协同过滤推荐研究综述*

张 瑶，陈维斌，傅顺开
(华侨大学 计算机科学与技术学院，福建 厦门 361000)

摘 要：推荐技术是目前在很多领域中广泛使用的技术之一。而协同过滤推荐算法是应用在推荐技术中很成功的算法。主要介绍了协同过滤推荐技术，总结了当前推荐算法的传统方法、改进算法以及性能评测方法。同时，分析了协同过滤推荐算法中的问题以及相应的解决办法。最后阐述了协同过滤推荐系统中仍需解决的问题和未来可能的发展方向。

关键词：推荐系统；协同过滤推荐算法；稀疏性；冷启动；性能评测
中图分类号：TP311 **文献标识码**：A **文章编号**：1674-7720(2013)06-0004-03

Collaborative filtering recommendation research

Zhang Yao, Chen Weibin, Fu Shunkai
(College of Computer Science and Technology, Huaqiao University, Xiamen 361000, China)

Abstract: Recommendation technology has been one widely applied technology in many fields, and collaborative filtering recommendation algorithm is one of the most recognized progress in this community due to its non-content-based nature. In this article, we give a comprehensive review of the research work on collaborative filtering based recommendation, including theory basis, evaluation methods, successful and representative applications and open issues. It can be a reference for academic researchers as well as applicants.

Key words: recommender system; collaborative filtering recommendation algorithm; sparsity; cold-start; performance evaluation

每天，人们都要面对很多的选择，通常会以周围人的意见作出选择。然而，面对海量的网络资源时，要从中作出最适合的选择就变得非常困难。电子商务系统作为目前典型的成功互联网应用，提供了更丰富的物品。为帮助大家更快速准确地定位自己的需求，推荐系统在电子商务系统中应运而生，例如知名的亚马逊等平台[1]。

目前，应用于推荐系统的算法主要分三类：基于内容的过滤推荐算法、协同过滤推荐算法及混合推荐算法。

基于内容的过滤推荐算法[2]是对用户的兴趣进行分析，构成用户配置文件，并将其和文件集中的文件用共同的特征变量表示。最后比较两者的相似度来为用户进行推荐。随后，通过用户的反馈信息，不断更新用户配置文件，以此来动态地为用户推荐感兴趣的信息[3]。

协同过滤推荐算法是通过用户对项目的评分数据，找到与目标用户或项目相似的对象作为候选推荐。当前主要的协同过滤推荐算法有两种：基于用户(user-based)和基于项目(item-based)的协同过滤推荐算法。基于用户的协同过滤推荐算法认为，如果用户对一些项目的评分比较相似，那么他们对其他项目的评分也比较相似；基于项目的协同过滤推荐算法认为，项目间的评分具有相似性，可以通过用户对目标项目的若干相似项目的评分来估计该项目的分值[4-5]。

混合推荐算法[6]是将基于内容的过滤和协同过滤相结合的方法，既保留了用户配置文件来代表用户的兴趣，同时又根据该配置文件来寻找相似的用户，两种方法互补完成推荐。

本文主要介绍协同过滤推荐算法，因为目前它的应用最为普遍。

1 协同过滤推荐技术

协同过滤推荐技术之所以得到广泛的应用，主要得益于它独立于被推荐对象的内容，只依靠用户对项目的评分或是喜好就可以为用户推荐。同时，还可以帮助用户发掘潜在的兴趣。这些优点使得它几乎适用于所有领域。

1.1 当前的协同过滤推荐算法

当前最常用的协同过滤推荐算法是基于用户和基于项目的算

*基金项目：中央高校基本科研业务费专项基金项目(11JJ0263)；华侨大学引进人才科研启动费项目(11Y0274)；福建省重大科技创新平台建设资助项目(2012H2002)；华侨大学科研基金资助项目(12HJY18)

法。基于用户的协同过滤技术[7]首先获取用户对于项目的评分，然后通过用户间的相似性寻找目标用户的最近邻居，最后利用预测函数和邻居用户的评分来完成推荐；基于项目的协同过滤技术[5,8]则根据项目间相似性得到目标项目的相似项目集合，然后通过预测函数和相似项目集合来产生推荐列表。

