

A Research Summary of Collaborative Filtering Recommendation Algorithm Based on E-Commerce

Liqun Yu, Fudong Wang

The Glorious Sun School of Business and Management, Donghua University, Shanghai
E-mail: Markylq@163.com

Received: Dec. 20th, 2018; accepted: Jan. 2nd, 2019; published: Jan. 9th, 2019

Abstract

The collaborative filtering recommendation algorithm is the most abundant and most used recommendation algorithm in the current e-commerce recommendation system. This paper first introduces three different collaborative filtering recommendation algorithms in detail and compares their respective advantages and disadvantages. Then, the current research status of data sparseness, cold start and scalability of collaborative filtering recommendation algorithm is summarized. And this paper analyzes the merits and deficiencies of the existing methods. Finally, the research hotspot of future collaborative filtering algorithm is proposed, which provides a reference for the future development of collaborative filtering algorithm.

Keywords

Collaborative Filtering Recommendation Algorithm, E-Commerce, Personalized Recommendation

基于电子商务的协同过滤推荐算法综述

俞立群, 王扶东

东华大学旭日工商管理学院, 上海
E-mail: Markylq@163.com

收稿日期: 2018年12月20日; 录用日期: 2019年1月2日; 发布日期: 2019年1月9日

摘要

协同过滤推荐算法是目前电子商务推荐系统中应用场景最丰富也是使用最多的推荐算法。本文首先详细

介绍了三种不同的协同过滤推荐算法并比较了各自的优势与不足, 然后对协同过滤推荐算法目前存在的数据稀疏性、冷启动和可拓展性问题现有的研究现状进行了总结, 并分析了现有的方法的可取与不足之处。最后提出了未来协同过滤算法的研究热点, 为协同过滤算法未来发展提供参考。

关键词

协同过滤算法, 电子商务, 个性化推荐

Copyright © 2019 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着电子商务和因特网技术的快速发展, 人们对于信息的需求进入到信息过载的时代。在面对纷繁复杂的信息时, 人们无法迅速找到自己所需要的信息, 而推荐系统就能很好地解决这个问题。它首先搜集用户信息, 进而分析用户偏好, 最后向用户进行个性化推荐。而推荐算法则是推荐系统最重要的一个部分, 各种推荐算法中协同过滤算法因其高效准确被广泛应用。

2. 协同过滤推荐算法

基于协同过滤的推荐是依据与目标用户相似度高的用户的喜好向其推荐项目。协同过滤算法基于以下假设[1]: 对于项目评分相似度高的用户其兴趣或偏好也是相似的即对同一个项目的喜好程度是一致的。Tapestry [2]是最早包含协同过滤算法思想的推荐系统, 该系统主要用于电子邮件的分类筛选, 解决 Xerox 公司资讯过载问题。但是需要用户手动输入一些信息系统进行查询, 然后再给出推荐结果, 所以是非智能自动推荐。1994 年明尼苏达大学创建的“GroupLens”自动化新闻推荐系统标志着推荐系统开始独立成为一个新的研究内容[3], 该系统建立了一个形式化的推荐模型, 为今后推荐系统的研究奠定了基础。基于协同过滤的推荐又可以分为两类[4]: 基于记忆的推荐(Memory-based Recommendation)和基于模型的推荐(Model-based Recommendation)。基于记忆的协同过滤算法又可分为基于用户的推荐(User-based Recommendation)和基于项目的推荐(Item-based Recommendation)。

2.1. 基于用户的协同过滤推荐算法

GroupLens 系统[3]所使用的就是基于用户的协同过滤算法。它的基本假设就是“对于项目评分相似度高的用户其兴趣或偏好也是相似的即对同一个项目的喜好程度是一致的”。该算法让用户根据自己的兴趣对购买过的项目进行打分, 然后分析比较该用户与其他用户的评分数据, 进而计算他们评分之间的相似度, 相似度越高表明他们偏好也就越相似。最后利用找到的与目标用户相似的用户评分来预测目标用户对其未评分项目的喜好程度完成推荐。皮尔森系数与余弦相似度的计算公式分别如下:

$$\rho_{x,y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}}$$

$$\cos \theta = \frac{\sum_n^1 (A_i \times B_i)}{\sqrt{\sum_n^1 (A_i^2)} \times \sqrt{\sum_n^1 (B_i^2)}}$$

