FIT5190 Assignment 3 Presentation

Web page classification based on content extraction

Group: No. 5
Teammates: Hu Wanling 2819****
Zheng Xuan 2819****

Contents

1 Introduction

2 Content Extraction

3 Conclusion

1

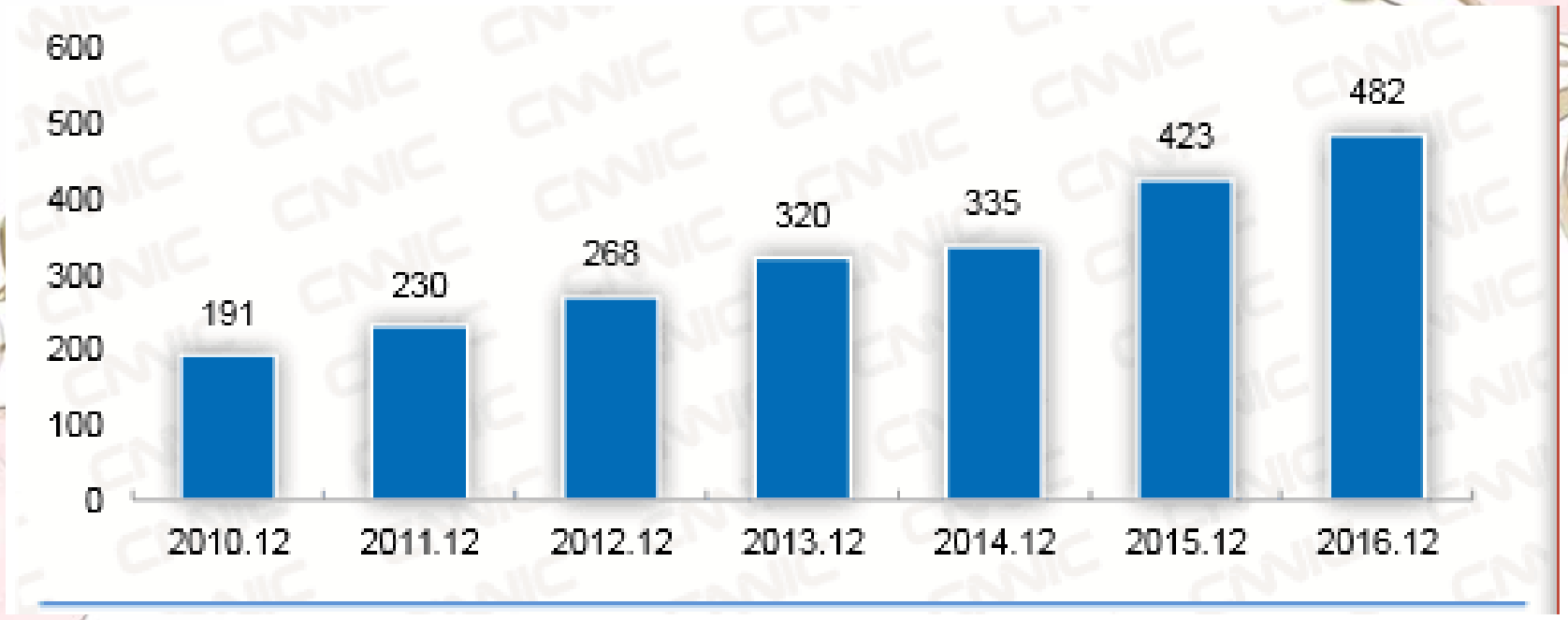
Introduction

(Hu Wanling)

