法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

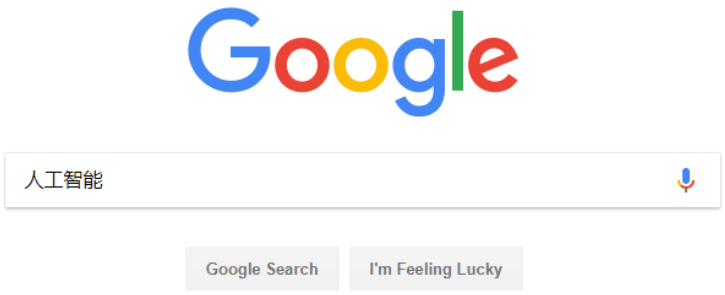
关注 小象学院

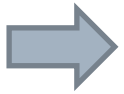
数据挖掘在搜索引擎中的应用

林沐

搜索引擎

一个**搜索事件**:





无论**搜索引擎**如何升级与迭代，无论用**何种手段**，搜索引擎就是在做**三件事**:

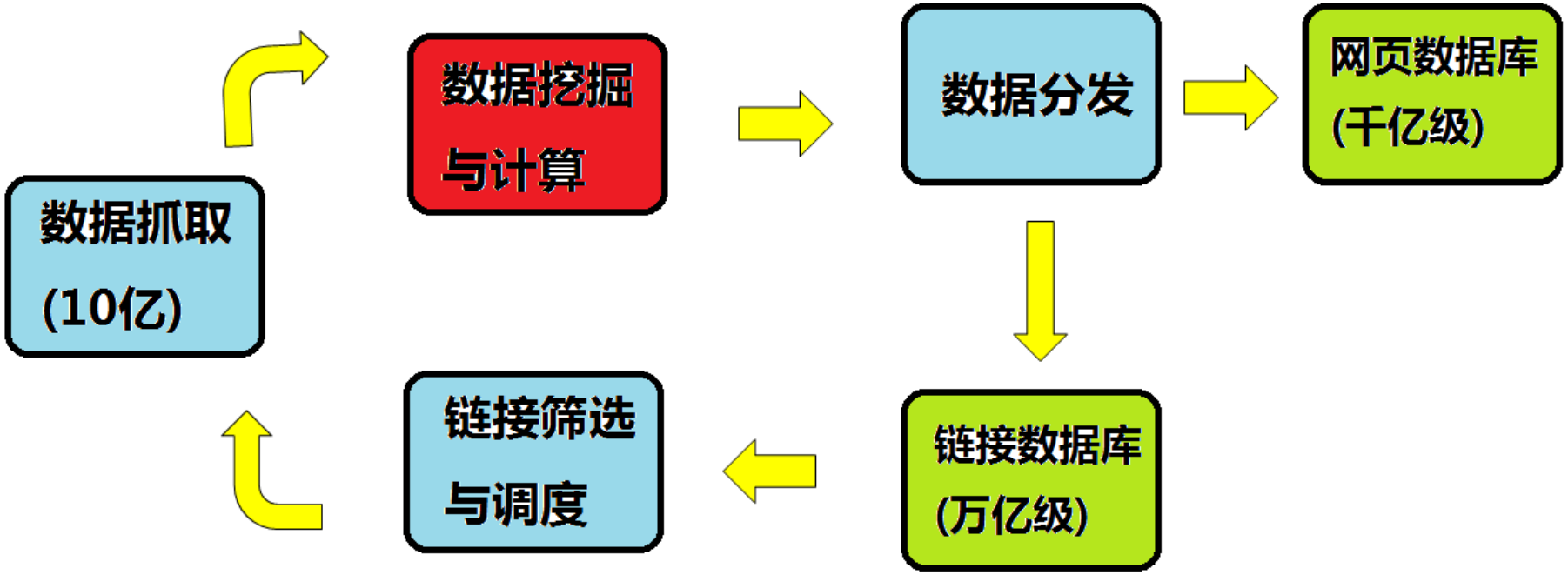
理解用户行为:语音、文字、图片等等输入

收录并理解互联网数据:网页的收录、**理解、分析**

将用户行为与互联网数据**建立联系**，并推荐结果

网页收录与网页数据挖掘

搜索引擎最基础的系统:**spider数据收录与数据挖掘**系统，spider系统需要**足够全**并**足够快**的收录网页数据。一个**全网搜引擎**(google、baidu、sogou)的**spider系统**每日至少**调度抓取**十亿级链接，整个spider数据库会收录存储**千亿级网页**与**万亿级链接**，它是一个**环状系统**。



网页数据挖掘引入

接入**互联网**后，我们会浏览新闻、微博、论坛、博客等等**各式各样**的网页，但网页到底是什么?**观察**如下网页:

思考两个问题:

1.在你观察的**网页中有什么**?

2.从整体来看，该网页**是怎样的网页**，有什么特征?

网页中有什么?


