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。

- 课程详情请咨询
  - 微信公众号：大数据分析挖掘
  - 新浪微博：ChinaHadoop

分布式爬虫

大纲

- 环境搭建
- HTTP 协议
- 宽度与深度抓取
- 不重复抓取策略
- 网站结构分析

环境搭建

基础环境 – Python Unix

- python 2.7
- pip，并设置 pip 源

配置 pip conf，自动设置源

```
# mkdir ~/.pip/
# vim ~/.pip/pip.conf
[global]
index-url=https://pypi.tuna.tsinghua.edu.cn/simple
```

也可以每次安装的时候制定 source

```
# pip install –i https://pypi.tuna.tsinghua.edu.cn/simple lxml
```

基础环境 - Python Windows

- 直接下载 Anaconda，很多比较难以安装的源都已经包含了
- 仍然配置 pip 源，各个系统的默认 pip.ini 位置不同，需要根据实际情况设置

官网：　　https://anaconda.org/

下载主页：https://www.continuum.io/downloads

HTTP 协议

TCP/IP 四层 与 OSI 七层

HTTP 协议

- 物理层：电器连接
- 数据链路层：交换机，STP，帧中继
- 网络层：路由器，IP 协议
- 传输层：TCP、UDP 协议
- 会话层：建立通信连接，网络拨号
- 表示层：每次连接只处理一个请求
- 应用层：HTTP、FTP

HTTP 协议

- 应用层的协议
- 无连接：每次连接只处理一个请求
- 无状态：每次连接、传输都是独立的

HTTP HEADER

REQUEST 部分的 HTTP HEADER

- Accept: text/plain
- Accept-Charset: utf-8
- Accept-Encoding: gzip, deflate
- Accept-Language: en-US
- Connection: keep-alive
- Content-Length: 348
- Content-Type: application/x-www-form-urlencoded
- Date: Tue, 15 Nov 1994 08:12:31 GMT
- Host: en.wikipedia.org:80
- User-Agent: Mozilla/5.0 (X11; Linux x86_64; rv:12.0) Gecko/20100101 Firefox/21.0
- Cookie: $Version=1; Skin=new;

keep-alive

HTTP是一个**请求<->响应**模式的典型范例，即客户端向服务器发送一个请求信息，服务器来响应这个信息。在老的HTTP版本中，每个请求都将被创建一个新的**客户端->服务器**的连接，在这个连接上发送请求，然后接收请求。这样的模式有一个很大的优点就是，它很简单，很容易理解和编程实现；它也有一个很大的缺点就是，它效率很低，因此Keep-Alive被提出用来解决效率低的问题。

Keep-Alive功能使客户端到服务器端的连接持续有效，当出现对服务器的后继请求时，Keep-Alive功能避免了建立或者重新建立连接。

HTTP/1.1

默认情况下所在HTTP1.1中所有连接都被保持，除非在请求头或响应头中指明要关闭：Connection: Close

HTTP HEADER

RESPONSE 的 HTTP HEADER

- Accept-Patch: text/example;charset=utf-8
- Cache-Control: max-age=3600
- Content-Encoding: gzip
- Last-Modified: Tue, 15 Nov 1994 12:45:26 GMT
- Content-Language: da
- Content-Length: 348
- ETag: "737060cd8c284d8af7ad3082f209582d"
- Expires: Thu, 01 Dec 1994 16:00:00 GMT
- Location: http://www.w3.org/pub/WWW/People.html
- Set-Cookie: UserID=JohnDoe; Max-Age=3600; Version=1
- Status: 200 OK

