



“十三五”科学技术专著丛书

基于链路预测的推荐系统—— 原理、模型与算法

朱旭振 编著

Recommendation System Based on Link
Prediction — Principle, Model and Algorithm



北京邮电大学出版社
www.buptpress.com



“十三五”科学技术专著丛书

基于链路预测的推荐系统

——原理、模型与算法

朱旭振 编著



北京邮电大学出版社
www.buptpress.com

内 容 简 介

本书从复杂网络角度出发,研究基于相似性链路预测的协作推荐算法。本书主要面向广大的推荐算法研究者,希望能通过本书的介绍,帮助更多研究者步入推荐算法的研究之门。本书分为4部分:第1部分介绍复杂网络的基础知识以及网络分析软件 Pajek 的基本使用方法;第2部分介绍复杂网络上链路预测研究的一般方法、实验数据和性能指标,并给出笔者的几个研究实例;第3部分介绍基于链路预测的推荐算法研究,将一般网络上的链路预测研究思路扩展到二部图网络,基于物质扩散理论实现推荐系统建模,同时给出了笔者的几个研究实例;第4部分对本书进行总结,并对未来可能的研究方向进行展望。

本书不仅讲解了整体思路、单个问题的建模方法以及实验方法,还介绍了推荐系统建模的研究过程,抛砖引玉,注重引导新手入门。本书同时给出了大量实验数据、编程方法以及重要模块的代码,以期能铺石引路,以飨读者。

图书在版编目(CIP)数据

基于链路预测的推荐系统:原理、模型与算法 / 朱旭振编著. -- 北京:北京邮电大学出版社, 2018.9

ISBN 978-7-5635-5486-7

I. ①基… II. ①朱… III. ①互联网络—数据处理 IV. ①TP393.4

中国版本图书馆 CIP 数据核字(2018)第 140667 号

书 名: 基于链路预测的推荐系统——原理、模型与算法

著作责任者: 朱旭振 编著

责任编辑: 徐振华 孙宏颖

出版发行: 北京邮电大学出版社

社 址: 北京市海淀区西土城路 10 号(邮编:100876)

发 行 部: 电话: 010-62282185 传真: 010-62283578

E-mail: publish@bupt.edu.cn

经 销: 各地新华书店

印 刷: 北京玺诚印务有限公司

开 本: 720 mm×1 000 mm 1/16

印 张: 12.5

字 数: 244 千字

版 次: 2018 年 9 月第 1 版 2018 年 9 月第 1 次印刷

ISBN 978-7-5635-5486-7

定 价: 38.00 元

· 如有印装质量问题,请与北京邮电大学出版社发行部联系 ·

前 言

飞速发展的计算机、互联网和 Web 技术改变了人们的生活,人们在虚拟社区中结交好友,在新闻网站中浏览新闻,在视频网站中观看电影,在虚拟图书馆中查阅书籍,在电商平台中购买物品。但是,人们在享受多彩生活的同时也感受到了信息爆炸带来的不便,即人们无法在海量数据中快速有效地找到最相关的信息。电影、书籍、网页等信息的数据量动辄以千万级,这些数据信息的增长速度已经远远超过了人类的自然处理能力。在这种大数据的背景下,用户获取所需信息的代价越来越大,仅仅依靠传统人力的方式已经无法评价和选择这些物品。在这种情况下,有效过滤海量信息的最有吸引力的方法就是个性化推荐技术。它利用用户个人信息,如用户活动的历史记录,发现用户喜好,然后根据用户喜好进行推荐,例如,Amazon 利用用户的购买历史记录向用户推荐书籍,AdaptiveInfo 利用用户的阅读历史向用户推荐新闻,TiVo 数字视频系统根据用户的观看模式和评分记录向用户推荐电视节目。

在广泛经济价值和社会意义的吸引下,众多研究者提出了多样的推荐算法,如基于内容的推荐、基于知识的推荐、基于关联规则的推荐、基于效用的推荐等。基于协作的推荐算法由于其对新奇兴趣的发现不需要领域知识,推荐个性化、自动化程度高,并且能处理复杂的非结构化对象,受到了广泛关注。而在协作推荐算法中,基于链路预测相似性的协作推荐算法由于其简单性、高效性和准确性得到了广泛关注。

