# MobileRec：用于移动应用推荐的大规模数据集

M.H. Maqbool 中佛罗里达大学，美国佛罗里达州奥兰多
hasanmaqbool@knights.ucf.edu

Umar Farooq 独立研究员 美国加利福尼亚州山景城
ufarooq.cs@gmail.com

Adib Mosharrof 肯塔基大学 美国肯塔基州列克星敦
amo304@g.uky.edu

A.B. Siddique 肯塔基大学 美国肯塔基州列克星敦
siddique@cs.uky.edu

Hassan Foroosh 中佛罗里达大学，美国佛罗里达州奥兰多
hassan.foroosh@ucf.edu

arXiv:2303.06588v1 [cs.IR] 12 Mar 2023

## 摘要

推荐系统已融入我们的数字生活，应用场景无处不在，从在电子商务网站上推荐产品，到在流媒体平台上推荐电影和音乐。亚马逊产品评论和 MovieLens 等现有推荐数据集，极大地推动了各自领域推荐系统的研究与发展。过去十年间，移动用户和应用程序（也称应用）的数量呈指数级增长，但移动应用推荐系统的研究却受到严重制约，主要原因是缺乏高质量的基准数据集，而产品、电影和新闻推荐领域则不存在这一问题。为了推动应用推荐系统的研究，我们推出了一个名为 MobileRec 的大规模数据集。我们根据用户在 Google Play 商店中的活动构建了 MobileRec 。 MobileRec 包含 19.3 百万次用户交互（即用户对应用的评论），涵盖超过 10K 个独特应用，分属 48 个类别。 MobileRec 记录了总计 0.7 百万名不同用户的序列活动。每位用户至少与五个不同的应用进行过交互，这与以往每位用户仅记录一次交互的移动应用数据集形成鲜明对比。此外， MobileRec 提供了用户对已安装应用的评分和情感信息，并为每个应用提供丰富的元数据，例如应用名称、类别、描述和总体评分等。我们通过对多种最先进的推荐方法进行比较研究，证明 MobileRec 可作为应用推荐的优秀测试平台。定量结果可作为基线，供其他研究人员对比其研究结果。

MobileRec 数据集可在以下网址获取：

https://huggingface.co/datasets/recmeapp/mobilerec.

## 关键词

序列推荐，GooglePlay 数据集，应用推荐数据集。

在满足以下条件的情况下，允许免费为个人或课堂使用而制作本作品全部或部分内容的数字副本或纸质副本：副本不得用于或分发以谋取利润或商业利益，且副本首页须附有本声明和完整引文。必须尊重本作品中由 ACM 以外的其他方拥有版权的部分。注明出处后可以摘录。以其他方式复制、再版、发布到服务器或分发给列表，均须事先获得明确许可和/或支付费用。请通过 permissions@acm.org 申请许可。*SIGIR'23*，*2023*年*7*月*23*日至*27*日，中国台湾台北 © 2023 美国计算机协会。ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00 https://doi.org/10.1145/nnnnnnn.nnnnnnn




**图 1：用户活动序列示例。基于用户过去的交互（例如应用安装），应用推荐系统会推荐新的应用供用户安装。**

**ACM 引用格式：**

M.H.Maqbool、Umar Farooq、Adib Mosharrof、A.B. Siddique 和 Hassan Foroosh。2023 年。MobileRec：用于移动应用推荐的大规模数据集。载于《第 *46* 届 *ACM* 信息检索研究与发展会议论文集》（*SIGIR'23*）。ACM，纽约州纽约市，共 10 页。https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 引言

过去十年间，移动应用呈指数级增长，超过 50 亿用户因各种原因使用它们 [8]，包括社交媒体、娱乐、新闻、效率和网约车等。随着这一热潮，Google Play [16] 和 Apple App Store [5] 分别上架了超过 350 万和 220 万款应用。 [6]。日益拥挤的应用市场给用户有效发现符合自身偏好的应用带来了重大挑战。个性化应用推荐能够缓解用户的认知过载，改善应用安装体验。如图 1 所示，应用推荐系统能够根据用户以往安装的应用及其交互行为，向用户推荐新的应用。尽管 Google Play 和 App Store 采用应用推荐技术向用户推荐应用，并可能利用内部收集的用户数据，但有关应用推荐的研究几乎空白。

**表1：现有移动应用数据集与 MobileRec 的比较。**

| 数据集特征 | Top 20 Apps [3] | RRGen [15] | AARSynth [13] | Srisopha 等人 [38]* | PPrior [14]† | MobileRec |
|---|---|---|---|---|---|---|
| 评论数量 | 200K | 309K | 2.1M | 9.3M | 2.1M | 19.3M |
| 应用数量 | 20 | 58 | 103 | 1,600 | 9,869 | 10,173 |
| 应用类别数量 | 9 | 15 | 23 | 32 | 48 | 48 |
| 单个用户的多条评论 | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| 应用元数据 | ✗ | ✗ | ✓ | ✗ | ✗ | ✓ |
| 评论评分 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 评审文本 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 评审时间戳 | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ |

* [38] 不公开提供。† [14] 仅包含负面用户评论。

推荐系统在众多领域都展现出显著成效，例如新闻 [46]、电影 [10]、产品 [42, 48]、时尚 [21]、健身 [34]、玩具 [18, 31, 33]、美妆 [37] 和 CD 等。特定领域数据集的可用性，以及其中包含的用户-物品交互 [1, 26, 30]，对推荐系统的发展和改进起到了关键作用。早期研究通常使用唯一标识符来表示物品和用户，并利用相应的交互（例如评分）来学习有助于推荐任务的有效关联。这些关联通常通过协同过滤技术或类似方法来学习 [26]。近年来，将用户序列交互纳入推荐系统的开发中，显著提升了系统性能 [23]。通过考虑用户与物品交互的序列模式 [33]（例如用户浏览或购买物品的顺序），这些方法能够更好地捕捉用户偏好，并提供更准确的推荐 [23]。此外，循环神经网络 (RNNs) [11, 19]和基于注意力的模型 [43]等能够有效处理序列数据的新技术，在推荐任务中表现出色 [42, 48]。

有几个值得关注的移动应用数据集，例如 Top20apps 数据集 [3], RRGen [15], AARSyth [13], Srisopha 等人[38], 和 PPrior[14], 等。然而，Top 20 Apps 和 RRGen 仅包含数量有限的评论，通常只有几十万条。此外，这些数据集覆盖的应用不到 100 个，涉及的类别不足 20 个。尽管 AARSynth、Srisopha 等人和 PPrior 等数据集包含数千个应用的数百万条评论，规模更大，但由于缺少唯一的用户标识符，它们并不适合用于构建有效的应用推荐系统。此外，Srisopha 等人的数据集 [38] 并未公开，PPrior 则仅提供负面用户评论。表1展示了移动应用数据集的比较。本研究旨在填补这一研究空白，提供一个大型、内容丰富且多样化的基准数据集，并将其命名为 MobileRec，以助力研究人员开发应用推荐系统。据我们所知，这是唯一一个提供用户与应用之间的序列交互及多次交互记录的移动应用数据集。我们还在表2中将所提出的数据集与多个领域中成熟推荐数据集的最新版本进行了比较。