HTTP 请求方法

HTTP Method	RFC	Request Has Body	Response Has Body	Safe	Idempotent	Cacheable
GET	RFC 7231	No	Yes	Yes	Yes	Yes
HEAD	RFC 7231	No	No	Yes	Yes	Yes
POST	RFC 7231	Yes	Yes	No	No	Yes
PUT	RFC 7231	Yes	Yes	No	Yes	No
DELETE	RFC 7231	No	Yes	No	Yes	No
CONNECT	RFC 7231	Yes	Yes	No	No	No
OPTIONS	RFC 7231	Optional	Yes	Yes	Yes	No
TRACE	RFC 7231	No	Yes	Yes	Yes	No
PATCH	RFC 5789	Yes	Yes	No	No	Yes

HTTP 响应状态码

- 2XX 成功
- 3XX 跳转
- 4XX 客户端错误
- 500 服务器错误

HTTP 响应状态码 300

- 300 Multiple Choices 存在多个可用的资源，可处理或丢弃
- 301 Moved Permanetly 重定向
- 302 Found 重定向
- 304 Not Modified 请求的资源未更新，丢弃

一些 Python 库，例如 urllib2 已结对重定向做了处理，会自动跳转；动态网页处理的时候，也是自动跳转，所以不需要单独处理

HTTP 响应状态码 400、500

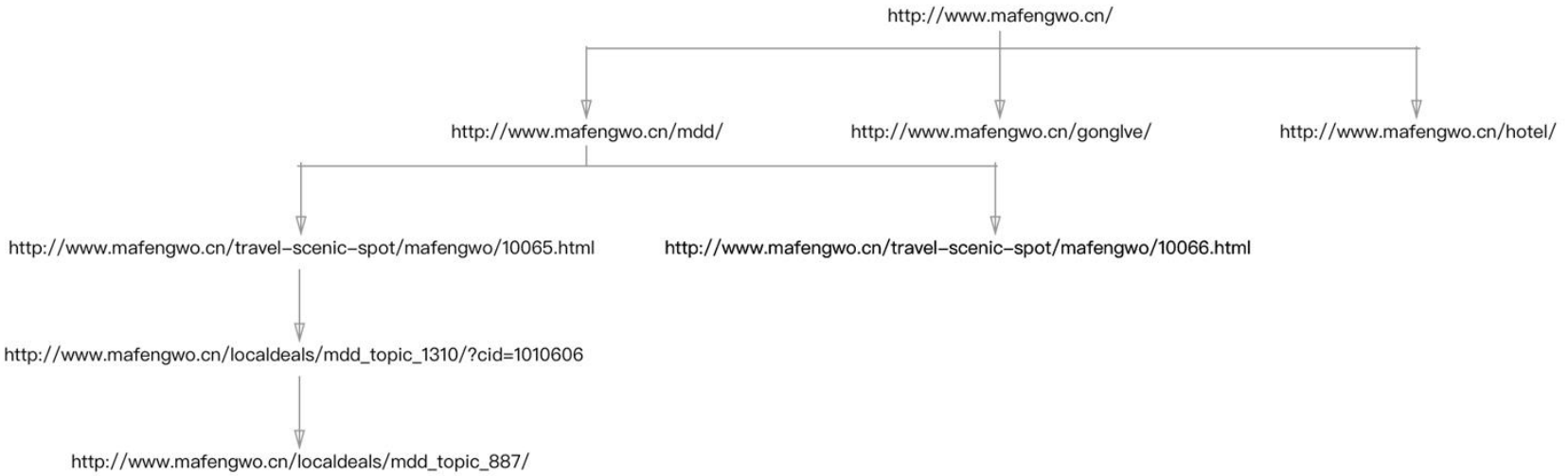
- 400 Bad Request 客户端请求有语法错误，不能被服务器所理解
- 401 Unauthorized 请求未经授权，这个状态代码必须和WWW-Authenticate报头域一起使用
- 403 Forbidden 服务器收到请求，但是拒绝提供服务
- 404 Not Found 请求资源不存在，eg：输入了错误的URL
- 500 Internal Server Error 服务器发生不可预期的错误
- 503 Server Unavailable 服务器当前不能处理客户端的请求，一段时间后可能恢复正常

