



大连理工大学

信息检索研究室

*Information Retrieval Laboratory of DUT*

# 搜索引擎

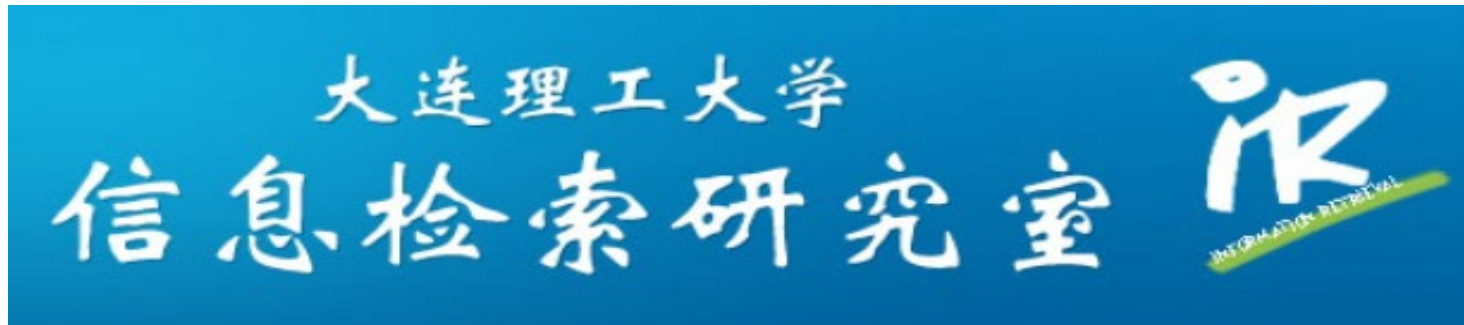
电信学部-计算机科学与技术学院

博士研究生 张晓堃

导师 林鸿飞教授

- 2012-2016 本科 大连理工大学 计算机科学与技术
- 2016-2018 创业公司工作
- 2020 硕博连读 - 博士三年级
- 导师：林鸿飞
- 方向：推荐系统 -> 序列/会话推荐

学术主页: <https://zhang-xiaokun.github.io/>



实验室网址: <https://ir.dlut.edu.cn/>

公众号: 大连理工大学信息检索研究室

研究方向: 自然语言处理、情感计算和幽默计算、信息检索与信息推荐、生物医学文本挖掘、知识图谱与问答系统、智慧司法等

**欢迎真正想做科研的同学加入我们!**

**保研&科研助手**

- 理解搜索引擎工作原理（计算机网络/web构建等）
- 搭建一个可运行的实验系统(**DEMO**)
  - ◆ 通过**亲自动手**（可借助开源工具）搭建一个**完整、可运行**的小型全文检索实验系统
  - ◆ 爬取信息、建立索引、用户界面

- 尽量通俗易懂地给大家讲解相关知识
- 让大家有所收获
  - ◆ 以后刷微博、看新闻、追剧的时候 “呀！这个我知道！”
  - ◆ 实用性、以后的科研、生活中会用（python数据处理，网络数据获取等）
- 与大家共同学习

## 搜索皆智能 智能皆搜索

加拿大两院院士-裴健

Google  
谷歌

ASK 爱问

bing  
必应 Beta

Dogou 狗狗

YAHOO!  
中国雅虎

Baidu 百度

SOSO 搜搜  
腾讯旗下搜索网站

Sogou 搜狗

中搜  
zhongsou.com

有道 youdao  
网易旗下搜索

Baidu 百科

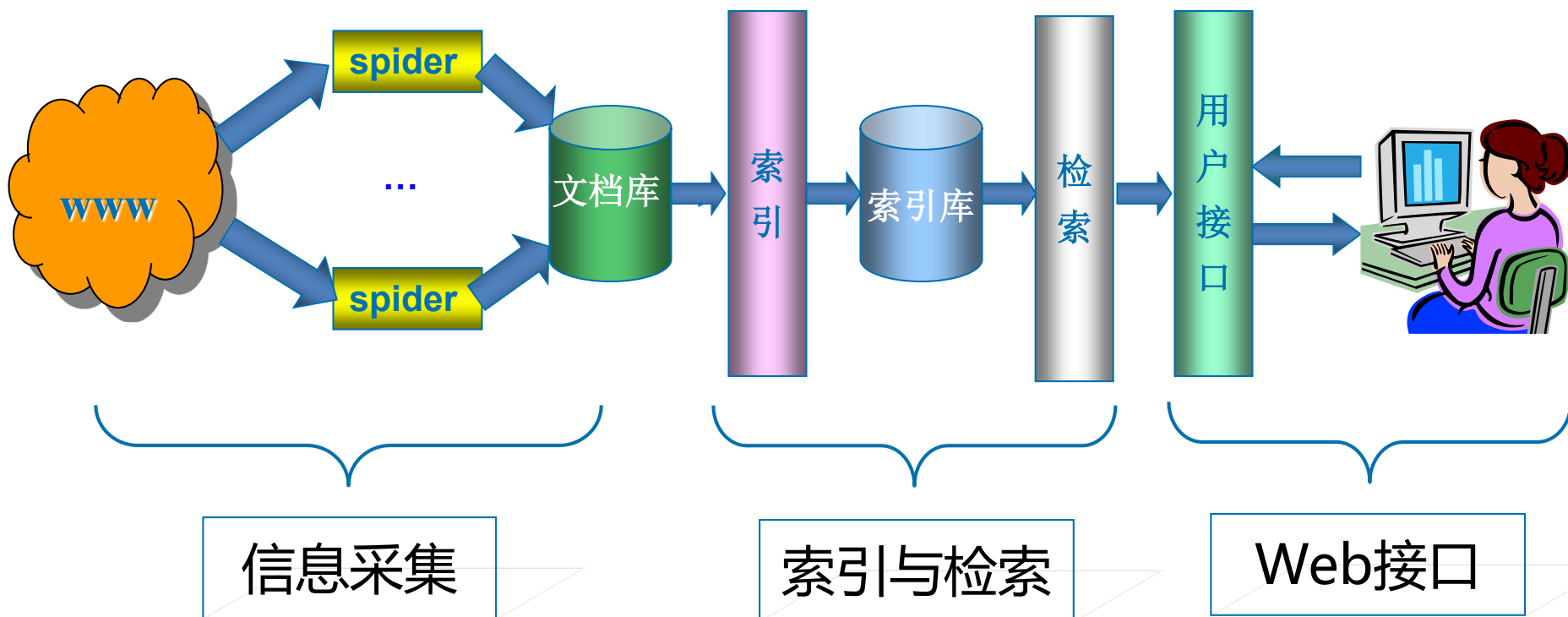
# 搜索引擎基本框架



大连理工大学  
信息检索研究室



Information Retrieval Laboratory of DUT



- **Web信息的搜集**
- **基于Lucene的索引与检索**
- **提供Web服务**



hao123

设为首页

大连 切换  
七日天气雷阵雨  
17 ~ 11°C晴  
18 ~ 12°C10月16日 周五 农历查询  
宜 破屋 坏垣 星座运势推荐: 生鲜神券, 点击领取!  
邮箱: 

登录 | 网盘

Baidu 百度

网页 | 搜你想搜

百度一下

● 大学老师茶叶中夹带毒品网络贩卖

链接到另一个网址(html)

青岛此次疫情由患者共用CT室引发热 广州新增1例无症状感染者热 解放军军机15小时内两度巡台...  
成都大学党委书记遗体已被找到热 浙江嘉兴公布新冠疫苗价格 第40个世界粮食日荐

hao123新闻 人民网 新华网 央视网 国际在线 中国日报 中国网 中经网 光明网 央广网 求是网 中青网 网信网 大连市政府

百度 腾讯 网易 百度地图 hao123影视 免费游戏 淘宝网 京东 苏宁易购  
携程旅行网 好看视频 爱奇艺 淘宝网 凤凰网 知网 哔哩哔哩 知乎 虎扑

提示: 您的自定义网址存在丢失风险, 建议马上登录来保存。

同步网址到账号

58同城 唯品会 苏宁秒杀 天涯社区 AcFun弹幕 百度学术 学信网 虎扑社区 中国大学MOOC  
网易云音乐 12306 企查查 安居客房产 工商银行 人事考试网 QQ邮箱 豆瓣网 彩票·走势图

点击展开

精选 社会 娱乐 生活 体育



重磅! 特朗普: 本拉登还活着 奥巴马骗人

浙江刚刚官宣! 新冠疫苗价格曝光 新闻  
郑州闹市区现大型墓群? 官方回应 热点

推荐 视频 国内 国际 娱乐 小视频 体育 科技 财经 汽车 游戏 女性 历史

续写更多“春天的故事”——习近平总书记出席深圳经济特区建立40周年庆祝大会并在广东考察纪实...



澳大利亚涉华“阴谋论”制造者

据最新统计, 目前该研究所约550万美元的年度预算中, 只有43%来自国防部。该研究所年度报告显示, 除国防部外, 该机构主要有三方面的资助者: 第一类是洛克希德-马丁公司、诺斯罗普-格鲁...

国际在线 1天前更新

hao123

设为首页

大连 切换  
七日天气

雷阵雨  
17 ~ 11°C

晴  
18 ~ 12°C

10月16日 周五  
宜 破屋 坏垣  
农历查询  
星座运势

推荐: 生鲜神券, 点击领取!  
邮箱:

登录 | 网盘

Baidu 百度

网页 搜你想搜

3

百度一下

今日油条回应被今日头条起诉

点击跳转到另一个网页

青岛此次疫情由患者共用CT室引发 **热** 广州新增1例无症状感染者 **热** 解放军军机15小时内两度巡台...  
成都大学党委书记遗体已被找到 **热** 浙江嘉兴公布新冠疫苗价格 第40个世界粮食日 **停**

hao123新闻 人民网 新华网 央视网 国际在线 中国日报 中国网 中经网 光明网 央广网 求是网 中青网 网信网 大连市政府

百度 腾讯 网易 百度地图 hao123影视 免费游戏 淘宝网 京东 苏宁易购  
携程旅行网 好看视频 爱奇艺 淘宝网 凤凰网 知网 哔哩哔哩 知乎 虎扑

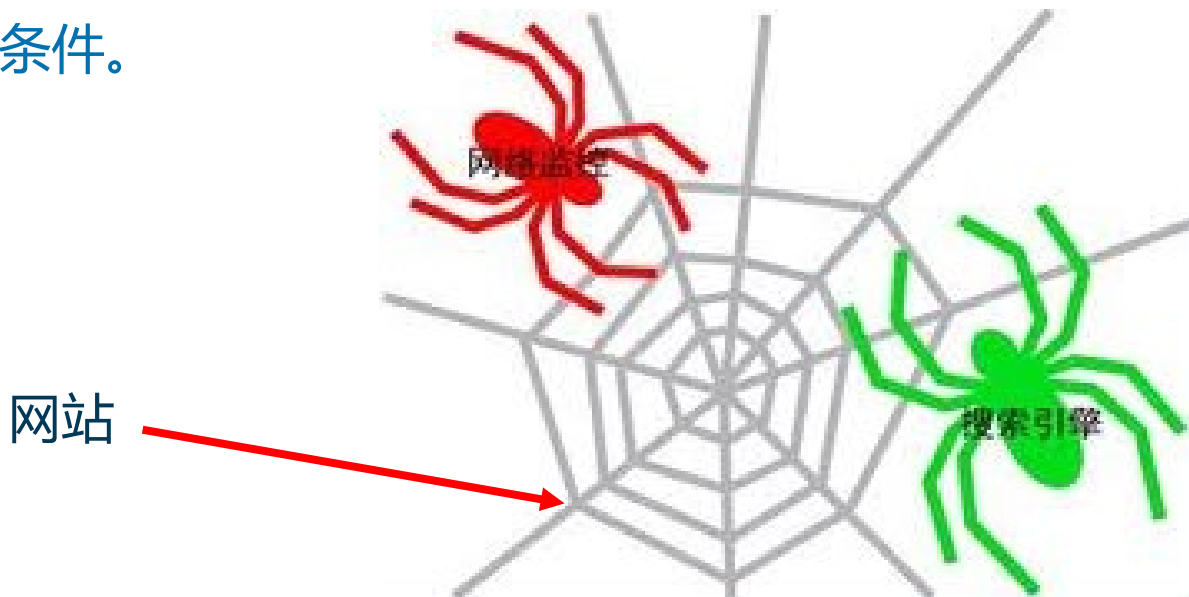
提示: 您的自定义网址存在丢失风险, 建议马上登录来保存。 同步网址到账号

58同城 唯品会 苏宁秒杀 天涯社区 AcFun弹幕 百度学术 学信网 虎扑社区 中国大学MOOC  
网易云音乐 12306 企查查 安居客房产 工商银行 人事考试网 QQ邮箱 豆瓣网 彩票·走势图