导航栏
首页 | 新闻 | 体育 | 娱乐 | 财经 | 科技 | 视频 | 微博 | 博客 | 读书 | 教育 | 时尚 | 育儿 | 健康 | 星座 | 收藏 | 女性 | 汽车 | 房产 | 更多
注册 | 登录
sina 新浪科技
新浪科技 > 业界 > 硅谷新视野专题 > 正文
网页位置
新闻 Pebble 搜索
标题
Pebble CEO：正考虑如何使产品变得比谷歌好
时间
2014年03月21日 09:42 新浪科技 微博 我有话说(15人参与) 收藏本文
主图
Pebble智能手表
新浪科技讯 北京时间3月21日早间消息，谷歌(1120.15, 5.87, 0.53%)本周发布了面向
头条推荐
【新闻】台媒：菲军队与中国海警船在南海对峙
【体育】英超曼城平阿森纳 曼联4-1逆转 切尔西负
【娱乐】传文章出轨恋姚笛 经纪人独家回应
【女性】女明星Instagram肉搏大战
【育儿】女童误把药片当糖果吃目前仍在抢救(图)
【专栏】孙永杰：HTC见证奇迹的时刻到了？
推荐链接
新闻排行 评论排行 博文排行
1 “人人快递”应用在湖北被叫停：无经营许可证
2 雷军：明年小米手机销量将达1亿台
3 男子买3辆特斯拉只展示不开：30富豪排队想摸
4 360发布4G版随身WiFi 售价299元
5 网络公开课席卷美国:传统大学成互联网受害者？
6 湖南国科微电子与中科院成立联合实验室
7 美“网络部队”将扩编至6000人
8 马云李小加同时出席2014中国IT领袖峰会
9 联想前高管吕岩出任PPTV CEO
10 黑莓BBM月活跃用户数达8500万

整体来看，它是个怎样的网页?

从**整体**来看这是个新闻网页，它**描述了**一个科技产品，该网页**质量较高**，有描述内容的主图、排版清晰，大段文本。该网页**所在站点**质量较高(新浪科技新闻频道)等等。
不同类型的网页还有许多:

网页数据挖掘

计算并提取网页的百余个**网页属性**字段，其中使用**多种技术**方法，如机器学习的分类、聚类、回归、自然语言处理、规则聚合、主题模型等等。最终的目标即**充分的理解网页**，为搜索引擎排序提供**准确的网页属性**。

网页属性提取

网页编码：GB18030

网页语言：中文

内容提取：标题、时间、主图、正文...

网页分类：

新闻、博客、论坛、视频、商品...

高级属性：

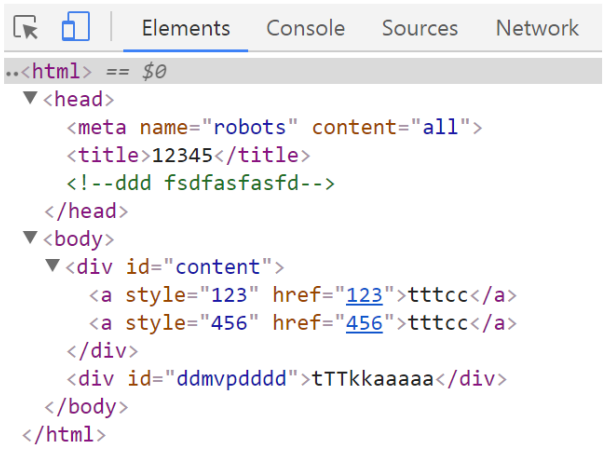
站点价值、网页丰富度、整洁度...

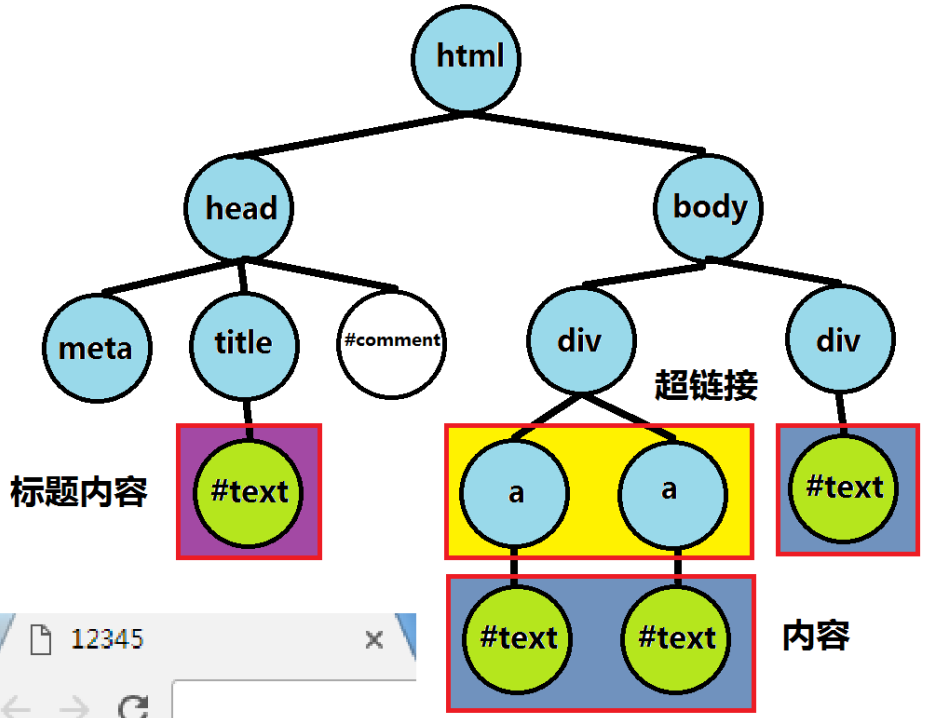
网页数据挖掘基础：网页与HTML解析

网页数据挖掘最为基础的环节就是**网页(HTML)解析**，**网页解析**即将网页源码**HTML字符串**转为计算机可以理解的数据结构，**HTML树**，即dom树。

