Data Center手册

所有数据按网站类别分成：新闻、论坛、微博、博客、搜索结果

* 数据处理过程

分为4步。

Step1. 数据收集

将指定目录下的iget6数据格式的zip文件，按字段tp分文件夹转移到另一个目录。

为了保证制定目录下的文件是完整的，程序获得文件名后，等待10秒才开始读取这个文件，如果指定目录是通过ftp上传上来的，传一个文件时间超过10秒，请不要直接放在指定目录下，可以用一个file\_divert做一个目录拷贝，file\_divert可以设置更长的等待时间。

Step2. 排重

排重标准是，一段时间内，一个字段或多个字段合集的内容为排重依据。

Bbs排重字段：host，topicid，作者，发帖时间，内容

其他数据排重字段：url

Step3. 网民行为分析

以作者url为标识，统计每个人的日发帖数。

如果某人发帖数超过N，不重复的帖子数为m。当日他的总发贴数> N\*m 则这个人是发帖机器人。

Step4. 标签数据/数据转化/浏览数跟帖数拷贝

数据过滤：发帖时间旧，发帖字数少的会被丢弃。

标签数据：

|  |  |
| --- | --- |
| 字段 | 算法 |
| CH 频道 | 来自信息源名字，iget6. CH |
| AT作者类型 | 对于**论坛/微博**的计算方法：  作者URL在过去（robot.useHistory.days）天内，曾经被认定为发帖机器人，则为机器人  读取（calctype.usedays）日的作者url/发帖数，作为集合A。再向前读取(checknew.usedays)日的作者url，作为集合B。B和A数据不重复。如果作者没有在B集合中出现过，则视为新用户。  作者在集合A中，平均日发帖数为P  若 p>=10 为活跃分子  若1=<P<10 为一般参与者  若P<1 为偶尔发帖  作者类型，每日计算，同一个人在不同的时期可能分属不同类型。  搜索结果数据的作者类型统一为：搜索结果  博客的作者类型统一为：博客  新闻的统一为：新闻 |
| LC地域信息 | step4.location\_keywords.txt文件中保存关键词与地域信息的对应关系。  信息中如果含有关键词，则信息会标上地域信息标签。  一个信息可以标多个地域标签。  文件step4.location\_keywords.txt的数据格式为  [关键词]=[地域信息] |
| EM正负面 | step4.emotion\_keywords.2.txt是正负面词文件。  判断逻辑为   * 在包含不重复负面词数量>=6 且 信息长度/300<不重复负面词数量 的条件下，若 正面词数量不足负面词数量的一半 则信息为负面；若 正面词出现总次数 不足负面词出现总次数的一半，则为负面；除以上2种情况，信息为有正有负。 * 在包含不重复正面词数量>=6 且 信息长度/300<不重复正面词数量 的条件下，若 负面词数量不足正面词数量的一半 则信息为正面；若 负面词出现总次数 不足正面词出现总次数的一半，则为正面；除以上2种情况，信息为有正有负。 * 上面都不符合，为无正负 |
| MT信息类型 | 根据iget6.TP字段。  BBS要考虑PN字段区分主贴跟帖 |
| MD媒体类型 | 根据iget6.TP字段。 |
| DL网站级别 | 遵从step4.domain\_level.txt中的定义。若无定义为iget6.DL（信息源的高级设置domain\_level）  step4.domain\_level.txt的格式为  01ny.cn=地方媒体 |
| CC频道分类 | 1.新闻：  遵从step4.news\_channel\_cate.txt中的定义，若无定义为iget6.CC（信息源的高级设置channel\_category）  step4.news\_channel\_cate.txt格式为  [domain]\t[host]\t[category]  2.bbs  遵从step4.channel\_cate.txt中的定义，若无定义为iget6.CC（信息源的高级设置channel\_category）  step4.channel\_cate.txt格式为  [iget6.CH]\t[host]\t[category]  3.blog  若 iget6.CC不为空，为iget6.CC（信息源的高级设置channel\_category）  否则为”博客”  4.微博  若 iget6.CC不为空，为iget6.CC（信息源的高级设置channel\_category）  否则为”微博”  5.搜索  若 iget6.CC不为空，为iget6.CC（信息源的高级设置channel\_category）  否则为”搜索结果” |
| TH论坛话题ID | Iget6.TH @ domain |

