SEO大小站关系

lucifaer / 2017-10-16 05:03:00 / 浏览数 3509 技术文章 技术文章 顶(0) 踩(0)

对于在黑产seo研究中遇到的大小站关系的一点看法，如有疑问和其他想法，望不吝交流：D

最近在对出入链分析的时候，我提出了这么一个基础的问题到底什么样的站是大站，什么样的是小站。我们都知道腾讯，新浪这样的站是大站；政府网站也是大站；类似360

上面的例子中，我提到了三个向量的网站：高流量，高用户；政府网站；企业官网。他们之中，有流量高的，有出入链高的，有比重高的，有搜索权威认定的。那么在做分析

接下来我们就针对网站的各个向量来范式的对网站进行分析。

0x00 PR值

PR值算是一个比较老的评判标准了，我们不再概述PR值到底是什么了，我们关注一下PR值为我们分析seo提供了什么样的维度。

1. PR值现在的实际意义

了解过PR算法的人应该明白，PR值实际上表现了不同内容网站间的相对值，也就是说它最重要的是表现了不同站之间的相对重要性。从严谨的角度上来说，PR值是一个相对

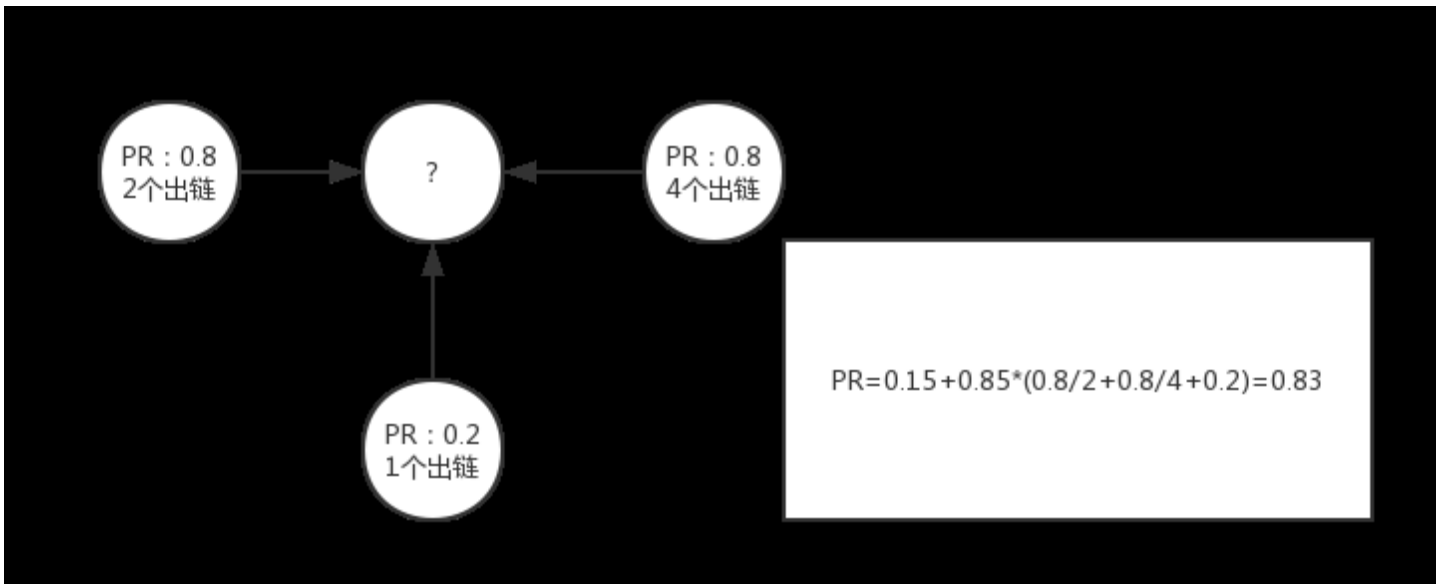
从上面我们可以说PR值并不能成为判别网站大小的依据，它的缺陷在于并没有考虑到流量对于排名的影响。

