

个性化推荐算法综述

石媛

(西安石油大学 计算机学院, 西安 710065)

摘要: 为了更好地解决信息超载的问题, 个性化推荐系统应运而生, 有效地解决了这一复杂问题。目前, 个性化推荐系统已经在国内得到了广泛的应用和关注。本文系统地研究了算法分类并总结了国际学术界的各种个性化推荐系统算法, 重点介绍了其优缺点和不同的推荐算法性能及其评价指标。提出了今后的研究重点、难点和推荐系统的一些热点问题。
关键词: 信息超载; 个性化推荐; 性能评价

A survey of personalized recommendation algorithms

SHI Yuan

(School of Computer Science, Xi'an Shiyou University, Xi'an 710065, China)

【Abstract】 In order to better solve the complex problem of information overload, personalized recommendation system comes into being and effectively solves this complex problem. At present, personalized recommendation system has been widely used and concerned in China. This paper systematically studies and summarizes various personalized recommendation system algorithms in the current international academic circles, summarizes the existing personalized recommendation system algorithms in the world, and focuses on its advantages and disadvantages, different recommendation algorithm performance and evaluation indicators. Finally, this paper focuses on the future research focus, difficulties and some hot issues of recommendation system.

【Key words】 information overload; personalized recommendation; performance evaluation

0 引言

随着互联网技术的普及和迅猛发展, 越来越多的数据积聚。然而面对数量巨大、种类繁多的信息, 要从中快速高效的利用和获取得到这些数据和信息简直就是大海捞针。这些反映了互联网在为广大用户提供海量数据和信息的同时所带来的问题和弊端。与此同时利用海量数据和信息的效率也有所降低, 这就是所谓的信息过载。

现如今, 从社交网络上直接获取海量信息资源已经非常容易, 但使用者面对海量的资源、信息的同时, 也受到了信息分散和对信息负担过重的因素影响。这也就意味着在许多情况下很难准确地找到那些能够真正满足实际需要的资料和信息。即使诸如搜狗、360、雅虎等众多知名搜索引擎已经可以很好地满足日常搜索需求, 但在进行搜索时也容易眼花缭乱, 很难准确地获取到自己所需要的资料和信息。目前, 个性化推荐算法系统是解决使用者信息过度负担和分散等问题的主要方法之一。

1 推荐系统的定义

学术界关于推荐系统的概念和定义研究已有很多版本。1997年Resnick和Varian提出“它是利用电子商务网站向客户提供商品信息和建议, 帮助用

户决定应该购买什么产品, 模拟销售人员帮助客户完成购买过程”^[1]。

文献[2-3]中对推荐系统的定义为: 设 C 是所有用户的集合, S 是所有可以推荐给用户的对象集合。设效应函数 $u(\cdot)$ 可以计算对象 S 对于一个用户 C 而言的产品推荐度和作用取决于一个产品的用户推荐度和产品可用性。即 R 通常是指在一定的推荐度数值范围内的一个具有全序非零或负值的实数。推荐度需要研究的关键问题是如何准确找到在推荐度一定范围内比 R 大的哪些人为用户, 该对象被称为用户 S^* , 如公式(1)所示:

$$\forall c \in C, S^* = \arg \max_{s \in S} u(c, s). \quad (1)$$

根据以上定义, 初步总结出推荐算法的核心有三大模块: 用户建模模块、推荐对象建模模块和推荐算法模块。

2 个性化推荐算法

个性化推荐系统包括的算法众多, 本文主要介绍了3种个性化推荐算法: 协同过滤推荐算法(Collaborative Filtering, 简称CF)、基于内容的推荐算法(Content-Based Recommendations CB)和基于社交网络推荐算法。针对3种推荐算法进行比较, 给出了各推荐算法的优缺点, 提出了个性化推荐算法的特点

作者简介: 石媛(1993-), 女, 硕士研究生, 主要研究方向: 数据挖掘。

收稿日期: 2020-06-11

哈尔滨工业大学主办 ● 系统开发与应用

以及个性化推荐系统相应的算法评价指标。

2.1 协同过滤推荐算法

协同过滤算法是普遍运用的一种算法,以协同过滤算法的比较对象为划分依据,能够将其划分为以下两种:一是基于用户的推荐算法。基于用户的推荐算法是一种过滤算法,用于根据目标用户的偏好进行定位和找到与目标用户感兴趣的用户群;另一种则是基于产品的推荐算法。基于产品的推荐算法包括计算产品的相似性和比较相似性,根据目标产品与现有用户的关系和现有行为数据来计算产品的相似性和比较相似性,最终为用户推荐提供高相似性的产品。

