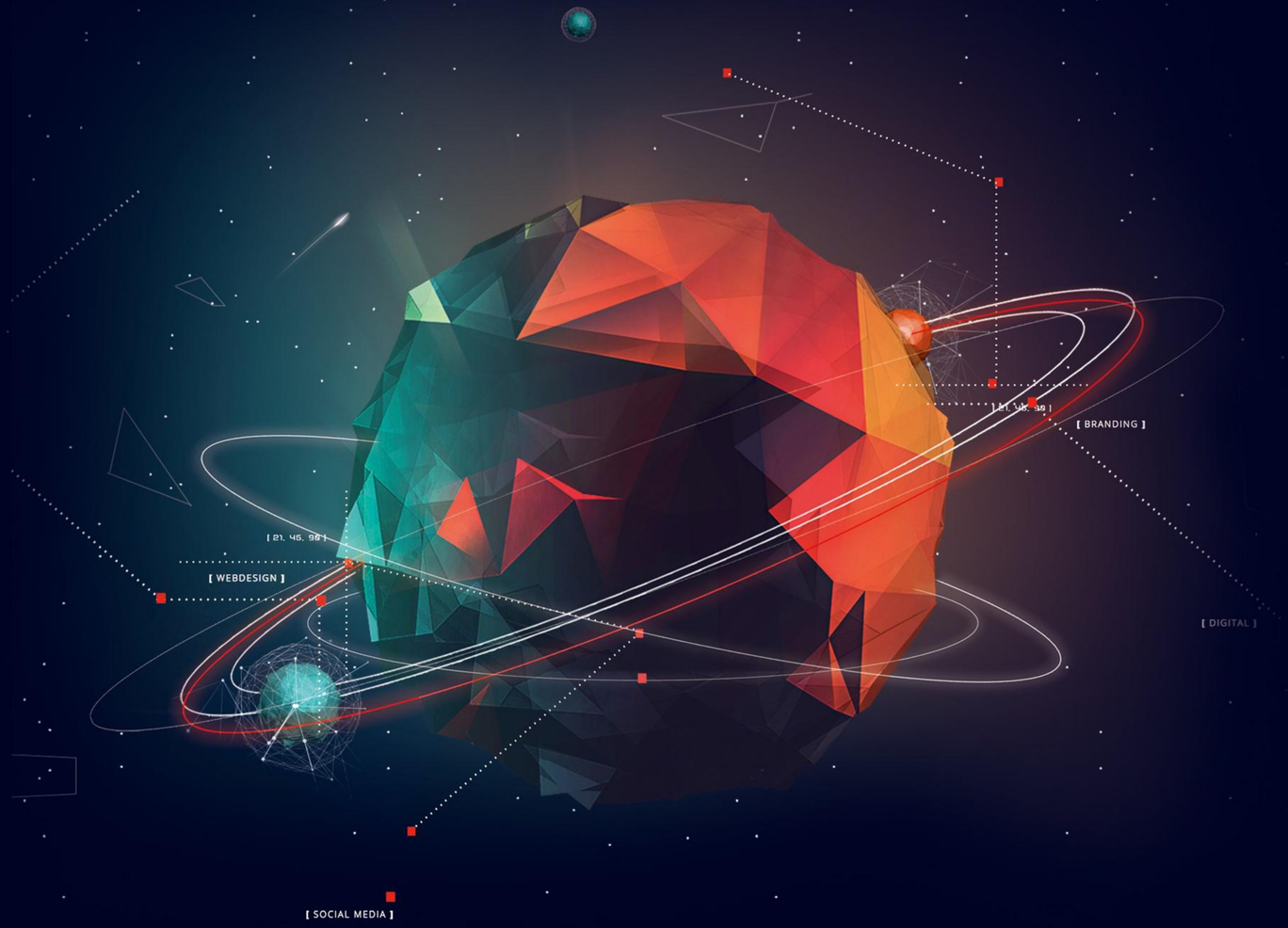
中国高校计算机大赛
CHINA COLLEGIATE COMPUTING CONTEST