Data from 39th report of CNNIC

The number of Chinese websites

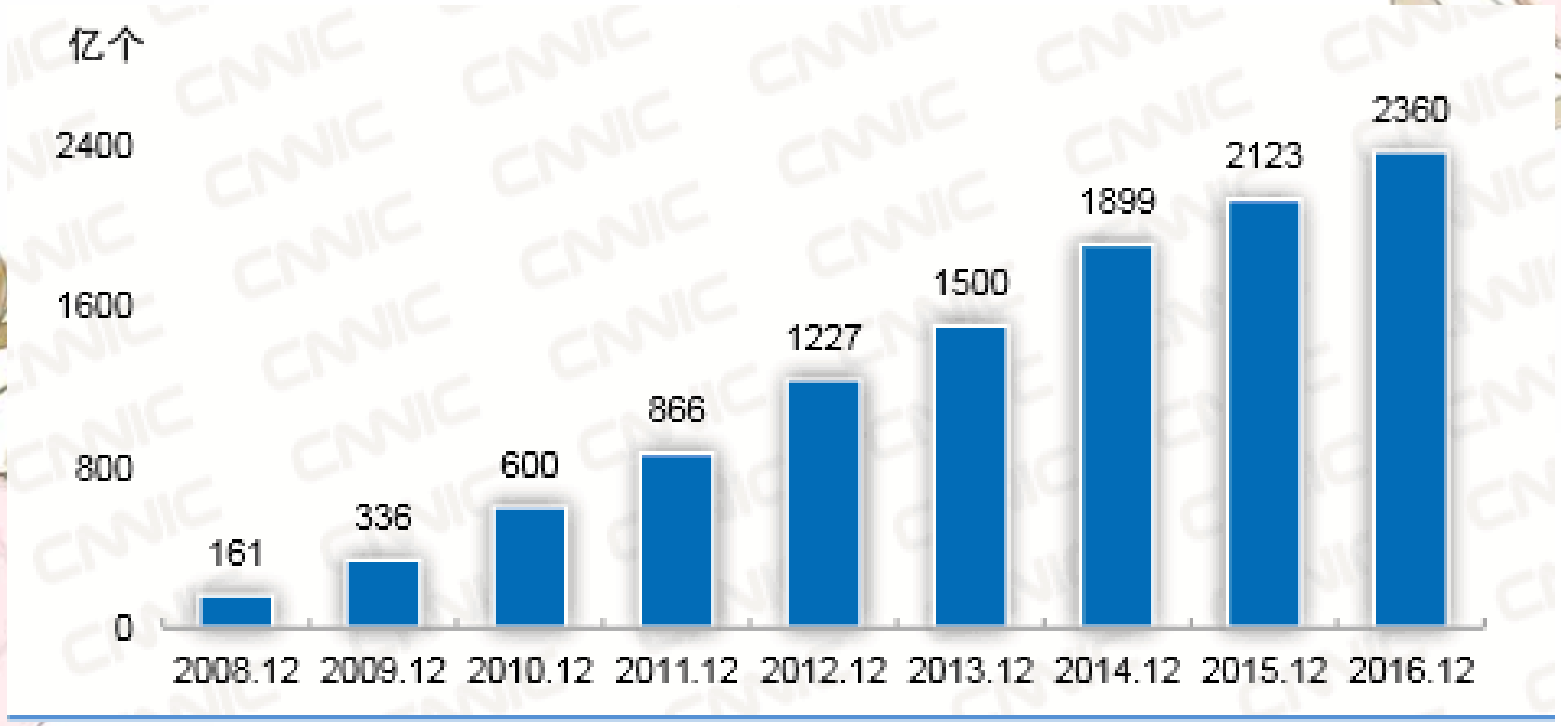
Unit: Ten thousand



By 2016 December, the number of Chinese websites is 4.82 millions, and the annual growth rate of 14.1 per cent.