它具有的实际意义是反映了某个网站的是否为一个较为“权威的网站”。

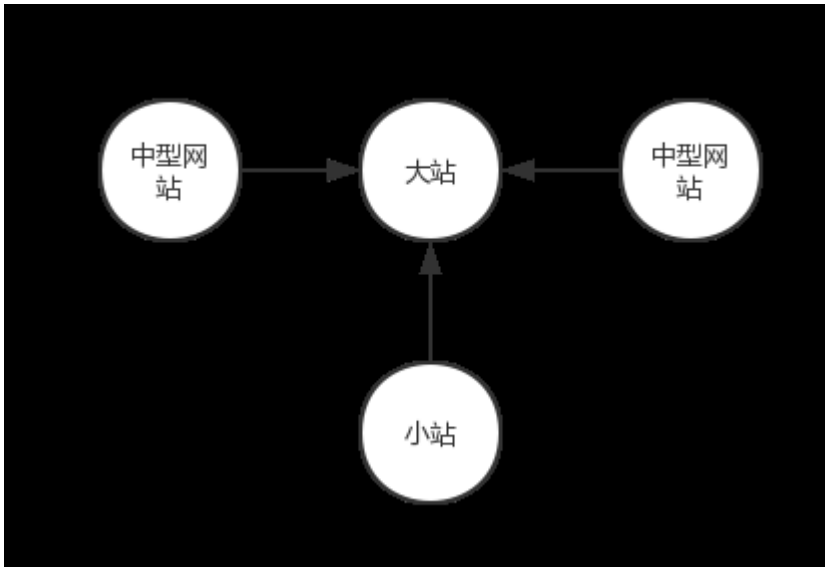
2. 根据PR值所得的相对结果

PR值分析的相对结果是由于PR值本身的相对性来说的。具体可以总结为下面四种模型：

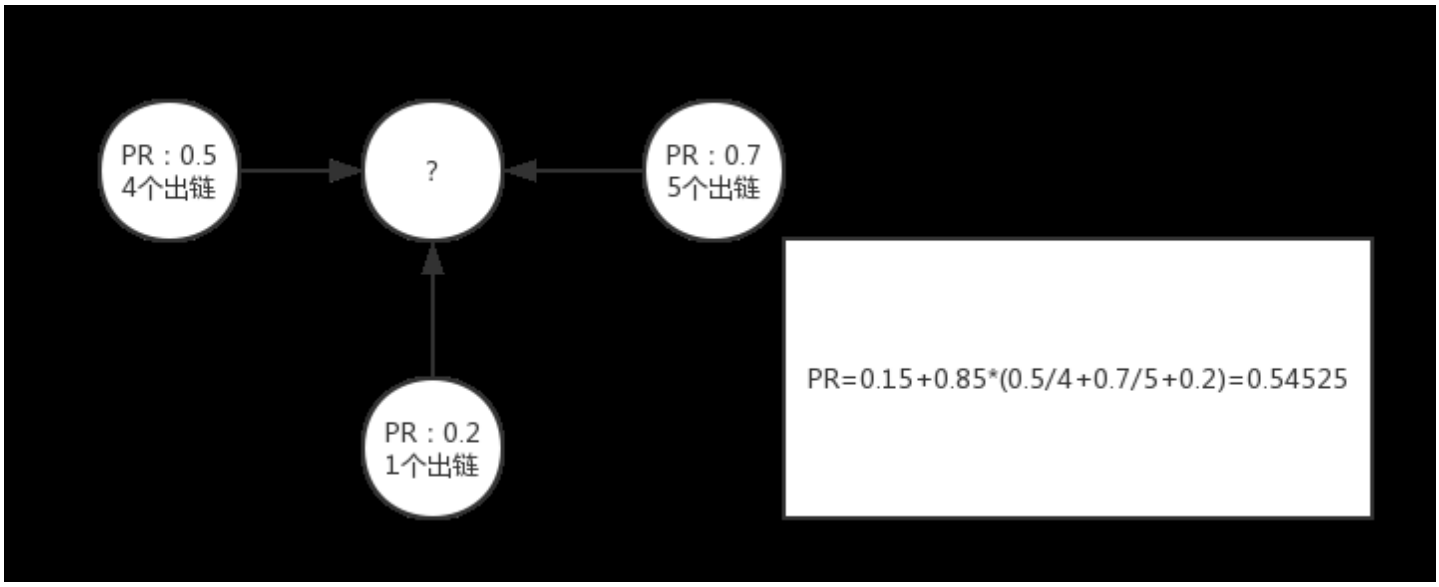
1. 权重大



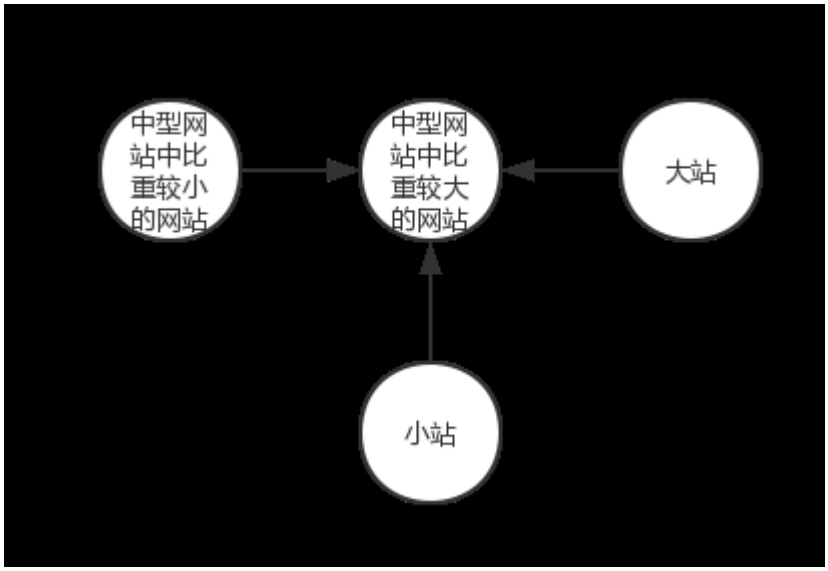
中间网站的PR值相对于其他的三个网站的PR值较大，所以可以说在这四个站的关系中，中间的站为权重大的站：



