
新浪微博互动预测大赛答辩
2015天池大数据竞赛
TIANCHI天池

队伍介绍

一步一步往上爬

来自中科院计算所的三位树蛙小矿工

岳智磊

分布式系统

云计算

初显奇

数据挖掘

机器学习

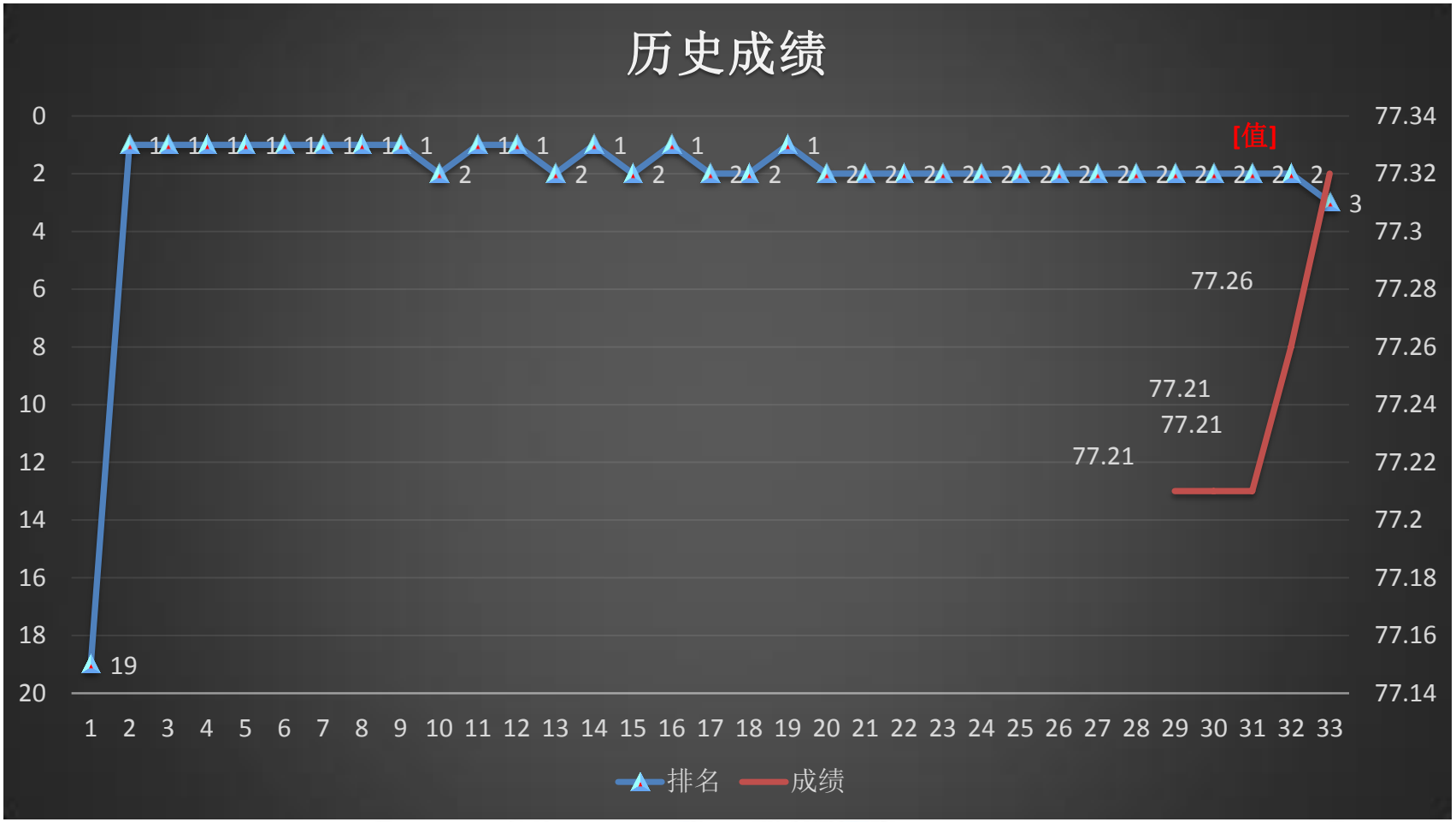
陈绍毅

机器学习

分布式系统

队伍介绍

一步一步往上爬 ???



