第8章 物联网搜索引擎

张 策

嵌入式系统研发中心
山东省嵌入式系统工程技术研发中心

第8章 物联网搜索引擎

本章内容

一、概述

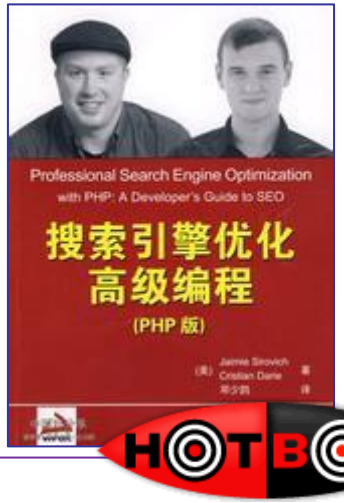
- 提供“**普适性的数据分析与服务**”的**搜索引擎**才能诠释出物联网“更深入的智能化”的内涵。

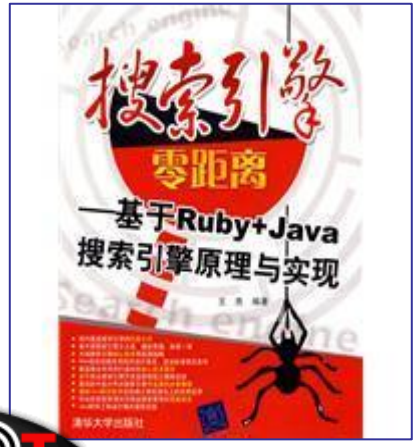
1. Web搜索引擎

- 一个能够在合理响应时间内，根据用户的查询关键词，返回一个包含相关信息的结果列表（hits list）服务的综合体。

- 传统的Web搜索引擎是基于查询关键词的，对于相同的关键词，会得到相同的查询结果。
- 常见Web搜索引擎：

bing
bing
altavista
Ask.com
有道youdao
即刻
必应bing Beta
Baidu百度
Google谷歌
SOSO搜搜
YAHOO! 中国雅虎
Sogou搜狗





HOTBOT

一、概述

2. 搜索引擎的发展

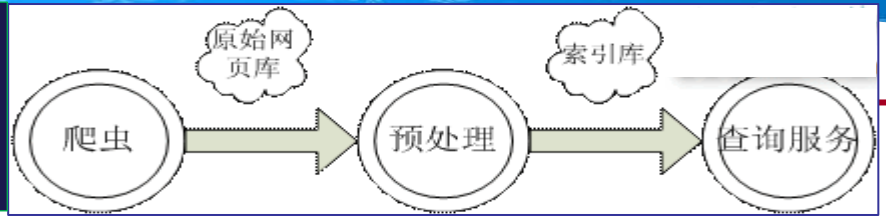
- 搜索引擎的起源可追溯到1992年，由NCSA维护的“What’s NEW!”页面。
- 第一个原始搜索引擎W3Catalog（1993.9）
- 第一个Web机器人程序“World Wide Web Wanderer”（1993.6 MIT）
- 里程碑：WebCrawler（1994），Lycos（1994）商用
- Google的建立(1998)：斯坦福博士生Larry Page(拉里·佩奇)和Sergey Brin(谢尔盖·布林)创立了Google

1993-2010 Web搜索引擎一览表	
1993	W3Catalog Aliweb JumpStation
1994	WebCrawler Infoseek Lycos
1995	AltaVista Open Text Web Index Magellan Excite SAPO
1996	Dogpile Inktomi HotBot Ask Jeeves
1997	Northern light Yandex
1998	Google
1999	AlltheWeb GenieKnows Naver Teoma Vivisimo
2000	Baidu Exalead
2001 | 2007	Info.Com Yahoo!Search A9.com Sogou MSNSearch Ask.com GoodSearch SearchMe WikiSeek Quaero LiveSearch ChaCha Guruji.com Sproose WiKiaSearch Blackle.com
2008	Powerset Picollator Viewzi Cuil Boogami LeapFish Forestle VADLO Sperse!Search DuckDuckGo
2009	Bing Yebol Mugurdy Goby
2010	YandexGlobal