如下图，蓝色的是**标签节点(Element)**，绿色的是**文本节点(Text)**。标签节点标识了HTML**文档的逻辑结构**，文本节点是HTML**文档的内容**，内容可以**显示**在浏览器中。在HTML文档中，有很多重要的内容用于分析网页属性，例如HTML**文档标题**，HTML文档中的**超链接**，HTML文档中的**文本内容**。

```
1  <html>
2  <head>
3  <meta name="robots" content="all" />
4  <title>12345</title>
5  <!--ddd fsdfasfasfd-->
6  </head>
7  <body>
8  <div id="content">
9  <a style="123" href="123" >tttcc</a>
10 <a style="456" href="456" >tttcc</a>
11 </div>
12 <div id="ddmvpdddd">tTTkkaaaaa</div>
13 </body>
14 </html>
```



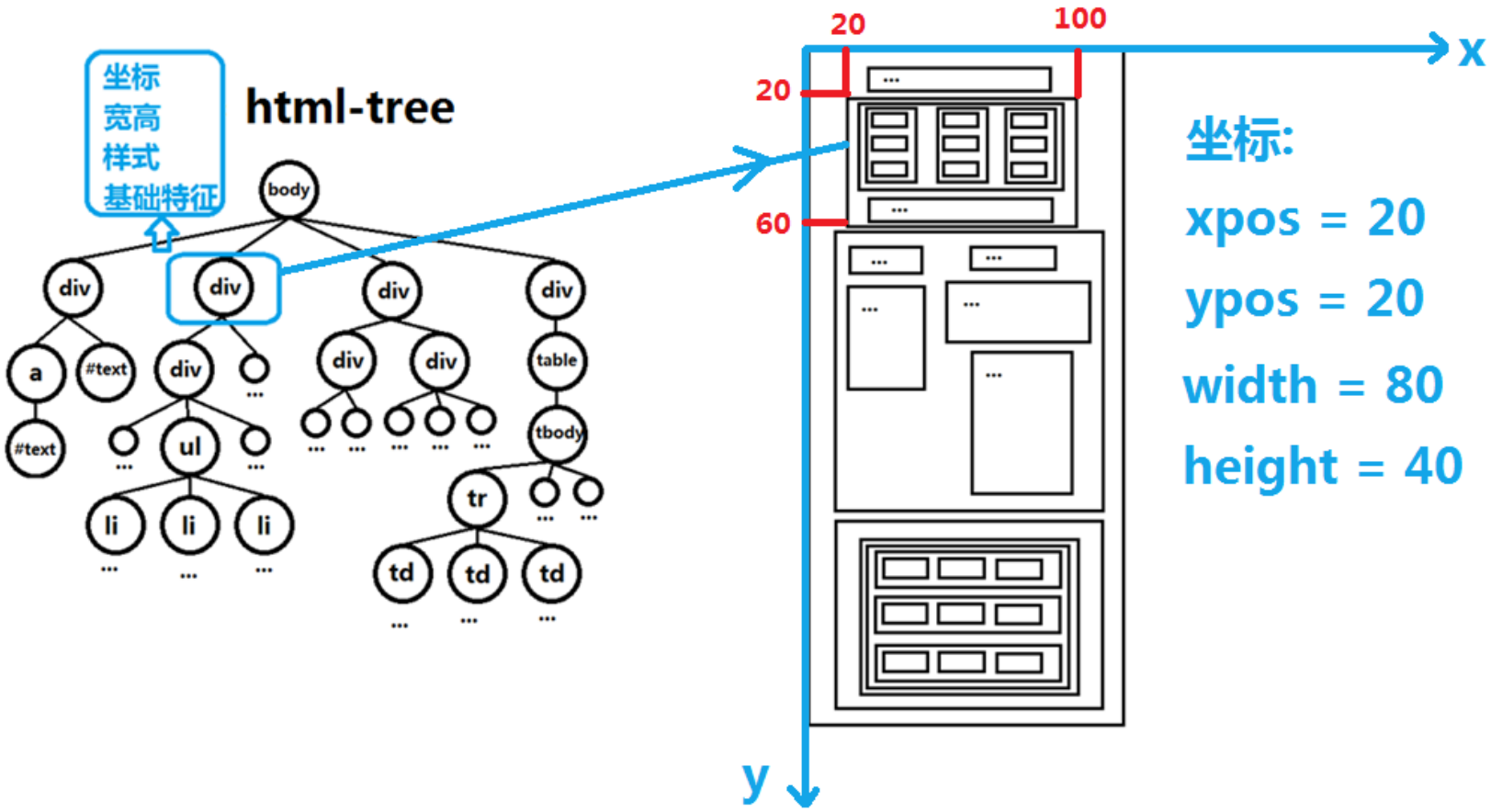


网页的真正理解，平面分析算法

平面分析算法核心思想是将网页树转化为可以理解的**平面数据结构**，从而可以将网页**精确的**划分出区域。它是更**高级的**网页分析数据结构，依赖这个数据结构，可以对网页进行**更深层次的**理解，如主体边框识别、标题、关键内容识别。