MobileRec 是一个大型数据集，基于用户在 Google Play 商店上的活动构建，包含 1.930 万次用户交互（即用户评论及其评分），涵盖 48 个类别中的 1 万个独立应用。该数据集记录了 0.7 百万名独立用户的序列活动，每位用户至少与五个应用进行过交互。数据集中的每次用户交互都包含已安装应用的关键信息，例如用户评分、文本评论和评论日期。此外，MobileRec 中的每个应用都配有丰富的元数据，例如应用标题、类别、长篇文本描述、总体平均评分、开发者信息、内容分级及其他相关详情。表3展示了该数据集的重要特征。为推动应用推荐系统研究，MobileRec 的水平可与其他领域中成熟且广受认可的数据集相媲美。除了应用推荐之外，我们的综合数据集还能为多种研究和数据分析提供重要见解和宝贵资源，例如情感分析、应用开发与优化、市场研究以及欺诈检测等。

我们还通过对多种前沿推荐技术进行比较评估，表明 MobileRec 可作为应用推荐研究的宝贵测试平台。具体而言，我们使用 MobileRec 对多个现有的通用和序列推荐系统进行了评估，包括 Pop（即物品流行度）、SASRec [23], ELECRec [7], BERT4Rec [39], HGN [28],SINE [40], LightSANs [12], GRU4Rec [41], 和 GCSAN [47]。我们对这些方法在 MobileRec 上所得结果的分析揭示了一些有趣的趋势。值得注意的是，我们发现，在 Amazon-Beauty 数据集上，ELECRec [7] 的表现显著优于 GRU4Rec [41]，在 Hit@5 指标上提升超过 300%。然而，在 MobileRec 上测试时，ELECRec 在同一指标上的表现明显不如 GRU4Rec。类似地，我们观察到，相较于 SASRec [23]，LightSANs [12] 在 Hit@10 和 NDCG@10 指标上分别提升了 3.91% 和 2.65%。但在 MobileRec 上，SASRec [23]的表现优于 LightSANs [12]。这表明 MobileRec 数据集中用户-物品交互和用户兴趣具有内在的动态性和短暂性，从而形成复杂的用户交互历史。这一观察对最先进的推荐模型具有重要意义。本文所呈现的数值结果也可作为其他研究人员评估其

**表 2：MobileRec与不同领域知名推荐数据集最新版本的比较。我们使用了2018年亚马逊评论数据集、Yelp 的 2022 年版本，以及 2019 年 12 月发布的 ML-25M 最新版本来生成这些统计数据。**

| 数据集 | 亚马逊评论 | Yelp | ML-25M | MobileRec（我们的方法） |
|---|---|---|---|---|
| 交互总数（百万） | 233.1 | 6.99 | 25.0 | 19.3 |
| 至少有五次交互的用户数（百万） | 10.6 | 0.29 | 0.16 | 0.70 |
| 至少有 15 次交互的物品数（千） | 2072 | 77.58 | 20.59 | 10.17 |
| 单个用户的最大交互数 | 446 | 3048 | 32202 | 256 |
| 单个用户的最小交互数 | 1 | 1 | 20 | 5 |
| 每位用户的平均交互数 | 5.32 | 3.52 | 153.81 | 27.56 |
| 单个项目上的最大交互数 | 13,560 | 7,673 | 81,491 | 14,345 |
| 单个项目的最小交互数 | 1 | 5 | 1 | 20 |
| 每个项目的平均交互数 | 15.45 | 46.49 | 423.39 | 1896.88 |

自身结果并客观评估不同方法的质量和有效性的基线。展望未来，诸如先进文本表示方法和预训练语言模型 [9, 35], 等自然语言处理技术，有望为利用 MobileRec 开展应用推荐研究开辟新的方向。这些技术在提升应用推荐系统的质量和有效性方面前景广阔，因此尤其受到研究界的关注。

具体而言，本研究做出了以下贡献：

- 我们推出了 MobileRec，这是迄今为止规模最大的序列化用户-应用交互集合，包含超过 1900 万次交互，涵盖来自 Google Play 的 1 万多款不同应用，横跨所有类别。值得注意的是，这是唯一一个记录了每位用户多次交互的移动应用数据集。
- 我们的实验研究考察了用户-应用交互的动态特征，并通过采用多种最先进的推荐系统，为研究社区建立了基线结果，展示了 MobileRec 的实际效用。

## 2 相关工作

### 2.1 应用数据集与推荐

如表1所示，现有多个用于移动应用用户评论的数据集。在早期收集用户评论的工作中，Iacob 和 Harrison [20]收集了 161 款移动应用的 3,279 条评论，并分析了用户提出的功能请求。Khalid 等人 [25] 和 [24]聚焦于 iOS 应用，整理了一个包含 20 款应用、共 6,390 条用户评论的数据集。McIlroy 等人 [32] 使用 12,000 款移动应用的 601,221 条用户评论，研究应用商店中的负面评论。Maalej 和 Nabil [29] 收集了规模更大的数据集，其中包含 1,186 款应用的 130 万条评论。这些数据集主要关注用户投诉和对用户与开发者对话的理解。此外，这些数据集并未公开提供。

Top 20 Apps [3] 可在 Kaggle 上获取，其中包含来自 9 个类别、涵盖 20 个应用的 20 万条评论。该数据集提供评分和评论文本。RRGen [15] 包含来自 58 个应用的 30.9 万多条评论。与 Top 20 Apps 类似，RRGen 仅提供评分和评论文本。这两个数据集均未提供应用元数据、评论时间戳以及用户的唯一标识符。

AARSynth [13] 提供了 200 多万个用户评论，涵盖 100 多个应用，并包含应用元数据。与上述数据集类似，该数据集的评论也缺少关键信息。Srisopha 等人 [38] 的数据集包含来自 1,600 个应用的 900 多万条用户评论。该数据集包含评论时间戳，有助于结合时间段背景理解评论。然而，该数据集不包含用户的唯一标识符和应用元数据。请注意，Srisopha 等人 [38] 的数据集并未公开。

近期，PPrior [14] 数据集提供了来自 Google Play、涵盖 48 个类别的 9,000 多款应用的 200 多万条评论。该数据集包含评分、评论文本和评论时间戳。不过，它未提供交互（即评论）中的用户标识符和应用元数据。此外，PPrior 仅提供负面用户评论（即仅包含评分为 1 和 2 的评论）。因此，PPrior 不适合用于构建稳健的应用推荐系统。在本研究中，我们提供了一个名为 MobileRec 的数据集，其规模大于上述所有移动应用数据集，并包含每位用户的唯一标识符以及交互的时间戳。此外，每位用户至少有 5 次与应用的交互，且所有纳入的数据集中应用均至少有 15 次交互，这使其成为移动推荐系统的理想测试平台。

### 2.2 现有推荐数据集

推荐数据集涵盖了众多不同领域，从亚马逊等电子商务平台，到包含电影相关数据集的娱乐行业，甚至延伸至在线游戏平台及其他领域。这些多样化的数据集为研究人员提供了丰富的机会，使他们能够针对广泛的应用场景，研究并提升推荐系统的有效性。

亚马逊产品评论 [18, 31, 33] 是一个大型数据集，在2018年更新版本中包含2.331亿条评论。亚马逊产品评论数据集包含29个类别。请注意，该数据集的早期版本于2014年发布，包含1.428亿次交互和24个类别。该数据集提供评论文本、评分和有用性投票，以及描述、类别信息、价格和品牌等产品元数据。亚马逊产品评论数据集包含多个类别，包括 All Beauty、图书、CD 和黑胶唱片。

