

Web Page Segmentation

FIT5190 Introduction to IT Research Methods

Group 1:

- Yang Yan 2759****
- Xianqiang Gao 2731****

26 May 2016

Southeast University-Monash University Joint Graduate School

Objectives

CONTENTS

- 1** How to segment a web page
- 2** How to identify important blocks

Web page segmentation

新闻 体育 NBA 娱乐 财经 股票 汽车 科技 手机 数码 女人 论坛 视频 旅游 房产 家居 教育 读书 游戏 健康

网易科技 移动互联 网易 > 网易科技 > 移动互联

请输入关键词

首页 移动互联 智能硬件 中概股 通信 IT 滚动 原创 专栏 投稿 专题

《热血传奇》版权纷争 WEMADE涉诉讼致股价大跌

网易科技讯 5月25日消息，今日，就网游《热血传奇》的版权争议，在盛大游戏方面回应后，韩国公司WEMADE再度发声，否认盛大游戏的“独占运营权”，并表示已经向中国法院提起诉讼。据了解，这是盛大游戏与WEMADE之间因《热血传奇》第二次爆...[阅读更多]

网易科技报道 2016-05-25 23:03:59

分享 收藏 评论 举报

中国概念股周三早盘涨跌互现 阿里巴巴跌3.7%

Symbol	Last price	Change
NYES	174.90	+0.90 (0.51%)
SINA	48.17	+0.23 (0.47%)
BIDU	59.86	-0.29 (0.48%)
SOHU	176.90	+1.04 (0.59%)
CYDQ	17.86	-0.01 (0.06%)
CTRP	45.01	+0.09 (0.20%)
YONG	27.55	-0.29 (1.03%)
QIHU	73.74	-0.11 (0.14%)
NBA	23.45	+0.11 (0.47%)
WUBA	58.74	+0.46 (0.78%)
QUNR	33.60	+0.38 (1.13%)

网易科技讯 5月25日消息，美股周三高开，在前一日的基礎上繼續上漲。有報告指原油庫存下滑，油價接近每桶50美元。能源與原材料類股票受提振。截至发稿，道瓊斯工業平均指數報17,765.63點，上漲59.58點，漲幅為0.34%。標準普爾500...[阅读更多]

网易科技报道 2016-05-25 21:44:32

分享 收藏 评论 举报

阿里正遭美国SEC调查 涉及双11数据及会计操作



网易科技讯 5月25日消息，据美国媒体报道，阿里巴巴周三透露，美国证券交易委员会（SEC）对其双11购物节的会计核算和各种合并操作是否违反美国证券法进行调查。受此影响，阿里巴巴股票在盘前交易中下跌3%。阿里巴巴称：“今年早些时候，美国证券交...[阅读更多]

热门标签

请输入关键词

华为营收

英特尔

创业板

2014 CES

4G全国资费

Facebook

谷歌

专栏



新闻 体育 NBA 娱乐 财经 股票 汽车 科技 手机 数码 女人 论坛 视频 旅游 房产 家居 教育 读书 游戏 健康 彩票 车险 海淘 理财 酒香

网易科技 移动互联 网易 > 网易科技 > 移动互联

请输入关键词

首页 移动互联 智能硬件 中概股 通信 IT 滚动 原创 专栏 投稿 专题 企业库 数码 手机

《热血传奇》版权纷争 WEMADE涉诉讼致股价大跌

网易科技讯 5月25日消息，今日，就网游《热血传奇》的版权争议，在盛大游戏方面回应后，韩国公司WEMADE再度发声，否认盛大游戏的“独占运营权”，并表示已经向中国法院提起诉讼。据了解，这是盛大游戏与WEMADE之间因《热血传奇》第二次爆...[阅读更多]

网易科技报道 2016-05-25 23:03:59

分享 收藏 评论 举报

中国概念股周三早盘涨跌互现 阿里巴巴跌3.7%

Symbol	Last price	Change
NYES	174.90	+0.90 (0.51%)
SINA	48.17	+0.23 (0.47%)
BIDU	59.86	-0.29 (0.48%)
SOHU	176.90	+1.04 (0.59%)
CYDQ	17.86	-0.01 (0.06%)
CTRP	45.01	+0.09 (0.20%)
YONG	27.55	-0.29 (1.03%)
QIHU	73.74	-0.11 (0.14%)
NBA	23.45	+0.11 (0.47%)
WUBA	58.74	+0.46 (0.78%)
QUNR	33.60	+0.38 (1.13%)