点击展开

```
data-status="1" title="好看视频"><div class="inline-block-wrapper"><a class="sitelink icon-site" href="
https://haokan.baidu.com/" style="background-image: url(
https://dgss0.bdstatic.com/5bVWsj_p_tVS5dKfpU_Y_D3/res/r/image/2019-07-01/3a117368c5bcfaf958ee74b0aec56287.png)"
data-title="好看视频">好看视频</a></div></li><li class="js_site-item site-item" data-id="4435044" data-title=
"爱奇艺" data-icon="
https://dgss1.bdstatic.com/5bVXsj_p_tVS5dKfpU_Y_D3/qiusuo_icon/24c7e207a280974a518b1290a25bce4e.png" data-status
="1" title="爱奇艺"><div class="inline-block-wrapper"><a class="sitelink icon-site" href="http://www.iqiyi.com/"
style="background-image: url(
https://dgss1.bdstatic.com/5bVXsj_p_tVS5dKfpU_Y_D3/qiusuo_icon/24c7e207a280974a518b1290a25bce4e.png)" data-title
="爱奇艺">爱奇艺</a></div></li><li class="js_site-item site-item" data-id="4435045" data-title="淘宝网"
data-icon="
https://dgss0.bdstatic.com/5bVWsj_p_tVS5dKfpU_Y_D3/res/r/image/2020-04-09/9a95d3783ba0e6dea8bd386e2d0ad67f.png"
data-status="1" title="淘宝网"><div class="inline-block-wrapper"><a class="sitelink icon-site" href=
https://www.taobao.com/" style="background-image: url(
https://dgss0.bdstatic.com/5bVWsj_p_tVS5dKfpU_Y_D3/res/r/image/2020-04-09/9a95d3783ba0e6dea8bd386e2d0ad67f.png)"
data-title="淘宝网">淘宝网</a></div></li><li class="js_site-item site-item" data-id="4435046" data-title=
"凤凰网" data-icon="
https://dgss0.bdstatic.com/5bVWsj_p_tVS5dKfpU_Y_D3/qiusuo_icon/d7fd0fcc2e428773bf1c105caa851de0.ico" data-status
="1" title="凤凰网"><div class="inline-block-wrapper"><a class="sitelink icon-site" href="https://www.ifeng.com/
" style="background-image: url(
https://dgss0.bdstatic.com/5bVWsj_p_tVS5dKfpU_Y_D3/qiusuo_icon/d7fd0fcc2e428773bf1c105caa851de0.ico)" data-title
="凤凰网">凤凰网</a></div></li><li class="js_site-item site-item" data-id="4435049" data-title="知网" data-icon=
"https://dgss0.bdstatic.com/5bVWsj_p_tVS5dKfpU_Y_D3/res/r/image/2020-02-11/636035c9a022d2c06b334e0cb27a5fcc.png"
```

- **网络爬虫**又称网络蜘蛛，网络机器人，robot, spider, crawler。
- 网络爬虫是一个**自动提取网页**的程序，它为搜索引擎从万维网上下载网页，是搜索引擎的重要组成。爬虫一般从一个或若干初始网页的URL开始，在抓取网页的过程中，不断从当前页面上抽取新的URL放入队列，直到满足系统的一定停止条件。

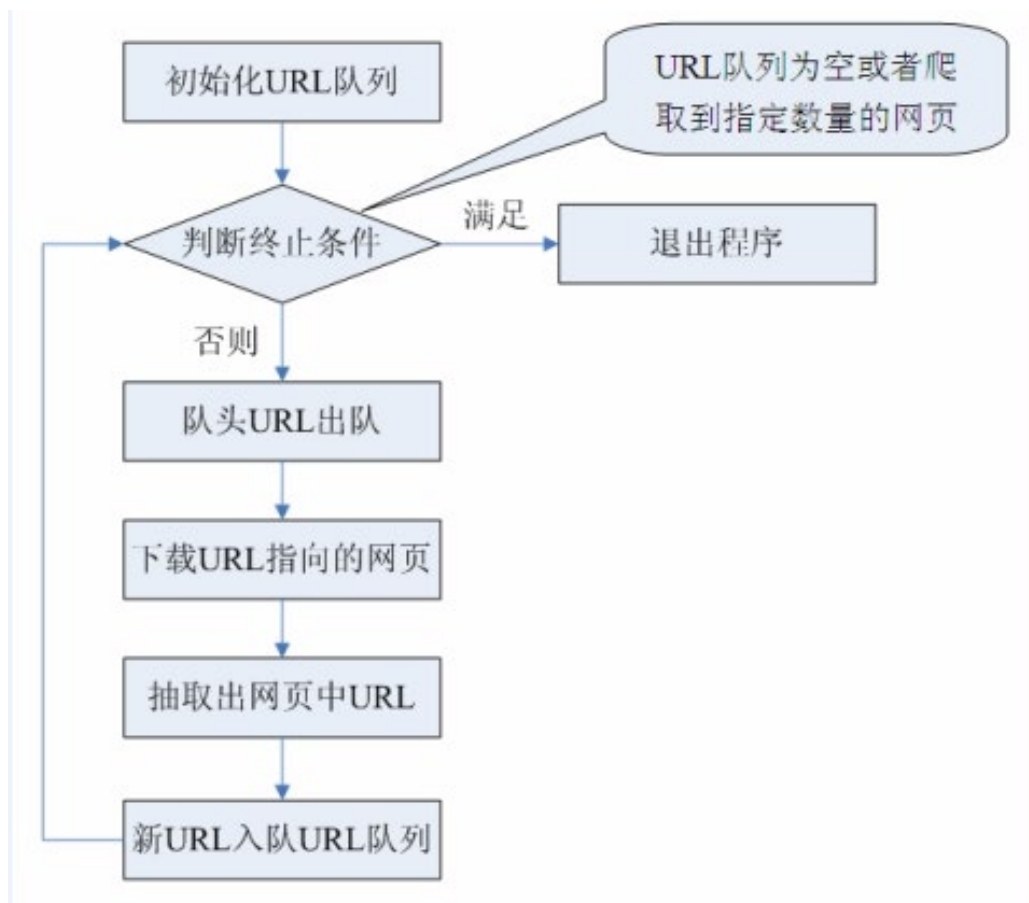


## ● 非定向爬虫

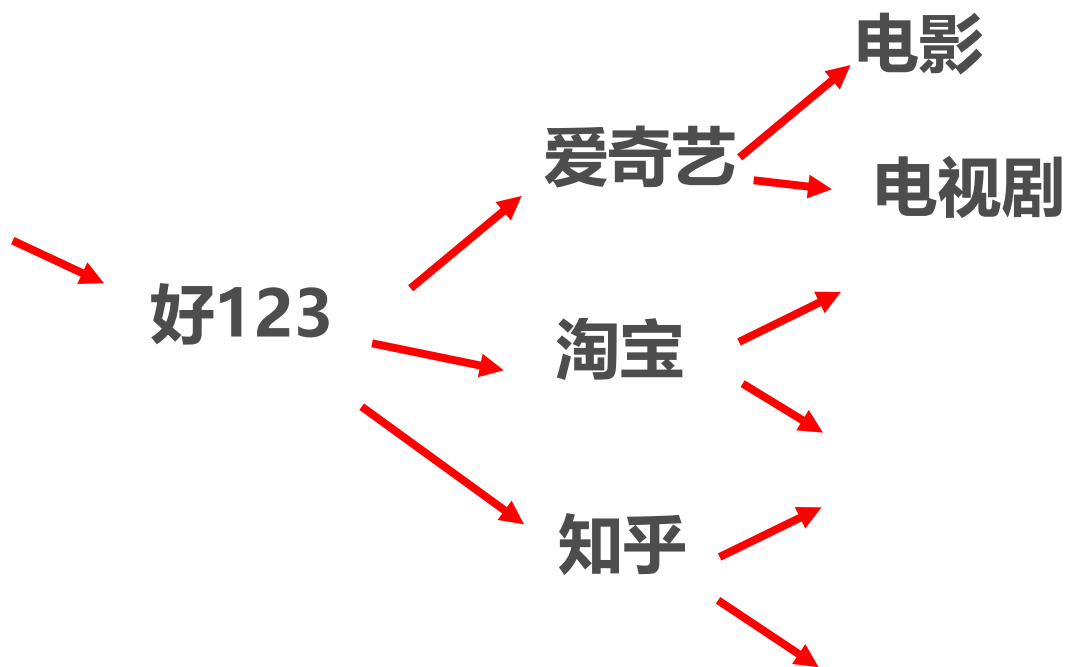
- ◆ 爬取互联网上任何基于Http协议的内容
- ◆ 工具：Larbin、WebMagic、Heritrix、Nutch、Scrapy...

## ● 定向爬虫

- ◆ 根据网站自身的属性采用特定的爬取策略
- ◆ 工具包：HttpClient (java) , requests (python)

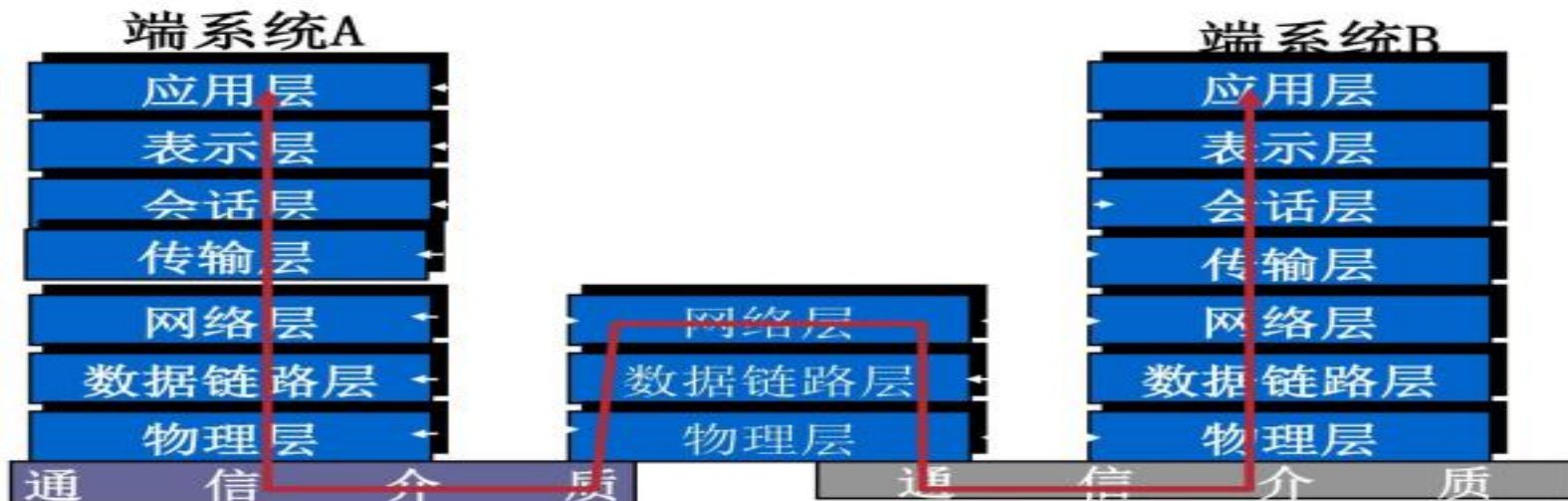


- 种子的选取
- 终止条件的设定
- 网页更新去重
- 爬取网页的策略
- 网页内容异步加载
- 多线程并发爬取
- .....





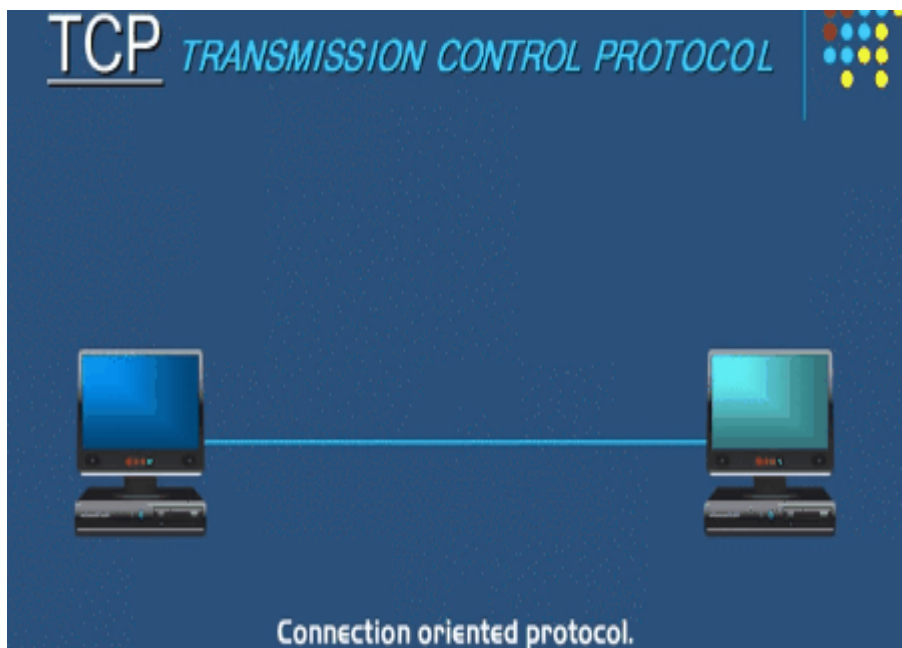
- HTTP: Hyper Text Transfer Protocol (超文本传输协议)
- 万维网协会和Internet工作小组, 1999年6月发布了RFC 2616, 定义了今天普遍使用的HTTP/1.1
- HTTP协议是用于从WWW服务器传输超文本到本地浏览器的传送协议, 属于应用层协议, 由请求和响应构成, 是一个标准的客户端服务器模型



- HTTP协议通常承载于TCP协议之上，有时也承载于TLS或SSL协议层之上（这就是所说的HTTPS）
- HTTPS = HTTP + SSL/TLS  <https://www.baidu.com>
- SSL (Secure Socket Layer, 安全套接字层): 为数据通讯提供安全支持
- TLS (Transport Layer Security, 传输层安全): SSL的升级版
- 默认HTTP端口为80, HTTPS端口为443
- HTTPS是HTTP协议的安全版本, HTTP协议的数据传输是明文的, 不安全的, HTTPS使用了SSL/TLS协议进行了加密处理 (谍战片里的密码本)
- TCP 与 UDP



