TUNGEE 探迹®

方楠的转正述职报告

智能拓客·产品研发部·招投标组

www.tungee.com

2023年10月11日

目录

CONTENTS

一、试用期指标及主要工作情况

试用期指标

团队目标：对比千里马反向对比达到95%

实现路径：

- 一种召回新源的方法
- 一种可以提高至少10倍开发效率的开发方案
- 若干个数据采购商机线索

个人指标：

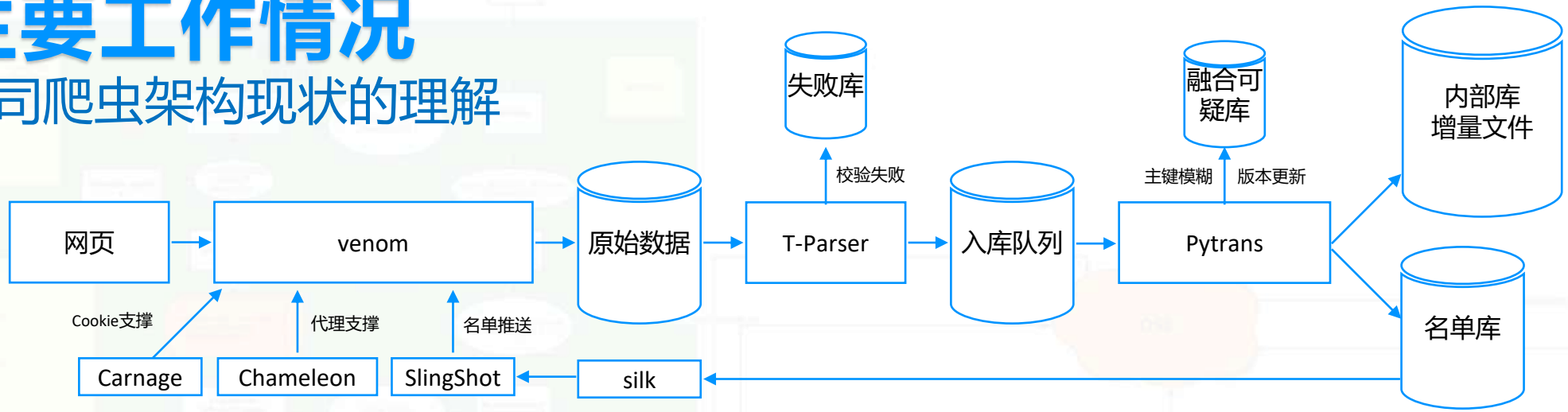
- 【**完成**】招投标采集系统开发（目标8000）：已完成模版生成9176个源（入职时为1000，全为重点源）
- 【**完成**】招投标数据采购或攻坚（目标合计4条）：已贡献4条有效的数据采购商机，未有攻坚动作
- 【**未完成**】千里马反向对比95%：团队最高周反向对比82.31%，现正向对比已超过反向对比

禅道：

- 7月：延期率80%、饱和度60.87%、代码量0、bug0（本月因账号绑定问题统计异常）
- 8月：延期率41.7%、饱和度121.3%、代码量318、bug4、需求难度2.5
- 9月：延期率13.3%、饱和度110.89%、代码量605、bug0、需求难度0、全员排名第二（2/85）
- 【延期率高】的原因主要是前期工作方向路径不确定，难以合理分解，攻坚较为困难。随着方案日趋成熟稳定，计划逐渐可控，延期率下降。

主要工作情况

公司爬虫架构现状的理解



Venom

采集系统。有Scheduler、 Fetcher、 Processor、 ResultHandler、 LogHander五个组件，【Chameleon】【Carnage】【SlingShot】【Silk】支撑组件

相比于过往公司的不同之处在于设计了一个**基于OSS的mutipart存储方式（crawl_id）**，解耦了爬虫系统多页面合并采集，降低了开发时的思维负担，提高了开发效率。另外就是直接用OSS+回调接口的方式进行下级通信，相比传统的kafka时效性较差，但是非常有利于模块解耦和数据管理，云计算成本也较低。