本书从单一节点网络上的链路预测研究入手,研究端点间影响相似性的拓扑因素,并进一步基于超图理论和物质扩散理论,将研究结果扩展至二部图网络上,对二部图网络物品间的链路预测进行建模,发现物品间的相似性,结合协作技术完成推荐。本书首先介绍基础知识,使得读者对复杂网络有基本的认识,并介绍复杂网络分析工具 Pajek;其次介绍一般网络上单一节点间的链路预测研究;再次介绍二部图网络上基于链路预测的协作推荐研究;最后进行总结并展望未来的研究方向。

本书采用问题描述、理论建模、数据仿真、性能计算的方法介绍各个实例的研究思路,通过笔者的研究举例,针对每个研究点介绍研究方法,并给出此项研

究的参考文献,同时引导读者思考未来可能的研究思路。通过介绍各个研究案例,可以帮助读者快速进入未来的研究课题。

本书的撰写目的是通过理论介绍、实例讲解和代码分析引导感兴趣的读者尽快入门,同时希望解决读者存在的疑惑。本书适合于专注链路预测、推荐算法研究和理论建模的广大研究工作者、算法工程师和软件开发人员使用。希望本书能成为读者的进步阶梯和良师益友。由于笔者水平有限,本书难免存在偏颇之处,敬请读者多提宝贵意见。

朱旭振
北京邮电大学

目 录

第 1 部分 基础知识

第 1 章 绪论	3
1.1 研究背景	3
1.1.1 推荐系统的发展现状及特征分析	3
1.1.2 推荐系统的国内外研究现状	7
1.2 相关理论基础	10
1.2.1 复杂网络理论基础	10
1.2.2 链路预测理论	13
1.2.3 基于链路预测的协同推荐理论	13
1.3 复杂网络下基于链路预测推荐所面临的问题及研究意义	14
1.3.1 面临的问题	14
1.3.2 研究意义	17
1.4 研究思路	17
1.5 本书的主要内容	18
本章参考文献	21

第 2 部分 复杂网络上的链路预测方法

第 2 章 网络分析软件 Pajek	29
2.1 Pajek 软件介绍	29
2.1.1 高速计算	30
2.1.2 可视化	30
2.1.3 抽象化	30
2.2 Pajek 软件使用基础	30
2.3 Pajek 软件分析网络属性	32
2.3.1 度的计算	33

2.3.2	两点间的距离	33
2.3.3	k 近邻	34
2.3.4	聚类系数	35
2.4	Pajek 软件抽取极大连通子图	36
2.5	Pajek 软件网络画图	36
2.5.1	绘制复杂网络图	36
2.5.2	绘制不同类节点的复杂网络图	37
2.5.3	绘制不同大小节点的复杂网络图	38
2.5.4	绘制不同权值边的复杂网络图	38
2.6	网络文件 .net 简介	38
2.6.1	Pajek 网络文件的一般结构	39
2.6.2	具体参数的意义和取值	39
2.6.3	文件举例	41
2.7	本章小结	43
	本章参考文献	44
第 3 章	基于相似性的链路预测研究	45
3.1	链路预测的研究方法	45
3.2	链路预测的典型研究成果	45
3.3	链路预测的实验数据	46
3.4	链路预测的实验方法	47
3.4.1	数据集划分方法	47
3.4.2	链路预测的度量指标	47
3.5	链路预测重要代码讲解	48
3.5.1	数据集划分代码讲解	48
3.5.2	关键测试指标代码讲解	51
3.6	基于拓扑相似性链路预测的思考	53
3.7	本章小结	53
	本章参考文献	53
第 4 章	基于弱关系的链路预测算法	56
4.1	研究背景	56
4.2	问题描述	56
4.3	基于弱关系的优化链路预测模型	57
4.3.1	CN 算法、AA 算法和 RA 算法介绍	57

