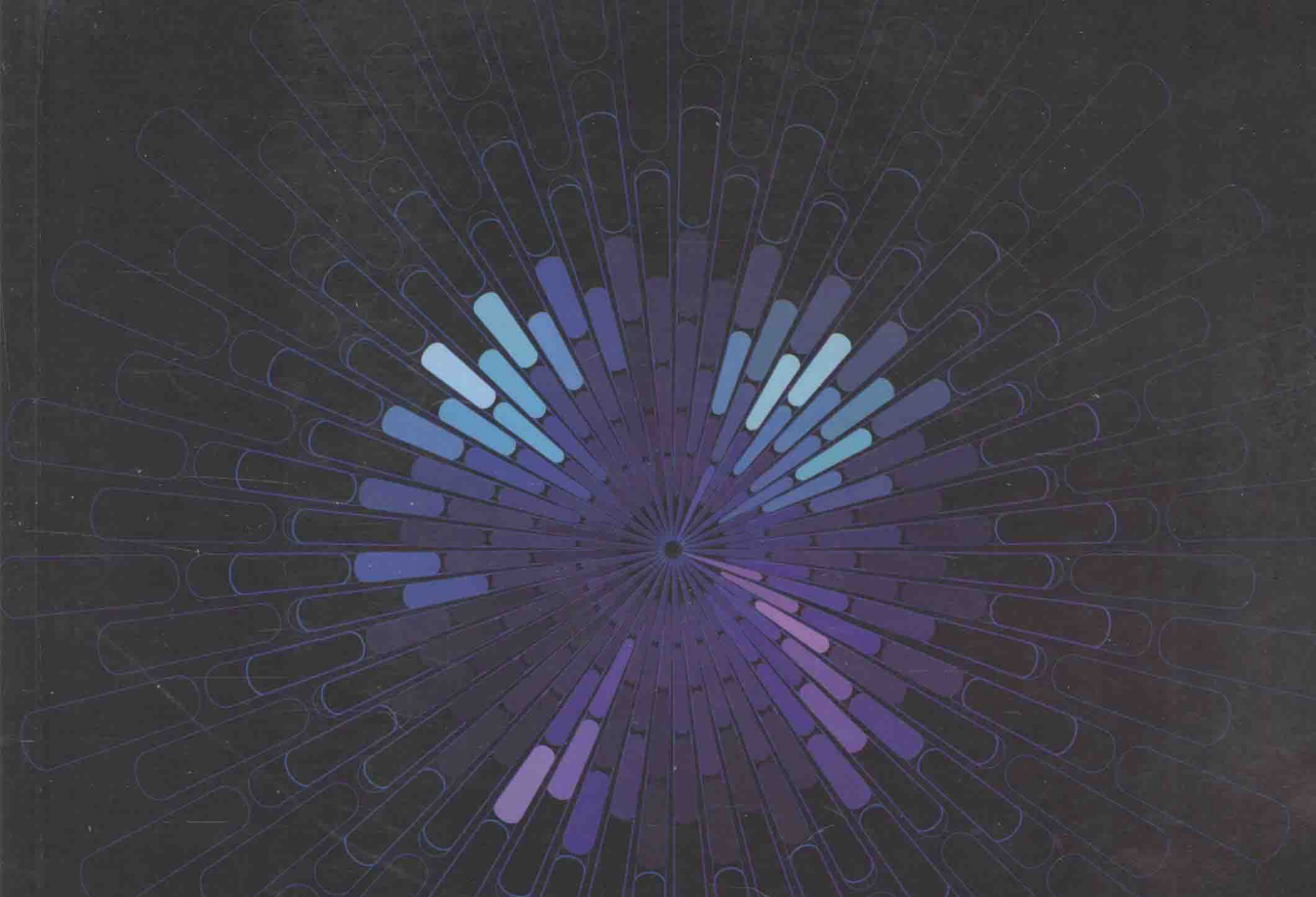

用户网络行为画像

大数据中的用户网络行为画像分析与内容推荐应用

牛温佳 刘吉强 石川 等著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

用户网络行为画像

大数据中的用户网络行为画像分析与内容推荐应用

牛温佳　刘吉强　石　川
康翠翠　童恩栋　诸　峰　著
覃毅芳　管洋洋　[澳]李刚

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

如何能牢牢地黏住老用户、吸引新用户、读懂用户的偏好兴趣和喜怒哀乐，这都是对企业发展至关重要甚至关乎生死存亡的问题，解决这个问题的方法就是推荐系统。本书分为上中下三篇，共 13 章，上篇为用户画像知识工程基础，包括表征建模、画像计算、存储及各种更新维护等管理操作；中篇为推荐系统与用户画像，包括传统协同过滤等经典推荐算法的介绍，以及涉及用户画像的推荐方法；下篇为应用案例分析，包括 Netflix、阿里等数据竞赛的经典数据案例，以及在具体工程开发过程中的具体案例，分别从系统需求、总体结构、算法设计、运行流程及测试结果等五个方面提供详细案例指导。

本书适合从事互联网工作的人员阅读，也作为相关专业的教材。

图书在版编目（CIP）数据

用户网络行为画像：大数据中的用户网络行为画像分析与内容推荐应用 / 牛温佳等著. —北京：电子工业出版社, 2016.3

ISBN 978-7-121-28070-2

I. ①用… II. ①牛… III. ①互联网络—研究 IV. TP393.4

中国版本图书馆 CIP 数据核字（2016）第 010289 号

责任编辑：田宏峰
印　　刷：北京京师印务有限公司
装　　订：北京京师印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编　100036
开　　本：787×980　1/16　印张：14.75　字数：330 千字
版　　次：2016 年 3 月第 1 版
印　　次：2016 年 3 月第 1 次印刷
印　　数：3 000 册　　定价：59.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

前　言

