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 Schoo

Objectives

CONTENTS

1 How to segment a web page

2 How to identify important blocks

Web page segmentation

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_网易科技频道</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.2nuQL3zzaxem.19.css" rel="stylesheet" type="text/css" />
<script src="http://img1.cache.netease.com/cnews/js/ntes_jslib_1.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>
```

Vision-based Page Segmentation

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

Our proposed tool

Important blocks

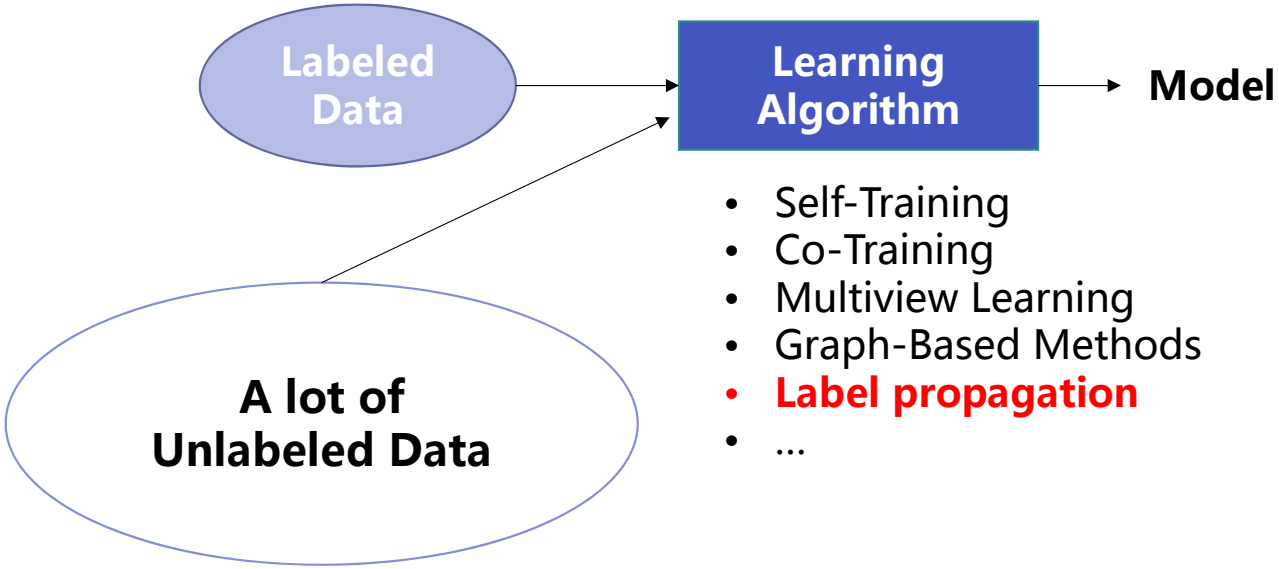
Literature review

- **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

Block Set

Spatial Feature
Visual Feature
Content Feature
Etc.

Train Set
(labeled Block +unlabeled Block)

Test Set
(ublabeled Block)

Learning Model

Block is important or not

Conclusion

VIPS

+

Semi-supervised learning model

THANK YOU!