下文将详细地介绍和讨论以上两种算法及其一些改进算法。

1.2 基于用户的协同过滤推荐算法

基于用户的协同过滤算法是早期较传统的一种自动协同过滤技术。它主要依靠系统中的用户评分矩阵来计算用户间的相似性，再根据相似性来计算目标用户对于项目的预测评分。

定义 用户评分矩阵

一个 $m\times n$ 的矩阵。m 和 n 分别代表矩阵中的用户数和项目数。矩阵中的元素代表用户对项目的评分或偏好。

1.2.1 相似性计算方法

协同过滤推荐算法本质上是基于近邻的搜索算法，其核心是计算用户/项目间的相似性。常用的相似性度量方法主要有三种[5,8]：余弦相似性度量方法、皮尔森相关相似度量方法、修正的余弦相似性度量方法。

余弦相似性度量方法：该方法把用户对项目的评分看作是一个 n 维向量，则用户间的相似性就由两个用户向量间的余弦夹角来决定。

修正的余弦相似性度量方法：该方法是对余弦相似性度量方法的改进算法。它通过减去用户对于项目的平均分来弥补不同用户对项目的评分尺度不同的问题。

皮尔森相关相似性度量方法：在该方法中，用户间的相似性的计算如下：

$$\mathrm{sim}(i,j)=\frac{\sum_{C\in I_{ij}}(R_{i,c}-\overline{R}_i)(R_{j,c}-\overline{R}_j)}{\sqrt{\sum_{C\in I_{ij}}(R_{i,c}-\overline{R}_i)^2}\sqrt{\sum_{C\in I_{ij}}(R_{j,c}-\overline{R}_j)^2}} \tag{1}$$

其中，I_{ij} 表示用户 i 和 j 共同评分过的项目集合，$R_{i,c}$、$R_{j,c}$ 表示用户 i、j 对于项目 c 的评分，$\overline{R}_i$、$\overline{R}_j$ 表示用户 i、j 对于所有项目的平均评分。

1.2.2 评分预测

预测评分可由相似度作为权重，通过计算邻居用户对未评分项目的加权平均值来得出。

该方法的不足之处在于：(1)如果用户评分的项目较少，推荐效果会比较差。对于一个新加入系统的用户，这个方法完全无法“工作”。(2)为了反应用户最新状态，需要在线计算详尽用户，这可能导致在线响应速度低下。

1.3 基于项目的协同过滤推荐算法

基于项目的协同过滤算法[5]依赖的是项目间的相似性，这对更新的实时性要求较低，因此也是目前应用较多的一种算法。该算法与基于用户的协同过滤算法在过程上类似，只是在相似度上稍有不同。

在相似性度量方法上，1.2.1 中提到的方法仍适用，只是此时计算相似度的对象不同。除此之外，针对项目间相似性的度量，还可使用条件概率度量方法[9]。之所以可以用它来计算相似度是因为，如果选择一个项目的用户中也有很多用户选择另一个项目，则两个项目相似。

1.4 协同过滤推荐改进算法

1.4.1 基于项目评分预测的协同过滤推荐算法

参考文献[8]提出了一种对未评分项目预测评分的改进算法。该算法先找出用户 i 和 j 分别评分的项目集合的并集 I_{ij}，然后计算用户 i 和 j 对于 I_{ij} 中各自未评分项目的评分，最后利用基于用户的协同过滤算法来计算用户 i 和 j 的相似度，并产生相应的推荐列表。

该算法在一定程度上解决了稀疏性的问题，使得两个用户共同评分的项目有所增加，同时在计算两个用户的相似性时，也不会出现用户未评分项目为 0 的状况，提高了推荐精度。但是，由于采用的是传统的相似度量方法，算法的冷启动问题仍未得到根本的解决。

1.4.2 基于项目聚类的协同过滤推荐算法

参考文献[4]是将项目以 K 均值方法进行聚类，然后计算目标项目与聚类中心的相似性，并选择相似度大于 e 的聚类来作为搜索邻居的项目空间，最终完成推荐。

该算法选择若干个聚类来搜索项目的邻居，可以减少搜索空间，提高搜索效率，但相似性阈值 e 的选取需进行权衡。用户需要不断调整 e 来进行聚类数量的控制，以此来调整精度。