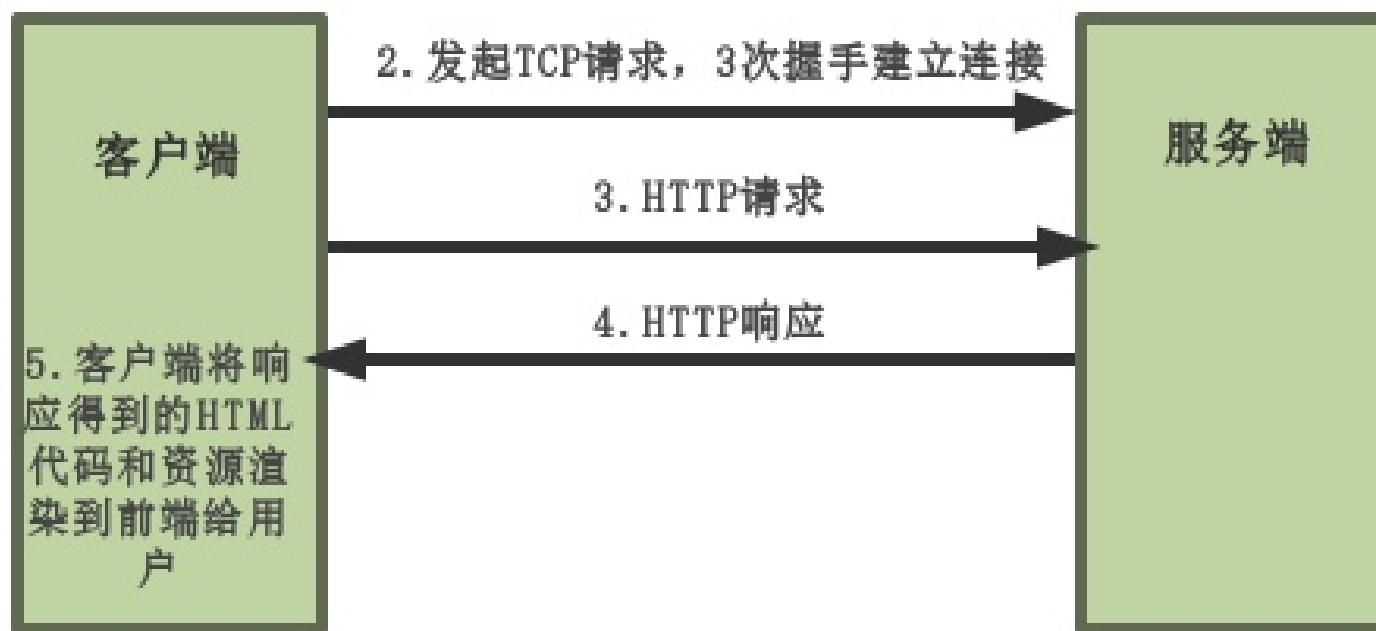
- TCP(传输控制协议), 面向连接的、一对一、可靠的;三次握手协议
- UDP(用户数据报协议), 面向无连接, 有单播, 多播, 广播的功能, UDP是面向报文的, 不可靠性



# HTTP访问示例



1. www.baidu.com DNS域名解析为服务器IP



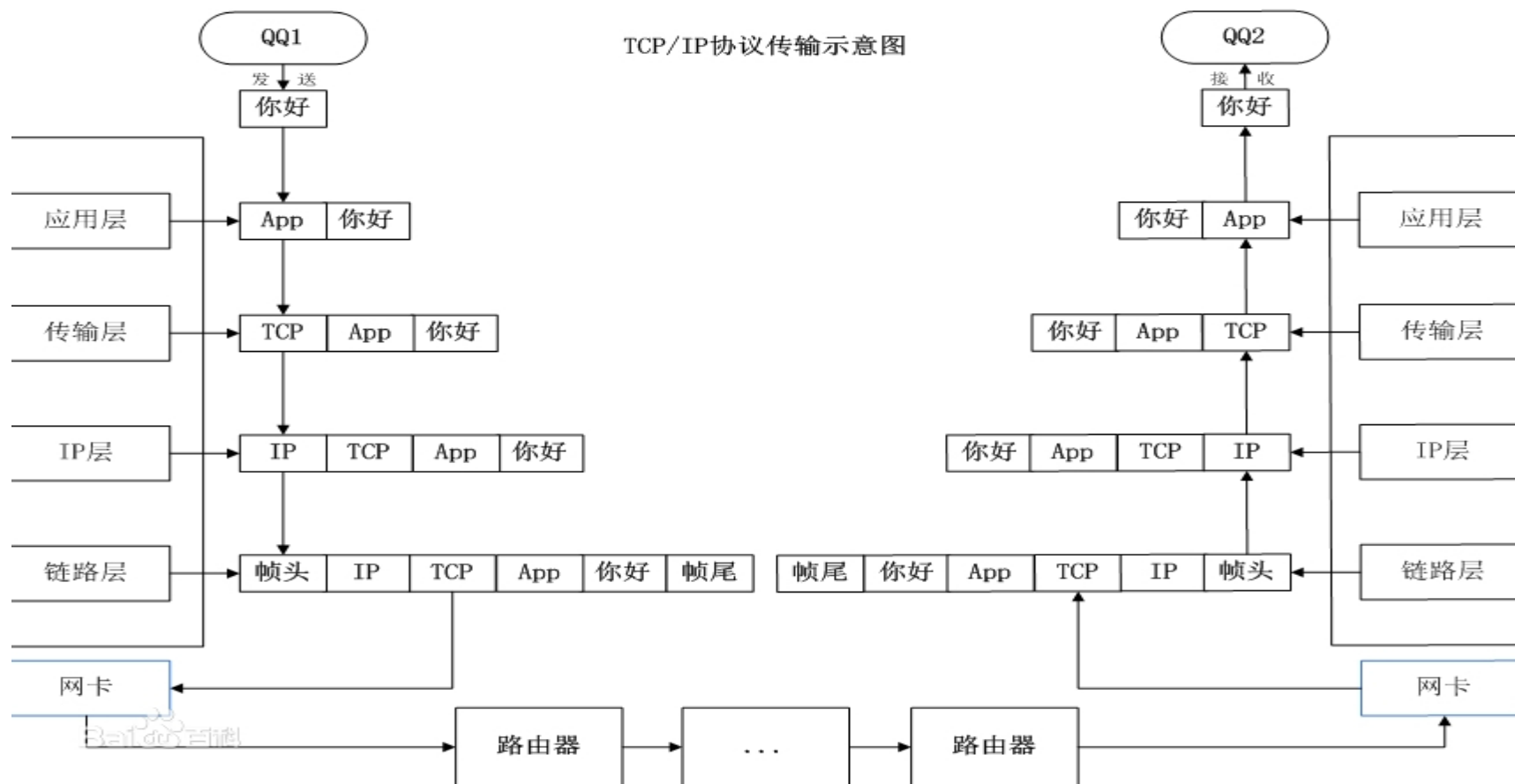
知乎 @文同

- (1) 首先客户端与服务器需要建立连接（只要单击某个超链接，HTTP的工作就开始了）
- (2) 建立连接后，客户机向服务器发送请求
- (3) 服务器接收到请求后，给予相应的相应信息
- (4) 客户端接受服务器所返回的信息通过浏览器显示在用户显示屏上，  
然后客户端与服务器断开连接

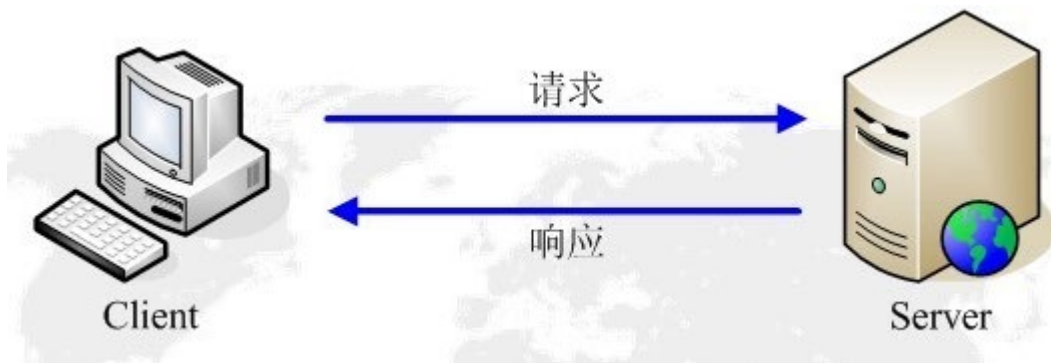
# TCP/IP协议传输示意图



TCP/IP协议传输示意图



- HTTP协议永远都是客户端发起请求，服务器回送响应（无法推送）
- HTTP协议是一个无状态的协议，同一个客户端的这次请求和上次请求没有对应关系，却不利于客户端与服务器保持会话连接
- 为了弥补这种不足，产生了两项记录http状态的技术，一个叫做Cookie，一个叫做Session



- Cookies是**客户端**保存状态的一种方案
  - ◆ 会话性质的cookie，存放在浏览器内存
  - ◆ 持久化的cookie，存放在硬盘上
- Cookies可以记录你的用户ID、密码、浏览过的网页、停留的时间等信息。当你再次来到该网站时，网站通过读取Cookies，得知你的相关信息，就可以做出相应的动作（如在页面显示欢迎你的标语，或者让你不用输入ID、密码就直接登录等等）
- **清除缓存快捷键** Ctrl+Shift+Delet

- Session机制是一种**服务器**端保存用户状态的机制，服务器使用一种类似于散列表的结构来保存信息。
- 客户端维护Session ID的方式
  - ◆ Cookie
  - ◆ URL重写
  - ◆ 表单隐藏字段

- HTTP请求由三个部分组成：**请求行、消息报头、请求正文**

**请求行**

POST https://dl.req.163.com/dl/1 HTTP/1.1  
Host: dl.req.163.com

**请求头**

Connection: keep-alive  
Content-Length: 411  
Origin: https://dl.req.163.com  
User-Agent: Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/75.0.3770.100 S  
Content-Type: application/json  
Accept: \*/\*  
Referer: https://dl.req.163.com/webzi/v1.0.1/pub/index\_d12\_new.html?cd=https%3A%2Fmimo.127.net%2Fep%2Ffreena  
Accept-Encoding: gzip, deflate, br  
Accept-Language: zh-CN,zh;q=0.9  
Cookie: 1555555555-NTXZDL-YD6aQ3IUXGfDI87Z6h6jsvhiI17hgZH2W5CB9dKnwj4rRrW5Cxp9Luvbgpy90GLUjijcyrjK/95552zmrAV

**请求正文**

"un": " ", "pw": "TP5JtoMqBdyV1veX3X7Anzhe00h5PjNngduk/fqAusMXGoLbySeNVC3nR1Ev"



# HTTP协议之请求 – 请求行



- 请求行: **Method Request-URI HTTP-Version CRLF**

例: GET /index.jsp HTTP/1.1 (CRLF)

方法	含义
----	----

GET	请求获取Request-URI所标识的资源
-----	-----------------------

POST	在Request-URI所标识的资源后附加新的数据
------	---------------------------

HEAD	类似于get请求, 但返回的响应中没有具体的内容, 用于获取响应消息报头
------	--------------------------------------

PUT	请求服务器存储一个资源, 并用Request-URI作为其标识
-----	---------------------------------

DELETE	请求服务器删除Request-URI所标识的资源
--------	--------------------------

TRACE	请求服务器回送收到的请求信息, 主要用于测试或诊断
-------	---------------------------

CONNECT	保留将来使用
---------	--------

OPTIONS	请求查询服务器的性能, 或者查询与资源相关的选项和需求
---------	-----------------------------

- GET:请求指定的页面信息，并返回实体主体。
- POST:向指定资源提交数据进行处理请求（例如提交表单或者上传文件）。数据被包含在请求体中
- GET 提交的信息直接拼在URL里，多用于查询，长度有限

`url='http://12. X. X. X:1/solr/patent/summary?keyword=' +keyword`

- POST数据被包含在请求体中，多用来提交（密码放入body中），长度原则上没有限制

# HTTP协议之请求消息



按照内容类型排列的 **Mime** 类型列表

类型/子类型	扩展名
application/envoy	evy
application/fractals	fif
application/futuresplash	spl
application/hta	hta
application/internet-property-stream	acx
application/mac-binhex40	hqx
application/msword	doc
application/msword	dot
application/octet-stream	*
application/octet-stream	bin
application/octet-stream	class
application/octet-stream	dms
application/octet-stream	exe
application/octet-stream	lha
application/octet-stream	lzh
application/oda	oda
application/olescript	axs
application/pdf	pdf
application/pics-rules	prf
application/pkcs10	p10
application/pkix-crl	crl
application/postscript	ai

- Accept: 浏览器可接受的媒体类型
- Accept-Charset: 浏览器可接受的字符集
- Accept-Encoding: 浏览器可接受的编码
- Accept-Language: 浏览器可接受的语言
- Authorization: 授权
- Connection: 表示请求的连接类型
- Content-Length: 内容长度
- Cookie: 这是最重要
- Host: 初始URL中的主机名
- Referer: 跳转前URL
- User-Agent: 浏览器标识

- HTTP响应由三个部分组成：**状态行**、**消息报头**、**响应正文**



The screenshot shows an HTTP response with three parts highlighted by red boxes and labeled with red text:

- 状态行 (Status Line):** HTTP/1.1 200 OK
- 响应头 (Response Header):** Server: nginx  
Date: Sun, 07 Jul 2019 16:22:38 GMT  
Content-Type: text/plain; charset=UTF-8  
Content-Length: 60  
Connection: keep-alive  
Set-Cookie: JSESSIONID=82E23ED94048E70E2D3258A33DC3378; Path=/contact163/; HttpOnly  
Content-Language: zh-CN
- 响应正文 (Response Body):** {"code":200,"data":{"contacts":[],"groups":[]},"msg":"S\_OK"}

- 状态行: **HTTP-Version Status-Code Reason-Phrase CRLF**

**例: HTTP/1.1 200 OK (CRLF)**

状态代码有三位数字组成，第一个数字定义了响应的类别，且有五种可能取值：

**1xx:** 指示信息--表示请求已接收，继续处理

**2xx:** 成功--表示请求已被成功接收、理解、接受

**3xx:** 重定向--要完成请求必须进行更进一步的操作

**4xx:** 客户端错误--请求有语法错误或请求无法实现

**5xx:** 服务器端错误--服务器未能实现合法的请求

- 常见状态代码、状态描述、说明：

- ◆ **200** OK //客户端请求成功

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，如URL错误，资源名拼写错误

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

# HTTP协议之响应 – 消息报头



- **Location:** 用于重定向接受者到一个新的位置
- **Server:** 服务器用来处理请求的软件信息

# HTTP相关知识点 – 压缩



- HTTP压缩是在传输过程中对数据进行压缩。
- HTTP压缩传输的文件。
- HTTP压缩采用GZIP或DEFLATE算法。
- 网页压缩情况

请输入要查询网址：

113.10.161.28/Manage/MainFrame.aspx

查询

网址 [113.10.161.28/Manage/MainFrame.aspx](http://113.10.161.28/Manage/MainFrame.aspx) 检测结果如下：