随着大数据时代的到来，互联网企业的竞争已经到了寸土必争和群雄逐鹿的时代，如何能牢牢地黏住老用户、吸引新用户、读懂用户的偏好兴趣和喜怒哀乐，这都是对企业发展至关重要甚至关乎生死存亡的问题。而为了解决这个问题，学界和业界一直是蛮拼的，积极从各个领域汲取理论，从人工智能、智能信息处理，细化到数据挖掘和机器学习，再后来就有了一个更加专用的术语——推荐系统。通俗地讲，推荐就是发掘用户集合和对象集合的语义关系，为用户提供语义最相关的 TOP-N 对象集合，而语义关系就是能读懂用户偏好兴趣的核心。因此，从业务来看，推荐系统是面向具体业务的交叉研究，无业务讲推荐系统，感觉言之无物；从技术来讲，没有永远一招鲜的技术，不同的数据、不同的场景就会有不同的结果；而从知识面上讲，涉及的技术非常广泛，可以大胆预言，推荐系统的研究还会包含更多其他领域的技术，因此是无止境的。

推荐系统在很多实际应用中已经被成功的开发和利用，例如 Amazon 和淘宝的猜你喜欢买、优酷等视频网站的猜你喜欢看，甚至你在某网站搜索二手房后，你在逛某个论坛时，这些房屋的广告都会追过来，推荐系统及相关技术如影随形。很多大公司专门建立一个独立的推荐系统研发或者数据分析师团队，旨在提高服务智能化和拓展企业利润空间，同时也可以大大增加用户的满意度。而小公司或者初创企业，其实也迫切需要推荐系统，但是往往会遇到投入成本过大的问题。这也是本书的一个初衷，希望可以帮助一些小企业技术人员快速地理解和部署简单的推荐系统，以用户画像为核心，对相关算法有个初步理解和入门。而对于一些高深的推荐算法研究，我们不敢在国内外顶级学者的算法研究面前班门弄斧，更希望从推荐服务提供商的角度多畅谈一下对用户画像的理解，对常见算法有个普及型和稍微深入的介绍，就已经达到本书的目的了。但实际上，用户画像其实是一个比较抽象的概念，粒度如何控制？是给一群人打上文艺男的标签，还是直接给单个人打上文艺微胖男或者文艺知性女的标签？标签间的关系是什么？一直喜欢看文艺电影的，此时此刻就一定想看文艺电影，是否一定要推荐文艺电影，还是推荐排行榜的美国大片效果更好？如何追求大客户和小客户的体验差异化权衡（大客户小个体模型，小客户大群体模型）？这些都是特别有意思且值得深入研究的地方。但是从通用业务的角度，只要在统计方面发现用户的黏性增加，广告的单击率和转化率提升，这就算一个上线产品的基本成功点了，已经具备可以继续深入优化的基础。