Data from 39th report of CNNIC

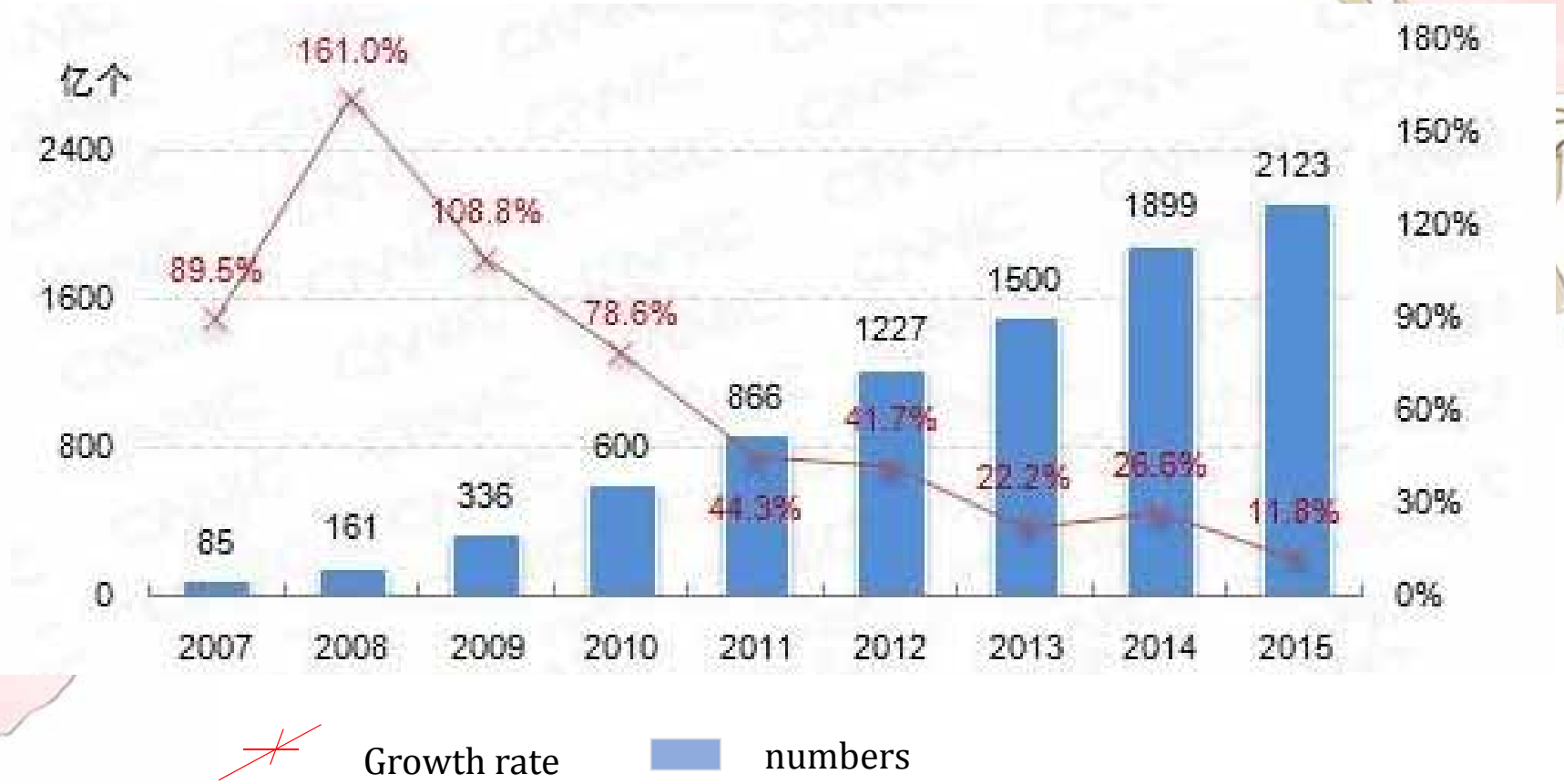
The number of Chinese web pages



By 2016 December, the number of Chinese web pages is 236 billions, and the annual growth rate of 11.2 per cent.

Data from 39th report of CNNIC

The growth rate of Chinese web pages



Data from 39th report of CNNIC

The Statistics table

表 4　中国网页数

	单位	2015 年	2016 年	增长率
网页总数	个	212,296,223,670	235,997,583,579	11.2%
静态网页	个	131,447,834,396	176,083,292,929	34.0%
	占网页总数比例	61.9%	74.6%	20.5%
动态网页	个	80,848,389,274	59,914,290,650	-25.9%
	占网页总数比例	38.1%	25.4%	-33.3%
网页长度（总字节数）	KB	14,815,932,917,365	13,539,845,117,041	-8.6%
平均每个网站的网页数	个	50,197	48,922	-2.5%
平均每个网页的字节	KB	70	57	-18.6%

The number of static web pages is 176.1 billions and account for 74.6% of the total.
The number of dynamic web pages is 59.9 billions and account for 25.4% of the total

classification

Information age

Massive data

An important way to deal with massive data is to categorize them.
The page classification can be described as categorizing pages based on the information carried by the page.

Webpage classification

- The page save in the corresponding database according to the page type, it is good for user to quickly and accurately find their aim pages.
- Improve the recall and precision.
- Improve the efficiency of search engines.
- Key technology of network security management

Strategies

1、Manual classification

- Classify web pages manually by experts
- Example: www.yahoo.com
- Accuracy is high, but efficiency is low. cannot meet the needs.

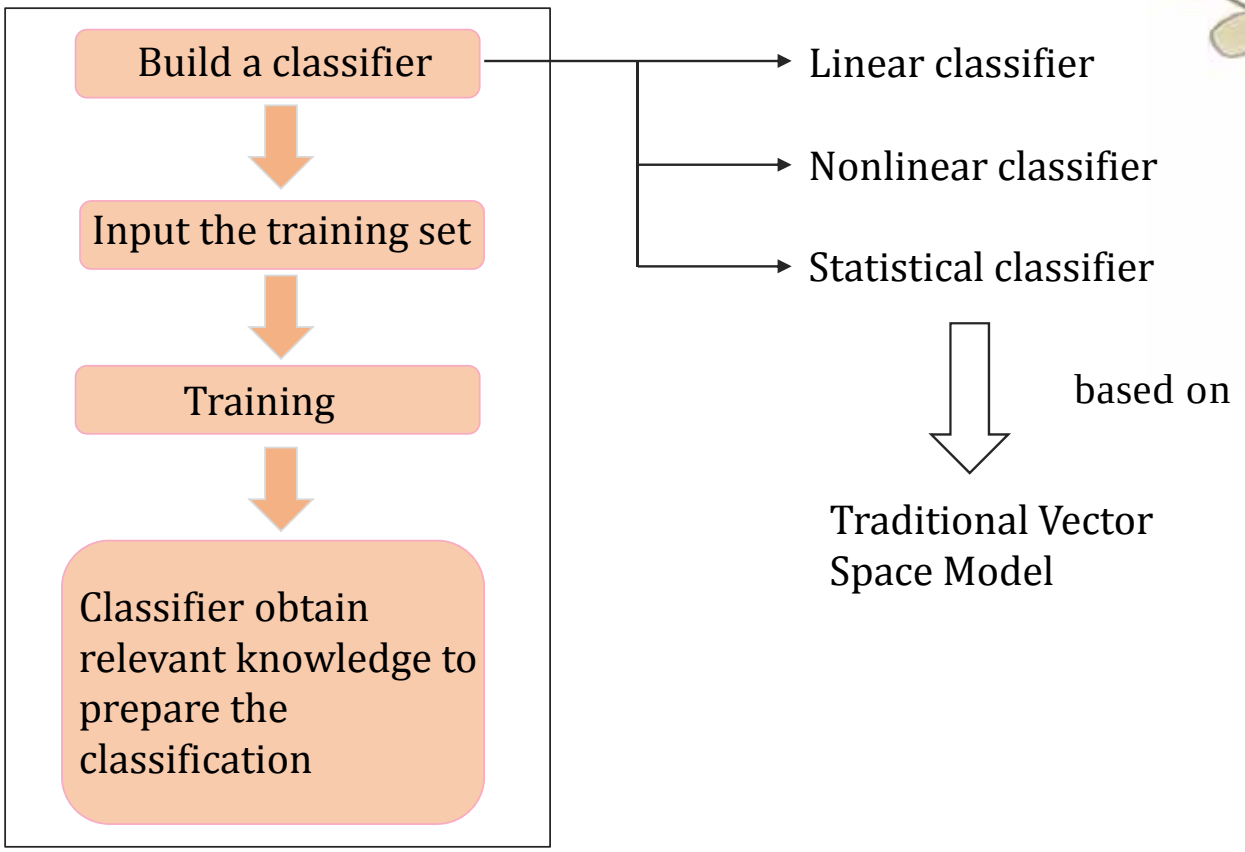
Algorithms based on statistics and machine learning