outline

1 赛题分析

2 数据分析

3 特征工程

4 算法框架

5 可用性分析

赛题分析

抽样用户的历史原创博文数据

预测

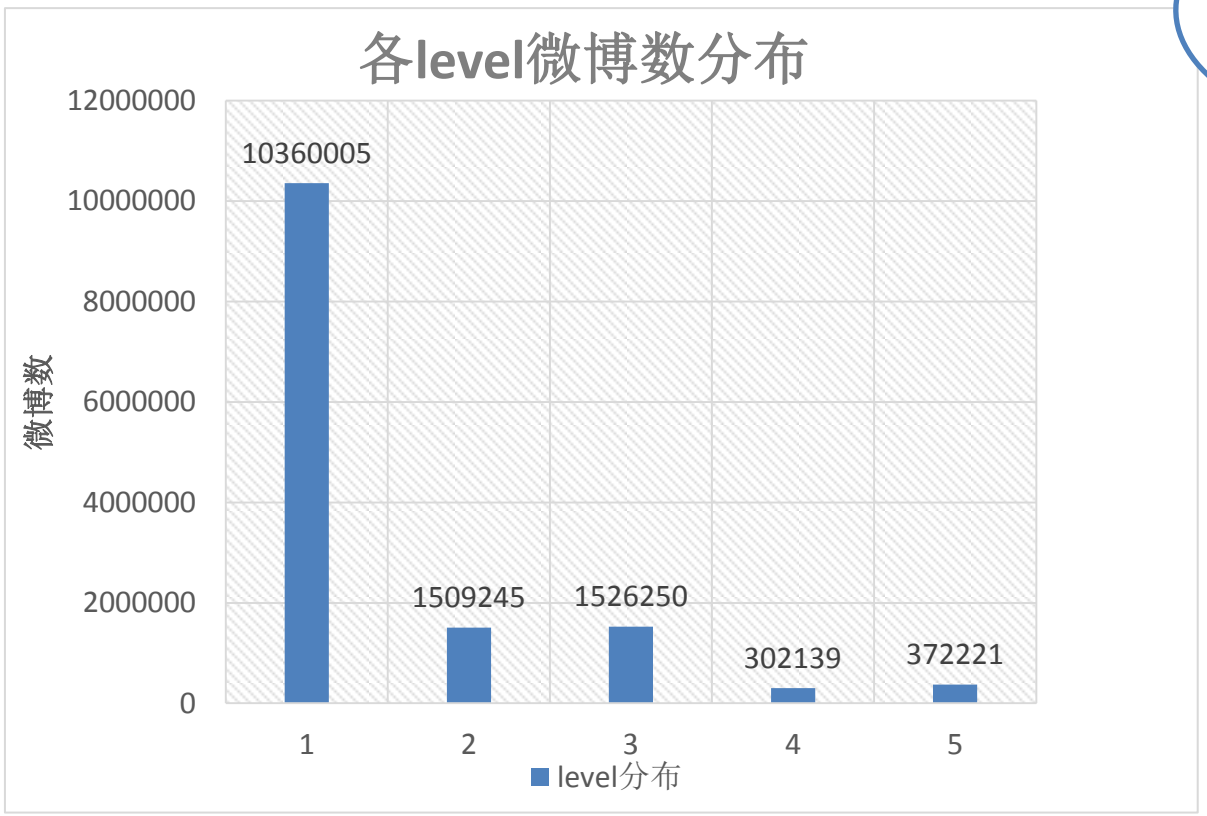
新博文发表一天后的互动情况

赛题分析

预测效果

- 用户历史发博互动情况
- 用户人际关系网络
- 微博条目本身（微博内容、发博时间）
- 准确率计算公式

数据分析



样本极度倾斜

采样 or other？

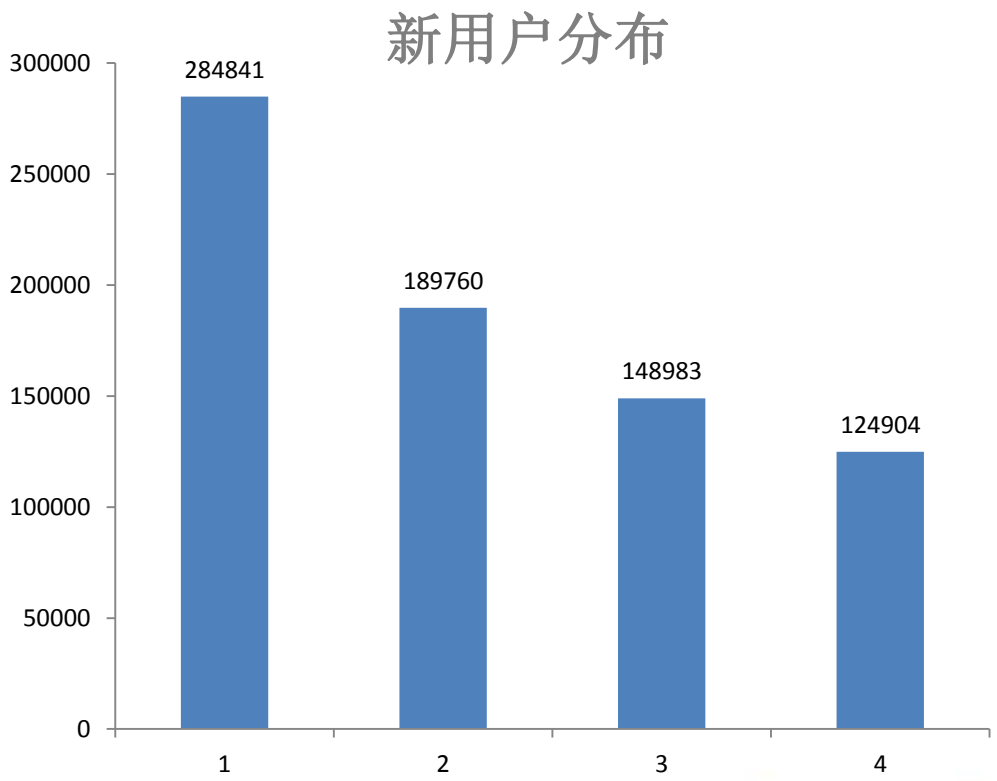
Lv1:lv2:lv3:lv4:lv5 = 34.28 : 5 : 5.05 : 1 : 1.23

数据分析

发博数	用户数
小于400	1929453
大于400且小于800	23640
大于800且小于等于1200	5340
大于1200且小于等于1600	2143
大于1600且小于等于2000	1237
大于2000且小于等于2400	689
大于2400且小于等于2800	634
大于2800且小于等于10000	2010
大于10000	536