Tungee-Parser

解析系统。有Schduler、 Processor、 ResultWriter、 TaskChecker四个组件，另一些支撑是以接口和继承形式提供的

相比于过往公司的不同之处在于完全解耦了“下载”和“解析”。**缓存了一周的贴源数据**，容错性更强。另外有一个维护较好的TaskChecker模块，极大的降低了因开发失误导致的脏数据的录入。此外使用gevent协程框架，对于CPU密集型的解析性能更强。

Pytrans

入库系统。有Preprocess、 Import_inter、 Import_silk、 combine_inter四个核心组件，Preporcess中有主键确定和hbase版本管理两个流程。

相比于过往公司的不同之处在于对多源->多字段多表->的场景进行了更加“细致”的管理，对数据融合的过程进行了抽象。将入库流程全环节都抽象成接口或基类，增加了数据安全性，降低了犯错的概率。

主要工作情况
公司爬虫技术的实际体验

可靠性

偶尔假死、偶尔解析队列堵死，解决速度快，线上无感

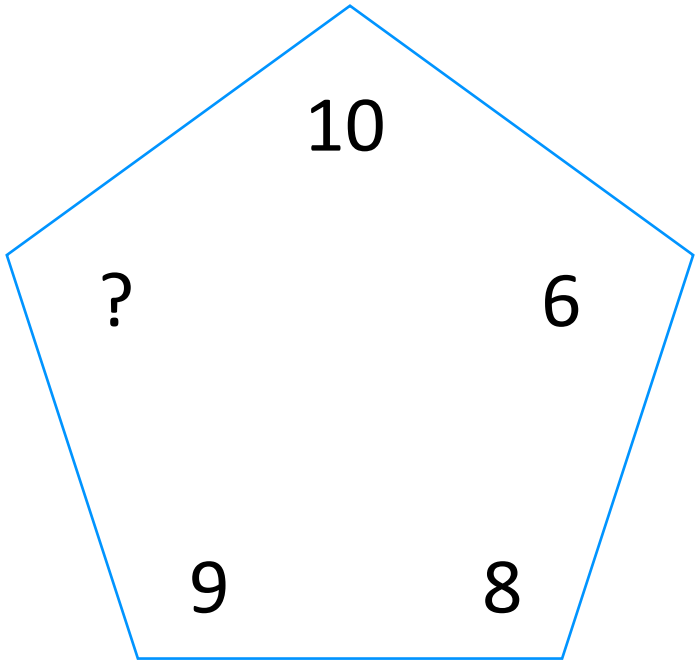
基于OSS架构一周内贴源缓存，脏数据快速覆盖

入职一个月后
我确定了
【提高公司的开发效率】
作为我的工作重点

机器效率与性能

不太清楚具体数据，从架构上来