是否压缩	是
压缩类型	gzip
原始文件大小	2547 字节
压缩后文件大小	925 字节
压缩率（估计值）	63.68%

## Header信息

Cache-Control	private
Date	Wed, 20 Mar 2013 11:26:49 GMT
Content-Type	text/html; charset=utf-8
Server	Microsoft-IIS/6.0
X-Powered-By	ASP.NET
X-AspNet-Version	2.0.50727
Content-Encoding	gzip
Vary	Accept-Encoding
Transfer-Encoding	chunked



# HTTP相关知

- JSON 即 JavaS  
，非常适合于服
- JSON 是基于纯  
的，因此，JSO  
String, Numk  
Object 对象。

## JSON示例

```
{
  "statuses": [
    {
      "created_at": "Tue May 31 17:46:55 +0800 2011",
      "id": 11488058246,
      "text": "求关注。",
      "source": "<a href='http://weibo.com' rel='nofollow'>新浪微博</a>",
      "favorited": false,
      "truncated": false,
      "in_reply_to_status_id": "",
      "in_reply_to_user_id": "",
      "in_reply_to_screen_name": "",
      "geo": null,
      "mid": "5612814510546515491",
      "reposts_count": 8,
      "comments_count": 9,
      "annotations": [],
      "user": {
        "id": 1404376560,
        "screen_name": "zaku",
        "name": "zaku",
        "province": "11",
        "city": "5",
        "location": "北京 朝阳区",
        "description": "人生五十年，乃如梦如幻；有生斯有死，壮士复何憾。",
        "url": "http://blog.sina.com.cn/zaku",
        "profile_image_url": "http://tp1.sinaimg.cn/1404376560/50/0/1",
        "domain": "zaku",
        "gender": "m",
        "followers_count": 1204,
        "friends_count": 447,
        "statuses_count": 2908,
        "favourites_count": 0,
        "created_at": "Fri Aug 28 00:00:00 +0800 2009",
        "following": false,
        "allow_all_act_msg": false,
        "remark": "",
        "geo_enabled": true,
        "verified": false,
        "allow_all_comment": true,
        "avatar_large": "http://tp1.sinaimg.cn/1404376560/180/0/1",
        "verified_reason": "",
        "follow_me": false,
        "online_status": 0,
        "bi_followers_count": 215
      }
    },
    ...
  ],
  "previous_cursor": 0,
  "next_cursor": 11488013766,
  "total_number": 81655
}
```

// 暂未支持  
// 暂未支持

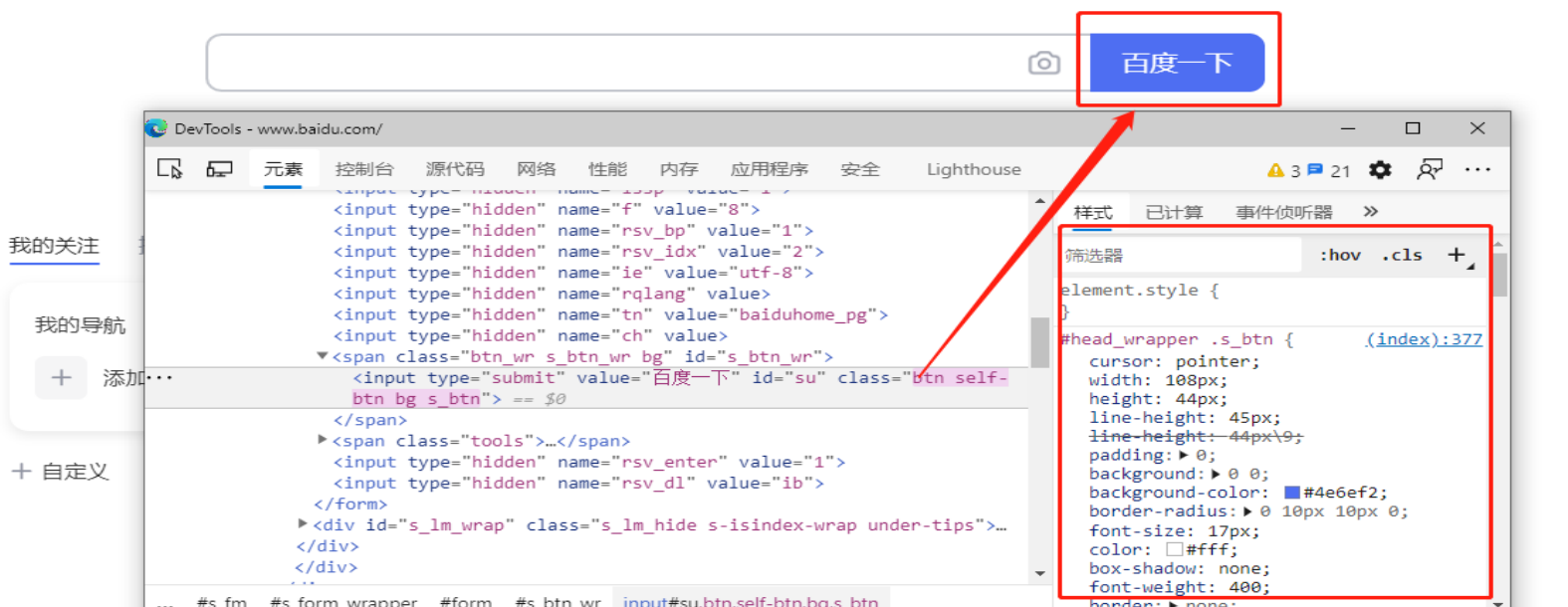
## ● HTML/xml/css/js

新闻 hao123 地图 视频 贴吧 学术 更多

大连 21°C 优 设置



Baidu 百度



The screenshot shows the Baidu homepage with the search bar and the '百度一下' button highlighted by a red box. Below the search bar, the DevTools 'Elements' panel is open, showing the HTML structure of the page. A red arrow points from the '百度一下' button to the corresponding HTML element in the DOM tree. The DOM tree shows the following structure:

```
<input type="hidden" name="f" value="8">
<input type="hidden" name="rsv_bp" value="1">
<input type="hidden" name="rsv_idx" value="2">
<input type="hidden" name="ie" value="utf-8">
<input type="hidden" name="rqlang" value=">
<input type="hidden" name="tn" value="baiduhome_pg">
<input type="hidden" name="ch" value=">
<span class="btn_wr s_btn_wr bg" id="s_btn_wr">
  <input type="submit" value="百度一下" id="su" class="btn self-
  btn bg s_btn">
</span>
<span class="tools">...</span>
<input type="hidden" name="rsv_enter" value="1">
<input type="hidden" name="rsv_dl" value="ib">
</form>
<div id="s_lm_wrap" class="s_lm_hide s-isindex-wrap under-tips">...

The 'Styles' panel on the right shows the CSS rules for the selected element, including the following styles:

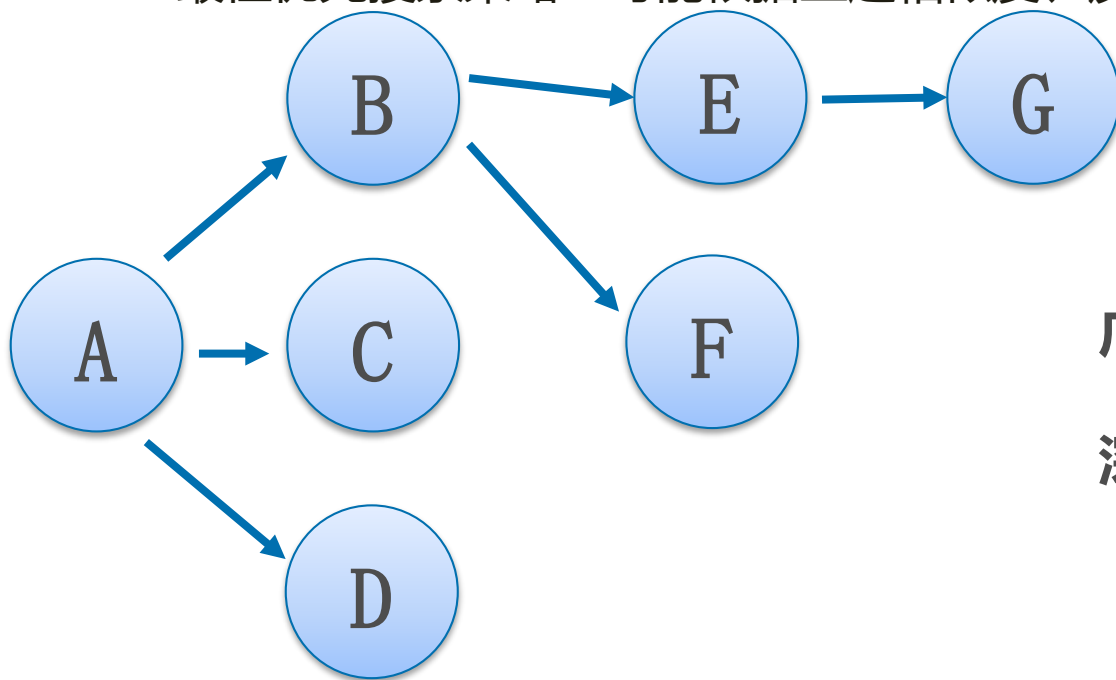


```
element.style {
}
#head_wrapper .s_btn {
  cursor: pointer;
  width: 108px;
  height: 44px;
  line-height: 45px;
  line-height: 44px\9;
  padding: 0;
  background: 0 0;
  background-color: #4e6ef2;
  border-radius: 0 10px 10px 0;
  font-size: 17px;
  color: #fff;
  box-shadow: none;
  font-weight: 400;
  border: none;
```


```

## ● 广度优先 vs. 深度优先 vs. 最佳优先搜索策略

- ◆ 广度优先：先采集完同一层的网页，再采集下一层网页
- ◆ 深度优先：先沿一条路径采到叶节点，再从同层其他路径进行采集
- ◆ 最佳优先搜索策略：可能根据主题相似度、反向链接数、PR值等策略



广度优先：ABCDEFGG -> 队列

深度优先：ABEGFCD -> 栈

## ● 网站采集 vs. 全局URL采集

- ◆ 网站采集：一个网站一个网站采集
- ◆ 全局URL采集：将所有URL放入一个URL池，从中使用某种方法进行选择
- ◆ 网站采集在采集效率上可能不如全局URL采集，通常的搜索引擎采用全局URL采集的方法。

## ● 孤立站点

- ◆ 用户提交

- **正则表达式**

- ◆ 可以过滤非正规的网址、无需下载的文件（后缀名）或特定域名下的网页

- **建立IP规则库**

- ◆ 如若建立校内搜索引擎，则在爬取时将所有非校内IP过滤掉

## ● 批量搜集

- ◆ 每次搜集替换上一次的内容

## ● 增量搜集

- ◆ 开始时搜集一批
- ◆ 往后：1、搜集新出现的网页；2、搜集在上次搜集后有改变的网页；3、删除上次搜集后不存在的网页

## ● 比较：

- ◆ 定期批量重采非常简单，但是浪费带宽，周期也长；
- ◆ 增量采集可以节省带宽，网页更新周期相对较短，但是系统的复杂性增大。

- **历史参考策略**

- ◆ 据页面以往的历史更新数据，预测该页面未来何时会发生变化。

- **用户体验策略**

- ◆ 根据用户点击信息优先爬取质量较高/关注度高的页面

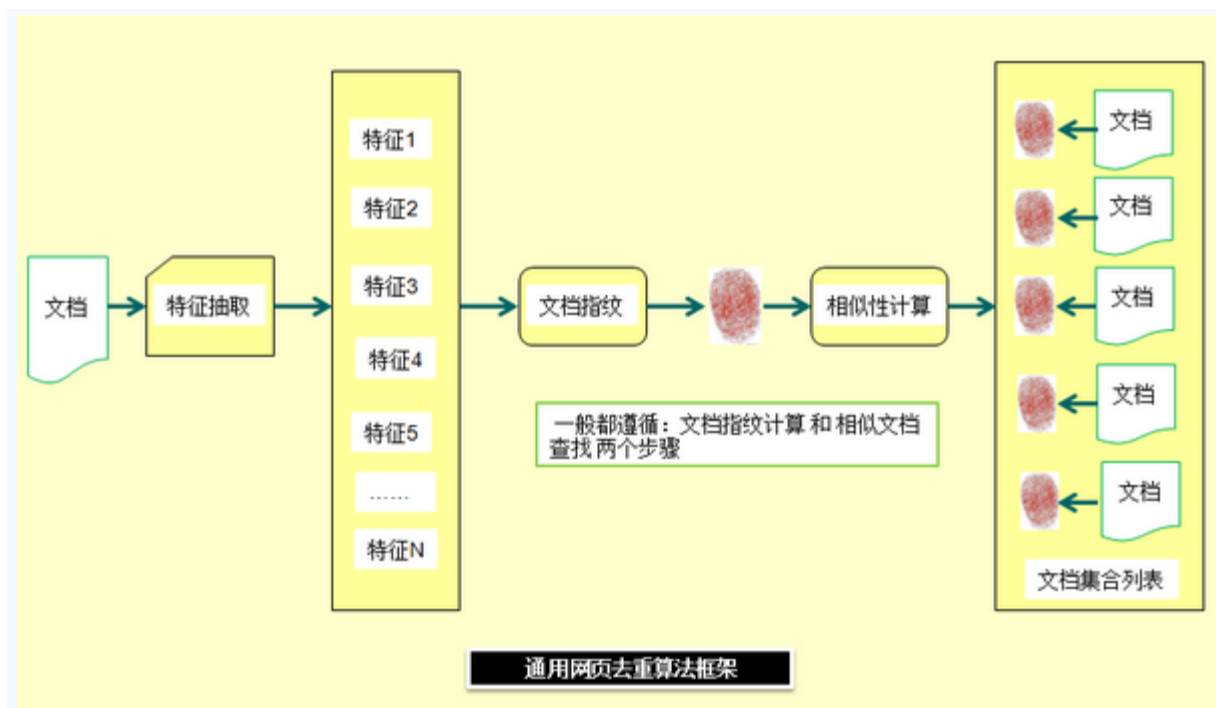
- **聚类抽样策略**

- ◆ 无需保存历史信息，解决冷启动问题（无历史信息的网页）

- MD5值比较法

- ◆ 缺点：精确匹配才算重复

- 网页指纹法





## ● 主要内容抽取

- ◆ TIKA, 可抽取HTML, PDF, MS-\*, Image(元数据), XML等
- ◆ Lucene提供工具包抽取HTML (较粗糙, 容易出错)
- ◆ cx-extractor, 基于行块分布函数的通用网页正文抽取算法 (哈工大)