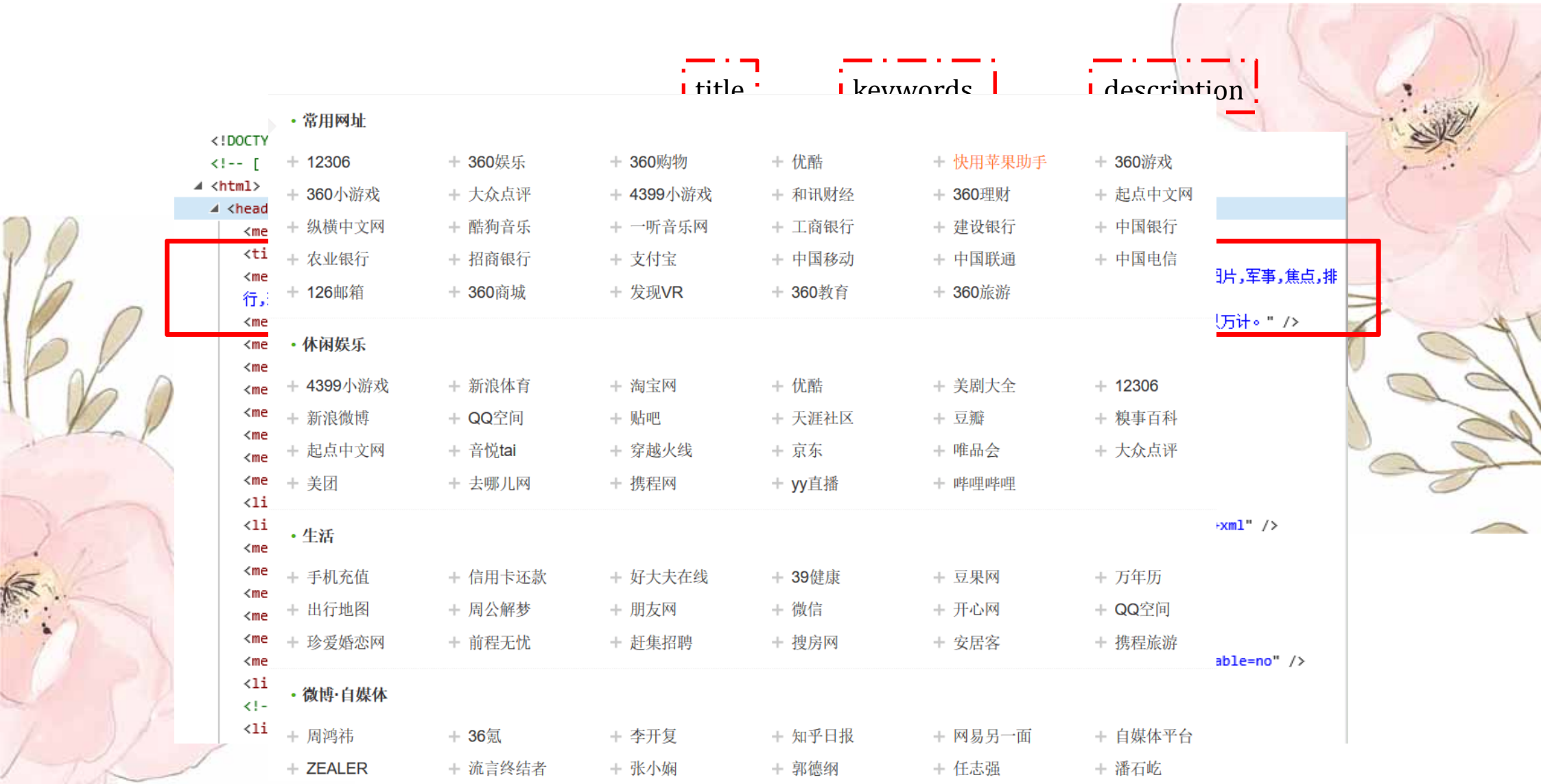
Tab.1 The advantages and disadvantages compared five classification algorithm

