

文章编号:1007-6735(2014)01-0031-03

二阶有向相似性对协同过滤算法的影响

石珂瑞, 刘建国

(上海理工大学 复杂系统科学研究中心, 上海 200093)

摘要: 考虑用户的二阶相似性信息,提出了一种改进的协同过滤个性化推荐算法. 实证统计发现,经典的基于产品映射的用户相似性定义中包含很多流行产品的信息,因此,无法准确度量用户的兴趣关联,通过引入有向的二阶相似性,算法可以有效降低大众主流喜好对目标用户相似性定义的影响. Movielens 数据集上的实验结果显示,算法的准确度可以达到0.080 8,相对于经典的协同过滤算法,其准确性提高了 22.08%,且当推荐列表长度 $L=50$ 时,推荐列表的多样性可以达到 0.775,较经典的协同过滤算法提高了 10.87%. 研究表明,二阶有向相似性信息对个性化推荐算法有很大影响.

关键词: 系统分析与集成; 个性化推荐; 协同过滤算法; 二阶有向相似性

中图分类号: TP 274.2 **文献标志码:** A

Effect of Directive Second-order Similarity on Collaborative Filtering Recommender

SHI Ke-rui, LIU Jian-Guo

(Research Centre of Complex Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: An improved collaborative filtering(CF) recommendation algorithm based on the second-order similarity between users was introduced. According to the empirical statistics,it was found that the definition of classical item-projection-based similarity contains so plenty of information of popular items that the users' interests or preferences are hard to be measured. Owing to considering the directive second-order similarity,the new algorithm can effectively depress the influence of mainstream preferences on target user. The numerical results on one benchmark dataset Movielens show that the accuracy can reach 0.080 8, which is improved by 22.08% comparing with the standard CF. Correspondingly, the diversity reaches 0.775 and could be improved by 10.87% in the optimal case when the recommendation list equals to 50. The study indicates that the directive second-order similarity are crucial in recommendation algorithm.

Key words: systems analysis and integration; personalized recommendation; collaborative filtering; directive second-order similarity

协同过滤算法是应用最广泛的一种推荐算法^[1-6]. 它的核心思想是利用朋友的历史选择信息

收稿日期: 2013-04-17
第一作者: 石珂瑞(1986-),女,硕士研究生. 研究方向: 推荐算法. E-mail: shikerui2010@gmail.com
通讯作者: 刘建国(1979-),男,教授. 研究方向: 网络科学、商务智能、知识管理. E-mail: liujg004@ustc.edu.cn

预测目标用户的喜好程度. 由于其具有较好的准确率, 因此, 得到了广泛的应用, 成为电子商务系统中解决信息超载的有效工具之一^[7-8]. 最近, 基于物质扩散和热传导的方法已经被成功地引入到个性化推荐算法的研究中, 用来度量用户之间的相似性行为, 并且取得了不错的效果^[3,9-10]. Liu 等通过改变用户相似性方向, 提出了一种有向相似性的协同过滤算法^[11], 该算法发现, 增加度小用户向度大目标用户推荐的强度, 可以极大地提高算法准确性和推荐列表多样性. 已有的统计结果表明, 用户的二阶相似性可以大幅度提高推荐准确性^[10]. 本文构造二阶有向相似推荐算法, 综合考察用户的有向相似性和二阶信息的影响, 实验的数值结果表明, 利用邻居用户到目标用户的二阶相似性可以为目标用户提供更加准确的推荐结果.

1 经典协同过滤算法

协同过滤算法的思想是: 首先, 计算用户之间的相似性; 其次, 根据用户之间的相似选择行为对目标用户进行推荐. 定义产品集合为 $O = \{o_1, o_2, \dots, o_m\}$, 用户集合为 $U = \{u_1, u_2, \dots, u_n\}$, 那么, 推荐系统可以表示为矩阵 $A = \{a_{ij}\} \in R^{m \times n}$, 其中, $a_{ij} = 1$, 表示用户 u_i 选择了产品 o_j ; 否则, $a_{ij} = 0$. 在经典的协同过滤算法中, 用户 u_i 和用户 u_j 的相似性根据夹角余弦系数或 Pearson 系数进行计算. 受到物质扩散过程的启发, Liu 等^[12]提出了基于用户-产品二部分图的用户 u_i 和 u_j 相似性 S_{ij} 的计算方法.

$$S_{ij} = \frac{1}{k(u_j)} \sum_{a=1}^m \frac{a_{ia} a_{ja}}{k(o_a)} \quad (1)$$

式中, $k(u_j)$ 表示用户 u_j 的相似性; $k(o_a)$ 表示产品 o_a 的度, $k(o_a) = \sum_{i=1}^n a_{ia}$; a_{ia}, a_{ja} 表示产品的标号.

