协同过滤算法研究综述

翁星星
合肥工业大学管理学院，安徽合肥　230009

摘　要 本文在介绍传统协同过滤算法的基础上，分析其存在的一些弊端，文章着重介绍了协同过滤算法的研究情况，目的是为协同过滤算法改进提供引导作用。

关 键 字 协同过滤；个性化推荐；稀疏性
中图分类号 TP39　　**文献标识码** A　　**文章编号** 1674-6708（2013）97-0232-02

0 引言

随着网络和电子商务的迅猛发展，用户可以在网上随意寻找自己感兴趣的商品，但随着信息爆炸式增长，用户在这过程中浪费了很多时间，个性化推荐系统对电子商务网站的业绩有很深的影响，其主要作用表现在以下几方面：可以把随意浏览网站的潜在客户转变为实际购买者；提升电子商务网站交叉销售能力；提升客户对网站的忠诚度。其中协同过滤技术是目前运用最广泛的个性化推荐技术。

1 协同过滤算法

协同过滤技术是通过收集整理过去用户产生的数据来寻找邻居用户，其基本原理是根据相似用户的兴趣来推荐当前用户没有参与但是很有可能会感兴趣的项目，所基于的假设是如果两个用户兴趣类似，那么很有可能当前用户会喜欢另一个用户所喜欢的项目。协同过滤推荐技术分为 3 个阶段：评分数据表示；最近邻居形成；推荐项目集产生

1）评分数据表示：将用户对于项目的评分收集整理后描述成一个的用户－项评分矩阵，其中 m 表述用户数，n 表式项目数。矩阵中元素表述用户 对项目的评分；

2）最近邻居形成：指根据项目评分矩阵来发现目标用户的最近邻居。协同过滤技术是通过计算用户之间的相似性来找到目标用户的最近邻，所以算法的关键就在于如何准确找到目标用户的最近邻。常用的用户之间的相似度算法有 Pearson 相关系数和余弦相似性；

3）推荐项目集产生：目标用户的最近邻居集产生后，可以得出目标用户对未评分项的预测分，将分值按照高低排列，产生 TOP-N 的推荐项目集合；

这就导致了协同过滤技术过分依赖于用户评分，但目前电子商务网站的用户和商品数量一直在上升，同时用户对商品项的评分却非常稀少，通常在 1% 以下，使得用户－项目评分矩阵过于稀疏，导致个性化推荐质量下降：

1）评分矩阵稀疏使得寻找最近邻的准确度降低；

2）冷启动（cold-start）问题，此问题是稀疏性的极端情况，指当新用户或新项目进入到推荐系统中时，由于没有历史数据，导致无法产生推荐集。

针对评分矩阵稀疏性问题许多研究人员对协同过滤算法提出了改进，本文系统的归纳和分析了各算法的研究情况，同时为协同过滤算法提供了几点研究方向。

2 改进的协同过滤算法综述

2.1 结合项目相似性和时间函数的协同过滤算法

刘芳先等分析传统协同过滤算法的局限于以下三点：

1）传统算法对于用户之间的相似度是通过两用户共同给予的项目评分来计算的，却没有考虑项目是否相关，如一用户对于某书籍的兴趣可能跟他看过的书有关，而跟他评价过的服装没关系；2）随着时间变化用户的兴趣也会变化的，这点传统算法却没有考虑到；3）传统的协同过滤算法在计算项目间相似性，没能将项目特征考虑在内，导致相似性度量不够准确。

在此基础上刘芳先提出来改进算法，其主要思想是将项目的相关性引入到用户相似性的计算公式中，同时在预测新目标项的得分时引入了时间加权函数，时间加权函数能反映出用户对最近点击的项目兴趣较大，新数据对于预测得分影响大，而旧数据体现的是用户之前的兴趣，所以在预测上占权重较小。

这种改进算法在计算用户相似性的时候引入项目相似度，这样可以在一定程度上减少不相关的项目对于推荐结果的影响，同时将时间函数引入了预测得分的公式中，一定程度上反映出随用户趣变化得到推荐集也不同。但是这算法依然对用户－项目评分矩阵依赖性太大，不利于解决数据稀疏性问题。