4.3.2 改进优化算法模型	58
4.4 实验结果与分析	59
4.4.1 数据集	59
4.4.2 度量指标	60
4.4.3 结果与分析	60
4.5 本章小结	63
4.6 研究思考	64
本章参考文献	64
第5章 基于路径异构性的链路预测算法	66
5.1 研究背景	66
5.2 问题描述	67
5.3 基于路径异构性的链路预测建模	67
5.3.1 SP 模型	68
5.3.2 对比算法	69
5.4 实验结果与分析	70
5.4.1 数据集	70
5.4.2 评估准则	71
5.4.3 结果与分析	71
5.5 本章小结	74
5.6 研究思考	74
本章参考文献	75
第6章 基于端点影响力的链路预测算法	80
6.1 研究背景	80
6.2 问题描述	81
6.3 基于端点影响力建立链路预测模型	82
6.3.1 EP 模型	82
6.3.2 对比算法	83
6.4 实验结果与分析	84
6.4.1 数据集	84
6.4.2 评估准则	86
6.4.3 结果与分析	86
6.5 本章小结	90
6.6 研究思考	91
本章参考文献	91

第3部分 基于链路预测的推荐算法研究

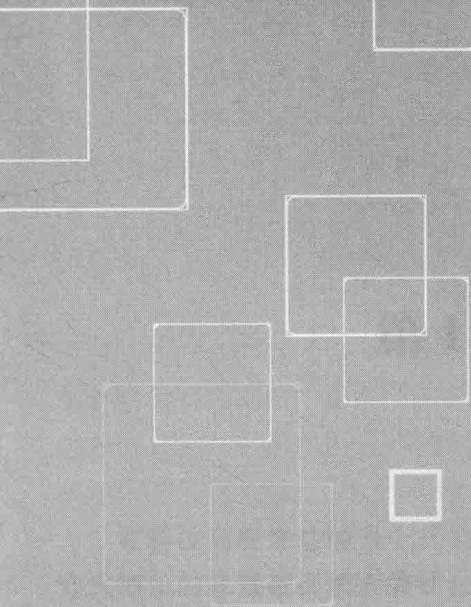
第7章 推荐模型的研究方法	99
7.1 推荐模型常见研究方法	99
7.2 基于链路预测的推荐模型研究方法	100
7.3 推荐技术的典型研究成果	101
7.4 推荐技术的研究数据介绍	101
7.5 推荐实验方法	102
7.5.1 数据集划分方法	102
7.5.2 推荐算法的度量指标	102
7.6 推荐算法重要代码讲解	104
7.6.1 数据集划分代码讲解	105
7.6.2 推荐算法关键指标代码讲解	108
7.7 基于二部图推荐算法的研究思路	111
7.8 本章小结	112
本章参考文献	112
第8章 基于修正相似性的协作推荐算法	114
8.1 研究背景	114
8.2 问题描述	115
8.3 基于修正相似性的推荐算法 CSI	116
8.3.1 基于二部图网络的经典相似性算法	117
8.3.2 相似性修正模型 CSI	117
8.3.3 对比算法	118
8.4 实验结果与分析	119
8.4.1 数据集	120
8.4.2 评价准则	120
8.4.3 结果与分析	122
8.5 本章小结	125
8.6 研究思考	126
本章参考文献	126
第9章 基于一致性的协作推荐算法	131
9.1 研究背景	131

9.2	问题描述	132
9.3	基于一致性的推荐算法 CBI	133
9.3.1	基于网络的因果性推荐算法 NBI	134
9.3.2	基于一致性的推荐算法 CBI 和 UCBI	134
9.3.3	对比算法	135
9.4	实验结果与分析	136
9.4.1	数据集	137
9.4.2	评价准则	137
9.4.3	结果与分析	139
9.5	本章小结	143
9.6	研究思考	144
	本章参考文献	144
第 10 章	基于一致性冗余删除的协作推荐算法	148
10.1	研究背景	148
10.2	问题描述	148
10.3	修正冗余删除推荐算法	149
10.3.1	相似性估计偏差现象	149
10.3.2	相似性冗余问题	150
10.3.3	修正冗余删除相似性指标 CRE	150
10.3.4	对比算法	151
10.4	实验结果与分析	153
10.4.1	数据集	153
10.4.2	评价准则	154
10.4.3	结果与分析	156
10.5	本章小结	160
10.6	研究思考	161
	本章参考文献	161
第 11 章	一致性下基于惩罚过度扩散的推荐算法	165
11.1	研究背景	165
11.2	问题描述	166
11.3	对称和过度扩散惩罚算法模型	166
11.3.1	非对称扩散问题	167