网易科技报道 2016-05-25 21:44:32

分享 收藏 评论 举报

阿里正遭美国SEC调查 涉及双11数据及会计操作



网易科技讯 5月25日消息，据美国媒体报道，阿里巴巴周三透露，美国证券交易委员会（SEC）对其双11购物节的会计核算和各种合并操作是否违反美国证券法进行调查。受此影响，阿里巴巴股票在盘前交易中下跌3%。阿里巴巴称：“今年早些时候，美国证券交...[阅读更多]

热门标签

请输入关键词

华为营收

英特尔

创业板

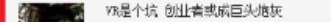
2014 CES

4G全国资费

Facebook

谷歌

专栏



Web page segmentation

```
<html id="ne_wrap">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=gbk">
<meta name="model_url" content="http://tech.163.com/special/0009rt/tech_hlw.html" />
<title>移动互联网科技频道</title>
<script>if((/_touchall=1/.test(location.search))||!/auto|house|home|bbs|blog/.test(location.host))&&!/_dy.163.com/_v2/.test(location.href)&&!
(document.documentElement&&document.documentElement.getAttribute("phone"))&&/163.com/.test(location.host)&&!/pc=1/.test(location.search)&&/android.*?
mobile|ipod|blackberry|bb\d+|phone/i.test(navigator.userAgent))document.write("<meta name='viewport' content='width=device-width,initial-scale=1,maximum-scale=1,user-scalable=no'></div
style='position:fixed;width:100%;height:100%;background:#fff'><div style='position:absolute;top:50%;left:0;width:100%;height:40px;margin-top:-40px;text-
align:center;background:url(http://img1.cache.netease.com/utf8/endpage/image/loading.gif) no-repeat top center;padding-top:40px;color:#666'>页面加载中 ...</div></div><script
src='http://img1.cache.netease.com/f2e/system/touchall/collect/foot/MP300CSqjUId.js'"+' defer><'+'/script><plaintext style='display:none'>');</script>
<link href='http://img2.cache.netease.com/f2e/tech/common/css/tech.2m6QL3zzaxem.19.css' rel='stylesheet' type='text/css' />
<script src='http://img1.cache.netease.com/cnews/js/ntes_jslib_l.x.js' type='text/javascript' charset='gb2312'></script>
<script src='http://img1.cache.netease.com/f2e/lib/js/ne.js'></script>
<meta http-equiv="Expires" content="0" />
<meta name="keywords" content="网易科技,通信新闻" />
<meta name="description" content="网易科技,通信新闻" />
<meta name="robots" content="index, follow" />
<meta name="googlebot" content="index, follow" />
<link rel="stylesheet" href='http://img1.cache.netease.com/f2e/tech/article_list/article_list.670847.css' />
<base target="_blank" />
</head>
<body>
<div class="article_list_wrap" id="article_list_wrap">
  <div class="ntes_nav_wrap">
    <div class="ntes-nav" id="js_N_nav">
      <div class="ntes-nav-main clearfix">
        <div class="c-fl">
          <div class="ntes-nav-select ntes-nav-select-wide ntes-nav-app js_N_navSelect c-fl" tabindex="0">
            <a href="http://www.163.com/#f=topnav" class="ntes-nav-select-title ntes-nav-entry-bgblack" data-module-name="n_topnavapp">应用<em class="ntes-nav-select-arr"></em></a>
            <div class="ntes-nav-select-pop">
              <ul class="ntes-nav-select-list clearfix">
                <li data-module-name="n_topnavapplist_t_0">
                  <a href="http://m.163.com/newsapp/#f=topnav"><span><em class="ntes-nav-app-newsapp">网易新闻</em></span></a>
                </li>
                <li data-module-name="n_topnavapplist_t_1">
                  <a href="http://music.163.com/#f=topnav"><span><em class="ntes-nav-app-msc">网易云音乐</em></span></a>
                </li>
                <li data-module-name="n_topnavapplist_t_2">
                  <a href="http://yuedu.163.com/#f=topnav"><span><em class="ntes-nav-app-yuedu">网易云阅读</em></span></a>
                </li>
              </ul>
            </div>
          </div>
        </div>
      </div>
    </div>
  </div>
</div>
```