一、概述

3. Web搜索引擎的结构

搜索引擎三段式工作流程



❶**网络爬虫模块**：主要功能是通过对Web页面的解析，根据Web页面之间的连接关系抓取这些页面，并储存页面信息交给索引模块处理。

前一部分得到的数据结果为后一部分提供原始数据。

❶爬虫

❷**索引模块**：主要完成对于抓取的数据进行预处理建立关键字索引以便搜索模块输出。

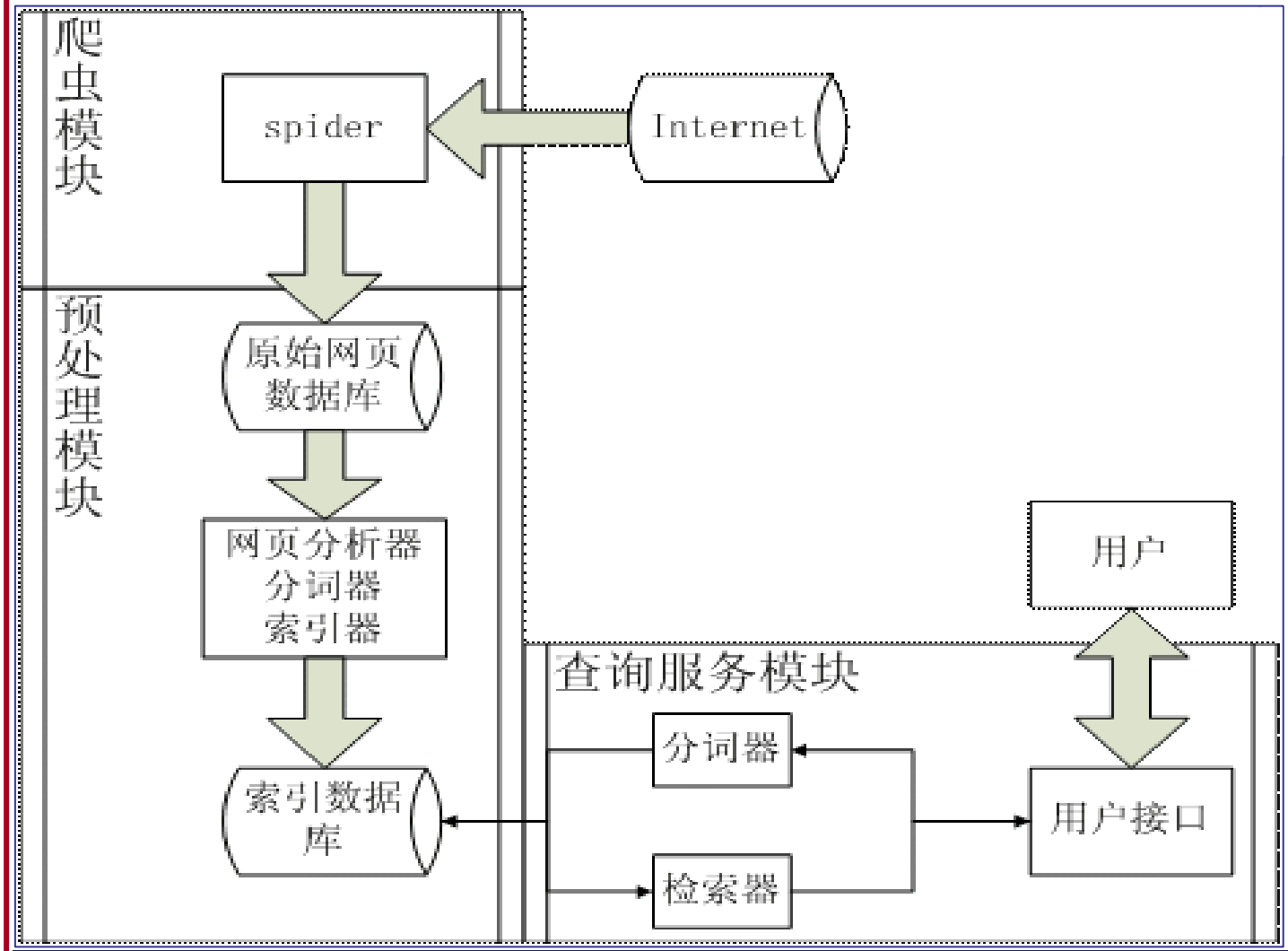
❷索引

❸搜索

❸**搜索模块**：对于用户的关键词，根据数据库的索引知识给出合理的搜索结果。