分类算法	KNN	NB	ANN	ID3	SVM
优点	算法简单,性能良好	速度快,适于高维数据	自学习自适应能力强	算法简单,学习能力强	非线性,适于高维数据
缺点	实时性弱,计算开销大	条件假设,准确性低	计算量大,计算时间长	对噪声很敏感	不适合大规模训练样本

Strategies

2、Automatic classification




title
keywords
description
常用网址
12306
360娱乐
360购物
优酷
快用苹果助手
360游戏
360小游戏
大众点评
4399小游戏
和讯财经
360理财
起点中文网
纵横中文网
酷狗音乐
一听音乐网
工商银行
建设银行
中国银行
农业银行
招商银行
支付宝
中国移动
中国联通
中国电信
126邮箱
360商城
发现VR
360教育
360旅游
休闲娱乐
4399小游戏
新浪体育
淘宝网
优酷
美剧大全
12306
新浪微博
QQ空间
贴吧
天涯社区
豆瓣
糗事百科
起点中文网
音悦tai
穿越火线
京东
唯品会
大众点评
美团
去哪儿网
携程网
yy直播
哔哩哔哩
生活
手机充值
信用卡还款
好大夫在线
39健康
豆果网
万年历
出行地图
周公解梦
朋友网
微信
开心网
QQ空间
珍爱婚恋网
前程无忧
赶集招聘
搜房网
安居客
携程旅游
微博·自媒体
周鸿祎
36氪
李开复
知乎日报
网易另一面
自媒体平台
ZEALER
流言终结者
张小娴
郭德纲
任志强
潘石屹

2

Content extraction

(Zheng Xuan)