## ● 特定内容抽取

```
Lexer lexer = new Lexer(html);  
Parser parser = new Parser(lexer);  
  
NodeFilter divFilter = new AndFilter(new TagNameFilter("div"), new HasAttributeFilter("id", "photo"));  
  
NodeList divNodes = parser.Parse(divFilter);
```

## ● 多机分布式并行

- ◆ 局域网联接多机进行并行采集
- ◆ 广域网分布式采集

## ● 单机多程序并行

- ◆ 多进程并行
- ◆ 多线程并行

## ● 多线程中主要问题

- ◆ 网络带宽
- ◆ 服务器对爬虫请求频率的限制
- ◆ 异常处理（多次爬取、日志记录）

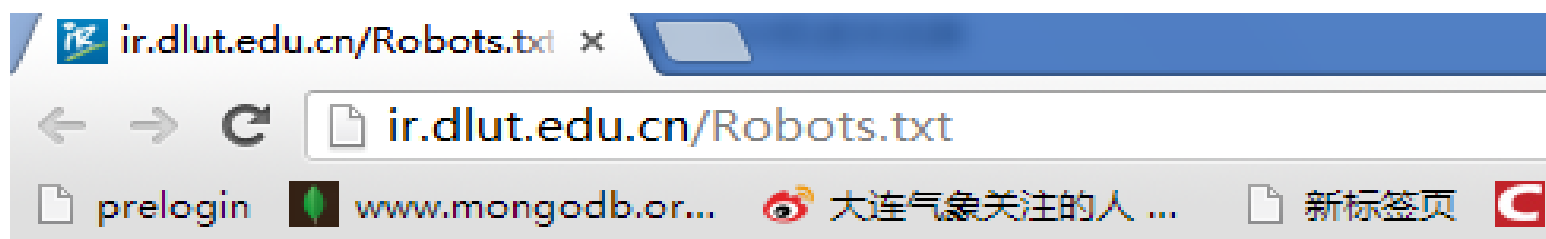
## ● 减少重复页面的采集

- ◆ URL重复的检测和排除
- ◆ 内容重复的检测和排除

## ● 保证重要页面的高优先级

- ◆ 入度高的网页相对重要 指向该网站的网站多，  
社交圈，1w人，有9000个是其好友，这个人很重要
- ◆ URL浅的网页相对重要  
好123

- robots.txt (统一小写) 是一种存放于网站根目录下的ASCII编码的文本文件, 它通常告诉网络蜘蛛, 此网站中的哪些内容是不应被搜索引擎的漫游器获取的, 哪些是可以被获取的。



```
User-agent: *  
Disallow: /App_Code/  
Disallow: /App_Data/  
Disallow: /SiteManager/  
Disallow: /UserControls/
```

- **全功能的网页爬虫**
- **性质: MIT Licence**
- **网址: <http://weblech.sourceforge.net/>**
- **版本: V0.0.3(2002-06-14)**
- **平台: Java**
- **特点:**
  1. 代码是用纯Java写的, 可以在任何支持Java的平台上均可运行
  2. 支持多线程下载网页
  3. 可维持网页间的链接信息
  4. 可配置性强

## ● 使用方法:

- ◆ 1) 按需求修改配置文件Spider.properties
- ◆ 2) 运行run.bat开始爬行
- ◆ 3) 如果程序中断, 运行resume.bat继续爬行

- **saveRootDirectory = c:/weblech/sistes** 设置文件的存放路径，默认为当前文件夹
- **mailtoLogFile = mailto.txt** 设置邮件链接(mailto links)的存放文件
- **refreshHTMLs = true**
- **refreshImages = false**
- **refreshOthers = false** //设置如果本地硬盘已经存在待爬取的文件，是否重新载入文件
- **htmlExtensions = htm,html,shtm,shtml,asp,jsp,php** 设置spider要下载资源的扩展名
- **imageExtensions = 同上**
- **startLocation = http://ir.dlut.edu.cn/** 设置spider爬行的起始页面
- **depthFirst = false** 设置进行广度优先爬行或深度优先爬行
- **maxDepth = 5** 爬行的最大深度（第一个页面深度为0，其链接的深度为1）
- **urlMatch =** 基本的URL过滤。下载的网页的网址中必须包括urlMatch串
- **interestingURLs= google.cn** 优先级最高的网页
- **boringURLs=baidu.com** 优先级最低的网页
- **basicAuthUser = myUser**
- **basicAuthPassword = 1234** 设置需要验证的网站的用户名和密码
- **spiderThreads = 15** 爬行的线程数
- **checkpointInterval = 30000** 设置写断点的时间间隔（ms）





## ● 主要类

- ◆ SpiderConfig.java
- ◆ HTMLParser.java
- ◆ DownloadQueue.java
- ◆ URLGetter.java
- ◆ URLObject.java
- ◆ Spider.java

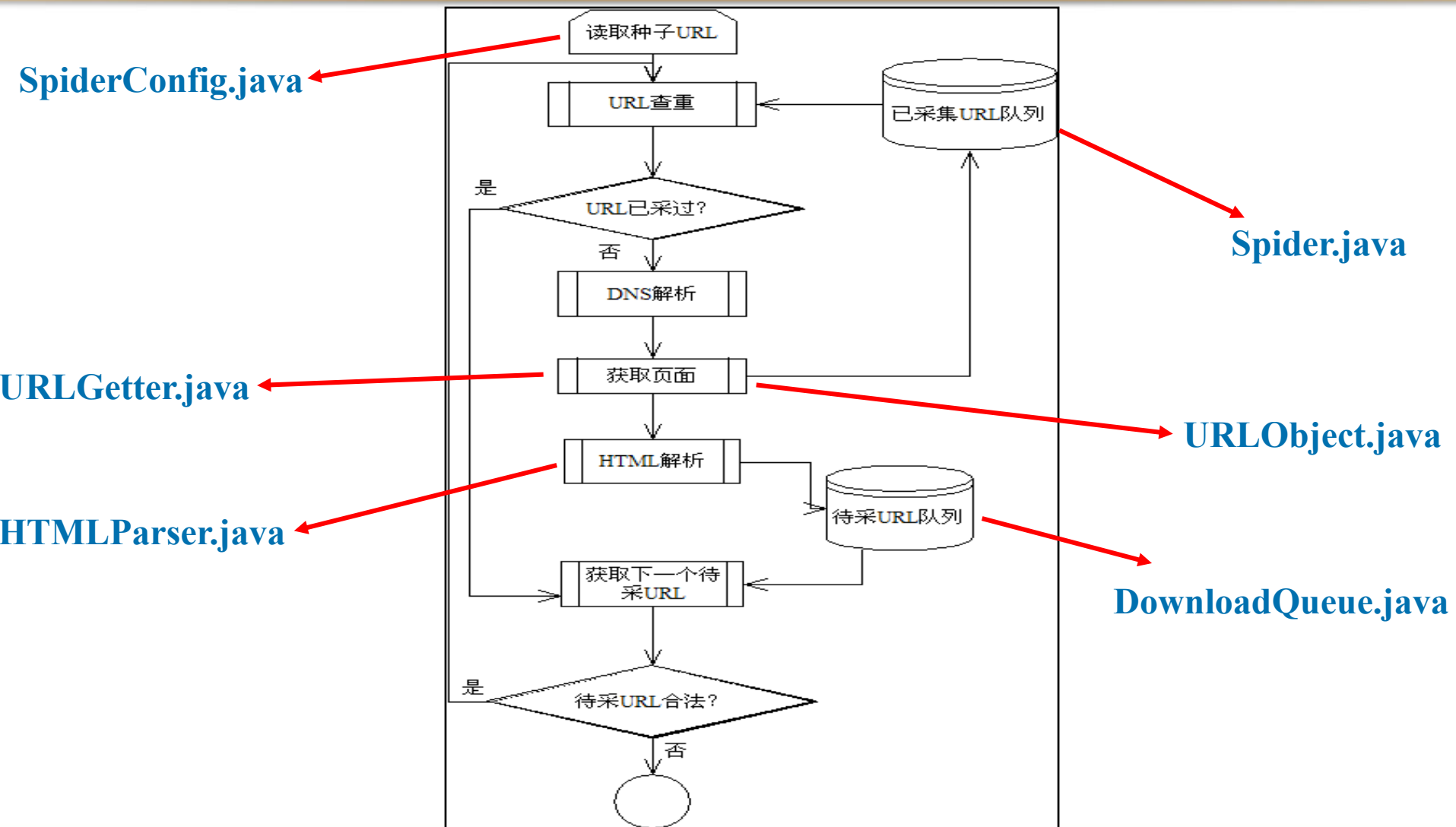
# Weblech——流程图



大连理工大学  
信息检索研究室



Information Retrieval Laboratory of DUT



## ● 如何写一个简单的Spider程序？

- ◆ 1、种子URL队列 (List 1) ；
- ◆ 2、Http协议，获得url的内容content，url → URL完成队列 (List 2) ；
- ◆ 3、存储content；
- ◆ 4、解析content → URLs；
- ◆ 5、判断URLs是否在List 2中，如果否，加入到List 1中；
- ◆ 6、如果List 1非空，执行2；
- ◆ 7、如果List 1为空，完成。

## ● 要求：

- ◆ 利用Spider程序，完成网页爬取工作。
  - 1、开发一个spider程序，多线程，支持断点续爬。
  - 2、利用开源工具。

- Web信息的搜集
- **基于Lucene的索引与检索**
- **提供Web服务**

- 索引基本原理
- 检索模型
- Lucene简单介绍

- **为加快搜索速度，建立特定的数据结构（类似字典）**
  - ◆ 不可能是逐个文档扫描(太慢)
  - ◆ 反向索引、B+树、散列索引等
- **大规模数据的索引常常用反向索引（倒排）**
  - ◆ Inverted file
  - ◆ 所有的搜索引擎都用倒排索引
  - ◆ 速度快

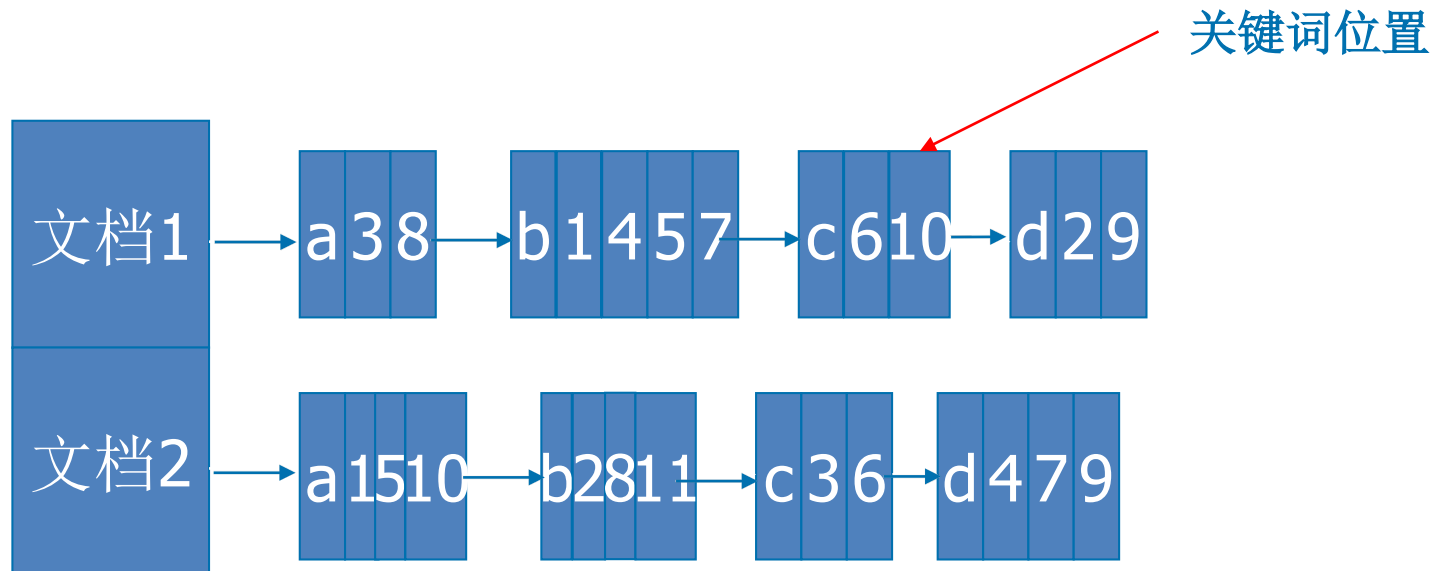
# 前向索引(Forward index)