式(1)表示的就是从邻居用户 u_j 到目标用户 u_i 的相似性. 对于用户-产品 (u_i, o_a) 来说, 如果 u_i 没有选择产品 o_a (即 $a_{ia} = 0$), 那么, 目标用户 u_i 对产品 o_a 的预测打分

$$v_{ia} = \sum_{j=1}^m S_{ji} a_{ja} / \sum_{j=1}^m S_{ji} \quad (2)$$

式中, S_{ji} 表示从邻居用户 u_j 到目标用户 u_i 的相似性, 即利用目标用户 u_i 指向邻居用户 u_j 的相似性来预测 u_i 对产品 o_a 的喜好程度.

2 改进的协同过滤算法

在考虑了二阶相似信息的协同过滤算法中, 用户 u_i 和用户 u_j 的二阶相似性可以用如下形式来表示:

$$H_{ij} = S_{ij} + \lambda S_{ij}^2 \quad (3)$$

式中, H 是目标用户 u_i 和其邻居用户 u_j 的二阶相似性矩阵; $S = \{S_{ij}\}$ 是由式(1)所得到的一阶相似性定义; λ 是一个可调参数.

考虑二阶相似性和用户相似性方向, 作者设计一种改进的算法. 对于一个给定的用户 u_i , 改进的协同过滤算法可以描述为: 首先利用式(3)计算用户之间的二阶有向相似性, 进而对目标用户进行推荐.

3 数值实验

使用标准数据集 Movielens 测试新算法的性能表现. 该数据集包括 943 个用户和 1 682 个产品 (电影). 用户会对自己看过的产品按照 5 分制进行打分, 1 分表示最不喜欢, 5 分表示最喜欢. 在此假设用户打分大于 2 分的就看作他选择了该产品, 进而构建用户-产品二部分图. 删除部分连边后数据集剩余 943 个用户, 1 574 个产品和 82 520 条边. 将数据集分为训练集和测试集这两部分. 利用数据集的 90% 作为训练集预测用户对未选择产品的喜好程度, 利用其余 10% 作为测试集度量算法的表现.

3.1 评价指标

a. 准确度. 一个好的算法应该是要将训练集中已知的用户喜欢的产品排在比较靠前的位置, 本文采用平均排序打分度量推荐算法的准确性. 若有 $L_i = 100$ 个产品没有被 u_i 选择, 而 o_a 在推荐列表中排名第 10 个, 即 o_a 的位置是 10/100, 其排序打分记为 $r_{ia} = 0.1$. 平均排序打分 $\langle r \rangle$ 越小, 说明产品在推荐列表中的位置越靠前, 算法的准确性就越高, 反之亦然.

b. 流行性和多样性. 推荐产品的平均度 $\langle k \rangle$ 是评价推荐算法的重要指标. 如果算法趋向于推荐流行的产品, 那么, 被推荐产品的平均度会很高; 反之, 如果算法趋向于推荐不太流行的产品, 则其平均度会很低. 此外, 利用平均海明距离度量推荐列表的多样性. 用户 u_i 和 u_j 推荐列表的平均海明距离被定义为 $p_{ij} = 1 - Q_{ij}/L$. 其中, L 为推荐列表的长度, Q_{ij} 为推荐给用户 u_i 和 u_j 的 2 个推荐列表中相同产品的个数. 推荐列表的多样性定义为 p_{ij} 的平均值 $P, S = p_{ij}$. P 越大, 算法可以提供更多样化的推荐, 反之亦然.

3.2 结果分析

当采用 90% 的数据作为训练集时, 从图 1 中可以看出, 在 Movielens 数据集上, 经典的二阶协同过滤算法的平均排序打分 $\langle r \rangle$ 在 λ 为 -0.815 附近达到最小值 0.083, 而本文算法的 $\langle r \rangle$ 在 λ 为 -0.726 附近可以达到 0.080 8. 图 2 表示推荐产品的平均度和推荐列表的多样性随着参数 λ 的变化趋势. 从图

2 中可以得到,推荐产品的平均度 $\langle k \rangle$ 和 λ 正相关,也就是说降低用户所选择主流产品的影响会给不太流行产品更多的被推荐机会,当推荐列表长度 $L = 50$ 时,在最优点 $\lambda_{\text{opt}} = -0.7266$ 处,产品的平均度 $\langle k \rangle = 191$,相对于经典的基于物质扩散的协同过滤(CF)算法,流行性降低了10.32%.从图2中也可以发现,算法的平均海明距离 P 和 λ 负相关,表明考虑了邻居用户对目标的二阶相似性后使得推荐列表更为离散,并且当推荐列表长度 $L = 50$ 时,在 $\lambda_{\text{opt}} = -0.7266$ 处,产品的平均海明距离 P 大约为0.775,多样性提高了10.87%.因此,这种新算法在增加CF算法的精确性方面具有很大优势,并且也可以较好地帮助目标用户发现一些隐藏的暗信息,为目标用户提供更为多样化的推荐.