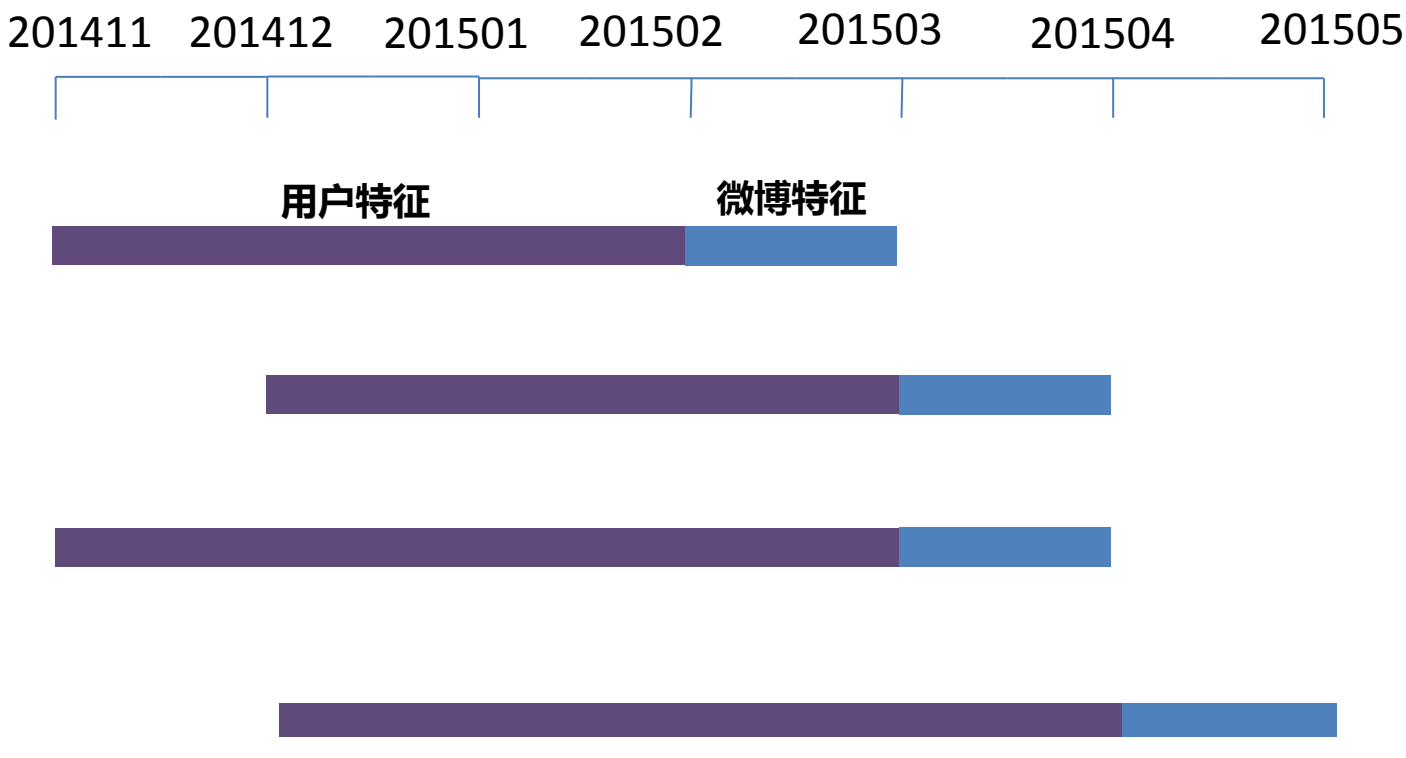
数据集构造

用户的特征 + 微博内容、时间特征 —预测→ 互动level



新用户随着用户特征统计的区间扩大而显著减少

数据集构造


201411
201412
201501
201502
201503
201504
201505
用户特征
微博特征
线下训练集
线下验证集
线上训练集
线上测试集

数据集构造

$$\text{precision}=\frac{\sum_{i=1}^{5}(weight_i \times count\_r_i)}{\sum_{i=1}^{5}(weight_i \times count_i)}$$