2019高校大数据挑战赛答辩

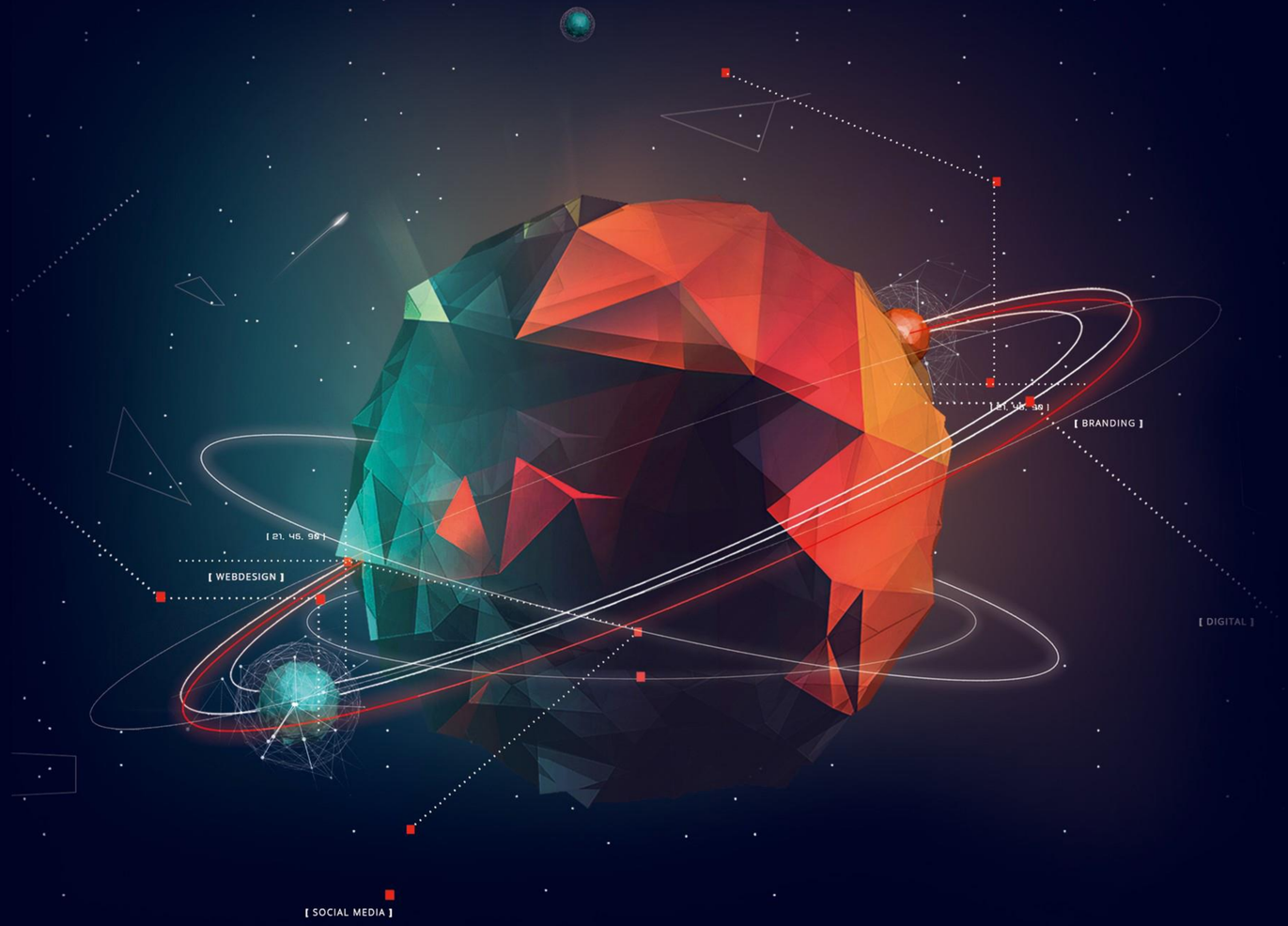
答辩团队：蜗牛本牛

中国高校计算机大赛
CHINA COLLEGIATE COMPUTING CONTEST



目录

中国高校计算机大赛
CHINA COLLEGIATE COMPUTING CONTEST



赛题及数据分析

赛题及数据分析

（一）赛题分析——文本点击率预估

搜索中一个重要的任务是根据query和title预测query下doc点击率，本次大赛参赛队伍需要根据脱敏后的数据预测指定doc的点击率。训练集：10亿 测试集1：2kw 测试2：1亿

评价指标

$$qAUC = \frac{sum(AUC_i)}{query\_num}$$

其中AUC_i 为同一个query_id下的AUC

分析：AUC属于排序指标，可以考虑用ranking模型去做；数据量大，计算资源有限，如何合理的选择数据进行快速迭代是提分的关键

query_id	query_title_id	label/preds	
1	1	*	AUC_1
	2	*	
	3	*	
	4	*	
2	1	*	AUC_2
	.	*	
	.	*	
	10	*	
…			
N	1	*	AUC_N
	.	*	
	.	*	
	20	*	