错误处理

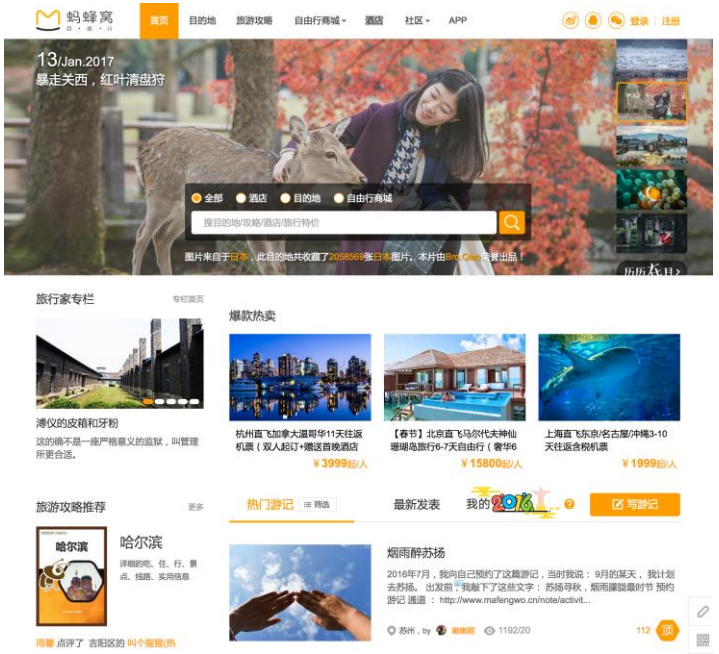
- 400 Bad Request 检查请求的参数或者路径
- 401 Unauthorized 如果需要授权的网页，尝试重新登录
- 403 Forbidden
  - 如果是需要登录的网站，尝试重新登录
  - IP被封，暂停爬取，并增加爬虫的等待时间，如果拨号网络，尝试重新联网更改IP
- 404 Not Found 直接丢弃
- 5XX 服务器错误，直接丢弃，并计数，如果连续不成功，WARNING 并停止爬取

