中国高校计算机大赛

2021 中国高校计算机大赛 —— 微信大数据挑战赛

通　知

2016 年，教育部高等学校计算机类专业教学指导委员会、教育部高等学校软件工程专业教学指导委员会、教育部高等学校大学计算机课程教学指导委员会、全国高等学校计算机教育研究会联合创办了“中国高校计算机大赛”（China Collegiate Computing Contest，简称 C4），第五届（2020 年）“中国高校计算机大赛”继续由全国高等学校计算机教育研究会主办。大数据挑战赛是其中的一项重要赛事，在 2018-2020 年均入选全国普通高校学科竞赛排行榜，获得社会各界的高度关注和广泛好评。

2021 中国高校计算机大赛——微信大数据挑战赛（以下简称“大赛”）是由清华大学和腾讯微信事业群联合举办，腾讯云提供竞赛平台和资源支持，以企业真实场景和实际脱敏数据为基础，面向全球开放的高端算法竞赛。大赛旨在通过竞技的方式，提升人们对数据分析与处理的算法研究与技术应用能力，探索大数据的核心科学与技术问题，尝试创新大数据技术，推动大数据的产学研用。

本次大赛面向全球开放，不限年龄国籍，高等院校在校学生（包括高职高专、本科生、研究生）以及科研机构和企业从业人员均可报名参赛。参赛队伍根据赛题要求设计相应的算法进行数据分析和处理，比赛结果按照指定的评价指标使用在线评测数据进行评测和排名，得分最优者获胜。

请各学校积极配合，按照通知和大赛章程做好宣传和组织工作，为在校生和毕业生参与竞赛提供必要的条件和支持。

竞赛详情见附件（2021 大数据挑战赛竞赛规程）。

全国高等学校计算机教育研究会

2021 年 4 月

2021 中国高校计算机大赛——微信大数据挑战赛
竞赛规程（2021.5.11 更新)

2016 年，教育部高等学校计算机类专业教学指导委员会、教育部高等学校软件工程专业教学指导委员会、教育部高等学校大学计算机课程教学指导委员会、全国高等学校计算机教育研究会联合创办了“中国高校计算机大赛”（China Collegiate Computing Contest，简称 C4），目前“中国高校计算机大赛”继续由全国高等学校计算机教育研究会主办。大数据挑战赛是其中的一项重要赛事，在 2018-2020 年期间均入选全国普通高校学科竞赛排行榜，获得社会各界的高度关注和广泛好评。

2021 中国高校计算机大赛——微信大数据挑战赛（以下简称“大赛”）由清华大学和腾讯微信事业群联合举办，由腾讯云提供大赛资源支持。本次大赛是以企业真实场景和实际脱敏数据为基础、面向全球开放的高端算法竞赛。大赛旨在通过竞技的方式，提升人们对数据分析与处理的算法研究与技术应用能力，探索大数据的核心科学与技术问题，尝试创新大数据技术，推动大数据的产学研用。

一、参赛对象

本次大赛面向全球开放，不限年龄国籍，高等院校在校学生（包括高职高专、本科、研究生）以及科研机构和企业从业人员均可参赛。具体要求如下：

- 可以自由组队参赛，具体组队要求见赛道相关说明；
- 参赛选手应保证报名信息准确有效，如队伍中的选手信息不符合要求，组委会有权取消整个队伍的参赛资格及奖励。

为了保证大赛的公平性，将禁止以下类型人员报名参赛：

- 大赛主办和技术支持单位如有机会接触赛题和相关数据的人员不允许参赛。
- 赞助企业的在职人员（不含实习生）不允许参赛。

二、赛制说明

本次大赛分为报名&组队、初赛、复赛和决赛等四个阶段，其中初赛阶段由参赛队伍下载数据在本地进行算法设计和调试，并通过大赛报名官网提交结果文件；复赛阶段要求参赛队伍在大赛官网平台上进行数据处理、算法调试和生成结果，数据不可下载，可使用平台提供的计算资源和工具包；决赛要求参赛者进行现场演示和答辩。本次大赛所提供数据均为脱敏数据。

1. 报名&组队（4 月 30 日 – 6 月 28 日)

参赛选手须在大赛官网报名并且组队参赛（即使单人参赛也要组建单人队伍），大赛不收取任何报名费用。大赛报名系统开放时间为北京时间 2021 年 4 月 30 日 10:00，截止时间为北京时间 2021 年 6 月 28 日中午 12:00。