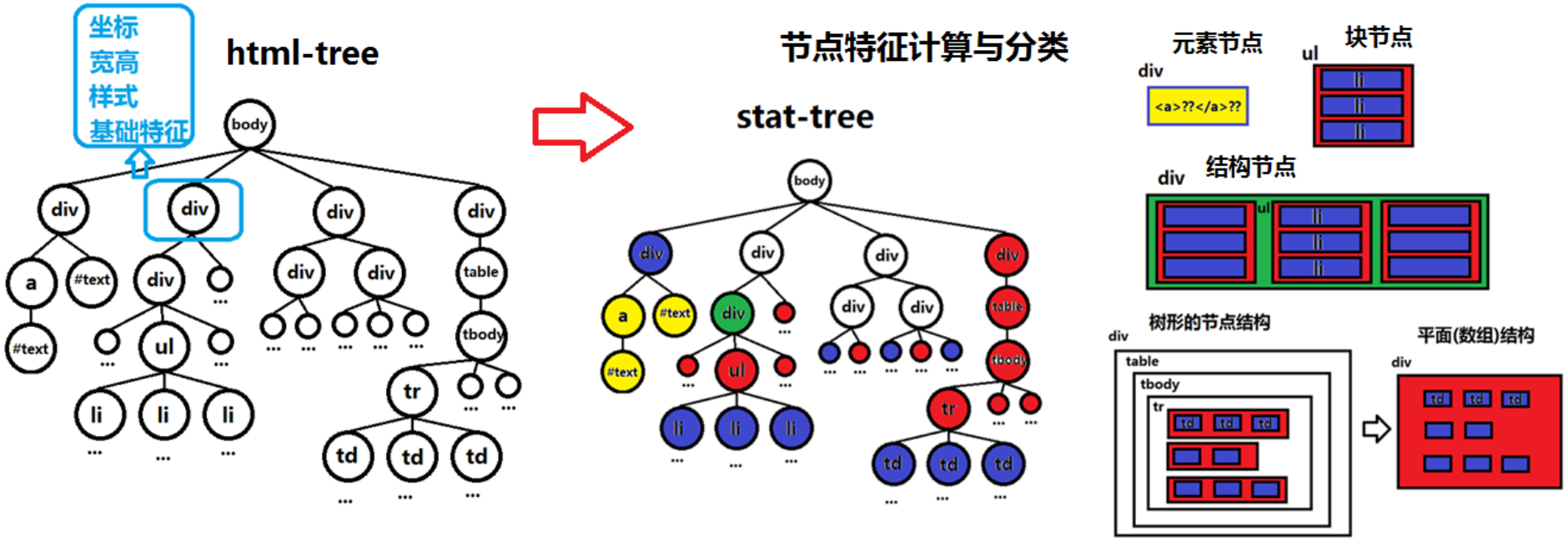
统计树:stat-tree

依赖HTML树构建统计树，**统计树**为HTML解析树增加基础的**统计属性**，例如**坐标、宽高、样式**等基础信息。



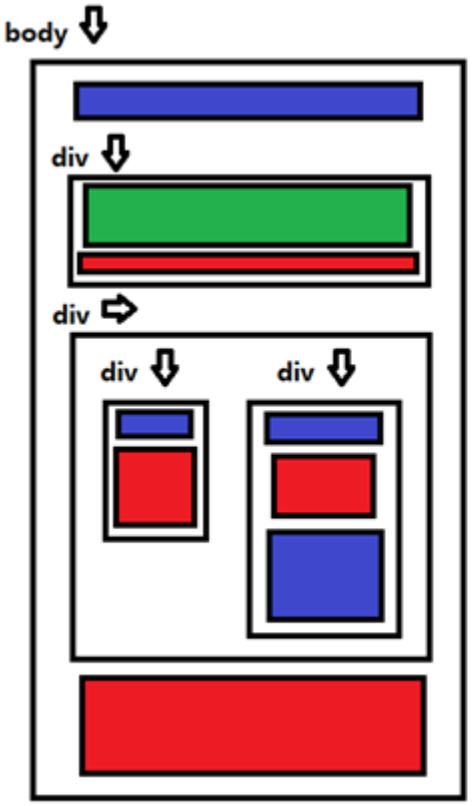
节点特征计算与分类

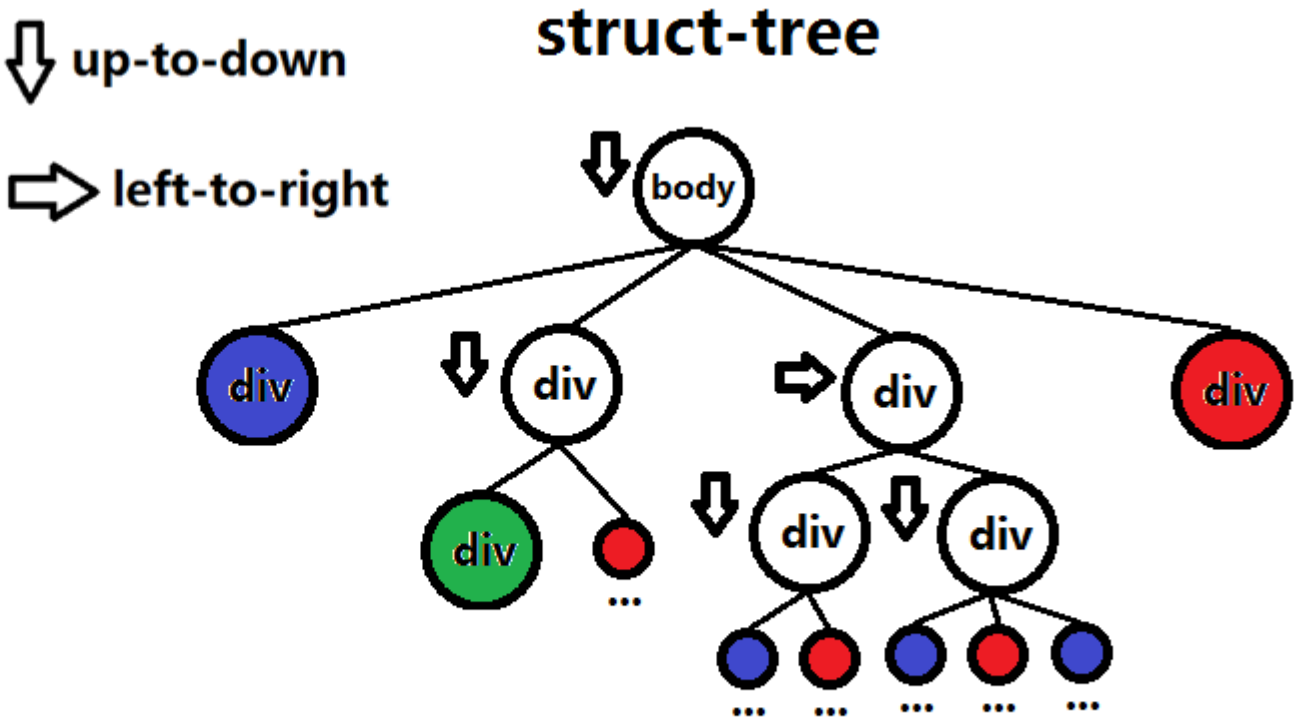