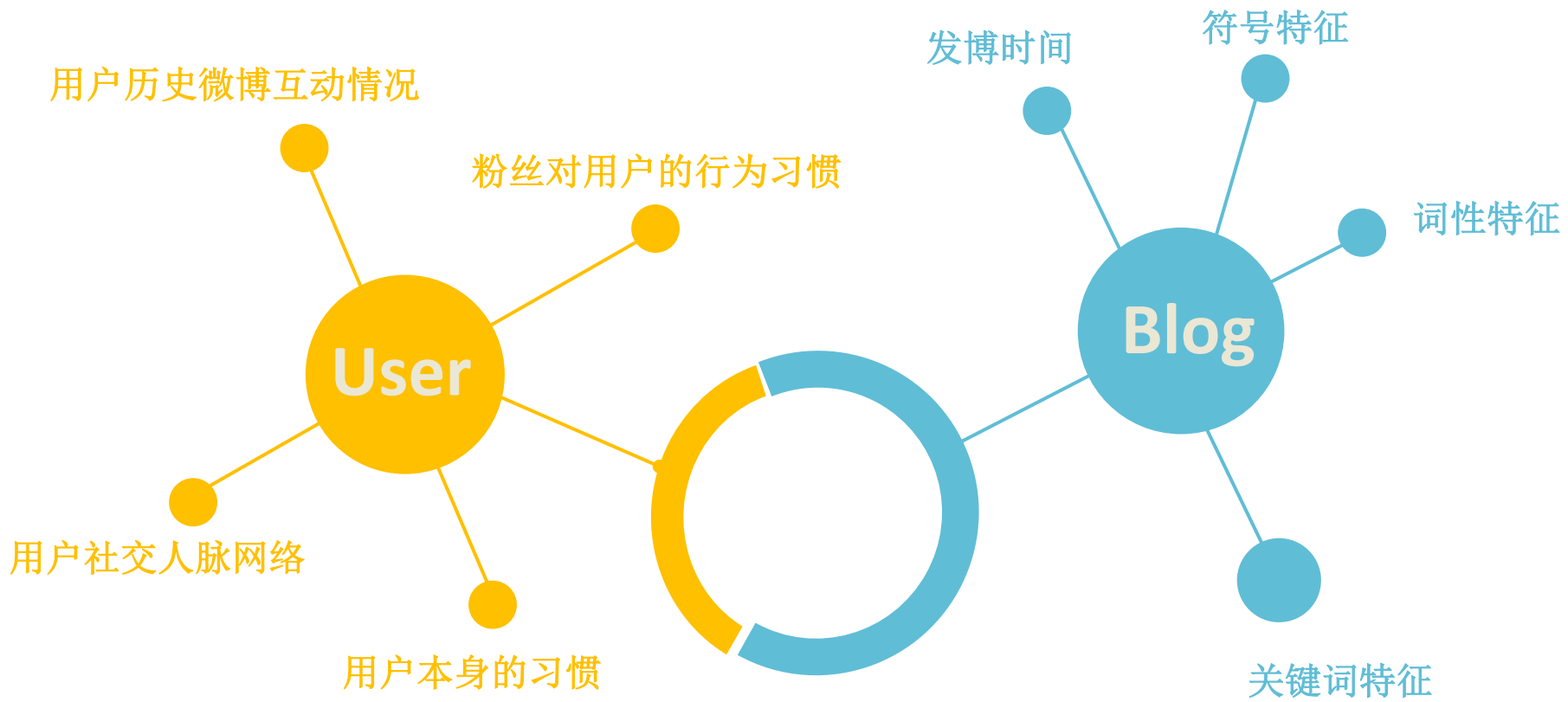
简单有效的解决权重问题，效果显著

大量数据未利用，效果极差

方案一：
数据按权重抽样

方案二：
数据按权重复制

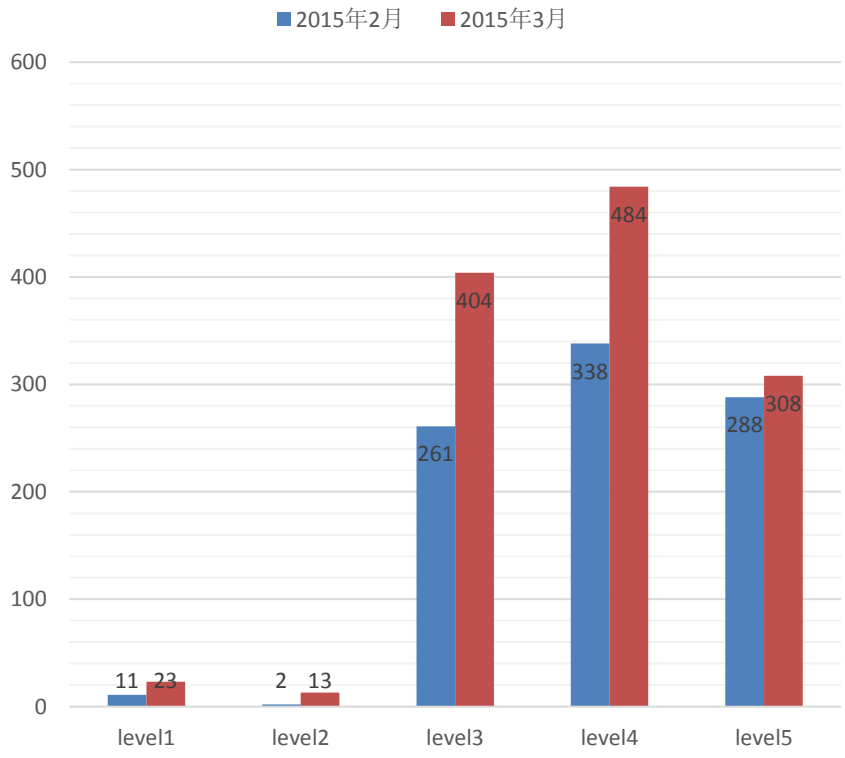
特征工程


用户历史微博互动情况
粉丝对用户的行为习惯
User
用户社交人脉网络
用户本身的习惯
发博时间
符号特征
词性特征
Blog
关键词特征

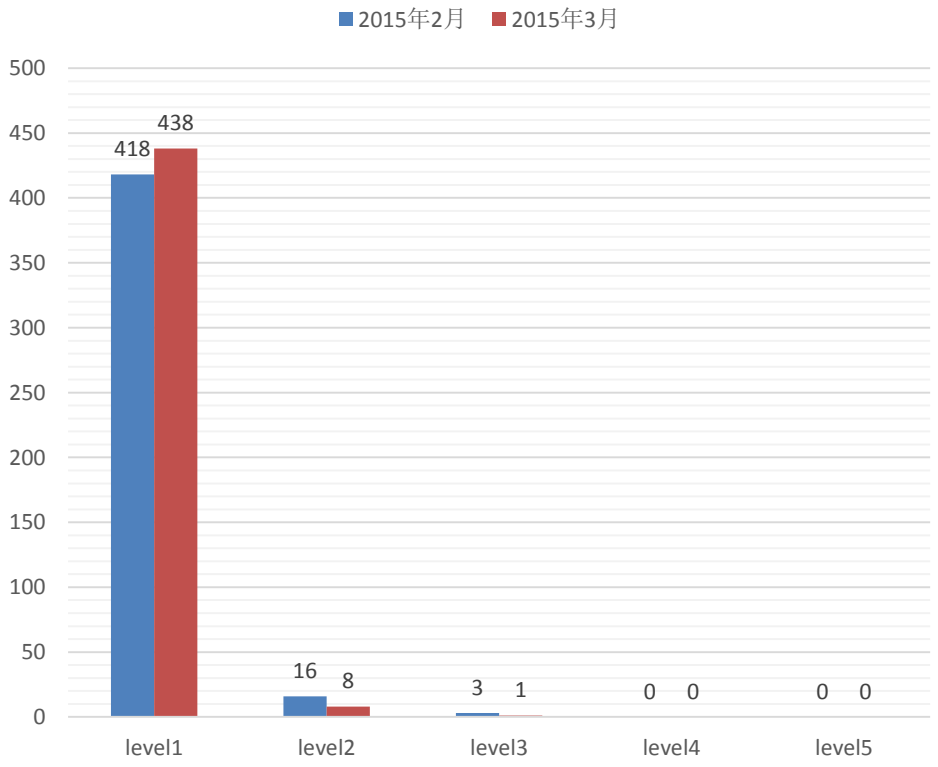
特征工程

用户历史微博特征

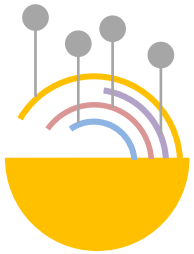
用户A微博分布



用户B微博分布



特征工程



用户历史微博特征

计数类：

5个level的微博数目

比值类：

5个level的微博数目/用户总微博数

分布类：

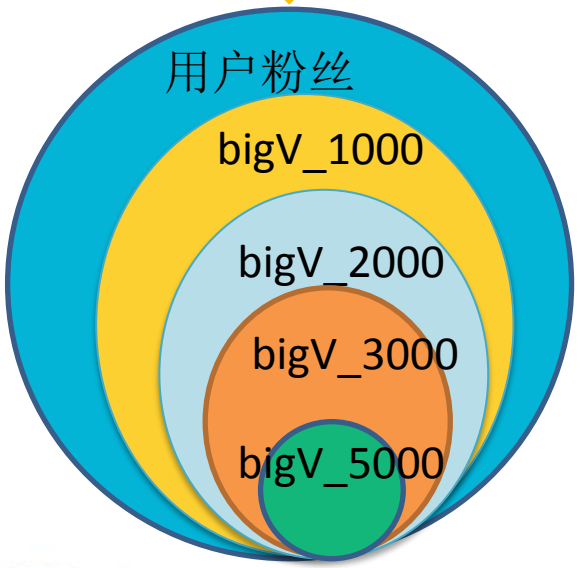
每种level微博的行为数avg、sigma

特征工程

拥有粉丝数	用户数
0-50	41017132
51-100	94770
101-200	39648
201-500	22067
501-1000	8020
1001-2000	4296
2001-3000	1595
3001-5000	1440
>5000	2946

用户社交人脉关系网

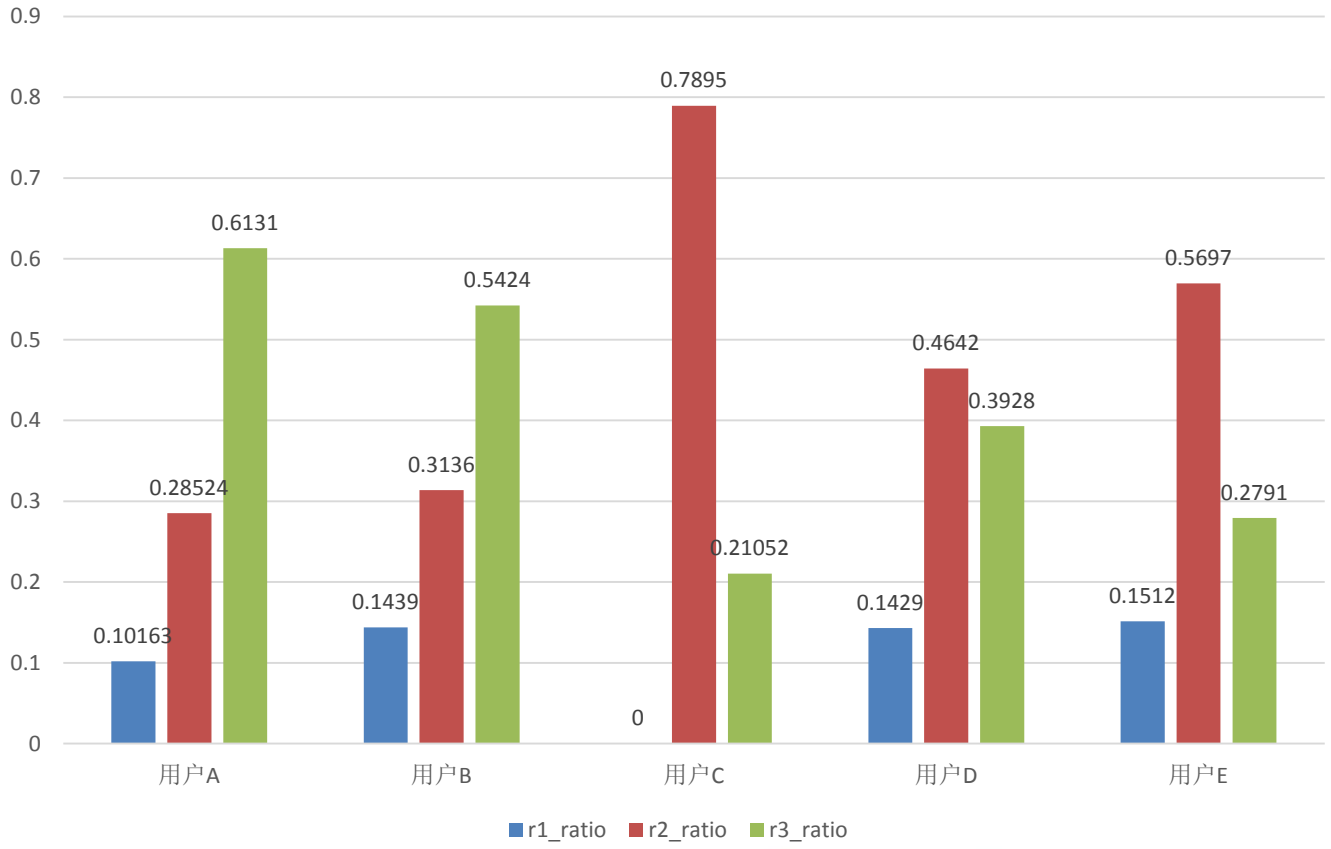
分级别大V用户	条件
bigV_1000	>1000
bigV_2000	>2000
bigV_3000	>3000
bigV_5000	>5000

