Chapter 12

Search engine

Introduction to Internet of Things

内容提要

Only **Search Engines** that provide "pervasive data analysis and services" can interpret the connotation of "deeper intelligence" of the Internet of things.

This chapter introduces the basic concepts, architecture and related technologies of search engine.

Review

Chapter 11 introduces mass information storage technology and data center

- The Internet of things needs massive data storage
- Three network storage architectures
- Basic concepts of data centers and typical data centers
- How to reduce the cost of data center

This chapter focuses on the basic concepts of the search engine, including the development history, architecture (information collection, indexing technology, search services), and a brief analysis of search engine like google Web .

Content

12.1 Introduction to search engines

12.2 Search engine architecture

12.3 Internet of things search engine

What are the search engines?

What are the three major modules?

12.1 Search engines introduction

Web search engine

A combination of hits list services that return relevant information based on the user's query keywords in a reasonable response time.

Traditional Web search engines are based on query keywords, for the same keywords, you will get the same query results.

Common Web search engines

The development of search engines

- The origins of search engines can be traced back to 1992, when NCSA maintained "What's NEW!" page.
- First original search engine W3Catalog (1993.9)
- The first Web robot program "World Wide Web Wanderer" (1993.6 MIT)
- Milestones: WebCrawler (1994), Lycos (1994) commercially
- Founding of Google: Larry Page and Sergey Brin, Stanford doctoral students, founded Google

1993-2010 Web Search Engine	
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