* 配置文件
* 数据目录结构



数据目录(data.path)：数据保存的根目录。

分类目录(savefolder)：根目录下一级目录，可以有多个来保存不同的数据。

时间目录：分类目录下，以年月日的形式，当日处理的数据存于当日目录下

压缩文件：保存于时间目录下。

* 总配置文件：iget6\_im4.conf



* Step1 数据收集配置文件：



* Step2配置文件：step2.removedup.bbs.conf; step2.removedup.blog.conf; step2.removedup.mb.conf; step2.removedup.news.conf



remove\_dup.useday=100 排重的时间范围，例子中是100天内排重。

remove\_dup.basis.fieldlist 排重依据字段，这些字段的内容都一致才算重复。

注意：搜索和新闻是一起排重的，其他的bbs，blog，微博都是单独排重。

* Step3 ：

step3.author\_postnum.conf，统计当日作者url和发帖数。



step3.author\_robot.conf 发帖机器人判定



min\_post\_num\_perday=50 判定标准：每天至少50个相同内容的帖子

* Step4 ：

step4.data\_im4.conf





datelimit.begin 这个时间前的信息丢弃

datelimit.delaydays=10 10天前的信息丢弃

output.textlength.limit=10 不重复字数少于10个字的信息丢弃

step4.author\_sorter.conf



step4.copy\_bbsrv.conf 跟帖量浏览量的拷贝



* 统计程序

在目录E:\IGet6\dc\_stat下

* stat1.bat 按字段统计数量。时间范围在.bat中修改。

java -cp IGet6IM4\_fat.jar -Xmx800m -Xms800m com.iget6.stat.DataStat conf\data\_stat.conf yesterday 1

yesterday 表示从昨天开始统计1天内的数据。

或者将昨天换成 20120101等具体日期。

配置文件是data\_stat.conf

如果选定的字段是时间，按天进行汇总。

可以使用fieldfilter过滤。

* cube1.bat 按2个字段cube计算。基本用法同stat1.bat
* rd\_check.bat 重复检查。

配置文件



export.duptime.limit=3 只要要3次以上重复才输出。

输出文件为



其中md5字段。下划线前是排重内容的md5值。下划线后是重复次数。通过搜索下一个相同的md5字段值，可以找到重复的记录

* 注意事项
* 关闭

请在出现sleep。。。等字样时，杀掉程序。在运行过程中，请等待执行完成。

* bbs分组排重的设置办法

当bbs数据量比较大，需要对bbs数据先按域名或者入口分组，组内排重，以提高排重速度。

设置办法如下。

* 配置文件 iget6\_im4.conf

group.bbs.size=32

将bbs分成32个组

* 配置文件 step1.MultiPathCollector.conf

group.data\_type=bbs

bbs.group.size=%group.bbs.size%

bbs.group.field=ho

bbs.group.exception=bbs.tianya.cn

bbs.group.field.bbs.tianya.cn=ch

按ho不同分组，对于bbs.tianya.cn是例外的，他是按ch（频道名，即入口）分组。

如果有多个例外，用,分开，然后写多组bbs.group.field.XXX=字段。如：

bbs.group.exception=bbs.tianya.cn,abc.com

bbs.group.field.bbs.tianya.cn=ch

bbs.group.field.abc.com=ch

* 其他配置文件不需要变动，例子见\datacenter\conf\group\_conf\_demp.zip

如果取消分组，需要将上述配置中

group.bbs.size=32

group.data\_type=bbs

注销掉即可