1.4.3 基于项目类别相似性的协同过滤推荐算法[10]

参考文献[10]同时考虑评分和类别相似性，并将类别表达成一颗类别树(如图 1 所示)来计算类别相似性，然后通过权重系数 α 将两者进行加权组合，得出综合相似性。

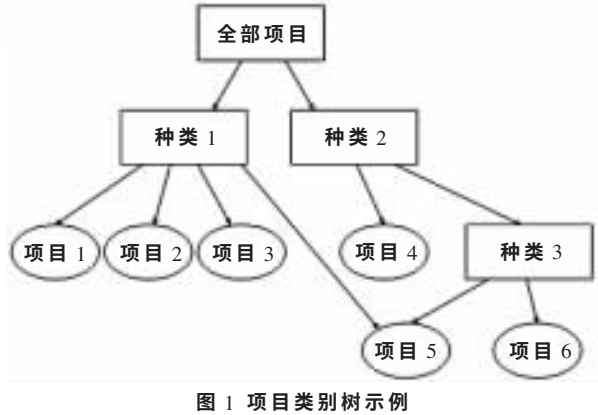

图 1 项目类别树示例

该算法计算的是综合相似性，从而在一定程度上提高了推荐精度，缓解了稀疏性问题。但它也有一定的缺点：(1)类别标示一般需要人工标注，工作量太大，并会造成一定的主观偏见；(2)权重系数的选取一般只能靠经验来定。

1.4.4 其他的改进算法

随着研究的深入，更多的改进推荐算法被提出。如参考文献[11]是在1.4.1中所提到算法的基础上，通过引入类别相似性来改进；参考文献[12]是综合评分相似度和类别相似度，并引入阈值来调整评分相似性和最近邻的选取；参考文献[13]也利用“组合”的思想，但不再是相似度组合，而是将用户加权评分和项目加权评分进行组合来达到与1.4.1中算法一样的目的。这些算法都改进了精度，缓解了稀疏性、冷启动等问题。

2 系统性能的评测

对推荐系统来说，系统的性能直接影响用户对系统的使用程度。因为推荐出来的项目不符合用户的兴趣，用户对系统的信任度就会下降。常用的精度测评方法有平均绝对误差MAE、规范化的平均绝对误差NMAE方法和召回率—精度等方法。以前大家关注较多的是精度问题，现在评测一个系统的性能，除了精度之外，新颖性、多样性、覆盖率等更多指标越来越受重视[14]。

多样性可以分为系统用户间多样性和用户内多样性[15]。前者反映了系统为不同用户推荐不同项目的能力，可以通过计算用户对之间的汉明距离，然后取均值来得出；后者反映了系统为一个用户推荐不同种类项目的能力，可以通过计算某一用户推荐列表中的项目间相似度的差距，最后取均值来得出。

新颖度指系统可以为用户推荐的非流行项目的能力。较常使用的是计算一个用户推荐列表中商品的平均度[16]，商品的平均度越低证明商品的新颖度越高。

覆盖率可以体现出系统为用户推荐的项范围的大小，若覆盖率太低，推荐范围太窄，用户满意度就可能会下降。较常用的为推荐覆盖率[14]，用来计算推荐项占全部项的比例。

3 推荐系统面临的问题

3.1 稀疏性问题

推荐系统中，由于用户浏览的项目有限，导致用户评分矩阵非常稀疏，从而使得推荐精度不高。参考文献[4,8,11]的出发点都是希望缓和矩阵稀疏性问题。因为如果两个用户没有共同评分的项目，那将无法计算他们的相似度。同样的，如果两个项目共同评分的用户交集为空，则项目之间的相似度也无法得出。目前解决稀疏性问题效果较好的是奇异值分解(SVD)的技术[17]。在推荐时，可以先使用SVD方法计算出项目评分，然后再计算相似度最终得出预测。SVD的方法在评分预测上不依赖相似度方法进行计算，并且通过维数化简来得到密集的矩阵。虽然解决了数据稀疏性问题，但是在化简过程中丢失掉了一些数据信息，精度也会受到一定的影响。