目前市场上的相关书籍，将用户画像的描述或者隐藏在具体的算法中，或者简单以用户偏好的形式带过，往往不是从单独系统性的角度阐述的，或多或少导致用户知道用户画像的意思，但是一方面理解起来深度不够（如用户每个时段的观影稳定性定量是多少），另一方面不知道如何存储、表示和实际使用。因此，本书希望言之有物，以视频网站的用户画像为切入点，在广度上也会覆盖主流常见的推荐算法原理和技术介绍，给出了如何使用面向用户画像的高级推荐算法，并且通过具体案例的详细描述和数据测试流程，对读者的理解与实践产生积极的指导意义。

本书侧重针对视频的个性化推荐系统相关技术，重视对以用户画像为核心的牵引，重视实际操作，点面结合，尤其是借鉴了我们在产业界做的一些具体线上项目流程和实施代码，力求对推荐系统的持续发展提供借鉴和参考价值，贡献绵薄力量。特别需要指出的是，在实践部分，我们不会特别纠结算法的准确率（因为有了基础推荐系统后可以对插件化的算法不断改进和优化），而是重点叙述用什么开源模型，怎么快速搭建起来，有哪些基本配置和模块，关键画像模块怎么构建；很多基本数据，怎么接入系统，怎么用；推荐怎么输出，输出数据是什么，怎么用；结合我们的服务器时间，对数据处理规模和推荐时间性能给出基本的参考。

本书分为上中下三篇，共 13 章。上篇为用户画像知识工程基础，包括表征建模、画像计算、存储及各种更新维护等管理操作；中篇为推荐系统与用户画像，包括传统协同过滤等经典推荐算法的介绍，以及涉及用户画像的推荐方法；下篇为应用案例分析，包括 Netflix、阿里等数据竞赛的经典数据案例，以及我们在具体工程开发过程的具体案例，分别从系统需求、总体结构、算法设计、运行流程及测试结果五个方面提供详细案例指导。

最后，本书虽涉及视频推荐系统的关键技术和相应的详细应用分析，仍难以详尽叙述理论和工程实现的方方面面。由于作者水平有限，不足之处在所难免，敬请广大读者批评指正，欢迎及时与出版社或作者联系，我们将会及时在下一版中予以更新及补充。

本书由中科院信息工程研究所郭莉、谭建龙担任顾问，在编写过程中得到了中科院信息工程研究所刘萍、胡玥、王斌、刘庆云、时金桥、熊刚，北京交通大学闫子淇，北京邮电大学刘军、郑静，澳大利亚 Deakin 大学 Tianqing Zhu、Shaowu Liu，360 奇虎科技的燕凯等各位学者和工程师的帮助和支持，在此向他们表示由衷感谢。在实验环境方面，感谢北京云量数盟给予的实验支持与帮助，感谢辛苗、牛奕涵对本书内容的启发与指导，需要感谢的人太多，也特别感谢和致敬该领域的著名专家学者项亮。此外，本书中的部分内容参考了相关互联网电商企业的推荐系统公开技术资料，再次感谢他们的精彩分享。

作　者

2016 年 1 月

目　录

上　篇

第 1 章　用户画像概述……3
1.1　用户画像数据来源……3
1.1.1　用户属性……5
1.1.2　用户观影行为……5
1.2　用户画像特性……5
1.2.1　动态性……5
1.2.2　时空局部性……6
1.3　用户画像应用领域……6
1.3.1　搜索引擎……6
1.3.2　推荐系统……7
1.3.3　其他业务定制与优化……7
1.4　大数据给用户画像带来的机遇与挑战……8

第 2 章　用户画像建模……9
2.1　用户定量画像……9
2.2　用户定性画像……10
2.2.1　标签与用户定性画像……10
2.2.2　基于知识的用户定性画像分析……12
2.2.3　用户定性画像的构建……16
2.2.4　定性画像知识的存储……22
2.2.5　定性画像知识的推理……26
2.3　本章参考文献……29

