小象学院

网站结构及常见爬虫策略

主　讲:　杨真

Part 1 基础

- 环境搭建
- HTML 基础
- 第一个10行代码的爬虫
- 内容抽取及解析
- HTTP 协议
- 辅助工具

Part 2 爬虫

- 网站结构及常见爬虫策略
- BBS网站结构分析及方案
- 控制节奏

Part 3 进阶

- MySQL 数据库
- 多线程
- 并行抓取
- 网站服务架构
- 表单、登录及Cookie处理

Part 4 实战

- 网站结构分析
- 网页抓取方案
- 数据提取
- 存储方案

Robots.txt

- 网站对爬虫的限制
- 利用 sitemap 来分析网站结构和估算目标网页的规模

http://www.mafengwo.cn/robots.txt

```
User-agent: *
Disallow: /music/
Disallow: /travel-photos-albums/
Disallow: /lushu/
Disallow: /hc/
Disallow: /hb/
Disallow: /insure/show.php
Disallow: /myvisa/index.php
Disallow: /booking/discount_booking.php
Disallow: /secrect/
Disallow: /gonglve/visa.php
Disallow: /gonglve/visa_info.php
Disallow: /gonglve/visa_case.php
Disallow: /gonglve/visa_seat.php
Disallow: /gonglve/visa_readme.php
Disallow: /gonglve/insure.php
Disallow: /gonglve/insurer.php
Disallow: /gonglve/hotel.php
Disallow: /gonglve/hotel_list.php
Disallow: /gonglve/flight.php
Disallow: /gonglve/traffic.php
Disallow: /gonglve/scenery.php
Disallow: /insure/tips-*.html
Disallow: /skb-i/
Disallow: /weng/pin.php?tag=*
Disallow: /rank/
Disallow: /hotel/s.php
Disallow: /photo/mdd/*_*.html
Disallow: /photo/poi/
Disallow: /hotel/*/?sFrom=*

Sitemap: http://www.mafengwo.cn/sitemapIndex.xml
```