2.2. 基于项目的协同过滤推荐算法

随着用户与项目的不断增多, 评分矩阵稀疏性加大, 计算用户之间相似性的时间显著增长以及困难性加大。Sarwar 等[5]人提出了基于项目的协同过滤推荐算法。它的基本假设就是“用户可能会对与他之前偏爱的项目相似的项目感兴趣”, 简单来说就是用户收藏购买或者关注了某项目后, 就将与该项目相似的其他项目推荐给该用户。基于项目的协同过滤推荐算法第一步是计算项目的相似性, 在评分矩阵中, 将每个项目看成一个列向量, 计算某个项目和其他所有项目列向量与它的相似度, 再进行预测推荐。

2.3. 基于模型的协同过滤推荐算法

基于模型的协同过滤推荐算法就是借助机器学习等技术利用训练集数据训练出一个计算模型, 然后利用训练出的模型和目标用户对于部分项目的评分数据, 预测目标用户对于其尚未打分的项目的评分。基于模型的协同过滤算法主要有基于关联规则的协同过滤算法、基于逻辑回归的协同过滤算法、基于神经网络的协同过滤算法等[6]。

2.4. 三种协同过滤推荐算法比较

下表 1 从优点、缺点及应用场景三个方面对以上三种协同过滤推荐算法进行了比较, 比较结果如图所示

Table 1. Comparison of advantages and disadvantages of three collaborative filtering algorithms and application scenarios

表 1. 三种协同过滤算法优缺点及应用场景比较

	基于用户的协同过滤算法	基于项目的协同过滤算法	基于模型的协同过滤算法
优点	准确率较高、隐式特征的挖掘	计算简单、实时响应、可解释性好	线下计算、可实现快速实时响应
缺点	可拓展性差、新用户的冷启动问题	未考虑用户之间的差别、数据稀疏性、冷启动问题	模型增量训练问题
应用场景	用户远多于项目的社交网络网站	项目远多于用户的非社交网站	大量计算, 实时响应

3. 协同过滤算法存在的问题及解决方法

协同过滤推荐技术在个性化推荐系统中取得了巨大的成功, 并得到了广泛的应用。然而, 互联网及电子商务规模的急速扩张、推荐内容形式的更加多样、用户及项目的数量急剧上升以及数据的爆炸式增长, 给推荐系统带来了严重的数据稀疏性、冷启动及拓展性问题。

3.1. 数据稀疏性问题

协同过滤推荐算法中用户相似度计算是依据用户对项目的打分, 所以用户 - 项目矩阵评分数据越丰富, 用户相似度计算也就越准确, 最终的推荐效果就会越好。然而在电子商务领域, 一方面用户的精力或购买的项目总是有限的, 另一方面如今项目的数量远远大于用户的数量, 这就导致用户不可能对所有项目都有评价, 出现大量的项目没有被用户所评分, 而且这类情况往往占有绝大部分比例, 相关研究指出这个比例在 1% 以下[7]。绝大多数评分数据的缺失就使得用户-项目评分矩阵成为了一个极其稀疏的矩阵, 导致无法计算用户相似度和项目相似度或计算得到的结果准确度大幅下降, 导致系统的推荐质量急剧下降。在这种情况下影响最终推荐效果的关键问题就是数据稀疏性问题。

针对数据稀疏性带来的推荐质量下降的影响, 学者们提出了许多解决方法来缓解数据稀疏性问题。大致可以分为两类: 一是对用户 - 项目评分矩阵缺失部分进行填充, 另一类是对用户 - 评分矩阵降维。

