1. 将位置信息进行细化，由原先的5个转换成100个。通过统计正文pic在页面的百分比的分布，然后将其划分为3类（正文pic区域，正文pic区域前部，正文pic区域后部）
2. img\_block\_a\_density，统计img所在block中a标签数和body中a标签数的比例。这个当body中a标签数为0时，**需要考虑使用什么值进行处理，**img\_block\_li\_density同样需要处理。
3. var\_surr\_text和img\_density区分度不高。
4. 上一页和下一页的获取时，需要考虑a标签的title/alt属性值等于上一页和下一页。目前PA中有关于上一页和下一页的关键词集合。判断a标签。
5. 对于关键词content，如果页面中存在，而图片没有命中是有问题的，但是如果页面中不存在content关键词，可能就没有问题。
6. 对导航栏的识别，就是距离导航栏的远近。
7. 获取a标签最多的block，然后计算图片距离该block的位置。在上面的为正，在下面为负之类的。
8. 增加位置，简单利用block的位置，而不是tag标签的位置。避免网页顶部有的存在全国地区等有很多a标签的情况。
9. 时间密集度，如果时间密集，可能是因为评论区，不过万一页面只有一个时间，怎么处理。可以手动设置一个区间，类似是否大于5个，这个采用众数，或者平均数，根据训练数据统计

10、推荐关键词的使用：

Img所在Block含有的推荐相关的关键词占页面的比例

距离位置：img所在block距离推荐关键词block的距离（仅计算最近的）

Img元素距离关键关键词的距离。（仅计算最近的）（给一个平均数或者众数，这个数值是统计出来的，暂时先不这样操作，而是额外增加一个特征，表示页面是否存在。）

页面中是否存在推荐相关的关键词。

Img所在block是否存在推荐相关的关键词。（和比例不重复，比例对于不存在时，不能处理）

Img是否在推荐关键词的下方（有多个推荐关键词，怎么办，暂时不考虑这个，因为有距离可以反应一定的情况。）

推荐关键词所在的位置，有html位置和block位置。

11、可能需要额外考虑下，评论距离正文的位置，或者时间密度比较高的block在正文中的block，如果过分向评论倾斜，论坛可能会提取不成功。

**对于分母为0的情况，可以考虑众数/平均数，即分子和分母都增加众数或者平均数**

**可以只观察某一种类别，根据该特征是为了证明是正例还是负例来选择。**

1. 将display为none的情况，直接过滤掉
2. // 统计下爬取网页失败的page url情况

统计下进行相除操作的特征数

利用weka观察现有特征数据分布，优化特征设计，增加类别区分度。

目前对推荐类的图片识别不好，可以增加推推荐关键词的识别，就是所在的模块是否含有表示推荐的关键词。

评论

回复

引用

发表

我有话说

我也来说两句

发布

分享

转发

收藏

举报

跟帖

时间密集度，如果时间密集，可能是因为评论区，不过万一页面只有一个时间，怎么处理。可以手动设置一个区间，类似是否大于5个

我的浏览记录

我的历史记录

**有的表示上一页和下一页的在正文图片上方。**

前一篇/后一篇

上一图集/下一图集

上一组图/下一组图

上一幅图/下一幅图

PrePage/nextPage

首页/尾页

上一页/下一页

上一个/下一个

推荐关键词的使用：

Img所在Block含有的推荐相关的关键词占页面的比例

距离位置：img所在block距离推荐关键词block的距离（仅计算最近的）

Img元素距离关键关键词的距离。（仅计算最近的）（给一个平均数或者众数，这个数值是统计出来的，暂时先不这样操作，而是额外增加一个特征，表示页面是否存在。）

页面中是否存在推荐相关的关键词。

Img所在block是否存在推荐相关的关键词。（和比例不重复，比例对于不存在时，不能处理）

Img是否在推荐关键词的下方（有多个推荐关键词，怎么办，暂时不考虑这个，因为有距离可以反应一定的情况。）

推荐关键词所在的位置，有html位置和block位置。

对于分母为0的情况，可以考虑众数/平均数，即分子和分母都增加众数或者平均数

将display为none的情况，直接过滤掉

// 统计下爬取网页失败的page url情况

**这些关键词，不能简单使用是否在上方或者下方，有的确实存在中间推荐图片的。主要是商品类。**

相关/更多

您可能在找：

您可能感兴趣

更多相关图片

其他作品

相关的作品

推荐

热门图集

24小时热文

本文相关推荐

推荐阅读

相关图片，相关检索

相关商品推荐

热门商品推荐

相关博文

猜你喜欢

人气搜索

排行榜

最新资讯

头条爆料

热点新闻

关注

相似商品推荐

热销

热门关注

猜你喜欢

来自XX组品

其他作品

同风格的图片

热搜词

换一换

大家都爱看

为您推荐

相关作品

精品推荐

精选阅读

延伸阅读

相关图集推荐

美图欣赏

娱乐八卦 / 新闻看点

相似作品

查看更多

猜你喜欢

精华文章

图文推荐

今日导读

评论排行

类似文章

热榜

热门

热图

热文

看过该视频的还喜欢

大家围观的直播

火爆话题

重点关注

点击今日

热评话题

猜你感兴趣

最新文章

都在看

相关内容

你可能还需要这些

更多作品

八卦

聚焦

关注

热评

2015-06-10 09:24

2017.07.07

2015年11月27日 12:37

2010/12/30

19:55

32秒前

2分钟前

2小时前

2天前

2月前

7月7日 11:52

今天 12:43

15-06-26 20:59

将&nbsp;转换成空格

2015-06-10 09:24

2015年11月27日 12:37

统计下推荐词相关的密度

统计下推荐词所在位置信息

目前surr text的方差计算以及图片密度计算，暂时是采用3个text和2个pic

进行相除操作的，即小于1的值

relative\_pos

relative\_block\_pos

img\_block\_a\_density

img\_block\_li\_density

img\_block\_time\_density

img\_path\_digital\_char\_ratio

img\_alt\_html\_title\_similarity，标题为空时，该值为1，需要考虑

img\_width\_height\_ratio

img\_width\_level

img\_height\_level

img\_density

不进行归一化：

img\_body\_a\_cnt

img\_body\_li\_cnt

img\_body\_time\_cnt

img\_path\_digital\_cnt，看下分布

img\_subpath\_digital\_cnt

img\_subpath\_cnt

img\_alt\_title\_chinese\_cnt

// h，p，div等的个数，应该是和全局（body）比较，还是和最大值进行比较？

文本数和长度相关的应该进行平滑