- 概念：由文档到关键词

文档1: b d a b b c b a d c

文档2: a b c d a c d b d a b

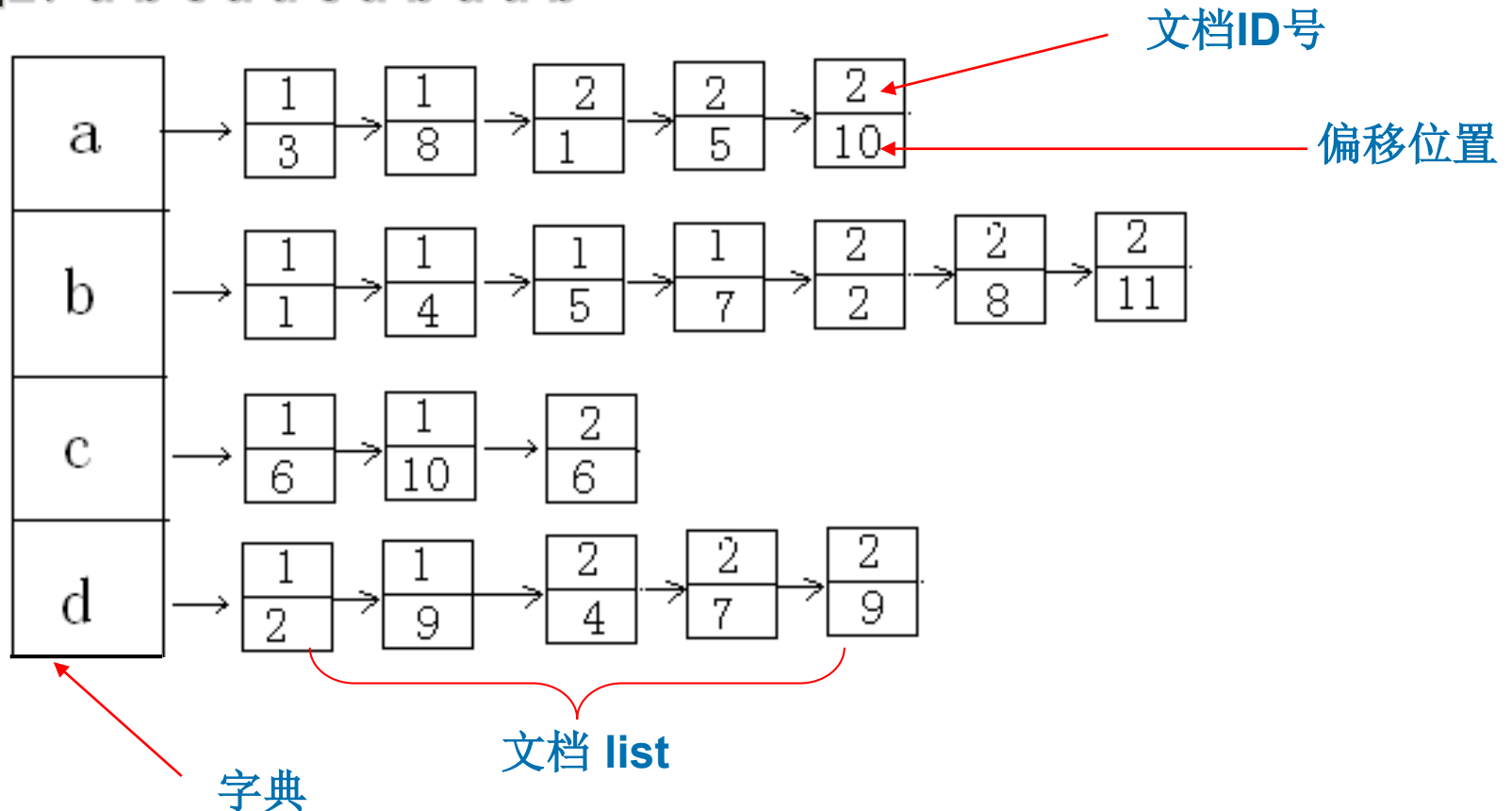




# 反向索引(Inverted index)



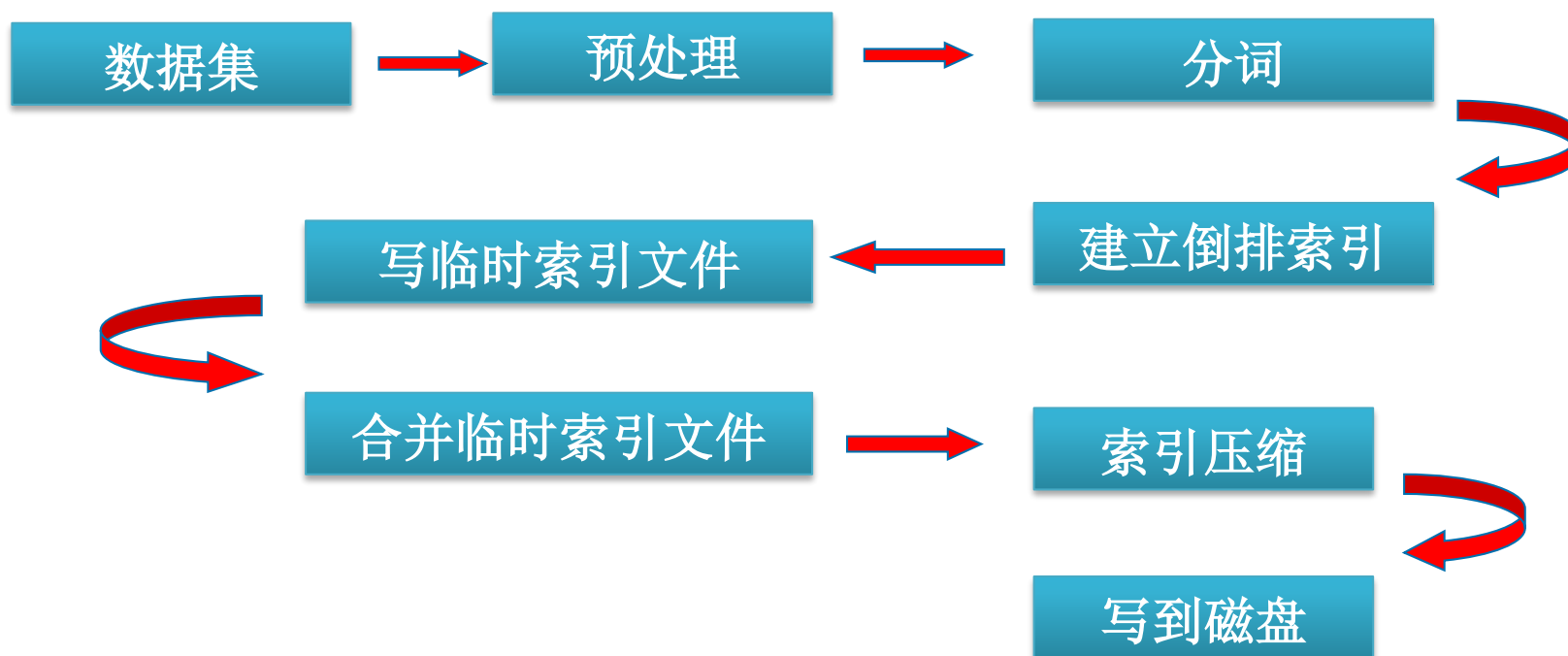
- 文档1: b d a b b c b a d c
- 文档2: a b c d a c d b d a b



## ● 实际应用中:

- ◆ 对字典排序
- ◆ 把字典读入内存
- ◆ 如果字典太大，对字典建立二级索引，把字典的索引读入内存

# 建立倒排索引的流程



## ● 对于中文，分词的作用实际上是要找出一个个的索引单位

◆ 例子：李明天天都准时上班

## ● 索引单位

◆ 字：李/明/天/天/都/准/时/上/班

- 索引量太大，查全率百分百，但是查准率低；
- 比如，查“明天”这句话也会出来

◆ 词：李明/天天/都/准时/上班

- 索引量大大降低，查准率较高，查全率不是百分百，而且还会受分词错误的影响；
- 比如，上面可能会切分成：李 明天 天都 准时 上班

◆ 二字串：李明/明天/天天/天都/都准/准时/时上/上班

- **进行词根还原: stop/stops/stopping/stopped → stop**
  - ◆ 好处: 减少词典量, 提高查全率;
  - ◆ 坏处: 按词形查不到, 词根还原还可能出现错误
- **不进行词根还原:**
  - ◆ 好处: 支持词形查询;
  - ◆ 坏处: 增加词典量
- **开源工具包:**
  - ◆ Snowball

## ● 停用词(Stop words):

- ◆ 指那些出现频率高但是无重要意义；通常不会作为查询词出现的词，如“的”、“地”、“得”、“都”、“the”等等
  - 消除：通常是通过查表的方式去除，
    - 好处----大大减少索引量，
    - 坏处----有些平时的停用词在某些上下文可能有意义
  - 保留：索引空间很大

## ● 什么叫检索?

- ◆ 用户提交一个查询 (Query) , 查找与该查询相关结果的过程。
- ◆ E.g. 关键词->多个文档, 涉及文档排序,
- ◆ 百度, 一个关键词, 返回很多结果, 如何排序

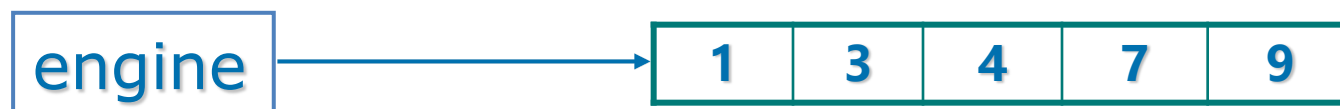
## ● 检索模型:

- ◆ 布尔模型
- ◆ 向量空间模型
- ◆ 概率模型
- ◆ 统计语言模型
- ◆ .....



- 简单的检索模型，建立在集合论和布尔代数的基础上。
- 遵循两条基本规则：
  - ◆ 每个索引词在一篇文档中只有两种状态：出现或不出现，对应权值为 0或1。
  - ◆ 查询是由三种布尔逻辑运算符 and, or, not 连接索引词组成的布尔表达式。
- 优点：
  - ◆ 简单，易于实现，能够保证较高的查全率。
- 缺点：
  - ◆ 只能精确判断文档是否出现某一查询词，但并没有给出每个词的重要程度，不能给出相关性排序

# 布尔模型



## ● 查询和文档都转化成标引项(Term)及其权重组成的向量表示

- ◆ 康奈尔大学 Salton 1970年代提出并倡导，原型系统SMART
- ◆ 例如：
  - 文档1: ( $\langle 2006, 1 \rangle, \langle \text{世界杯}, 3 \rangle, \langle \text{德国}, 1 \rangle, \langle \text{举行}, 1 \rangle$ ),
  - 文档2: ( $\langle 2002, 1 \rangle, \langle \text{世界杯}, 2 \rangle, \langle \text{韩国}, 1 \rangle, \langle \text{日本}, 1 \rangle$ )
- ◆ 查询: ( $\langle 2006, 1 \rangle, \langle \text{世界杯}, 1 \rangle$ )
- ◆ 查询和文档进行向量的相似度计算：夹角余弦或者内积
  - 文档1:  $1 * 1 + 3 * 1 = 4$
  - 文档2:  $0 * 1 + 2 * 1 = 2$
- ◆ 优点：简洁直观，效果好，可以应用到很多其他领域。
- ◆ 缺点：理论上不够完善，标引项之间的独立性假设与实际不符

## ● 权重影响因子TF (Term Frequency) :

- ◆ 局部，一个文档内
- ◆ 词频：即一个单词在一个文档中出现的次数。一般在文档中反复出现的单词往往反映主题，故一个单词的出现频率越高，相应权值越高。
- ◆ 最简单的形式：词频
- ◆ 变体： $W_{tf} = 1 + \log(tf)$  数字1用于平滑，log机制用于抑制过大差异

## ● 权重影响因子DF (Document Frequency)

- ◆ 全局，所有文档
- ◆ 词的文档频率，DF越高区分度越低，包含该词的文档数/文档总数
- ◆ 比如的，the/a，几乎在每个文档都会出现，文档频率很高，则区分度低，代表性差，

## ● 权重影响因子IDF (Inverse DF)

- ◆ 逆文档频率，DF的“倒数”
- ◆ 衡量不同单词对文档的区分能力。反映了一个特征词在整个文档集合中的分布情况，特征词出现过的文档数目越多，IDF值越低，这个词区分不同文档的能力越差。
- ◆  $IDF = \log \left( \frac{N}{n} \right)$  N代表文档集合中的文档总数，n代表特征词在其中多少文档中出现过

## ● 权重影响因子TF-IDF:

- ◆  $\text{Weight}(\text{word}) = \text{TF} * \text{IDF}$
- ◆ 考虑某一word在本文档和所有文档当中的出现情况
- ◆ 越大越能代表这一文档

文档1: ( $\langle 2006, w_1 \rangle, \langle \text{世界杯}, w_2 \rangle, \langle \text{德国}, w_3 \rangle, \langle \text{举行}, w_4 \rangle$ ),

文档2: ( $\langle 2002, w_1 \rangle, \langle \text{世界杯}, w_2 \rangle, \langle \text{韩国}, w_3 \rangle, \langle \text{日本}, w_4 \rangle$ )

- 对用户的查询进行扩充：比如用户输入“计算机”，我们扩充一个词“电脑”
- 同义词扩展：
  - ◆ 同义词词典
  - ◆ 通过统计构造的同义词词典
- 相关词扩展：
  - ◆ 相关词：“2006世界杯”与“德国”
  - ◆ 基于全局分析的查询扩展：对文档集合进行分析得到某种相关词典
  - ◆ 基于局部上下文的查询扩展
  - ◆ 基于概念的查询扩展
- 查询重构：对用户的初始查询进行修改(可以是加词、减词，或者对于向量模型表示的初始查询进行权重的修改等等)，是比查询扩展更泛的一个概念