一、概述

3. Web搜索引擎的结构

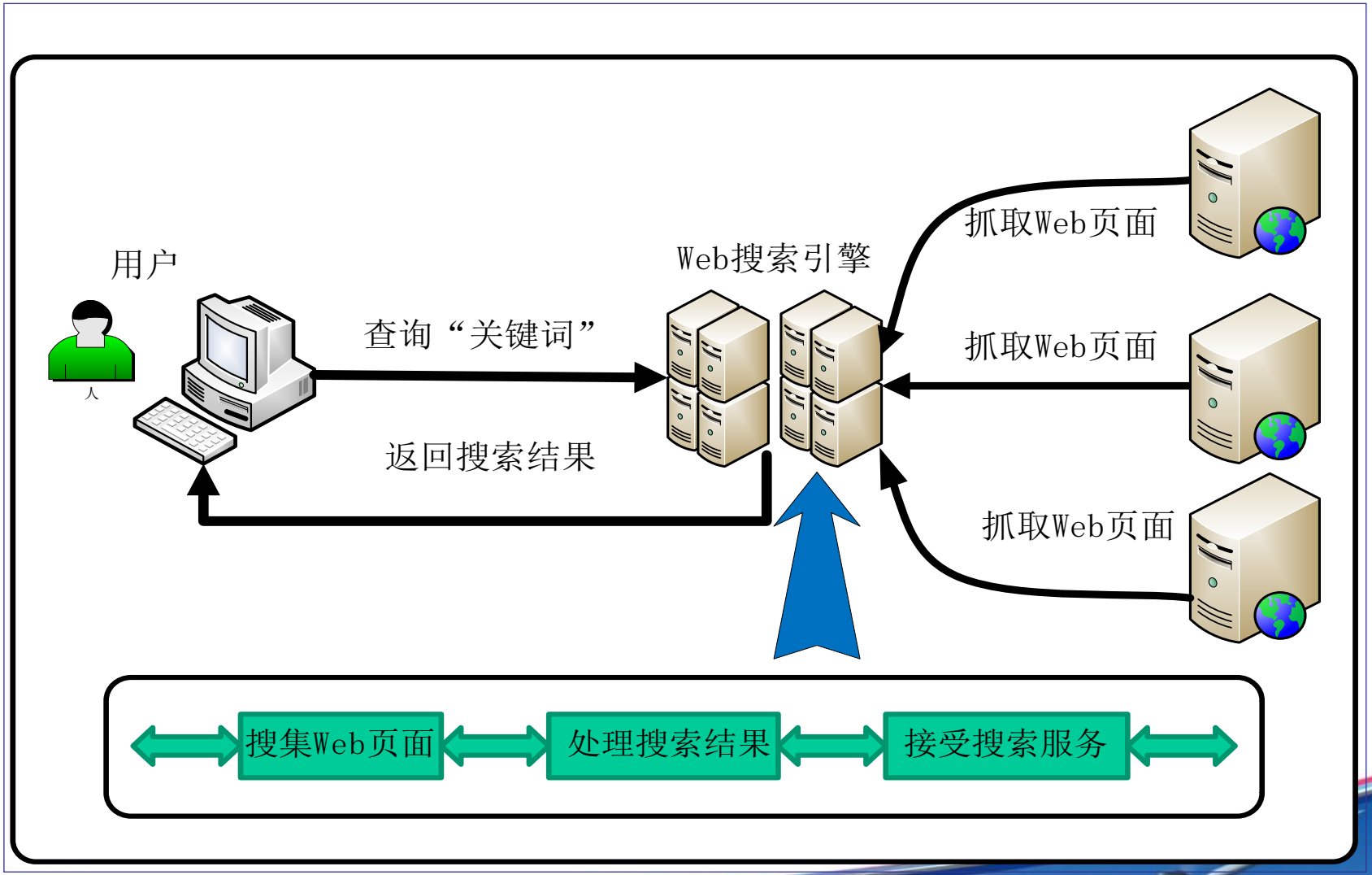


搜索引擎整体结构

- **爬虫**从 Internet 中爬取众多的网页作为原始网页库存储于本地；
- 然后**网页分析器**抽取网页中的主题内容交给**分词器**进行分词，得到的结果用**索引器**建立正排和倒排索引，这样就得到了**索引数据库**；
- 用户查询时，再通过**分词器**切割输入的查询词组并通过**检索器**在索引数据库中进行查询，得到的结果返回给用户。

