浙江大學

硕 士 研 究 生 读 书 报 告

题目	协同过滤推荐技术综述
作者姓名	李林
作者学号	21651048
指导教师	李启雷
学科专业	移动互联网与游戏开发技术
所在学院	软件学院
提交日期	二〇 一六 年 十二 月

Survey of Recommendation Based on Collaborative Filtering

Major Subject: Software Engineering

Advisor: liqilei

By

Lilin

Zhejiang University, P。R。 China

2016

摘要

协同过滤是推荐系统中广泛使用的推荐技术,研究人员对如何完善协同过滤推荐技术开展大量工作,但是相应的研究总结较少。文中对协同过滤的相关研究进行全面回顾,首先阐述协同过滤的内涵及其存在的主要问题,包括稀疏性、多内容及可扩展性,然后详细介绍国内外学者针对以上问题的解决方案,最后指出协同过滤下一步的研究重点。文中介绍一个相对完整的协同过滤知识框架,对理清协同过滤的研究脉络,为后续研究提供参考,推进个性化信息服务的发展具有一定意义。

关键词:个性化服务,推荐系统,协同过滤

Abstract

Collaborative filtering is a widely used technique in recommender systems。Extensive studies are carried out on collaborative filtering. However ,systematic summary of this field is scarce. In this paper, research of collaborative filtering is reviewed. The meaning and key issues of collaborative filtering , including sparsity , multiple-content and scalability are described firstly , and then the solutions to the above key issues are introduced in detail. Finally , the future work of collaborative filtering is pointed out. The knowledge framework of collaborative filtering is introduced. It makes the research clues of collaborative filtering clear provides a reference to other scholars , and improves the performance of personalized information services.

Keywords: Personalized Service，Recommender System，Collaborative Filtering

协同过滤推荐技术综述

1、引言

协同过滤是推荐系统中广泛使用的推荐技术，研究人员对如何完善协同过滤推荐技术开展大量工作，但是相应的研究总结较少。文中对协同过滤的相关研究进行全面回顾，首先阐述协同过滤的内涵及其存在的主要问 题，包括稀疏性、多内容及可扩展性，然后详细介绍国内外学者针对以上问题的解决方案，最后指出协同过滤下一步的研究重点。文中介绍一个相对完整的协同过滤知识框架，对理清协同过滤的研究脉络，为后续研究提供参考，推进个性化信息服务的发展具有一定意义。

在随着移动互联网、物联网、云计算等技术的快速 发展，全球数据量呈爆炸式增长，大数据时代已经到来。 全球数据量级从 TB 发展至 PB 乃至 ZB，可称为 海量、巨量甚至超量。 相对于以往便于存储的以文本 为主的结构化数据，音频、视频、图片、地理位置信 息等非结构化数据的比例逐步提升，达到 80%左右。 互联网上信息的急剧增长，一方面使人们获 取的信息资源越来越丰富，给人们带来极大的便利; 另一方面，面对海量的信息资源，人们不得不花费更 多的时间和精力去搜寻对其有帮助的信息，因此 “信息超载”[1]现象越来越 严重。

推荐系统是解决信 息超载问题的有效方案，它根据用户特征推荐满足用户需求的对象，实现个性化服务。 推荐系统的优点在于：它能主动收集用户的特征资料，通过对用户个性 、习惯 、偏好的分析，为用户定制并提供其感兴趣的信息; 同时能及时跟踪用户的需求变化，根据变化自动调整信息服务的方式和内容。 与搜索引擎提供 的“一对多”式的信息服务不同，推荐系统输出的结 果更符合用户的个性化需求，实现“一对一”式的信息服务，同时用户的参与程度也更低，

从而大为降低 用户搜寻信息的成本。 推荐系统作为一种新近的智能信息服务方式，在电子商务、社会网络、数字化图书馆、视频/音乐点播等领域得到广泛应用。 特别是在电子商务领域，其作用尤为突出，主要表现为以下几点：1) 将电子 商务网站浏览者转变为购买者; 2) 提高电子商务网站的交叉销售能力; 3) 建立客户忠诚度。