## ● Lucene简介

- ◆ 完整、高效、易用、易扩展的开源全文检索**工具包**
- ◆ 性质: Apache License
- ◆ 作者: Doug Cutting
- ◆ 网址: <http://lucene.apache.org/>
- ◆ 版本: Lucene 7.5.0
- ◆ 平台: 跨平台

# Lucene的其他语言版本

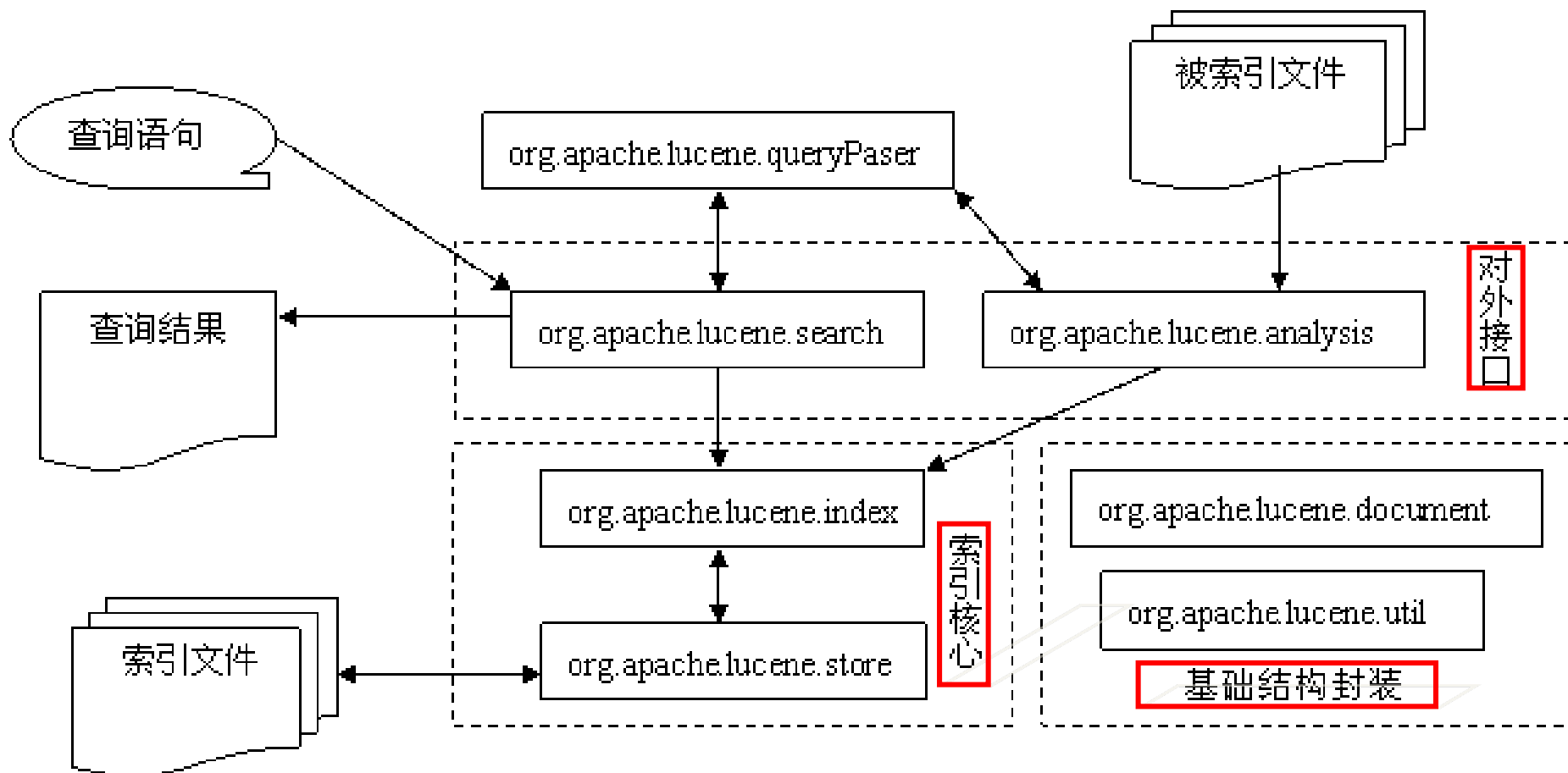


- [CLucene](#) - Lucene implementation in C++
- [dotLucene](#) - Lucene implementation in .NET
- [Lucene4c](#) - Lucene implementation in C
- [LuceneKit](#) - Lucene implementation in Objective-C (Cocoa/GNUstep support)
- [Lupy](#) - Lucene implementation in Python
- [NLucene](#) - another Lucene implementation in .NET (out of date)
- [Zend Search](#) - Lucene implementation in the Zend Framework for PHP
- [Plucene](#) - Lucene implementation in Perl
- [KinoSearch](#) - a new Lucene implementation in Perl
- [PyLucene](#) - GCJ-compiled version of Java Lucene integrated with Python via SWIG
- [MUTIS](#) - Lucene implementation in Delphi
- [Ferret](#) - Lucene implementation in Ruby

- 建立索引|150GB/小时，普通PC
- 增量索引的速度和批量索引相当
- 索引大小约为索引文本的20%-30%
- 100G左右索引查询ms级
- 最优秀的源码之一，建议阅读源码

- 结果排序 - 相关性最好结果优先
- 强大的查询表达式处理功能 - 短语、通配符、模糊查询等
- 分字段检索
- 指定日期范围检索
- 根据字段排序
- 支持多索引检索与结果合并
- 支持更新与检索同时进行

# Lucene系统的组织结构



- **segments**: 存储索引的各个segment的信息
- **.del**: 已删除文档信息
- **.fnm**: 域信息 (域名域标志等)
- **.fdt**: 域数据, 存储文档的各种属性数据, 例如文档路径, 文档长度 (按文档标号顺序组织)
- **.fdx**: 文档域数据指针, 每文档一个
- **.tis**: 索引词 (term) 信息, 即词典
- **.tii**: 存储.tis中每IndexInternal个Term, 这个文件装入内存以加快检索速度(二级索引)
- **.frq**: 存放索引词 (term) 的词频信息
- **.prx**: 索引词 (term) 的位置信息
- 其它文件



- **Elasticsearch**
- **Solr**
- **Lucene**
- **whoosh (python)**
- **Ferret**

- Web信息的搜集
- 基于Lucene的索引与检索
- **提供Web服务**



## ● 用户查询请求处理程序

- ◆ Python、Java、ASP.NET、PHP等
- ◆ 功能
  - (1) 获取用户查询式：把用户输入的查询关键词发送给检索服务器
  - (2) 显示结果：从检索服务器获取结果，分页呈现给用户

## ● 例：

- ◆ Java: Tomcat + JSP(或spring mvc)
- ◆ Python: Flask/**Django**/Tornado
- ◆ Vue.js、AngularJS、React

## ● 参考

### ◆ django 教程

[https://www.cnblogs.com/Leo\\_wl/p/5824541.html#\\_labelTop](https://www.cnblogs.com/Leo_wl/p/5824541.html#_labelTop)

### ◆ django创建新的项目，总体逻辑

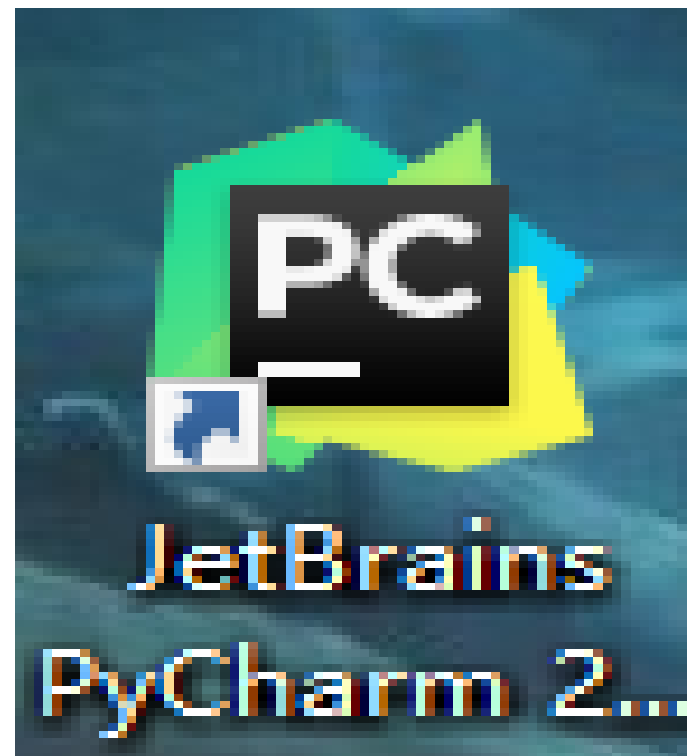
◆ [https://blog.csdn.net/tang\\_jin2015/article/details/81193943](https://blog.csdn.net/tang_jin2015/article/details/81193943)

### ◆ 将数据输出到页面内

◆ <https://www.cnblogs.com/ybf-yyj/p/8120286.html>

## ● 工具&环境:

- ◆ windows
- ◆ IDE: pycharm(专业版!!!)
- ◆ python3.5.1
- ◆ Django



# 例子1用户查询界面



## ● Web浏览器界面

### ◆ 输入查询表达式（结构化）

搜索结果	包含以下 <b>全部</b> 的字词	<input type="text"/>	10 项结果 ▼	搜索
	包含以下 <b>任何</b> 一个字词	<input type="text"/>		
	<b>不包括</b> 以下字词	<input type="text"/>		

### ◆ 输入查询表达式（自然语言形式）

DUTBIO

<input type="text" value="mad cow disease"/>	搜一下
<input checked="" type="radio"/> 中英混合 <input type="radio"/> 中文 <input type="radio"/> 英文	

# 例子1结果显示界面



大连理工大学  
信息检索研究室



Information Retrieval Laboratory of DUT

## DUTBIO



[医学搜索](#) [交互扩展](#) [医学问答](#) [关系抽取](#)

mad cow disease

搜一下

☒ 中英混合 ☐ 中文 ☐ 英文

查询翻译

搜索: mad cow disease 结果数: 118189 当前结果: 21 ~ 28

[第一页](#) [上一页](#) [下一页](#) [最后一页](#)

### **疯牛病**免疫组织化学诊断试剂盒

作者: 王志亮

出处: 中国技术专家网

该课题为**疯牛病**免疫组织化学试剂盒的研究。该试剂盒是利用抗原抗体的免疫反应和ABC显色放大系统的原理研制而成, 它的核心试剂是PrP单克隆抗体4C11。**疯牛病**的免疫组织化学实验中发现, 该试剂盒的阳性检出率

### **疯牛病**免疫组织化学诊断试剂盒

作者: 王志亮

出处: 中国技术专家网

该课题为**疯牛病**免疫组织化学试剂盒的研究。该试剂盒是利用抗原抗体的免疫反应和ABC显色放大系统的原理研制而成, 它的核心试剂是PrP单克隆抗体4C11。**疯牛病**的免疫组织化学实验中发现, 该试剂盒的阳性检出率

### **Prion diseases: a typical Kuhnian abnormality in a molecular paradigm.**

Authors: Silvestri G|Baldassarre F Department of Biochemistry and Biotechnology, Federico II University, Naples, Italy.

FAU - Silvestri, G

Source: Med Hypotheses

As a new class of pathogens with unusual properties, prions have been implied in several spongiform encephalopathies mainly affecting farm animals (scrapie, **mad-cow disease**) and humans (kuru, Creutzfeldt-Jacob **disease**, fatal familial insomnia) (1).

### **[Prion diseases in men]**

Authors: Keohane C Department of Pathology (Neuropathology), Cork Regional Hospital, Wilton, Eire. FAU -

Keohane, C

Source: Arch Anat Cytol Pathol

[疯牛病]



# 例子1结果显示界面



大连理工大学  
信息检索研究室



Information Retrieval Laboratory of DUT

## DUTBIO



医学搜索 交互扩展 医学问答 关系抽取

mad cow disease

搜一下

☒ 中英混合 ☐ 中文 ☐ 英文

查询翻译

搜索: mad cow disease 结果数: 118189 当前结果: 21 ~ 28

[第一页](#) [上一页](#) [下一页](#) [最后一页](#)

### **疯牛病**免疫组织化学诊断试剂盒

作者: 王志亮

出处: 中国技术专家网

该课题为**疯牛病**免疫组织化学试剂盒的研究。该试剂盒是利用抗原抗体的免疫反应和ABC显色放大系统的原理研制而成, 它的核心试剂是PrP单克隆抗体4C11。**疯牛病**的免疫组织化学实验中发现, 该试剂盒的阳性检出率

### **疯牛病**免疫组织化学诊断试剂盒

作者: 王志亮

出处: 中国技术专家网

该课题为**疯牛病**免疫组织化学试剂盒的研究。该试剂盒是利用抗原抗体的免疫反应和ABC显色放大系统的原理研制而成, 它的核心试剂是PrP单克隆抗体4C11。**疯牛病**的免疫组织化学实验中发现, 该试剂盒的阳性检出率

### **Prion diseases: a typical Kuhnian abnormality in a molecular paradigm.**

Authors: Silvestri G|Baldassarre F Department of Biochemistry and Biotechnology, Federico II University, Naples, Italy.

FAU - Silvestri, G

Source: Med Hypotheses

As a new class of pathogens with unusual properties, prions have been implied in several spongiform encephalopathies mainly affecting farm animals (scrapie, **mad-cow disease**) and humans (kuru, Creutzfeldt-Jacob **disease**, fatal familial insomnia) (1).

