数据集 > 资讯内容用户行为数据集

收藏 5

资讯内容用户行为数据集

比赛

大量的用户行为数据和资讯内容数据，帮助你设计算法，发掘用户的兴趣爱好，为每个用户推荐最精准匹配的资讯内容。

上传者： 小科

挂载目录： datagrand

数据类型： Kesci 精选

数据来源： “达观杯”大数据算法竞赛

更新时间： 2017-05-04

数据大小： 63.2MB

创建项目

概述 项目 3 评论 0

背景介绍

手机资讯APP已经成为大部分人获取资讯最重要的渠道。在信息爆炸的时代，能在有限的屏幕内给用户展示最感兴趣的资讯才能留住用户。因此，精准的个性化推荐已经成为每个资讯APP的标配。

此次比赛的目的，是为了从大量的用户行为数据中，发掘用户的兴趣爱好，并为每个用户推荐最精准匹配的资讯内容。

达观作为领先的大数据公司，掌握前沿的数据挖掘、智能推荐、搜索引擎和自然语言处理等技术，每天为上亿用户提供数亿次的个性化推荐和搜索服务。

意见反馈

下载与作品提交说明

数据集下载

下载地址：“达观杯”大数据算法竞赛
(https://cdn.kesci.com/datagrand_update.zip)

提交说明

本次比赛需要选手下载数据集，按照下载文件中的提交样例格式在**比赛页面-【团队提交】**中提交作品。

同时为了选手能更方便地通过K-Lab创建项目，向社区分享自己的创意，算法，学习记录等任何有价值的代码文档，或者通过fork和评论别人的项目来进行比赛和学习上的交流，我们也提供了在线的比赛数据集，使用方法为：

创建项目后：
Python用户，输入 `ls ../input/datagrand/` 查看数据路径
R用户，输入 `list.files("../input/datagrand/")` 查看数据路径
使用相关包读取数据。

数据描述

本次比赛选取了一批用户（candidate.txt），以及一批候选资讯内容数据（news_info.csv）用以推荐给用户。同时提供了这批用户在某一天（记为第N天）对资讯内容的多种行为数据，包括点击、完整阅读、评论、收藏、分享等，作为训练数据。比赛目标是针对这批用户，预测每个用户在第二天（记为第N+1天）会产生行为的资讯列表。

计算私有（private）排行榜的数据包含全部的用户id，从中选取50%的用户id用于计算公开（public）排行榜。我们另外提供了一份公开测试数据供下载测试，公开测试数据不包含计算排行榜用到的用户id

注意候选资讯内容数据中，有一部分是第N+1天才新增的资讯，因而不会出现在训练数据中，但用户在第N+1天有可能对其产生行为，选手需要能处理这类新增资讯内容。

数据集包含以下数据文件：

- `train.csv` ：训练数据集。
- `test.txt` ：测试数据集。
- `news_info.csv` ：候选资讯内容数据。
- `candidate.txt` ：待推荐的用户ID，每行一个ID。
- `samplesubmission.txt` ：提交结果样例

数据字典

1. `train.csv` 数据大小：56.2MB

列名	描述	数据类型
user_id	用户唯一ID	string
item_id	资讯唯一ID	string
cate_id	资讯类别ID	string
action_type	用户行为类型，包括view、deep_view、share、comment、collect等	string
action_time	行为发生时间，秒级时间戳	int

2. `news_info.csv` 数据大小：901KB

列名	描述	数据类型
item_id	资讯唯一ID	string
cate_id	资讯类别ID	string

3. tes.txt 和 sample_submission.txt

每行是一个用户的 点击/推荐 结果，为每个用户推荐5个最可能有行为的itemid。格式为“userid,itemid1 itemid2 itemid3 itemid4 itemid5”。
注意提交结果务必按照sample_submission.txt的格式，每个用户推荐5个最可能有行为的itemid即可，itemid不可重复，否则该用户推荐结果视为无效。
test.txt中每个用户点击的资讯个数不定，但个数大于等于1个。

数据预览

1. train.csv

user_id	item_id	cate_id	action_type	action_time
0365F7AE-5048-42B3-BB2C-8E637A380A3E	557082	1_3	view	1487423564
0365F7AE-5048-42B3-BB2C-8E637A380A3E	557166	1_3	view	1487424075
0365F7AE-5048-42B3-BB2C-8E637A380A3E	555824	1_1	view	1487424241
0365F7AE-5048-42B3-BB2C-8E637A380A3E	554390	1_10	view	1487424395
0365F7AE-5048-42B3-BB2C-8E637A380A3E	557166	1_3	deep_view	1487424175
0365F7AE-5048-42B3-BB2C-8E637A380A3E	555824	1_1	deep_view	1487424345
0365F7AE-5048-42B3-BB2C-8E637A380A3E	557082	1_3	deep_view	1487423680
0365F7AE-5048-42B3-BB2C-8E637A380A3E	554390	1_10	deep_view	1487424422
06068254-792D-4AFE-AC6C-DE43DB15D735	556134	1_3	view	1487417354
06068254-792D-4AFE-AC6C-DE43DB15D735	556134	1_3	deep_view	1487417411

10 of 1199730 rows, 5 columns

2. news_info.csv 候选资讯内容数据 901KB

item_id	cate_id
287144	1_6
378467	1_17
378464	3_7
378465	1_10
287140	1_16
287141	1_16
287142	3_7
287143	3_2
287148	1_1
287149	1_1

10 of `56342` rows, `2` columns

> 如果您对数据集或比赛有疑问或建议，请前往比赛页面下论坛板块进行提问和建议，谢谢！

更新记录

版本	更新时间	更新内容
当前版本	5月7日	1.团队人数由1人改成最多5人2.精简 `news_info.csv` 。3.更新赛事文档说明

关于 Kesci

(/apps/home_log/index.html#!/about/index)

联系我们

(/apps/home_log/index.html#!/about/contact)

加入我们

(/apps/home_log/index.html#!/about/job)

隐私政策

(/apps/home_log/index.html#!/about/privacy_policy)

产品介绍

数据竞赛

(/apps/home_log/index.html#!/about/competition)

K-Lab (/apps/home_log/index.html#!/about/lab)

K-学院

(/apps/home_log/index.html#!/about/college)

合作伙伴

上海交大网络信息中心 (http://net.sjtu.edu.cn/)

上海大数据联盟 (http://www.shbigdata.org.cn/)

上海云基地 (http://www.cloud-valley.com/)

星红桉 (http://www.star-v.com.cn/)

达观数据 (http://www.datagrand.com/)

BDP (https://www.bdp.cn/)

七牛云 (http://www.qiniu.com/)

UCloud (https://www.ucloud.cn)

DaoCloud (https://www.daocloud.io/)

聚合数据 (https://www.juhe.cn/)

城市数据派 (http://www.udparty.com/)

微信公众平台