3 Content extraction

Topic specific web pages

3-1 Pre-processing

Pre

HTML label

<title> </title>

<script> </script>
<noscrip> </noscript>

<style> </style>

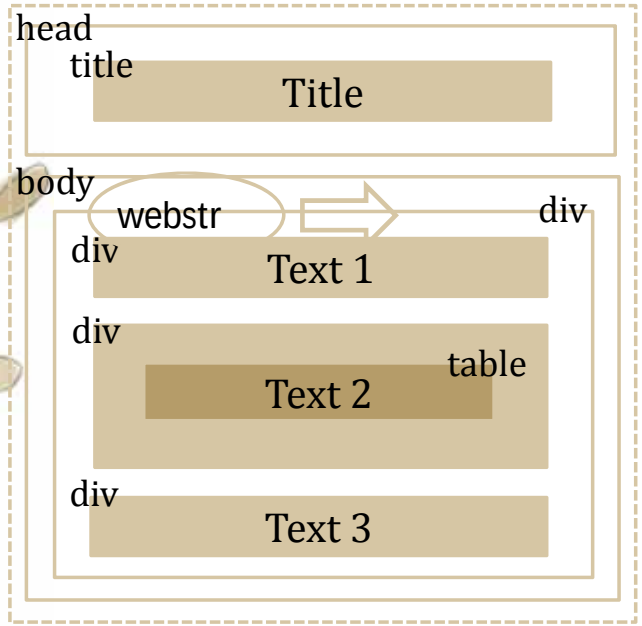
<!--***-->

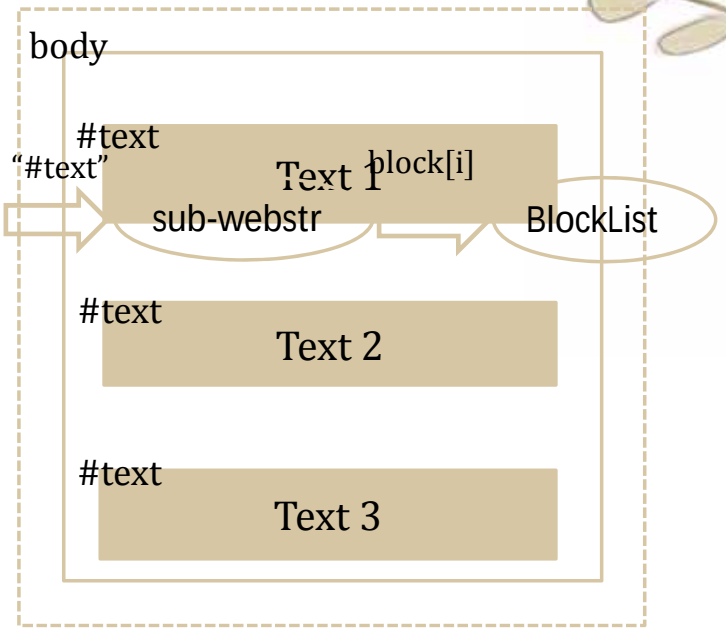
```
<!DOCTYPE html>
<!-- [ published at 2017-05-24 21:21:37 ] -->
<html>
  <head>
    <meta http-equiv="Content-type" content="text/html; charset=utf-8" />
    <title>新闻中心首页_新浪网</title>
    <meta name="keywords" content="新闻,时事,时政,国际,国内,社会,法治,聚焦,评论,文化,教育,新视点,深度,网评,专题,环球,传播,论坛,图片,军事,焦点,排行,环保,校园,法治,奇闻,真情" />
    <meta name="description" content="新浪网新闻中心是新浪网最重要的频道之一，24小时滚动报道国内、国际及社会新闻。每日编发新闻数以万计。" />
    <meta name="robots" content="noarchive" />
    <script src="http://open.weather.sina.com.cn/api/weather/warn_pic/?city=%E8%8B%8F%E5%B7%9E&callback=homeWeatherWarnFun__"></script>
    <script src="http://ip.house.sina.com.cn/sina_sanshou_2010.php" type="text/javascript" charset="gb2312"></script>
    <script src="//news.sina.com.cn/js/694/2012/0830/realtime.js?ver=1.5.1" charset="gb2312"></script>
    <style type="text/css">.tn-title-login-custom{float:l…</style>
    <style type="text/css">.real-time-window .rtw2-lt,.re…</style>
    <style type="text/css">.outlogin_layerbox_bylx, .outl…</style>
    <link id="14956322597562" href="http://i.sso.sina.com.cn/css/outlogin/v1/outlogin_skin_finance.css" rel="stylesheet" type="text/css" media="screen" charset="utf-8" />
    <style type="text/css">.ui-outlogin-shake {-webkit-an…</style>
  </head>
    <link href="http://i0.sinaimg.cn/dy/news3.png" rel="apple-touch-icon" />
    <!-- id="news_web_index_v2015_style" -->
    <link href="http://news.sina.com.cn/js/268/index2015/index.css" rel="stylesheet" type="text/css" />
```

