法. 进一步, Zhou 等^[13]基于物质扩散与热传导原理提出了一种混合推荐算法(HHM).

HHM 算法虽然能够同时提高算法的准确性与多样性, 但是, 必须采用全部的用户信息^[14]. 然而用户的早期行为不能很好地反映其目前的兴趣偏好, 也就是说应该考虑时间信息对于推荐效果的影响. 近几年来, 许多学者尝试将时间因素融入到算法中来提升推荐效果. 例如, Liu 等^[15]提出了一种基于时间因素的推荐算法. 另外, Zhang 等^[16]将基于时间和拓扑结构这两种方法混合起来, 提出了一种用来抽取信息骨架的方法. 虽然这一方法只需要处理部分信息, 但是, 却缺乏时间窗口对推荐效果影响的研究, 这对于降低计算复杂性至关重要. 本文提出了一种基于有限时间窗口的改进混合推荐算法. 首先采用标准数据集 Netflix, 通过逐渐增大时间窗口的方式生成一系列训练集, 然后将每一个训练集作为已知的数据来预测用户的兴趣偏好, 最后利用测试集来检验推荐算法的效果. 实验结果表明, 采用部分用户近期数据能够同时提升推荐的准确性和多样性, 而且新改进的算法适用于不同活跃程度的用户. 表明本文的方法可以极大地降低计算复杂性, 非常具有实践价值.

1 模型与方法

1.1 基于二部分网络的混合推荐算法

用户-产品二部分网络包括一组由集合 $U = \{u_1, u_2, \dots, u_n\}$ 表示的用户节点, 一组由集合 $O = \{o_1, o_2, \dots, o_m\}$ 表示的产品节点, 以及连接这两组节点的连边, 由集合 $E = \{e_1, e_2, \dots, e_p\}$ 表示. 其中, 如果用户 u_j 选择过产品 o_i , 那么就在 u_j 和 o_i 之间连接一条边 $a_{ij} = 1$; 否则, $a_{ij} = 0$.

标准的热传导算法最初由 Zhang 等^[7]提出. 假设每个产品都具有一个初始资源, 并且它可以在用户-产品二部分网络上传递, 使得所有产品都会获得一个最终的资源. 这一资源传递的过程可以表示为

$$f' = Wf \quad (1)$$

式中, W 代表资源转移矩阵; f 表示产品的初始资源; f' 则表示最终资源.

热传导算法^[7]在推荐列表多样性上具有优势, 但是, 因为对冷门产品分配过多的资源而导致准确性很差. 物质扩散算法^[10]却因为更加关注流行产品, 可以表现出很高的推荐准确性. 为了综合上述两种方法的优势, Zhou 等^[13]提出了一种混合推荐算法(HHM)可以同时提高推荐结果的准确性和多样性. 资源转移矩阵可以表示为

$$W_{\alpha\beta, \text{HHM}} = \frac{1}{k_{o_\alpha}^{1-\lambda} k_{o_\beta}^\lambda} \sum_{j=1}^n \frac{a_{\alpha j} a_{\beta j}}{k_{u_j}} \quad (2)$$

式中, k_{o_α} , k_{o_β} 和 k_{u_j} 分别代表产品 o_α , o_β 以及用户 u_j 的度; $a_{\alpha j}$, $a_{\beta j}$ 分别代表用户 u_j 与产品 o_α , o_β 的连边; λ 为混合参数, $\lambda = 0$, 代表标准的热传导算法, 而 $\lambda = 1$, 则表示物质扩散算法.

当混合参数 λ 调节到一个最优值时, 该算法在准确性和多样性两方面都可以得到一个更好的推荐结果.