**表3：MobileRec 数据集中各项重要特征的描述。**

| 特征 | 描述 |
|---|---|
| uid | 由 16 个字符组成的字母数字用户唯一标识符可唯一标识用户，同时也能使用户匿名化。<br>示例：Aj0Sm6myfh6YN3Rn、3pZUhksFIcLjEXtl、dvx0dqXTKtHUmY3O |
| 应用_包 | 表示应用程序的 Android 包名，并用于唯一标识该应用程序。<br>示例：com.google.android.calculator、com.king.crash、org.wikipedia |
| 应用_名称 | 应用在 Google Play 上显示的标题。<br>示例：糖果传奇、大富翁 - 经典棋盘游戏 |
| 应用_类别 | 应用的类别。<br>示例：娱乐、金融、效率 |
| 评论 | 用户对应用程序撰写的文字评论 |
| 评分 | 用户为应用程序给出的数值评分。<br>示例：5、2、1、4 |
| 投票数 | 认为这条评论有帮助的用户数量。<br>示例：1、···、6、··· |
| date | 特定用户/物品交互的日期，即评论日期。<br>示例：2018年10月21日、2021年11月4日、2021年1月16日 |
| 已格式化_date | 特定用户/物品交互的日期。（评论日期，格式为 YYYY-MM-DD）。<br>示例：2018-10-21、2021-11-04、2021-01-16 |
| unix_时间戳 | 将评论日期转换为 Unix 时间戳。<br>示例：1.540094e+09 1.547788e+09, 1.610773e+09 |

这些类别各自包含数千至数百万条评论。例如，All Beauty 有 0.37 百万条评论和 32K 个独立产品；图书有 5100 万条评论和 200 万个产品；CD 和黑胶唱片则包含 100 万条评论和 544K 个产品。5-core 是亚马逊产品评论数据集的一个精简版本，其中每位用户至少有 5 次交互，因此适用于构建稳健的推荐系统。由于亚马逊产品评论数据集规模庞大，现有的一些研究[7, 23]仅采用该数据集中的部分类别（如图书、Beauty、CD 和游戏）来评测其提出的方法。

Yelp [4] 数据集有多个版本，发布时间从2018年到2022年。最新版本 Yelp2022 拥有190万名唯一用户、15万个不同物品和690万条评论。类似地，MovieLens [1]提供了 ML-25M [17], ，其中包含2500万条评分，以及针对6.2万部电影的100万次标签应用。该数据集的早期版本包括 ML-100k 和 ML-1M。

除了上述大规模数据集外，来自不同领域的各类数据集也在推动各自领域的研究方面发挥了重要作用。下面我们简要介绍其他推荐数据集。Steam 包含从 stream [23] 抓取的评论和游戏信息。Book-Crossing [50] 和 GoodReads [44, 45] 提供了用户-物品交互，其中交互类型为评分。DIGINETICA[49]包含 0。2 百万用户、184K 件物品和约 1 百万次交互，其中交互类型为用户点击。该数据集由从电商搜索引擎日志中提取的用户会话汇编而成。Twitch [36] 提供了 1500 万用户在 600 万件物品上的 4.74 亿次交互，交互类型为用户点击。这些数据集推动了各自领域中多个推荐系统
[10, 18, 21, 31, 33, 34, 37, 42, 48] 的发展。我们期待 MobileRec 发挥类似作用，推动构建稳健的应用推荐系统的研究。

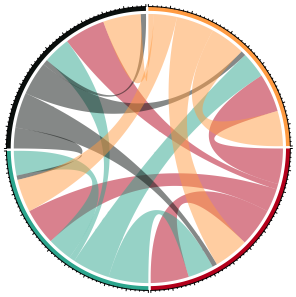
# 3 MOBILEREC数据集

## 3.1 数据集构建

Google Play [16] 是安卓用户安装应用并表达对应用看法的默认应用商店。因此，它通过用户评论承载了海量用户交互。用户评论会在滚动网页时动态加载。因此，传统网页抓取工具无法有效下载大规模用户评论数据。为了自动滚动页面并执行其他点击操作（例如，较长的评论默认不会完整显示），我们使用了 Selenium Web-Driver [2]。

为获取数据，首先通过浏览所有应用类别页面和排行榜，收集可直接从 Google Play 访问的所有应用的软件包名称（即唯一标识符）。随后，我们使用这些信息下载应用元数据。应用元数据包含应用名称、开发者、类别和文本描述等信息。我们通过访问每个应用的用户评论版块并递归向下滚动页面，提取评论文本、用户对应用的评分、用户信息（即随后进行匿名化处理）以及时间戳，从而获取与各应用相关的用户评论，直到该应用没有更多可用评论。表3展示了该数据集的重要特征。我们使用了多种筛选条件和排序机制（例如按最新时间、评分排序），以便尽可能多地下载用户评论。

为确保数据质量，我们设置了多项检查，以清除重复条目、错误信息以及爬取过程中可能出现的其他错误。在进一步处理数据并将其转换为5-core之前，我们先删除了数据中的重复项。我们还为用户分配了一个由16个字符组成的字母数字用户唯一标识符，以实现匿名化。我们只



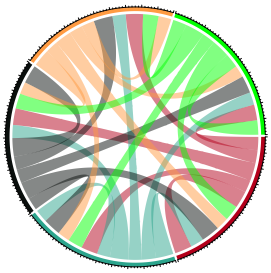
**(a) 基于内容_评分的用户序列交互趋势。**



**(b) 基于应用价格的用户序列交互趋势。**



**（c）基于应用评分的用户序列交互趋势。**



**（d）前五大类别中用户的序列交互趋势。**

**图 2：用户序列交互趋势与内容_评分、价格层级、评分以及前五大类别的关系 .**

保留评论数超过5条的用户。评论数少于15条的物品则从最终数据集中移除。最终，我们获得了一个大规模的应用用户评论5核数据集（即带时间戳的交互），涵盖谷歌Play上的应用，为了解用户对各类应用的意见和体验提供了宝贵见解。表1和表2列出了该数据集的重要统计信息。

## 3.2 数据集特征

我们介绍数据集中的各项特征（也见表3），这些特征有助于讨论数据集中的趋势。用户唯一标识符 是用户唯一的16位字母数字ID，也用于对用户进行匿名化，因为出于隐私原因，MobileRec 不提供实际的用户ID。评论 是用户撰写的实际评论文本。应用_包名 是应用程序唯一的 Android 包名。评分 表示用户对应用程序（由应用_包名标识）给出的实际评分。格式化_日期 用于按时间顺序对用户交互进行排序。更多信息请参阅表3。

## 3.3 数据集分析

本节介绍 MobileRec 数据集中的各类趋势，例如：用户如何根据基于评论的交互从一个类别转移到另一个类别；用户的行为如何随应用程序定价而变化；用户如何根据应用程序的内容分级在不同应用程序之间转移（即 Mature 17+、青少年级、全年龄级 10+ 和全年龄级）。

我们在图2(a)、2(b)、2(c)和2(d)中捕捉了用户动态行为的交互快照。为了捕捉迁移行为，我们考察了用户的前两次交互。例如，在图2(a)中可以看到，在给定时间步 $t$于成熟级 17+ 类别中进行交互的大多数用户，在时间步 $t+1$的交互中最终转而与全年龄级和青少年级内容评级的应用互动

。类似地，图 2(b) 展示了基于价格的用户序列交互，并揭示了一种预期趋势。图 2(b) 指出，在时间步 $t$ 与免费应用交互的用户中，只有极少数会在其时间步 $t+1$ 与付费应用交互。相反，许多与付费应用交互的用户会迅速转向免费应用。此外，从图2(c)可以看出，与评分最高的应用交互的用户

