个性化推荐算法研究综述

张志威

（广州大学华软软件学院 软件工程系，广东 广州 510990）

摘 要：数据大爆炸时代的到来，面对信息过载问题，如何让人们在海量的数据面前能够获取到自己想要的信息，于是设计出了推荐系统。个性化推荐算法作为系统的核心，目前主要有以下 5 个分类：基于内容的推荐、基于关联规则的推荐、基于知识的推荐、协同过滤推荐和混合推荐 [1]。因此，对个性化推荐算法进行综述，介绍个性化推荐算法常用的 5 个评估指标，并给出未来可能研究的热点和方向。

关键词：个性化推荐算法；协同过滤；混合推荐；算法评估指标

中图分类号：TP391.3 **文献标识码：**A **文章编号：**1003-9767（2018）17-027-03

A Survey of Personalized Recommendation Algorithms

Zhang Zhiwei

(Department of Software Engineering, South China Institute of Software Engineering.GU, Guangzhou Guangdong 510990, China)

Abstract: With the advent of the data explosion era, when faced with the problem of information overloading, how can people get the information they want in front of massive data? The recommendation system aims at solving this situation. The personalized recommendation algorithm is the core of the recommendation system, which currently has the following five categories: content-based recommendation, association rule-based recom-mendation, knowledge-based recommendation,collaborative filtering rec-ommendation and hybrid recommendation. Next, the research will state the above personalized recommendation algorithms. Finally, this paper will introduce the five evaluation indicators which are frequently used in personalized recommendation algorithms as well as the possible research points and directions in the future.

Key words: personalized recommendation algorithm; collaborative filtering; hybrid recommendation; algorithm evaluation index

1 引言

数据大爆炸时代的到来，需要人们正确处理之前“信息匮乏”的问题和现在“信息过载”的问题 [1]。在浩瀚的数据面前很难准确有效地获取到自己需要的信息，因此，就需要用到信息过滤技术。信息过滤技术目前主要是分为检索和搜索引擎技术与推荐系统技术。分类检索和搜索引擎缓解了信息过载问题，当信息分类不准确或者用户输入的关键词过少等问题，会增加用户的检索时间及影响检索结果。个性化推荐算法作为推荐系统的核心，通过收集用户之前的一些历史记录等信息，分析用户的偏好，对用户产生推荐。现如今，个性化推荐算法已经在电子商务、教育及旅游服务等领域有着深入的研究与应用。

2 个性化推荐算法

个性化推荐算法目前主要有以下 5 个分类：基于内容的推荐、基于关联规则的推荐、基于知识的推荐、协同过滤推荐和混合推荐 [2]。

2.1 基于内容的推荐

基于内容的推荐算法是一个简单但又重要的推荐思想 [3]。首先，该算法需要项目的各个属性特征，构建项目属性向量。接着分析用户历史行为记录，构建用户兴趣偏好向量，通过计算用户兴趣偏好向量与未评价的各个项目自身的属性向量的相似度大小比较，对目标用户产生项目预测评分或 top-N 推荐。假设用户兴趣偏好向量为 u，未评价某项目特征向量为 p，$u=(t_1,t_2,\cdots,t_i,\cdots,t_n),i=1,2,\cdots,n$，$p=(f_1,f_2,\cdots,f_i,\cdots,f_n),i=1,2,\cdots,n$，可以利用相似度计算公式如余弦相似度得出用户兴趣偏好向量与项目属性向量的相似度，相应的计算公式如 2.1 式所示：

基金项目：广东省特色创新类项目（自然科学类）“智慧博物馆移动应用公众服务平台”（项目编号：2015KTSCX176）。

作者简介：张志威（1990-），男，四川广汉人，硕士研究生。研究方向：智能信息处理、软件测试。

$$\mathrm{sim}(u,p)=\cos(u,p)=\frac{u\cdot p}{|u|\times|p|}=\frac{\sum_{i=1}^{n}t_i\times f_i}{\sqrt{\sum_{i=1}^{n}t_i^2}\times\sqrt{\sum_{i=1}^{n}f_i^2}} \quad (2.1)$$

计算的相似度值越大，说明用户对该未评分项目的特征偏好程序越高。推荐的结果与其他用户无关，只与项目本身的属性特征信息和该目标用户对若干项目产生历史行为数据有关。该算法优点是推荐结果直观，缺点是存在数据的稀疏性、复杂的项目难以提取出项目属性特征向量和项目属性之间的重要程度权重值分配等问题。因此，文献 [4] 研究了特征权重的选取方法对推荐效果的影响；文献 [5] 将项目语义应用于个性化推荐；文献 [6] 通过分析项目属性关系将项目粒度化，提出了一种基于内容的加权粒度序列推荐算法；文献 [7] 提出基于内容和兴趣漂移模型应用于电影推荐算法中。