一、概述

4. Web搜索引擎的工作模式



二、搜索引擎的体系结构

1. Web搜索引擎的3个重要问题

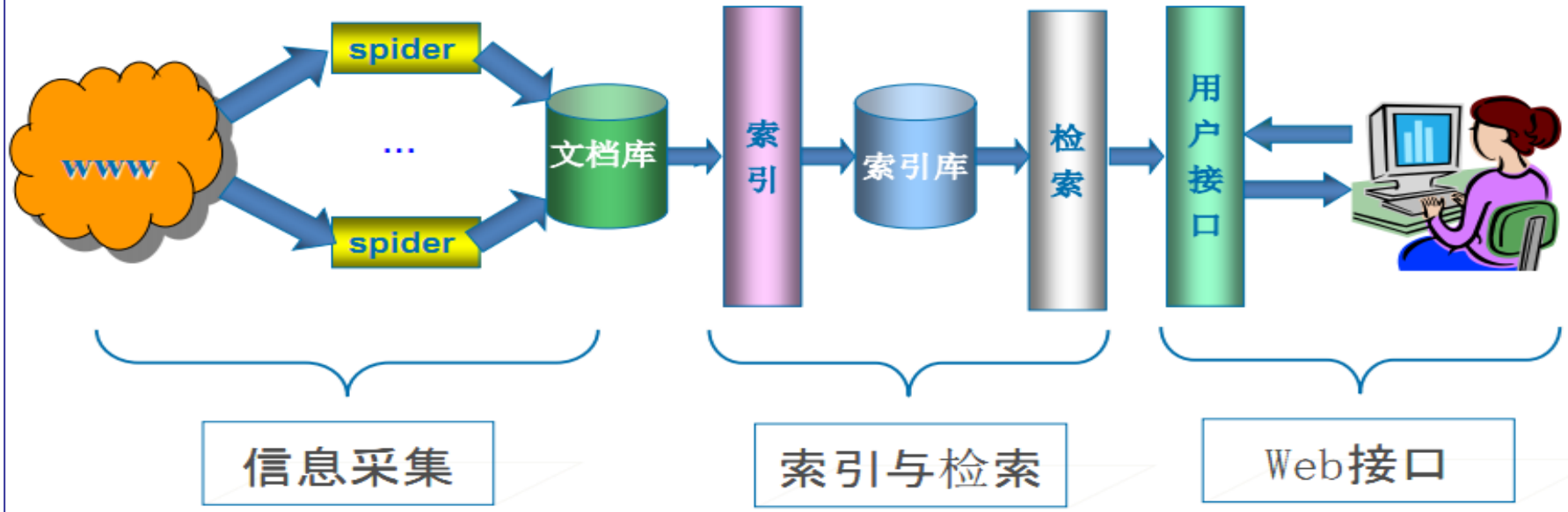
❶**响应时间**：一般来说合理的响应时间在秒这个数量级

❷**关键词搜索**：得到合理的匹配结果

❸**搜索结果排序**：如何对海量的结果数据排序

2. 搜索引擎的体系结构

❶信息采集、❷索引技术、❸搜索服务



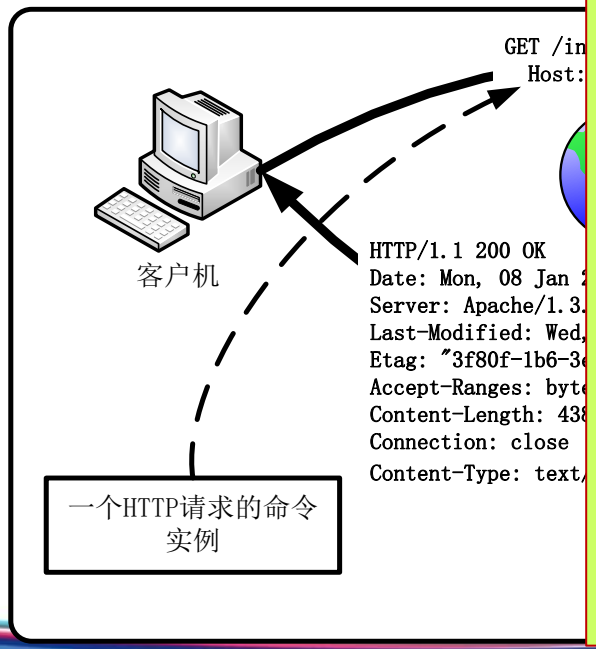
二、搜索引擎的体系结构

2. 搜索引擎的体系结构

❶信息采集、❷索引技术、

（1）信息采集

- Web搜索引擎的信息采集
  - 主要功能：Web上
  - 基于超文本传输协议
- 典型的基于超文本传输



❖ HTTP/1.1协议中共定义了八种方法（有时也叫“动作”）来表明Request-URI指定的资源的不同操作方式：

❶OPTIONS ——返回服务器针对特定资源所支持的HTTP请求方法。也可以利用向Web服务器发送'*'的请求来测试服务器的功能性。

❷HEAD ——向服务器索要与GET请求相一致的响应，只不过响应体将不会被返回。这一方法可以在不必传输整个响应内容的情况下，就可以获取包含在响应消息头中的元信息。

❸GET——向特定的资源发出请求。注意：GET方法不应当被用于产生“副作用”的操作中。

❹POST——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。POST请求可能会导致新的资源的建立和/或已有资源的修改。

❺PUT—— 向指定资源位置上传其最新内容。

❻DELETE——请求服务器删除Request-URI所标识的资源。

❼TRACE——回显服务器收到的请求，主要用于测试或诊断。

❽CONNECT——HTTP/1.1协议中预留给能够将连接改为管道方式的代理服务器。

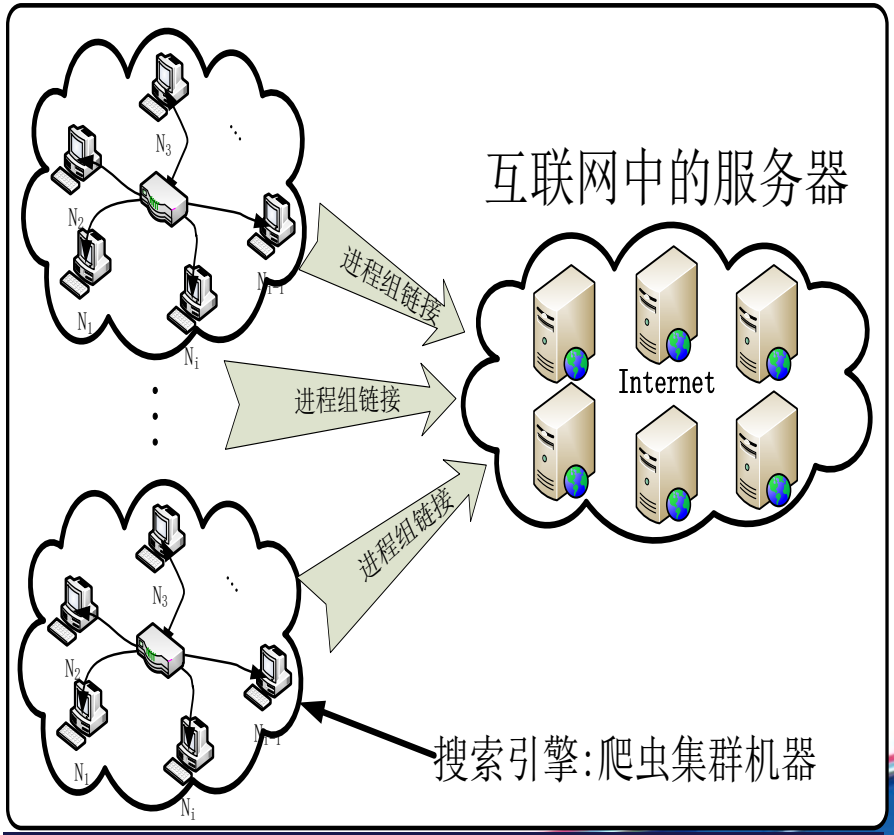
二、搜索引擎的体系结构

2. 搜索引擎的体系结构

❶信息采集、❷索引技术、❸搜索服务

（1）信息采集 → ❶网络爬虫程序的工作模式

- **网络爬虫程序**根据HTTP协议，发送请求，并通过TCP连接接受服务器的应答。
- 由于Web搜索引擎需要抓取数以亿计的页面，所以建立快速分布式的网络爬虫程序才能满足搜索引擎对性能和服务的要求，其物理实现可能是一组终端。



