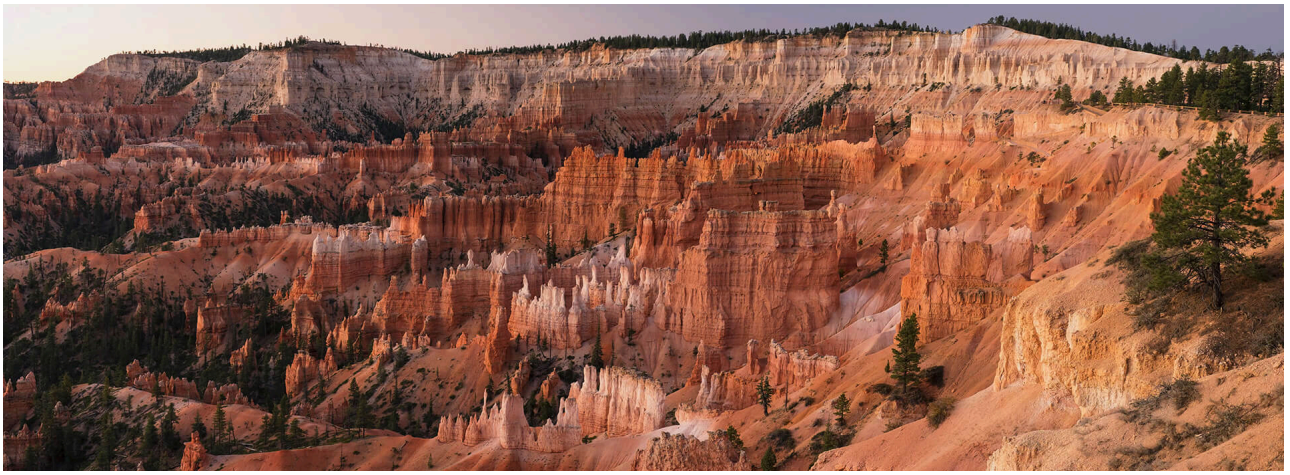


AISpider 盘源工具

写在前面

盘点市场数据量级、找采集源.....对于产品来言，真是一个一言难尽，很苦逼的事情

但是现在不一样了，以后 AISpider 帮你找源，找源只需要提个需求就可以了

数据

20240611 更新

种类	源总数量
9	277,621

- 采购盘它.xlsx 4.16 MB
- 群友盘它.xlsx 750.01 KB
- 法院盘它.xlsx 3.02 MB
- 人民政府盘它.xlsx 576.84 KB
- 招聘盘它.xlsx 454.99 KB
- 展会盘它.xlsx 1.62 MB
- 门户网站盘它.xlsx 1.14 MB
- 药监局盘它.xlsx 21.4 KB
- 招投标盘它.xlsx 279.77 KB

数据源不全怎么办?