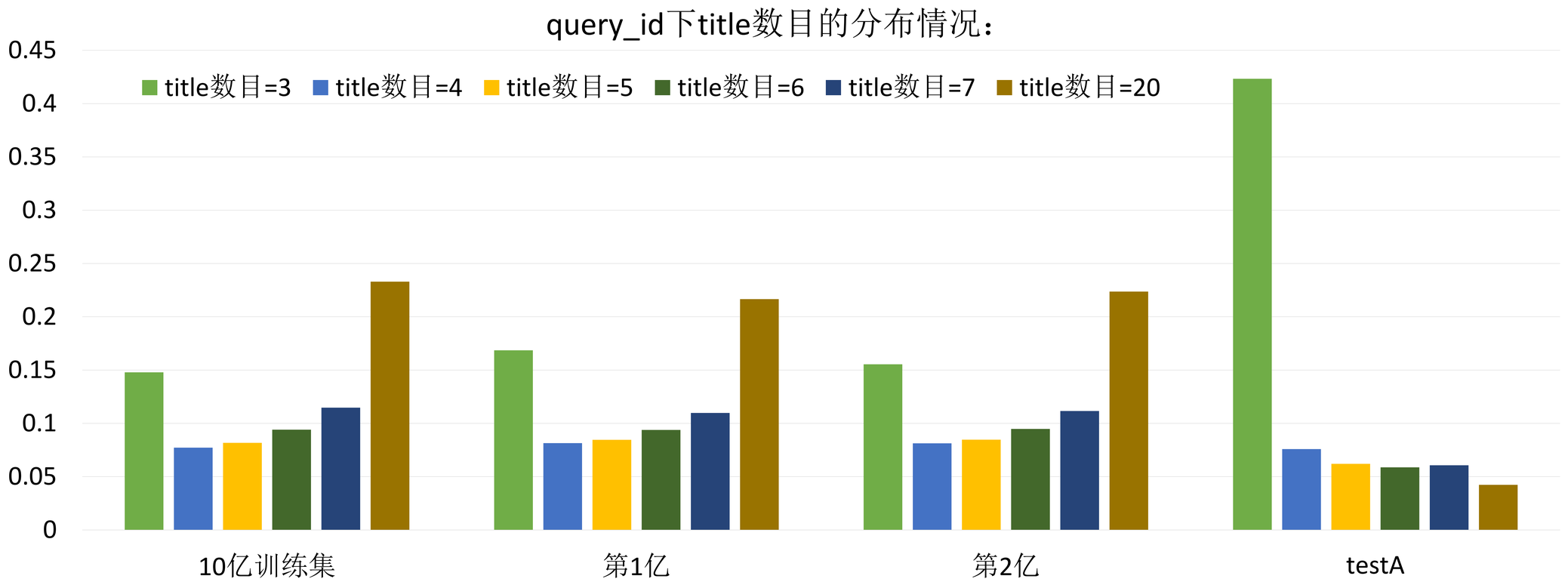
赛题及数据分析

训练集和A榜测试集的query_id下title数目分布不一致造成：线上线下得分差异大

B榜测试集和训练集分布相近，因此B榜和A榜的分数有较大的差异

（二）数据分析

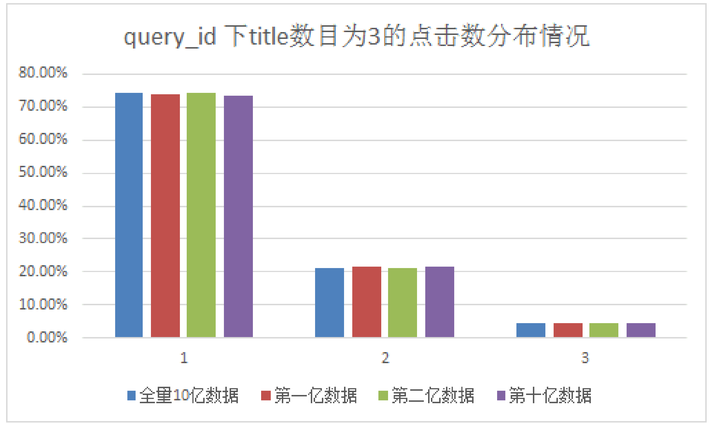
2.1 query_id下title数目的分布情况：

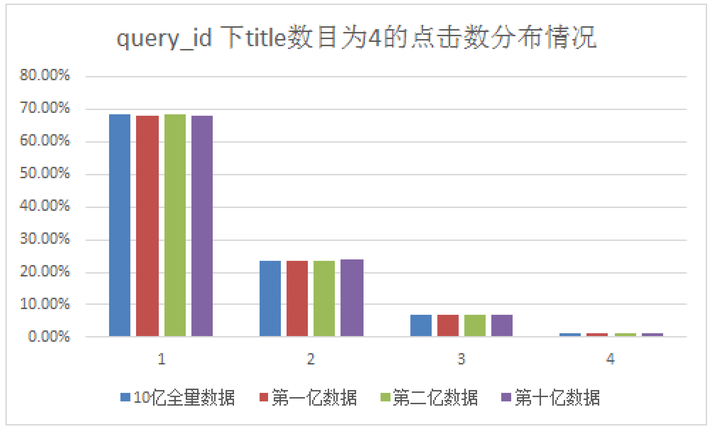


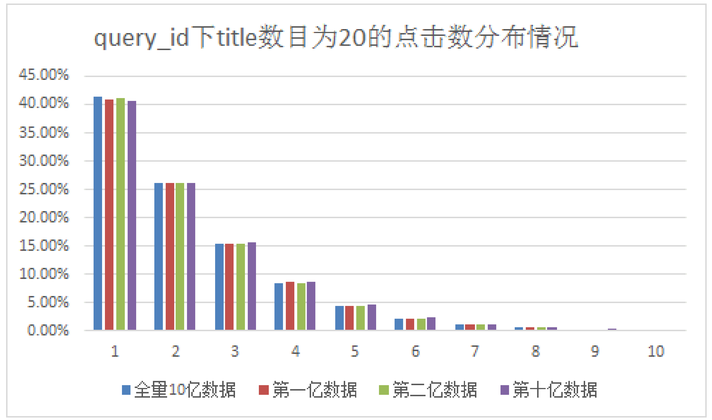
赛题及数据分析

（二）数据分析

2.2 点击数分布情况：





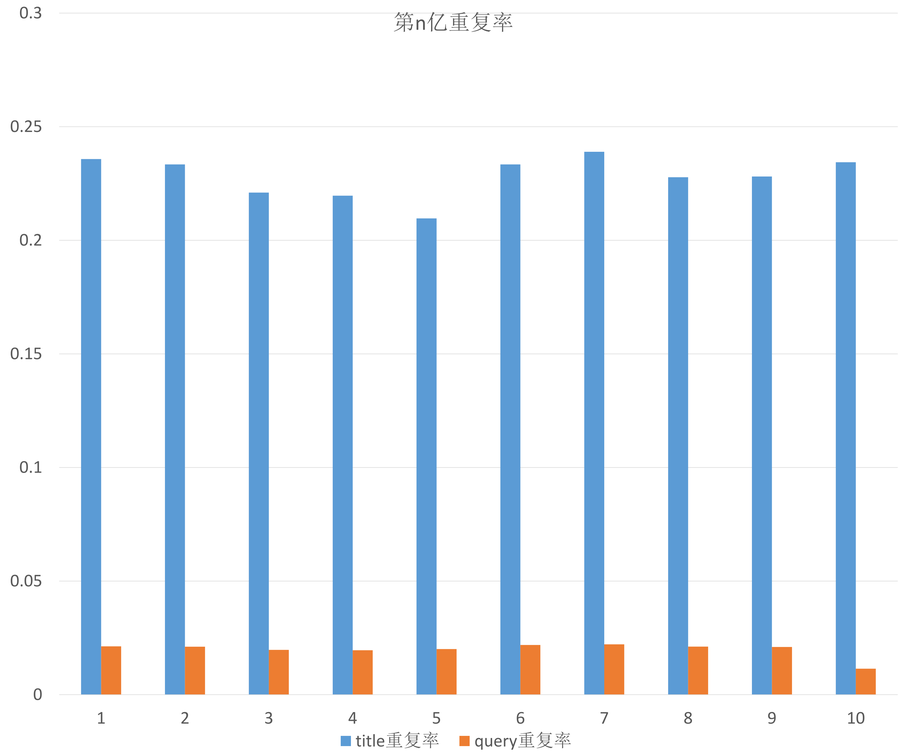


任意连续的1亿数据点击数的数据分布和整体10亿数据的分布大体一致

赛题及数据分析

2.3 title重复分布情况：

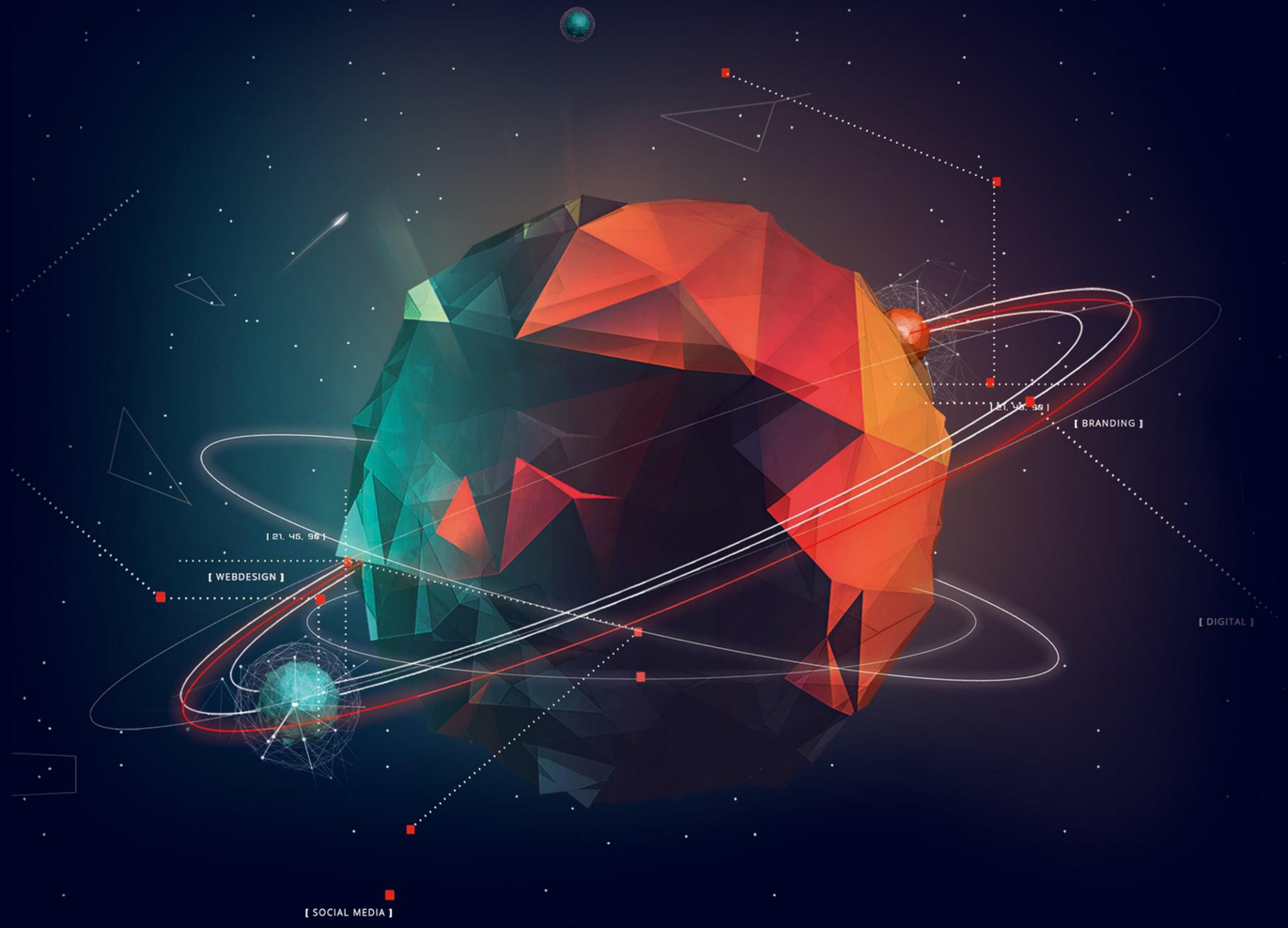
testA 2kw数据与train重复的query占比39.25%，title有71.18%，query-title对: 29.189%



1）对于重复title：文本信息+历史信息

2）对于新的title：文本和搜索领域相关特征比较重要

中国高校计算机大赛
CHINA COLLEGIATE COMPUTING CONTEST



特征工程

特征工程

（一）总体特征

文本特征

词粒度匹配特征:

1、匹配词开始、结束位置
2、匹配词数量、最大连续匹配长度
3、Proximity
4、query、title词长度及词比率、和、差
5、不同词数、相同词数、不同词差
.....