第 3 章　群体用户画像分析……31
3.1　用户画像相似度……32
3.1.1　定量相似度计算……32
3.1.2　定性相似度计算……34

3.1.3 综合相似度计算 …… 35
3.2 用户画像聚类 …… 36

第 4 章 用户画像管理 …… 41
4.1 存储机制 …… 41
4.1.1 关系型数据库 …… 42
4.1.2 NoSQL 数据库 …… 43
4.1.3 数据仓库 …… 45
4.2 查询机制 …… 46
4.3 定时更新机制 …… 47
4.3.1 获取实时用户信息 …… 47
4.3.2 更新触发条件 …… 48
4.3.3 更新机制 …… 49

中 篇

第 5 章 视频推荐概述 …… 55
5.1 主流推荐方法的分类 …… 56
5.1.1 协同过滤的推荐方法 …… 56
5.1.2 基于内容的推荐方法 …… 57
5.1.3 基于知识的推荐方法 …… 59
5.1.4 混合推荐方法 …… 60
5.2 推荐系统的评测方法 …… 61
5.3 视频推荐与用户画像的逻辑关系 …… 61

第 6 章 协同过滤推荐方法 …… 65
6.1 概述 …… 65
6.2 关系矩阵及矩阵计算 …… 67
6.2.1 U-U 矩阵 …… 67
6.2.2 V-V 矩阵 …… 70
6.2.3 U-V 矩阵 …… 72
6.3 基于记忆的协同过滤算法 …… 74
6.3.1 基于用户的协同过滤算法 …… 75
6.3.2 基于物品的协同过滤算法 …… 78
6.4 基于模型的协同过滤算法 …… 81

6.4.1 基于隐因子模型的推荐算法 82
6.4.2 基于朴素贝叶斯分类的推荐算法 85
6.5 小结 88
6.6 本章参考文献 88

第 7 章 基于内容的推荐方法 91
7.1 概述 91
7.2 CB 推荐中的特征向量 94
7.2.1 视频推荐中的物品画像 94
7.2.2 视频推荐中的用户画像 96
7.3 基础 CB 推荐算法 97
7.4 基于 TF-IDF 的 CB 推荐算法 99
7.5 基于 KNN 的 CB 推荐算法 102
7.6 基于 Rocchio 的 CB 推荐算法 104
7.7 基于决策树的 CB 推荐算法 106
7.8 基于线性分类的 CB 推荐算法 107
7.9 基于朴素贝叶斯的 CB 推荐算法 109
7.10 小结 111
7.11 本章参考文献 111

第 8 章 基于知识的推荐方法 113
8.1 概述 113
8.2 约束知识与约束推荐算法 114
8.2.1 约束知识示例 114
8.2.2 约束满足问题 115
8.2.3 约束推荐算法流程 117
8.3 关联知识与关联推荐算法 118
8.3.1 关联规则描述 118
8.3.2 关联规则挖掘 121
8.3.3 关联推荐算法流程 123
8.4 小结 124
8.5 本章参考文献 124

第 9 章 混合推荐方法 125
9.1 概述 125

9.2 算法设计层面的混合方法 …… 126
9.2.1 并行式混合 …… 126
9.2.2 整体式混合 …… 129
9.2.3 流水线式混合 …… 131
9.2.4 典型混合应用系统 …… 133
9.3 混合式视频推荐实例 …… 136
9.3.1 MoRe 系统概览 …… 136
9.3.2 MoRe 算法介绍 …… 137
9.3.3 MoRe 算法混合 …… 139
9.3.4 MoRe 实验分析 …… 140
9.4 小结 …… 142
9.5 本章参考文献 …… 142

第 10 章 视频推荐评测 …… 145
10.1 概述 …… 145
10.2 视频推荐试验方法 …… 146
10.2.1 在线评测 …… 147
10.2.2 离线评测 …… 149
10.2.3 用户调查 …… 150
10.3 视频离线推荐评测指标 …… 151
10.3.1 准确度指标 …… 151
10.3.2 多样性指标 …… 159
10.4 小结 …… 161
10.5 本章参考文献 …… 162