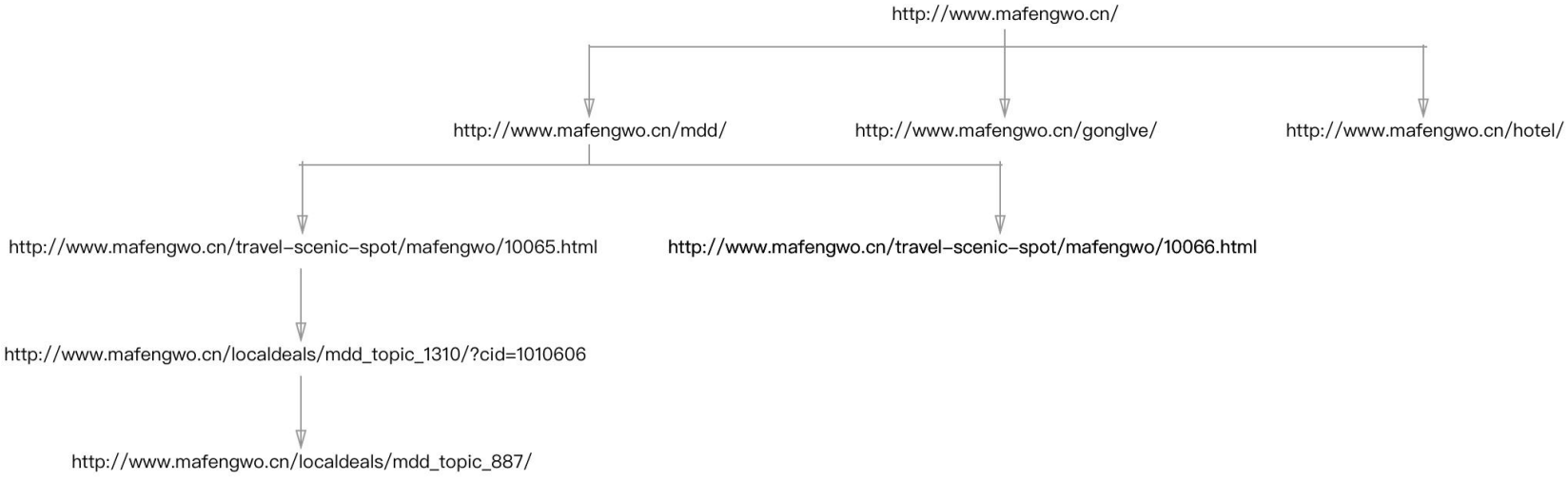
Sitemap

http://www.mafengwo.cn/sitemapIndex.xml

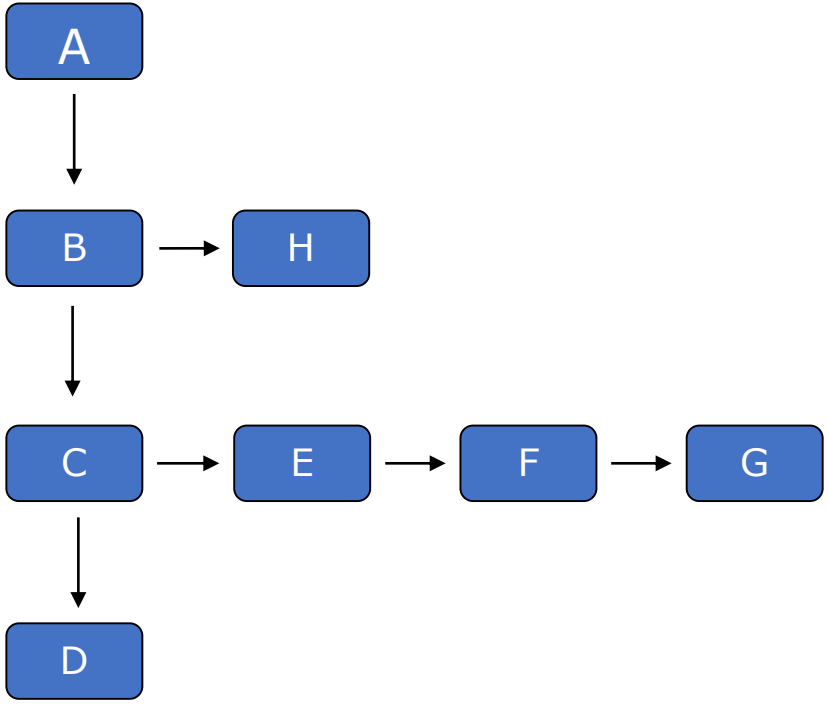
```
<sitemapindex xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <sitemap>
    <loc>http://www.mafengwo.cn/article-0.xml</loc>
    <lastmod>2018-12-31</lastmod>
  </sitemap>
  <sitemap>
    <loc>http://www.mafengwo.cn/article-1.xml</loc>
    <lastmod>2018-12-31</lastmod>
  </sitemap>
  <sitemap>
    <loc>http://www.mafengwo.cn/article-2.xml</loc>
    <lastmod>2018-12-31</lastmod>
  </sitemap>
  <sitemap>
    <loc>http://www.mafengwo.cn/article-3.xml</loc>
    <lastmod>2018-12-31</lastmod>
  </sitemap>
  <sitemap>
    <loc>http://www.mafengwo.cn/articleList-0.xml</loc>
    <lastmod>2018-12-31</lastmod>
  </sitemap>
  <sitemap>
    <loc>http://www.mafengwo.cn/articlePhoto-0.xml</loc>
    <lastmod>2018-12-31</lastmod>
  </sitemap>
  <sitemap>
    <loc>http://www.mafengwo.cn/articlePhoto-1.xml</loc>
    <lastmod>2018-12-31</lastmod>
  </sitemap>
  <sitemap>
    <loc>http://www.mafengwo.cn/articlePhotoDetail-0.xml</loc>
    <lastmod>2018-12-31</lastmod>
  </sitemap>
  <sitemap>
    <loc>http://www.mafengwo.cn/books-0.xml</loc>