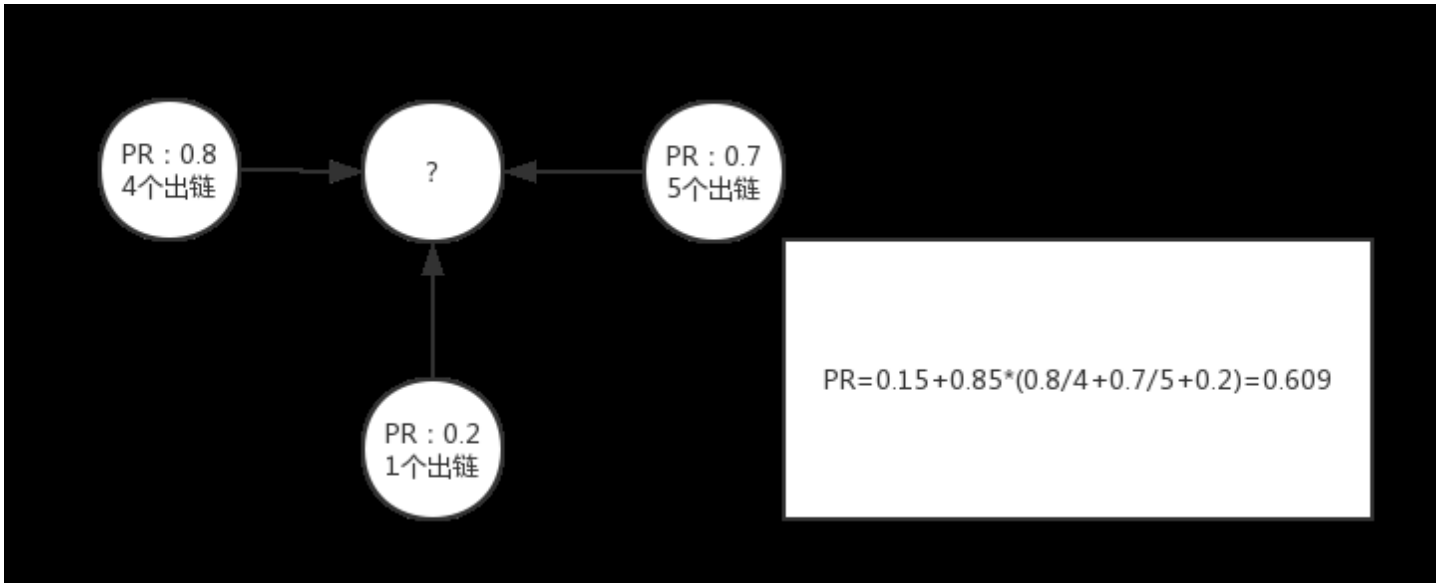
2. 权重较大



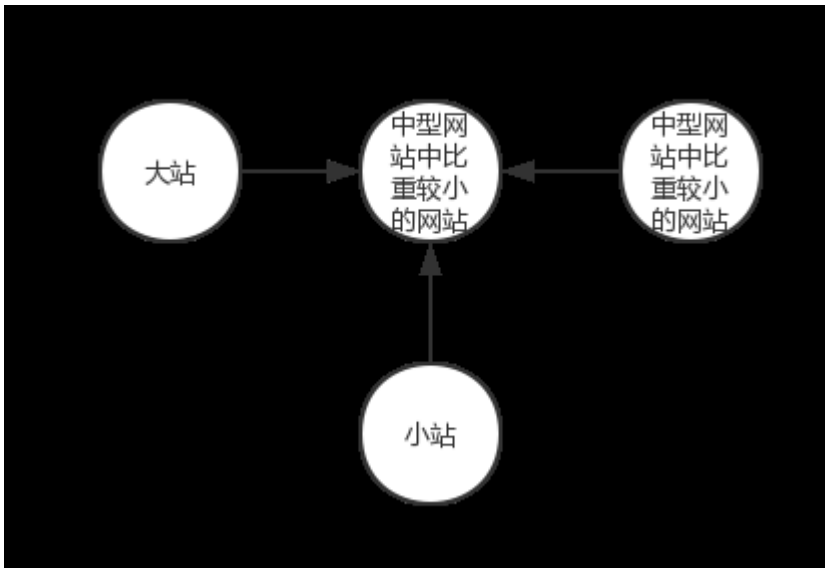
同理，中间的网站相较于两边的网站处于中型网站中所占比重较大的网站，所以其关系可以用下面的图来说明：