下 篇

第 11 章 系统层面的快速推荐构建 …… 165
11.1 概述 …… 165
11.2 本章主要内容 …… 166
11.3 系统部署 …… 166
11.3.1 Hadoop2.2.0 系统部署 …… 166
11.3.2 Hadoop 运行时环境设置 …… 169
11.3.3 Spark 与 Mahout 部署 …… 175

11.4 Mahout 推荐引擎介绍 181
11.4.1 Item-based 算法 181
11.4.2 矩阵分解 185
11.4.3 ALS 算法 187
11.4.4 Mahout 的 Spark 实现 190
11.5 快速实战 193
11.5.1 概述 193
11.5.2 日志数据 194
11.5.3 运行环境 196
11.5.4 基于 Mahout Item-based 算法实践 201
11.5.5 基于 Mahout ALS 算法实践 205
11.6 小结 208
11.7 本章参考文献 208
第 12 章 数据层面的分析与推荐案例 211
12.1 概述 211
12.2 本章主要内容 212
12.3 竞赛内容和意义 212
12.3.1 竞赛简介 212
12.3.2 竞赛任务和意义 213
12.4 客户-商户数据 215
12.4.1 数据描述 215
12.4.2 数据理解与分析 217
12.5 算法流程设计 219
12.5.1 特征提取 219
12.5.2 分类器设计 220
12.5.3 算法流程总结 222
12.6 小结 222
12.7 本章参考文献 223

上　篇

第1章 用户画像概述

什么是用户画像？从中文概念来讲，用户画像与用户角色非常相近，是用来勾画用户（用户背景、特征、性格标签、行为场景等）和联系用户需求与产品设计的，旨在通过从海量用户行为数据中炼银挖金，尽可能全面细致地抽出一个用户的信息全貌，从而帮助解决如何把数据转化为商业价值的问题。而从英文概念角度，用户画像（User Portrait）、用户角色（User Persona）、用户属性（User Profile）这三个概念其实都是各有侧重和容易混淆的。用户角色更倾向于业务系统中不同用户的角色区分，如学校教务管理系统，老师审核、设置选课，学生查看选课和成绩。那么老师、学生就是不同的用户角色。用户画像更倾向于对同一类用户进行不同维度的刻画，对同一个电商的买家进行用户画像设计，就是将买家进一步细分和具象，如闲逛型用户、收藏型用户、比价型用户、购买型用户等。用户属性则更倾向于对属性层面的刻画和描述，特别是基本属性的内涵居多，包括性别、年龄、地域等。根据以上描述，对于视频推荐的业务来说，我们将遵循以下概念使用，用户画像近似等同于用户角色，统一称为中文概念的用户画像，而用户属性则是用户画像的子集。

用户画像的应用是非常广泛的，很多领域和行业都有用户画像这个概念，它在视频推荐领域也得到广泛应用。其中一个主要原因是，用户画像是一种能将定性与定量方法很好结合在一起的载体，定性化的方法，通过对用户的生活情境、使用场景、用户心智进行分析来对用户的性质和特征做出抽象与概括；定量化可以对特征做精细的统计分析与计算，获得对于用户较为精准的认识，便于在数值排序的基础上实现核心用户的发掘与突出。

1.1 用户画像数据来源

图 1-1 为一种常规的用户画像计算引擎示意图，虽然用户画像是一个最终的整体结果，但是它是由各个子画像综合计算而来的，而这些子画像作为中间结果并不会删除，

而是作为重要的画像解释和应用数据保存下来。拿视频推荐来说，子画像包括演员偏好画像、导演偏好画像、电影风格偏好画像，以及用户的基本属性等。需要注意的是，属性相对于其他子画像更加不易变化，因此在图中并没有特别强调该部分画像更新模块。用户注册等基本属性信息往往用于刻画相对静态的画像；而丰富的大量的用户行为日志，则是用于捕捉动态画像的重要数据来源。