目前推荐系统所采用的推荐技术主要包括关联规则[2]、基于内容的推荐 [3] 、协同过滤[4] 和 混 合 推 荐 [5] 。协同过滤根据其他用户的偏好向目标用户推荐，它首先找出一组与目标用户偏好一致的邻居用户，然后分析该邻居用户，把邻居用户喜欢的项目推荐给目标用户。协同过滤的优势在于以下四点: 1) 不需考虑被推荐项目的内容; 2) 可为用户提供新异推荐; 3) 对用户访问网站时的干扰较小; 4) 技术易于实现。因此它成为一种较流行的推荐技术。

2，协同过滤及其存在的问题

如推荐系统的根源最早可追溯到认知科学、近似 理论、信息检索、管理科学等领域的研究，自 20 世纪 90 年代期第一个推荐系统 Tapestry 提出以来，推荐系统开始成为一个独立的研究领域，并一直保持着较高的研究热度。 众多组织和学者对推荐系 统展开广泛、深入的研究，并从不同角度定义推荐系统。 本文综合众多文献的描述，将推荐系统定义为: 推荐系统采用统计分析和机器学习技术识别用户的兴趣偏好，向用户提供满足其需求的信息、商品和服务的推荐，从而减少用户的搜寻成本，增加网站运营 商的收益。

从20世纪90年代初，出现了很多推荐系统，推荐系统各有各的优劣势，下面主要介绍的是关联规则、基于内容的推荐、混合推荐、协同过滤四个技术的优劣。如下表所示。

推荐技术	主要优点	主要缺点
关联规则	自动化程度高	个性化程度低 规则难以提取 规则质量难以保证
基于内容的推荐	个性化程度高 自动化程度高 结果直观易解析	有限的内容分析
混合推荐	个性化程度高 自动化程度高 推荐准确性高	有限的内容分析 技术实现难度高 执行效率低
协同过滤	个性化程度高 自动化程度高 推荐领域广 技术易实现	稀疏性问题 多内容问题 可扩展性问题

从这张表可以看出，协同过滤技术相对一般的技术性能上还是很优异的，协同过滤通常可分为两类: 基于记忆的协同过滤和基于模型的协同过滤。 基于记忆的协同过滤利用整个用户 - 项目评分数据集进行计算，每个用户都是评分预测过程的组成部分。 基于记忆的协同过滤为目标用户选择一部分兴趣相近的邻居用户，根据邻居用户的评分预测目标用户对项目的评分值。 典型的基于记忆的协同过滤有最近邻协同过滤及其改进算法。 基于模型的协同过滤根据训练集数据学习得出一个复杂的模型，然后基于该模型和目标用户已评分数据，推导出目标用户对未评分项目的评分值。 典型的基于模型的协同过滤有基于聚类技术的协同过滤、基于概率方法的协同过滤、基于矩阵分解的协同过滤等。

协同过滤技术具有其他推荐技术无法比拟的优势，具体如下。

1) 推荐对象可为任何类型的资源。 由于协同过滤基于相似用户的兴趣偏好产生推荐，因此它的关键是找出与目标用户偏好相似的邻居用户，而不必分析、提取项目内容信息，这使得协同过滤技术可将任何类型的资源项推荐给用户。

2) 产生新异推荐。 协同过滤不对比项目内容与 用户描述间的相关性，所推荐项目不完全局限于用 户的历史偏好范围内，这将有助于发现用户的潜在 兴趣，实现跳跃式推荐。

3) 对用户干扰较小。 协同过滤一般使用评分向量作为用户偏好的表示方式，用户登录系统后只需对一些项目评分，便可获得推荐服务，用户使用系统 的整个过程中受到的干扰较小。