2.2 基于关联规则的推荐

关联规则问题是 Agrawal 等人于 1993 年提出来的 [8]。为寻找数据库数据项之间的关系，提出频繁项集概念，再根据频繁项集定义关联规则。关联规则推荐过程：第一，用户或专家指定最小支持度阈值和最小置信度阈值；第二，从数据库中找到不低于最小支持度的频繁项集；第三，利用第二步中得到的高频项集来产生满足最小置信度的强规则；第四，根据强规则实施个性化推荐。当数据集比较大时，比较影响算法效率。文献 [9] 利用关联规则挖掘的特性，挖掘用户属性与项目之间的关联，提出了基于关联规则挖掘的分类随机游走算法；文献 [10] 中将关联规则挖掘应用与商品推荐中，通过用户历史记录分析挖掘，得到不错的推荐效果。

2.3 构建知识网络

基于知识推荐，需要建立知识网络图谱，分析用户已有的知识和需求的知识之间的关联，为用户推荐新知识。文献 [11] 考虑将知识系统以知识网络进行表达，引入最近邻优先的候选知识选择策略，提出一种基于建构主义学习理论的个性化知识推荐方法 - 建构推荐模型，通过知识网络的知识关联结构挖掘用户知识需求，并推荐出最具建构学习价值的新知识。针对不同的应用场景，如何构建一个最为合理的知识网络是一个关键性的问题。

2.4 协同过滤推荐

协同过滤推荐是 Goldberg 等人于 1992 年提出来的 [12]，是个性化推荐算法中研究和应用最为广泛的推荐算法。协同过滤推荐算法只需要用户对项目进行评分，不需要用户特征和项目属性。主要有以下两种推荐算法：基于内存的协同过滤推荐算法和基于模型的协同过滤推荐算法。基于内存的协同过滤推荐又可以分为基于用户的协同过滤和基于项目的协同过滤 [13]。

2.4.1 基于用户的协同过滤

基于用户的协同过滤主要有 4 个步骤：第一步，根据用户 - 项目评分矩阵，进行用户相似度计算，得到用户相似度矩阵；第二步，通过用户相似度矩阵，运用 kNN 最近邻居算法，通过用户相似度大小选择与目标用户最近的 k 个邻居；第三步，利用目标用户的最近邻居集合，计算出目标用户的项目预测评分；第四步，选择预测评分值较大的 N 个进行目标用户的 top-N 推荐。

2.4.2 基于项目的协同过滤

基于项目的协同过滤和上述算法的计算过程类似，只是在第一步和第二步计算方法上有所区别，基于项目的协同过滤是通过计算项目之间的相似度大小，形成项目相似度矩阵，对目标用户的未评分项目，分别计算项目相似度矩阵中的每个项目的相似度大小并形成最近邻项目集合。

基于用户的协同过滤和基于项目的协同过滤都仅仅依赖于整个用户 - 项目评分矩阵，不需要根据用户本身的特征进行建模，也不根据项目本身的属性进行建模。以上两种经典的协同过滤算法在开源库如 Mahout 算法库中已经实现。

2.4.3 基于模型的协同过滤

基于模型的协同过滤通过数据挖掘、机器学习等知识建立用户 - 项目的评分预测模型，主要有以下两种方法：（1）基于矩阵的奇异值分解模型。将整个数据集的用户 - 项目评分矩阵，通过矩阵分解为用户特征矩阵和项目特征矩阵，从而降低矩阵的维度，进行相似度的计算并推荐；（2）基于聚类模型。如通过 Canopy 聚类，K 均值算法等算法，对用户或项目进行聚类，从而对目标用户所属某一类的评分进行预测并推荐。

协同过滤推荐算法适用难于提取出项目属性特征的半结构化或非结构化数据，如音频、视频等数据。随着系统规模的增大，协同过滤推荐算法将面临着数据稀疏性问题、冷启动（用户冷启动，项目冷启动）问题和算法实时性、扩展性等问题，影响推荐系统的推荐准确度和效率。

2.5 混合推荐

混合推荐是将若干种个性化推荐算法通过某种方式进行结合。针对不同的应用场景，考虑各单一的个性化推荐算法的优缺点，增强推荐效果。常用的有以下 5 种方式进行结合：加权、变换、混合、特征组合和层叠等。文献 [14] 利用数据挖掘知识和协同过滤算法，提出一种结合用户聚类和评分偏好的混合推荐算法；文献 [15] 提出一种基于用户特征聚类和 Slope One 填充的协同过滤混合推荐算法，试图解决数据稀疏性、可扩展性问题；文献 [16] 提出搭建 Hadoop 分布式平台，并结合 Mahout 工具进行组合推荐算法的设计与实现，应用于电影推荐；文献 [17] 提出协同过滤推荐算法在融合大数据技术、社会网络分析技术及关键用户分析技术方面的研究。可见，结合机器学习、数据挖掘等知识的混合推荐算法，是今后研究与应用的主要方向。

3 推荐算法评估

个性化推荐算法的运行效果需要有算法评估指标来

衡量。目前主要有以下 5 个算法评估指标：平均绝对误差 MAE、均方根误差 RMSE、准确率 Precision、召回率 Recall 及 F-Measure 测量值 [18]。