爬虫程序物理设备架构图

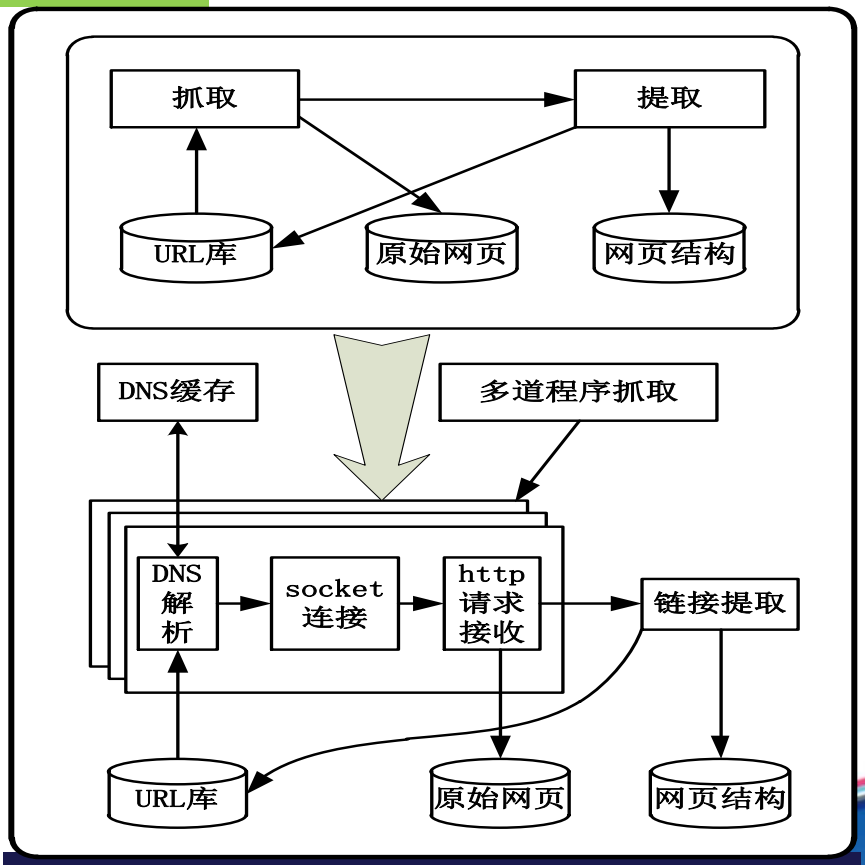
二、搜索引擎的体系结构

2. 搜索引擎的体系结构

❶信息采集、❷索引技术、❸搜索服务

（1）信息采集 → ❷网络爬虫程序的基础结构

- 首先**网络爬虫程序**从URL链接库读取一个或多个URL作为初始输入并进行域名解析
- 然后根据域名解析结果（IP）访问**Web服务器**，**建立TCP连接**，**发送请求**，**接受应答**，储存接收的数据，并分析提取链接信息（URL）放入URL连接库里。
- **爬虫程序递归**执行该过程直到URL链接库为空。



爬虫程序的基础结构图

二、搜索引擎的体系结构

2. 搜索引擎的体系结构

❶信息采集、❷索引技术、❸搜索服务

（1）信息采集 → ❸信息采集优化

- **网络连接优化策略**
  - 持久性连接
  - 多进程并发设计
- **域名系统的缓存策略**：由于网络爬虫程序会频繁调用域名系统，域名系统缓存(即DNS缓存)可提高爬虫程序性能。
  - **LRU**（Least Recently Used：近期最少使用，或最近最少使用）算法
  - **LFU**（Least Frequently Used：最近最不常用）算法
  - FIFO（First-In, First-Out）算法

回忆《计算机组成原理》中讲解过的Cache
道理都是一样的，只是用在了不同地方而已

二、搜索引擎的体系结构

2. 搜索引擎的体系结构

❶信息采集、❷索引技术、❸搜索服务

（1）信息采集 → ❹网页抓取算法

ⓘ深度优先算法

在Web收集页面信息时，使用一个或一组预定义URL地址开始，然后根据页面内容中的超链接深度抓取页面，直到搜索结束（没有新的URL）。

ⓘ广度优先算法

在Web收集页面信息时，使用一个或一组预定义URL地址开始，然后根据页面内容中的超链接广度抓取页面，抓取下一层的URL直到这一层的URL完全被抓取，直到搜索结束时返回。

ⓘ基于内容的算法

根据**关键字、主题文档的相似度**和**链接文本（Linked texts）估计**链接值，并确定相应搜索策略的算法。链接文本是包含对URL链接解释说明和内容摘要的文字信息。

ⓘ基于HITS(Hypertext-Induced Topic Search)的算法

主要思想：在抓取Web页面时，采用Authority/Hub抓取策略。**Authority**表示该页面被其他页面所引用的次数（页面入度值，in-degree value）。**Hub**表示一个Web页面指向其它页面的数量（页面出度值，out-degree value）。

二、搜索引擎的体系结构

2. 搜索引擎的体系结构

怎么理解?

❶信息采集、❷索引技术、❸搜索服务

（1）信息采集 → ❹网页抓取算法

ⓘPageRank（Google的传奇技术）

定义PageRank： 我们假设有$T_1...T_n$个页面指向页面A（即引用）。参数d是一个阻尼因子，其取值区间属于(0,1)，我们通常取值为0.85。C(A)定义为页面A指向其他页面的链接数(即链出数)，页面A的PageRank或PR(A)值可以通过下面的公式得到：

$$PR(A)=(1-d)+d(\frac{PR(T_1)}{C(T_1)}+...+\frac{PR(T_n)}{C(T_n)})$$

☺ 注意：PageRank值是Web页面的概率表示，所以所有Web页面的PageRank值的和是**1**。

二、搜索引擎的体系结构

2. 搜索引擎的体系结构

❶信息采集、❷索引技术、❸搜索服务

（2）索引技术

- Web爬虫抓取回来的页面信息，需要放入**索引数据库**里。
- 索引建立的好坏对于搜索引擎有很大的影响，优秀的索引能够显著的提高搜索引擎系统运行的效率及检索结果的品质。
- **文本分析技术**是建立数据索引信息的支撑技术。

二、搜索引擎的体系结构

2. 搜索引擎的体系结构

❶信息采集、❷索引技术、❸搜索服务

（2）索引技术 → ❶索引建立：预处理

- 当Web搜索引擎获得数据信息以后，首先需要对数据进行预处理，如将句子切分成有意义的词汇。由于中文的特殊性在切分句子时会产生二义性，如何合理的切分词汇是一个**技术难题**。
- **中文分词**完全不同于英文分词，英文行文中，单词间以空格分隔；而中文只有字/句/段有明显分隔符，唯独**词**没有形式上的分隔符存在。

愿动者健康长寿

二、搜索引擎的体系结构

浮云长长长长长长长消

海水朝朝朝朝朝朝朝落

海水朝（ch á o）,朝（zh ā o）朝（zh ā o）朝（ch á o）,朝（zh ā o）朝（ch á o）朝（zh ā o）落;

浮云长（zh ǎ ng）,长（ch á ng）长（ch á ng）长（zh ǎ ng）,长（ch á ng）长（zh ǎ ng）长（ch á ng）消.