4) 技术易于实现。 协同过滤的用户偏好信息搜 集容易，算法程序并不复杂且可扩展性较好，因此整个协同过滤系统的实现较为容易。

尽管协同过滤在个性化推荐方面取得较大成 功，但其本身存在的一些关键问题制约着其进一步 发展，本文将这些问题概括如下。

1) 稀疏性[6]。 实际的网站中用户和项目数量庞大，而用户通常只对一小部分项目评分，可用于计算用户/项目间相似性的数据非常有限，使得最近邻搜寻不够准确，推荐质量较差。 经常可看到的现象是两个用户/项目间没有任何共同评分项，导致相似性无法计算。 即使有的用户 / 项目间相似性可计算，可靠性也难以保证。

2) 多内容[7]。 传统的协同过滤没有考虑项目类别的影响，当网站中项目类别的内容完全不同时，传统的协同过滤算法搜寻出的最 近邻往往与目标用户仅在个别项类上偏好相似，导致推荐结果不够合理。

3) 可扩展性[8]。 网站中用 户和项目的数量庞大(如 Amazon 商城拥有数以百万计的商品) ，且不断增加，这使得用户 - 项目评分 矩阵成为高维矩阵，由此产生协同过滤的可扩展性问题 ,即随着用户和项目数量的增多 ,算法的计算复 杂度急剧增加，严重影响系统推荐的实时性。

3，协同过滤的现状

国内外研究人员针对协同过滤中存在的稀疏性、可扩展性、多内容等问题进行广泛而深入的研究。

其中稀疏性问题主要有几种解决方案：空值填补、新相似性方法、结合基于内容的推荐 、推荐结果融合、图论。各种解决方案的优劣如下表所示。

	空值填补	新相似性研究	结合基于内容的推荐	推荐结果融合	图论
稀疏性改善程度	高	低	高	低	低
算法推荐质量	低	中	高	中	中
算法实现难度	中	低	高	低	高

其中可扩展性问题主要有几种解决方案：数据集缩减、聚类、矩阵分解、主成分分析、增量更新。各种解决方案的优劣如下表所示。

	数据集缩减	聚类	矩阵分解	主成分分析	增量更新
可扩展性改善程度	高	高	高	高	中
算法推荐质量	低	中	中	低	高

算法 实现难度	低	中	高	中	低

其中多内容问题主要由Yu等提出,并且由他们提出解决方案,他们指出传 统的协同过滤算法没有考虑项目类别的影响，当网 站中项目类别的内容完全不同时，传统的协同过滤 算法搜寻出的最近邻往往不够合理，导致推荐质量 较差。针对这一问题，他们首先计算目标项目与其他 项目的相似性，确定目标项目所属类别，然后在目标 项目所属类别中搜寻目标用户的最近邻。 由于所有 最近邻对被预测项目的内容都较熟悉，因此推荐精 度较高。

但是 Yu[9] 的方法仅在一个项目类别中搜寻最近邻，实际网站中用户的评分非常稀疏，一个项目类别 中的评分更稀疏，使得他们的方法在实际的应用中 效果很差。 文献针对以上问题提出一种基于项 类偏好的协同过滤推荐算法，首先为目标用户找出 一组项类偏好一致的候选邻居，然后在候选邻居中 搜寻目标用户的 k 最近邻，从整体上提高最近邻搜 寻的准确性。

对于多内容问题，国内外学者的研究尚处于探索阶段，并不像稀疏性问题、可扩展性问题的研究那 样成熟。在多内容问题研究方面,学者们更多采用单 独项类搜寻最近邻的方法，单独项类评分数据异常 稀疏，这在现实系统中是不可行的。 可考虑设计一些 方法,如首先计算用户间的项目类别偏好相似性,然 后在项目类别偏好相似性较高的用户中搜寻目标用 户的最近邻。或结合基于内容的推荐,对目标用户感 兴趣项类中的特定项目进行内容分析,找出与这些 内容匹配度较高的备选项目，然后根据目标用户最 近邻的意见，从备选项目中产生top-N[10]推荐集。