宽度及深度抓取

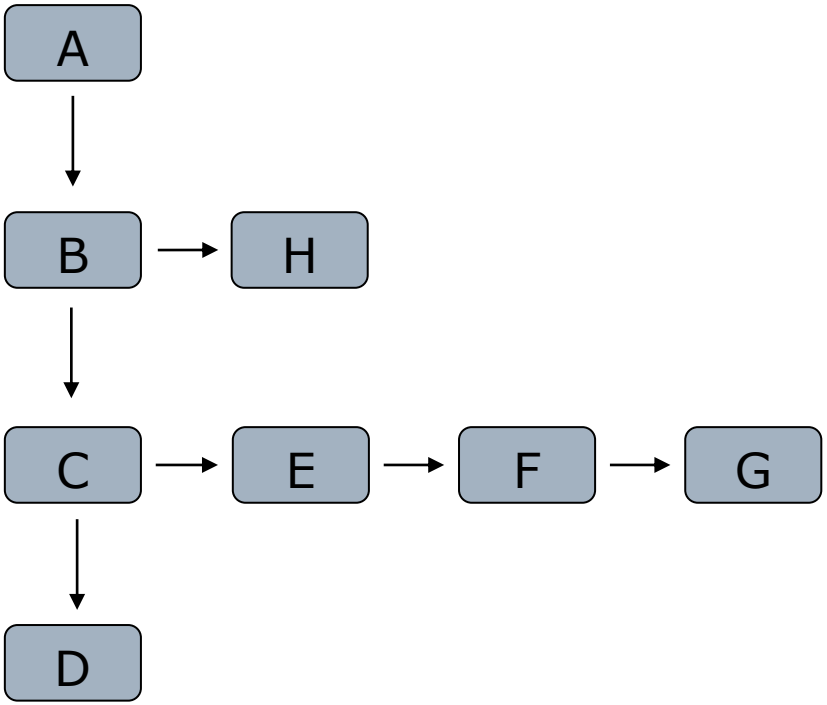
网页抓取原理


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/

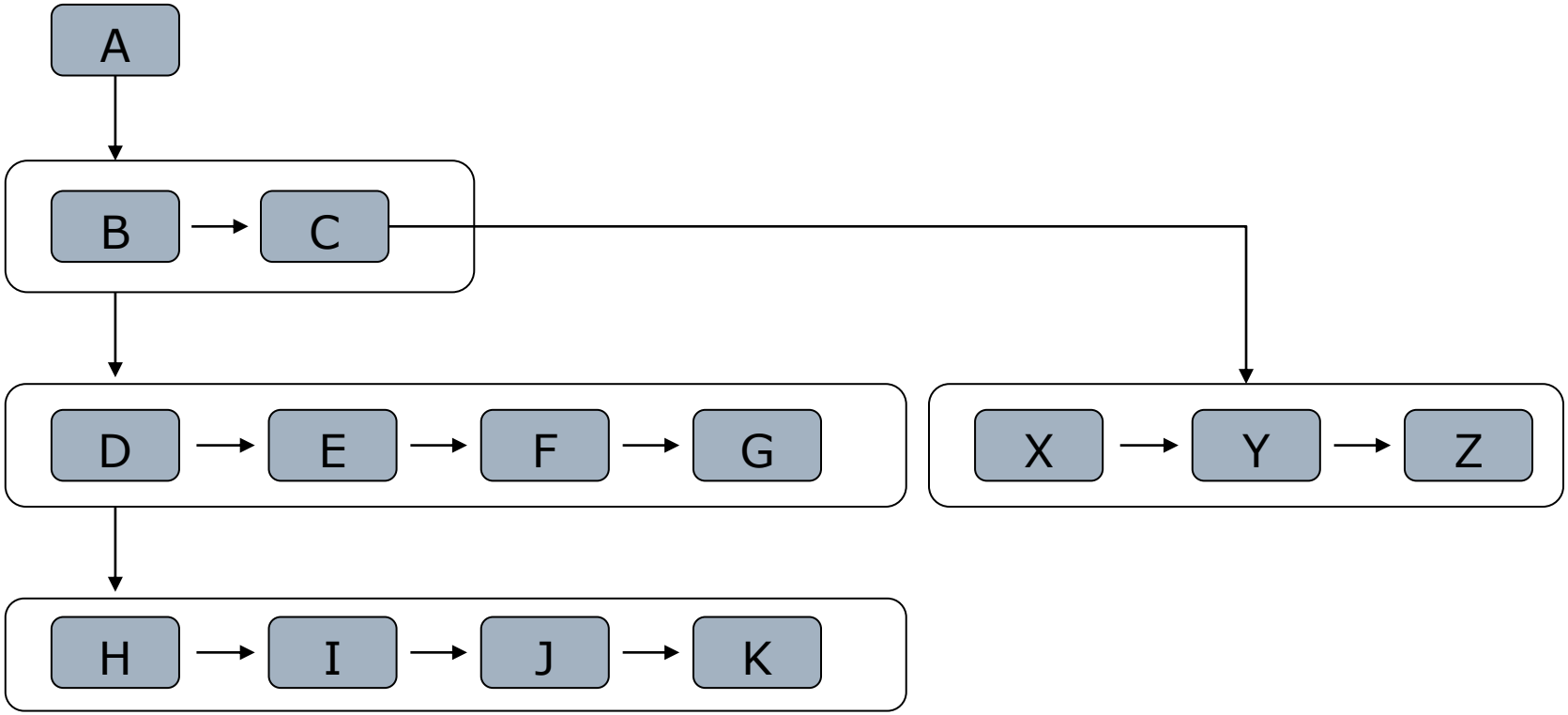
Mafengwo 的结构



深度优先策略



宽度优先策略



选择哪种策略？

- 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

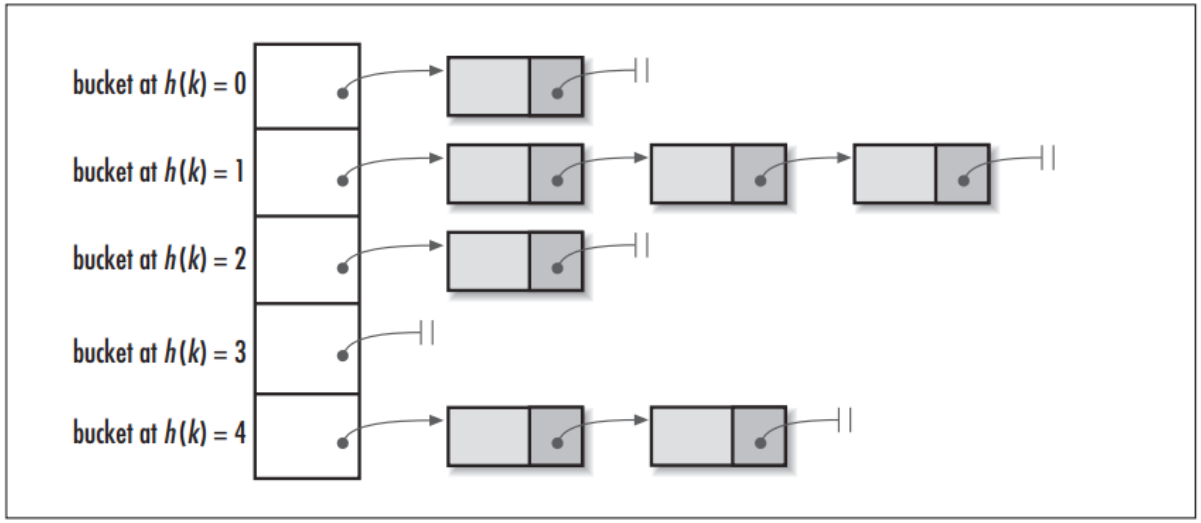
不重复抓取策略

如何记录抓取历史？

1. 将访问过的URL保存到数据库 效率太低
2. 用HashSet将访问过的URL保存起来。那只需接近O(1)的代价就可以查到一个URL是否被访问过了。消耗内存
3. URL经过MD5或SHA-1等单向哈希后再保存到HashSet或数据库。
4. Bit-Map方法。建立一个BitSet，将每个URL经过一个哈希函数映射到某一位。

MD5 函数