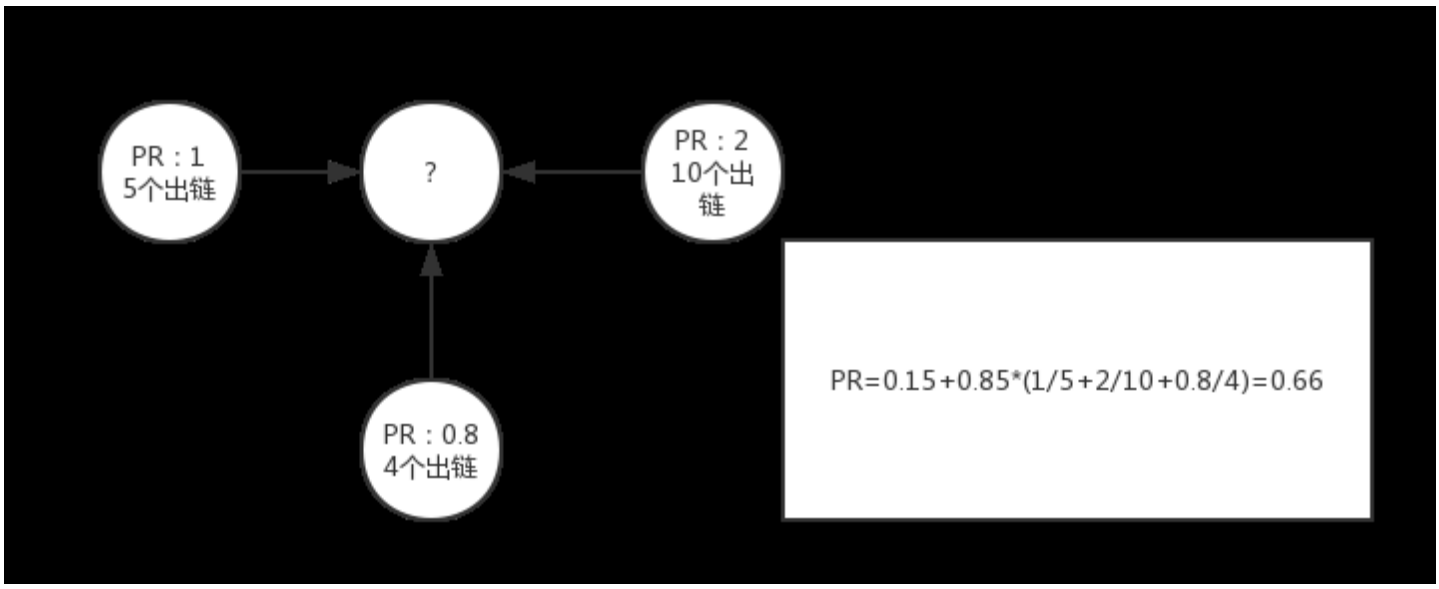
3. 权重较小



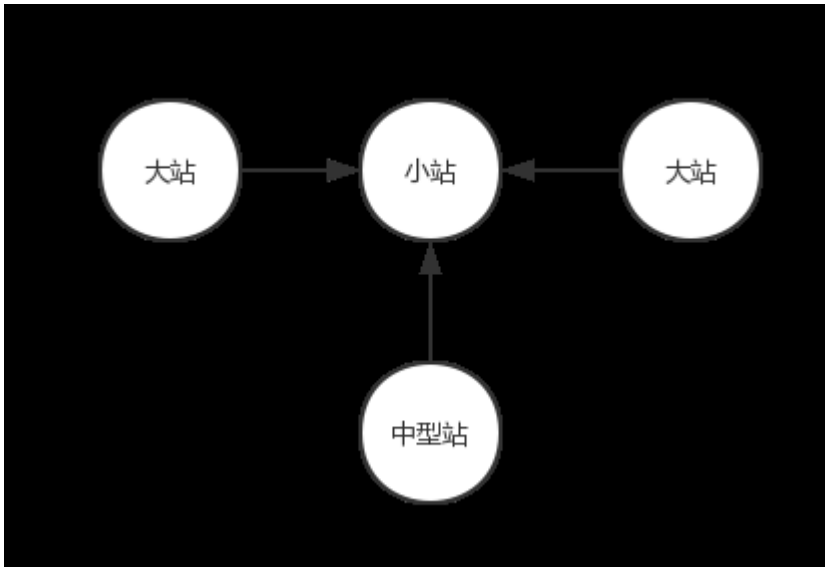
这个道理和上面相同，只不过因为得到的PR值较小，而成为了中型网站中权重较小的网站。



4. 权重小



这样的情况原理和上面两种的原理相同，其关系为：



3. PR值局限

通过上面的分析，我们可以明显的看到PR值的局限：PR值关注站与站之间的关系，所以其本质来说是一种相对的值的关系，通过PR值来确定大小站关系，本来就是不科学与

0x01 Alexa排名

Alexa排名对于做seo优化的人来说，是一个比较重要的判断向量。以下我们从Alexa指数的角度来看一下网站。

1. Alexa排名的意义

Alexa网站排名的计算是以网站的每天平均使用人数、人均访问页面数、与其他网站的链接和曝光数、网友所留言讨论的消息篇数等信息为基础，并以比重不明的加权平均数

从上述的介绍中，我们可以看出，Alexa排名主要完成的工作是评估网站的受欢迎指数。同时由于其数据是从Google
Chrome、Firefox、IE来获取数据的，可以反映一般用户流量的情况，以及网站的权威程度。