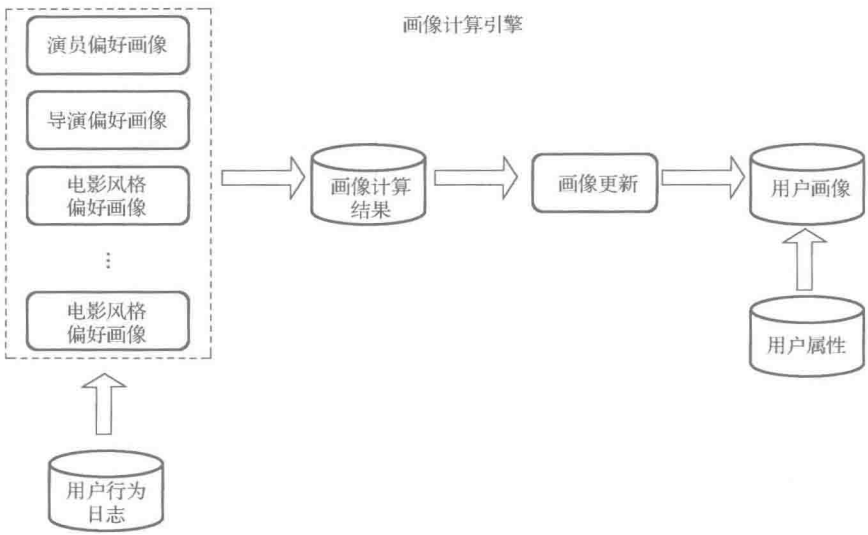

图 1-1　用户画像计算引擎示意图

从数据的角度看，用户画像就是一个对原始数据二次计算重构后的新数据，对计算增加了负担，对存储也增加了负担。所以一开始必须经过逻辑设计，从而才能确定数据结构方面的设计。

图 1-2　用户画像入口模式示意图

从可视化的角度来看，不同于以往的传统统计模式（如某个视频某个月的观看量按时间轴统计图），用户画像可能会开启一个以用户为核心牵引的新的入口呈现模式，如图 1-2 所示，沿着可解释路线，每个标签单击进去都是详细记录和细节，从抽象到细节逐步去体现用户画像数据结构，而这对于服务商来说，更加直观和更有帮助。

1.1.1　用户属性

用户属性用来描述一个用户的“个性”，从而与其他用户加以区分。因此，为实现精准及个性化的推荐，系统通常对每个用户都有一个用户属性的建模，其中包括用户的基本信息，如用户的性别、年龄、年收入、兴趣爱好、活跃时间、所在城市等。用户属性可以用于基于人口统计学的推荐，例如，系统会根据用户的属性计算不同用户间的相似度，如果计算得到用户 A 的属性和用户 B 相似度较高，那么系统会认为用户 A 和用户 B 是相似用户，在推荐引擎中，可以称他们是“邻居”。进一步讲，基于“邻居”用户群的观影喜好推荐给当前用户一些电影。此外，用户属性还可以用于对推荐结果进行过滤和排序，从而优化推荐结果。

1.1.2　用户观影行为

在推荐系统中，简单地使用用户属性存在以下问题：首先，用户属性是相对静态的数据，实时性不够；其次，基于用户属性的推荐结果过于粗糙，因为用户较难与具体的推荐内容之间建立联系（如我们很难断定某商品一定不会被某年龄段的人喜欢）。针对用户属性存在的局限性，推荐系统通常会部署特定的模块来捕捉用户的观影习惯、记录用户观影记录，来建立兴趣模型，从而针对用户的爱好进行个性化视频推荐。

用户的一次观影行为包括人物、时间、地点、事件等要素。每一次的用户观影行为本质上是一次随机事件，可以描述为：什么用户，在什么时间，在什么地点，观看了什么电影。“什么用户”涉及对用户的标识；时间则包括两个重要信息，时间戳与时间跨度，其中，时间戳标识用户行为的发生点，时间跨度则标识了用户行为的持续时间；地点体现了用户观影的渠道，便于做推荐结果的推送；观看的电影内容则直接指示用户的观影偏好，对于精准推荐至关重要。相比于静态的用户属性，用户观影行为能够更为准确地描述用户特征，是推荐系统中设计用户画像最为重要的数据来源。