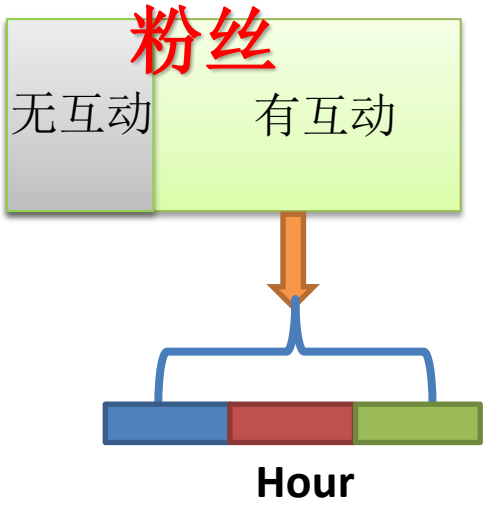


特征工程

粉丝对用户行为习惯

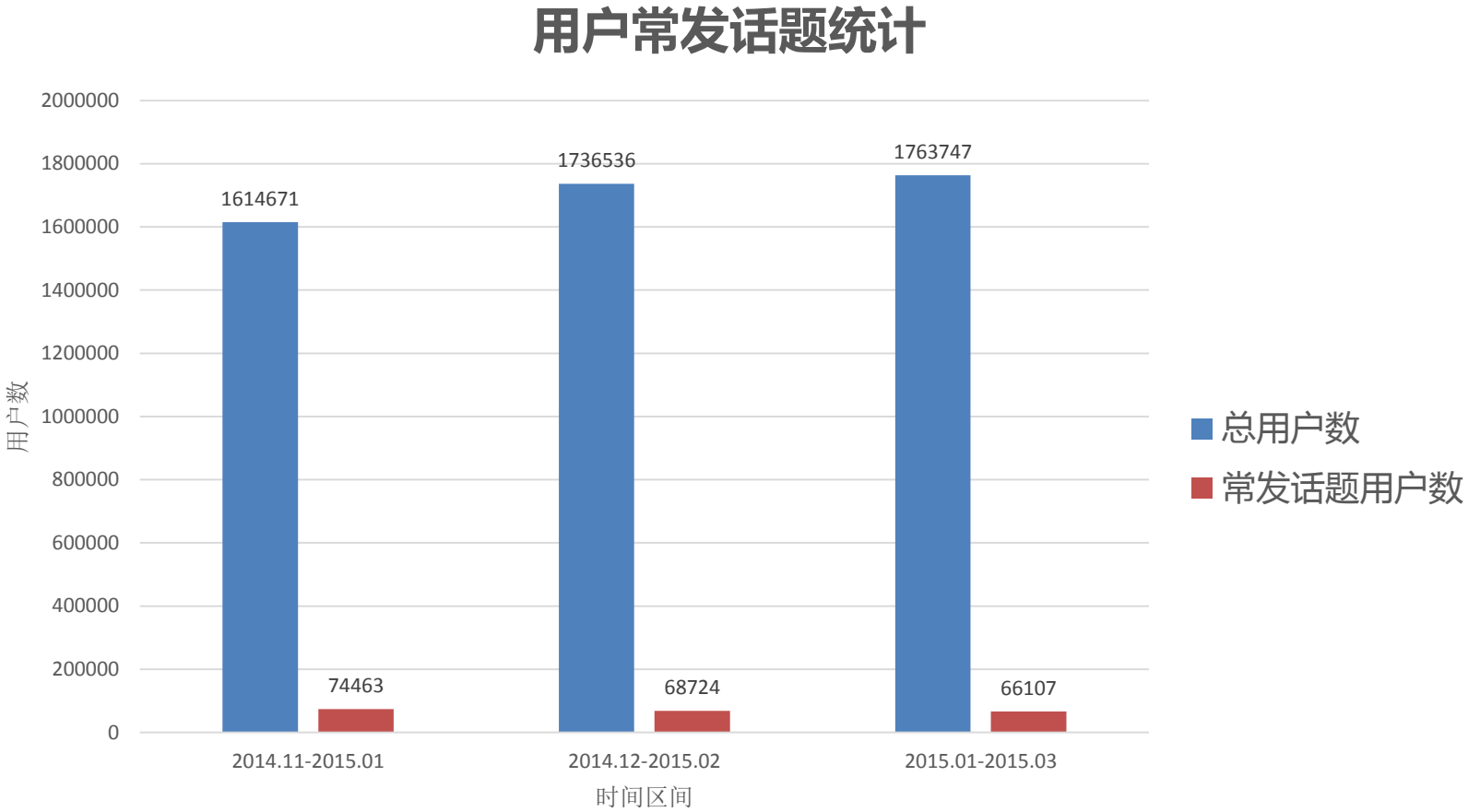
不同用户在不同时间段其粉丝行为分布





特征工程

用户本身习惯


用户常发话题统计
2000000
1800000
1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
用户数
1614671
74463
1736536
68724
1763747
66107
2014.11-2015.01
2014.12-2015.02
2015.01-2015.03
时间区间
总用户数
常发话题用户数