2. Alexa排名的缺陷

如上所说，Alexa排名的优点，其实就是它自己的缺陷。从Alexa本身的统计手法就可以看出，Alexa排名对于个人站点、小型企业站点来说，其统计的数据是不准确的，同时

除非是纯互联网公司（比如做互联网广告的，盈利性论坛什么的），否则，该排名指数是不会影响到实际业务的。

其次，Alexa排名只是统计了Google
Chrome、Firefox、IE的数据，对于国外的网站来说，数据统计是较为准确的，但对于国内网站来说，准确性又要低一个层次。

3. 一些可以看到的现象

在中小型网站中，我们发现了一些非常有意思的现象：相当一部分高PR值高流量的网站，Alexa排名很低。这样的现象产生的原因我不再赘述，通过这样的现象而得出的结论

0x02 出入链流量

顾名思义，无论网站如何进行seo优化，其最终的目的就是为了提高搜索引擎收录量及该网站的流量。其seo的手法都是可以从流量上表现出来的。接下来我们从出入链流量

1. 统计出入链流量的数据源

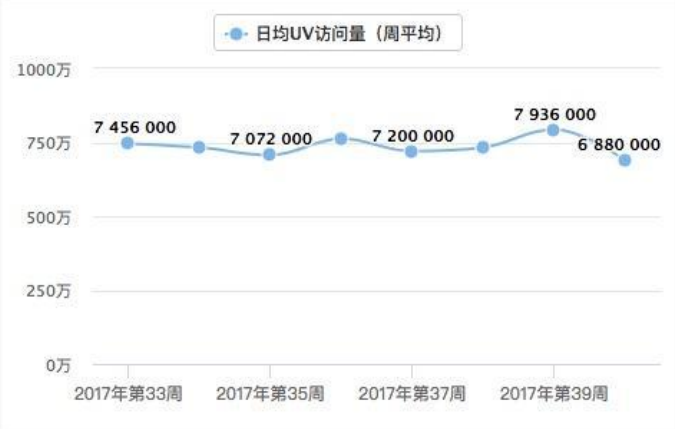
出入链流量的数据表现在日均uv以及日均pv值，相对来说，uv值是比较准确的，但是缺点是对于中小型网站的uv数据需要自己进行收集，网络上并没有现成的统计数据。

比如`bilibili.com`：

网站 bilibili.com 的全球网站排名与 UV & PV 值 以下UV&PV数据为估算值，非精确统计，仅供参考

周期	全球网站排名	变化趋势	日均UV	日均PV
当日	223	↑ 19	8320000	79040000
周平均	257	↑ 10	6848000	68137000
月平均	244	↑ 5	7296000	64642000
三月平均	247	↑ 11	7296000	65372000

网站 bilibili.com 的预估流量 以下UV&PV数据为估算值，非精确统计，仅供参考



看日均uv和pv为百万级，算是一个非常大的站点了，再看看我的博客...

网站 lucifaer.com 的全球网站排名与 UV & PV 值 以下UV&PV数据为估算值，非精确统计，仅供参考

周期	全球网站排名	变化趋势	日均UV	日均PV
当日	-	0	-	-
周平均	-	0	-	-
月平均	-	0	-	-
三月平均	-	0	-	-

网站 lucifaer.com 的预估流量 以下UV&PV数据为估算值，非精确统计，仅供参考

站点太小了，干脆就没有收录。

2. 从出入链流量角度分析问题的缺陷

1. 首要的缺陷，就是没有一个确定的标准。具体来说就是没有一个准确的uv值或pv值来指明该网站的大小。

举个例子：相较于我的个人博客，bilibili算是一个大站，但是bilibili相较于qq.com...

网站 qq.com 的全球网站排名与 UV & PV 值 以下UV&PV数据为估算值，非精确统计，仅供参考

周期	全球网站排名	变化趋势	日均UV	日均PV
当日	9	0	212160000	833788000
周平均	9	0	207360000	806630000
月平均	9	0	207968000	808995000
三月平均	9	0	211360000	830644000

网站 qq.com 的预估流量 以下UV&PV数据为估算值，非精确统计，仅供参考

就根本不是一个数量级上的问题了。这个时候我们会理所应当的认为qq.com是我们所说的大站，而bilibili是一个小站。那....

网站 google.com 的全球网站排名与 UV & PV 值 以下UV&PV数据为估算值，非精确统计，仅供参考

周期	全球网站排名	变化趋势	日均UV	日均PV
当日	2	0	1262720000	9609299000
周平均	1	0	1322240000	12045606000
月平均	1	0	1341856000	11982774000
三月平均	1	0	1333792000	11657342000

网站 google.com 的预估流量 以下UV&PV数据为估算值，非精确统计，仅供参考

所以说，从这个角度看，我们不好说哪个是大站，哪个是小站。

1. 其次，对于一些小的网站（个人博客，或者说是没有插入相关统计代码的网站）很难得到其uv值与pv值。
2. cdn与cname解析的域名会对统计结果造成很大的影响。