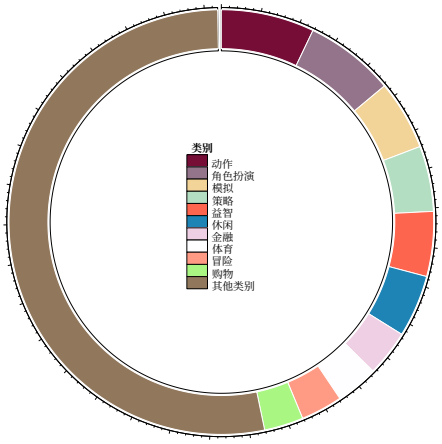



**图3：评论数量排名前十的类别。刻度盘周围的数字表示百分比，例如，动作类别中的应用所产生的评论占 %** MobileRec**数据集总评论数的** $\approx 7$。

通常不会在应用质量（即应用的整体平均评分）方面妥协，这一点从这些用户极少转向低评分应用可以看出。显然，喜欢高评分应用的用户也可能转向评分介于 *3.5-4.5* 之间的应用。图 2(d) 展示了极其动态的用户迁移模式。用户会在前五大类别之间动态地从一个类别迁移到另一个类别。这种高度波动的模式体现了在推荐系统中对用户行为建模的复杂性，尤其是在冷启动场景下，如此动态的行为会带来很大挑战。值得注意的是，MobileRec 包含 48 个类别，而图 2(d) 仅展示了评论数量最多的前五大类别。如果考虑所有类别，这种动态波动的迁移模式可能会更加复杂。

图 3 总结了前十类别的评论数量。可以看出，在 MobileRec 数据集的全部 48 个类别中，动作类别的评论数量最多。另外值得注意的是，前十类别

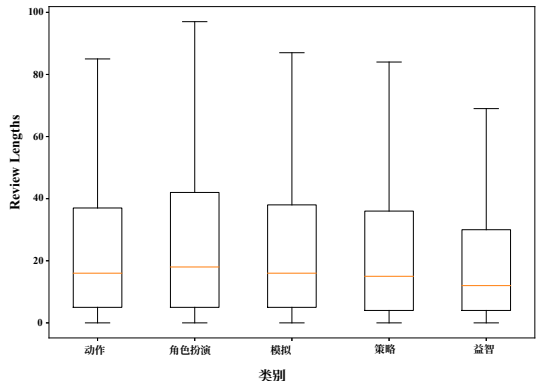



**图 4：前五大类别的评论长度，依据评论数量。**

的评论数量约占其余类别评论总数的 50%。

图 4 汇总了前五大类别（动作、角色扮演、模拟、策略和益智）的评论长度。为便于查看，我们在图 4 的箱线图中省略了须线，但提供了各类别离群值的一些统计数据。例如，在动作类别中，共有 1,373,484 条评论，其中 10,357 条评论的长度超过 100 词。因此，离群值约占动作类别评论总数的 0.754%。类似地，角色扮演类别共有 1,321,861 条评论，其中有 12,187 条评论的长度超过 100 词，离群值占比为 0.922%。模拟类别共有 1,023,366 条评论，其中 8,056 条评论的长度超过 100 词，占该类别评论总数的 0.787%。策略类别中，评论长度超过 100 词的占 0.771%，即该类别 963,360 条评论中有 7430 条符合这一条件。最后，益智类别共有 954,909 条评论，其中 4,832 条评论的长度超过 100 词，仅占该类别评论总数的 0.506%。从图 4 可以看出，前五大类别中的大多数评论长度都不足 100 词。

## 3.4 其他应用场景

除了构建推荐系统外，这个大规模数据集还可用于多种用途。从应用开发和优化到市场研究和欺诈检测，这些数据所提供的见解可以帮助企业和研究人员做出明智的决策，并制定更有效的策略。在这项工作中，我们专注于为各种推荐系统提供基线结果，并将该数据集的其他用途留给研究界探索。

与任何数据集一样，该数据集也存在潜在的滥用情形。因此，数据中不包含任何有关用户的信息。具体而言，我们对数据进行了去标识化和匿名化处理，以防止识别出个人。为求完整，我们认为有必要指出可能出现的滥用情形。例如，公司可能付费购买虚假评论或评分，人为抬高其应用的排名；也可能操纵应用元数据，使其应用更频繁地被某个推荐系统推荐。同样，攻击者也可能利用这些数据来找出更有可能被推荐的应用，随后创建包含恶意代码的虚假版本。

综合来看，我们认为该数据集带来的益处大于其潜在危害。例如，该数据集有望为个人和组织提供许多有用的见解与益处。具体而言，我们预计该数据集将促进应用推荐系统等领域的研究与开发。

# 4 基线

我们考虑了多种推荐系统进行基准测试，并为 MobileRec 数据集建立了基线结果。

## 4.1 通用推荐

**Pop.** 这是一种简单的基于流行度的模型。在该模型中，会记录物品的流行度，并向用户推荐最受欢迎的物品。例如，在 Youtube 上使用该模型时，会向用户推荐观看次数最多的视频。

## 4.2 序列推荐

**SASRec [23].** 马尔可夫链假设，用户的下一步动作可以由其近期动作（仅一个或少数几个近期动作）决定。另一方面，基于循环神经网络的方法能够考虑长期的用户-物品交互历史，以揭示用户的潜在兴趣。SASRec 通过采用注意力机制，在统一的设计中结合了基于马尔可夫链（MC）的方法和基于循环神经网络（RNN）的方法各自的优势。涉及稀疏数据集时，基于 MC 的方法可能更有效；而在涉及更稠密数据集的情况下，基于 RNN 的方法则表现良好。SASRec 提出平衡基于 MC 的方法和基于 RNN 的方法的能力，从而兼得两者之长。

**ELECRec [7]。** 下一物品预测任务通常被建模为生成任务。ELECRec 提出在推荐系统中为下一物品预测任务采用判别器。判别器负责判断下一个采样物品是否为真实目标物品。生成器 $G$ 和判别器 $D$ 联合训练。生成器 $G$ 接受训练，为判别器生成高质量样本。

**BERT4Rec [39]。** 基于序列神经网络效果不够理想且存在局限这一观察。其局限性可归因于对用户行为进行受限的单向（从左到右）编码，以及其对用户交互序列有序性的假设。BERT4Rec 是一种序列推荐模型，采用双向自注意力对用户的交互序列进行编码。该基线将完形填空目标用于序列推荐，并通过联合利用双向上下文来预测掩码物品。**HGN [28]。** 考虑到近期用户-物品交互在时间顺序上的重要性，分层门控网络提出了一种结合贝叶斯个性化排序（BPR）的方法，以捕捉用户的长期和短期兴趣。该方法由特征门控模块、物品门控模块和物品-物品乘积模块组成。门控模块负责决定从特征层和实例层向下游层传递哪些特征。物品-物品乘积模块

捕捉用户过去交互过以及未来将要交互的物品之间的物品关系。

**SINE [40]**。 认识到用户行为序列中多个概念不同的物品的重要性后，SINE 指出，为用户行为序列生成统一嵌入，主要会受到用户最近交互的影响。基于这一观察，SINE 提出使用多个嵌入来捕捉用户交互行为的各个方面。由于概念池可能很大，SINE 能够从大型概念池中推断出一组稀疏概念。借助多个嵌入，SINE 采用兴趣聚合模块来预测用户的当前意图，该模块也用于对下一物品预测进行建模。