文本位置特征:

query每个词在title中的位置

.....

语义相似度特征

1、句向量的余弦、欧式等距离特征
2、编辑距离、 jaccard等
3、使用NN训练出来的有监督的语义特征

历史信息特征

count特征

1阶计数特征，如title共计出现次数

nunique特征

类别变量的nunique特征，
如title对应多少个query

CTR特征

1、tilte转化率
2、tag转化率

Title质量分数特征

使用以上特征lgb排序模型得分0.626，线上A榜TOP8

特征工程

（二）文本特征挖掘

1、细粒度的相似度特征

具体做法	对于相似度的计算只取query前n个词和title前m个词进行相似度计算 (比如计算Jaccard相似度，取query前10个词和title前10个词进行计算)

2、词位置特征

具体做法	query前10个词在title中的位置

3、SIF Embedding[1] (Smoothed Inverse Frequency Embedding)

主要思想	1）对组成sentence vec的word vec进行线性加权，权重为$a / (a + p_w)$。(其中p_w 是词频，a是超参数) 2）使用SVD移除语义无关的内容
提升效果	对比tf-idf weighted sentence vec 线上提升约2个千

[1] S. Arora, Y. Liang, and T. Ma. A simple but tough-to-beat baseline for sentence embeddings. In ICLR, 2017

特征工程

（三）历史信息特征挖掘

1、CTR特征

使用转化率特征的时候如何避免信息泄露？

对于CTR，使用了贝叶斯平滑

$$\hat{r} = \frac{C + \alpha}{I + \alpha + \beta}$$

CTR使用方式调研

方案	分数
5折不抽样	0.6012
5折0.5抽样	0.5958
10折不抽样	0.6014

特征工程

（三）历史信息特征挖掘

2、title质量分数特征

一个query中用户只点击了1篇title和一个query中用户点击了所有的title，二个query中被点击的title对于其query的相关性等价吗？

假设前提

1）一个query中用户点击了 i 篇title和另一个query中用户点击了 j（j！=i）篇title，二者被点击的title对于query的重要程度不一样。
2）一个query中用户有 i 篇title未点击和另一个query中用户有 j（j！=i）篇title未点击，二者未点击的title对于query的重要程度也不一样。

计算方式

1）对于点击的title，该title质量得分等于 1 - 该query下的title点击率
2）对于未点击的title，该title质量得分等于0 - 该query下的title点击率
3）title的最终得分是包含该title的所有query下该title的得分总和

计算公式

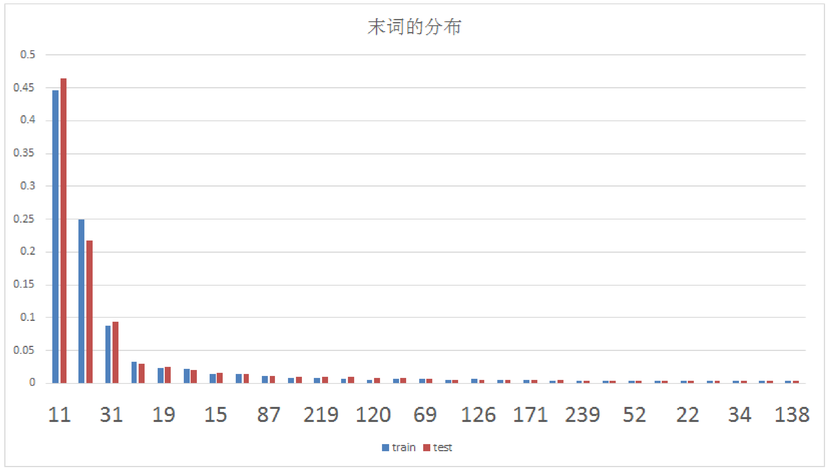
$$\text{Score}_{title} = \sum_{\text{包含}title\text{的}query} lable_{query} - \text{点击率}_{query}$$

效果情况

	初赛qAUC	复赛qAUC
title 质量分数特征的值作为预测结果	0.570615	0.572866

特征工程

列名	类型	示例
query_id	int，一个query的唯一标识	1
query	字符string，term空格分割	"字节跳动"
title	字符string，term空格分割	"字节跳动 百科"
label	int，取值{0, 1}，有点击为1，无点击为0	1



（四）末词（tag）的信息数据挖掘

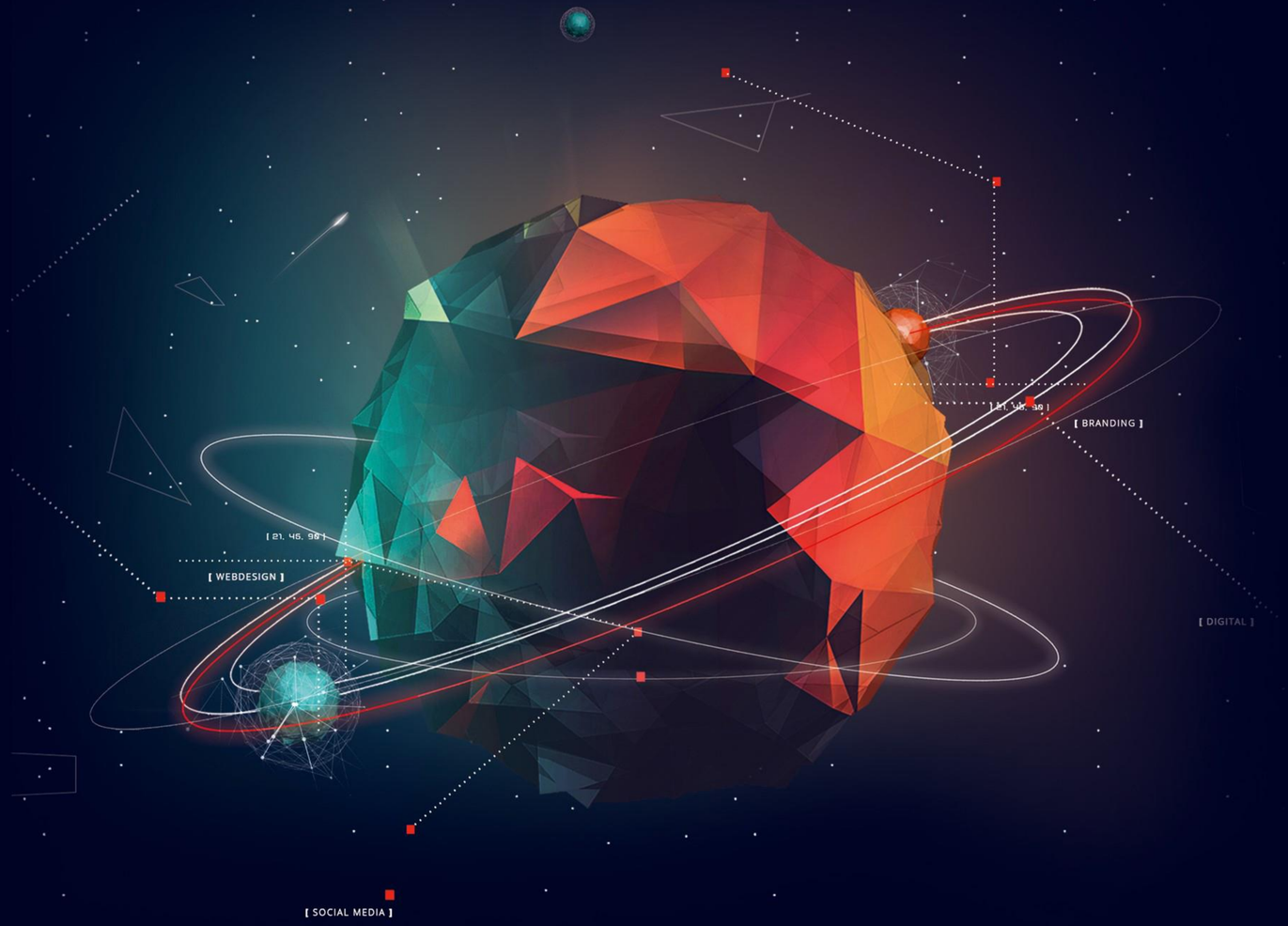
特征	特征解释
tag	tag特征
tag_appr_flag	tag是否在query中
tag_convert	tag转化率

