Hub页识别的改进调研报告

|  |  |
| --- | --- |
| 撰写人 | 彭石 |
| 文档提交日期 | 2012-11-14 |

百度在线网络技术（北京）有限公司

(版权所有,翻版必究)

修改记录

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | 修改后  版本号 | 修改内容简介 | 修改日期 | 修改人 |
| 1 | V1.0 | 全文 | 2012-11-15 | 彭石 |
| 2 |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

目 录

[1 摘要 1](#_Toc340771265)

[2 背景和目标 1](#_Toc340771266)

[3 hub页未识别改进 1](#_Toc340771267)

[3.1 捞回属性考虑 2](#_Toc340771268)

[3.1.1 索引页的特征 2](#_Toc340771269)

[3.1.2 内容页中HUB页特征 3](#_Toc340771270)

[3.1.3 内容页中普通链接的特征 3](#_Toc340771271)

[3.1.4 捞回属性 3](#_Toc340771272)

[3.2 调研程序 4](#_Toc340771273)

[3.2.1 调研的indexCheck 4](#_Toc340771274)

[3.2.2 调研的hubdig 4](#_Toc340771275)

[4 调研结果评估 4](#_Toc340771276)

[4.1 捞回比例 4](#_Toc340771277)

[4.2 识别准确率评估 5](#_Toc340771278)

[4.3 新增Hub页的流量估计 5](#_Toc340771279)

[5 后继工作 5](#_Toc340771280)

[5.1 rank\_in 和rank\_out的值问题 5](#_Toc340771281)

[5.2 参数确定方法 6](#_Toc340771282)

# 摘要

本次调研主要解决因为pagetype=0导致hub页未识别的情况。

目前hub页识别的步骤：通过indexCheck从链接库中选取新入库的链接，与linkpair数据融合，计算前链的强度谱。然后再从链接库中选取已抓取的、浅层降权链接或非降权链接、高质量索引页或高外链页面或浅层低质量索引页输出一个链接集合。只有属于这个链接集合的，并且其强度谱也符合要求，才会被indexCheck输出到hubdig进行进一步计算。这第二步选取的链接绝大部分是pagetype=1的索引页。

历次时效性分析结果中的HUB页未识别CASE，有29%所以是因为pagetype=0造成的。本次调研就是针对pagetype=0的链接进行hub页挖掘。

通过调研程序过实际数据，在全库pagetype=0的链接中，新获取20万左右的符合hub页条件的链接。针对于之前汇总的11个因为pagetype=0造成的hub页未识别的case，在获取的新hub页中，召回了6个，占比例54.5%。新增hub页的准确率大概在66%左右。

# 背景和目标

汇总2012年时效性链接发现CASE中，得到37个hub页未识别的CASE。其中有11个属于因pagetype=0未识别hub页的情况，占29%。下面列出这些链接，并给出其它属性，考虑通过其它属性捞回。

本次调研主要工作方式：根据indexCheck和hubdig代码，写一个线下调研程序，过实际数据，分析新得到的hub页的各种指标。如数量，准确率多少，对badcase能召回多少等。

# hub页未识别改进

汇总2012年的时效性评估中，hub页未识别的CASE如下表所示，绿色的page\_type=1，红色page\_type=0 。红色的CASE是本次调研处理的对象，占比29%。



## 捞回属性考虑

### 索引页的链接属性

索引页链接的抽查，其属性如下：



### 内容页中符合HUB标准的链接属性

这部分链接就是本次工作需要挖掘出来的链接。他们由于是内容页，所以被indexCheck忽略，不能被识别为hub页。



### 内容页中普通链接的属性

Pagetype=0的，不应该被识别为hub页的内容页链接属性示例



### 捞回属性

从属性上看，hub页的rank\_in rank\_out值通常较大，且link\_depth和dir\_depth较小 。而纯粹的内容页，link\_depth和dir\_depth普遍大于2，且rank\_in和rank\_out较小。所以主要从link\_depth、dir\_depth以及rank\_in、rank\_out角度来考虑捞回，针对于上述CASE的情况，预计可以捞回9个，占81%左右。



## 调研程序

### 调研的indexCheck

在db-spi-pbtest3.vm上面部署一个调研的indexCheck，路径：/home/spider/indexCheck\_psdy。

1：修改链接选取条件，仅仅选取这次预期捞回的pagetype=0的hub页链接。 [脚本fetch\_linkbase.sh、filter\_linkbaseB.sh]

3：用新选取链接，与上次线上indexCheck产生的stat\_link\_value属性产生强度谱数据[gen\_spectrum.sh]

4：对强度谱数据增加几个属性 [gen\_spectrum\_prop.sh]

### 调研的hubdig

在db-spi-pbtest3.vm上面部署一个调研的hubdig，路径：/home/spider/hubdig。

1：下载上一步骤中调研indexCheck程序产生的强度谱数据

2：去除人工配置hub页

3：去除从其它机器下载的博客、新闻、论坛页

4：接下来正常跑hubdig，跑出一份hub页数据

# 调研结果评估

## 捞回比例

针对于之前汇总的11个因为pagetype=0造成的hub页未识别的case，在获取的新hub页中，召回了6个，占比例54.5%。



对于其中5个未召回的，给出原因如下：

www.tudou.com/home/\_10076271 不符合链接选取条件，未召回

blog.sina.com.cn/s/articlelist\_1348303482\_0\_3.html 不符合链接选取条件，未召回

list.pptv.com/sort\_list/4.html 强度谱不够，未召回

list.pptv.com/sort\_list/4-1468.html 强度谱不够，未召回

ent.163.com/special/23rdgma/ 专题页，一段时间产生很多链接，之后没有新链接产生，强度谱不够，未召回

## 识别准确率评估

采用随机抽查30条新产生的链接，在浏览器中看其是否真是hub页。抽查结果如下表，其中20个是符合hub页标准的，所以本次识别的hub页的准确率在66%左右。



## 新增Hub页的对系统的影响

新增hub页的数量为20万个左右，而目前线上hub页总共1000万左右，所以本次挖掘，能够会在原有基础上，新增2%左右。

累计新增hub页的density\_day数据，总量482万次，对目前系统影响不大。

# 后继工作

## rank\_in 和rank\_out的值问题

这两个值随着系统的长时间运行，一直不断的增长，可能在以后需要定期调整。

## 参数确定方法

目前设定的链接选取条件：page\_type=0&&(link\_depth<=2 || dir\_depth<=2) && (rank\_in > 10000 || rank\_out > 2000)。其中的数值是根据抽样，人为推定的，是不是可以通过统计方法更有效确定值？

## 考虑更多的属性

可以继续考虑其它属性捞回hub页的可能性。