特征工程

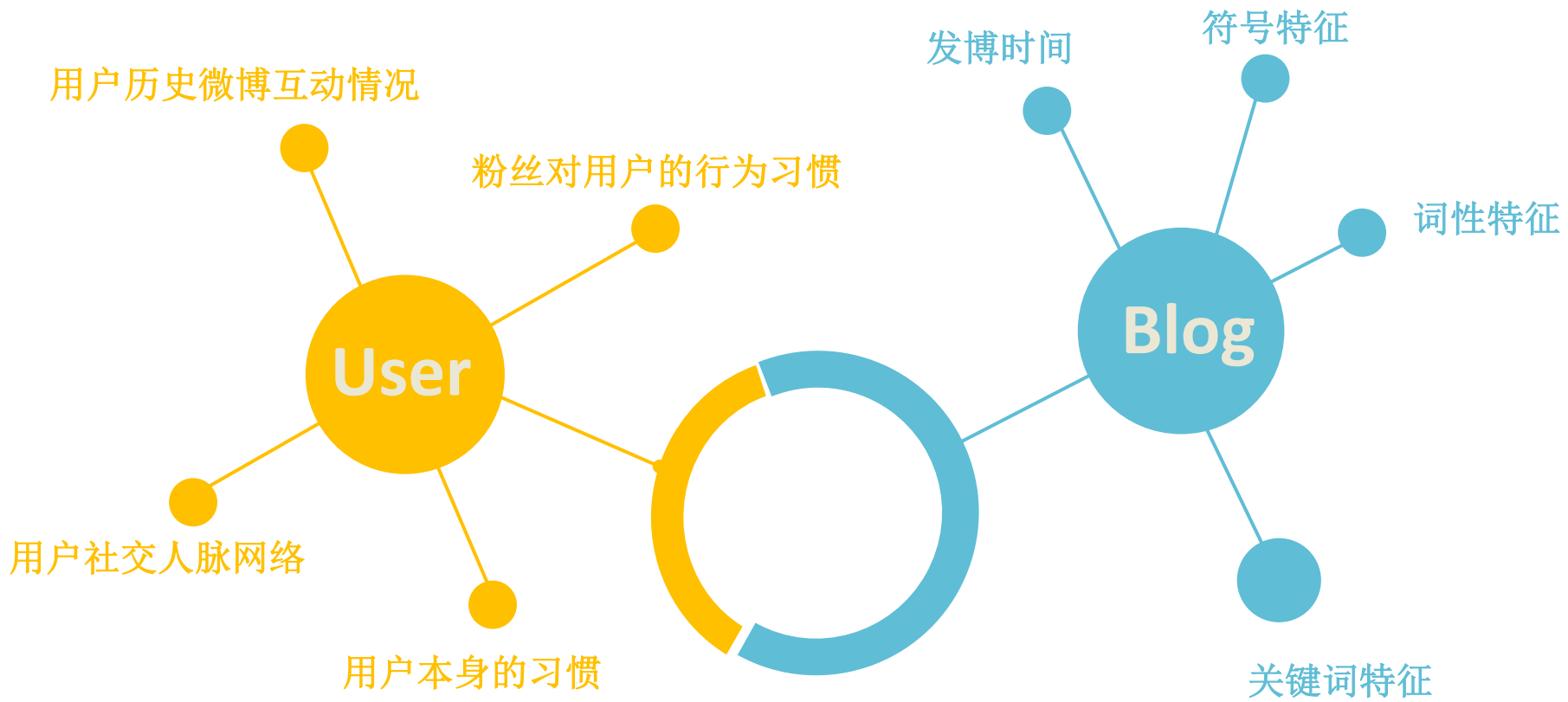
用户本身习惯

Uid: 001b4831c4f69cf139a6ece74f4a3c5c
2014.11-2015.01常发话题：#弘道·早安语录#
数目：92　level1:100%

2015.02发该话题28个，level1:100%

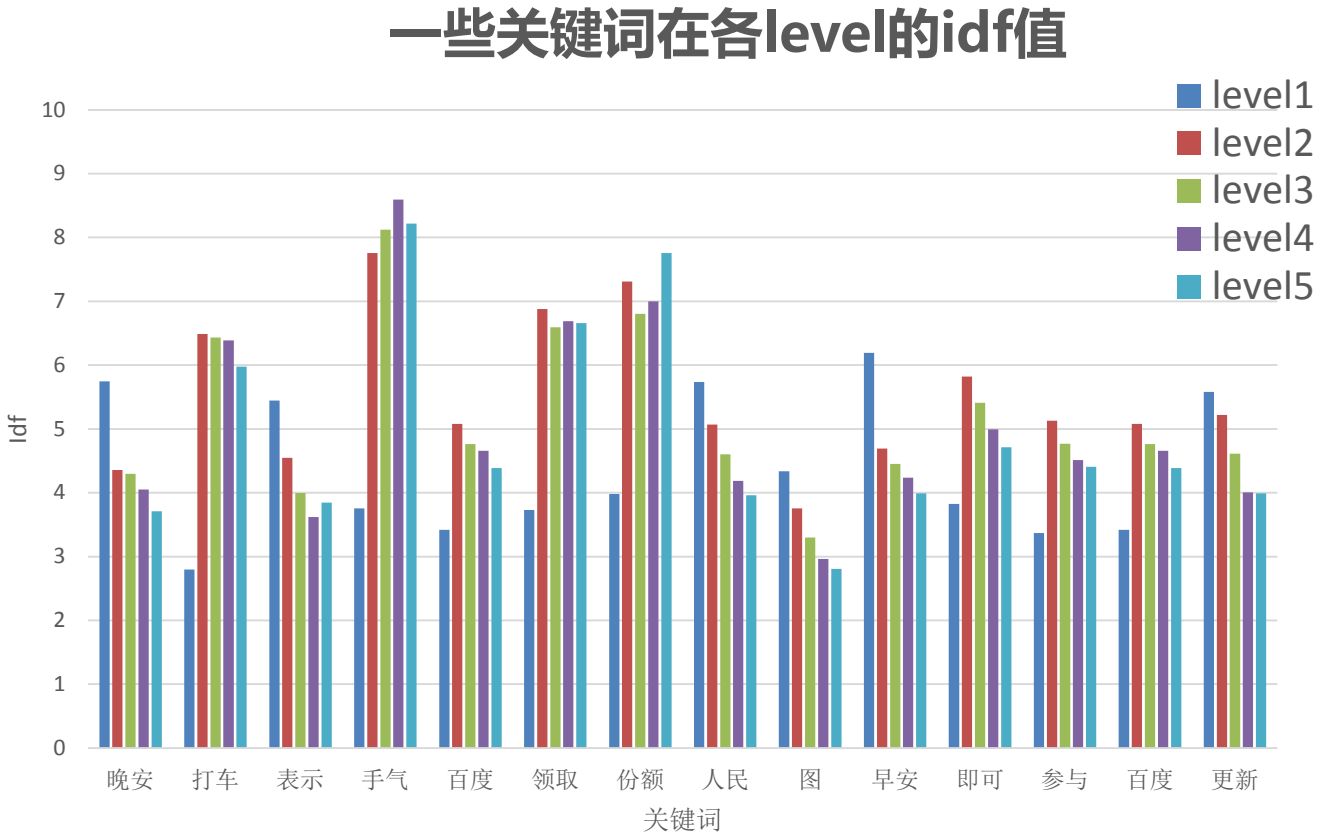
blog_time	blog	farword_num	comment_num	like_num	ac
2015-02-04 07:48:10	#弘道·早安语录# 让我们将事前的忧虑，换为事前的...	0	0	0	0
2015-02-23 07:39:10	#弘道·早安语录#舍与得，无非一种轮回，看破了，...	0	0	0	0
2015-02-05 07:22:10	#弘道·早安语录# 更新你的思想，你就能获得新生。	0	0	0	0
2015-02-12 07:34:07	#弘道·早安语录#人生的脚步常常走得太匆忙，舍取...	0	0	0	0
2015-02-10 07:30:08	#弘道·早安语录#有时，虽然理性上很着急，很想改...	0	0	0	0
2015-02-24 07:44:12	#弘道·早安语录#这个世上最不开心的，就是那些懂...	0	0	0	0
2015-02-02 07:38:11	#弘道·早安语录#成功需要成本，时间也是一种成本...	0	0	0	0
2015-02-14 07:35:15	#弘道·早安语录#心若自在，身在顺境；心若不安，...	0	0	0	0
2015-02-01 07:33:11	#弘道·早安语录#只有不断找寻机会的人才会及时把...	0	0	0	0
2015-02-17 07:35:13	#弘道·早安语录#当你心累的时候，可以换个观念看...	0	0	0	0
2015-02-15 07:40:14	#弘道·早安语录#努力做一个可爱的人，不埋怨谁，...	0	0	0	0
2015-02-08 07:35:09	#弘道·早安语录#闲看花开，静待花落，冷暖自知，...	0	0	0	0
2015-02-19 07:45:13	#弘道·早安语录#人生就像坐飞机，飞多高不重要，...	0	0	0	0