备注：转化率为加贝叶斯平滑十折有序计算的点击率

提升效果

初赛时线上A榜提升约2个千

中国高校计算机大赛
CHINA COLLEGIATE COMPUTING CONTEST



算法模型方案

LGB模型方案

LGB二分类模型和排序模型得分对比情况

模型名称	使用数据	线下
LGB (二分类)	2kw	0.65028
LGB (PairWise)	2kw	0.65534 （5k提升）

深度学习模型方案

ESIM[2] (Enhanced LSTM for Natural Language Inference)

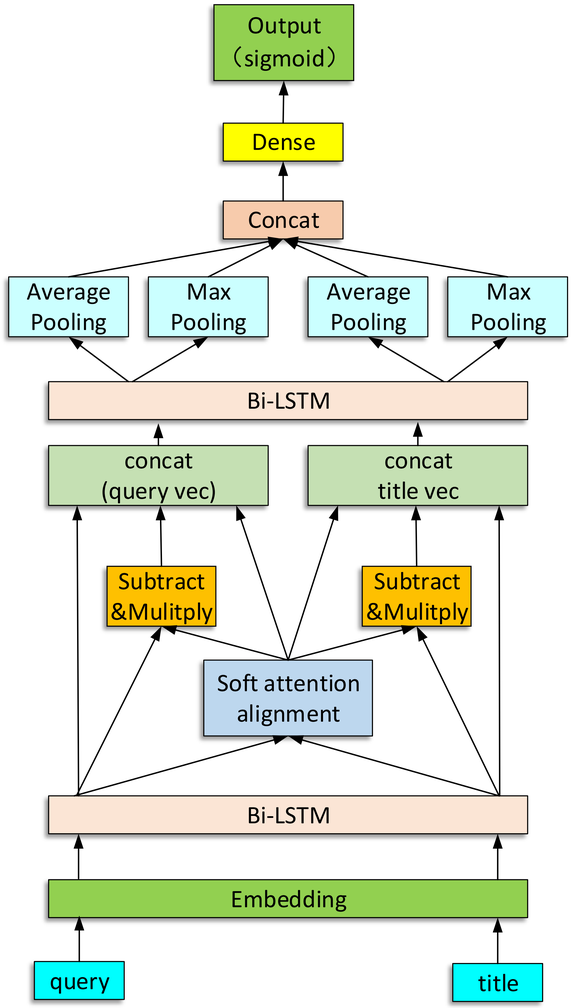
训练数据

为了防止过拟合以及增大差异性，ESIM模型采用第二个1亿的数据作为训练集，和LGB模型的训练数据形成区分。

效果情况

1、与LGB模型预测结果直接加权融合，融合线上提升5个千

2、预测结果作为LGB的深度语义匹配特征，LGB线上提升1个百

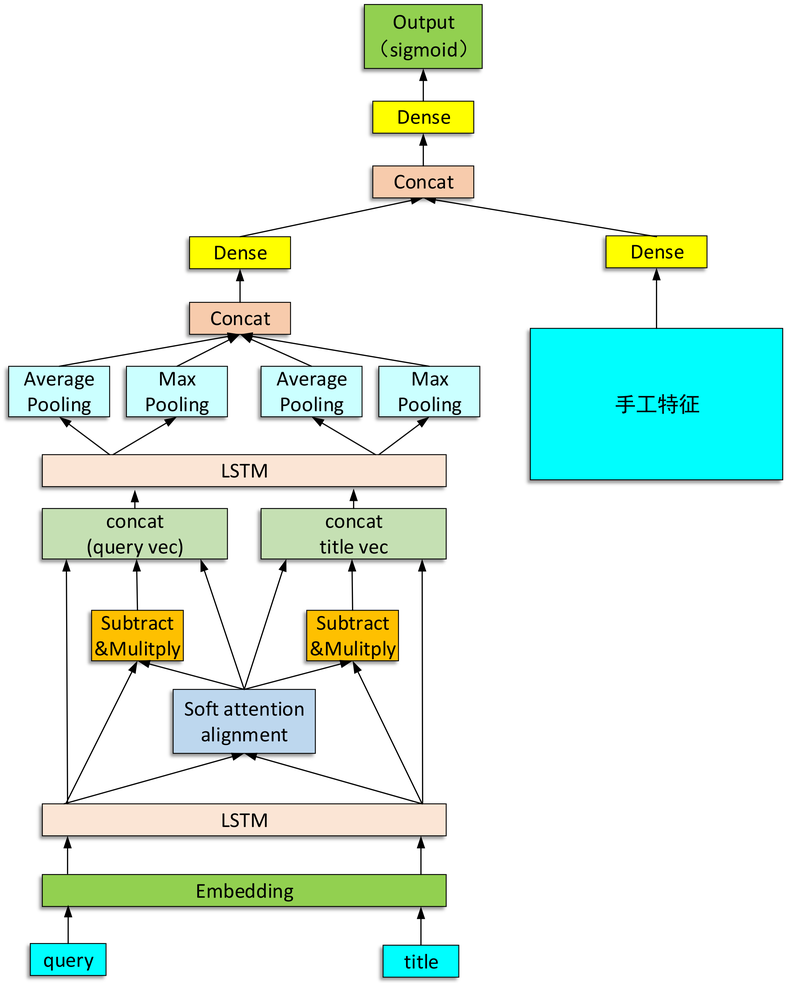


[2] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced LSTM for natural language inference. 2017