MD5 签名是一个哈希函数，可以将任意长度的数据量转换为一个固定长度的数字（通常是4个整型，128位）。计算机不可能有2的128那么大内存，因此实际的哈希表都会是URL.MD5再%n（即取模）。现实世界的URL组合必然超越哈希表的槽位数，因此碰撞是一定存在的，一般的HASH函数，例如Java的 HashTable 是一个HASH表再跟上一个链表，链表里存的是碰撞结果



提高效率？

- 评估网站的网页数量
- 选择合适的HASH算法和空间阈值，降低碰撞几率
- 选择合适的存储结构和算法

评估网页数量


Bai du 百度
site:www.mafengwo.cn
百度一下
网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»
所有网页 时间不限 所有网页和文件 www.mafengwo.cn 清除
蚂蜂窝
自·由·行
www.mafengwo.cn的站点信息
备案方：北京蚂蜂窝网络科技有限公司
该网站共有 26,433,119 个网页被百度收录 到站长平台分析收录量>>
站长工具
链接提交 死链提交 流量与关键词 移动适配
旅游攻略,自由行,自助游攻略,旅游社交分享网站 - 蚂蜂窝
蚂蜂窝
自·由·行
在做攻略的时候非常希望能看到大家的分享,这样能在出行中有计划的玩着。看到大家分享辣么多... 8 台湾,by april 190/4 塞班岛,一个不得不去的海岛,清新...
www.mafengwo.cn/ V3 - 百度快照 - 477条评价