盲猜阿里云的折扣应该不高（4折-5折？）

去重系统形同虚设，重复请求率较高，浪费带宽



开发效率

Python2框架过于老旧，研发成就感很低

表面复杂度太高，难学，培养成本高

完善的IP池服务和验证码服务

标准化的开发动作，包括爬虫和解析

查问题因为封装较深导致困难

监控与管理

完善的代码管理体系，易找易继承

安全的上线流程，但流程导致中层工作繁重辛苦

离职人员的代码维护上面表现好

公司专门为每个项目组设计数据周会，重视程度高

各条线非线上数据常用文件管理，知识较为混乱

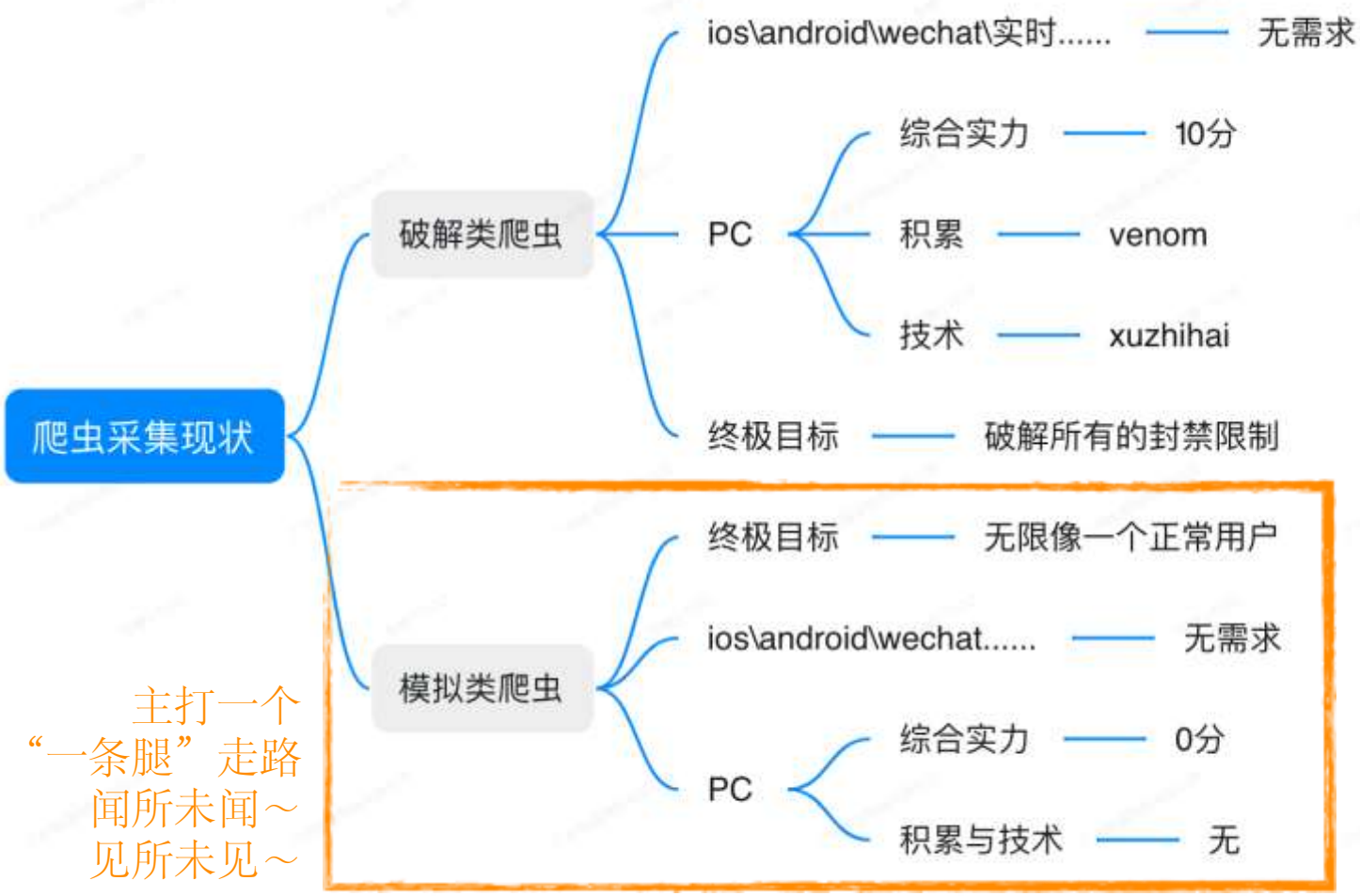
可配置性

性能可配置性较高，但优先级形同虚设

通用爬虫适应面广，开发质量很高

主要工作情况

针对公司爬虫架构现状的分析、理解与动作



主打一个
“一条腿”走路
闻所未闻～
见所未见～

	破解类爬虫	模拟类爬虫
S	性能✖10	开发效率✖10 维护✖10
W	框架老旧 学习成本高	无研发基础 无既往案例
O	稳定为王	招聘情况差 外包人力不足
T	/	“数据科学家”

最终，我的动作是做了一个【**链接**】
将【**模拟类爬虫**】开发效率高的优势
打破【**隔阂**】运行在现有的框架上
从而在【**长尾数据**】这个典型场景上
达到了几十倍的提升人效的效果

