Hubdig问题梳理

彭石

2013-5-30

目录

[配额问题 3](#_Toc357635370)

[site\_level的定义 3](#_Toc357635371)

[打分和排序、配额突破策略 4](#_Toc357635372)

[新闻，博客，BBS版面页的补充数据 5](#_Toc357635373)

[无效参数问题 5](#_Toc357635374)

[Hub页生效问题 7](#_Toc357635375)

[理论上的收益 8](#_Toc357635376)

[历史CASE回顾 9](#_Toc357635377)

[附录：分析数据 10](#_Toc357635378)

[配额变化的结果： 10](#_Toc357635379)

[策略变化的结果 11](#_Toc357635380)

中文hubdig的代码已经很久没有升级，各种遗留的飞线，遗留的策略需要重新评估一下。数据在变化，很多策略需要设定。现在对hubdig进行梳理，总结问题如下。

# **配额问题**

之前有一份site\_need配额数据。但已经没有产出。近一年都是用了那份数据，肯定会存在问题。一是站点在变化，站点配额肯定也在变化。

Hubdig产生的hub页在spider环上不受限制，有多少尽量抓多少。但是，Hubdig还是需要站点配额：一是要从总量上限定hub页数量，二是要在各站点之间协调分配比例。

之前hubdig有一个配额分配环：从cc取site\_need数据，据此产生hub页结果集合。然后再统计站点site\_need数据与实际站点上hub页集合的差异，产生一个需求调整文件给CC，然后CC再据此重新计算site\_need。CC那边早就不计算site\_need了，这个逻辑环早就断开。

目前CC算给大环时效性chk的配额，考虑了站点规模，评级，时效性收益计算的结果，以及hubdig产出的需求调整量。改成用大环时效性chk的配额来指导hubdig在各站点间hub页的分配，是合理的。

已经解决。

# site\_level的定义

site\_level是hubdig内定义的站点质量评级，站点分类3级别site\_level，0 1 2。

0级别最低，大致对应siterate 0 1 2

1级别，大致对应siterate 3 4 5

2级别，大致对应 6 7 8

但是有些调整策略，

当wdn小时库一天入库量 < 3，则将site\_level==1的置为0。这个策略在大部分情况下是合理的。但是当site\_leve==1，site\_rate==5时，会误伤一些还不错的站点。比如top.ku6.com等评级为5，但是被这个策略伤了。

已经解决。

# 打分和排序、配额突破策略

目前hubdig在配额限制步骤中，对URL的排序以：站点（同一个站点在一起）/来源类型（人工排前面，indexCheck结果在后面）/去重标记(正常的在前面)/score（高的在前面）四项依次为准，排序方式是合理的。

但score计算和配额突破条件可能存在问题：

计算公式

score = 10 \* ext\_link + 10 \* sobar / 1000000 + 10 \* (sobar / 1000 % 1000) + 10 \* pr + inner\_link + 10 \* all\_new + self\_new;

配额突破条件：

if (site\_use + self\_new > limit) {

if (is\_firstpage || is\_ext || (score >= 100 && dup\_flag == 0 && prefix\_level > 0 && pd < 300)) {

突破配额限制

} else {

不突破限制

}

}

1：目前给url打分的公式是当时拍下的，各影响因子的权重未必合理。selfnew的权重在计算score分数的时候很低。导致很多self\_new较高的被限制住。

2：pd<300这种策略有硬伤。如果某个pattern模式大于300，则全部删除。这个策略硬伤太大。

解决方法：

结合CASE的情况，重新考虑打分公式，并调整配额突破的条件。从CASE的情况看，对于self\_new比较大的可以考虑突破，对站点级别高的可以放松条件。

分析具体数据可以看到：被配额限制的部分绝大部分分值很低，分值>50的已经很少，所以考虑放宽阈值。

对pd的策略，改成对pattern下url进行计数。超过阈值再放弃，至少可以保证捞回部分的符合条件的链接。

已经解决。

# 新闻，博客，BBS版面页的补充数据

目前使用的新闻频道数据，博客飞线，BBS版面页飞线比较旧，来源混乱，而且量比较很少，有些模块已经停止更新了。

现在WDM统一挖掘新闻频道页，博客，和BBS版面页，加入了点击展现等数据。用这份数据与现在的hubdig结果集合进行对比，有几万个hubdig没有挖掘出来的小新闻站点。此类网站新闻发布频率比较低，但还是有收益。