盘源是有成本的，不是免费的。因此每次只会盘到满足需求就会停止（比如需求只要市级，区级就不会盘）

但不代表盘源器就不能继续盘了，并不代表收敛了，只是一个成本问题。

盘源原理

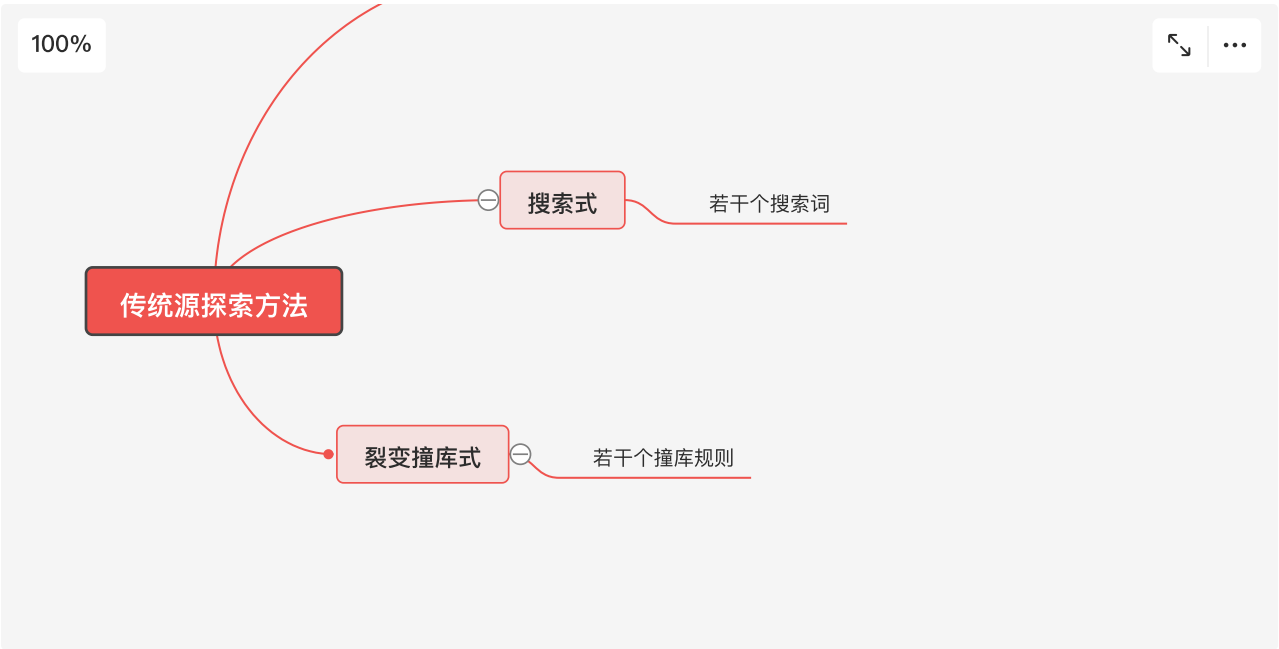
概述

暂时没空写，每个类型都不太一样。

总之已经开发好的子模块有**百度爬虫**、**站长工具爬虫**、**ICP 备案爬虫**、**SEO 爬虫（正反链）**、**Conac 撞库爬虫**、**探迹数据爬虫**、**竞品爬虫**、**一些开源、采购的名单**等

之所以 AISpider 盘源器效果这么好，主要是因为每个召回方法都有其局限性，但是一旦联立使用，效果可能达到 **1+1>10** 的水平，以下图举例。

传统盘源方法：问题→各自独立



AISpider 盘源方法：关键词→交叉与收敛

轮次	名单爬虫	搜索式爬虫	撞库爬虫
1	若干种子源		
1		若干种子搜索词	
1			过滤、去重规则
1–end	使用去重过滤规则，生成结果源名单		

2		将名单注入重新搜索，找到新源 + 新数据，运行到成本/新源数达到收敛条件	
2	新源，召回新数据，运行到新源/成本达到收敛条件		
2–end	使用去重过滤规则，生成结果名单		
3–pre	是否满足产品需求，是则输出 不满足产品需求→调整收敛函数和去重过滤规则，用新的名单接着裂变去跑		
3		名单重新搜索，找到新源 + 新数据，运行到成本/新源数达到收敛条件	
3	新源，召回新数据，运行到新源/成本达到收敛条件		
...	重复以上流程，第一轮大约是 1 工时，第二轮大约是 8 工时，第三轮大约是一周，现在还没有跑到第四轮还不客观收敛的需求		
end	到达交付标准 或重调收敛函数和去重过滤规则 或达到客观收敛（下一轮搜索爬虫刚开始运行就达到收敛条件，无法继续运行）		