填充法主要有以下两种: 一是直接进行填充。直接填充是指填充一个固定的缺省值或是其他用户对

于此项目评分的平均值或众数值。直接填充法简单快速,但是准确率很低。因为每个用户对于每个项目的评分都是不一样的且相对独立的,所以现在一般不会采用直接填充法;二是预测填充。预测填充是指利用机器学习等技术对于用户对各项目的打分情况进行预测。主要有贝叶斯分类方法[8]或者BP神经网络方法[9]。该方法对噪声数据有较强的抗干扰能力,但是机器学习模型训练时间长且可能存在过拟合或欠拟合问题。

在评分矩阵稀疏度不变的情况下将评分矩阵分解到更低的维度,让评分矩阵不再稀疏。矩阵分解技术主要包括奇异值分解技术[10](SVD)、隐语义模型[11](Funk_SVD)和SVD++算法[12]等。矩阵分解技术保留了用户在特征项目上的评分数据,使得用户-项目矩阵的评分密度有所上升。但是在进行矩阵分解的同时也会丢掉部分信息导致推荐质量不够高。

3.2. 冷启动问题

协同过滤推荐算法的冷启动问题可分为两种,一是新用户的冷启动问题,一是新项目的冷启动问题。针对冷启动问题,一般可以直接推荐最热门的项目或是让用户在一开始的时候填一些基本信息,但用户对一些隐私问题会有所保留,所以最终的推荐效果不理想。有学者[13]就提出引入社会标签来解决冷启动问题,该标签可以代表用户根据个人兴趣对项目的分类,该方法不需要用户备什么特殊的技能,也不需要用户过分特意参与到系统中,因此该方法在社区境中取得了广泛的应用。但是采取固定标签选择的方式,有可能所给范围太小,而如果采取用户自定义的方式,可能会有模糊不清的表述。针对新用户问题李聪[14]等提出了一种基于n序访问解析逻辑的冷启动消除方法,该方法主要通过解析用户WEB访问顺序来计算用户之间的相似度。针对新项目问题郭晓波[15]等提出一种混合推荐算法,提出一种基于相似度的关联推荐模式,实现一种新的结合关联规则推荐与协同过滤推荐方法。于洪[16]等结合用户、项目属性、用户评价时间信息和标签提出解决新项目冷启动问题的个性化推荐算法。

3.3. 可拓展性问题

随着电子商务规模急速扩张及用户数不断增多,系统所需的计算量大大增加,推荐实时性也会越来越差,而响应速度又极大地影响了用户体验,因此解决协同过滤推荐算法的可拓展性问题也是十分重要。

针对可拓展性问题,最简单的办法是提高硬件设施的效果,但是无法从根本上解决这个问题。有学者[10]提出利用奇异值分解技术降维,从而提高算法的可拓展性问题。但该方法会丢掉部分信息导致推荐质量不够高。此外,还有学者[17]提出利用聚类的方法提高可拓展性,聚类的方法包括对用户的聚类、对项目的聚类、同时对用户和项目进行聚类。此外,有学者提出基于模型的协同过滤推荐算法,在离线状态下对于模别进行建模和预估,生成用户的偏好模型,在线时使用建好的偏好模型根据用户实时喜好信息进行预测,为用户进行推荐,以最快响应用户需求。这种方式虽然可以大大缓解算法的可拓展性问题,但是用户偏好模型建立过程要求高,并且根据用户的兴趣变化修改训练得到的模型难度大,即增量训练有困难。

4. 总结与展望

本文首先详细介绍了三种不同的协同过滤推荐算法并比较了各自的优势与不足,然后对协同过滤推荐算法目前存在的数据稀疏性、冷启动和可拓展性问题现有的研究现状进行了总结,并分析了现有的方法的可取与不足之处。随着因特网及电子商务的不断发展,协同过滤推荐算法也不断有新的技术提出,未来有以下几个方面的研究热点:一是借助大数据技术实现分布式计算;二是结合社会化网络信息进行推荐,三是综合评分以及文本评论信息进行推荐。总之,作为一个充满挑战又发展迅速的新兴科研领域,协同过滤推荐算法必将迎来更好地发展。