已经生效。

# 无效参数问题

**背景**：

无效参数，是指url带有的参数不影响页面的内容。比如news.qq.com/index.shtml ， news.qq.com/index.shtml?cal=1&date=20，这个两个URL，其实是一个页面。

某些站点下，带有无效参数的URL大量挤占站点配额，并且重复chk同一个页面，降低了hub页集合的质量。这种问题容易爆发在某个站点上。

以news.qq.com/index.shtml为例，一个URL，因为无效参数，变成了50个hub页。

news.qq.com/index.shtml

news.qq.com/index.shtml?cal=1&date=20

news.qq.com/index.shtml?cartid=8&cartnam

(…省略几十行…)

news.qq.com/index.shtml?act=add...4

另外一例子

bbs2.dl.net.cn/forum.php

bbs2.dl.net.cn/forum.php?mod=vie

bbs2.dl.net.cn/forum.php?fromuid=15906

bbs2.dl.net.cn/forum.php?fromuid=37266

……

bbs2.dl.net.cn/forum.php?mod=

hubdib有通过DC的model\_func词典来去除无效参数。但hubdig获取的索引页，都是从链接库中取的，已经是经过dc策略的。如果还带有参数还是无效的，则是DC的策略也去除不了的。所以通过DC的model\_fun词典去不掉已经存在的无效参数。

解决方案：

可以考虑后链的重复度来去除无效参数。

比较无参数的URL（news.qq.com/index.shtml），与其对应的有参数的URL（news.qq.com/index.shtml?cal=1&date=20），如果后链重复度非常高，表示这个参数是无效的。

IndexCheck在计算强度谱之前，有保留前链以及其follow出来的后链集合，可以考虑通过这个步骤的数据，来挖掘无效的参数。

这个可以用俊伟准备做的站点内后链覆盖度数据来优化，正在跟进。

# Hub页生效问题

当前hub页生效及退场的策略如下：

1. Hubdig每天挖掘hub页，回灌linkcache中多次抓取
2. 每天回灌的hub页词典加载到DC，DC的策略如下：如果前链存在hub页词典中，后链hubdepth值+3；如果前链不存在词典中且hubdepth为0-2，则更新为7，否则保持不变，此时后链赋值默认值7
3. Saver生效策略：非缺省值一律替换缺省值；同为0—2区间取最新值；同为3—6区间取小值；0—2区间值可替换3—6区间值；对已抓链接，新链接为3-7(缺省值包括)可替换0—2区间值；其他情况，保留原值

**存在问题**

看策略1,2，存在问题如下：

Hub页回灌linkcache与加载到DC的时间耦合性太强。如果某条链接A，在第一天没有挖掘为hub页，但是第2天挖掘为hub页，但是回灌linkcache后，DC词典尚未更新，那么该hub页调度一次后，hubdepth就会变为7而被删除，所以要求先加载词典到DC，再回灌linkcache。但这样也有个问题：

* 1. 如果某个hub页，在第一天挖掘为hub页，第二天没有，提前加载DC词典后，该hub页还没有调度完成就被删除，不符合策略预期。但是这个影响不大，因为第二天未识别为hubub页，表明质量以及不行了，只是提前几个小时被删除而已。
  2. DC的词典加载需要加载100台机器，需要一定的时间，且存在加载超时等情况，造成词典加载不成功（先前出现过），这种情况导致的误伤很大。