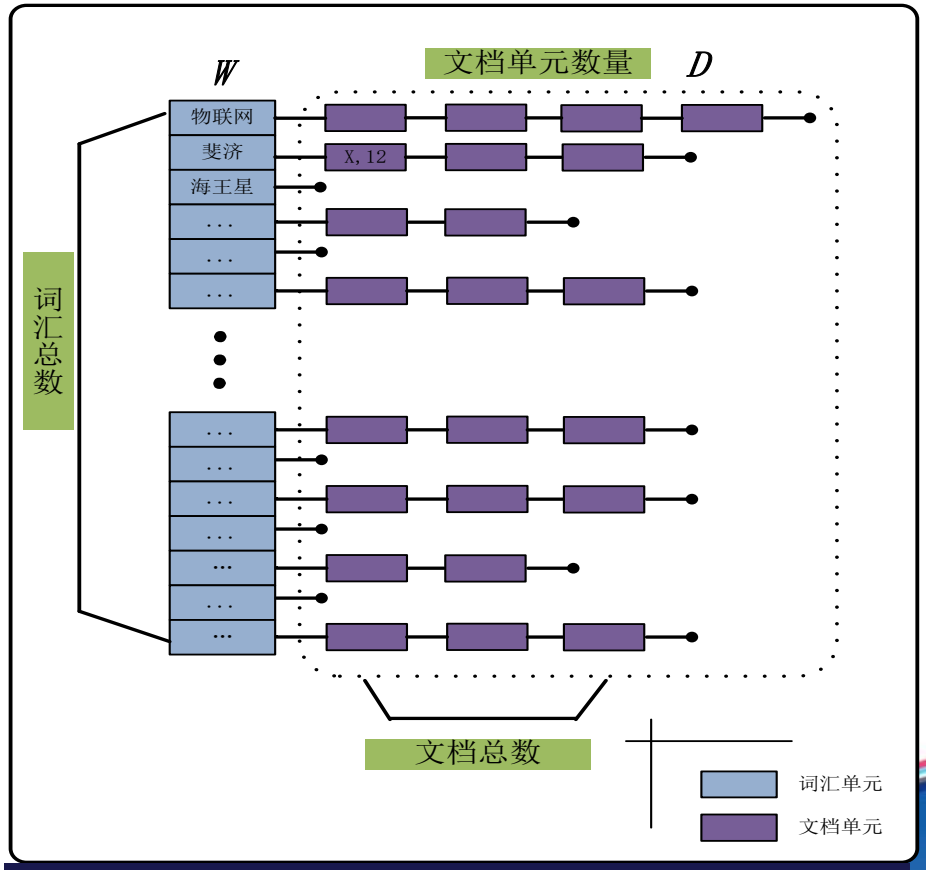
二、搜索引擎的体系结构

2. 搜索引擎的体系结构

❶信息采集、❷索引技术、❸搜索服务

（2）索引技术 → ❷索引建立：倒排文件模型

- **倒排文件**（inverted file），是指一个**词汇集合W**和一个**文档集合D**之间对应关系的**数据结构**。
- 建立倒排文件索引是建立**索引数据库**的核心工作。



索引模块架构

二、搜索引擎的体系结构

2. 搜索引擎的体系结构

❶信息采集、❷索引技术、❸搜索服务

（3）搜索服务

- 搜索服务是Web搜索引擎工作流程的最后一步，根据用户提交的查询关键字展开搜索，将匹配结果返回给用户。
- 搜索服务的好坏直接影响Web搜索引擎的用户满意程度。

（3）搜索服务 → ❶结果显示

- 接受用户的输入，提交用户搜索请求。
- 根据搜索结果列表合理的展示给用户。
- 在保护隐私的前提下，记录用户使用行为的详细信息，以便提高下次服务的满意度。

（3）搜索服务 → ❷网页快照

- Web上的数据每时每刻都在变化着，所以随时存在着检索到的页面信息已经不存在的可能。
- Web搜索引擎为了提高服务质量，需要对搜索到的页面信息进行**快照**，以便在原来页面信息失效的情况下，保证用户能够通过快照功能查看页面。

