1. 看下原PA的标题如何进行识别，通过增加图片在第一个h\*上方还是下方，来避免采集h\*上方的图片。两种选择：单纯选择第一个；不过第一个h\*的位置属性，和标题相似性，以及上下周围是否有时间等文本。

面包屑上方下方，怎么判断是否是面包屑，通过h2上下是否包含时间，是否包含面包屑

1. 增加一下具有跳转链接的密度块，统计块级别的，主要是a标签的href不是本页面的，分strict和松散的
2. 增加一下pic url name的正常概率值
3. 增加一下图片宽度大于小于某一个阈值，而不是用区间范围
4. 增加display属性的判断，为none的，src为空的，直接过滤（通过设置属性，对这些样例不进行输出）
5. block中超链接的文本密度，而不是周围文本
6. 父结点的解析宽度和高度也算进去
7. 宽度和高度如果是100%，不用算值，就当是不存在
8. 专门针对头像附近的图片，被误识别，进行处理（这个暂时不进行处理，对block的平滑策略修改后，看下效果，再尝试进行处理。）
9. 平滑处理，如果只在block之间进行处理，如何（这个最后进行实验）
10. 只在block中，对于路径一样的图片进行平滑，<http://www/xxx1/xxx2/xxx3/1.html，只比较xxx1/xxx2/xxx3>
11. 验证下（有logo字样，不过是大写的，比较时，是否都转换成小写了）

13、pic url是否含有潜在的时间字符串

14、文件名中含有\_，大写字母的数量（这个特征比较弱）

15、增加pic在块中的位置。这个是用来处理页面中分块逻辑有问题时使用

| >（&gt;），>> >>>这些符号需要单独进行处理

1 sohu.com 责任编辑

2 sina.com.cn 我要评论

3 sina.com.cn 网友评论

4 163.com 网友发表了看法

5 qq.com 复制链接

6 qq.com 我要评论

7 qq.com 责任编辑

8 qq.com 已有

9 lady.qq.com 发表评论(

10 ifeng.com 共有评论

11 ifeng.com 凤凰网友参与评论

12 xinhuanet.com 分享到

13 xinmin.cn 新民网编辑

14 che168.com 版权所有。未经许可，不得转载。

15 yesky.com 条网友评论

1 sohu.com 责任编辑

2 sina.com.cn 我要评论

3 sina.com.cn 网友评论

4 163.com 网友发表了看法

5 qq.com 复制链接

6 qq.com 我要评论

7 qq.com 责任编辑

8 autos.cn.yahoo.com 分享到：

9 sports.qq.com 分享到：

10 lady.qq.com 发表评论(

11 ifeng.com 共有评论

12 ifeng.com 凤凰网友参与评论

13 xinhuanet.com 分享到

14 xinmin.cn 新民网编辑

15 che168.com 版权所有。未经许可，不得转载。

16 yesky.com 条网友评论

17 nipic.com 查看此会员上传的更多图片

18 verycd.com 条资讯　ｒ

19 ontour.com.cn/news/ 更新日期：

"分享到"

"评论排行",

"自动登录",

"新闻报料",

"监督举报",

"【免责声明】",

"［免责声明］",

"免责声明：",

"版权声明：",

"相关声明：",

"著作权声明：",

"版权相关声明：",

"［版权声明］",

"【版权声明】",

"版权与免责声明",

"中证网声明：",

"腾讯声明：",

"版权所有：",

"手机看报，随时掌握天下！",

"手机读报，精彩随身",

"华讯财经推荐一只股价已经横盘调整了近两个月",

"任何商业、非商业站点，未经书面许可不得转载。",

"【党史频道“网友园地”栏目向全国网友征",//http://dangshi.people.com.cn/n/2013/0218/c85037-20515447.html

"版权申明",

"您对其他相关新闻感兴趣，请在这里搜索",

"版权与免责声明\xa1\xa0：",

"版权与免责声明：",

"请注意：",

"请您注意",//http://ent.people.com.cn/n/2013/0225/c1012-20585102-2.html

"发表言论",//http://www.cq.xinhuanet.com/2013-02/25/c\_114792602.htm

"（免责声明：",

"版权及免责声明",

"版权及免责声明：",

"中证网声明：",

"订东北网彩信手机报，移动发",

"核心提示：手机看新闻，请订阅漯河手机报",

"移动用户发送ＨＮＺＢ到",

"浏览过此新闻的网友",

"手机读报，精彩随身",

"中铝网声明：",

"（来源：",

"声明：",//http://sh.sina.com.cn/citylink/jk/t\_xa/2013-02-21/1336168512\_2.html

"新浪独家稿件声明：",//http://news.sina.com.cn/w/2013-02-23/054226335765.shtml

"中国日报网版权说明：",//http://ent.chinadaily.com.cn/dyrihan/2013-02-22/content\_8327106\_2.html

//"责任编辑：",//http://war.163.com/13/0225/07/8OHUA7J700014OMD.html //http://mobile.pconline.com.cn/409/4091958.html

"本文来源：",//http://war.163.com/13/0225/07/8OHUA7J700014OMD.html

"点击进入",//http://leaders.people.com.cn/n/2013/0307/c58278-20714541.html

"了解更多",//http://hn.qq.com/a/20130224/000077.htm

"【背景资料】",//http://leaders.people.com.cn/n/2013/0307/c58278-20714541.html

//"编辑：", //http://www.fxingw.com/syhd/2013-12-13/52685.html

"欢迎订阅",

"新浪乐居专业购房服务",//http://xinxiang.house.sina.com.cn/scan/2013-06-24/14552261352.shtml

"将本文分享到",//http://review.cnfol.com/130506/436,1594,15023779,00.shtml

"中金在线声明",

"您已浏览完所有图片",//http://mil.cnr.cn/gmws/tpxw/201305/t20130503\_512501227.html

"本文全部内容均由用户上传", //http://slides.discovery.tom.com/vw/343175-14.html

"欢迎在新浪微博上关注我们", //http://mobile.yesky.com/245/35331245.shtml

//"微信号：",//http://sports.sohu.com/20130619/n379244136.shtml

"版权声明",

"免责声明",

"来源：",

"时间：",

"作者：",

"编辑：",

"点击：",

"记者：",