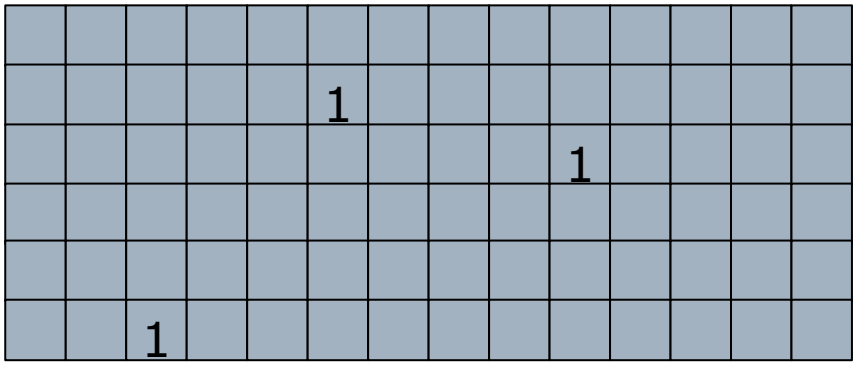
评估网页数量


Google
site:www.mafengwo.cn/gonglve/
全部 图片 新闻 地图 更多 设置 工具
找到约 13,200 条结果 （用时 0.47 秒）
Google 推广
尝试使用 Google Search Console
www.google.com/webmasters/
您对 www.mafengwo.cn/gonglve 是否具有所有权？请从 Google 获取索引和排名数据。
Google
site:www.facebook.com
全部 图片 新闻 地图 更多 设置 工具
找到约 1,960,000,000 条结果 （用时 0.39 秒）
Facebook - Log In or Sign Up
https://www.facebook.com/ ▾ 翻译此页
Create an account or log into Facebook. Connect with friends, family and other people you know. Share photos and videos, send messages and get updates.

BITMAP 方式记录

将URL的MD5值再次哈希，用一个或多个BIT位来记录一个URL：

1. 确定空间大小 e.g. facebook 1.5Gb
2. 按倍增加槽位 e.g. 16GB
3. HASH 算法映射(murmurhash3, cityhash) Python: mmh3 bitarray



BITMAP 方式记录

pip install murmurhash3 bitarray

```
from bitarray import bitarray
import mmh3

offset = 2147483647 // 2^31 - 1
bit_array = bitarray(4*1024*1024*1024)
bit_array.setall(0)

# mmh3 hash value 32 bit signed int
# add offset to make it unsigned int 0 ~ 2^32-1
b1 = mmh3.hash(url, 42) + offset
bit_array[b1] = 1
```

BITMAP 方式记录

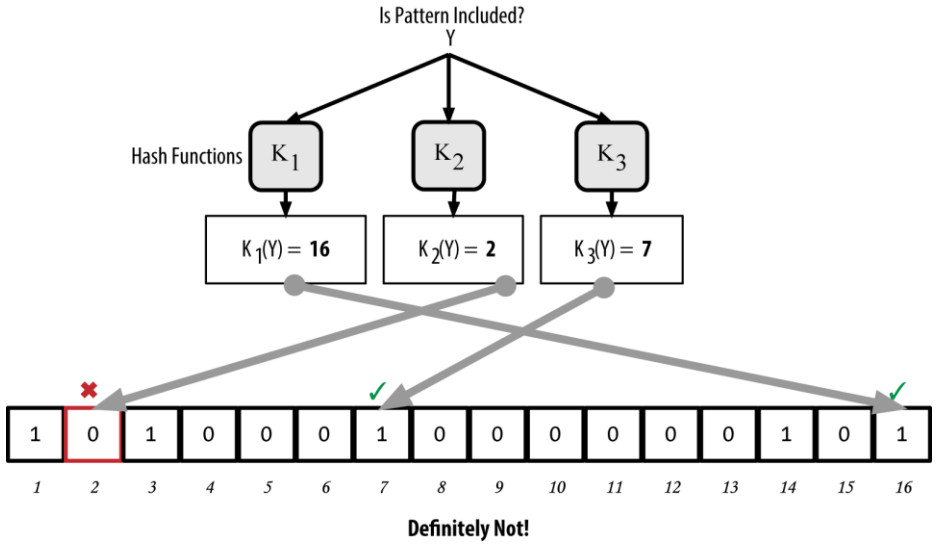
- 优势：对存储进行了进一步压缩，在MD5的基础上，可以从128位最多压缩到1位，一般情况，如果用4bit或者8bit表示一个url，也能压缩32或者16倍