- 报名方式：登录比赛官网，完成个人信息注册，即可报名参赛；
- 每个选手可单人成队或 2-3 人组队参赛，且每人只能参加一支队伍。

大赛官方渠道主要包括：

- 大赛官网：https://algo.weixin.qq.com/
- 大赛邮箱：data@tsinghua.edu.cn
- 大赛 QQ 群：762146461 / 304389749 / 616504318

报名截止之后，不再允许添加或更改任何队伍成员。如有中途退出情况，只允许在参赛队伍内部更换队长或删除队员。参赛队伍须应在决赛开始前向大赛组委会提交成员更换申请，由参赛队伍全部成员亲笔签名，经由大赛组委会审核后变更生效。

2. 初赛（5 月 20 日 – 6 月 30 日)

参赛队伍可从大赛官方网站下载数据，在本地进行算法调试，并在线提交结果。

5 月 20 日 12:00 开始，选手可以从竞赛平台下载初赛训练数据集，用于参赛队伍训练模型以及制定预估策略；同时，平台提供测试数据集，用于参赛队伍在比赛中的模型评估和排名。

初赛采用 AB 榜形式：

- 初赛 A 阶段：5 月 20 日 10:00 – 6 月 29 日 20:00，每个参赛队伍每天可以有 3 次提交结果机会，系统实时评测并返回成绩。排行榜每小时更新，将选择参赛队伍在本阶段的历史最优成绩，按照评测指标从高到低排序。
- 初赛 B 阶段：6 月 30 日 10:00-20:00。系统将在 6 月 29 日 21:00 更换测试数据，参赛队伍需再次下载数据文件。本阶段提供 2 次提交结果的机会，系统进行实时评测并返回成绩。排行榜每小时进行更新，并选择参赛队伍在本阶段的历史最优成绩进行排名展示。

初赛提交的截止时间是 6 月 30 日 20:00，初赛以 B 榜成绩作为初赛成绩依照，要求 TOP120 团队提交代码审核，具体要求届时通知，代码提交截止时间 7 月 2 日 12:00。组委会将审核并取消存在人工标注、相互抄袭等行为队伍的比赛资格，晋级空缺名额后补。初赛成绩符合要求且通过实名认证的排名前 100 名的参赛队伍将进入复赛，所有通过审核队伍将获得初赛名次证书。

3. 复赛（7 月 12 日 – 8 月 9 日)

复赛参赛队伍需要在大赛官网平台上完成数据处理、建模、算法调试、生成结果等，所有比赛数据不可下载，可使用平台提供的计算资源和工具包。

复赛采用 AB 榜形式：

- 复赛 A 阶段：7 月 12 日 12:00 – 8 月 6 日 12:00，每个参赛队伍每天可以有 3 次提交结果机会，系统实时评测并返回成绩。排行榜每小时更新，将选择参赛队伍在本阶段的历史最优成绩，按照评测指标从高到低排序。

- 复赛 B 阶段：8 月 2 日 12:00-8 月 9 日 12:00。8 月 2 日 12:00 开始，竞赛平台提供最终成绩的测试数据集，此数据集仅用于复赛队伍的最终成绩评比和排名，不参与比赛过程中每天排名。参赛队伍可以随时提交该数据集的结果，如有多次提交则以最后 2 次提交为准。8 月 9 日 13:00，系统对参赛队伍提交的最终测试集结果进行评测，并根据 2 次评测结果取高分，公布所有复赛队伍的最终成绩和排名。

复赛提交的截止时间是 8 月 9 日 12:00，复赛以 B 榜成绩作为复赛成绩依照，TOP35 团队需要提交代码审核，具体要求届时通知，代码提交截止时间是 8 月 10 日 21:00。组委会将审核并剔除只靠人工标注而没有算法贡献的队伍，晋级空缺名额后补，最终通过复赛成绩审核的前 6 名队伍将晋级决赛。

4. 决赛（8 月下旬）

决赛将以现场答辩会的形式进行，具体要求和安排另行通知。受邀参加决赛的选手在决赛期间的食宿由大赛组委会安排，往返交通费及其他费用自理。

晋级决赛团队需提前准备答辩材料，包括路演 PPT、参赛总结、算法核心代码。在决赛答辩会上，每支队伍面对评委有 20 分钟的路演时间和 10 分钟的答辩时间。评委将根据选手的技术思路、理论深度和现场表现进行综合评分。