    <lastmod>2018-12-31</lastmod>
  </sitemap>
```

```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://www.mafengwo.cn/yj/10010/</loc>
    <lastmod>2018-12-31 02:10:03</lastmod>
    <changefreq>daily</changefreq>
    <priority>0.8</priority>
  </url>
  <url>
    <loc>http://www.mafengwo.cn/yj/10010/1-0-2.html</loc>
    <lastmod>2018-12-31 02:10:03</lastmod>
    <changefreq>daily</changefreq>
    <priority>0.7</priority>
  </url>
  <url>
    <loc>http://www.mafengwo.cn/yj/10010/1-0-3.html</loc>
    <lastmod>2018-12-31 02:10:03</lastmod>
    <changefreq>daily</changefreq>
    <priority>0.7</priority>
  </url>
  <url>
    <loc>http://www.mafengwo.cn/yj/10010/1-0-4.html</loc>
    <lastmod>2018-12-31 02:10:03</lastmod>
    <changefreq>daily</changefreq>
    <priority>0.7</priority>
  </url>
  <url>
    <loc>http://www.mafengwo.cn/yj/10010/1-0-5.html</loc>
    <lastmod>2018-12-31 02:10:03</lastmod>
    <changefreq>daily</changefreq>
    <priority>0.7</priority>
  </url>
  <url>
    <loc>http://www.mafengwo.cn/yj/10011/</loc>
    <lastmod>2018-12-31 02:10:03</lastmod>
    <changefreq>daily</changefreq>
    <priority>0.8</priority>
  </url>
```