Vision-based Page Segmentation

- HTML DOM tree
- Visual features
- Semantic blocks
- Heuristic rules

Our proposed tool



Important blocks

VIPS Tool

File Edit Help

Total: 9 Current: 0 Remain: 8

URL: http://tech.163.com/internet/

Save Up Down

▼ VIPSPage

- LayoutNode ID:1
 - LayoutNode ID:1-1
 - LayoutNode ID:1-2
 - LayoutNode ID:1-2-1
 - LayoutNode ID:1-2-1-1
 - LayoutNode ID:1-2-1-2
 - LayoutNode ID:1-2-1-2-1
 - LayoutNode ID:1-2-1-2-1-1
 - LayoutNode ID:1-2-1-2-1-2
 - LayoutNode ID:1-2-1-2-2

应用 网易首页 登录 注册免费邮箱 网易考拉 请输入关键词

网易科技 移动互联 网易 > 网易科技 > 移动互联

首页 移动互联 智能硬件 中概股 通信 IT 滚动 原创 专栏 投稿 专题 企业库

《热血传奇》版权纷争 WEMADE涉诉讼致股价大跌

网易科技讯 5月25日消息，今日，就网游《热血传奇》的版权争议，在盛大游戏方面回应后，韩国公司WEMADE再度发声，否认盛大游戏的“独占运营权”，并表示已经向中国法院提起诉讼。据了解，这是盛大游戏与WEMADE之间因《热血传奇》第二次爆...[\[阅读更多\]](#)

网易科技报道 2016-05-25 23:03:59

中国概念股周三早盘涨跌互现 阿里巴巴跌3.7%

Symbol	Last price	Change
NTES	174.90	+0.90 (0.51%)
SINA	48.17	+0.29 (0.59%)
SOHU	39.86	-0.03 (-0.08%)
BIDU	176.95	+1.04 (0.59%)
CYOU	17.86	-0.01 (-0.06%)
CTRP	46.01	+0.59 (1.30%)
YOKU	27.53	-0.01 (-0.02%)
QIHU	73.74	-0.10 (-0.14%)
WB	23.45	+0.11 (0.47%)
WUBA	50.74	+0.46 (0.91%)
QUNR	33.60	+0.36 (1.08%)

网易科技讯 5月25日消息，美股周三高开，在前一日的基礎上继续上涨。有报告指原油库存下滑，油价接近每桶50美元。能源与原材料类股票受提振。截至发稿，道琼斯工业平均指数报17,765.63点，上涨59.58点，涨幅为0.34%。标准普尔500...[\[阅读更多\]](#)

网易科技报道 2016-05-25 21:44:32

热门标签

请输入关键词

华为营收 打车APP

英特尔 五道口沙龙

创业汇 成人用品电

2014 CES 诺基亚安

红米二代 4G全国资

特斯拉 Facebook

智能家居 谷歌

Key Value

BgColor	transparent
ContainImg	11
ContainP	0
ContainTable	false
Content	《热血传奇》版权纷争...
DOMIdNum	20
DoC	6
FontSize	20
FontWeight	400
FrameSourceIndex	0
ID	1-2-1-2-1-1
IsImg	false
LinkTextLen	574
ObjectRectHeight	5370
ObjectRectLeft	16
ObjectRectTop	329
ObjectRectWidth	620
SRC	<UL class=newsList ...
SourceIndex	386
TextLen	3828
order	8