- 缺陷：碰撞概率增加

Bloom Filter

Bloom Filter使用了多个哈希函数，而不是一个。创建一个m位BitSet，先将所有位初始化为0，然后选择k个不同的哈希函数。第i个哈希函数对字符串str哈希的结果记为h（i，str），且h（i，str）的范围是0到m-1。

只能插入，不能删除！！



pybloomfilter

安装

pip install pybloomfilter (可能运行时会crash）

git clone https://github.com/axiak/pybloomfiltermmap.git
python setup.py install

构造函数

class pybloomfilter.BloomFilter(capacity : int, error_rate : float[, filename=None : string][, perm=0755])

并不实际检查容量，如果需要比较低的error_rate，则需要设置更大的容量

Sample

```
>>> fruit = pybloomfilter.BloomFilter(100000, 0.1, '/tmp/words.bloom')
>>> fruit.update(('apple', 'pear', 'orange', 'apple'))
>>> len(fruit) 3
>>> 'mike' in fruit
False
>>> 'apple' in fruit
True
```

官方文档

https://media.readthedocs.org/pdf/pybloomfiltermmap3/latest/pybloomfiltermmap3.pdf

如何有效记录抓取历史？

- 多数情况下不需要压缩，尤其网页数量少的情况
- 网页数量大的情况下，使用 Bloom Filter 压缩
- 重点是计算碰撞概率，并根据碰撞概率来确定存储空间的阈值
- 分布式系统，将散列映射到多台主机的内存

网站结构分析

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

```
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

```xml
<sitemapindex xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
    <sitemap>
        <loc>http://www.mafengwo.cn/article-0.xml</loc>
        <lastmod>2017-02-01</lastmod>
    </sitemap>
    <sitemap>
        <loc>http://www.mafengwo.cn/article-1.xml</loc>
        <lastmod>2017-02-01</lastmod>
    </sitemap>
```

```xml
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
    <url>
        <loc>http://www.mafengwo.cn/i/6153755.html</loc>
        <lastmod>2017-02-01 02:05:28</lastmod>
        <changefreq>daily</changefreq>
        <priority>0.9</priority>
    </url>
    <url>
        <loc>http://www.mafengwo.cn/i/6153749.html</loc>
        <lastmod>2017-02-01 02:05:28</lastmod>
        <changefreq>daily</changefreq>
        <priority>0.9</priority>
    </url>
```

有效率抓取特定内容

➢ 利用 sitemap 里的信息

直接对目标网页 .html 进行抓取

➢ 对网站目录结构进行分析

大多数网站都会存在明确的 top-down 的分类的目录结构，我们可以进入特定目录进行抓取

对于 www.mafengwo.cn 这个网站，所有旅游的游记都位于 www.mafengwo.cn/mdd 下，按照城市进行了分类，每个城市的游记位于城市的首页。

城市的首页： /travel-scenic-spot/mafengwo/10774.html

游记的分页格式： /yj/10774/1-0-01.html

游记的页面：/i/3523364.html

疑问

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

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：大数据分析挖掘
- 新浪微博：ChinaHadoop