Concat-ESIM

在ESIM中以concat的方式引入其他特征

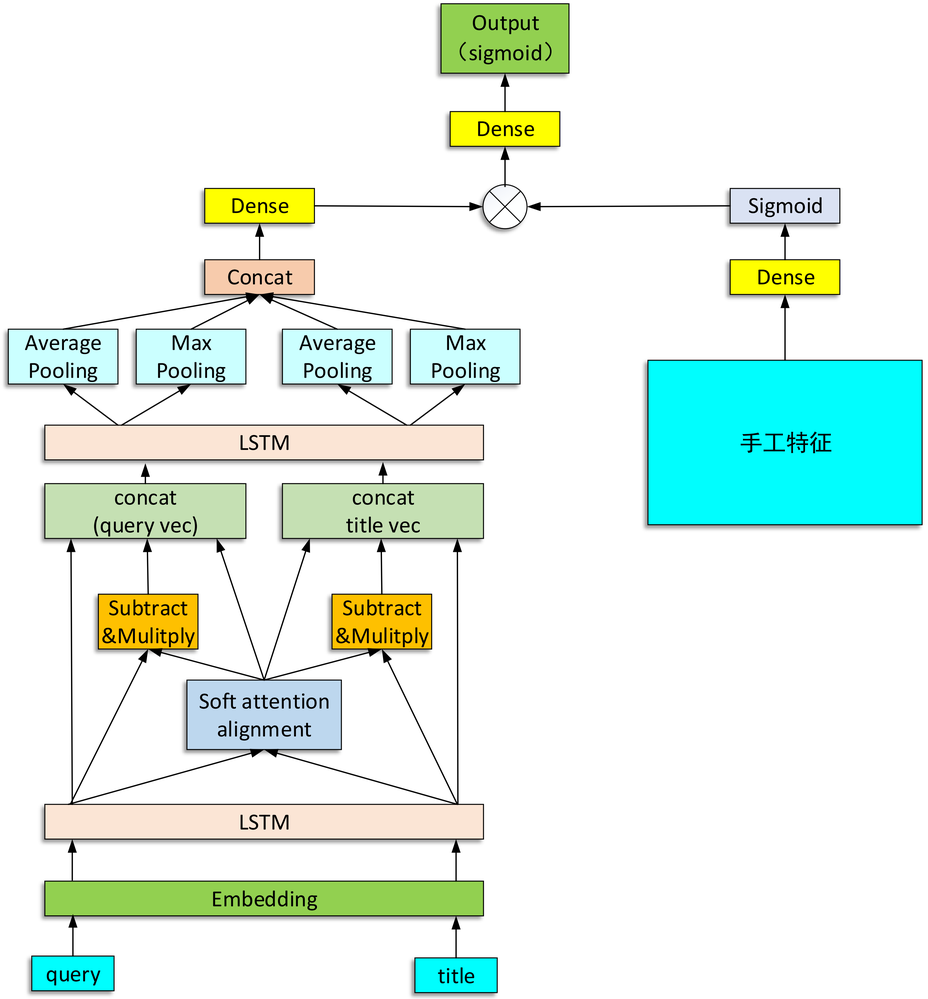
concat-esim加入10维统计特征，对比esim模型，线上提升约3个百



Gate-ESIM

Gate机制在Deep learning中是非常常见的，例如LSTM中的输入门，遗忘门；经过查阅相关资料，最终使用gate方案代替concat方案。

	concat-esim	gate-esim	gate-esim	gate-esim
数据量	2kw	2kw	1亿	1亿
手工特征数	10维	10维	34维	34维
LSTM隐层大小	32	32	128	128
线下评分	0.595702	0.6027(7个千)	0.6761	0.6778(1.7k)
线上评分	-	-	0.629903 (单模TOP3)	0.631+(预计)



深度学习训练Tricks

1. 学习率

Cyclical Learning Rate[3](CLR) 学习率在base_lr 和 max_lr之间进行周期性的变化，通过衰减函数控制

2. 正则

(1) 词向量embedding之后加dropout

(2) gate-esim特征FC层 使用1e-4的L2正则

3. 优化算法

Stochastic Weight Averaging(SWA)[4]

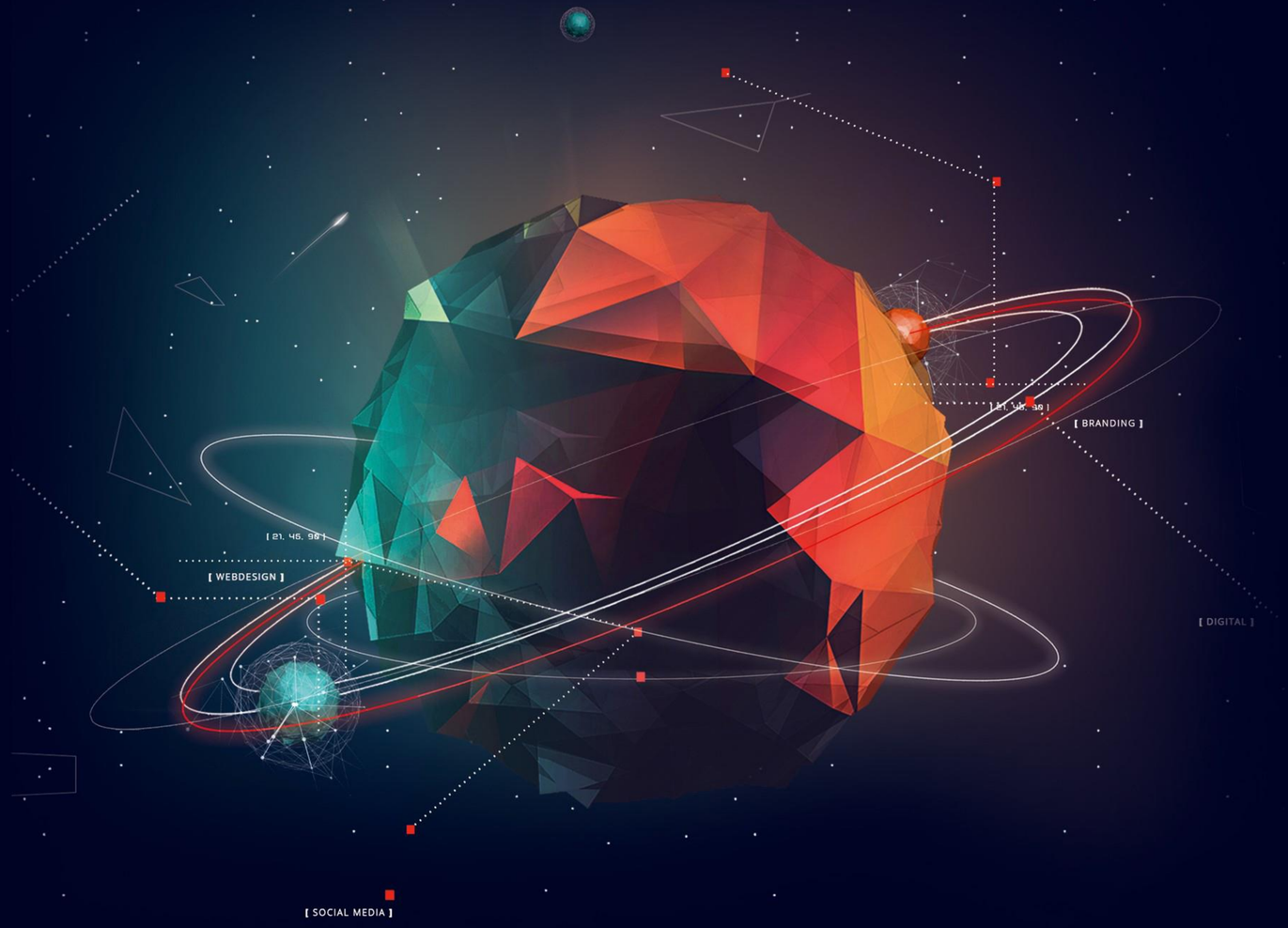
每次学习率循环结束时产生的局部最小值趋向于在损失面的边缘区域累积，这些边缘区域上的损失值较小。通过对几个这样的点取平均，得到一个甚至更低损失的、全局化的通用解。