0x03 那该怎么做？

说了这么多，我们到底该怎么解决这个问题？我用我手头上的数据，提出一个判定的思路。以下截图为一个demo数据，并不具有准确及实际意义。

手头上的数据

1. Alexa排名top n（基本上没有什么用）
2. 目标域名的出入链情况
3. 目标域名及出入链的PR值
4. 出入链的uv值

要解决的问题

识别目标域名中有哪些域名做了seo

初步的设想

1. 通过出入链关系对目标域名进行分组：分组为源域名-出入链域名：

id	source_domain	relation_domain	date	type
1	http://000000a.com/pr.jsp?	http://jz.faisco.com/?_ta=4	0809	out
2	http://000000a.com/pr.jsp?	http://wpa.qq.com/msgrd?menu=yes&site=qq&uin=2111641845&v=3	0809	out
3	http://000000a.com/pr.jsp?	http://www.faisco.com/ts.html?a=a17086024562&t=3	0809	out
4	http://000000df.cn/news/18069.html	http://www.xqhzzorf.ga/news/68341.html	0810	in
5	http://000000df.cn/news/25087.html	http://www.sittwb.tk/news/74423.html	0810	in
6	http://000000df.cn/news/26081/94331.html	http://soemcdor.gq/news/37800.html	0810	in
7	http://000000df.cn/news/39475.html	http://xj3el04.cn/news/10055.html	0810	in
8	http://000000df.cn/news/48022.html	http://6cnhfbp.cn/news/17846.html	0810	in
9	http://000000df.cn/news/55040.html	http://vukaqqvv.ga/mfvoy/	0810	in
10	http://000000df.cn/news/77381.html	http://www.imvovjjs.ga/news/19051.html	0810	in
11	http://000000df.cn/ub1.html	http://fere1uy.cn/news/16986.html	0810	in
12	http://00004t.cn/viewspace-284.html	http://393845.ass9.cn/	0811	in
13	http://00004t.cn/viewspace-284.html	http://394.ius3.cn/	0811	in
14	http://00004t.cn/viewspace-284.html	http://949732.fhs3.cn/	0811	in
15	http://00004t.cn/viewspace-284.html	http://aea.igs6.cn/	0811	in
16	http://00004t.cn/viewspace-284.html	http://edf.ifs8.cn/	0811	in
17	http://00004t.cn/viewspace-284.html	http://ehmdrf.ccz56.cn/	0811	in
18	http://00004t.cn/viewspace-284.html	http://fco.fys3.cn/	0811	in
19	http://00004t.cn/viewspace-284.html	http://fsohax.aps7.cn/	0811	in
20	http://00004t.cn/viewspace-284.html	http://gsg.fds3.cn/	0811	in
21	http://00004t.cn/viewspace-284.html	http://hng.ffs6.cn/	0811	in
22	http://00004t.cn/viewspace-284.html	http://huicui.dus9.cn/	0811	in
23	http://00004t.cn/viewspace-284.html	http://mej.ius9.cn/	0811	in
24	http://00004t.cn/viewspace-284.html	http://mhv.cks5.cn/	0811	in
25	http://00004t.cn/viewspace-284.html	http://osw.cds3.cn/	0811	in
26	http://00004t.cn/viewspace-284.html	http://pfw.00004i.cn/	0811	in
27	http://00004t.cn/viewspace-284.html	http://pvm.fgs3.cn/	0811	in
28	http://00004t.cn/viewspace-284.html	http://sgufli.00004f.cn/	0811	in
29	http://00004t.cn/viewspace-284.html	http://shishicaizhucesongcaijintuanduiqun.00004j.cn/	0811	in
30	http://00004t.cn/viewspace-284.html	http://shishicaizuojia.ils8.cn/	0811	in
31	http://00004t.cn/viewspace-284.html	http://tcp.kc83.cn/	0811	in
32	http://00004t.cn/viewspace-284.html	http://tgu.ihs8.cn/	0811	in
33	http://00004t.cn/viewspace-284.html	http://tur.ays2.cn/	0811	in
34	http://00004t.cn/viewspace-284.html	http://wenku.kis3.cn/	0811	in
35	http://00004t.cn/viewspace-284.html	http://wwsblz.dcs1.cn/	0811	in
36	http://00004t.cn/viewspace-284.html	http://www.168.fbs1.cn/	0811	in
37	http://00004t.cn/viewspace-284.html	http://www.502.cys7.cn/	0811	in
38	http://00004t.cn/viewspace-284.html	http://www.jiangxishishicaimianfeipojieruanjian.ags6.cn/	0811	in
39	http://00004t.cn/viewspace-284.html	http://xby.bis6.cn/	0811	in
40	http://00004w.cn/viewspace-100.html	http://277488.fvs5.cn/	0811	in
41	http://00004w.cn/viewspace-100.html	http://351147.irs1.cn/	0811	in
42	http://00004w.cn/viewspace-100.html	http://592836.its3.cn/	0811	in
43	http://00004w.cn/viewspace-100.html	http://689952.ccs3.cn/	0811	in
44	http://00004w.cn/viewspace-100.html	http://dso.igs8.cn/	0811	in
45	http://00004w.cn/viewspace-100.html	http://eaqgwv.ccz45.cn/	0811	in
46	http://00004w.cn/viewspace-100.html	http://egzjpe.fas6.cn/	0811	in
47	http://00004w.cn/viewspace-100.html	http://ewb.dxs9.cn/	0811	in
48	http://00004w.cn/viewspace-100.html	http://ezwsli.dos6.cn/	0811	in
49	http://00004w.cn/viewspace-100.html	http://fsokkh.axs8.cn/	0811	in
50	http://00004w.cn/viewspace-100.html	http://gtg.iss1.cn/	0811	in