刘勇在分析了计算项目相似度时碰到的问题：当两项目只有很少用户给予评分，同时给予评分的用户所关注的项目特征可能不是目标用户所关注的特征，这会导致推荐质量下降。基于这类问题，刘勇提出了改进的相似度计算公式：

$$sim(i,j)=\frac{mutual\_num}{item\_num}\times\frac{\cos(\vec{i},\vec{j})}{\|\vec{i}\|*\|\vec{j}\|}$$

Mutual_num 表示对于项目 i、j 都评分的用户数目，item_num 表示对项目 i，j 中任何一个有评分的用户集合数目。

2.2 降维处理

文献［7］为了降低项目评分矩阵的稀疏性，提升推荐精度，提出了一种基于主成分降维技术和 K-means 聚类的混合协同过滤新算法。算法先对用户－项目矩阵进行缺失值填充，然后运用主成分分析技术提取主成分因子，在降低矩阵的维数同时保证大部分信息没有损失，在降维后的向量空间上进行 K-means 聚类，找到目标用户的最近邻，最后得到目标用户对于未评分项目的预测值，从而产生推荐集。该算法在一定程度上缓解超高维空间寻找最近邻问题。

文献［8］提出了基于项目聚类的协同过滤，算法主要思想是结合项目评分与项目属性的项目相似度，再对项进行聚类。聚类可以通过一些聚类算法将项和用户聚成若干子类，再在各小类中产生推荐集。张娜等先计算项目相似度再用 k 划分聚类算法进行项目聚类，产生 k 个用户－项目子矩阵，然后对已有的项目聚类结果用 k 划分算法进行客户聚类，最后在目标用户所在的几个矩阵中寻找最近邻。

2.3 结合基于内容推荐的协同过滤算法

文献［10］在分析了传统协同过滤在处理新项目和新用户问题上的瓶颈提出了结合基于内容推荐的协同技术。协同过滤算法过分依赖于用户评分，而对于新项目和新用户没能产生评分数据，推荐集中就不会出现，但基于内容的推荐算法对于每

个用户都有用户描述，其中记录了用户感兴趣的内容。可以根据用户喜好和项目的特征信息，推荐给与目标用户特征相似的项目，这就能较好的解决这一问题。

虽然这算法可以一定程度上解决“新项目”问题，但也存在一定的局限：用户或项目特征提取能力有限，目前只能进行简单的提取，对于项目特征不能做到准确的定位，基于内容的推荐现阶段只能对文本内容提取，而对于一些影像，图像很难做到提取特征。

2.4 结合基于关联规则的协同过滤算法

文献[11]提出了一种结合关联规则和协同过滤的算法，其主要思想是：先通过关联规则在商品项中找到频繁项，再将这些频繁项捆绑在一起对目标用户进行推荐，这就可以更好更多的产生推荐集了。但是目前这方面算法研究还处于初级阶段，可以从以下几方面进行进一步的研究：1）如何将Web日志预处理更好的融入到协同过滤中去；2）面对数据快速更新速度，如何剔除无用的信息，保证推荐及时性和准确性；3）如何更好的将这一推荐技术应用到实践中。

2.5 其他的一些改进算法

傅鹤岗[12]等在分析了传统协同过滤算法在用户数量快速增长的时代下所需要付出的代价很大，提出了基于模范用户的协同过滤算法。其主要思想是：用户的兴趣常集中在某几个特定区域，可以先对用户进行聚类，使得类内相似度高而类间相似度低，再在这基础上产生推荐集。施凤仙[13]等提出了结合项目区分用户兴趣度的协同过滤算法，其主要思想是在计算用户相似度时对于不同的项目所占的权重不同，因为用户对于很多大众流行产品评分很高但不能真正反映用户的兴趣度，

3 总结与展望

随着电子商务迅速发展，用户及商品项都呈现爆炸式增长，同时用户对商品项的评分又过于稀少，导致数据过分稀疏，对于未来个性化推荐系统发展来说这是个瓶颈。本文总结了大量研究人员提出的改进算法，这些算法在一定程度上能解决数据稀疏性问题。但这一问题一直都存在，因此对该算法如何改进还需要进一步研究探讨，下一步的工作可以从以下几方面进行：

1）建立一套完善的评分激励制度。这可以从根本上解决数据稀疏性问题，完善的激励制度可以使得用户愿意客观的去给予商品项评分，通过这项制度，可以得到更多准确，可信度高的评分项，从而利于推荐系统产生推荐集；

