1. 信息源高级设置部分

BBS.TITLEID.SCOPE=SITE/CATEGORY

主要设置bbs主题的排重范围，SITE 标识信息源内部排重，默认选项。CATEGORY是信息源所在分类一级内排重。主要用于去掉全局置顶贴的下载。

Sleep\_After\_Each\_Downlad.millisecond=0

线程没下载结束后休息毫秒数

reserve\_undownload\_url=true

是否保留未下载的链接，并在下一次下载时，继续下载。默认是false，即所有链接重新分析。设置该选项主要用于短时间完成网站遍历。

default\_cookie=

设置默认的cookie

1. BBS下载模板配置
2. 主题列表页面

类：com.zhongsou.ispider.text.ana. BBSTopicAna\_ExprMatch0

模板说明：

/\*\*

\*模板规则

\* expression=sdgas[\*]dgs[$]dgsdgsdghsd

\* expression.prefix= //定义内容的总起始位置（可忽略）

\* expression.suffix= //定义内容的总结束位置（可忽略）

\* expression.firstUnique=true //表达式第一部分不能出现在任何一个通配串中

\* expression.matchMaxLen=2000 //通配串的最大长度,0代表无限制

\*

\* bbs.topic=sddf<$1$>sdg //[主题]

\* //例如：url=http://new.abx.com/show?id=123&id1=456的主题编号为123,456

\* bbs.topicid=id,id1 //若有<$0$>标记，表示来自0变量，否则来自url

\* bbs.sub=<$2$> //[跟贴数]; bbs.sub=<$2$>,0 表示第0个数字。负数表示倒数第几个数,例如2通配=15/32 则sub＝15

\* bbs.view=<$3$>//[人气数] bbs.view=<$2$>,0 表示第0个数字。负数表示倒数第几个数

\*

\* bbs.createtime=<$4$> //[创建时间]可省略;

\* bbs.createtime.style=dateformat/millisecond/second //三选一，如果不定义为dateformat.

\* bbs.createtime.zone=8 //仅time.style＝millisecond/second时有效。

\* bbs.createtimeformat=yyyy-MM-dd

\* bbs.createtime.invalidChar= //时间字符需去掉的字符（可以不定义）

\* bbs.createtime.autoAdjust=false //时间字符串按格式自动调整，去掉两头多余字符

\* bbs.lastmodified=<$4$> //[最后更新时间];

\* bbs.lastmodified.style=dateformat/millisecond/second //三选一，如果不定义为dateformat.

\* bbs.lastmodified.zone=8 //仅time.style＝millisecond/second时有效。

\* bbs.lastmodifiedformat=yyyy-MM-dd

\* bbs.lastmodified.invalidChar= //时间字符需去掉的字符（可以不定义）

\* bbs.lastmodified.autoAdjust=false //时间字符串按格式自动调整，去掉两头多余字符

\* bbs.url=<$5$>

\* //适用的规则为帖子翻页链接是首页链接＋页码变量

\* bbs.article.numOfPage=15 //每页贴数量(含主题贴) 0为不分页或需要到跟贴页面分析下一页.

\*#方法一

\* bbs.article.pageKey=page //url中页数的变量

\* \*#方法二 <$PageNo$>中PageNo为保留字，不能出现在elements中，可以根据startPageNo，PageStep变化。

\* bbs.article.pageExpr=http://www.qqoo.net/thread.php?fid-<$1$>-search--page-<$PageNo$>.html

\*

\* bbs.article.startPageNo=1 //起始编号 从首页开始

\* bbs.article.PageStep=1 //每页跟贴数量

\*

\*subNum.includeTopicPost =false //跟贴数是否包含主题贴

\*

\* bbs.elite=<$6$> //精华贴标记 bbs.elite=<$6$>,2 表示取前2个字

\* bbs.elite.true=1 //若A6==1则为精华贴

\* bbs.elite.include= //若包含字符串则为精华，bbs.elite.true与bbs.elite.include定义一个即可

\* bbs.intop=<$7$> //固顶贴标记 bbs.intop=<$7$>,2 表示取前2个字

\* bbs.intop.true=1//若A7==1则为固顶贴

\* bbs.intop.include= //若包含字符串则为固定贴，bbs.intop.true与bbs.intop.include定义一个即可

\*

\* 方法1//用于分析不包含在TABLE中的重要链接，如下一页等

\* pagelink.format=http://[\*].com.cn?id=[$]

\* 方法2//为script函数

\* pagelink.script=nextpage([$],[\*])

\* pagelink.script.href=http://sdgsdg.com.cn/servlet?sd=<$0$>&df=<$1$>//<$0$>第0个参数

\* 方法3//模板匹配

\* pagelink.spidermatch=rrrr[$]=[\*]

\* pagelink.spidermatch.href=http://sdgsdg.com.cn/servlet?sd=<$0$>&df=<$1$>//<$0$>第0个通配符的内容