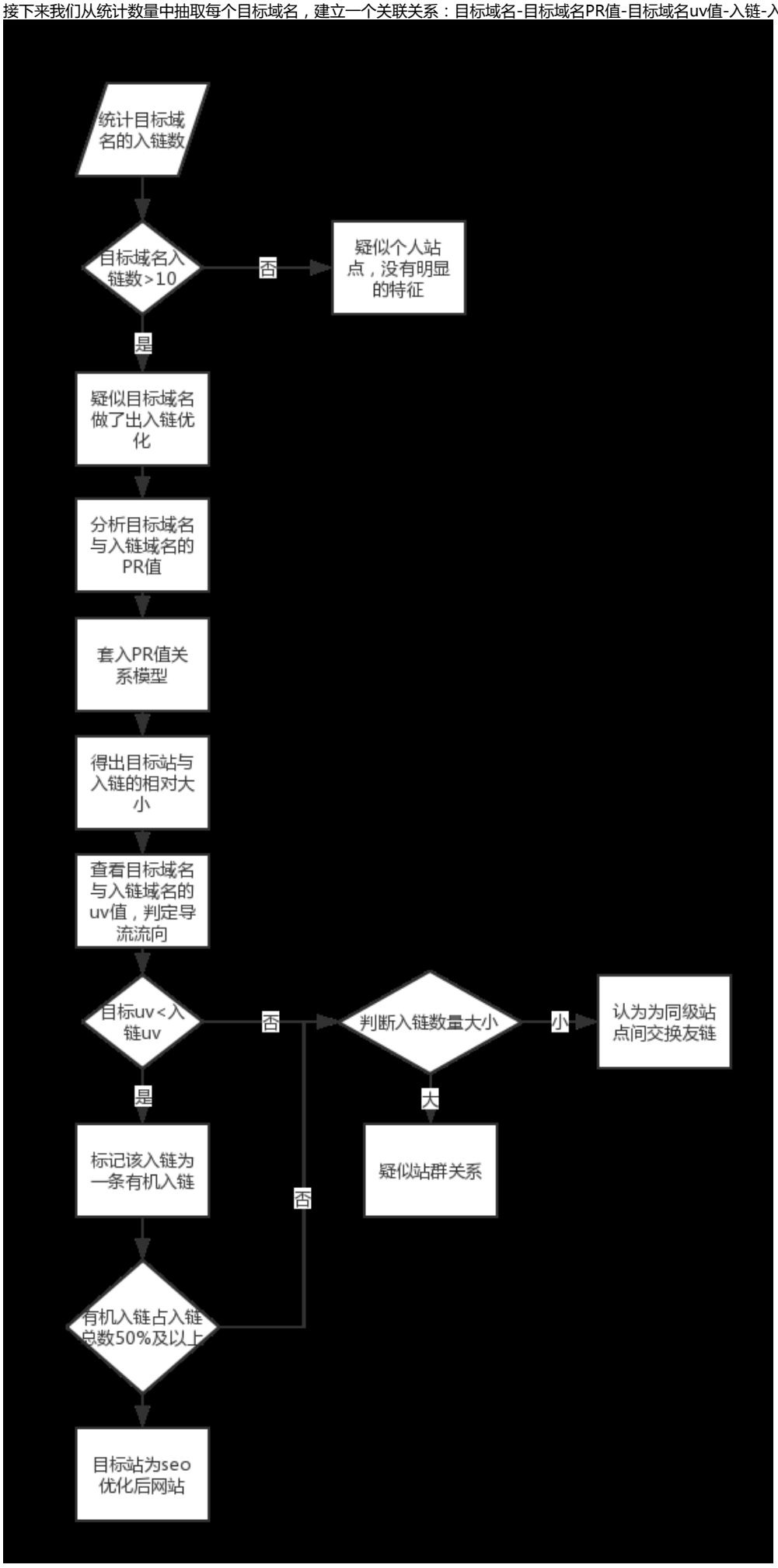
1. 给分组后的所有域名进行PR标识。这边需要注意的一个问题就是，很多网站是没有PR值的，这对于我们解决问题是一个非常大的阻碍。