除以上问题外，研究人员还研究协同过滤中存 在的同义词、托攻击、冷启动、推荐信任度、显示跟踪 等问题。

另外还有研究人员指出协同过滤系统在进行项 目推荐时，不仅要根据预测评分的高低选择项目，还应考虑预测评分的可信度。 Schafer 等的研究表 明，推荐项目时综合考虑预测评分和信任度比单纯 采用预测评分更合理，他们分别给出基于用户和基 于项目协同过滤的信任度计算方法。

4，未来研究方向

尽管研究人员研究协同过滤的稀疏性问题、多内容问题和可扩展性问题，并取得一些成果，然而对 于协同过滤的研究还有待不断完善，今后的进一步 研究工作包括如下方面。

1) 新用户偏好模型。 用户偏好信息是对用户提 供个性化推荐服务的基础，系统只有充分掌握用户 偏好，才有可能提供高质量的推荐。目前协同过滤推 荐系统主要让用户评分项目，以评分信息反映用户 的偏好。 评分方式简单、直接，易于推荐模型的建立， 同时对用户操作系统时的干扰较小，但评分数据所 包含的信息量有限，仅依靠评分难以提供令用户满 意的推荐。 另外很多项目不适合用纯定量化的方式 评价，这给用户表达偏好 / 系统分析用户偏好增加难 度。因此有必要尝试构建新的用户偏好模型，可从用 户对项目的评论信息中提取用户偏好，建立基于文 本信息的用户偏好模型。 借鉴文献中的 PRCUM，挖掘用户评论项目的文本信息，采 用特征词模式、特征词权重表示用户偏好，将用户偏 好表达拓展到文本信息空间。 用户感兴趣的特征词 数量远大于评分项目数量，可降低偏好数据的稀疏 性。 词语等文本信息与评分值相比具有更丰富的含 义，能提高偏好表达的准确性。

2) 改进评价方法。 评价方法可用于评估协同过 滤系统的性能，确定系统是否很好地满足用户需求。 好的评价方法可跟踪用户使用系统满意程 度，及时发

现系统不足，不断完善协同过滤系统。 目 前绝大多数协同过滤算法的评价采用离线分析的方式，即将用户已评分的项目划分为训练集和测试集， 采用训练集中的数据预测用户测试集中的偏好，将 测试集中的真实偏好与预测偏好对比评估算法优 劣。 该方法以用户评价过的项目为基础评估算法，是 对真实情况的一种模拟。 离线分析的优势在于它能 在任何数据集上快速、经济地评价算法，尤其是规模 很大的数据集。但该方法有 2 点明显不足: (1) 评价 时所用项目仅局限于用户已有评分的项目 ,由于数 据稀疏性 ,多数项目不能用于评估算法; (2) 仅局限 于对预测结果的客观评价，不能反映用户对协同过 滤系统的主观偏好。一种变通的方法是在线对用户 访 问 、调 查 ,收 集 用 户 对 推 荐 项 目 集 合 的 评 论 信 息 ， 或直接获取用户使用系统的满意程度。 在线评 价通过与用户互动得到其真实感受，较离线分析更 准确，是未来评价协同过滤系统的主要趋势。

另外要扩展系统的评价准则。 最常用的评价准 则是准确性准则，如评价预测准确性的平均绝对误差、均方根误差等 ,评价推荐准确性的 Recall、Precision、F1 等。 准确性反映的是系统的预测 评分能力，预测准确性高并不能保证系统用户有一 个满意的使用经历。 协同过滤系统不仅要提供准确 的预测，更要提供有用的推荐。因此用于衡量系统有 用性的准则需不断被采用 ,如覆盖率准则、新颖性准则、学习率准则、满意度准则等。