在统计树上计算**节点的各种特征(面积、颜色、文本信息等)**，依赖于节点特征，对节点进行**分类**，将节点标记**元素、块、结构**等类型，并将树形的节点转换为**平面数组**形式，方便后续分析。



结构树:struct-tree

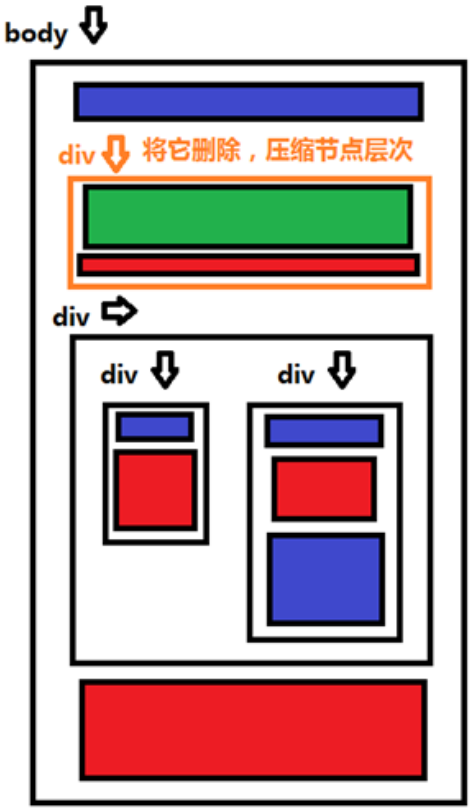
依赖统计树再构造**结构树**，主要保留有效的**用于布局的结构节点**，将统计树缩减为一颗**更简单的**结构树，同时计算内部节点**排列方式**。