11.3.2	扩散冗余问题	168
11.3.3	基于对称的过度扩散惩罚模型	169
11.3.4	对比算法	170
11.4	实验结果与分析	172
11.4.1	数据集	172
11.4.2	评价准则	172
11.4.3	结果与分析	174
11.5	本章小结	178
11.6	研究思考	178
	本章参考文献	179

第 4 部分 总结与未来展望

第 12 章	总结和展望	185
12.1	总结	185
12.2	未来研究展望	188



第1部分

基础知识

第1章 绪 论

在科技高度发达的当今社会,随着计算机、互联网和智能终端的广泛应用,爆炸式增长的海量数据使人类社会进入了大数据时代。新的时代让人们的生活和工作方式发生了重大变化,尤其是在现实生活和工作中,网络环境占据了重要地位,帮助人们结识好友、在线交易等,给人们带来了极大便利,但同时也给人们的工作和生活带来了挑战。不断累积的信息已经远远超出了人们的处理能力,导致了一个尴尬的困境:人们面对着丰富的资源,却无法快速有效地找到急需的信息。幸运的是,人们的活动信息和交互信息已经实现了有效存储,利用存储的信息,研究人员可以研究高效的信息推荐系统,但是在研究推荐算法时,仍然存在很多问题:是否可以利用用户历史信息构建复杂网络,并借用一般网络下的链路预测方法发现相似性,进而完成推荐;利用何种链路预测方法可以更加有效地发现相似性;由于用户数据的稀疏性和非对称性,相似性估计是否存在偏差,如何修正;基于用户历史进行相似性推荐,对因果性的假设是否合理;基于相似性的推荐模型中是否存在相似性冗余;对个性化推荐而言,用户的多样化偏好和物品的高度流行性之间有什么关系,等等。这些都是研究高效推荐系统所必须解决的重要问题。

本章首先介绍了推荐系统的研究背景以及当前国内外研究现状;然后针对目前链路预测和基于链路预测的推荐研究,分析了所面临的挑战,并简要介绍了与本书相关的研究基础和理论方法,包括复杂网络理论、基于复杂网络的链路预测理论以及基于链路预测的推荐理论;最后介绍了作者在研究中所发现的重要问题、研究内容及研究思路,并给出了本书的组织结构和章节顺序。

1.1 研究背景

1.1.1 推荐系统的发展现状及特征分析

1. 推荐系统的概念

根据百度百科定义^[1]:“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的系统。”推荐系统由3个重要部分组成:用户建模部分、物品建模部分和推荐算法部分。但是从本质上讲,推荐系统是在研究两个物品对象相关性的基础上,对未来用户选购可能性做出预测的系统。

根据对象节点是否单一,网络可分为:单一类别对象网络(例如传感器网络、通信网络、Internet网络、航空网络、电站网络、科学家合作网络、生物链网络、蛋白质网络等)和两重类别的二部图网络(例如常见的用户-物品网络等)。在前者网络中,基于单一类别对象间相关性,预测对象间发生连接可能性的算法被称为链路预测算法;而在后者网络中,基于物品和用户间相关性,预测未来用户选购物品可能性的算法被称为推荐算法。这两种算法之间存在着紧密关联,可以基于单一类别网络中的链路预测算法,提出更加有效的推荐算法。

2. 推荐系统的发展及现状

推荐系统起源于当今迅猛发展的计算机技术、互联网技术和Web技术。这些技术的快速发展不断改变着人们的生活,也累积了海量的数据:数以百万计的电影和音乐,数以十亿计的在线商品,数以万亿计的网页等。人们逐渐地从信息匮乏走向了信息过载^[2],进入了大数据时代。