所以，两者之间的时间耦合对策略的效果存在很大的影响。

**解决方案**

1. DC不根据词典来对前链的hubdepth做设定，而是根据下发的包头中传递的值来确定当前链接的hubdepth，以及计算其后链的hubdepth
2. 由于saver存在该策略：对已抓链接，新链接为3-7(缺省值包括)可替换0—2区间值，所以对于hub页来说，可以自然退场；对于新挖掘的hub页，只需要每天回灌linkbase即可，而这个操作时间上相对宽松，减少策略的耦合。

开始推进解决

# 理论上的收益

1：Hubdig的配额与CC的时效性配额形成循环。

目前，Hubdig根据cc算的时效性配额分配站点内的hub页，并产生差异调整量，反馈给CC，以便计算下一轮的时效性配额。

这个循环能够动态调整站点配额，自动优化站点配额。

2：site\_level

当wdn小时库一天入库量 < 3，则将site\_level==1的置为0。这个策略在大部分情况下是合理的。但是当site\_leve==1，site\_rate==5时，会误伤一些还不错的站点。比如top.ku6.com等评级为5，但是被这个策略伤了。

3: 优化配额突破策略

（1）捞回的阈值，从100降为50，能够捞回更多高质量的hub页。

（2）原来同pattern下url数>300，全部砍掉的策略，现在改为计数，当超过阈值，后面的抛弃，前面的保留。这样避免把某pattern的hub页全部砍掉。如某一类型的论坛版面页等。

4：wdm那边的新闻BBS论坛数据参考了点击展现数据，有一定的补充作用。

diff了数据，有几万个新闻站点页补充进来。

5：无效参数问题

与军伟继续沟通解决方案

6：hub页生效问题

开始推动相关方一起解决，安排在6月中旬解决。

# 历史CASE回顾

之前的一批CASE，现在基本都能够解决。

1：bbs.csdn.net/forums/CSharp

2：www.kuvcd.com/List/List\_v\_c\_31\_pg\_\_s\_\_ty\_1.Html

这两个链接之前因为inner\_link, sobar extern\_link pagerank这些值都为0，所以突破只能靠all\_new和self\_new。之前报CASE的时候，self\_new和all\_new不够，没有突破配额成功，现在all\_new，self\_new已经逐渐增大，hubdig已经挖掘出来。目前在保持hub页总量大致不变的情况下，下降了配额突破的score的阈值，也能更快挖掘出这种hub页。

3：www.sxdaily.com.cn/GB/44/45/ 这个CASE因为link\_depth>10,被indexCheck标记了filter\_type，然后在 hubdig 中被过滤掉。暂时无法解决

 3：wei100.cn/

这个站点下一些前链，其selfnew值接近100，但是无all\_new。现在可以挖掘出来了

[www.51edu.com/daxue/](http://www.51edu.com/daxue/)news

extern\_link很高，self\_new all\_new 3 0 ，实际页面基本少量产生新链接，分值超过100，，但是因为过wdn筛选的小时级统计<3，导致site level=0 ，没有突破。现在可以解决。

top.ku6.com/v\_t1d1c101000p1.html，被site\_level策略误伤，现在可以解决。

[www.awotuan.com/feed.php?ename=beijing](http://www.awotuan.com/feed.php?ename=beijing)

此前无法加载到linkcache，现在解决。

总结：从CASE的情况看，优化一些策略细节后，能够解决很多历史上的CASE。

# 附录：分析数据

Hubdig的变化

1：配额数据的变动

2：专题页挖掘数据的补充。

3：过滤条件的优化。

抽样分析的站点

t.163.com <site\_stat.29:> 4903 <site\_stat.09:> 25261

www.xcar.com.cn <site\_stat.29:> 774 <site\_stat.09:> 627

www.alibaba.com <site\_stat.20:> 23988 <site\_stat.09:> 24008

www.qjy168.com <site\_stat.29:> 14891 <site\_stat.09:> 12394

v.ku6.com <site\_stat.26:> 2893 <site\_stat.09:> 13411

photo.blog.sina.com.cn <site\_stat.26:> 75108 <site\_stat.09:> 14862

news.sina.com.cn <site\_stat.29:> 2544 <site\_stat.09:> 632

news.qq.com <site\_stat.29:> 1753 <site\_stat.09:> 638

## 配额变化的结果：

1：某些站点配额变大很多，导致其下的url，不需要经过配额突破策略，大部分低质量的前链也成为hub页，以photo.blog.sina.com.cn为代表。这个变化会引入低质量的hub页，无负面影响。