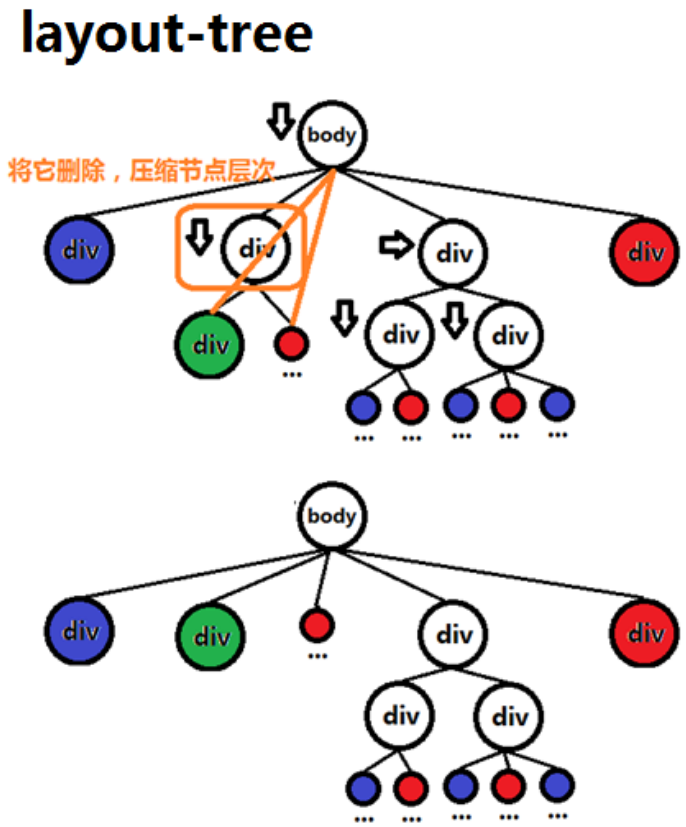




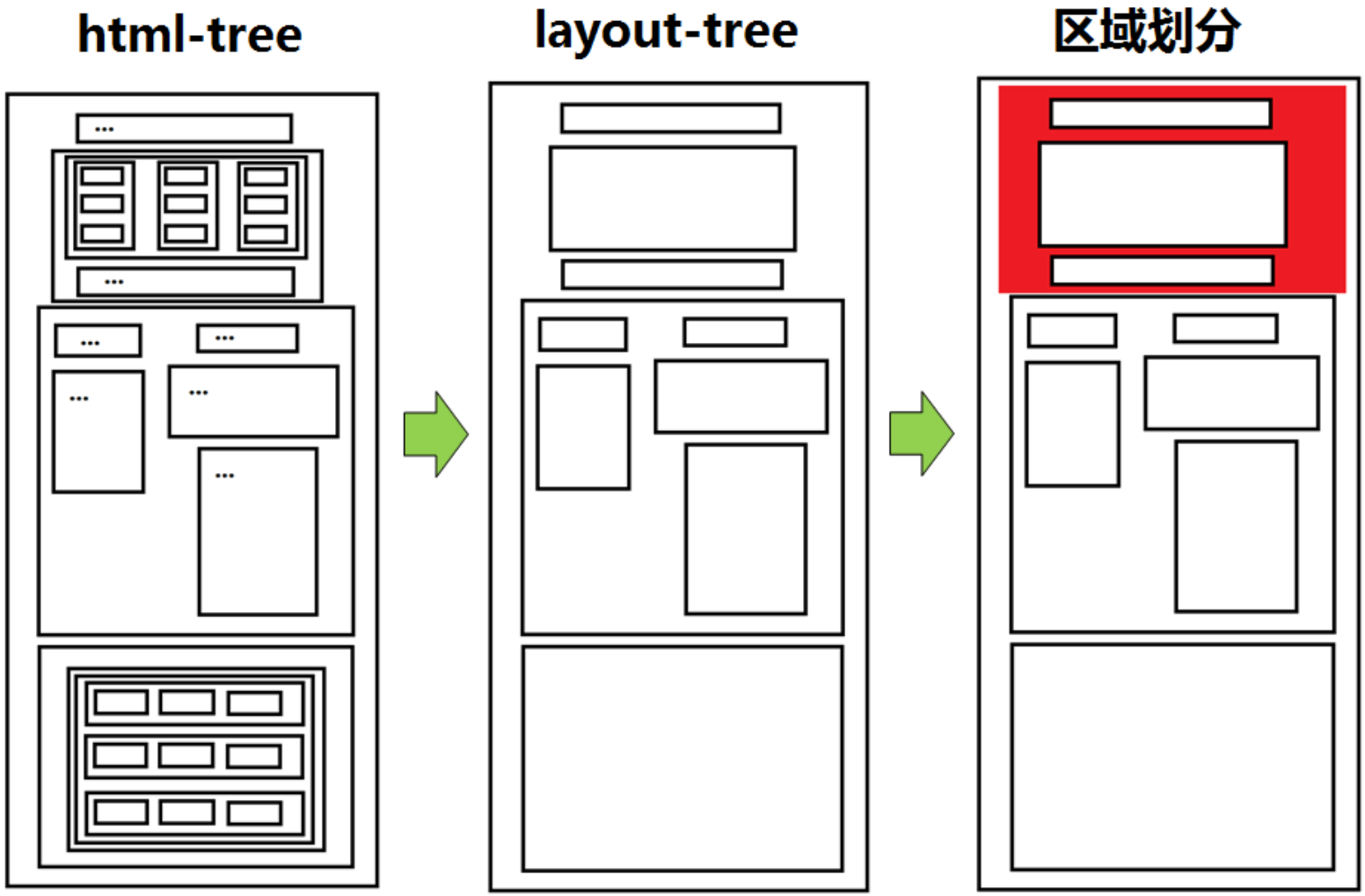
布局树:layout-tree

依赖结构树再构造**布局树**，将节点层次按照**布局**方法进行**压缩**，保证每个节点的孩子的布局与父亲的布局排列**不同**，使得树的结构更加**扁平**。