为了解决信息过载问题,科学家和学者们提出了众多有效的解决方案,其中最具有代表性的就是分类目录和搜索引擎。这两种解决方案分别催生了互联网领域两家著名公司——雅虎和谷歌。目前比较著名的分类目录网站有国外的Lply、国内的Hao123等^[3],这些分类目录网站将网站信息分门别类,从而方便用户根据类别查找网站。但是随着互联网规模不断扩大,分类目录网站也仅能覆盖少量热门网站,愈来愈不能满足用户的信息需求。此时,搜索引擎应运而生,以谷歌、百度为代表的搜索引擎,可以以关键词搜索的方式完成信息检索。但是若要搜索引擎实现有效检索,精准的关键词将是必要的前提。当用户无法找到准确的关键词,或者用户根本没有明确需求时,搜索引擎就显得力不从心。

鉴于搜索引擎能力的不足,研究者们设计出了新的信息获取工具——推荐系统。与搜索引擎相比,推荐系统也是一种帮助用户快速获取信息的工具。不同的是,推荐系统并不需要用户提供明确的需求,而是通过分析用户的历史行为,发现用户的兴趣爱好,从而主动地向用户推荐他们感兴趣的信息。因此,推荐系统和搜索引擎可谓是两个互补的工具,搜索引擎可以满足用户有明确目的的信息检索,而在用户没有明确目的时,推荐系统能够主动推荐信息,尤其在商品推销中,这种推荐的作用尤其显著。通常热销商品仅占商品总数的一小部分,而非热销商品数量巨大,造成了所谓的“长尾效应”^[4]。这些非热销商品被称为长尾商品,它们的总销售额不可小觑,甚至在很多时候远远超过热门商品,更重要的是长尾商品往往能满足用户的个性化需求。通过分析用户偏好,推荐系统

可以提高长尾商品的销售量。

推荐系统的雏形是在1995年3月美国人工智能协会上,由卡耐基梅隆大学的 Robert Armstrong 等人和斯坦福大学的 Marko Balabanovic 等人分别提出的个性化导航系统 Web Watcher 及个性化推荐系统 LIRA^[1]。到目前为止,推荐系统已广泛应用于众多领域:在电子商务领域,基于用户浏览记录,著名的 Amazon 和阿里巴巴公司都推出了高效的推荐系统;在视频推荐领域,基于用户对视频的大量浏览记录,Netflix、YouTube、优酷、爱奇艺等公司推出了主动式推荐服务;在个性化音乐推荐领域,著名的个性化音乐推荐软件有 Pandora、Last.fm、豆瓣电台、酷我音乐盒等,它们从用户听歌历史行为中得到用户的兴趣模型,从而向用户推荐歌曲;在社交网站领域,Facebook、Twitter、人人网、新浪微博等建立虚拟社交社区,利用推荐技术推荐好友、推荐物品,甚至是推荐会话,不仅如此,丰富的社交网络关系和用户偏好数据与其他领域的应用积极结合,以实现协作推荐,例如 Facebook 与 Amazon 的结合,阿里巴巴与新浪微博的结合;在个性化阅读推荐方面,著名的 GoogleReader、鲜果网等主动地向用户推荐感兴趣的文章,解决了在海量文章中,用户快速检索的问题,而且随着移动设备的广泛普及,推荐系统可以有效地实现在移动设备上的个性化阅读;在基于位置服务推荐方面,基于地理位置信息和用户的兴趣爱好,著名的 Foursquare 公司、玩转四方公司向用户推荐感兴趣的信息,例如餐馆、影院等;在个性化广告投放领域,在国外最成功的就是 Facebook,而国内比较知名的有百度,他们根据用户的兴趣爱好,变传统轰炸式广告投放为精准投放,使用户愉悦地阅读感兴趣的广告。