*上面这段文字，就是我三个月工作内容的总结

二、试用期成功项目展示

工作成绩

一种”低代码”爬虫—AISpider采集系统（背景）

*数据为方便表述而进行的大概估计	**核心源** 国家级政府、巨型企业牵头官方公布的一手数据公开网政策法律要求的数据汇集基地 **特点** 全、权威、网站访问卡、不收费、无法律风险	**竞品源** 大数据公司对数据进行收集、采集、加工后的有偿数据平台 **特点** 数据经过加工符合客户场景、收费、破解的开发成本高、且破解有法律风险	**长尾源** 官网和地方子站为了SEO、自身方便或法律要求，公开在自己网站上的数据 **特点** 零散、结构化的开发成本高、单网站的数据价值低	**预估总数**
常规爬虫的项目分布	90-100%	90%-99%	0-10%	1-30
招投标项目的分布	70%	80%	30%	2-5w
企服政策分布	50%	60%	50%	11w

结论

核心源是重中之重，核心源没做好之前全力做核心源数据采集

当核心源已经饱和后，采集竞品源可以补充长尾源数据

但是自采长尾源能力才是超越竞品的最终途径

工作成绩

一种”低代码”爬虫—AISpider采集系统（价值）

RPA

AISpider的三个能力

找到新源的能力

自动采集的能力

检查错误的能力

拓展数据源:18098+10098

拓展采集源:10669+9669

对比外包方案(加修复约100源/周/人):+747人天

对比外包方案(现有效率约200/周):10倍(2000/周)