1.2 改进的混合推荐算法

本文基于用户的近期行为能够更好地反映其潜在的兴趣偏好的思想, 提出了一种基于有限时间窗口的改进混合推荐算法. 首先, 通过采用标准的 HHM 算法, 可以得到一个最优的混合参数使得推荐结果的准确率最高, 记为 λ_{opt} . 然后, 将最近发生的前 10% 的记录作为测试集, 另外按照下面不断地扩大时间窗口的方法, 划分出一系列训练集. 在剩下的用户记录中最大的时间信息记为 t_0 , 假设在时间标 $T \in [t_0 - \eta \Delta t, t_0]$ 这一范围内的记录就构成了第 η 个训练集, 其中, Δt 代表时间间隔, η 则表示训练集的编号. η 的最小值为 1, 代表第一个训练集, 它包含从时间标 t_0 向前倒推了一个单位时间间隔内的全部记录. η 的上界是原始训练集的生命周期与 Δt 的比值. 需要注意的是, 随着 η 值的不断增大, 训练集所包含的数据量也在不断增大, 但是, 测试集却保持不变. 另外, 将每一组训练集作为已知的数据, 来预测用户对没有选择过的产品的喜好程度. 这里采用改进的混合推荐算法, 其资源转移矩阵表示为

$$W_{\alpha\beta, H}^\eta = \frac{1}{(k_{o_\alpha}^\eta)^{1-\lambda} (k_{o_\beta}^\eta)^\lambda} \sum_{j=1}^n \frac{a_{\alpha j}^\eta a_{\beta j}^\eta}{k_{u_j}^\eta} \quad (3)$$

式中, $k_{o_\alpha}^\eta$, $k_{o_\beta}^\eta$ 和 $k_{u_j}^\eta$ 分别代表第 η 个训练集中产品 o_α , o_β 以及用户 u_j 的度; $a_{\alpha j}^\eta$, $a_{\beta j}^\eta$ 分别代表第 η 个训练集中用户 u_j 与产品 o_α , o_β 的连边.

最后, 采用准确率(precision)、召回率(recall)

和平均汉明距离(average Hamming distance)这3种指标来衡量新算法的推荐效果.

1.3 衡量指标

准确率 $P^{[17]}$:表示用户对系统所推荐的产品喜欢的概率,也就是系统推荐的产品中用户喜欢的产品所占的比例,即

$$P = \frac{1}{n} \sum_{i=1}^n \frac{h_i}{L} \quad (4)$$

式中, L 代表推荐列表的长度; h_i 代表同时出现在用户 u_i 的测试集和其推荐列表中的产品数目.

通常来说,当推荐列表长度 L 给定的时候,准确率越高,表明推荐结果越准确.

召回率 $R^{[17]}$:表示用户喜欢的产品被推荐的概率,即

$$R = \frac{1}{n} \sum_{i=1}^n \frac{h_i}{n_{ip}} \quad (5)$$

式中, n_{ip} 则代表测试集中用户 u_i 喜欢的产品数目.

同样地,召回率越高,说明推荐效果越好.

平均汉明距离 $S^{[17]}$:推荐结果的多样性可以采用平均汉明距离 $S = [H_{ij}]$ 来衡量,具体表示为

$$H_{ij} = 1 - Q_{ij}(L)/L \quad (6)$$

式中, $Q_{ij}(L)$ 则代表用户 u_i 和 u_j 的推荐列表中相同产品的数目.

最大值 $S = 1$,表明两个用户的推荐列表没有重复的产品,也就说明推荐系统的多样性最高;反之,如果 $S = 0$,则表示两个用户的推荐列表完全一致.

2 实证结果分析

2.1 实验数据

采用标准数据集 Netflix 来检验新算法的推荐效果. Netflix 的数据包括 8 609 个用户对 5 081 部电影的打分情况,是 netflix.com 网站从 2001 年 2 月至 5 月期间收集得到的. 根据本文提出的方法,最近发生的 41 924 条记录就构成了测试集. 假设划分训练集的时间间隔 Δt 为 2 天,那么,就可以得到 45 个训练集. 经过实验,可以得到最优的混合参数 $\lambda_{opt} = 0.51$,如图 1 所示.