2）与政府及企业部门共享客户资料。目前的政府和企业都有一套完善的管理系统，其中包含了很多个人信息，如果可以将这些信息和电子商务网站上的客户信息整合，那数据稀疏性问题可以得到一定程度的解决；

3）如何将新的评价替代旧的评价。用户的兴趣会随着时间变化，用户对于某商品项的评价也会改变，在推荐系统中如何快速有效的用新评价来替代旧评价有待于进一步的研究。

参考文献

[1]赵亮，胡乃静，张守志.个性化推荐算法设计[J].计算机研究与发展，2002，39(8)：986-990.

[2]Sarwar BM.Sparsity，scalability，and distribution in recommender systems[D].Minneapolis，USA：University of Minnesota，2001.

[3]Park ST，Pennock D，Madani O，et al.Na? ve filterbots for obust cold-start recommendations[A].In：Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C].New York，USA：ACM Press，2006：699-705.

[4]刘芳先，宋顺林.改进的协同过滤推荐算法[J].计算机工程与应用，2011，47(8)：72-75.

[5]张丙奇.域知识的个性化推荐算法研究[J].计算机工程，2005，31(21)：7-9.

[6]刘勇.基于项目相似度计算改进的协同过滤算法[J].商场现代化，2007，520：84-85.

[7]郁雪，李敏强.一种结合有效降维和K-means聚类的协同过滤推荐模型[J].计算机应用研究，2009，26(10)：718-3720.

[8]鲁培.一种改进的基于项目聚类的协同过滤推荐算法[J].科技传播，2011，1：205-206.

[9]张娜，何建民. 基于项目与客户聚类的协同过滤推荐方法[J].合肥工业大学学报，2007，30(9)：1160-1162.

[10]Adomavicius G，Tuzhilin A.Toward t he Next Generation of Recommender Systems：A Survey of the St ate-of-the-art and Possible Extensions[J].IEEE Transaction on K now ledge and Data Engineering，2005，17(6)：734-749.

[11]裘立波，姜元春，林文龙.基于关联规则和协同过滤的网络商品捆绑方法研究[J].计算机与现代化，2009，9:169-172.

[12]傅鹤岗，彭晋. 基于模范用户的改进协同过滤算法[J].计算机工程，2011，39(3)：70-72.

[13]施凤仙，陈恩红. 结合项目区分用户兴趣度的协同过滤算法[J].小型微型计算机系统，2012，7(33)：1533-1536.

协同过滤算法研究综述

作者：翁星星
作者单位：合肥工业大学管理学院,安徽合肥,230009
刊名：科技传播
英文刊名：PUBLIC COMMUNICATION OF SCIENCE & TECHNOLOGY
年，卷(期)：2013(16)

参考文献(13条)

1. 赵亮.胡乃静.张守志 个性化推荐算法设计[期刊论文]-计算机研究与发展 2002(08)

2. Sarwar BM Sparsity,scalability,and distribution in recommender systems 2001

3. Park ST.Pennock D.Madani O Na ve filterbots for obust cold-start recommendations 2006

4. 刘芳先.宋顺林 改进的协同过滤推荐算法[期刊论文]-计算机工程与应用 2011(08)

5. 张丙奇 域知识的个性化推荐算法研究 2005(21)

6. 刘勇 基于项目相似度计算改进的协同过滤算法 2007

7. 郁雪.李敏强 一种结合有效降维和K-means聚类的协同过滤推荐模型[期刊论文]-计算机应用研究 2009(10)

8. 鲁培 一种改进的基于项目聚类的协同过滤推荐算法 2011

9. 张娜.何建民 基于项目与客户聚类的协同过滤推荐方法 2007(09)

10. Adomavicius G.Tuzhilin A Toward t he Next Generation of Recommender Systems:A Survey of the St ate-of-the-art and Possible Extensions 2005(06)

11. 裘立波.姜元春.林文龙 基于关联规则和协同过滤的网络商品捆绑方法研究 2009

12. 傅鹤岗.彭晋 基于模范用户的改进协同过滤算法[期刊论文]-计算机工程 2011(03)

13. 施凤仙.陈恩红 结合项目区分用户兴趣度的协同过滤算法 2012(33)