特征工程

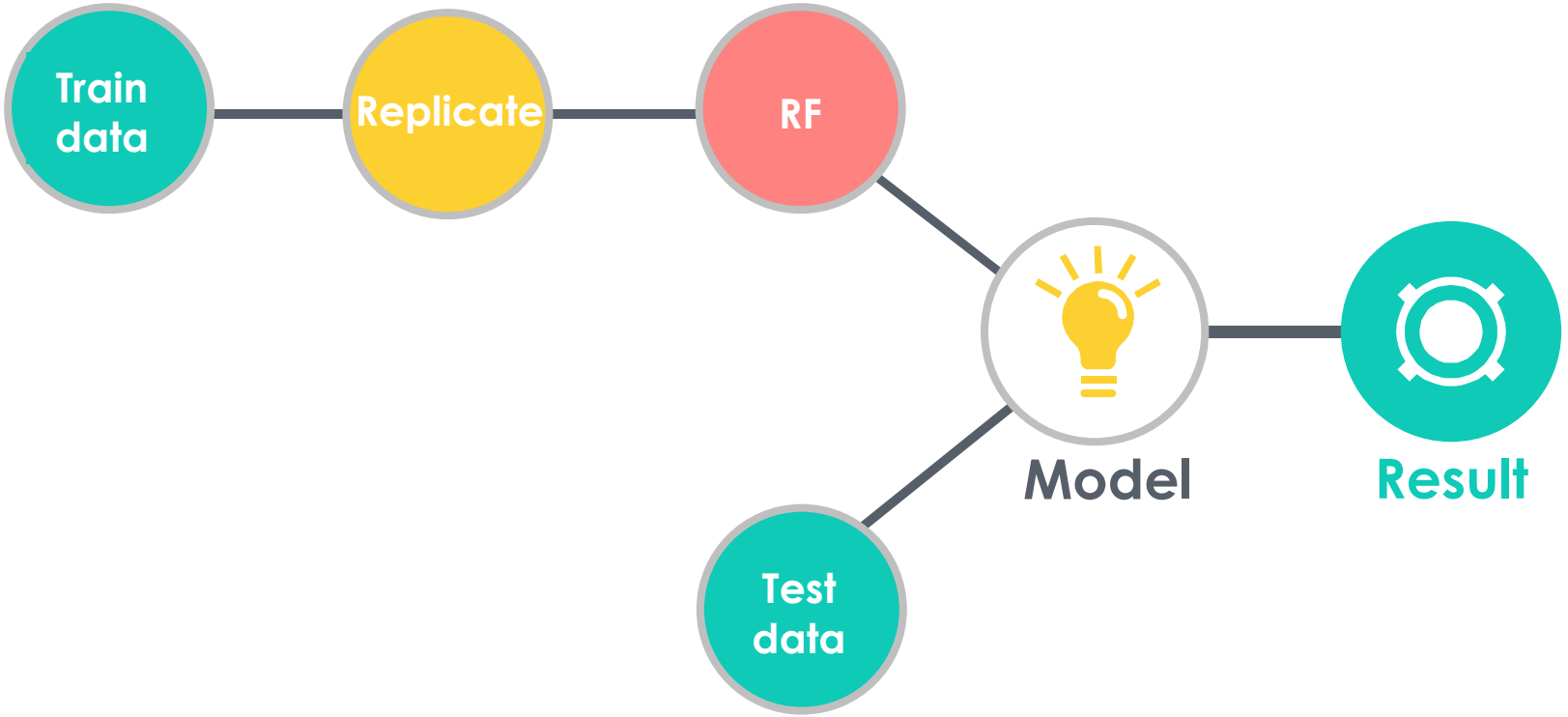

用户历史微博互动情况
粉丝对用户的行为习惯
User
用户社交人脉网络
用户本身的习惯
发博时间
符号特征
词性特征
Blog
关键词特征

特征工程

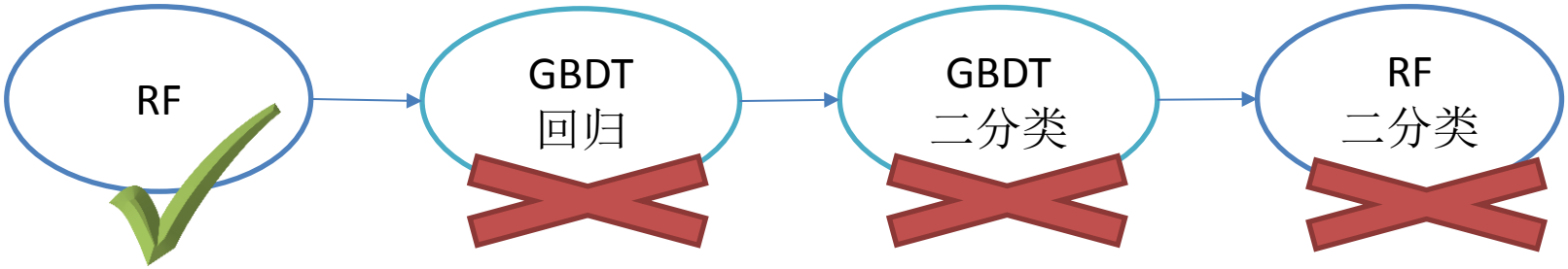
关键词特征


一些关键词在各level的idf值
level1
level2
level3
level4
level5
10
9
8
7
6
5
4
3
2
1
0
Idf
晚安
打车
表示
手气
百度
领取
份额
人民
图
早安
即可
参与
百度
更新
关键词