### 4.3 基于Transformer的推荐

**LightSANs [12]**。 自注意力网络（SAN）受到二次复杂度和易于过度参数化的限制。由于采用隐式位置编码，SAN 对物品间序列关系的建模不够准确，这也是其局限因素之一。为此，LightSAN 提出了一种低秩分解自注意力网络。低秩分解自注意力将用户的历史物品映射为少量潜在兴趣。物品与兴趣交互用于生成上下文感知表示。

### 4.4 基于会话的推荐

**GRU4Rec [41]**。 该基线采用基于RNN的方法进行基于会话的推荐。所提出的方法采用数据增强，并通过方法应对输入数据中的分布偏移。

### 4.5 基于GNN的推荐

**GCSAN [47]**。 该方法提出了一种图上下文化的自注意力模型，同时采用图神经网络和自注意力网络。在GCSAN中，图神经网络用于捕获丰富的局部依赖关系，而自注意力网络用于捕获长距离依赖关系。每个会话都表示为全局偏好与会话当前兴趣的组合。

## 5 实验设置

### 5.1 实验设置

对于序列推荐基线，我们统一使用 4096 的批次大小，并将最大交互长度设为 50。我们采用留一法策略进行验证和测试，并使用完整物品集进行评估。我们将早停耐心值设为 10。我们将序列推荐任务视为多类分类任务，并使用交叉熵损失训练模型。SASRec [23] 使用 Adam优化器训练，学习率为 0.001。层数和注意力头数均为 2。丢弃率为 0.5，并使用 GELU 作为激活函数。HGN [28] 和 SINE [40] 使用嵌入维度 64、学习率 0.001 和 Adam优化器进行训练。LigthSANs [12] 的潜在兴趣维度设为 5，使用 2 个注意力头和 2 个 Transformer层。训练时使用学习率 0.001 和 Adam优化器。GCSAN [47]具有 2 个 Transformer编码器层，2 个注意力头，隐藏状态特征维度为 64，前馈层的隐藏维度为 256，权重设为 0.6，图神经网络的层数为 1；使用 Adam 进行优化，学习率为 0.001。GRU4Rec [41] 使用 1 层进行训练，嵌入维度为 64，隐藏维度为 128，丢弃率设为 0.3。我们使用 RecBole [49] 为 MobileRec 建立基线。

### 5.2 评估指标

我们采用标准指标 Hit@K 和 NDCG@K，其中 $K \in \{1, 5, 10, 15, 20\}$，作为基准方法的评估指标。Hit@k 用于统计预测物品出现在前 $K$ 个列表中的次数，可表示为 [27]:

$$HR@K = \frac{1}{M}\sum_{u=1}^{M}\mathbb{1}_{R_u \le K} = \sum_{R=1}^{N} W_R \cdot \mathbb{1}_{R \le K}$$

$$W_R = \frac{1}{M}\sum_{u=1}^{M}\mathbb{1}_{R_u = R}$$

$\mathbb{1}_X$表示指示随机变量，$M$表示用户总数，$N = |I|$表示物品总数。R 是物品的整数排名位置，范围为 $[1, N]$。对于用户 $u$，$R_u$表示物品 $i_u$在 $I$个物品中的排名。$W_R$表示物品 $i_u$位于 $R$位置的用户。

$NDCG@K$ 可表示为 [22, 49]:

$$\frac{1}{|U|}\sum_{u \in U}\left(\frac{1}{\sum_{i=1}^{\min(|R(u)|,K)}\frac{1}{\log_2(i+1)}}\sum_{i=1}^{K}\delta(i \in R(u))\frac{1}{\log_2(i+1)}\right)$$

所有物品集均用于对预测结果进行排序。

## 6 结果与讨论

表4展示了各种基线模型在 MobileRec 上的性能分析。接下来，我们将讨论这些结果并提供更多细节。如前所述，Pop 是一种基于流行度的模型，依赖于物品流行度。Pop 是最朴素的模型，正如预期的那样，与几乎所有其他基线相比，其性能最差。此前的报告显示，根据 Hit@10 和 NDCG@10 性能指标，SASRec 的结果优于 Pop。在 SASRec 中，随机抽取了 100 个负样本物品，并以这 101 个物品（包括真实值）为基准计算 Hit@10 和 NDCG@10。使用 MobileRec 时，我们采用完整物品集进行评估，这对推荐系统而言是一项更严格的评估标准。我们观察到，在所有性能指标上，除 Hit@1 外，SASRec 的表现均优于 Pop。例如，与 Pop 相比，SASRec 的 Hit@10 提高了 19.86%。同样，Hit@15 的性能提高了 16.34%。类似地，在 MobileRec 上，SASRec 的 Hit@20 比 Pop 提高了 15.23%。

从归一化折损累计增益指标来看，在 Beauty 数据集上进行评估时，SASRec 报告的 NDCG@10 比 Pop 高 41.37%。而在 MobileRec 上，与 Pop 相比，NDCG@10 仅提升了 14.28%。我们认为，SASRec 与 Pop 在这两个数据集（Beauty 和 MobileRec）上的 NDCG@10 提升幅度存在差异，可以结合 Beauty 和

**表4：**MobileRec**上各类基线的性能分析。基线涵盖多个类别，例如通用基线（如 Pop）、序列基线（如 SASRec、ELECRec、BERT4Rec、HGN、SINE）、基于会话的基线（如 GRU4Rec、GCSAN）、基于图神经网络的基线（如 GCSAN）以及基于 Transformer 的基线（如 LigthSANs）。**

| 方法 ↓指标 → | Hit@1 | Hit@5 | Hit@10 | Hit@15 | Hit@20 | NDCG@5 | NDCG@10 | NDCG@15 | NDCG@20 |
|---|---|---|---|---|---|---|---|---|---|
| **Pop** | 0.0027 | 0.0086 | 0.0151 | 0.0208 | 0.0256 | 0.0056 | 0.0077 | 0.0092 | 0.0103 |
| **SASRec** | 0.0026 | 0.0098 | 0.0181 | 0.0242 | 0.0295 | 0.0061 | 0.0088 | 0.0104 | 0.0117 |
| **ELECRec** | 0.0020 | 0.0094 | 0.0174 | 0.0237 | 0.0293 | 0.0056 | 0.0082 | 0.0098 | 0.0112 |
| **BERT4Rec** | 0.0024 | 0.0083 | 0.014 | 0.0183 | 0.0221 | 0.0054 | 0.0072 | 0.0083 | 0.0092 |
| **HGN** | 0.0012 | 0.0054 | 0.0096 | 0.0132 | 0.0165 | 0.0033 | 0.0046 | 0.0056 | 0.0064 |
| **SINE** | 0.0022 | 0.0087 | 0.0163 | 0.0228 | 0.028 | 0.0054 | 0.0078 | 0.0095 | 0.0107 |
| **LightSANs** | 0.0024 | 0.0102 | 0.0172 | 0.0227 | 0.028 | 0.0062 | 0.0085 | 0.0099 | 0.0112 |
| **GRU4Rec** | 0.0021 | 0.0086 | 0.0153 | 0.021 | 0.0261 | 0.0053 | 0.0074 | 0.0089 | 0.0102 |
| **GCSAN** | 0.0024 | 0.0094 | 0.0161 | 0.0214 | 0.0266 | 0.0059 | 0.0081 | 0.0095 | 0.0107 |