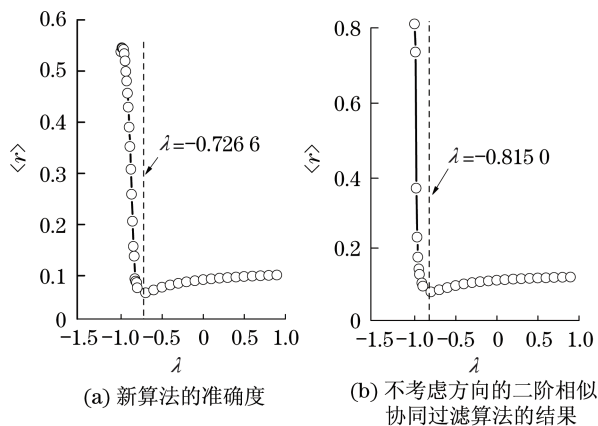


图1 Movielens数据集上平均排序打分随参数 λ 的变化趋势图

Fig.1 Correlation between the parameter λ and the average ranking score $\langle r \rangle$

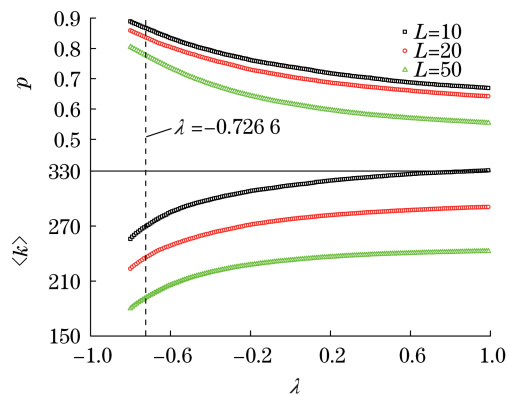


图2 Movielens数据集上推荐产品的平均度 $\langle k \rangle$ 及推荐列表的多样性 P 随参数 λ 的变化趋势

Fig.2 Correlation between the parameter λ and the average object degree $\langle k \rangle$ as well as the diversity P on dataset Movielens

4 结束语

首先提出了二阶有向协同过滤算法,并且研究

了用户的二阶有向相似性对协同过滤算法的影响.在标准数据集 Movielens 上的试验结果表明,利用邻居用户到目标用户的二阶相似信息可以提高推荐的准确性.当训练集的比例为90%时,算法的准确度可以达到0.0808,该结果较经典的CF提高了22.08%,这是已有的协同过滤算法中准确性效果最好的.

除准确性外,被推荐产品的平均度以及推荐列表的多样性也是量化推荐算法性能表现的2个重要指标,而数值结果显示,新算法在这2种指标下比经典的CF算法都要好.作者下一步的工作,将关注在二阶有向相似性下产品的度信息与产品打分的关系.

参考文献:

- [1] Herlocker J L, Konstan J A, Terveen L G, et al. Evaluating collaborative filtering recommender system [J]. ACM Trans Inform Syst, 2004, 22(1): 5-53.
- [2] Konstan J A, Miller B N, Maltz D, et al. GroupLens: applying collaborative filtering to Usenet news [J]. Commun ACM, 1997, 40(3): 77-87.
- [3] Liu J G, Wang B H, Guo Q. Improved collaborative filtering algorithm via information transformation [J]. Int J Mod Phys C, 2009, 20(2): 285-293.
- [4] 宣照国, 苗静, 党延忠. 基于扩展邻居的协同过滤算法 [J]. 情报学报, 2010, 29(3): 443-448.
- [5] 潘红艳, 林鸿飞, 赵晶. 基于矩阵划分和兴趣方差的协同过滤算法 [J]. 情报学报, 2006, 25(1): 49-54.
- [6] 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.
- [7] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems-a survey of the state-of-the art and possible extensions [J]. IEEE Trans Know and Data Eng, 2005, 17(6): 734-749.
- [8] Liu J G, Chen M Z Q, Chen J, et al. Recent advances in personal recommendation systems [J]. Int J Info & Sys Sci, 2009, 5(2): 230-247.
- [9] Zhou T, Ren J, Medo M, et al. Bipartite network projection and personal recommendation [J]. Phys Rev E, 2005, 76(4): 046115, 1-4.
- [10] Liu J G, Zhou T, Che H A, et al. Effects of high-order correlations on personalized recommendations for bipartite networks [J]. Physica A, 2010, 389(4): 881-886.
- [11] Liu J G, Shi K R, Guo Q. Solving the accuracy-diversity dilemma via directed random walks [J]. Phys Rev E, 2012, 85(1): 016118, 1-4.
- [12] Liu J G, Guo Q, Zhang Y C. Information filtering via weighted heat conduction algorithm [J]. Physica A, 2011, 390(12): 2414-2420.

(编辑:石 瑛)