2：某些站点配额变小，导致其下的url，需要走配额突破策略。我们认为时效性配额是可信的，所以这一变化也不会有太大的负面作用。(目前配额形成反馈环，如果目前配额小，会产生调整量反馈到CC，理论上后面会越来越大)

## 策略变化的结果

1：在hub页数量不变的情况下，配额突破的条件下降，pattern策略的优化，能够捞回更多高质量的前链

配额9613.900

t.163.com

相同3137

基本上都是减少（2万个）

ent.163.com/special/tbbts/ 最近无新链接

ent.163.com/special/yqwgz/ 最近无新链接

t.163.com/6084002812 不足一天一个。

t.163.com/7423494040 最近无新链接

增加的（4098）

t.163.com/9891535791 最近无新链接

http://t.163.com/1045295414 一个月前贡献很多链接

http://t.163.com/5272602508 最近有很多新链接

ent.163.com/special/selinazuopin/ 最近无心链接，但是是专题页，由专题页挖掘模块导入

ent.163.com/special/gaia/ 最近无心链接，但是是专题页，由专题页挖掘模块导入

t.163.com/5228059354 4月份有产生很多新链接

t.163.com/4211431015 每天调度一次，调度历史显示：mod\_cl。 （但是页面上没有啥新的链接产生）

因为配额大，变化的部分，都是分值很小，排在后面的，影响不大。

配额7505.518

www.xcar.com.cn

变化不大，质量差不多。

配额121954

www.alibaba.com

变化不大，质量差不多。

配额35933多

www.qjy168.com

变化不大，质量差不多。

配额： 2265

v.ku6.com

减少部分，类似于这样的。

v.ku6.com/u/21670888/rss

v.ku6.com/u/21712632/rss

v.ku6.com/u/21515503/rss

最近2个月都没有新链接

增加：少量

配额73290（site\_need.71）

photo.blog.sina.com.cn

相同458个

减少14404

photo.blog.sina.com.cn/amy4118 几个月内无新链接

photo.blog.sina.com.cn/csszsjb 一年内无新链接

增加33606

photo.blog.sina.com.cn/blogpiclist/u/1774617112 最近半月有新链接

photo.blog.sina.com.cn/blogpiclist/u/1886617272 最近2个月内有新链接

photo.blog.sina.com.cn/blogpiclist/u/1583517907 最近一年内无新链接

因为配额大，变化的部分，都是分值很小，排在后面的，影响不大。

配额（7484）site\_need.71

news.sina.com.cn

相同985

减少323

slide.mil.news.sina.com.cn/i\_list\_8\_864.html 一年内无新链接

news.sina.com.cn/z/2012nobel 老专题页

news.sina.com.cn/z/xiejiao 老专题页

news.sina.com.cn/z/dxxjwhn 老专题页

news.sina.com.cn/z/alftkgys 老专题页

增加2205

(news.sina.com.cn下主要是增加了topic\_hub挖掘的专题页，由于topic\_hub暂时会把旧专题页弄进来，所以增加了个旧专题页)

news.sina.com.cn/z/rwxzk51zj/ 老专题页（top\_hub弄进来的）

news.sina.com.cn/z/mtryhtszgy/

news.sina.com.cn/z/xianchao/

news.sina.com.cn/z/bscrash/

roll.mil.news.sina.com.cn/s\_boeing747-8\_all/4/index.shtml 是条老链接，专题页推送过来

roll.mil.news.sina.com.cn/phototj\_slide/741/index.shtml 页面上没有新链接，专题页推送过来

配额4386（site\_need.63）

news.qq.com

相同

664个

减少

260个

基本上都是专题页

增加

2180

基本上全是专题页