webstr

3-2 Structure division

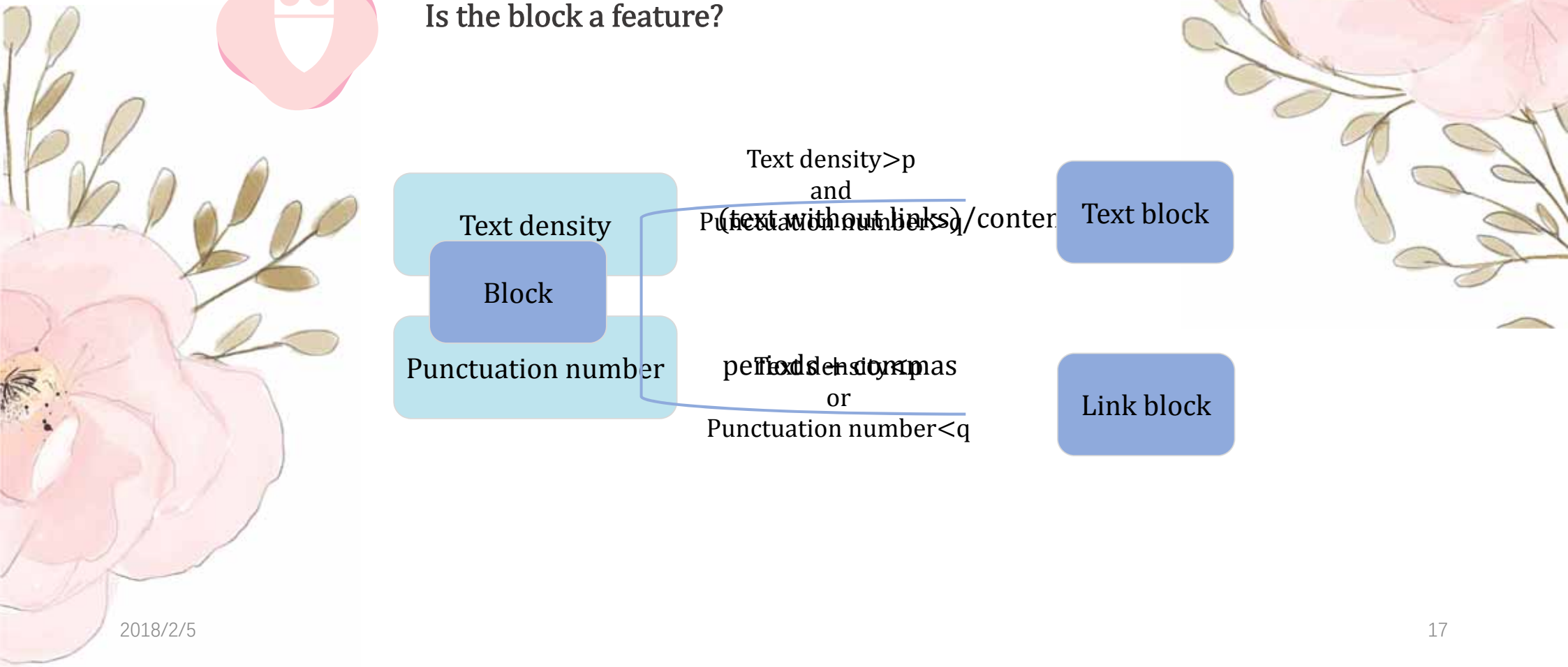
block Contain labels <table> </table>





3-3 Text characteristic

Is the block a feature?



3-4 Content extraction

Filter the “noise” blocks

- Text block
- Link block
- Text block
- Text block

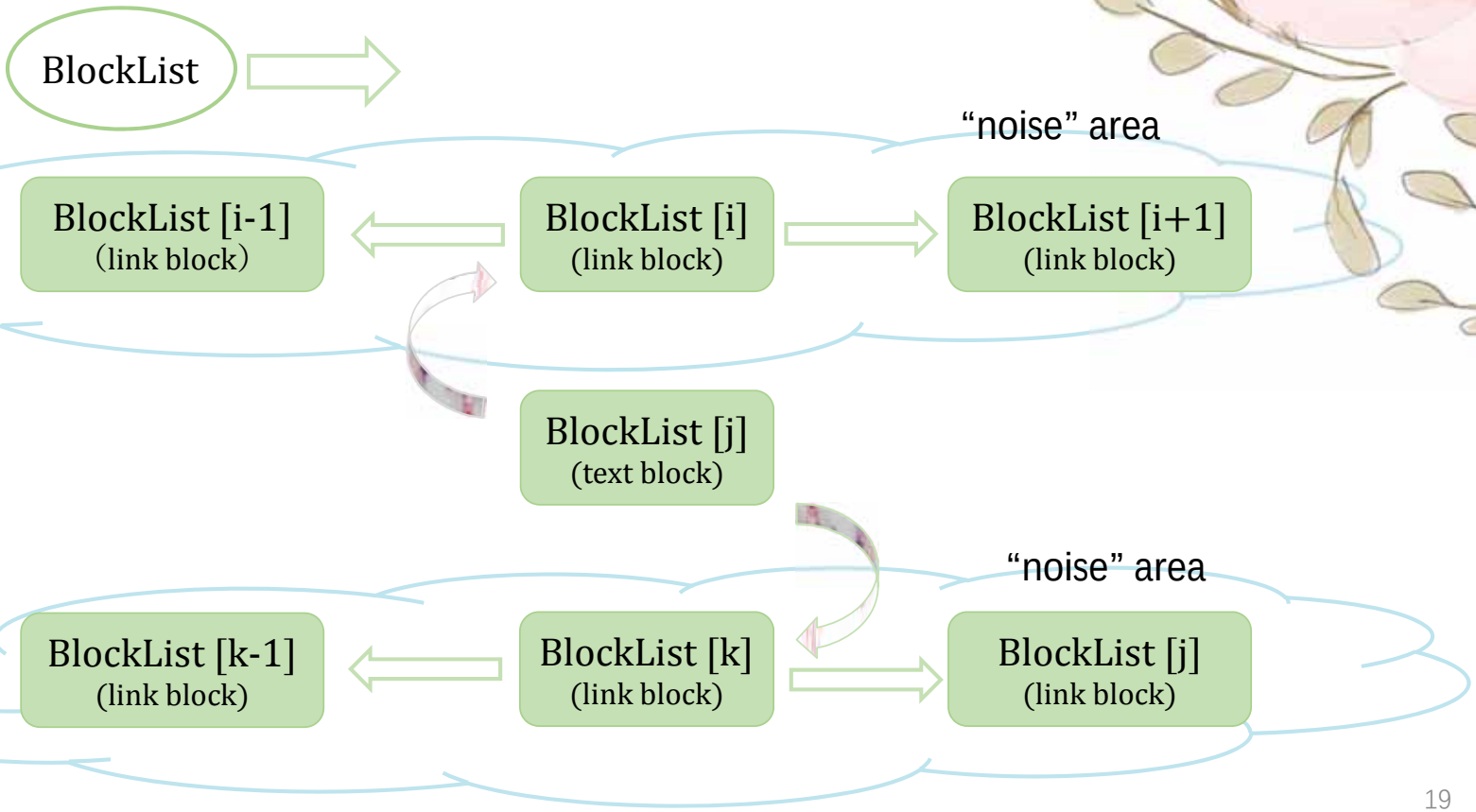
Related to the topic → **Useful!**

- Link block
- Link block
- Link block
- Link block

“noise” area → **Useless!**

3-4 Content extraction

Filter the "noise" blocks



3-4 Content extraction

Filter the “noise” blocks

BlockList

BlockList [i]