算法框架



模型选择

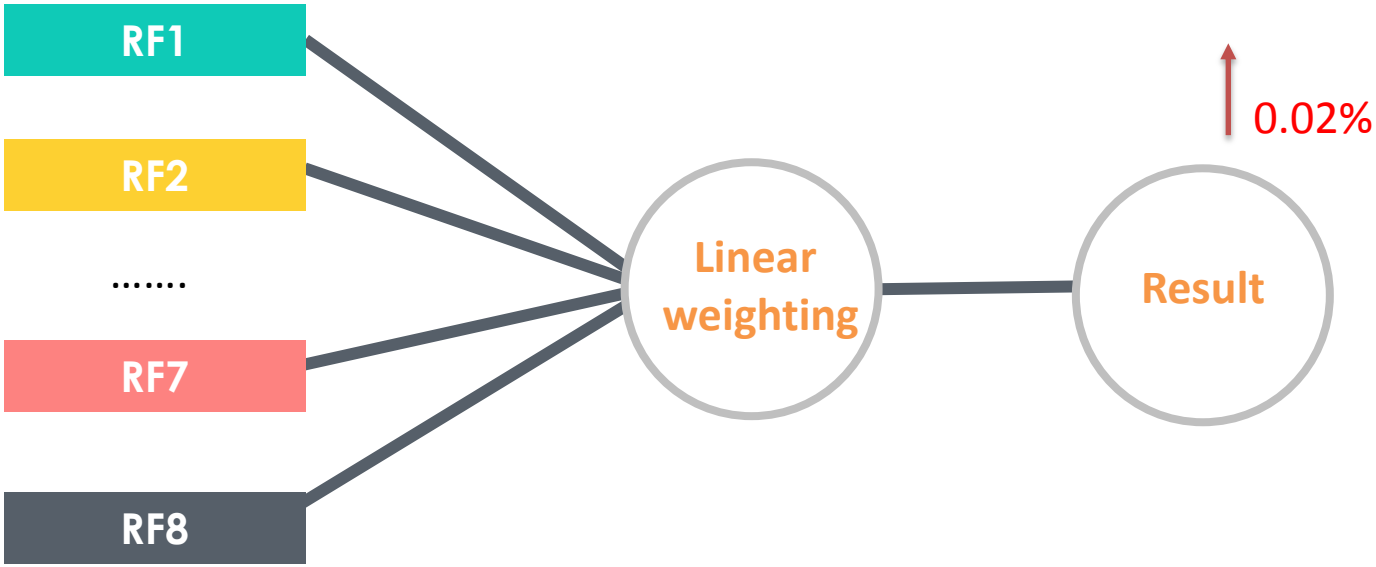


训练速度快：RF > GBDT回归 > RF二分类 > GBDT二分类

预测效果好：RF > RF二分类 > GBDT二分类 > GBDT回归

不易过拟合：训练过程中的数据抽样和特征选择

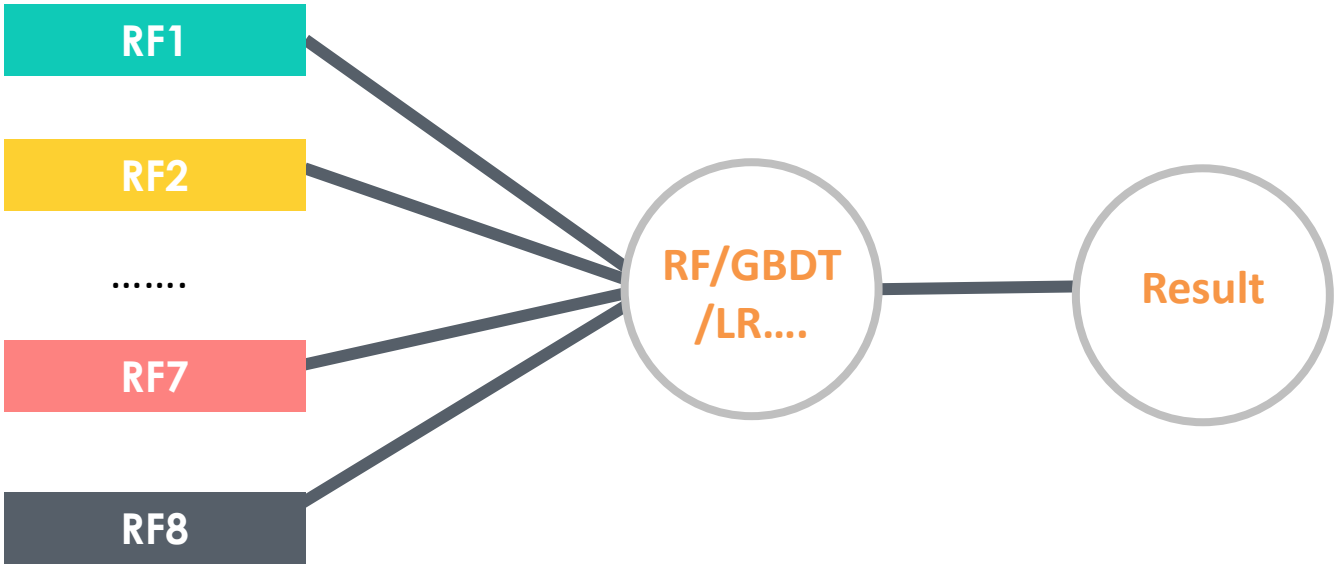
模型融合



模型融合

未完成.....

真正的融合方案



可用性分析

特征维度

- 维度少
- 统计快
- 所需资源少

单模型RF

- 部署简单
- 可并行化
- 执行效率高
- 训练速度快

致谢

感谢新浪微博和阿里举办这场比赛，让我们有机会接触到真实的数据

感谢阿里巴巴提供的数据平台，以及及时给我们解决各种问题的阿里人

感谢队友以及一路陪伴走过的数据达人

THANK YOU!

Contact us：
chuxianqi@ict.ac.cn