*预估新增数据量:124613/周+17801/天

数据判别器+以前没有

源判别器+以前没有

*预估新增数据量：现只正式灰度上线了1095个源，其他的正在走上线流程。数据是根据1095个源的周采集结果推算得出

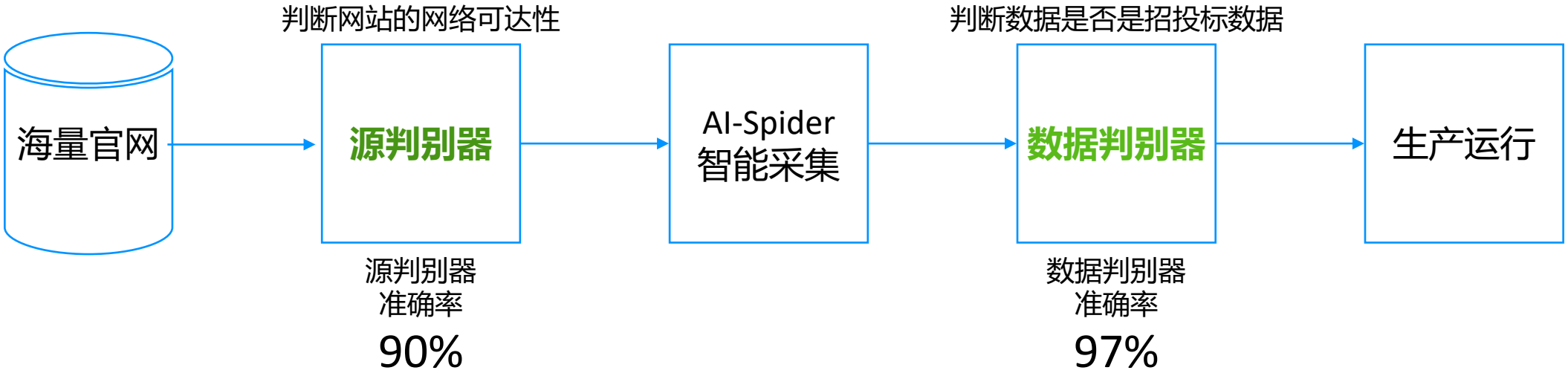
AISpider，就是一个针对 **长尾源** 数据采集的完整解决方案

接下来的PPT，会分别阐述这三个能力

工作成绩

一种”低代码”爬虫—AISpider采集系统（找到新源的能力）

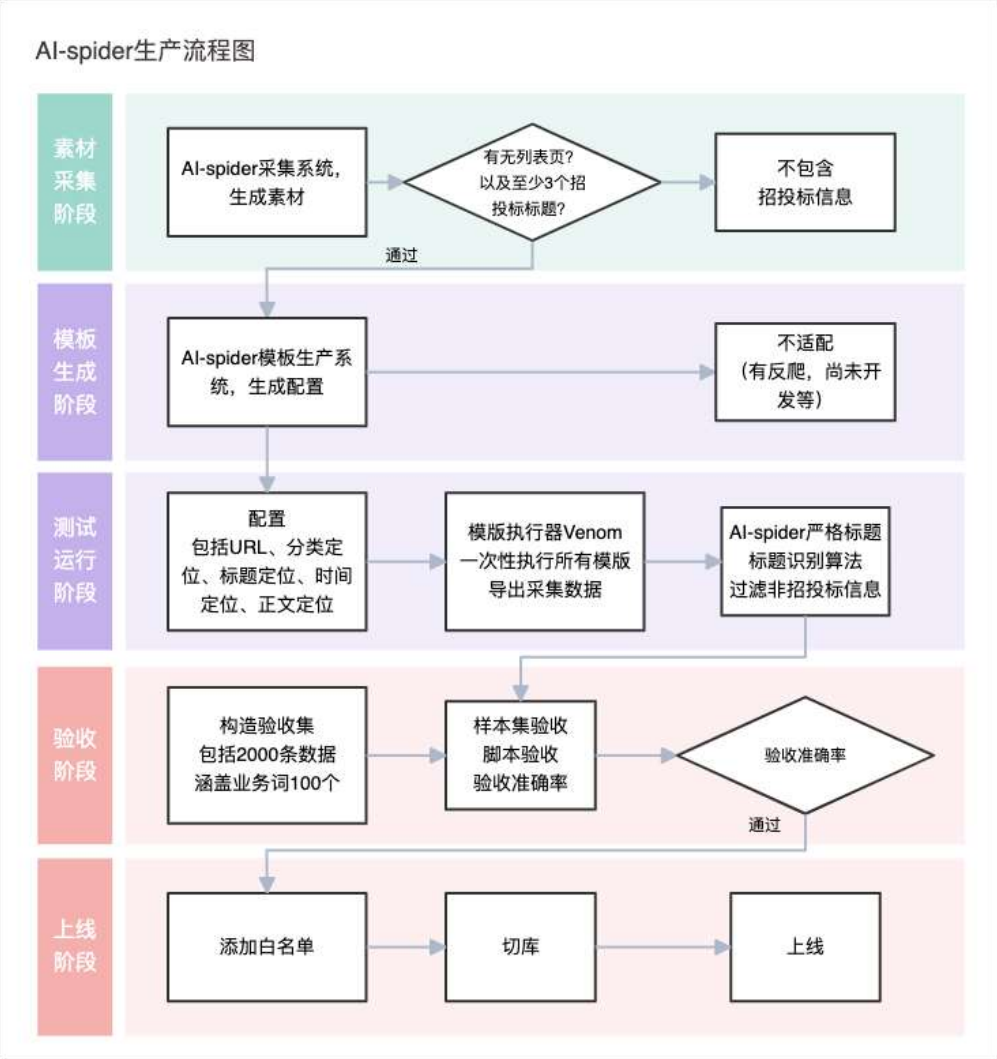
公司已有海量的公司官网数据，但是需要从这些官网中找到哪些含有招投标信息



已在招投标场景拓展数据源至18098(+10098)
只需要放入网站首页URL，就可以完成数据爬虫生产全流程！

工作成绩

一种”低代码”爬虫—AISpider采集系统（自动采集的能力）



AI-Spider
智能采集

源名单共计: 28238 漏斗比例:83.81%
可访问共计: 23483 漏斗比例:83.16%
有三个及以上招投标标题渠道共计: 18098 漏斗比例:77.07%
成功渠道总数: 11205

适配率(召回率):*61.91% (静态:10945 动态:260)