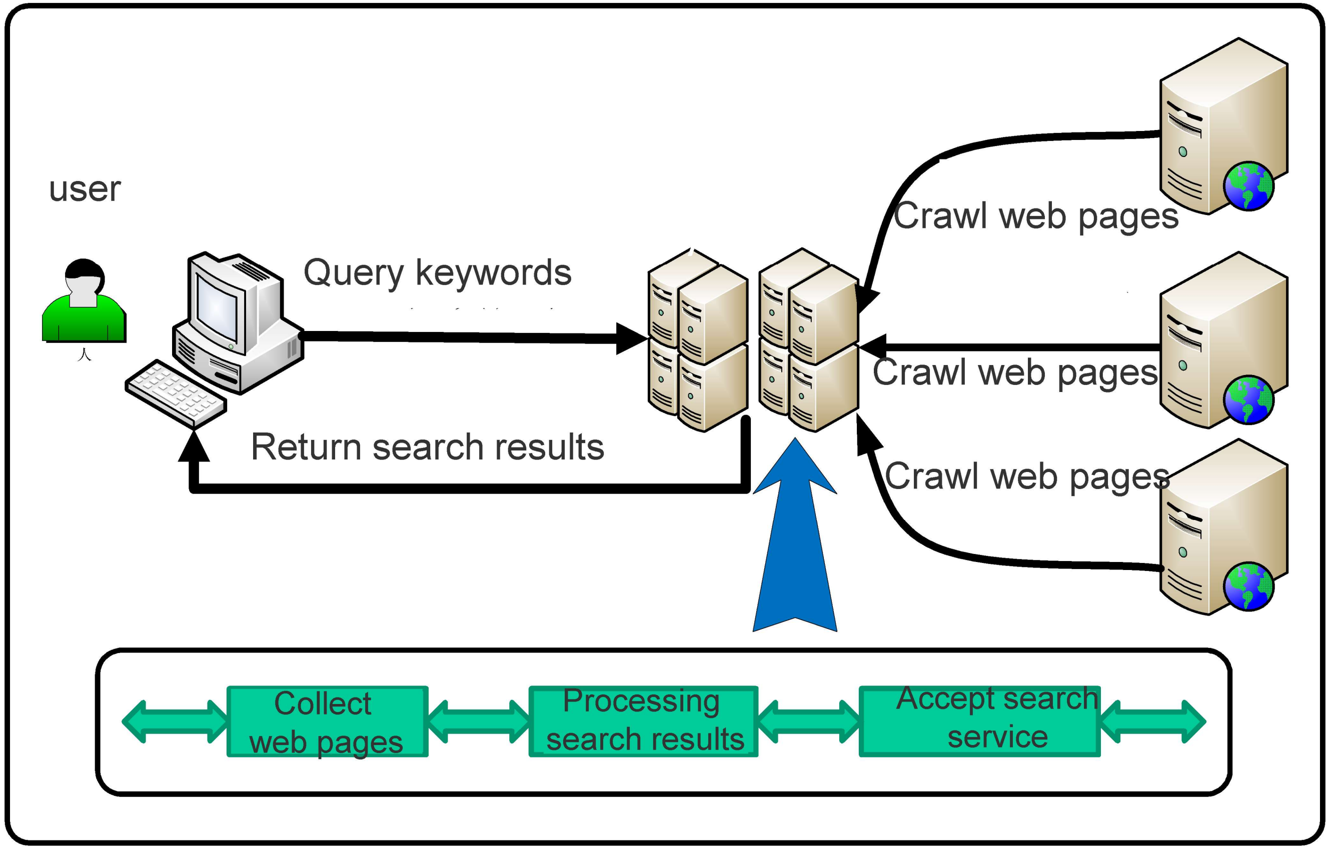
✔ The structure of Web search engines

Web crawler module: the main function is to parse Web pages, crawl these pages according to the connection between Web pages, and store page information to the index module for processing.

Index module: the main task is to preprocess the captured data and establish keyword index for searching module output.

Search module: for the user's keywords, according to the database index knowledge to give reasonable search results.

Web Search Engines Work Pattern



Content

12.1 Introduction to search engines

12.2 Search engine architecture

12.3 Internet of things search engine

What is the architecture and relevant technology of search engines?

Three important questions about Web search engines

- **Response time:** generally a reasonable response time is in the order of seconds
- **Keyword search:** get reasonable matching results
- **Search result sorting:** how to sort massive amounts of result data

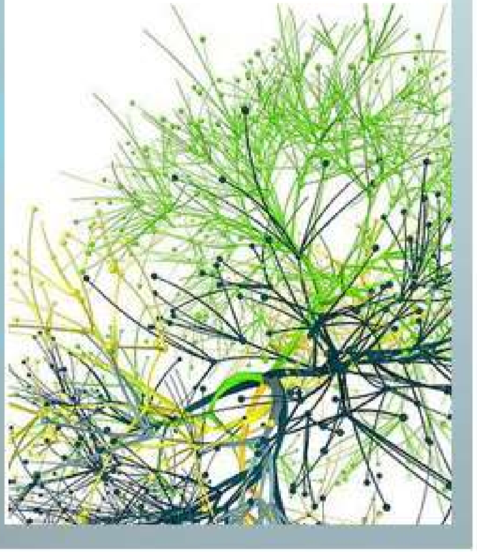
Search engine architecture

- ✓ Information collection
- ✓ Indexing technology
- ✓ Search service

Architecture: information collection

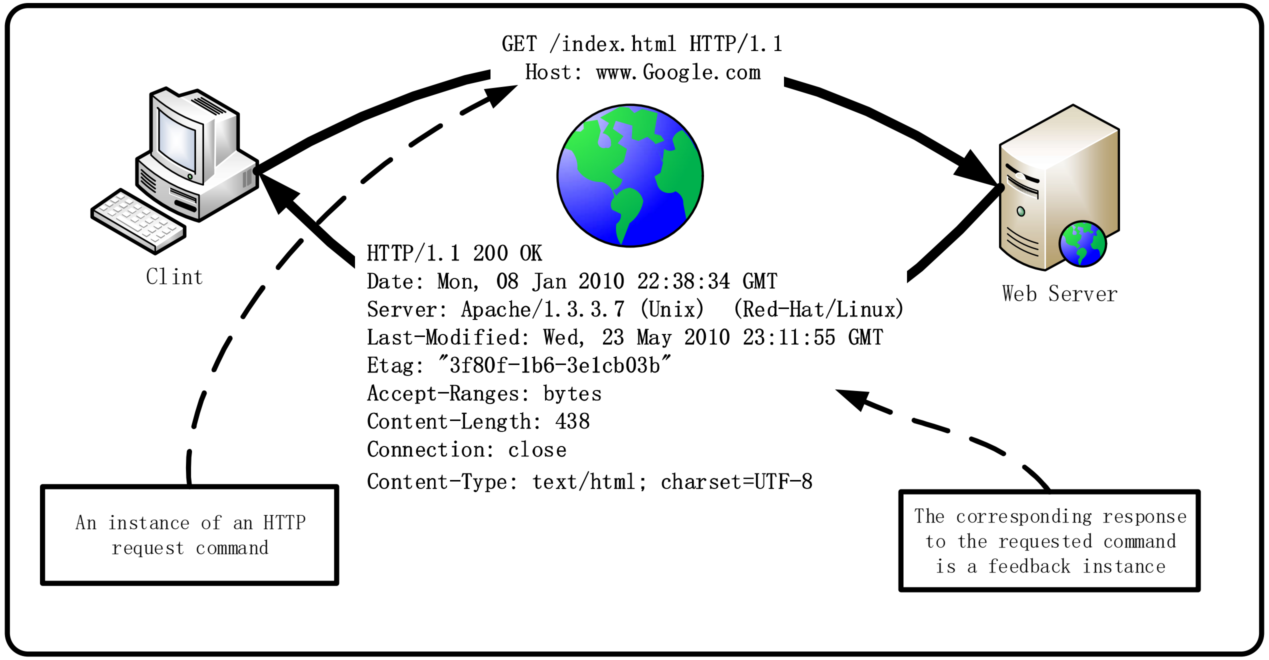
Web search engine information acquisition module