此时，检索的数据库中有，而物理页面已经没有了

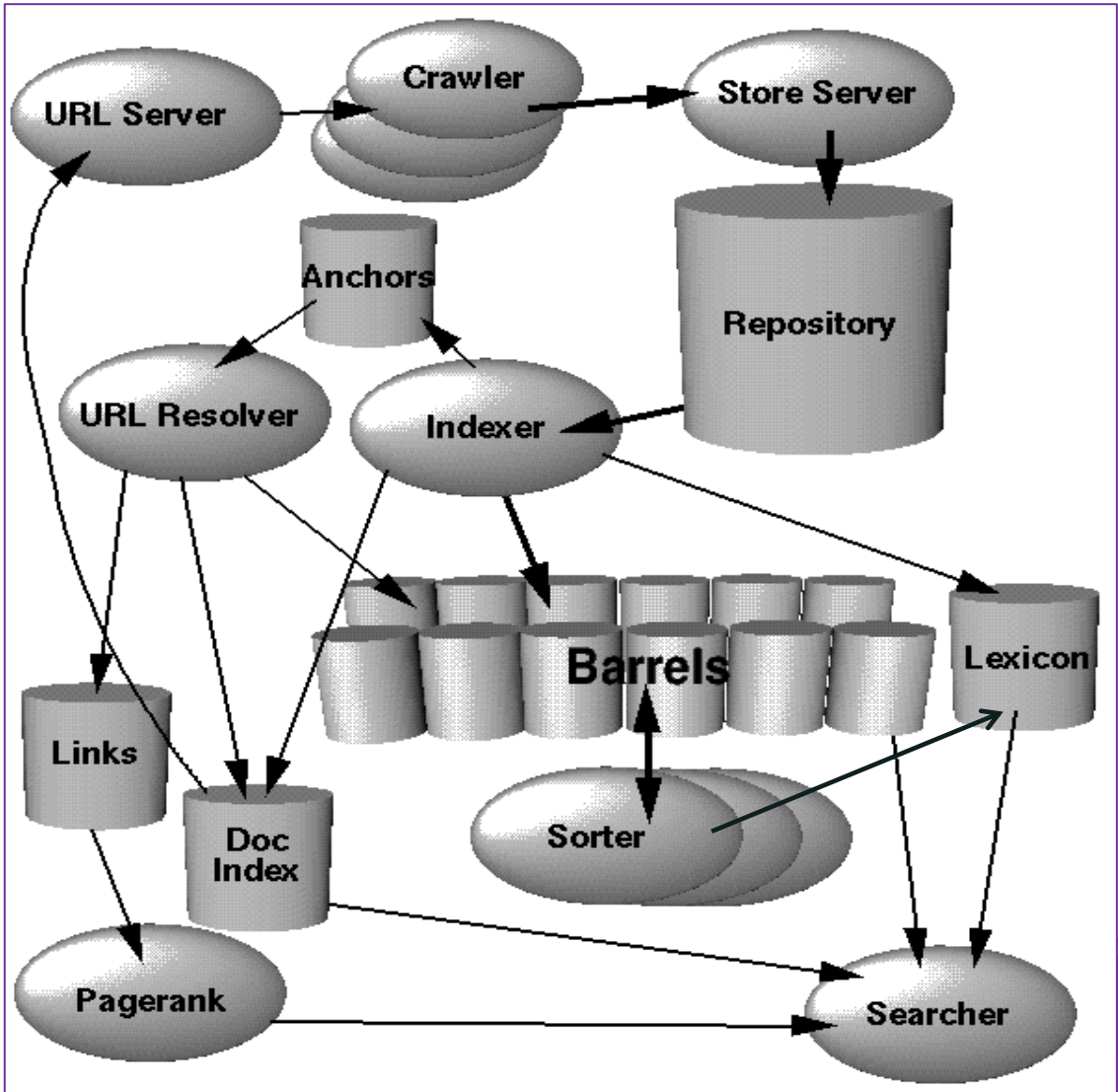
二、搜索引擎的体系结构

Google的大部分是用C或C++实现，可以在Solaris或者Linux下运行。

3. 类Google Web搜索引擎的实例分析*

类Google搜索引擎的架构

- URL服务器
- Web页面抓取器
- 存储服务器
- URL解释器
- 排序器
- Page Rank
- 搜索器



二、搜索引擎的体系结构

3. 类Google Web搜索引擎的实例分析*

Repository:53.5GB=147.8GB uncompressed

sync	length	compressed packet
sync	length	compressed packet

…

Packet(stored compressed in repository)