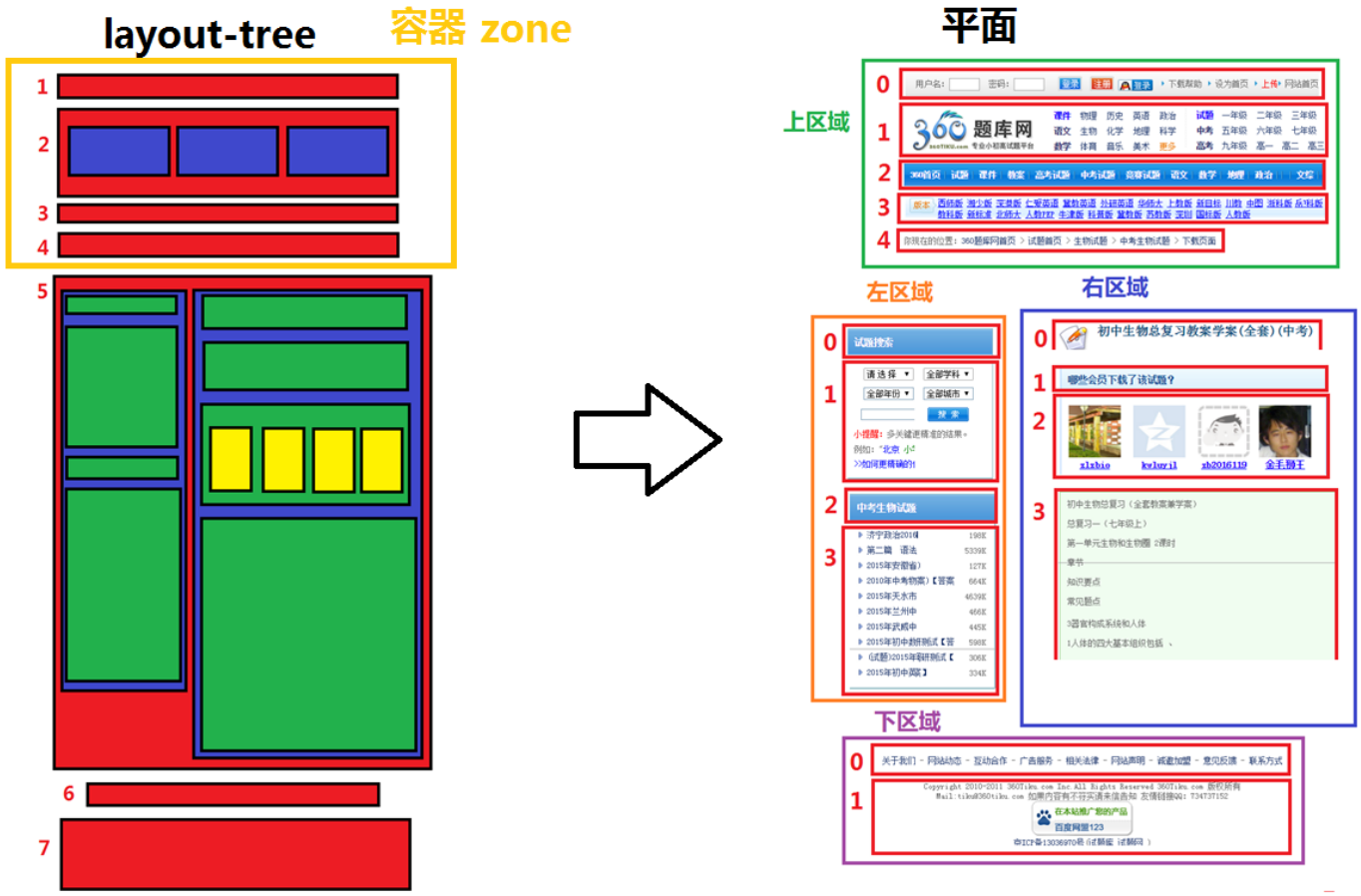


总结:为网页划分区域，构造平面数据结构做准备



平面构造算法

设置平面区域**容器zone**。**按层遍历**layout-tree，判断树上节点是否**可放入**容器zone中。**判断**节点**(机器学习模型)**是否可放入容器中，可能依赖节点的面积、宽高、叶结点个数等属性。若节点可**满足**放入条件，则将节点放入，否则该节点**拆分**为并排的**多个容器**，将该节点的**子节点**放入这些容器中。直到**第一层**节点全部完成遍历。