- Main function: collect page information on the Web, namely Web robot (crawler) program
- Based on Hypertext Transfer Protocol (HTTP)



Architecture: information collection

Typical network reply based on hypertext transfer protocol

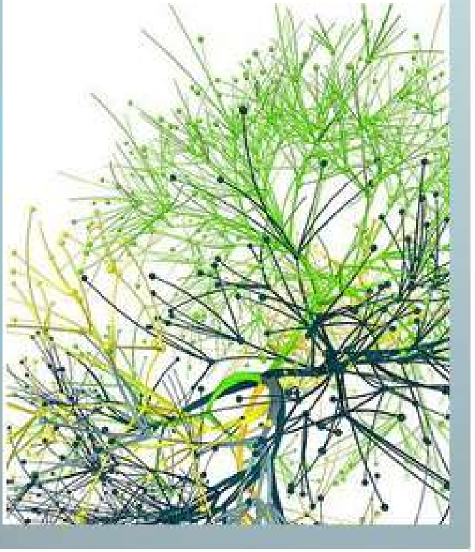


```
GET /index.html HTTP/1.1
Host: www.Google.com
```

```
HTTP/1.1 200 OK
Date: Mon, 08 Jan 2010 22:38:34 GMT
Server: Apache/1.3.3.7 (Unix)  (Red-Hat/Linux)
Last-Modified: Wed, 23 May 2010 23:11:55 GMT
Etag: "3f80f-1b6-3e1cb03b"
Accept-Ranges: bytes
Content-Length: 438
Connection: close
Content-Type: text/html; charset=UTF-8
```

An instance of an HTTP request command

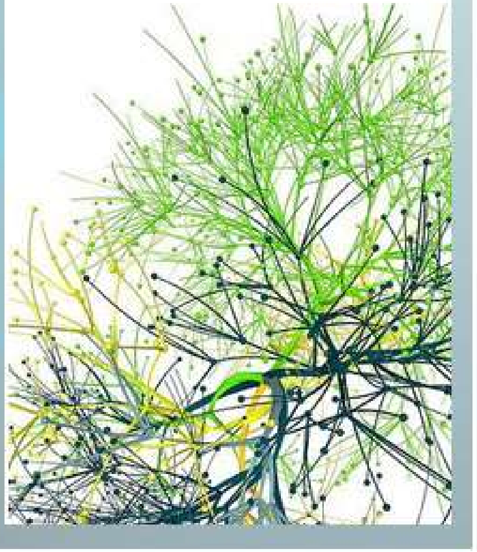
The corresponding response to the requested command is a feedback instance