docid	ecode	urllen	pagelen	url	page

Google数据仓库的结构

二、搜索引擎的体系结构

3. 类Google Web搜索引擎的实例分析*

查询评估流程

❶解析查询（Query）

❷把单词转化成wordID

❸从每个单词的短桶文档列表开始查找

❹扫描文档列表直到有一个文档匹配了所有的搜索词语

❺计算这个文档对应的查询的**评分**

❻如果到达短桶的文档列表结尾，从每个单词的全桶(full barrel)文档列表开始查找，跳到第4 步

❼如果没有到达任何文档列表的结尾，跳到第4步

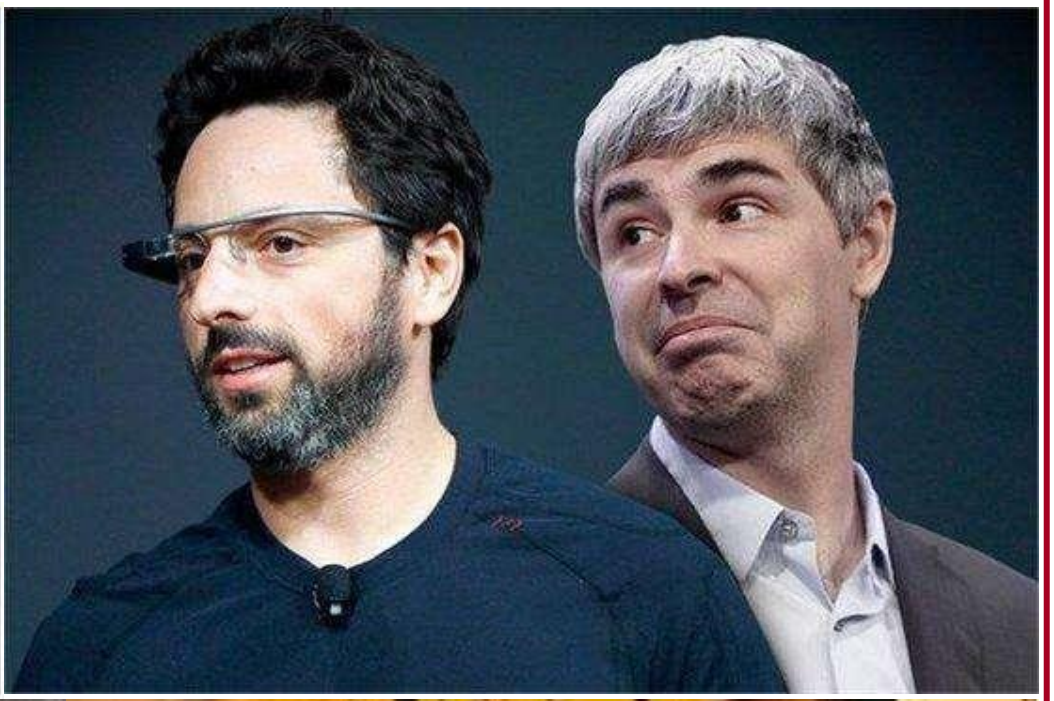
❽根据评分对匹配的文档排序，然后返回评分最高的k个

1. 物联网搜索引擎思考

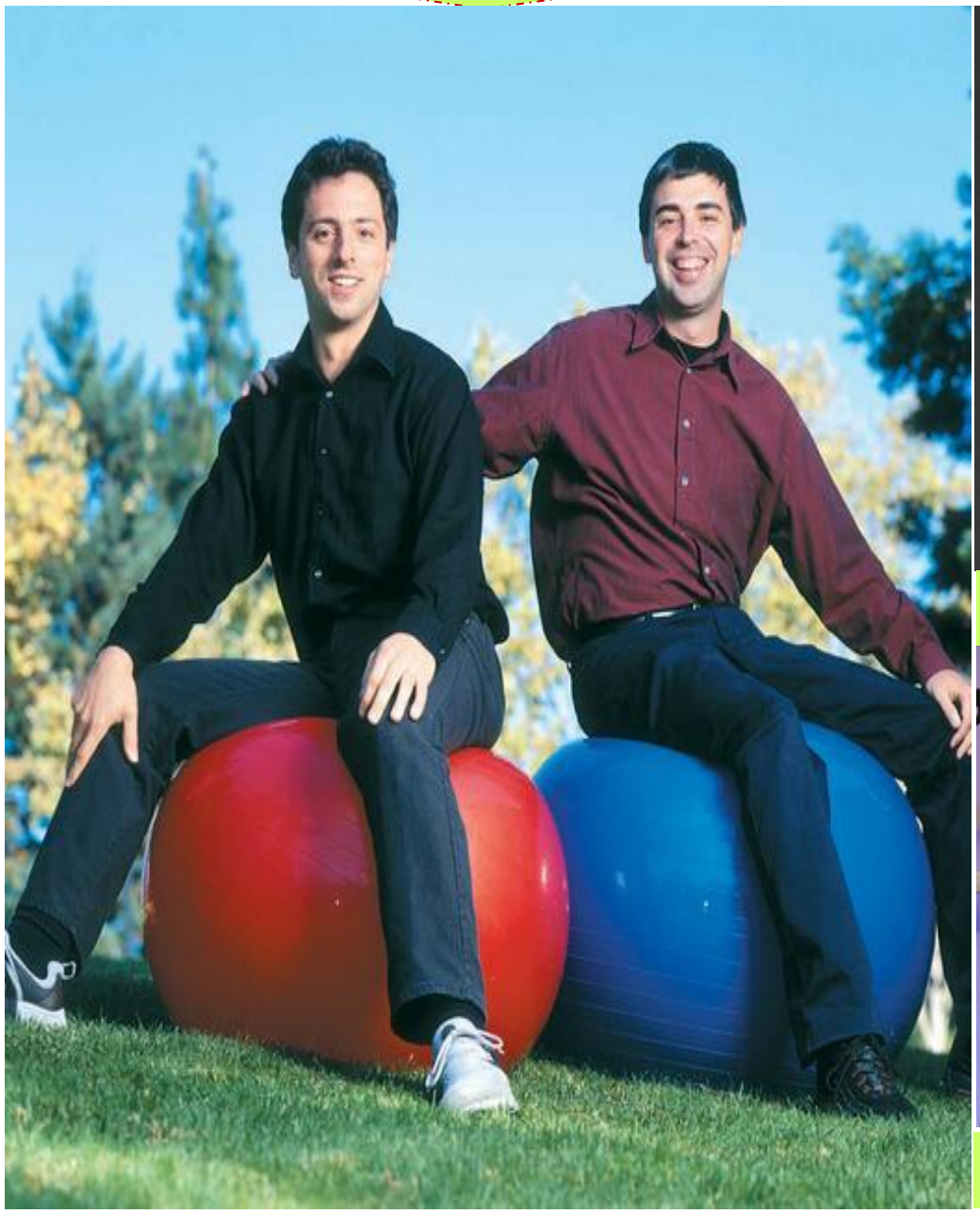
- 从智能物体角度思考搜索引擎与物体之间的关系，主动识别物体并提取有用信息。
- 从用户角度上的多模态信息利用，使查询结果更精确，更智能，更定制化。

物联网搜索引擎让你浏览网络摄像头

Google创始人：拉里·佩奇(左) 、谢尔盖·布林(右)



Google创始人：拉里·佩奇(右) 、谢尔盖·布林(左)



拉里·佩奇

谢尔盖·布林

拉里·佩奇

谢尔盖·布林

百度创始人： 李彦宏

（由左到右依次是扎克伯格、乔布斯、谷歌的两位创始人谢尔盖·布林（Sergei Brin）和拉里·佩奇）