id	source_domain	pr
1	007jp.5ah53g8.cn	0
2	00crku.dyzqx5k14.cn	0
3	00psjn.dyzqhhlyh.cn	0
4	010u1.579744665.cn	0
5	012.5761514.cn	0
6	019.342.273.jltxg.cn	0
7	01ixc.ogtng.cn	0
8	025805.pbdbh.cn	0
9	02660.lzbbh.cn	0
10	02dtqw.tfswh.cn	0
11	02xr7.china-mudiao.cn	0
12	031842.cn	1.81721
13	049508.com	0.071868
14	04uo3.gzrxbh.cn	0
15	0556wjw.com	4.97907
16	0565qc.lhzqhipos.cn	0
17	058.fg641k.cn	0.145883
18	05807.jfpbh.cn	0
19	059l5.zbkgpk.cn	0
20	05wdr9.afgdb.us	0.098817
21	060248.com	0.061197
22	0609kbr.766932100.cn	0
23	074908.com	0.07599
24	07947.qogsy.cn	0
25	0871kmlife.com	3.79339
26	08e.618814540.cn	0
27	08i.670036117.cn	0
28	097333.tdzkz.cn	0
29	09809.sycmcu.cn	0
30	0cf.608735646.cn	0
31	0cg5f0.lhzq2d59l.cn	0
32	0esmkk.zhangguixw8346.cn	0
33	0euqeq.rzruh.cn	0.121699
34	0ezfeq.lhzqxlzwe.cn	0
35	0fq.710981405.cn	0
36	0fvlq.pfydt.cn	0
37	0fwaq.696129741.cn	0
38	0h8zbb.15571759x.cn	0
39	0hjw9r.gaoqiws5716.cn	0
40	0j79x.pzrvr.cn	0
41	0j8lj.vtt777.info	0
42	0k4.zg7pde.cn	0
43	0kahp4.18jajtt.cn	0
44	0kesyy.vlebq.trade	0.018923
45	0koibq.672607.com	0.171655
46	0m0.pfxxo.cn	0.089515
47	0m0q0a.liuchefy5287.cn	0.213095
48	0p0fr9.pfkxt.cn	0
49	0qqey8.dyzqt0kh7.cn	0
50	0r6cp5.lhzqkqnjo.cn	0.14248

1. 统计原站与出入链网站的uv值，这边有着同上面一步一样的问题，那就是对网站的出入链uv值统计不完全，造成并不能建立一个强关系。（由于数据不方便展示，就不贴
2. 之后将目标站与入链数建立关系：目标站-入链数量：

id	domain	relation_num
3	000000a.com	2
4	000000df.cn	8
5	00004t.cn	28
6	00004w.cn	27
7	00004x.cn	29
8	00005c.cn	10
9	00005d.cn	5
10	0002008.com	1
11	0003q.cn	1
12	001011avvvav.cn	1
13	001011msssms.cn	1
14	001cl.net	6
15	001dr.com	3
16	0024xx68x20z.dginfo.com	1
17	003220.cn	1
18	005619.cn	2
19	005626.cn	2
20	005635.cn	1
21	005701.cn	1
22	005702.cn	2
23	005703.cn	1
24	005707.cn	3
25	005713.cn	1
26	005719.cn	3
27	005768.com	12
28	005800.com	2
29	005902.cn	3
30	006b.com	2
31	008418.cc	49
32	00ebbs.cn	1
33	00ji.cn	8
34	00rugov.cn	1
35	00zhgov.cn	1
36	00zq6c.cn	1
37	00zrifw2da.cn	1
38	01064697666.com	1
39	010dyzc.com	1
40	010goode.net	17
41	010gsbz.com	2
42	010kangfu.com	4
43	010yhzx.com	2
44	01122.co	1
45	012.pe	5
46	0146789.com	8
47	0150520.com	2
48	01cfgov.cn	1
49	01eigov.cn	1
50	01gegov.cn	2
51	01loft.com	4
52	01rtgov.cn	1

1. 接下来我们从统计数量中抽取每个目标域名，建立一个关联关系：目标域名-目标域名PR值-目标域名uv值-入链-入链数-入链PR-入链uv值。之后可以用下面的思路来进行



当然这个模型只是一个想法，还没有数据支撑。PR值关系模型在前文中已经有所提及。

1. 注意：
   这个模型中对于源数据的需求量较大，个人在做demo的时候调查了近4w的域名，经过引入其他向量后数据量约为500w左右。经过数据挖掘及处理后，得到的效果非常
   源数据的准确性要求较高，出现缺少PR值统计及uv值时，基本上是没有什么说服力的。
   * 大小关系本来是一种相对关系，在中间数据挖掘时会遇到各种各样的问题，尝试引入多种向量来解决。

0x04 展望

出入链对于seo来说是较为本质的数据，从中还有更多可挖掘的信息，接下来我抛砖引玉，提出自己的一个思路。

前段时间有人利用微博作为媒介，做了一个安全圈有多大的画像，我们可以把这个思路应用到出入链分析上。

出入链的优势是，从一个端点总能到达另外一个端点，那么我们就把我们“行走”的过程记录下来，是不是能对seo网站进行画像呢？

当然这是我的一个思路，这个实现起来是有难度的，但是我觉得是有意义的。

0x05 总结

通过建模，从理论上探讨了各项判定向量的可行性，并提出了一个初级的处理大小站关系的模型，并提出了对该工作的展望。希望对有同样问题的朋友有帮助。

上一篇：CTF线下赛AWD套路小结 下一篇：做为技术人员为什么要写文章分享？

1. 0 条回复
   - 动动手指，沙发就是你的了！