电子商务应用中的协同过滤算法综述

袁 洁

（山东科技大学 计算机科学与工程学院，山东 青岛 266590）

摘　要：电子商务的迅猛发展为用户提供了大量的信息，网购已经成为潮流，各种网购平台为用户提供了大量的信息，而如何在众多的电子商务网站和海量的商品中快速地找到用户需求的产品成为一个研究的重点。在此，推荐算法应运而生，，协同过滤推荐算法在电子商务系统中得到了广泛的应用。本文主要介绍两种协同过滤推荐技术在电子商务网站中的应用现状，并在此基础上介绍了一些改进的协同过滤算法的研究现状和推荐效果，以及算法未来可能的研究方向。

关键词：协同过滤；电子商务；稀疏性；相似性

1 概述

互联网的快速发展为电子商务网站提供了平台，依据中国互联网络信息中心发布的第38次《中国互联网络发展状况统计报告》可知截止到2016年6月，我国的网民规模进一步扩大，网民人数已经达到了7.1亿，网民规模较去年持续增长，其中网购用户的规模达到4.48亿。个性化推荐[1]逐渐成为电子商务中的一个重要研究内容。目前，几乎所有大型的电子商务系统，如当当、京东、Amazon、淘宝等都应用了各种形式的个性化推荐系统。

2 协同过滤推荐算法

协同过滤推荐以其应用简单并且推荐效果好的优势，在电子商务推荐系统中得到了广泛的应用。主要的协同过滤推荐算法有两种：基于用户的和基于项目的协同过滤。

2.1 基于用户(user-based)的协同过滤。在推荐系统中，基于用户的协同过滤推荐算法是最早应用的，其基本思想[2]是：针对用户的历史行为数据发现用户对商品或内容(如商品购买，收藏，内容评论或分享)的喜欢程度，根据不同用户对同一商品偏好程度计算用户之间的关系。并根据用户对不同商品的打分数据来计算出用户对其他产品的喜好程度，对有相同喜好的用户进行同类商品的推荐。基于项目的协同过滤推荐算法主要应用于社交网站中。

2.2 基于项目(item-based)的协同过滤。基于项目的协同过滤推荐算法基本思想[3]是：计算目标用户对商品的喜好程度，运用相似度算法，计算出各商品之间的相似度，得到与目标商品相似度高的项目的集合，将相似度高的商品作为目标商品的最近邻居，依据用户对所有商品的评分来对项目集合中的商品进行排名，并将商品推荐给用户。此算法主要应用于电子商务网站中。

2.3 相似性计算。电子商务领域中的协同过滤推荐算法的关键在于算法能够准确确定出目标用户的最近邻居，而确定最近邻居前提是先计算出各用户之间的相似性，目前研究者们研究使用的相似度计算主要有：余弦相似度，Jaccard相关度，Cosine相似度，wb-cos，基于云模型的相似度，修正的余弦相似度和相关相似度。其中，在电子商务中应用较广泛的是Jaccard相关度，wb-cos和相关相似度。

3 改进的协同过滤研究进展

协同过滤推荐算法一直是电子商务领域[4]的研究热点，到目前为止，协同过滤推荐算法的研究内容主要集中在以下三个主要的方面：用户相似度计算、用户信任计算和用户偏好计算。一些学者提出利用云模型、基于标签的大众标注系统协同推荐算法等来改善用户相似度的计算。Massa[4]、张中峰等分别提出的基于信任的和基于信任传播的协同过滤推荐方法来提高推荐系统的信任度，并提高了推荐的服务的质量和效率。严冬梅、秦光洁等分别提出了基于用户兴趣度和特征的优化和基于综合兴趣度的协同过滤推荐算法来优化用户偏好存在的问题，并分别采用了贝叶斯算法获取用户需求和综合用户的显性和隐形偏好对目标用户进行准确的定位和推荐。改进的协同过滤推荐算法使得电子商务的推荐系统更加高效、准确，但是随着数据量的增加，电子商务商户和用户的不断更新，系统的负荷也在不断加重，研究者们需要不断的针对当前的情形进行推荐算法的改进和推荐系统的扩展。

4 存在的问题和挑战

随着网络的发展，个性化推荐的普及所带来的内容复杂程度和用户人数等的不断增加，数据量的增加使得数据处理成为个性化推荐算法中比较棘手的难题，数据并行处理技术将是未来发展研究的重点，另外，数据更新存储技术的改进也是未来电子商务领域的研究重点。电子商务推荐系统的协同过滤算法技术仍需要我们不断的改进与扩展。

4.1 数据稀疏性问题。在电子商务推荐系统中，大量的商品呈现，这使得不同用户对于同一商品的选择的重叠性较小，购买相同商品的用户数量的稀疏导致了用户评分矩阵的稀疏，这就使得电子商务网站商品的推荐精度受到严重的限制，大多数的推荐算法在实际中的应用效果都没有想象中的好，为此，许多学者提出的改进算法都是为了缓和矩阵稀疏性问题，但目前，数据的稀疏性问题任是推荐系统面临的一个重要的问题。

4.2 冷启动问题。目前许多学者针对冷启动进行研究，其中一些研究者采用众数法和信息熵法。这两种方法能够帮助解决冷启动问题，但是，冷启动问题依然不能得到满意的效果，在这一问题上，依然是未来研究者需要重点解决的问题。

4.3 大数据处理的问题。近年来，电子商务系统的用户和商品不断的更新，并且不同的用户和不同的商品之间的交互也在不断的改变，系统的负载不断加重，算法运行效率受到限制，而近年来也有些学者提出了新的解决方案，并行化处理成为解决问题的突破口。有研究者在Hadoop平台设计分布式并行的的协同过滤推荐算法[5]，通过并行处理数据有效的提高了推荐系统的响应时间，为解决大数据问题提供了较好的解决方案，但是当系统提供的数据较稀疏的时，推荐算法的精度将受到严重影响。分布式并行计算的研究还刚刚起步，所以，分布式并行计算将会是电子商务网站推荐系统中算法研究的新的方向。

4.4 推荐效果评估。电子商务推荐系统虽然得到了广泛的应用，但是如何有效的评价推荐系统的效果没有较好的评估策略，推荐系统算法的评价指标主要是推荐算法的预测准确度和分类准确度，由于推荐系统的信息复杂性，使得推荐系统的推荐评估成为一个难题，所以推荐系统的效果评估仍然是研究领域面临的一个严峻的问题，也是学者们亟待解决的问题。

5 结论

本文通过介绍两种主要的协同过滤推荐算法的基本原理和算法的执行步骤，以及基于这两种算法的改进算法的特点，很多研究者提出的改进算法都是用来解决系统的稀疏性和冷启动问题，并且在一定程度上缓解了这些问题。可见稀疏性和冷启动问题对于推荐系统的精度影响尤为重要。另外，一个系统的推荐性能是否良好，也需要一定的策略来评估。这些方面也是继续改进的研究方向。

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] 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.

[3] 贺智明，王海超，高娟.电子商务协作过滤推荐技术的算法研究与改进[J].微型机与应用，2009，28(11):60-62.

[4] Massa P, Bhattacharjee B. Using trust in recommender system: An experimental analysis [J]. Proceedings of the 2nd Int'l Conf. on Trust Management, 2004,(6):221-235.

[5] 肖强，朱庆华，郑华，吴克文. Hadoop环境下的分布式协同过滤算法设计与实现[J].现代图书情报技术,2013,(1):83-89.