2.2 实验结果

利用准确率和召回率这两个指标来衡量推荐算法的准确性,而平均汉明距离则用于衡量推荐列表的多样性. 图 2 表示这 3 种指标随着 η 增大的变化情况. 可以看出准确率和召回率基本上都呈现出一种先上升后下降的趋势,而多样性基本上呈下降

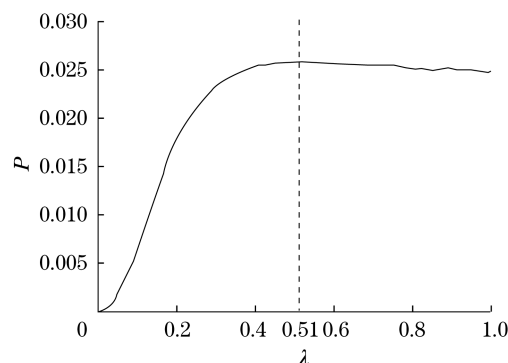


图1 混合参数与准确率的关系

Fig.1 Relation between hybridization parameter and precision

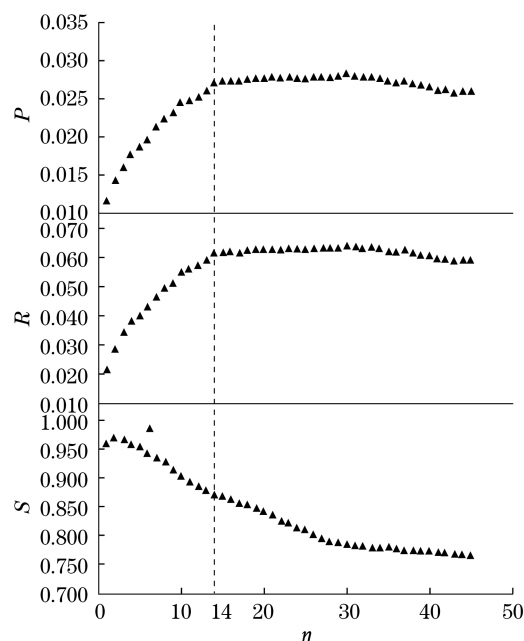


图2 推荐效果随训练集编号增大的变化情况

Fig.2 Relation between algorithmic performance and number of training set

趋势,也就是说,只采用部分近期数据就可以得到一个更准确且多样的推荐结果. 从图 2 中可以看出,只需要考虑 $14/45 \approx 31.11\%$ 的用户近期记录,所得到的推荐准确度可以平均提升 4.22%,而多样性可以提升 13.74%. 由此可见,为了生成一个更好的推荐结果而采用用户-产品二部分网络中的全部数据是没有必要的. 这一现象产生的原因可能在于用户的兴趣偏好是随时间动态变化的,考虑用户的早期行为会影响推荐结果的表现性. 由于最初训练集中的数据量太少,不能准确地反映出用户的兴趣偏好. 然而,当时间窗口不断增大的时候,已知的数据量越多,反而会干扰推荐的效果,也就是说,只考虑部分用户近期数据反而能够得到一个更好的推荐结果.

2.3 不同活跃程度的用户准确性

用户的活跃程度可以用他们的度来衡量. 数据集 Netflix 中用户的度分布情况近似呈现幂率形式^[2], 也就是说, 非常活跃的用户数量很少, 而绝大部分都是度小的用户. 为了研究本文提出的算法对不同活跃程度的用户是否都适用, 按照用户的度 k_u 将他们分为 5 类, 分别为 1~10, 11~20, 21~50, 51~100 和超过 100. 图 3(见下页)表示上述这 5 类用户的准确率随着 η 增大的变化情况. 从实验结果可以看出, 对于不同活跃程度的用户来说, 他们的准确率随着 η 的增大基本上呈现出一种先上升后下降的趋势, 也就是说, 对于不同活跃程度的用户来说, 本文提出的改进混合推荐算法都能够提高算法的准确性. 特别是对于新用户来说, 这里假设度不超过 10 的用户为新用户, 只需要考虑 31.11% 的用户近期记录, 准确率就可以提高 11%. 因此, 从提高推荐准确性的角度来看, 改进的混合推荐算法能够适用于不同活跃程度的用户.