MobileRec 的数据集统计信息来解释。Beauty 有 52,024 个用户，而 MobileRec 有 700,111 个用户。其次，Beauty 数据集包含 57,289 个物品，而 MobileRec 包含 10,173 个应用。Beauty 有 $0.4M$ 次交互，平均每位用户交互 7.6 次，每个物品平均交互 6.9 次。相比之下，MobileRec 有 $19.3M$ 次交互，平均每位用户交互 27.56 次，每个物品（即应用）平均交互 1,896.88 次。例如，MobileRec 平均每位用户有 27.56 次交互，是 Beauty 的 200 多倍。同样，与 Beauty 相比，MobileRec 的每个物品平均交互次数也显著更高。鉴于这些因素，MobileRec 呈现出更具动态性的推荐场景。MobileRec 中用户与物品之间的高度交互，可能是 SASRec 相较于其在 Beauty 上报告的相对 Pop 的提升，NDCG@10 提升幅度较小的原因。其次，由于我们在评估时采用完整物品集，而 SASRec 使用的是排序策略，这也可能是 SASRec 相较于 Pop 性能提升较小的潜在原因。在 Beauty 上，SASRec 的结果优于 GRU4Rec。我们观察到相同的趋势：在 MobileRec 上进行训练和评估时，SASRec 的表现也优于 GRU4Rec。

在 Beauty 数据集上，ELECRec 的 Hit@5、Hit@10、NDCG@5 和 NDCG@10 均优于 Pop。在 MobileRec 上进行训练和评估时，我们也观察到了相同的模式，唯一的例外是 Hit@1，此时 Pop 的表现优于 ELECRec。我们推测，用户-物品交互历史的高度动态性可能是原因。同样，ELECRec 在 Beauty 上相较于 GRU4Rec，Hit@5 提升了 329.87%，Hit@10 提升了 242.04%。此外，在 NDCG@5 和 NDCG@10 指标上，ELECRec 相较于 GRU4Rec 分别提升了 412.12% 和 331.38%。聚焦于在 MobileRec 上进行训练和评估时 ELECRec 相较于 GRU4Rec 的性能提升，ELECRec 的 Hit@5 和 Hit@10 分别提升了 9.30% 和 13.72%。同样，NDCG@5 和 NDCG@10 分别提升了 1.88% 和 10.81%。这一观察也表明，MobileRec 中的用户-物品交互具有高度动态性。两种模型都难以取得较高性能，因此与 ELECRec 中报告的性能差距相比，这两种方法之间的性能差距有所缩小。MobileRec 中有超过 1900 万次交互，而 Beauty 中只有 0.4M 次用户-物品交互

。Beauty 中的用户-物品交互具有如此高的动态性，使得模型的性能提升幅度较低。基于这一分析，MobileRec 应当成为现有推荐数据集的重要补充。

在 Beauty 数据集上，ELECRec 相较于 SASRec 同样展现出了性能提升。首先，我们将介绍 ELECRec 在 Beauty 数据集上的 Hit@5 、 Hit@10 、 NDCG@5 和 NDCG@10 指标方面相较于 SASRec 所取得的相对性能提升。随后，我们将讨论当训练和评估数据集为 MobileRec 时，ELECRec 相较于 SASRec 所实现的性能改进。首先，在 Hit@5 和 Hit@10 方面，ELECRec 相较于 SASRec 分别提升了 83.59% 和 59.47%。同样，在 NDCG@5 和 NDCG@10 方面，ELECRec 相较于 SASRec 分别提升了 103.61 % 和 84.11%。现在，让我们讨论在 MobileRec 上进行训练和评估时 ELECRec 与 SASRec 之间的性能比较。可以观察到，在 MobileRec 上，SASRec 在 Hit@5、Hit@10、NDCG@5 和 NDCG@10 指标上均优于 ELECRec。对 SASRec 相较于 ELECRec 在 Hit@5、Hit@10、NDCG@5 和 NDCG@10 方面的性能提升进行量化后，结果分别为 4.25%、4.02%、8.92% 和 7.31%。我们认为，ELECRec 使用了一个通过自然语言处理任务训练的生成器，而判别器负责判断序列中的下一个物品是否为正确的（真实）物品，还是虚假的下一个物品。判别器的判别能力取决于生成器生成的样本质量。由于 MobileRec 呈现出高度动态的用户-物品交互序列，且用户表现出的兴趣转瞬即逝，因此生成器可能难以生成高质量的训练样本。由于判别器捕捉真实物品相关性的能力取决于生成器生成的样本质量，低质量样本可能会导致判别器所带来的性能提升减弱。

值得回顾一下图 2，它展示了动态用户-物品交互序列的有趣快照。例如，从图2(b)可以看出，相当大比例的用户会从付费应用转向免费应用。图2(d)也呈现了类似的动态迁移模式，描绘了用户在排名前 5 的类别之间的迁移。我们认为，对于具有这种动态时间交互模式的 1,900 万次用户-物品交互，生成器要学习生成高质量样本可能会面临挑战。此外，ELECRec 报告了

其与 BERT4Rec 的比较性能。在原论文中，ELECRec 在 Beauty [7] 上的所有指标均优于 BERT4Rec。具体而言，ELECRec 报告称，在 Hit@5 和 Hit@10 上，其性能分别比 BERT4Rec 高出 100.85% 和 61.06%；在 NDCG@5 和 NDCG@10 指标上，则分别高出 131.50% 和 97%。我们发现，在使用 MobileRec 进行训练和评估时，ELECRec 在大多数评估指标上优于 BERT4Rec，这一模式与上述结果相同；但在 Hit@1 上，BERT4Rec 的表现更好。考虑到 Hit@1 是一项更严格的标准，我们认为，BERT4Rec 采用的训练目标——通过联合条件化双向上下文来预测序列中的掩码物品——可能更适合 Hit@1 这类严格评估指标。尤其是在 MobileRec 这类用户兴趣转瞬即逝的情境中，通过联合条件化双向上下文进行掩码物品预测，似乎有助于编码用户的动态行为。

图上下文自注意力模型 GCSAN 同时采用图神经网络和自注意力网络进行会话推荐。原论文报告了 GCSAN 与竞争方法之间的多项比较评估。由于我们也将 Pop 和 GRU4Rec 与 GCSAN 一同纳入基线，因此我们主要关注 GRU4Rec 和 Pop 相较于 GCSAN 的比较表征学习能力。在原论文中，使用 Amazon-Books数据集时，GCSAN 在 NDCG@5 和 NDCG@10 上的表现优于 Pop 和 GRU4Rec。同样，在不同的基准数据集上，与 Pop 和 GRU4Rec 相比，GCSAN 在 Hit@5 和 Hit@10 上也取得了更好的结果。我们观察到，在 MobileRec 上，GCSAN 在 NDCG@5、NDCG@10、Hit@5 和 Hit@10 上均持续优于 Pop 和 GRU4Rec，呈现出类似的趋势。我们注意到，GCSAN 在 NDCG@5 和 NDCG@10 上分别比 Pop 提高了 9.30% 和 6.62%。同样，GCSAN 在 Hit@5、Hit@10、NDCG@5 和 NDCG@10 上均优于 GRU4Rec。