由于人们生产生活的现实需求,推荐系统具有广泛的应用价值,在研究者们积极的推动下,推荐系统逐渐成为一个独立的研究领域。

3. 推荐系统的特征分析

(1) 以发现用户偏好为中心

推荐系统向人推荐,根本在于发现用户潜在的兴趣偏好,向用户推荐感兴趣的物品,这里用户偏好是推荐核心,用户偏好模型是推荐系统设计的关键。

(2) 主动式推荐

对于用户而言,用合适的关键词描述喜好有时比较困难,而且用户在很多时候是漫无目的的浏览,因此需要推荐系统提供主动式推荐。当今信息技术迅猛发展,为人们提供各种服务的网络应用层出不穷,人们也有了相比以往更大的选择空间,网络应用若要吸引用户并让用户产生忠诚感和信赖感,其推荐系统应具备发现用户兴趣并主动推荐的能力。用户浏览网络应用,并不一定会主动搜索其中的信息,若要增加用户的驻留时间,必须能发现用户的兴趣偏好,同时尽最大可能主动向用户推荐令其感兴趣的内容,唯有如此才能真正达到第一时间发现用户、第一时间吸引用户、第一时间服务用户,进而长久吸引用户的目的。

(3) 推荐的准确性依赖于信息的充分性和有效性

推荐系统无须用户通过关键词来寻找感兴趣的物品,而是依据用户浏览或购买历史,主动分析用户的兴趣爱好,估计物品相似性并实现协作推荐。当向用户进行推荐时,以用户行为记录为基础,估计用户尚未选购物品与以往购买物品之间的相似性,推断用户比较感兴趣的物品,进而按照相似程度推荐给用户。推荐的关键问题是在无用户偏好信息输入的情况下,主动发现偏好,在这种情况下,推荐系统推断用户的兴趣偏好就要以充分的历史信息为基础,例如,物品的描述信息,物品被多人同时购买的信息,物品与物品同时被购买的信息等。信息不仅要充分,而且还要有很高的质量,即没有冗余、缺失或错误。信息质量的好坏至关重要,可以想象,如果推荐系统缺乏足够的有效信息,又怎么能够做出合理的推荐呢?

(4) 应用领域广泛,实现技术与应用相关

从产生到现在,推荐系统已经历了漫长的发展过程,在众多领域都有广泛应用,例如,电影和电视剧推荐、音乐推荐、书籍推荐、好友推荐、广告推荐等。在不同领域,物品类别、描述方式、详尽程度和用户使用方式等都有较大差异,因此,在不同领域进行物品推荐,关键要因地制宜,具体问题具体分析,运用适当技术进行推荐,例如,对于电影、歌曲等抽象多媒体物品,通常缺乏足够的描述信息,经常采用协作式推荐;而对于文档等直接用文字描述的物品,含义直白明了,易于采用基于内容的精确推荐。虽然在不同领域,所要解决的推荐问题各有侧重,但都会遇到除用户选购记录外,缺乏其他相关信息的难题。此时可以利用普适的协作推荐技术来解决,并根据应用领域的相关特征,设计具有特征适应性的协作推荐算法。

(5) 多信息联合,综合推荐

前面已经提到,准确推荐需要充分有效的信息,但是在实际推荐应用中,还需要综合多种信息才能准确地完成推荐,例如,传统电影推荐系统是基于用户评分和浏览记录的协作推荐系统。如果在此基础上,能同时综合利用电影简介、评价或分类标签等信息,将能进一步准确地了解用户的兴趣爱好,提高推荐的准确性。因此,在研究实际推荐算法时,基于多样信息的发掘,实现综合推荐是设计推荐系统的有效思路。

(6) 多样的建模方法

通常推荐系统应用环境差异较大,在不同环境中,选择合适的方法进行建模就显得尤为关键。推荐系统由三部分组成:用户建模部分、物品建模部分和推荐算法部分。这3个部分的建模是整个建模方法的核心,与推荐环境和应用紧密相关。在推荐环境中,如果文字信息描述比较充分,则可以选择基于文本内容的推荐算法;如果物品之间、用户之间及用户与物品之间的关系描述比较明确,可以依据知识逻辑推断未来用户可能选购的物品,则可以采用基于知识的推荐算法;如果只有用户对物品的选购历史记录,则采用协作推荐比较恰当,并且协作推荐可以基于向