1.2　用户画像特性

1.2.1　动态性

从用户画像的数据来源分析，显然用户画像具有较强的动态性。其中，用户属性涉及人口统计特征，相对比较稳定，然而用户的观影行为则是随时间持续增加的，用户在

系统内的每次观看行为都使得现有的用户画像丧失时效性。此外，用户会受到周围环境、其他用户等的影响，从而改变其观影偏好。所有这些都将决定了用户画像不可能一成不变，而是实时动态变化的。这就要求我们设计合理有效的动态更新机制，从而精准地刻画用户。

1.2.2 时空局部性

用户画像的动态性使其不可避免地具有时空局限性。

首先，在时间上，用户画像的目标是通过精准的刻画用户，从而提供个性化的服务，因此，用户画像对于时效性非常敏感，某一时刻的用户画像对该时刻的推荐结果最为有效。距离时间越远，推荐结果的精确性越低，参考价值越差。

其次，在空间上，不同的应用领域有不同的侧重点，例如，营销领域的用户画像主要侧重用户的消费习惯，而在视频推荐领域，用户画像则主要侧重用户的观影喜好，因此需要针对各自的特点设计相应的用户画像，没有哪个用户画像一经构建就可以适用于所有的应用领域。

1.3 用户画像应用领域

1.3.1 搜索引擎

随着计算机技术和网络技术的飞速发展，互联网已在人们日常生活中发挥着越来越重要的作用。面对互联网用户数量的激增和信息的爆炸性增长，如何更好地利用互联网为用户快捷地提供所需服务是一个值得研究的问题。

通过采集用户注册信息、访问日志及查询信息，我们可以构建用户画像。从而在提供搜索服务时，根据用户输入的搜索关键字及已构建的用户画像，猜测该用户可能想要得到的信息，从而将该用户最可能需要的信息显示在最前面，提高用户的搜索体验。例如，Google 的 Kaltix 算法，其基本思路就将具有类似兴趣爱好的人归为一组，为属于不同组的用户给出不同排序的结果，同时还利用了 IP、位置等信息进行基于规则的过滤。

1.3.2　推荐系统

用户画像的主要应用领域即推荐系统。下面我们选取 Amazon 作为电子商务的代表、豆瓣作为社交网络的代表来做简单介绍。

Amazon 作为推荐引擎的鼻祖，已经将推荐的思想渗透在其系统的方方面面。Amazon 通过记录用户在站点上的行为，包括浏览物品、购买物品、将物品加入收藏夹和 wish list 等，同时 Amazon 还提供了评分等用户反馈的方式，这些共同构成了用户画像的数据来源，根据不同数据的特点对它们进行处理，并分成不同类别为用户推送推荐，包括：

（1）当日推荐：通常是根据用户近期的浏览记录或者购买记录，结合时下流行的物品给出一个综合的推荐。

（2）新品推荐：采取基于内容的推荐机制，将一些新到物品推荐给用户。在方法选择上由于新物品只有较少的用户喜好信息，所以基于内容的推荐能很好地解决这个新物品“冷启动”的问题。

（3）关联推荐：采用数据挖掘技术对用户的购买行为进行分析，找到经常被一起或被同一个人购买的物品集，从而进行关联推荐。这是一种典型的基于物品的协同过滤推荐机制。

（4）他人购买/浏览商品：这也是一个典型的基于物品的协同过滤推荐应用，通过社会化机制，用户能更快更方便地找到自己感兴趣的物品。

豆瓣是国内运营比较成功的社交网站，它以图书、音乐、电影和同城活动为核心，形成一个多元化的社交网络平台，其中推荐的功能是必不可少的。相比于 Amazon 的用户行为模型，豆瓣通过分析用户“看过”和“想看”列表获得用户的偏好信息，但这也使得他们的推荐结果更专注于用户的品位。

1.3.3　其他业务定制与优化

用户画像也常常应用在个性化业务定制领域。例如目前比较火的个性化阅读。新闻客户端根据用户画像，根据读者的行为习惯和阅读经历为其“定制”内容，为不同用户显示不同新闻，最大程度地满足用户的个性化阅读需求。这种机制还允许根据用户的实际行为来进行反馈调整，从而根据用户兴趣变化动态更新内容。