LightSANs 是一种用于下一物品推荐任务的 Transformer 变体，它采用低秩分解自注意力，将用户的历史兴趣投影到少数几个潜在兴趣中。LightSANs 报告了其在多个竞争基线上的 Hit@10 和 NDCG@10 结果。我们将重点关注 Pop、GRU4Rec、BERT4Rec 和 SASRec，以比较并分析 LightSANs 在 Amazon-Books数据集和 MobileRec 上的性能。在进一步比较 LightSANs 在 Amazon-Beauty 与 MobileRec 上的性能之前，我们先来看一下 Amazon-Books数据集。Amazon-Books数据集包含 19K 个用户、60K 个物品和 1.7 million 次交互。在 Amazon-Books数据集上，与 Pop 相比，LightSANs 的 Hit@10 和 NDCG@10 分别提升了 121.77% 和 172.43%。同样，与 SASRec 相比，Hit@10 和 NDCG@10 分别提升了 3.91% 和 2.65%。与 GRU4Rec 相比，Hit@10 和 NDCG@10 分别提升了 8.41% 和 5.72%。最后，与 BERT4Rec 相比，性能分别提升了 8.76% 和 4.03%。在 MobileRec 数据集上，与 Pop 相比，Hit@10 和 NDCG@10 分别提升了 13.90% 和 10.38%。在相同指标上，LightSANs 相较于 GRU4Rec 分别提升了 12.41% 和 14.86%。

与 ELECRec 类似，LightSANs 在 Hit@10 和 NDCG@10 指标上的表现也分别比 BERT4Rec 高 22.85% 和 18.05%。我们注意到，LightSANs 在 MobileRec 上延续了其在亚马逊图书数据集上的表现模式，优于其他竞争基线；不过，SASRec 是个例外，它在 Hit@10 和 NDCG@10 指标上的表现优于 LightSANs。LightSANs 相较于 BERT4Rec、Pop 和 GRU4Rec 展现出稳定的性能，这可归因于 LightSANs 更好的设计选择。例如，LightSANs 提出使用解耦的位置编码取代隐式位置编码。我们认为，这种隐式位置编码有助于 LightSANs 更准确地建模历史用户兴趣和用户-物品交互，从而使模型在 MobileRec 上取得更高的 Hit@10 和 NDCG@10。尽管如此，LightSANs 的表现仍不及 SASRec。SASRec 的性能优势可能源于其引入的可学习位置嵌入。

SINE 探讨了使用多个嵌入向量对用户兴趣进行编码的思路，这建立在其经验发现之上：用户的行为序列呈现出多个不同的兴趣。SINE 使用多个数据集对结果进行基准测试。我们考察 SINE 在亚马逊产品评论数据集上报告的性能，采用 Hit@10 和 NDCG@10 指标。我们的比较分析仅涉及 SASRec 和 GRU4Rec。SINE 报告称，在 Hit@50、Hit@100、NDCG@50 和 NDCG@100 上，性能分别优于 SASRec。与 GRU4Rec 相比，SINE 在 Hit@50、Hit@100、NDCG@50 和 NDCG@100 上也取得了更好的性能。使用 MobileRec 进行训练和评估时，SINE 在 Hit@20 和 NDCG@20 上分别以 7.27% 和 4.90% 的优势胜过 GRU4Rec；但在 Hit@20 和 NDCG@20 上，其表现仍不及 SASRec。首先，这可能是因为与 SINE 采用的 Hit@50、Hit@100、NDCG@50 和 NDCG@100 相比，Hit@20 和 NDCG@20 的评判标准更严格。其次，考虑到 SINE 致力于从用户的交互历史中捕捉其不同兴趣，我们认为，由于用户兴趣不断变化（如图 2 所示），在 MobileRec 中收敛到不同的用户兴趣具有挑战性。SINE 无法从交互序列中学习并嵌入不同的用户兴趣，这可能导致其性能不如 SASRec。

## 7 结论

本文介绍了 MobileRec，这是一个大规模的序列化用户-应用交互数据集。MobileRec 的独特之处在于，它记录了每位用户的多次交互，从而能够更全面地呈现用户行为。MobileRec 共包含 1.93 亿次用户-应用交互，涵盖来自 48 个类别的逾 1 万款不同应用，并涉及 0.7 亿名独立用户，每位用户至少与 5 款不同的应用发生过交互。MobileRec 以前所未有的细致程度，展现了用户在不同应用类别中的参与度。此外，MobileRec 中的每次用户-应用交互都包含丰富的上下文信息，例如用户评分、文本评论和评论日期。最后，MobileRec 中的每款应用都附带丰富的元数据，例如应用标题、类别、长篇文本描述、整体平均评分、开发者信息和内容分级等。我们还通过对多种最先进的推荐技术进行比较评估，展示了 MobileRec 数据集作为应用推荐研究实验测试平台的实用价值。这项评估也建立了基线结果，