### **[Prion diseases in men]**

Authors: Keohane C Department of Pathology (Neuropathology), Cork Regional Hospital, Wilton, Eire. FAU -

Keohane, C

Source: Arch Anat Cytol Pathol

[疯牛病]

# 例子2专利检索系统



大连理工大学  
信息检索研究室



Information Retrieval Laboratory of DLUT

思利

请输入文本或公开号进行快捷检索

检索

>>> 转至高级检索

# 例子2专利检索系统



大连理工大学  
信息检索研究室



Information Retrieval Laboratory of DUT

思利

机器人

检索



## 按域检索

### 申请人

华为技术有限公司 (38165)  
三星电子株式会社 (33957)  
中兴通讯股份有限公司  
(28369)  
浙江大学 (24873)  
松下电器产业株式会社  
(24777)  
清华大学 (18771)  
上海交通大学 (16280)  
哈尔滨工业大学 (15891)  
高通股份有限公司 (15859)  
LG电子株式会社 (15706)  
罗伯特·博世有限公司  
(15125)  
三菱电机株式会社 (15109)  
丰田自动车株式会社 (15028)  
皇家飞利浦电子股份有限公  
司 (14719)

## 范围检索

### 公开日/公告日

更早 (1357860)

2011-01-01 ~ 2011-12-31  
(277346)

2012-01-01 ~ 2012-12-31  
(425618)

主题分析

热点预测

统计分析

6233223 results found in 1161ms Page 1 of 311,662

专利公开号	专利名	申请人	公告日	
CN109129555A	一种六轴机器人手臂	王东哲; 张焜	2019-01-04	<a href="#">详情</a>
CN109129555A	一种六轴机器人手臂	王东哲; 张焜	2019-01-04	<a href="#">详情</a>
CN109129555A	一种六轴机器人手臂	王东哲; 张焜	2019-01-04	<a href="#">详情</a>
CN106915390A	四点支撑平衡机构	肖伟	2017-07-04	<a href="#">详情</a>
CN105773568A	机器人刹车系统	肖伟	2016-07-20	<a href="#">详情</a>
CN102681856A	基于示教器的机器人人机界面的实现系统及实现方法	上海发那科机器人有限公司	2012-09-19	<a href="#">详情</a>
CN102681856A	基于示教器的机器人人机界面的实现系统及实现方法	上海发那科机器人有限公司	2012-09-19	<a href="#">详情</a>
CN105564521A	一种履带式机器人	江苏优创数控设备有限公司	2016-05-11	<a href="#">详情</a>
CN105759640A	省流量机器人控制方法	重庆友武科技有限公司	2016-07-13	<a href="#">详情</a>
CN108654025A	一种智能乒乓球捡发球子母机器人及其工作方法	李昱伟	2018-10-16	<a href="#">详情</a>
CN106909139A	基于触觉的避障机器人	袁祖六	2017-06-30	<a href="#">详情</a>
CN105773639A	机器人应随方法	肖伟	2016-07-20	<a href="#">详情</a>
CN105771277A	机器人跟随系统	肖伟	2016-07-20	<a href="#">详情</a>
CN105759635A	机器人广范围交互方法	重庆友武科技有限公司	2016-07-13	<a href="#">详情</a>
CN108068908A	机器人脚板结构和人形机器人	深圳市优必选科技有限公司	2018-05-25	<a href="#">详情</a>



## 思利

### · 主要著作信息

专利名：一种六轴机器人手臂

专利公开(公告)号：CN109129555A

专利公开(公告)日：Fri Jan 04 00:00:00 UTC 2019

**摘要：**本发明公开了一种六轴机器人手臂，包括机器人一轴，机器人一轴的下方设有机器人一轴外壳，机器人一轴的上方设有机器人一轴盖板，机器人一轴的侧面设有信号电缆插座和动力电缆插座，机器人一轴盖板上设有机器人二轴，机器人二轴包括设置在底部的机器人二轴电机盒，机器人二轴电机盒的两侧设有机器人二轴力臂板和机器人二轴外壳，机器人二轴力臂板和机器人二轴外壳从内向外依次排列，机器人二轴电机盒上设有机器人三轴，机器人三轴与机器人二轴之间设有机器人三轴力臂，机器人三轴通过机器人三轴力臂与机器人二轴连接，机器人三轴力臂外侧设有机器人三轴力臂板和机器人三轴外壳，本实用的性能优良、动作灵敏，适宜在社会上推广使用。

**权利要求：**1.一种六轴机器人手臂，其特征在于：包括机器人一轴(1)，所述机器人一轴(1)的下方设有机器人一轴外壳(2)，所述机器人一轴(1)的上方设有机器人一轴盖板(3)，所述机器人一轴(1)的侧面设有信号电缆插座(4)和动力电缆插座(5)，所述机器人一轴盖板(3)上设有机器人二轴(10)，所述机器人二轴(10)包括设置在底部的机器人二轴电机盒(6)，所述机器人二轴电机盒(6)的两侧设有机器人二轴力臂板(7)和机器人二轴外壳(8)，所述机器人二轴力臂板(7)和机器人二轴外壳(8)从内向外依次排列，所述机器人二轴电机盒(6)上设有机器人三轴(9)，所述机器人三轴(9)与机器人二轴(10)之间设有机器人三轴力臂(11)，所述机器人三轴(9)通过机器人三轴力臂(11)与机器人二轴(10)连接，所述机器人三轴力臂(11)外侧设有机器人三轴力臂板(20)和机器人三轴外壳(21)，所述机器人三轴(9)上设有机器人三四轴转接件(12)，所述机器人三四轴转接件(12)的上方设有机器人四轴输出法兰(13)，所述机器人四轴输出法兰(13)上设有机器人五轴(14)，所述机器人五轴(14)上设有机器人五轴电机盒(15)，所述机器人五轴电机盒(15)的两侧依次设有机器人五轴力臂板(16)和机器人五轴外壳(17)，所述机器人五轴(14)的顶部设有机器人五六轴转接件(18)，所述机器人五六轴转接件(18)上设有机器人六轴输出法兰(19)。

专利名——翻译：Six-axis robot arm

**摘要——翻译：**The invention discloses a six-axis robot arm. The six-axis robot arm comprises a first robot shaft. A first robot shaft shell is arranged below the first robot shaft, a cover plate of the first robot shaft is arranged above the first robot shaft, a signal cable socket and a power cable socket are arranged on the side surface of the first robot shaft, the cover plate of the first robot shaft is provided with a second robot shaft, and the second robot shaft comprises a motor box of the second robot shaft arranged at the bottom. A force arm plate of the second robot shaft and a shell of the second robot shaft are arranged on two sides of the motor box of the second robot shaft. The force arm plate of the second robot shaft and the shell of the second robot shaft are sequentially arranged from inside to outside, the motor box of the second robot shaft is provided with a third robot shaft. A force arm of the third robot shaft is arranged between the third robot shaft and the second robot shaft. The third robot shaft is connected with the second robot shaft through the force arm of the third robot shaft. A force arm plate of the third robot shaft and a shell of the third robot shaft are arranged on the outer side of the force arm of the third robot shaft. The six-axis robot arm has excellent performance and sensitive action, and is suitable for popularization and use in society.

专利申请号：CN201811044616.3

专利申请日：Fri Sep 07 00:00:00 UTC 2018

申请人：王东哲、张煜

### · 分类信息

主分类号：B25J18/00

IPC：B25J18/00; B25J17/02; B25J9/06

国民经济分类：C34

# 例子2-echarts



## 专利趋势分析

- 申请趋势
- 公开趋势
- 申请-公开趋势

## 专利技术分析

- 技术构成
- 技术申请趋势
- 技术公开趋势
- 技术中国省市分布
- 中国专利类型
- 国民经济构成

## 申请人分析

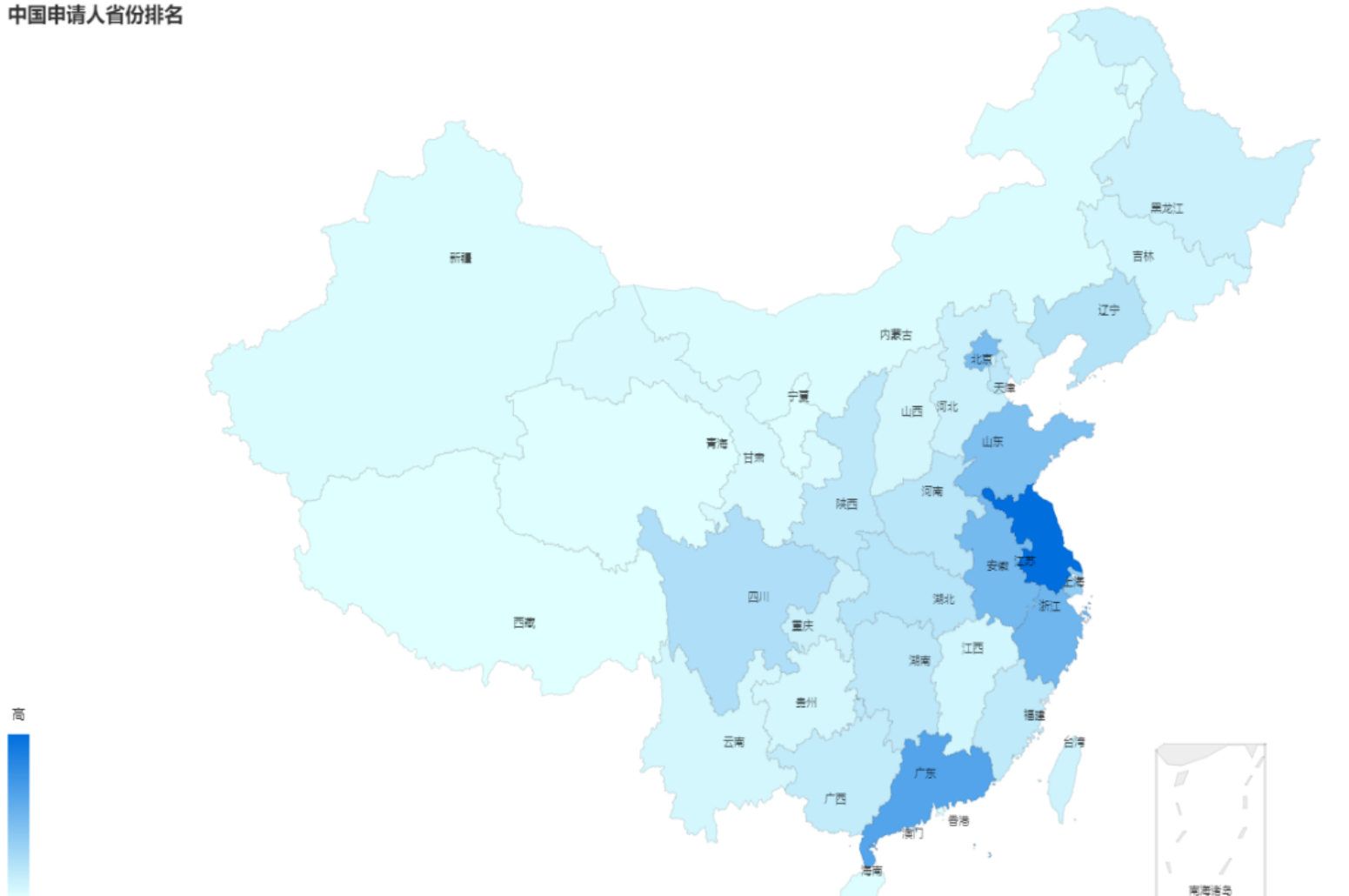
- 申请人排名
- 申请人国别
- 申请人构成
- 中国申请人省份排名

## 专利代理分析

- 代理人排名
- 代理机构排名

## 法律状态分析

中国申请人省份排名



## ● 动手搭建一个完整、可运行的小型全文检索实验系统

1. 爬取一定数目的网页
2. 预处理工作
3. 对数据建立倒排索引
4. 提供Web形式的检索服务
5. 尽量自己动手实现，少调用现成的模块化工具。哪怕逻辑不是很复杂，但是是我们自己做的！

- 需要提交**可运行的“工程代码”**和一份**“实验报告”**
  - ◆ “工程代码”部分：占实验成绩**40%**
  - ◆ “实验报告”部分：占实验成绩**60%**
- **实验分组：3人一组（至多）；不要使用订书器装订** 可使用回形针
  - ◆ 每组提交一份“工程代码”；
  - ◆ **每人都要提交一份实验报告。**
    - 写清同组人员
    - 写清本人所负责的工作

## ● 截止时间:

- ◆ 2022年12月31日21:00

## ● 提交方式:

- ◆ 以小组为单位, 将“工程代码” + “报告” 以附件形式一起发送;
- ◆ 发送到 [dutir\\_irtm@163.com](mailto:dutir_irtm@163.com)
- ◆ 邮件主题: 信息检索作业\_学号\_姓名\_学号\_姓名\_学号\_姓名
- ◆ 如信息检索作业\_20200001\_大明\_20200002\_小明\_20200003\_小小明
- ◆ 电子版“实验报告”的文件名: 学号\_姓名
- ◆ 打印版“实验报告”交到: 创新园大厦A0923室
- ◆ 有指示标志, 别交错位置了

- **报告不规整**

- ◆ 未写同组人员，未阐述各自工作
- ◆ 字体变换多端，缩进千变万化
- ◆ . . .

- **内容单薄，没有描述出所做工作**

- **实验工作胡编乱造**

- **报告中大篇幅与实验工作无关的内容**

- . . .

## 搜索引擎与文本挖掘实验报告

学 院（系）： \_\_\_\_\_  
专       业： \_\_\_\_\_  
学 生 姓 名： \_\_\_\_\_  
学       号： \_\_\_\_\_  
班       级： \_\_\_\_\_  
同 组 成 员： \_\_\_\_\_  
时       间： \_\_\_\_\_  
联 系 方 式： \_\_\_\_\_

**留下你的联系方式**

## ● 报告内容

- ◆ 实验目标
- ◆ 相关原理和工具
- ◆ 开发环境和运行环境
- ◆ 工作分工
- ◆ 总结（收获与感想）

## ● 一些限定

- ◆ 正文字体宋体、小四，一倍行间距
- ◆ 至少给出**一幅系统图片**（检索系统结果展示页）
- ◆ **尽量给出一些核心代码**（自己写的）
- ◆ 电子版页数不限；纸质版除封皮外**5页**（**卷的是能力，不是表明功夫**）





# 谢谢!