决赛分数将根据参赛队伍的算法成绩和答辩成绩加权得出，评分权重为复赛阶段 70%，决赛答辩 30%。

三、奖项设置

1. 初赛奖项

初赛 TOP120 且通过代码审核的团队将颁发初赛名次证书。

2. 复赛与决赛奖项

大赛奖金池总额为 66 万元人民币，所有奖金均为税前金额。

奖项名称	数量	对象	奖励办法
全国一等奖	1	决赛第 1 名	证书，奖金 30 万元
	1	决赛第 2 名	证书，奖金 10 万元
	1	决赛第 3 名	证书，奖金 7 万元
全国二等奖	3	决赛第 4-6 名	证书，奖金 3 万元
	4	复赛第 7-10 名	证书，奖金 1 万元

全国三等奖	20	复赛第 11-30 名	证书，奖金 3 千元
优胜奖	30	复赛第 31-60 名	证书

3. 周周星

在每个赛道的初赛阶段，设立周周星奖励。从初赛第三周开始，以每周一中午 12 点的排行榜为准，取前两名参赛队伍发放周周星纪念礼物；对于前面已经获得周周星的队伍，不重复发放，名额按名次顺延。

4. 其他激励

招聘绿色通道：复赛排名前 30 队伍的在校学生将获得 2021 年腾讯集团微信事业群校园招聘和实习招聘绿色通道资格，具体细则另行通知。

四、违规处理

参赛者应本着诚实、公平的态度参加比赛，如在以下情况出现违规，大赛组织委员会（简称“大赛组委会”）有权取消参赛者所在队伍的参赛资格，情节严重者将通报参赛者所在高校并追究其违法责任。

1. 账号使用：参赛者有义务保证账号信息的真实性和有效性，且账号仅限于参赛者本人使用；参赛者禁止使用多账号参赛，同一参赛者不可使用多个账号进行提交、刷分操作；如根据判断认为参赛账号存在异常或违背正常使用条例，组委会可以单方面暂停或终止该账号登录大赛平台。
2. 比赛成果：
    - 严禁参赛队伍之间相互抄袭。如不同参赛队伍提交结果高度相似，经判定存在抄袭行为的，组委会将取消相关参赛队伍的参赛资格，相关参赛成绩无效。
    - 参赛者应保证其在比赛过程中所产出的所有成果未侵犯任何第三方的知识产权、商业秘密及其他合法权益。如第三方因为参赛者侵权行为提出索赔、诉讼等，参赛者应承担由此产生的全部责任及损失。
    - 如大赛主办方及其关联公司有意取得参赛者在本次大赛中独立开发的依约定享有完整知识产权的研究成果，参赛者同意大赛举办方及其关联公司在同等条件下享有优先受让权，相关转让事宜由双方另行协商确定。
3. 数据使用：对于大赛提供的数据（数据集），参赛者须仅在比赛场景下使用，并应妥善保存已下载的数据（数据集），避免泄露；在完成比赛使用后应及时销毁已下载数据（数据集）；如使用比赛之外的任何数据应获得组委会许可。对于不提供下载的比赛数据，参赛者不得以任何形式擅自复制、下载或获取。参赛者如发现任何出现数据未授权访问的可能，应立即通知组委会并积极提供相关信息。

4. 代码分享：在大赛举办期间，未经组委会同意，参赛者禁止公开分享与赛事相关的数据、模型和代码；大赛结束之后，参赛者可以在拥有模型和代码的知识产权的情况下自行选择公开分享，但需要确保此类公开共享不会侵犯任何第三方的知识产权、商业秘密及其他合法权益。
5. 参赛者若在参赛过程中发现相关规则漏洞或技术漏洞，有义务及时告知组委会相关漏洞的信息，组委会将对提供相关信息的参赛者表示相关感谢；若参赛者利用相关漏洞进行参赛，经判断查证后，成绩将会被判断为无效成绩。