3.2 冷启动问题

冷启动问题[7,17]分为新项目和新用户问题。参考文献[10-13]的出发点就是为了解决冷启动问题。因为当一个系统增加一个新的项目并且该项目没有任何用户对其进行评分时，这个项目就无法被推荐出去，新项目问题就产生了。新用户问题的产生是由于系统的用户增多，当用户没有对系统中的项目进行评价时，系统无法分析出用户的偏好，也就无法产生推荐。

目前对于冷启动问题的解决有两种方法：众数法和信息熵法。

一般而言，大部分用户对于喜爱和不喜爱的项目评分较为相似，所以评分的数值分布在一定范围内的概率较大。因此可以使用众数法来对新项目或新用户对项目的评分进行预测。

对于冷启动问题也可采用信息熵法来解决。对于新项目问题，排序计算所得的信息熵，选取其中信息熵较大的用户，通过预测这些用户对该项目的平均评分来作为新项目的预测评分。对于新用户问题，排序用户信息熵值，然后选择其中较大熵值的用户来预测新用户对某一项目的评分。

本文介绍了协同过滤推荐算法的传统算法和改进算法。最后对于系统的性能评测做了详细介绍，对系统常见问题也进行了分析研究。

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] LINDEN G, SMITH B, YORK J. Amazon.com recommendations item-to-item collaborative filtering[J]. IEEE Computer Society, 2003(1-2):76-80.

[2] METEREN R V. M.v.S. using content-based filtering for recommendation[C]. In CML/MLNET Workshop on Machine Learning and the New Information Age, Baecelona, 2000: 47-56.

[3] 黄萱菁，夏迎炬，吴立德. 基于向量空间模型的文本过滤系统[J]. 软件学报，2003，14(3):435-442.

[4] 邓爱林，左子叶，朱扬勇. 基于项目聚类的协同过滤推荐算法[J]. 小型微型计算机系统，2004，25(9):1665-1670.

[5] Badrul Sarwar, G.K., KONSTAN J, et al. Item based collaborative filtering recommendation algorithms[C]. Proceedings of the 10th International World Wide Web Conference, 2001.

[6] 曹毅.基于内容和协同过滤的混合模式推荐技术研究[D]. 长沙：中南大学，2007.

[7] EKSTRAND M D, RIDEL J T, KONSTAN J A. Collaborative filtering recommender systems[M]. Hanover:Now Publishers Inc, 2010:81-173.

[8] 邓爱林，朱扬勇，施伯乐.基于项目评分预测的协同过滤推荐算法[J]. 软件学报，2003，14(9):1621-1628.

[9] 周军锋,汤显,郭景峰.一种优化的协同过滤推荐算法[J]. 计算机研究与发展, 1994,41(10):1842-1847.

[10] 李聪,梁昌勇,董珂.基于项目类别相似性的协同过滤推荐算法[J]. 合肥工业大学学报(自然科学版), 2008,31(3):360-363.

[11] 张忠平,郭献丽.一种优化的基于项目评分预测的协同过滤推荐算法[J].计算机应用研究, 2008,25(9):2658-2660.

[12] 汪静,印鉴. 一种优化的 Item-based 协同过滤推荐算法[J]. 小型微型计算机系统, 2010,31(22):2337-2342.

[13] 马丽. 基于组合加权评分的 Item-based 协同过滤推荐系统[J]. 现代图书情报技术, 2008(11):60-64.

[14] 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.

[15] Zhou Tao, Jiang Luoluo, Su Riqi, et al. Effect of initial configuration on network-based recommendation[J]. Europhysics Letters, 2008(81):58004-58007.

[16] Zhang Zike, Liu Chuang, Zhang Yicheng, et al. Solving the cold-start problem in recommender systems with social tags[J]. Europhysics Letters, 2010,92(2): 28002-28007.

[17] 孙小华.协同过滤系统的稀疏性与冷启动问题研究[D]. 杭州:浙江大学,2005.

(收稿日期:2012-12-20)

作者简介:

张瑶,女,1989 年生,硕士研究生,主要研究方向:数据仓库,数据挖掘。

陈维斌,男,1957 年生,教授,主要研究方向:数据库技术及应用,数据仓库,数据分析。

傅顺开,男,1978 年生,讲师,博士,主要研究方向:数据挖掘,信息检索。