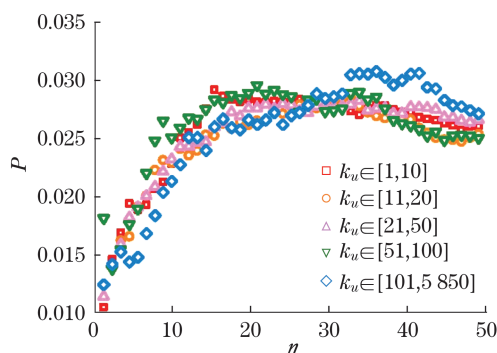


图3 不同活跃程度用户准确率的变化情况

Fig.3 Precision on users with different activeness

3 总结与展望

基于热传导和物质扩散原理的混合推荐算法^[13]能够同时提高推荐列表的准确性和多样性. 经典的方法是采用用户-产品二部分网络中的全部数据, 却忽略了时间窗口对于推荐算法效果的影响. 因此, 本文着重研究了时间窗口对于混合推荐算法的影响, 并且提出了一种基于有限时间窗口的改进混合推荐算法, 能够同时提高推荐结果的准确性和多样性. 首先采用标准数据集 Netflix, 通过逐渐扩大时间窗口的方法生成一系列训练集, 然后将每个训练集作为已知的数据来预测用户的兴趣偏好, 最后利用测试集来检验推荐算法的效果. 在 Netflix 数据

集上的实验结果表明, 只采用 31.11% 的近期数据, 所得到的推荐结果准确性可以平均提升 4.22%, 而多样性可以提升 13.74%. 另外还发现新提出的算法适用于不同活跃程度的用户. 本文的工作在理论和实践上都具有一定的价值. 在理论上, 本文的方法对于深入理解时间窗口对混合推荐算法的影响很有帮助; 在实践中, 能够降低大规模数据所引发的计算复杂性问题, 并且减少数据存储空间. 然而, 对于不同的数据集如何找到合适的时间窗口, 以及如何建立一个理论模型来解释为何采用部分近期数据所得到的推荐效果更好, 是作者未来的研究方向.

参考文献:

- [1] 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.
- [2] Lv L Y, Medo M, Yeung C H, et al. Recommender systems[J]. Physics Reports, 2012, 519(1): 1-49.
- [3] 陈华, 李仁发, 刘钰峰, 等. 个性化搜索引擎推荐算法研究[J]. 计算机应用研究, 2010, 27(1): 48-50.
- [4] 石珂瑞, 刘建国. 二阶有向相似性对协同过滤算法的影响[J]. 上海理工大学学报, 2014, 36(1): 31-33.
- [5] 李霞, 李守伟. 面向个性化推荐系统的二分网络协同过滤算法研究[J]. 计算机应用研究, 2013, 30(7): 1946-1949.
- [6] Ricci F, Nguyen Q N. Acquiring and revising preferences in a critique-based mobile recommender system [J]. IEEE, Intelligent Systems, 2007, 22(3): 22-29.
- [7] Zhang Y C, Blattner M, Yu Y K. Heat conduction process on community networks as a recommendation model [J]. Physical Review Letters, 2007, 99(15): 154301.
- [8] Guo Q, Leng R, Shi K, et al. Heat conduction information filtering via local information of bipartite networks[J]. The European Physical Journal B, 2012, 85(8): 1-8.
- [9] Liu J G, Zhou T, Guo Q. Information filtering via biased heat conduction[J]. Physical Review E, 2011, 84(3): 037101.
- [10] Zhou T, Ren J, Medo M, et al. Bipartite network projection and personal recommendation[J]. Physical Review E, 2007, 76(4): 046115.
- [11] Liu J G, Zhou T, Wang B H, et al. Effects of user's tastes on personalized recommendation [J]. International Journal of Modern Physics C, 2009, 20(12): 1925-1932.
- [12] 张子柯. 社会化标签系统的结构、演化和功能[J]. 上海理工大学学报, 2011, 33(5): 444-451.