五、申诉与仲裁

1. 参赛团队或选手对不符合大赛规定的设备、工具和软件，有失公正的评判和奖励以及工作人员的违规行为等，均可向大赛组委会提出申诉。组委会负责受理比赛中提出的申诉并进行调解仲裁，以保证大赛的顺利进行和大赛结果的公平公正。组织委员会作出的仲裁结果为终局决定。
2. 申诉报告应明确申诉内容，指定一名成员作为联系人，通过大赛邮箱以邮件发送，否则申诉将不予以受理。
3. 组织委员会将在收到申诉之日起 5 个工作日之内受理，并认真核查和处理。

六、其他说明

1. 在大赛举办过程中，竞赛规程可能会有少量的变更和调整，所有内容均以大赛官网为准。
2. 本大赛规程的最终解释权归“中国高校计算机大赛——微信大数据挑战赛”组织委员会所有。

“中国高校计算机大赛——微信大数据挑战赛”组织委员会

2021 年 4 月

附件：赛题描述——微信视频号推荐算法（2021.5.11 更新）

本次比赛基于脱敏和采样后的数据信息，对于给定的一定数量到访过微信视频号“热门推荐”的用户，根据这些用户在视频号内的历史 n 天的行为数据，通过算法在测试集上预测出这些用户对于不同视频内容的互动行为（包括点赞、点击头像、收藏、转发等）的发生概率。本次比赛以多个行为预测结果的加权 uAUC 值进行评分。

一、 竞赛数据

比赛提供训练集用于训练模型，测试集用于评估模型效果，提交结果 demo 文件用于展示提交结果的格式。所有数据文件格式都是带表头的.csv 格式，不同字段列之间用英文逗号分隔。初赛与复赛的数据分布一致，数据规模不同。初赛提供百万级训练数据，复赛提供千万级训练数据。

1. 训练集

（1）Feed 信息表

该数据包含了视频（简称为 feed）的基本信息和文本、音频、视频等多模态特征。具体字段如下：

字段名	类型	说明	备注
feedid	String	Feed 视频 ID	已脱敏
authorid	String	视频号作者 ID	已脱敏
videoplayseconds	Int	Feed 时长	单位：秒
description	String	Feed 配文，以词为单位使用空格分隔	已脱敏；存在空值
ocr	String	图像识别信息，以词为单位使用空格分隔	已脱敏；存在空值
asr	String	语音识别信息，以词为单位使用空格分隔	已脱敏；存在空值
description_char	String	Feed 配文，以字为单位使用空格分隔	已脱敏；存在空值
ocr_char	String	图像识别信息，以字为单位使用空格分隔	已脱敏；存在空值
asr_char	String	语音识别信息，以字为单位使用空格分隔	已脱敏；存在空值
bgm_song_id	Int	背景音乐 ID	已脱敏；存在空值
bgm_singer_id	Int	背景音乐歌手 ID	已脱敏；存在空值
manual_keyword_list	String	人工标注的关键词，多个关键词使用英文分号”;”分隔	已脱敏；存在空值
machine_keyword_list	String	机器标注的关键词，多个关键词使用英文分号”;”分隔	已脱敏；存在空值

manual_tag_list	String	人工标注的分类标签，多个标签使用英文分号";"分隔	已脱敏；存在空值
machine_tag_list	String	机器标注的分类标签，多个标签使用英文分号";"分隔	已脱敏；存在空值
feed_embedding	String	融合了 ocr、asr、图像、文字的多模态的内容理解特征向量	512 维向量

说明：

- 训练集和测试集涉及的 feed 均在此表中；
- description，orc，asr 三个字段为原始文本数据以词为单位使用空格分隔和脱敏处理后得到的。例如：文本“我参加了中国高校计算机大赛“经过处理后得到类似“2 32 100 25 12 89 27”的形式（此处只是一个样例，不代表实际脱敏结果）。此外，我们还提供了以字为单位使用空格分隔和脱敏的结果，对应的字段分别为 description_char、ocr_char、asr_char。
- machine_tag_list 字段比 manual_tag_list 字段增加了每个标签对应的预测概率值（取值区间[0,1]）。脱敏后的标签和概率值之间用空格分隔。例如：“1025 0.32657512;2034 0.87653981;35 0.47265462”。
- manual_keyword_list 和 machine_keyword_list 共享相同的脱敏映射表。如果原先两个字段都包含同个关键词，那么脱敏后两个字段都会包含同个 id。
- manual_tag_list 和 machine_tag_list 共享相同的脱敏映射表。如果原先两个字段都包含同个分类标签，那么脱敏后两个字段都会包含同个 id。
- feed_embedding 字段为 String 格式，包含 512 维，数值之间用空格分隔。