3.1 平均绝对误差 MAE

MAE 是用户预测评分值与用户实际评分值差值的绝对值的平均值大小，是最常用的评估指标。用户预测评分集合为 $\{p_i\}$，用户实际评分集合为 $\{r_i\}$，集合元素个数为 N，MAE 计算公式如公式（1）所示：

$$MAE=\frac{\sum_{i=1}^{N}|p_i-r_i|}{N} \quad (1)$$

3.2 均方根误差 RMSE

RMSE 也是较为常用的评估指标之一，计算公式如公式（2）所示：

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(p_i-r_i)^2}{N}} \quad (2)$$

3.3 准确率 Precision

在对于要给出 top-N 的推荐系统中，经常用准确率 Precision、召回率 Recall 和 F-Measure 测量值来进行系统性能的评估。TP 表示正确的推荐结果，FP 表示错误的推荐结果，FN 表示过滤的正确结果，准确率 Precision 表示正确的推荐结果与推荐的结果总数的比值，计算公式如公式（3）所示：

$$Precision=\frac{TP}{(TP+FP)} \quad (3)$$

3.4 召回率 Recall

召回率 Recall 指的是正确的推荐结果与正确的结果总数的比值，计算公式如公式（4）所示：

$$Recall=\frac{TP}{(TP+FN)} \quad (4)$$

3.5 F-Measure 测量值

F-Measure 测量值，要用到上面介绍的准确率 Precision 和召回率 Recall 两个评估指标，计算公式如公式（5）所示：

$$F_\alpha-Measure=\frac{(\alpha^2+1)\times Precision\times Recall}{\alpha^2\times Precison+Recall} \quad (5)$$

特别地，当 α 为 1，即 F_1-Measure 值，计算公式如公式（6）所示：

$$F_1-Measure=\frac{2\times Precision\times Recall}{Precison+Recall} \quad (6)$$

4 结 语

通过对个性化推荐算法的核心算法进行综述，介绍常用的 5 个算法评估指标。处于大数据时代，通过对用户特征建模、对项目属性建模、对用户行为建模、利用机器学习和数据挖掘等知识，结合物联网、云计算及大数据等技术而形成的混合推荐算法可能是未来的研究热点和方向。

参考文献

[1] 徐瑞朝，曾一昕 . 国内信息过载研究述评与思考 [J]. 图书馆学研究 .2017(18):21-25.

[2] 申辉繁 . 协同过滤算法中冷启动问题的研究 [D]. 重庆：重庆大学 ,2015:10-18.

[3]Lops P, de Gemmis M, Semeraro G Content-based recommender systems:State of the art and trends[M] // Recommender Systems Handbook. Springer US,2011:73-105.

[4]Debnath S, Ganguly N, Mitra P. Feature weighting in content based recommendation system using social network analysis［C］// Proceedings of the 17th international conference on World Wide Web. ACM,2008.1041-1042.

[5]Di Noia T, Mirizzi R, Ostuni V C, et al.Linked open data to support content-based recommender systems［C］// Proceedings of the 8th International Conference on Semantic Systems. ACM,2012.1-8.

[6] 王光，张杰民，董帅含，等 . 基于内容的加权粒度序列推荐算法 [J]. 计算机工程与科学 .2018,40(3):564-570.

[7] 吕学强，王腾，李雪伟，董志安 . 基于内容和兴趣漂移模型的电影推荐算法研究 [J]. 计算机应用研究 .2018,35(3):717-720.

[8]Agrawal R， Imieli ń ski T，Swami A． Mining association rules between sets of items in large databases［C］// Proceedings of the 1993 ACM SIGMOD international conference on management of data． New York: ACM，1993:207 － 216．

[9] 施海鹰 . 基于关联规则挖掘的分类随机游走算法 [J]. 计算机技术与发展 .2017,27(9):7-11.

[10] 张勇杰，杨鹏飞，段群，等 . 基于关联规则的商品智能推荐算法 [J]. 现代计算机 .2016(4):25-27.

[11] 谢振平，金晨，刘渊 . 基于建构主义学习理论的个性化知识推荐模型 [J]. 计算机研究与发展 .2018,55(1):125-138.

[12]Glodberg D,Nichols D,OkiB M,et al.Using collaborative filtering to weave an information tapestry[J].Communications of the ACM,1992,35(12):61-70.

[13] 吕杰，关欣，李锵，等 . 一种融合用户上下文信息和动态预测的协同过滤推荐算法 [J]. 小型微型计算机系统 .2016(8):1680-1685.

[14] 高茂庭，段元波 . 结合用户聚类和评分偏好的推荐算法 [EB/OL].[2017-07-21]. http://www.arocmag.com/article/02-2018-08-030.html.

[15] 龚敏，邓珍荣，黄文明 . 基于用户聚类与 Slope One 填充的协同推荐算法．计算机工程与应用 .2018.

[16] 韩皎 . 基于分布式平台的个性化推荐算法研究 [D]. 西安：长安大学 ,2017:43-60.

[17] 翁小兰，王志坚 . 协同过滤推荐算法研究进展 [J]. 计算机工程与应用 .2018,54(1):25-31.

[18] 朱扬勇，孙婧 . 推荐系统研究进展 [J]. 计算机科学与探索 ,2015,9(5):513-525.