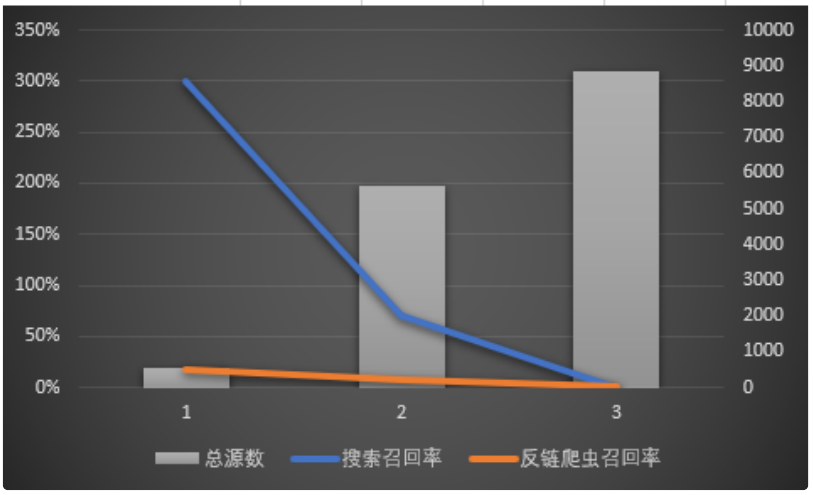
举个例子

我们需要盘点门户网站源，用来采集联系方式，提高我们拓客产品的联系方式数据竞争力。我们先给出了 10 个搜索词，比如黄页。给定了 10 个范例源。

1. 10 个搜索词百度搜索出 500 个源，去重过滤后产生 30 个新源。**（总源数 40）**
2. 10 个范例源 +30 个新源利用反链爬虫召回了 3000 个新源，去重过滤后产生 500 个新源。这轮的收敛分数是 16.67%**（总源数 540）**
3. 人工标注这 540 个新源各 10 个正负样本，我们根据这 20 个标注，调整了收敛函数和去重过滤规则。**（总源数 530–）**
4. 530 个源的标题分词产生 300 个新搜索词，百度搜索后，去重过滤后产生 210 个新源。**（总源数 640）**
5. 740 个源利用反链爬虫召回了 6w 个新源，去重过滤后产生 5000 个新源，这轮的收敛分数是 5k/6w=8.3%。**（总源数 5640）**
6. 人工标注新源各 10 各正负样本，我们根据这 20 个标注，调整了收敛函数和去重过滤规则。**（总源数 5630–）**
7. 5630 个源的标题未产生新的搜索词，搜索词召回方法收敛**（总源数 5630）**
8. 5630 个源正反链爬虫召回了 604300 个源，去重过滤后产生 3200 个新源，这轮的收敛分数是 0.529%，我们认为这个搜索代价太大了，认为其收敛**（总源数 8830）**

9. 我们完成的源探索工作，收敛曲线如下。

轮次	1	2	3
搜索召回率	300%	70.00%	0.00%
反链爬虫召回率	16.67%	8.30%	0.53%
总源数	530	5630	8830



盘完源后怎么办呢?

如何分析?

~~我开发了 data-valley 数据分评估系统（已上线），采集完了，可以查看源的采集数量、独有性，为你分析源的质量以及最终效果。~~

(因无人维护,20240923 在 data-valley 上回滚了所有改动)

如何开发?

1\~~俺正在开发一个与 AISpider 配套的低代码采集平台 venom-simple。它可以做到非技术人员配置采集爬虫，有点像八爪鱼采集器，相比八爪鱼，降低了一些通用性，从而更简单好用（再也不用到处找爬虫王程师求爷爷告奶奶啦）。使用文档点这里。~~*(最终因工作调动没有实现,争取在下次工作中把这个实现吧)*

2\对于结构相似的长尾源,我开发了 AISpider 爬虫模板生成工具,可以批量开发上万个源,在招投标场景进行了应用 https://workspace.dingtalk.com/cDLnN8ocfFvNcnp5DDpDM5

如何保证质量?

开发了【小刺猬质控平台】，能够分析我们的数据和竞品的比对情况，实时掌握数据与竞品的领先和差距，由菊姐负责使用。