准确率: *95%

*适配率测试时仅代表源召回率，暂无列表页召回率
*准确率为爬虫采集返回除正文外的准确率

工作成绩

一种”低代码”爬虫—AISpider采集系统（检查错误的能力）

检查错误的能力属于AISpider采集能力的伴生能力
因为AISpider的性能较差但几乎没有不能破解的网站
特别适合爬虫上线检查用例、爬虫稳定性测试用例的生成

原场景举例
需要运营从10个列表页的首页、中间页、尾页抽取30个标题case，从而检查爬虫采集完整性。
复制10*30=300个标题，需要工时30分钟-60分钟

抽取->60分钟
编写excel->10分钟
字段完整性->10分钟

节约时间81%

“
RPA模拟浏览器
在本机操作你的鼠标键盘
当你面运行
很难运行出错
”

新场景举例
通过AI-Spider快速生成case
5分钟设置，运行1小时（人不需要在）即可按要求生成检查case，并自动结合venom数据分析
不再需要编写excel函数做比对

编写AI-Spider脚本5分钟
字段完整性->10分钟

另支持全检、其他形态的抽检、开发前产品预检等多种检查方式

工作成绩

一种”低代码”爬虫—AISpider采集系统（难点和核心技术）

难点

1、在一定的性能效率要求下，与网站反采集策略无关的【一种通用的反爬破解手段】

2、元素定位算法，将人认知网站结构的“知识”模拟出来，从而自动反写生成配置模版

核心技术

1、undetected_driver模拟浏览器技术，基于内存核心级别的特征隐藏，使用到现在尚未遇到**无法自然破解的反爬(行为数据除外)。**

2、标题、正文、列表页位置位置Xpath反写算法，这是AISpider能代替人工的关键。利用正文、渲染位置、渲染大小等【**人类认知特征**】，代替【**程序特征**】，使开发人员从繁重的爬虫开发工作上解脱出来。

3、积累丰富的【黑白名单】和【网站结构】验证集数据，这种验证集使得我可以快速找到召回率最高的核心特征。并且能自由的根据需求，低成本**灵活的调节准确、召回率**的比例，满足产品需求快速应用。

三、试用期间工作的不足及改善措施

存在不足与改进方向

我个人认为自己的 **优点** 和 **不足** 是辩证的，具体表现出 **优势** 还是 **缺陷** 主要看 **工作需要**

优点	不足	表现	改进方案
具有创新性思维，快速适应环境，能够提出创新的解决方案	会经常改变自己的计划和决策 导致行动不稳定，难以持续追求一个目标	目标和执行方案 反复变化	更好的向上管理 及时同步进度和想法
能够迅速调整行动和计划以适应变化 不会因为困难而感到沮丧或无助	行动和决策常常是根据当前情况而变化的 因此很难预测下一步行动	禅道延期率高 给管理者带来很大的 不安全感	对禅道任务细化 风险任务先分解流程 “调研”+“执行”+“验收” 避免风险任务 导致延期
善于与他人合作，能够灵活地适应不同的团队和合作方式	缺乏稳定性 在一个没有挑战性和重要性的岗位上会陷入自我焦虑	招投标的方案正在落地中，又参与到企服的工作	参与新工作时 确保旧工作的完整落地
能够接受不同的观点和企业文化，能够快速合群	会因为对新鲜事物的追求而无法持续专注 对无法提高效率的重复性工作和制度有厌恶感，并难以管理	细致工作BUG率较高， 不认同的工作抗拒执行	以团队为重 积极表达且坚决执行