2.1.1 基于用户的推荐算法

基于用户的推荐算法需要分别计算用户之间的相似性和用户对当前产品的评分这两部分:其中计算用户间的相似性须要找到目前用户的最近邻。最近邻的寻找也有多种方法,目前常用的计算最近邻的算法有两种,一种是基于余弦算法来计算,另外一种就是基于关联算法进行计算^[4]。此外研究者们提出了主流加权预测和实例扩展等方法来实现推荐系统算法性能的提高。

2.1.2 基于产品的推荐算法

2003年,亚马逊推出了一种基于产品的推荐算法。该算法需要计算目标产品和该用户已购买过的或已评估的产品之间的相似性,并且预测用户感兴趣的产品。按照计算出来的产品相似性矩阵,将用户感兴趣的产品添加到推荐给用户产品的清单中。与基于用户的推荐算法相比,如果用户数量较少,使用基于用户的推荐方法预测效果会更好。尤其是在小型和普通型规模的电子商务网站表现优异。基于产品推荐算法的思想是只涉及产品不涉及用户,正因如此,基于产品的推荐算法可以根据产品的最近邻推荐提前计算好产品和产品之间的相似性,该算法可以提前在数据库中计算出相似产品的用户评分加权和,而后得出用户对产品的预测评分。当有用户需要推荐产品时,则在数据库找出与推荐过产品相似度高的产品,将其向用户推荐。在这种情况下,相似产品的数量和用户评分产品的数量有关,因为被用户评分过的产品数量通常相对较小,因此预测值可提前在线得出。

2.2 基于内容的推荐算法

该算法的基本思想如下:

步骤1 在具有目标用户的兴趣爱好信息的基础上,建立用户兴趣喜好模型。然后采用目前常用

的分类算法,例如支持向量机、决策树等分类器对用户的兴趣爱好信息进行分类

步骤2 对产品特征进行提取。依据推荐产品的特征,将产品的特征用 TF-IDF 算法(即词频-倒排文档频率法)进行提取。

步骤3 比较用户兴趣模型和产品特征,将其中相似度高的产品推荐给用户^[5]。

与基于用户的推荐算法比较,基于内容的推荐算法不需要计算用户之间的相似度和用户对产品的评分;与基于产品的推荐算法相比,则不需要计算产品之间的相似性。基于内容的推荐算法只需要学习用户兴趣模型,提取产品特征,比较用户与待推荐产品的匹配情况,推荐最适合的产品给用户即可^[6]。

2.3 基于社交网络的推荐算法

近来,互联网的快速发展让人类的社交方式由线下变成线上居多,智能手机的出现,让各个社交软件被广泛使用,不同行业、不同年龄、不同地区的人们利用社交软件形成了社交网络,使社交网络成为热点,基于社交网络的推荐已经成为研究的热门。该算法的核心思想是利用社交平台上记录到的用户信息,其中包括用户通讯录以及用户兴趣爱好等,根据社交网络上获取的信息向用户推荐好友、产品以及信息流的会话等。基于社交网络推荐算法可分为两类:一类是基于网络结构的社会化推荐,另一类是基于邻域的推荐^[7-8]。

基于邻域的推荐是利用社会网络将用户和朋友的关系数据、用户的历史浏览或购买产品轨迹以及用户的兴趣爱好结合起来,从而向目标用户推荐产品,而且推荐的产品是属于好友喜欢的产品集里。用户在通常情况下喜欢的产品大多都是自己熟悉的朋友所推荐的。因而算法必须要考虑用户和朋友的相熟度和喜好,用户 u 对产品 i 的兴趣 P_{ui} 可用公式(2)表示:

$$P_{ui} = \sum_{\mu \in f(u)} (f_{u\mu} + s_{u\mu}) r_{\mu i} \quad (2)$$

$f(u)$ 代表的是用户通讯录集; $f_{u\mu}$ 代表的是用户 u 与用户 v 之间的熟悉程度;代表用户 u 与用户 v 之间兴趣喜好的相似度; $s_{u\mu}$ 代表的是用户 v 对产品 i 的偏好。

基于网络结构的社会化推荐算法的表现方式可以分为两种:一种网络图是用户-产品二部图;另一种网络图是用户社交网络图。用户-产品二部图是用来表示用户对产品的行为;用户社交网络图是用来体现用户的社交网络行为。基于网络构造的社会

化推荐算法是将用户-产品二部图和用户社交网络图进行组合,通过收集到的数据(社交网络)组合成一个新的网络图^[9-10]。

按照用户与其朋友(用户对产品)的相熟程度和相似度,确定网络图中的权重系数,然后计算用户节点与产品节点的相关性,之后将相关性降序排列,把相关性高且用户没有选择过的产品生成产品列表

对用户进行推荐^[11]。

3 各类推荐算法的优缺点

每一种算法都有自身的优势和劣势,其中协同过滤推荐算法是目前运用最多的算法。协同过滤推荐算法是信息推荐领域中推荐效率最优的一种算法,该算法是基于内容对产品属性进行分析从而实现推荐。各类流行算法的优缺点见表 1。