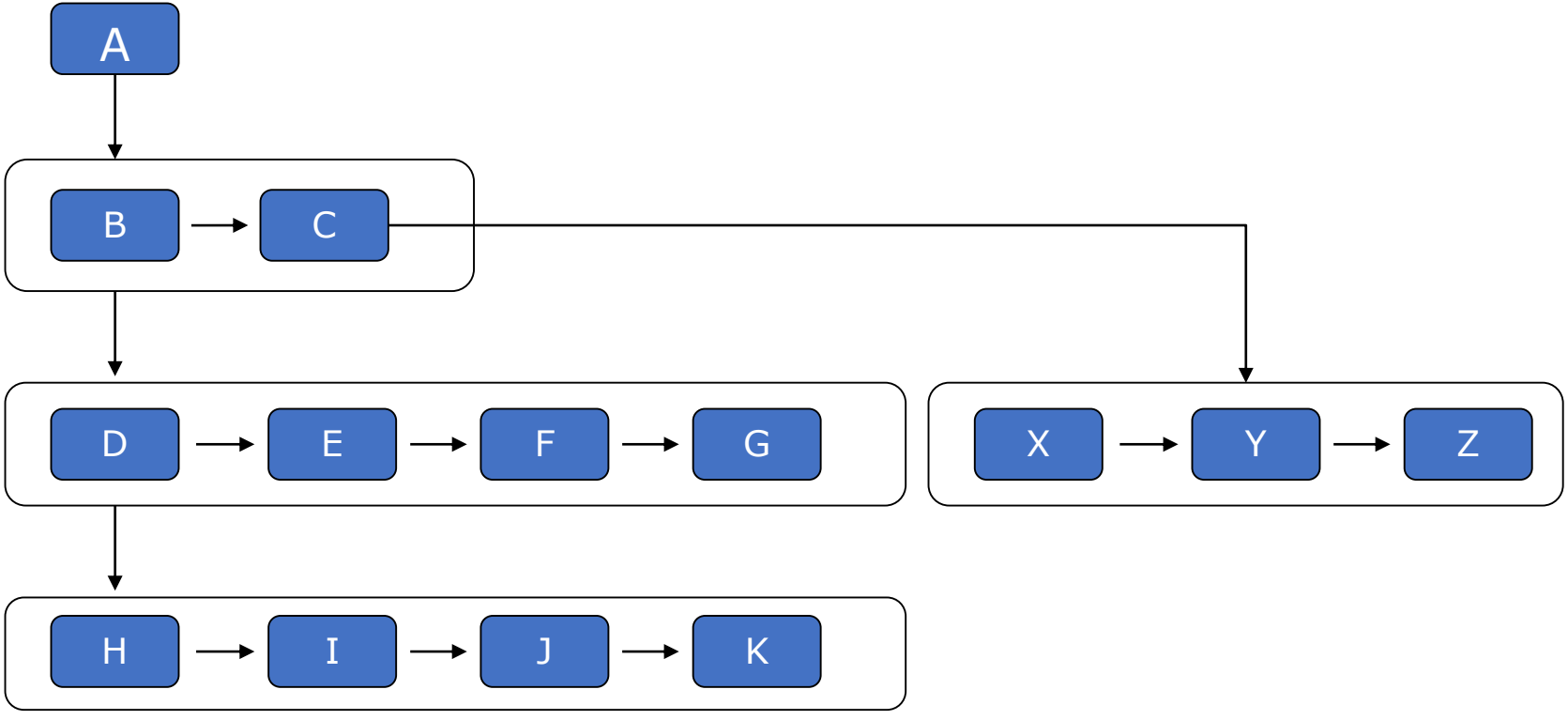
网站结构


http://www.mafengwo.cn/
http://www.mafengwo.cn/mdd/
http://www.mafengwo.cn/gonglve/
http://www.mafengwo.cn/hotel/
http://www.mafengwo.cn/travel-scenic-spot/mafengwo/10065.html
http://www.mafengwo.cn/travel-scenic-spot/mafengwo/10066.html
http://www.mafengwo.cn/localdeals/mdd_topic_1310/?cid=1010606
http://www.mafengwo.cn/localdeals/mdd_topic_887/

深度优先



宽度优先



宽度优先 vs 深度优先

- 重要的网页距离种子站点比较近
- 万维网的深度并没有很深，一个网页有很多路径可以到达
- 宽度优先有利于多爬虫并行合作抓取
- 深度限制与宽度优先相结合

定制抓取

- 按板块
- 按内容
- 网站特定结构
- 利用搜索

小象学院

Part 1 基础

- 环境搭建
- HTML 基础
- 第一个10行代码的爬虫
- 内容抽取及解析
- HTTP 协议
- 辅助工具

Part 2 爬虫

- 网站结构及常见爬虫策略
- BBS网站结构分析及方案
- 控制节奏

Part 3 进阶

- MySQL 数据库
- 多线程
- 并行抓取
- 网站服务架构
- 表单、登录及Cookie处理

Part 4 实战

- 网站结构分析
- 网页抓取方案
- 数据提取
- 存储方案

BBS 结构

- 首页：TOP10 热门文章
- 板块：获取所有的板块入口
- 文章：获取每个板块，每一页的文章列表
- 回帖：获取每一个帖子的主文，及所有跟帖

Board List

Format:

http://[domain]/nForum/section/[index]?ajax

注意：

存在子版块：汽车由 4 个子版块组成

Article List

Format:

http://www.newsmth.net/nForum/board/[BoardName]?ajax&p=[page]

注意：

http://www.newsmth.net/nForum/#!board/[BoardName]?p=[page] 这个请求只返回框架，没有数据

Post Detail

Format:

http://www.newsmth.net/nForum/article/[board_name]/[article_id]?ajax&p=[page]

注意：

http://www.newsmth.net/nForum/#!article /[BoardName]/page 这个请求只返回框架，没有数据

小象学院

Part 1 基础

- 环境搭建
- HTML 基础
- 第一个10行代码的爬虫
- 内容抽取及解析
- HTTP 协议
- 辅助工具

Part 2 爬虫

- 网站结构及常见爬虫策略
- BBS网站结构分析及方案
- 控制节奏

Part 3 进阶

- MySQL 数据库
- 多线程
- 并行抓取
- 网站服务架构
- 表单、登录及Cookie处理

Part 4 实战

- 网站结构分析
- 网页抓取方案
- 数据提取
- 存储方案

控制节奏

- 网站对爬虫的限制，最主要依赖于每个IP（或每个用户）的访问频次，过高频率的访问会被网站限制访问
- 控制节奏主要针对每个目标地址的访问频率

小象学院

只为遇见明天更优秀的你！