4. 词向量

skip-gram模型训练的 word2vec 和 fasttext 各100维进行拼接

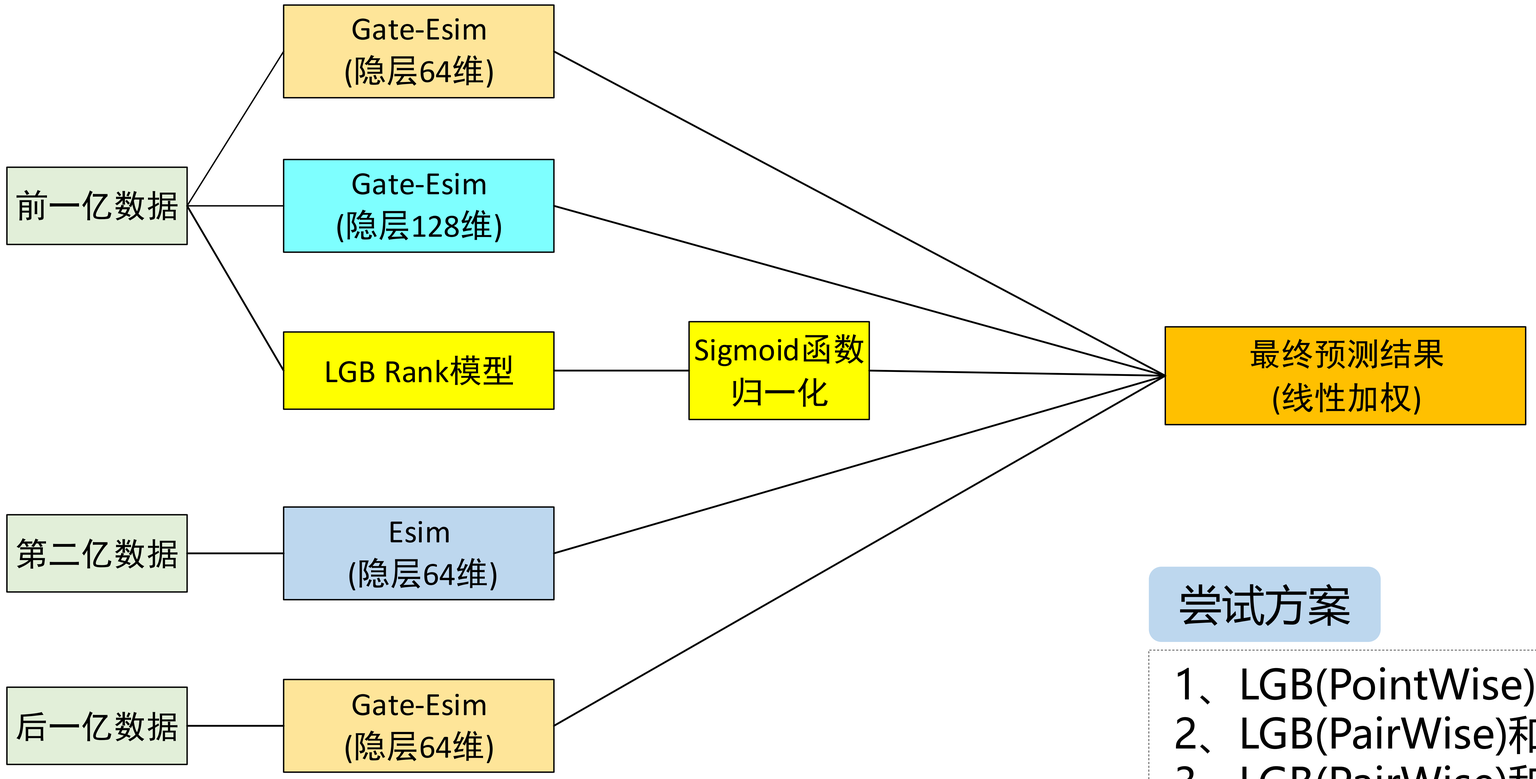
[3] Smith L N . Cyclical Learning Rates for Training Neural Networks. 2015.
[4] Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. Averaging Weights Leads to Wider Optima and Better Generalization. 2018