- [13] Zhou T, Kuscik Z, Liu J G, et al. Solving the apparent diversity-accuracy dilemma of recommender systems [J]. Proceedings of the National Academy of Sciences of the United States of America, 2010, 107 (10): 4511 – 4515.
- [14] Zeng A, Yeung C H, Shang M S, et al. The reinforcing influence of recommendations on global diversification [J]. Europhysics Letters, 2012, 97(1): 18005.
- [15] Liu J, Deng G S. Link prediction in a user-object network based on time-weighted resource allocation [J]. Physica A: Statistical Mechanics and its Applications, 2009, 388(17): 3643 – 3650.
- [16] Zhang Q M, Zeng A, Shang M S. Extracting the information backbone in online system [J]. PloS One, 2013, 8(5): e62624.
- [17] 朱郁筱, 吕琳媛. 推荐系统评价指标综述 [J]. 电子科技大学学报, 2012, 41(2): 163 – 175.
- (编辑: 石 瑛)

(上接第 326 页)

- [3] Chow S N, Mallet-Paret J, Shen W X. Traveling waves in lattice dynamical systems [J]. Journal of Differential Equations, 1998, 149(2): 248 – 291.
- [4] Keener J P. Propagation and its failure in coupled systems of discrete excitable cells [J]. SIAM Journal on Applied Mathematics, 1987, 47(3): 556 – 572.
- [5] 裴胜兵, 张卫国, 李想. 色散项系数为负的 MKdV-Burgers 方程的有界行波解 [J]. 上海理工大学学报, 2014, 36(3): 205 – 216.
- [6] 徐丽筱, 张天四, 黄晓鑫. 一类具有饱和发生率的随机 SIRS 模型全局正解的渐近行为 [J]. 上海理工大学学报, 2013, 35(6): 541 – 546.
- [7] Carr J, Chmaj A. Uniqueness of traveling waves for nonlocal monostable equations [J]. Proceedings of the American Mathematical Society, 2004, 132 (8): 2433 – 2439.
- [8] Guo J S, Wu C H. Existence and uniqueness of traveling waves for a monostable 2 – D lattice dynamical system [J]. Osaka Journal of Mathematics, 2008, 45(2): 327 – 346.
- [9] Pan S X. Asymptotic behavior of traveling fronts of the delayed Fisher equation [J]. Nonlinear Analysis Real World Applications, 2009, 10(2): 1173 – 1182.
- [10] Wang Z C, Li W T, Ruan S G. Traveling fronts in monostable equations with nonlocal delayed effects [J]. Journal of Dynamics and Differential Equations, 2008, 20(3): 563 – 607.
- [11] Yu Z X. Uniqueness of critical traveling waves for nonlocal lattice equations with delays [J]. Proceedings of the American Mathematical Society, 2012, 140(11): 3853 – 3859.
- [12] Yu Z X, Yuan R. Existence and asymptotics of traveling waves for nonlocal diffusion systems [J]. Chaos, Solitons & Fractals, 2012, 45(11): 1361 – 1367.
- [13] Yu Z X, Mei M. Asymptotics and uniqueness of travelling waves for non-monotone delayed systems on 2D lattices [J]. Canadian Mathematics Bulletin, 2013, 56(3): 659 – 672.
- [14] Widder D V. The laplace tranform [M]. Princeton: Princeton University Press, 1941.
- (编辑: 石 瑛)