3) 推荐解释。 协同过滤系统的界面多向用户展 示项目的推荐列表(有时伴随项目的预测评分) ,采用的是黑箱机制(Black Box) ,缺乏对推荐原理的介 绍。解释推荐产生的原因可使系统的工作机理更透 明，从而帮助用户了解系统的优缺点。 对推荐过程中的细节和推荐结果做出解释，使用户知道怎样获得 的 推

荐，为何如此推荐，用户会更加信任系统。Herlocker 等探索性研究协同过滤的推荐解释问。

5，结束语

推荐系统帮助用户在海量的信息资源中搜寻真 正有价值的信息，节约用户的搜寻成本，同时也提高 用户对网站的忠诚度，增加网站收益。 由于巨大的应用需求，推荐系统得到学术界和企业界的广泛关注。 美国计算机学会 ACM 多次把推荐系统列为研讨主题，国内外期刊 也纷纷将推荐系统作为研究专题。 推荐系统在电 子商务、社会网络、数字化图书馆、视频 / 音乐点播等 领域得到广泛应用，未来的社会化网站将由推荐系 统所驱动。

协同过滤是推荐系统中应用较成功的推荐技术，自 1992 年美国施乐公司 PARC 研究中心正式提 出协同过滤以来 ,国内外众多研究人员对如何完善 协同过滤推荐技术开展大量的工作。 本文对协同过 滤领域的相关研究进行系统归纳，介绍一个相对完 整的协同过滤知识框架。 这对于理清协同过滤的研 究脉络，为后续研究提供参考，推进我国个性化信息 服务的发展具有一定意义

参考文献

[1]Borchers A，Herlocker J，Konstan J，et al. Ganging up on Information Overload. Computer，1998，31(4) : 106 - 108

[2]Su X N，Yang J L，Deng S H，et al. Theory and Technology of Data Mining. Beijing，China: Scientific and Technical Documentation Press，2003 (in Chinese)(苏新宁，杨建林，邓三鸿，等. 数据挖掘理论与技术. 北京: 科学技术文献出版社，2003)

[3]Baeza-Yates R，Ribeiro-Neto B. Modern Information Retrieval. New York，USA: ACM Press，1997

[4]Li C. Research on the Bottleneck Problems of Collaborative Filtering in E-commerce Recommender Systems. Ph. D Dissertation. Hefei， China: Hefei University of Technology，2009 (in Chinese)(李聪. 电子商务推荐系统中协同过滤瓶颈问题研究. 博士学位 论文. 合肥: 合肥工业大学，2009)

[5]Xu H L ,Wu X ,Li X D ,et al. Comparison Study of Internet Re- commendation System. Journal of Software，2009，20(2) : 350 - 362 (in Chinese)(许海玲，吴潇 ,李晓东 ,等. 互联网推荐系统比较研究. 软件学报 ,2009 ,20(2) : 350 - 362)

[6]Sarwar B，Karypis G，Konstan J，et al. Item-Based Collaborative Filtering Recommendation Algorithms / / Proc of the 10th Interna- tional Conference on World Wide Web. Hong Kong，China，2001: 285 - 295

[7]Yu L，Liu L，Li X F. A Hybrid Collaborative Filtering Method for Multiple-Interests and Multiple-Content Recommendation in E-commerce. Expert Systems with Applications，2005，28(1) : 67 -77

[8] Albadvi A，Shahbazi M. A Hybrid Recommendation Technique Based on Product Category Attributes. Expert Systems with Appli- cations，2009，36(9) : 11480 - 11488

[9]Yu K，Schwaighofer A，Tresp V，et al. Probabilistic Memory-Based Collaborative Filtering. IEEE Trans on Knowledge and Data Engineering，2004，16(1) : 56 - 69

[10]KarypisG.EvaluationofItem-BasedTop-NRecommendationAlgo-rithms Proc of the 10th International Conference on Information and Knowledge Management. Atlanta，USA，2001: 247 - 254