（2）用户行为表

该数据包含了用户在视频号内一段时间内的历史行为数据（包括停留时长、播放时长和各项互动数据）。具体字段如下：

字段名	类型	说明	备注
userid	String	用户 ID	已脱敏
feedid	String	Feed 视频 ID	已脱敏
device	Int	设备类型 ID	已脱敏
date_	Int	日期	已脱敏为 1-n，n 代表第 n 天
timestamp_	Int	时间戳	已脱敏，并保持原来的先后顺序
play	Int	视频播放时长	单位：毫秒；若播放时长大于视频时长，则属于重播的情况
stay	Int	用户停留时长	单位：毫秒
read_comment	Bool	是否查看评论	取值{0, 1}，0 代表“否”，1 代表“是”
like		是否点赞	
click_avatar		是否点击头像	

favorite		是否收藏	
forward		是否转发	
comment		是否发表评论	
follow		是否关注	

2. 测试集

比赛 A/B 榜的数据量和数据分布一致。具体字段如下：

字段名	类型	说明	备注
userid	String	用户 ID	已脱敏
feedid	String	Feed 视频 ID	已脱敏
device	Int	设备类型 ID	已脱敏

3. 提交结果格式

- 初赛阶段：选手需要对测试集中每一行的 userid 和 feedid 的四种互动行为的发生概率进行预测，这四种行为包括：查看评论、点赞、点击头像、转发；
- 复赛阶段：选手需要对测试集中每一行的 userid 和 feedid 的七种互动行为的发生概率进行预测，这七种行为包括：查看评论、点赞、点击头像、转发、收藏、评论和关注。

具体格式说明如下：

字段名	类型	说明	赛段	备注
userid	String	用户 ID	初赛/复赛	已脱敏
feedid	String	Feed 视频 ID		已脱敏
read_comment	Float	是否查看评论		预测用户特定行为发生的概率，取值区间[0,1]。0 代表“否”，1 代表“是”；结果最多保留六位小数。
like	Float	是否点赞		
click_avatar	Float	是否点击头像		
forward	Float	是否转发		
favorite	Float	是否收藏	仅复赛	
comment	Float	是否发表评论		
follow	Float	是否关注		

说明：提交结果文件的行数需要与测试集的行数相同，且 userid-feedid 需要与测试集中一致（顺序可以不同）。

二、 评估标准

本次比赛采用 uAUC 作为单个行为预测结果的评估指标，uAUC 定义为不同用户下 AUC 的平均值，计算公式如下：

$$uAUC = \frac{1}{n}\sum_{i=1}^{n} AUC_i$$

其中，n 为测试集中的有效用户数，有效用户指的是对于某个待预测的行为，过滤掉测试集中

全是正样本或全是负样本的用户后剩下的用户。AUC_i为第 i 个有效用户的预测结果的 AUC（Area Under Curve）。AUC 的定义和计算方法可参考维基百科。

初赛的最终分数为 4 个行为（查看评论、点赞、点击头像、转发）的 uAUC 值的加权平均。复赛的最终分数为 7 个行为（查看评论、点赞、点击头像、转发、收藏、评论和关注）的 uAUC 值的加权平均。分数越高，排名越靠前。

在总分中，7 个行为的权重如下:

字段名	字段说明	权重
read_comment	是否查看评论	4
like	是否点赞	3
click_avatar	是否点击头像	2
forward	是否转发	1
favorite	是否收藏	1
comment	是否发表评论	1
follow	是否关注	1

加权 uAUC 的计算公式如下:

$$uAUC_{weighted} = \frac{\sum_{i=1}^{k} uAUC_i * W_i}{\sum_{i=1}^{k} W_i}$$

其中，k 为待预测的互动行为数，初赛 k=4，复赛 k=7。$uAUC_i$ 为第 i 个行为的 uAUC 值，W_i 为第 i 个行为的权重。

三、 其他说明

1. 本项比赛全程不允许使用外部数据集。
2. 允许使用开源的词典、embedding 和预训练模型，以上数据和模型需在 2021/07/12 日期前开源，且需通过邮件的形式向组委会报备开源链接地址和 md5，报备邮箱为 wechat_algo@tencent.com。
3. 复赛阶段允许使用初赛阶段的数据集。