四、未来工作规划及展望

工作目标

将AISpider对长尾源的处理能力赋能到各条线去（展望，量化计划见下页）

招投标：数据痛点：源少、数据少

已经完成落地，正在解决大量源涌入的善后问题（正在进行）

1、完善大量数据入库后的验收、清洗问题。确保增量数据真实赋能产品

企服：数据痛点：人工数据加工瓶颈、政策类长尾源数据少

正在赋能中，要先解决数据加工的瓶颈，才有落地的先决条件（正在进行）

1、自动化企服至少90%的“人工”数据加工流程，不降低“人工”参与度，AISpider就难以实施落地。

2、增加政策类数据的类型，增加政策类数据的总数

整体效率赋能：痛点：暂不清楚

尚未介入，正在探索落地场景（未进行）

1、降低爬虫工程师的工作门槛

2、提高爬出工程师的工作效率

工作目标

将AISpider对长尾源的处理能力赋能到各条线去（量化计划）

招投标：

1、22000有效源采集（现状10000个源）**2023Q4完成**

2、医院学校类：反向对比95%，正向对比50%（现状未单独统计过医院学校，整体反向对比82%，正向对比76%）**2023Q4完成**

企服：

1、自动化人工（外包、实习生）工作至少90%的工作量 **2023Q4完成**

2、政策分析或采集达到60000政府机构源（现状8000）**2023Q4完成**

3、政策分类增加至1000个（现状34）**2023Q4完成**

整体目标：

1、在增加源的基础上，召回率达到70%（现状60%） **2023Q4完成**

谢 谢
（后面的页是支撑材料）

附件：模版配置的历史情况（9月20日）

TUNGEE 探迹®

[10月10日]AI-Spider数据情况

共计11205源

周	当前	上周	上上周	历史3	历史4	历史5	历史6	历史7
采集渠道	**30418**	28238	24764	11890	8101	5304	3445	2265
可访问	**25318**	23483	20517	9357				
存在数据	**19146**	18098	16046	7634	4653	2433	1616	838
交付源	**11205**	10669	9176	2901	1731	1154	643	147

整体目标	**22000源采集** **医院学校，遥遥领先！** **增量反查95%，正查50%**
下次里程碑	标准化上线流程推进 上线9289源
上线方案	**测试集构建** **正文判别器** **根据正文判别器，修复badcase**
待办	标准化上线流程推进 上线9289源

附件：源增加的历史情况（9月20日）

TUNGEE 探迹®

[10月11日]AI-Spider-源拓展

新增潜在源预估4155个

周	第四周	第三周	第二周	第一周
思路	医院精细化	院校精细化	地域召回	高注册资金+国企
新增潜在源	**31163**	3212	4898	12640
累计	**64856**	33693	30481	25883
抽验有效性	**4/30**	4/10	10/10	7/20

序号	状态	策略集
1	完成	院校精细化 高筛-高注册资金-院校->官网 高筛-教育部高校名单->官网 高筛-教育部大专名单->官网
2	完成	医院精细化 高筛-高注册资金-医院->官网 高筛-评级医院名单->官网
3	未开始	招标机构精细化 高筛-招标代理机构->官网
4	未开始	重大民生精细化 高筛-国务院控股企业->官网
5	未开始	上市企业召回 高筛-上市企业名单->官网
6	未开始	外部势力制裁名单 高筛-美国制裁企业名单->官
7	未开始	房地产相关 高筛-拟在建承建名单->官网 高筛-地图-小区承建公司->官
8	完成	地域召回
9	未开始	政府源

源拓展【周期性工作】
季度总数目标22000个源
已完成识别16046/22000

源有效率识别工具
正样本准确率89%
负样本准确率95%

TODO
下周继续优化，目标双95%
分类识别工具

附件：政策源的潜在源列表（9月25日）

工作成绩

主要竞对的销售线索或情报—千里马、比地、剑鱼、天眼查

主要情报:

1、2022年标段总计2227239，累计金额104332.85亿。

2、采购价格：比地全库30w元、剑鱼全库20w元、千里马500w元、海外邓白氏编码千万元。

3、比地详细的数据库数据分布情报（逐年、逐分类详细数据），全库共计2.8亿条。

4、千里马以围标为KPI的商业运作模式，以及几家主要竞对的围标规模现状。

5、千里马、采招网独有的非公开数据的运作模式，解释了我们公开采集数据无法溯源的原因。

很可惜，最终对接后**未能采购**其中任何一家的数据（买得起的看不上，看得上的买不起）
这些情报对团队目标、方向、实施路径的制定存在间接价值

支撑材料：外包方案的生产流程图 和 ai-spider的生产流程图对比