\* //利用父链接拼接出主题的翻页链接

¡Á 方法4

\* pagelink.fromfatherlink.pageKey=page //url中页数的变量

\* pagelink.fromfatherlink.startPageNo=1 //起始编号 第二页开始

\* pagelink.fromfatherlink.PageStep=1 //页码步长

\* 方法5 //处理有下一页的情况

\* pagelink.nextPage.key=下一页

\* 方法6

\* pagelink.PageNoLoop.expr=<a href="[\*]-[\*]-[\*].html">[\*]</a>

\* pagelink.PageNoLoop.ignoreBlankWhenMatch=false

\* pagelink.PageNoLoop.url=<$0$>-<$12$>-<$PageNo$>\_right.html

\* pagelink.PageNoLoop.startPageNo=2

\* pagelink.PageNoLoop.PageStep=1

\*########

\* pagelink.setAnalysePageLink=false //若一次不能分析出所有翻页，则设为true，下一页接着分析,靠下载深度控制翻页数

\* pagelink.maxnum=5 //分析出翻页链接的最大数量

\*pagelink.postMethod=false//翻页链接用POST方法提交

\* linkanalyse=false //不再需要链接分析

\* ignoreBlankWhenMatch=false //匹配时是否忽略空格 false:匹配空格 true:不匹配空格

\*topicInEachPostsPage=false; //主题贴是否出现在每一页的跟贴列表页中

\* //其他论坛栏目入口

\* otherbbs.spidermatch=rrrr[$]=[\*]

\* otherbbs.spidermatch.href=http://sdgsdg.com.cn/servlet?sd=<$0$>&df=<$1$>//<$0$>第0个通配符的内容

\*/

1. 贴子显示页面

类：com.zhongsou.ispider.text.bbspost.BBSPostAna\_EM0

配置说明：

/\*\*

\*模板规则 title,content,time,author,ip,author\_url

\* //title 忽略则使用利用链接的title

\* //ip可忽略

\* expression=sdf[\*]sdfsdf[\*]

\* expression.prefix= //定义内容的总起始位置（可忽略）要求唯一;定义了以后，第一个元素的第一次匹配前缀以次为准

\* expression.suffix= //定义内容的总结束位置（可忽略）不要求唯一，只要出现就可以

\* title=<$0$>

\* content=<$1$>

\* time=<$2$>

\* time.style=dateformat/millisecond/second //三选一，如果不定义为dateformat.

\* time.zone=8 //仅time.style＝millisecond/second时有效。

\* time.format= //时间格式中有大写空格的，转为小写空格

\* time.invalidChar= //定义在时间格式中的无效字符， 仅time.style＝dateformat 有效

\* time.autoAdjust=false //时间字符串按格式自动调整，去掉两头多余字符， 仅time.style＝dateformat 有效

\* time.reserveTag=false //保留tag,当时间出现在tag里面，且配合time.autoAdjust=true使用， 仅time.style＝dateformat 有效

\* author=<$4$>

\* ip=<$5$>

\*author\_url=http://sadfs.com/userid=<$5$>

\*

\* ignoreBlankWhenMatch=false //匹配时是否忽略空格 false:匹配空格 true:不匹配空格

\* expression.firstUnique=true //表达式第一部分不能出现在任何一个通配串中

\* expression.matchMaxLen=2000 //通配串的最大长度,0代表无限制

\*

\* 对于贴数不固定的网站，如tianya，所有帖子页url不能一次算出，只能一页页计算，需要下面配置

\* 方法一，一页一页翻

\* nextpage.expr=<a href=[\*]>下一页</a>

\* nextpage.url=http://abc.com/<$0$>

\*

\* 方法二，页面上变量的数组形式，利用他可以直接翻到上次下载最后一页，然后一页页翻

\* nextPageValue是代指在匹配出的数组里面，下一页的变量值

\* nextpage.array.bodyexpr=

\* nextpage.array.body=<$1$>

\* nextpage.array.separator=,

\* nextpage.array.urlexpr=

\* nextpage.array.url=http://abc.com/<$1$>-<$nextPageValue$>.html

\*/

1. 帖子列表页中，主题贴与跟贴代码不一致。

类：com.zhongsou.ispider.text.bbspost.BBSPostAna\_EM0T

模板：

/\*\*

\*模板规则

\* 继承BBSPostAna\_EM0。增加下面的内容

\* topic.expression=sdf[\*]sdfsdf[\*]

\* topic.expression.firstUnique=true //表达式第一部分不能出现在任何一个通配串中

\* topic.expression.matchMaxLen=2000 //通配串的最大长度,0代表无限制

\* topic.title=<$0$>

\* topic.content=<$1$>

\* topic.time=<$2$>

\* topic.time.style=dateformat/millisecond/second //三选一，如果不定义为dateformat.

\* topic.time.zone=8 //仅time.style＝millisecond/second时有效。

\* topic.time.format= //时间格式中有大写空格的，转为小写空格

\* topic.time.invalidChar= //定义在时间格式中的无效字符， 仅time.style＝dateformat 有效

\* topic.time.autoAdjust=false //时间字符串按格式自动调整，去掉两头多余字符， 仅time.style＝dateformat 有效

\* topic.time.reserveTag=false //保留tag,当时间出现在tag里面，且配合time.autoAdjust=true使用， 仅time.style＝dateformat 有效

\* topic.author=<$4$>

\* topic.ip=<$5$>

\* topic.author\_url=http://sadfs.com/userid=<$5$>

\*/

1. 天涯网站下载配置说明

主题列表页面

bbs.article.numOfPage=0

pagelink.setAnalysePageLink=true

翻页数量靠信息源的下载深度控制，如果下载10页，深度为10+2;

帖子显示页面

设置 nextpage.array。。。