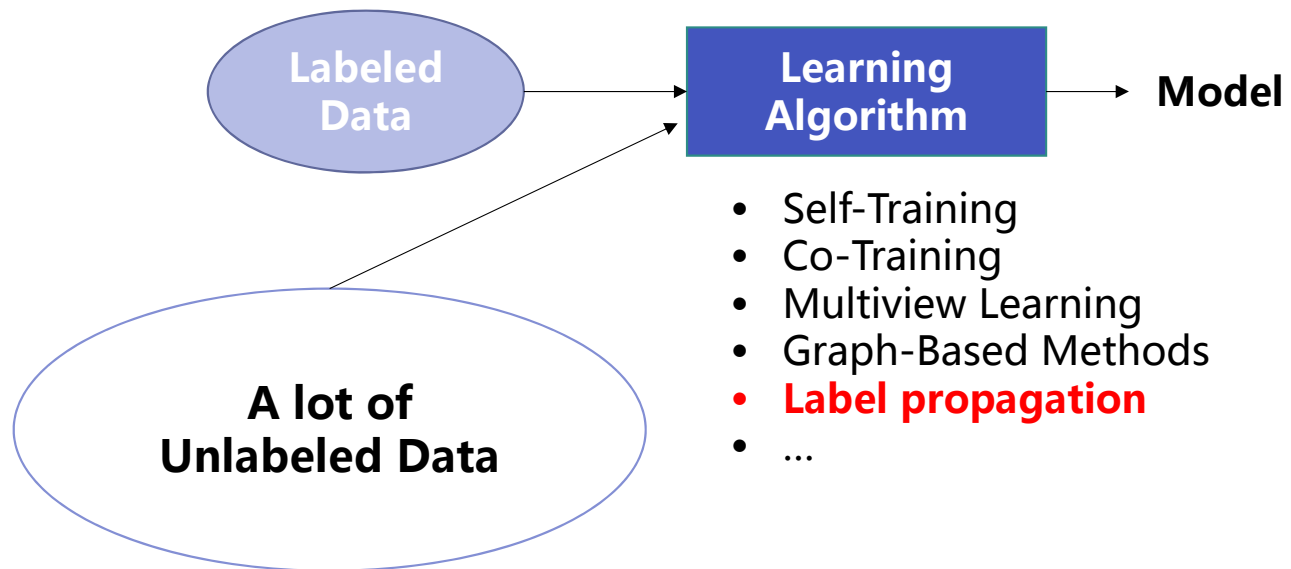
Literature review

- **VIPS + SVM/neural network**
- **VIPS + heuristic rules**

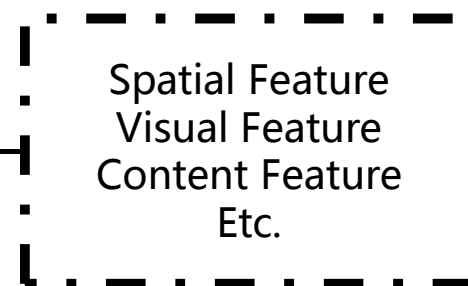
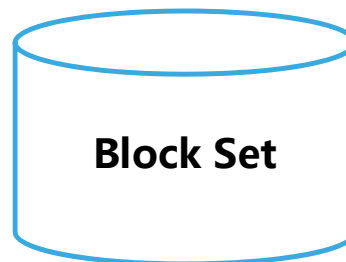
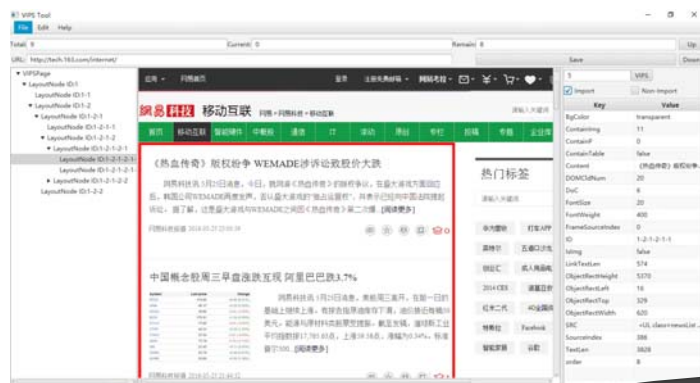
Identifying important blocks

- Semi-supervised learning model

The cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is relatively inexpensive

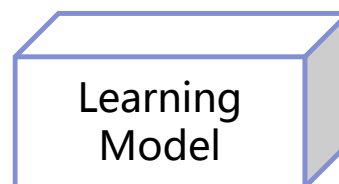


Processing



Train Set
(labeled Block + unlabeled Block)

Test Set
(unlabeled Block)



Block is important or not

Conclusion

VIPS
+
Semi-supervised
learning model

The background is a solid blue gradient. On the left side, there is a white curved shape with a fine halftone dot pattern. On the right side, there is another white curved shape, also with a halftone dot pattern. The text "THANK YOU!" is centered in the middle of the blue area.

THANK YOU!