参考文献

- [1] Herlocker, J., Konstan, J. and Riedl, J. (2000) Explaining Collaborative Filtering Recommendations. *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, ACM Press, New York, 241-250. <https://doi.org/10.1145/358916.358995>
- [2] Goldberg, D. and Nicols. D. (1992) Using Collaborative Filtering to Weave an Information Tapstry, *Communications of the ACM*, **35**, 61-70. <https://doi.org/10.1145/138859.138867>
- [3] Resnick, P., et al. (1994) GroupLens: An Open Architecture for Collaborative Filtering of Netnews. *Proceedings of ACM Conference on Computer Supported Cooperative Work, CSCW 1994*, Chapel Hill, 175-186. <https://doi.org/10.1145/192844.192905>
- [4] Breese, J.S., Heckerman, D. and Kadie, C. (1998) Empirical Analysis of Predictive Algorithms for Collaborative Filtering. *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, San Francisco, 24-26 July 1998, 43-52.
- [5] Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. (2001) Item-Based Collaborative Filtering Recommendation Algorithms. *Proceeding of the 10th international conference on World Wide Web*, Hong Kong, 1-5 May 2001.
- [6] 冷亚军, 陆青, 梁昌勇. 协同过滤推荐技术综述[J]. 模式识别与人工智能, 2014, 27(8): 720-734.
- [7] 孙小华. 协同过滤推荐系统的稀疏性与冷启动问题研究[D]: [博士学位论文]. 杭州: 浙江大学, 2006.
- [8] 焦明海, 陈晓芳, 陈旭, 等. 基于贝叶斯网络认知反馈的协同过滤推荐[J]. 控制工程, 2017, 24(7): 1310-1317.
- [9] 周朴雄, 张兵荣, 赵龙文. 基于 BP 神经网络的情境化信息推荐服务研究[J]. 情报科学, 2016, 36(3): 71-75.
- [10] 王建芳, 李骁, 武文琪, 等. 一种 SVD 和信任因子相结合的协同过滤算法[J]. 小型微型计算机系统, 2017, 38(6): 1290-1293.
- [11] Funk, S. (2006) Netflix Update: Try This at Home[EB/OL]. <http://sifter.org/~simon/journal/20061211.html>, 2017-9-15
- [12] Koren, Y. (2008) Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model. *Proceedings of International Conference on Knowledge Discovery and Data Mining*, 426-434.
- [13] Kim, H.N., Alkhaldi, A., Saddik, A.E., et al. (2011) Collaborative User Modeling with User-Generated Tags for Social Recommender Systems. *Expert Systems with Applications*, **38**, 8488-8496. <https://doi.org/10.1016/j.eswa.2011.01.048>
- [14] 李聪, 梁昌勇. 基于 n 序访问解析逻辑的协同过滤冷启动消除方法[J]. 系统工程理论与实践, 2012(7): 1537-1545.
- [15] 郭晓波, 赵书良, 牛东攀. 一种解决稀疏数据和冷启动问题的组合推荐方法[J]. 中国科学技术大学学报, 2015, 45(10): 804-812.
- [16] 于洪, 李俊华. 一种解决新项目冷启动问题的推荐算法[J]. 软件学报, 2015, 26(6): 1395-1408.
- [17] 吴湖, 王永吉, 王哲, 等. 两阶段联合聚类协同过滤算法[J]. 软件学报, 2010, 21(S): 1042-1054.

知网检索的两种方式:

1. 打开知网页面 <http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD>
下拉列表框选择: [ISSN], 输入期刊 ISSN: 2168-5843, 即可查询
2. 打开知网首页 <http://cnki.net/>
左侧“国际文献总库”进入, 输入文章标题, 即可查询

投稿请点击: <http://www.hanspub.org/Submission.aspx>

期刊邮箱: ecl@hanspub.org