中国高校计算机大赛
CHINA COLLEGIATE COMPUTING CONTEST



模型融合方案

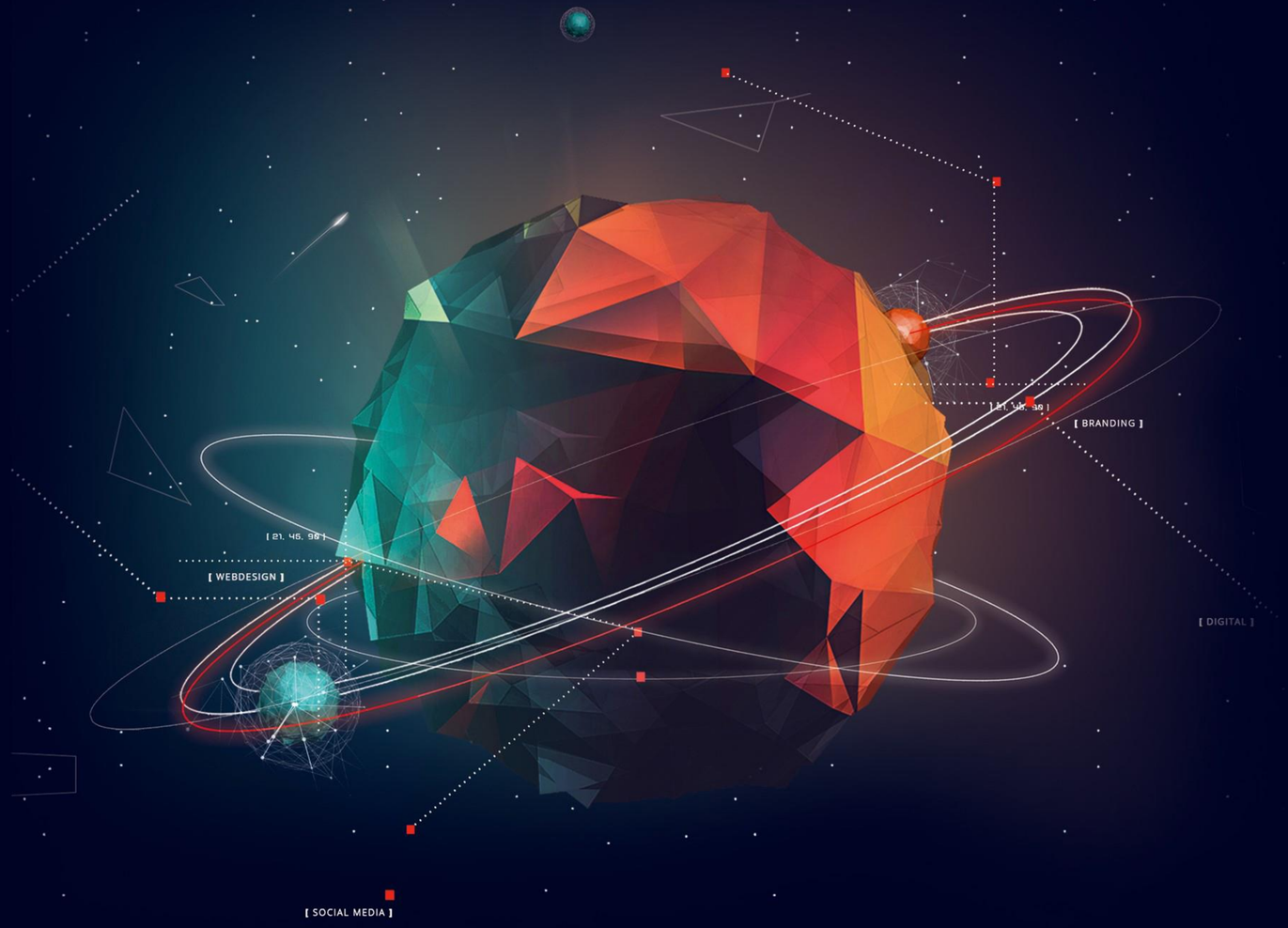
模型融合方案



尝试方案

1、LGB(PointWise)和NN直接加权融合
2、LGB(PairWise)和NN rank归一化后加权融合
3、LGB(PairWise)和NN 按组rank归一化后加权融合
4、LGB(PairWise)直接归一化后和nn加权融合
5、LGB(PairWise)sigmoid归一化后和nn加权融合

中国高校计算机大赛
CHINA COLLEGIATE COMPUTING CONTEST



参赛总结

参赛总结

优点

- 能够对数据和指标进行细致的分析，选择合适的模型
- 1亿数据下NN单模能取得Top3的成绩

不足

- 没能尝试出有效的构造pairwise数据集的NN模型，最后的NN还是采用了二分类的方案
- 没能构造出差异性较大的第二模型，导致总成绩没有很大提升

参赛总结

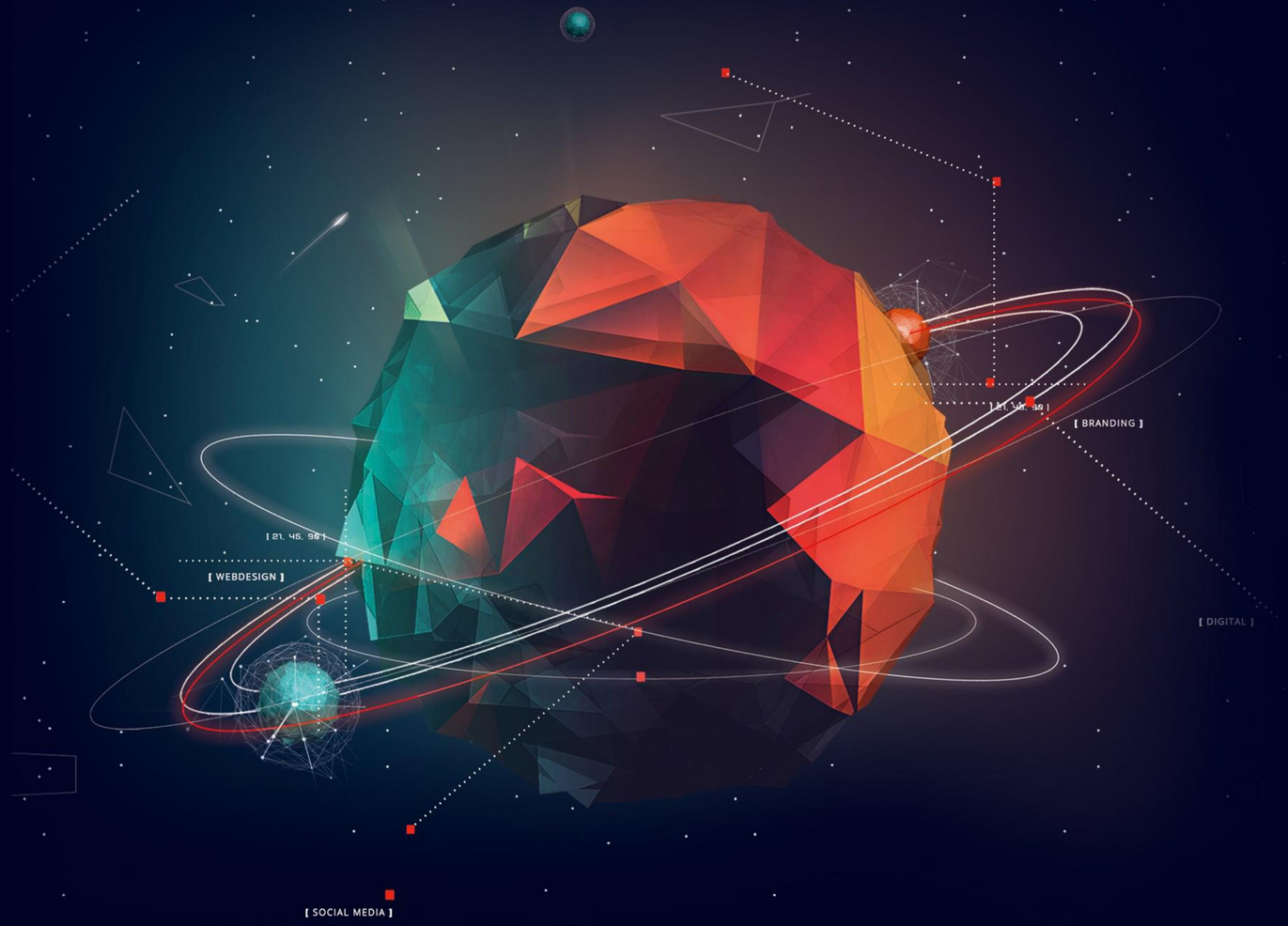
未来的改进方向

1.考虑构造特征时，引入时序因素，使得时间近的数据影响较大

2.尝试构造有效的pairwise learning的NN

尝试了采样构建: 一个正例随机采样三个负例，使用Rank Hinge Loss训练，最后因为效果不如二分类，所以没有采用；查阅资料发现pairwise对于负样本的要求比较高，所以判断是构建pairwise数据集的策略过于简单，由于时间关系最后没有深究。

中国高校计算机大赛
CHINA COLLEGIATE COMPUTING CONTEST



THANKS