网页元素提取

完成平面等高级的**网页分析数据结构**构造后，利用这些数据结构，将网页中有价值的节点**元素内容**进行提取，例如提取标题、正文、评论等内容。这些具有**相同定义**的内容可能来自于**不同的**网页，例如，分别从商品页、新闻页、视频页中提取**评论内容**。

网页分类

网页分类，对互联网中的网页进行**标记、识别**的过程。利用网页分类，筛选网页库中**最有价值**的网页。一般将网页转换为**网页特征向量**，利用**有监督的机器学习**(如随机森林、逻辑回归)进行训练模型，再利用**机器学习模型**识别网页的类型。网页分类系统为**多个二分类**模型叠加在一起。这样每个网页就可以属于多种类型了，每个类型相互**没有依赖**，升级**优化成本**较低。

网页

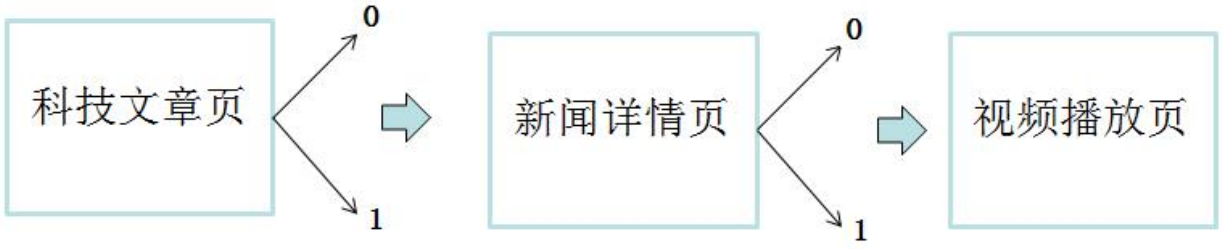
网页特征向量

a:1

b:20

c:0.6

...



关于计算机学习

1. 掌握一门**编译语言**
2. 掌握**算法与数据结构**
3. 掌握**一门脚本语言**
4. 掌握**开发环境**
5. 丰富其他**前沿知识**

关于技术面试

1. 听清并**搞懂**问题
2. **冷静思考**
3. 努力解决，**不**轻易**放弃**
4. 自信并**谦虚**

关于实习

1. **实习是**拿到BAT offer的**捷径**
2. 实习的**岗位有许多**:RD、QA、FE、OP、PM等
3. 尽量在一个实习岗位**坚持**半年
4. 实习还有**不菲的薪水**喔~
5. 关于学校课堂与上课，**自己把握**

关于求职与谈薪资

1.**表达对工作、职位、技术的热爱**:在谈薪资时，不管满意与否，也不管是否手里还有其他offer，一定要对招聘方表达你对这个岗位工作的兴趣，因为站在招聘者的角度，只有有热情的候选人将来才能做好工作。

2.表达对期望薪资的**客观理由**:

a.自身素质，既然已经通过了各项技术面试，那就说明已经胜任该工作岗位，所以再表达一下也是应该的。

b.竞品公司类似岗位，在谈薪资前，要对竞品公司的同职位的薪资有充分的了解，才能与更合理的与HR argue薪资。

c.应聘公司的相应级别或岗位的一般待遇，因为同是职称(title)，由于候选人个人素质不同，给的一般也不一样，要对该职称有整体了解。

3.**话不说死**:

在这一阶段，无论HR如何施压(不马上签就没了)，也不要马上答应或说死(不要马上对HR表达是否接受offer)，要留有余地，这个潜台词是我可能还在看别的机会，但是对于这个岗位我仍然保有很大兴趣。

4.**接offer**:

无论怎样，offer还是要接的，一般来讲，HR会给到他最终能给到的最高点(他也希望你入职，不然他本次的沟通就是无用的)，但是不代表一定会入职，如果真不太满意，可以拿着这个offer再去找别家，骑着骡子找马嘛。反正入职时间可以往后拖一拖。

5.态度:积极、乐观、谦虚、自信:无论你们对该职位满意与否，装也得装的这样。：)

关于职场

1. 正面看待工作中的各种问题，**积极面对**，让工作变得快乐
2. 遇到问题，**独立思考**，认真听取他人的建议，自己做决定
3. 技术与产品的发展在于创新，**敢于创新**，即使创新可能带来失败
4. 培养owner的精神，**勇于承担**，得到同事与上级的信任

关于人工智能

1. 人工智能**时代**的来临: PC桌面软件->互联网->移动互联网->**人工智能**
2. 人工智能的**学习**:打牢**基础**，接地气的解决问题，尝试新方法
3. 人工智能的**就业**:首选BAT等互联网行业巨头，不断**积累**，最终尝试创造新事物

关于创业

1. 认识**自我**，充分了解自己的**优点**与**缺点**
2. 建立**目标**，一个清晰且可执行的**梦想**
3. 寻找**资源**，**人与钱**是不可缺少的
4. 详细**计划**，越详细越好至少要明确一个月内的**milestone**
5. 迈出**第一步**，哪怕上面几点做的都不好，这一点是必须的

联系我们

小象学院：互联网新技术在线教育领航者

– 微信公众号：**小象学院**