这些结果将惠及研究界。我们希望该数据集能够激发更多研究，带来新的见解，并为未来的移动应用推荐系统铺平道路。

## 参考文献

[1] Movielens。https://grouplens.org/datasets/movielens/，2022年。访问日期：2022-11- 06。[2] Selenium WebDriver。https://www.selenium.dev/documentation/webdriver/，2022年。访问日期：2023-18-02。[3] Play 商店应用评论前 20 名。https://www.kaggle.com/datasets/odins0n/top-20-play-store-app-reviews-daily-update，2022年。访问日期：2022-12-09。[4] Yelp 开放数据集。https://www.yelp.com/dataset，2022年。访问日期：2023-18-02。[5] 苹果。Apple App Store。https://apps.apple.com/，2022年。访问日期：2022-11-06。[6] L. Ceci。主流应用商店中的应用数量。https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/。访问日期：2022-11-06。[7] Yongjun Chen、Jia Li 和 Caiming Xiong。ELECRec：将序列推荐器训练为判别器。arXiv 预印本 arXiv:2204.02011，2022年。[8] J. Degenhard。主流应用商店中的应用数量。https://www.statista.com/forecasts/1143723/smartphone-users-in-the-world。访问日期：2022-02-02。[9] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向变换器预训练。arXiv 预印本 arXiv:1810.04805，2018年。[10] Qiming Diao、Minghui Qiu、Chao-Yuan Wu、Alexander J Smola、Jing Jiang 和 Chong Wang。联合建模方面、评分和情感以进行电影推荐（jmars）。载于第20届 ACM SIGKDD知识发现与数据挖掘国际会议论文集，第193–202页，2014年。[11] Jeffrey L Elman。发现时间中的结构。《认知科学》，14(2):179–211，1990年。[12] Xinyan Fan、Zheng Liu、Jianxun Lian、Wayne Xin Zhao、Xing Xie 和 Ji-Rong Wen。更轻、更好：用于下一物品推荐的低秩分解自注意力网络。载于第44届 ACM SIGIR信息检索研究与发展国际会议论文集，第1733 1737页，2021 – 。[13] Umar Farooq、AB Siddique、Fuad Jamour、Zhijia Zhao 和 Vagelis Hristidis。面向应用的用户评论回复合成。载于2020年 IEEE大数据国际会议（Big Data），第699–708页。IEEE，2020年。[14] Moghis Fereidouni、Adib Mosharrof、Umar Farooq 和 AB Siddique。通过对比学习主动确定应用问题的优先级。载于2022年 IEEE大数据国际会议（Big Data），第535–544页。IEEE，2022年。[15] Cuiyun Gao、Jichuan Zeng、Xin Xia、David Lo、Michael R Lyu 和 Irwin King。自动生成应用评论回复。载于2019年第34届 IEEE/ACM自动化软件工程国际会议（ASE），第163–175页。IEEE，2019年。[16] 谷歌。Google Play 商店。https://play.google.com/store/apps，2022年。访问日期：2022-11-06。[17] F Maxwell Harper 和 Joseph A Konstan。Movielens 数据集：历史与背景。《ACM交互式智能系统汇刊》（tiis），5(4):1–19，2015年。[18] Ruining He 和 Julian McAuley。起伏变化：利用单类协同过滤对时尚趋势的视觉演变进行建模。载于第25届万维网国际会议论文集，第507–517页，2016年。[19] Sepp Hochreiter 和 Jürgen Schmidhuber。长短期记忆。《神经计算》，9(8):1735–1780，1997年。[20] Claudia Iacob 和 Rachel Harrison。从在线评论中提取并分析移动应用功能请求。载于2013年第10届软件仓库挖掘工作会议（MSR），第41–44页。IEEE，2013年。[21] Tomoharu Iwata、Shinji Watanabe 和 Hiroshi Sawada。利用时尚杂志照片的时尚搭配推荐系统。第二十二届国际人工智能联合会议，2011年。[22] Kalervo Järvelin 和 Jaana Kekäläinen。基于累积增益的信息检索技术评估。《ACM信息系统汇刊》（TOIS），20 446，2002 4 422( ): – 。[23] Wang-Cheng Kang 和 Julian McAuley。自注意力序列推荐。载于2018年 IEEE国际数据挖掘会议（ICDM），第197–206页。IEEE，2018年。[24] Hammad Khalid。识别 iOS 应用的用户投诉。载于2013年第35届国际软件工程会议（ICSE），第1474–1476页。IEEE，2013年。[25] Hammad Khalid、Emad Shihab、Meiyappan Nagappan 和 Ahmed E Hassan。移动应用用户会抱怨什么？《IEEE软件》，32(3):70–77，2014年。[26] Yehuda Koren。因子分解与邻域相遇：一种多面向的协同过滤模型。载于第14届 ACM SIGKDD知识发现与数据挖掘国际会议论文集，第426–434页，2008年。[27] Dong Li、Ruoming Jin、Jing Gao 和 Zhi Liu。关于 Top-k 推荐评估的采样。载于第26届 ACM SIGKDD知识发现与数据挖掘国际会议论文集，第2114–2124页，2020年。[28] Chen Ma、Peng Kang 和 Xue Liu。用于序列推荐的分层门控网络。载于第25届 ACM SIGKDD知识发现与数据挖掘国际会议论文集，第825–833页，2019年。[29] Walid Maalej 和 Hadeer Nabil。错误报告、功能请求，还是单纯的赞扬？自动分类应用评论。载于2015年 IEEE第23届国际需求工程会议（RE），第116–125页。IEEE，2015年。[30] Julian McAuley。亚马逊商品数据。http://jmcauley.ucsd.edu/data/amazon/，2022年。访问日期：2022-11-06。[31] Julian McAuley、Christopher Targett、Qinfeng Shi 和 Anton Van Den Hengel。基于图像的风格与替代品推荐。载于第38届 ACM SIGIR信息检索研究与发展国际会议论文集，第43–52页，2015年。[32] Stuart McIlroy、Nasir Ali、Hammad Khalid 和 Ahmed E. Hassan。分析并自动标注移动应用评论中提出的用户问题类型。《实证软件工程》，21:1067–1106，2016年。[33] Jianmo Ni、Jiacheng Li 和 Julian McAuley。利用远程标注评论和细粒度方面解释推荐结果。载于2019年自然语言处理实证方法会议暨第9届国际自然语言处理联合会议（EMNLP-IJCNLP）论文集，第188–197页，2019年。[34] Jianmo Ni、Larry Muhlstein 和 Julian McAuley。对心率和活动数据进行建模，以实现个性化健身推荐。载于 WWW，2019年。[35] Alec Radford、Jeff Wu、Rewon Child、David Luan、Dario Amodei 和 Ilya Sutskever。语言模型是无监督多任务学习器。2019年。[36] Jérémie Rappaz、Julian McAuley 和 Karl Aberer。直播平台上的推荐：动态可用性与重复消费。载于第15 399届 ACM推荐系统会议论文集，第2021 390页 – 。[37] Jérémie Rappaz、Julian McAuley 和 Karl Aberer。直播平台上的推荐：动态可用性与重复消费。载于 RecSys，2021年。[38] Kamonphop Srisopha、Daniel Link 和 Barry Boehm。开发者应该如何回复应用评论？预测开发者回复成效的特征。EASE 2021，第119–128页，美国纽约州纽约市，2021年。美国计算机协会。[39] Fei Sun、Jun Liu、Jian Wu、Changhua Pei、Xiao Lin、Wenwu Ou 和 Peng Jiang。BERT4Rec：利用来自Transformer的双向编码器表示进行序列推荐。载于第28届 ACM信息与知识管理国际会议论文集，第1441–1450页，2019年。[40] Qiaoyu Tan、Jianwei Zhang、Jiangchao Yao、Ninghao Liu、Jingren Zhou、Hongxia Yang 和 Xia Hu。用于序列推荐的稀疏兴趣网络。载于第14届 ACM网页搜索与数据挖掘国际会议论文集，第598–606页，2021年。[41] Yong Kiam Tan、Xinxing Xu 和 Yong Liu。用于基于会话推荐的改进型循环神经网络。载于第1届推荐系统深度学习研讨会论文集，第17–22页，2016年。[42] Mehrab Tanjim、Congzhe Su、Ethan Benjamin、Diane Hu、Liangjie Hong 和 Julian McAuley。用于下一物品推荐的潜在意图注意力序列模型。载于 WWW，2020年。[43] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力就是你所需的一切。《神经信息处理系统进展》，30，2017年。[44] Mengting Wan 和 Julian McAuley。单调行为链上的物品推荐。载于第12届 ACM推荐系统会议论文集，第86–94页，2018年。[45] Mengting Wan、Rishabh Misra、Ndapa Nakashole 和 Julian McAuley。从大规模评论语料库中进行细粒度剧透检测。arXiv 预印本 arXiv:1905.13416，2019年。[46] Fangzhao Wu、Ying Qiao、Jiun-Hung Chen、Chuhan Wu、Tao Qi、Jianxun Lian、Danyang Liu、Xing Xie、Jianfeng Gao、Winnie Wu 等。Mind：用于新闻推荐的大规模数据集。载于计算语言学协会第58届年会论文集，第3597–3606页，2020年。[47] Chengfeng Xu、Pengpeng Zhao、Yanchi Liu、Victor S Sheng、Jiajie Xu、Fuzhen Zhuang、Junhua Fang 和 Xiaofang Zhou。用于基于会话推荐的图上下文化自注意力网络。载于 IJCAI，第19卷，第3940–3946页，2019年。[48] An Yan、Chaosheng Dong、Yan Gao、Jinmiao Fu、Tong Zhao、Yi Sun 和 Julian McAuley。个性化互补商品推荐。载于 WWW，2022年。[49] Wayne Xin Zhao、Shanlei Mu、Yupeng Hou、Zihan Lin、Yushuo Chen、Xingyu Pan、Kaiyuan Li、Yujie Lu、Hui Wang、Changxin Tian、Yingqian Min、Zhichao Feng、Xinyan Fan、Xu Chen、Pengfei Wang、Wendi Ji、Yaliang Li、Xiaoling Wang 和 Ji-Rong Wen。RecBole：面向推荐算法的统一、全面且高效的框架。载于 CIKM，第4653–4664页。ACM，2021年。[50] Cai-Nicolas Ziegler、Sean M McNee、Joseph A Konstan 和 Georg Lausen。通过主题多样化改进推荐列表。载于

*14th* 国际 *l conf*关于世界的会议 *ld Wide Web*，第22–32页，2005年 .