BlockList [k]

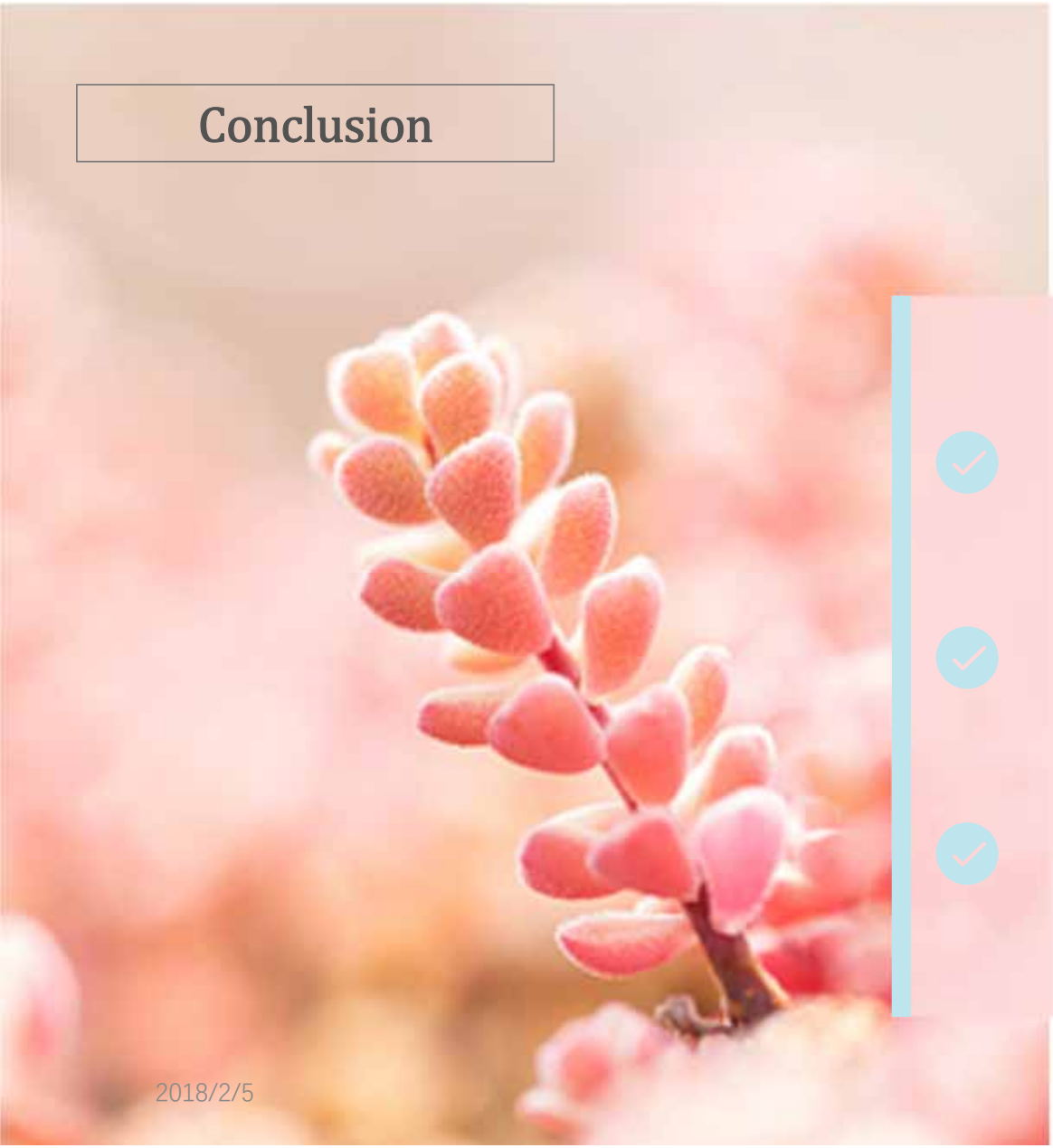
Collect all the blocks

3

Conclusion

(Zheng Xuan)

Conclusion



- Goal: Eliminate useless information
  Web page classification
- Combination: textual features
  structural characteristics
- Further work: handle images and videos

References

Hui Nie, Jinhua Zhang (2012). Content extraction of the theme page under block type layout[J]. *Technical Information*. (1), 31-39

Soucy, P., & Mineau, G. W. (2001). A simple KNN algorithm for text categorization. *IEEE International Conference on Data Mining* , 647-648.

Selamat, A., & Omatu, S. (2004). Web page feature selection and classification using neural networks. *Information Sciences*, 158(1), 69-88.

Thomas, A., & Oommen, B. J. (2013). The fundamental theory of optimal “anti-bayesian” parametric, pattern classification using order statistics criteria. *Pattern Recognition*, 46(1), 376-388.

Zhigang Zhang, Jing Chen, Xiaoming Li. (2004). An HTML web page purification method[J]. *Technical Information*, 23(4), 387-393.

Thanks for listening !
Q&A