表 1 各类推荐算法优缺点

Tab. 1 Advantages and disadvantages of various recommendation algorithms

推荐算法	优点	缺点
协同过滤推荐算法	推荐性能随时间的推移不断提高;能够向用户提供新的兴趣点;推荐个性化、自动化程度高;不需要领域知识,能够处理复杂的非结构化现象	用户-产品矩阵的稀疏性;可扩展性、冷启动问题;对用户的评分数据依赖性
基于内容的推荐算法	结果通俗易懂,可解释性强;不需要专业的领域知识背景;不需要用户评分数据	受新用户/新产品的限制;推荐结果缺乏惊喜“度”对推荐算法产生不良影响
基于社交网络的推荐算法	利用好友进行推荐增加了用户对推荐结果的信任度;有利于推荐长尾商品	在实际系统中难以获取用户好友数据

4 算法评价指标

常用的个性化评价指标为平均绝对误差、均方根误差、标准平均误差等。

(1) 平均绝对误差。平均绝对误差:用于衡量用户预测评分与实际评分之间的平均绝对误差,定义如式(3)所示:

$$M = \sum_{a=1}^n |p_{ia} - r_{ia}| / n. \quad (3)$$

(2) 均方根误差。均方根误差的定义如式(4)所示:

$$R_m = \sqrt{\sum_{a=1}^n |p_{ia} - r_{ia}|^2 / n}. \quad (4)$$

(3) 标准平均误差。标准平均误差的定义如式(5)所示:

$$N = M / (r_{\max} - r_{\min}). \quad (5)$$

其中, n 为平均用户 i 已经预测或测评过分的平均标准产品数量; M 是平均绝对误差; p_{ia} 与 r_{ia} 分别为已经预测的用户的平均绝对误差评分和真实的用户的平均绝对误差评分; R_m 为均方根误差; N 为平均标准误差; n_i 为系统中包含的用户与产品; r_{\max} 为用户评分的最大值; r_{\min} 为用户的平均绝对误差评分的最小值^[12]。

5 结束语

本文通过对目前流行的个性化推荐算法进行对比分析,介绍了不同推荐算法的优缺点并给出了相对应的评价指标。推荐系统的研究对经济和社会的重大意义为:

(1) 能够协助用户在知识过剩的网络环境中找到本身所需资料信息;

(2) 可以让电子商务网站通过增加合适的推荐算法网站进行完善。经过推荐系统,网站可以及时分析出用户感兴趣的产品并及时推荐给用户,不但留住了旧用户吸引了新用户同时也提高了网站销售。

推荐系统随着时间的推移逐渐出现了许多与推荐有关的问题,这些问题也引起了业界学者的注意。关于对推荐系统的进一步研究分别是:其一关于用户偏好获取方法和从推荐对象中提取特性的方法的研究;另外是关于推荐系统的安全性以及多维度推荐系统的多方面的研究。

参考文献

- [1] RESINICK P, VARIAN H R. Recommender systems[J]. Communications of the ACM, 1997, 40(3): 56-58.
- [2] ADOMAVICIUS G, TUZHLIN A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions[J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(6): 734-749.
- [3] SUNG H H. Helping online customers decide through Web personalization[J]. IEEE Intelligent Systems, 2002, 9: 34-43.
- [4] 崔春生, 吴祈宗, 王莹. 用于推荐系统聚类分析的用户兴趣度研究[J]. 计算机工程与应用, 2011, 47(7): 226-228.
- [5] 张帅, 贾年. 基于协同过滤的发型推荐算法[J]. 现代计算机(专业版), 2019(11): 49-52.
- [6] 石京京, 肖迎元, 郑文广. 改进的基于物品的协同过滤推荐算法[J]. 天津理工大学学报, 2019, 35(1): 32-36.
- [7] 时宇岑, 印莹, 赵宇海, 等. 基于多开发者社区的用户推荐算法[J]. 软件学报, 2019, 30(05): 1561-1574.
- [8] 徐吉, 李小波, 许浩. 基于用户信任的协同推荐算法研究与分析[J]. 数据通信, 2019(2): 29-34.
- [9] 刘辉, 郭梦梦, 等. 个性化推荐系统综述[J]. 常州大学学报(自然科学版), 2017, 29(3): 51-59.
- [10] 高凤丽, 孙连山. 个性化推荐系统概述[J]. 技术与市场, 2015, 22(2): 78-79.
- [11] 周万珍, 曹迪, 许云峰, 等. 推荐系统研究综述[J]. 河北科技大学学报, 2020, 41(1): 76-87.
- [12] 朱扬勇, 孙婧. 推荐系统研究进展[J]. 计算机科学与探索, 2015, (5): 513-525.