✓ Web Crawlers work pattern

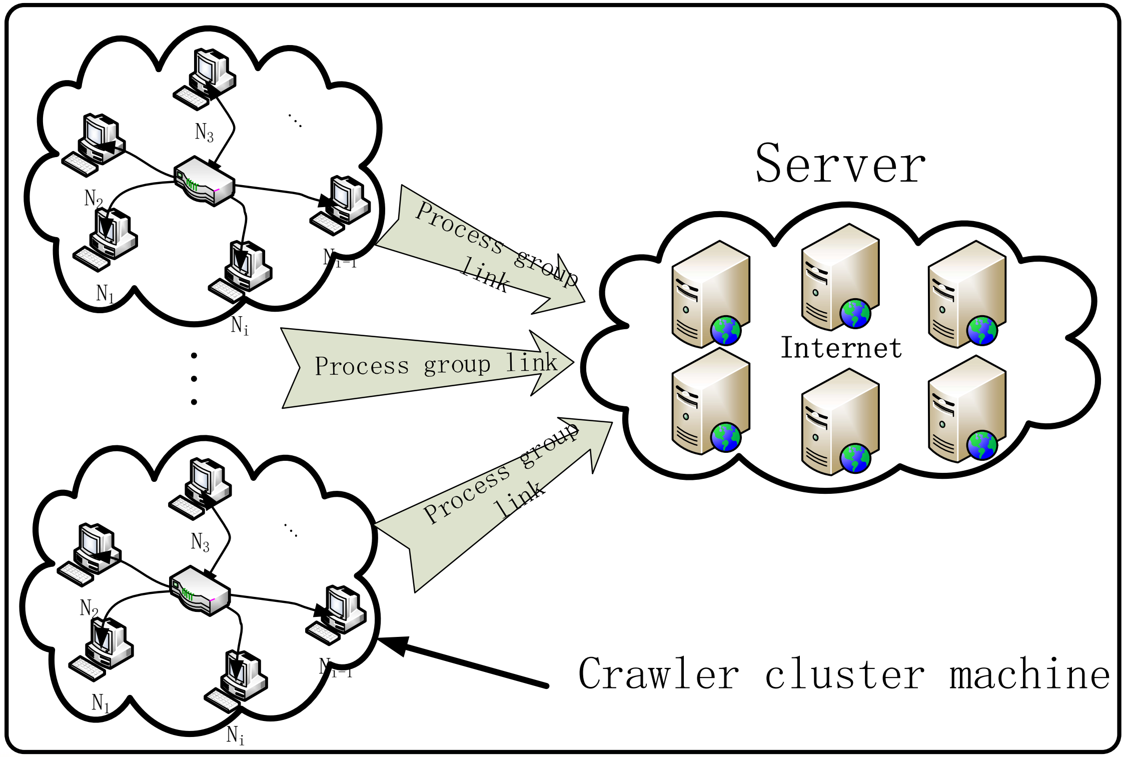
The web crawler sends the request according to the HTTP protocol and receives the reply from the server through the TCP connection.

Since Web search engine needs to crawl hundreds of millions of pages, it is necessary to establish a fast distributed Web crawler program to meet the requirements of search engine on performance and service. Its physical implementation may be a group of terminals.

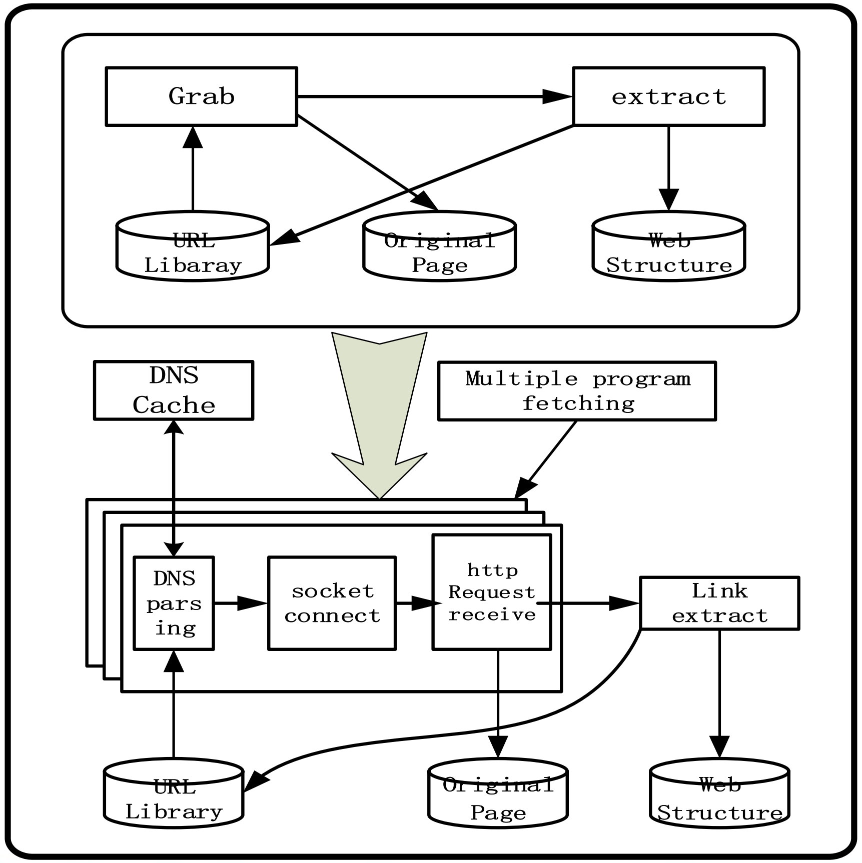


Web Crawlers work pattern

Architecture diagram of crawler physical device



The infrastructure of a web crawler



Grab

extract

URL Libaray

Original Page

Web Structure

DNS Cache

Multiple program fetching

DNS pars ing

socket connect

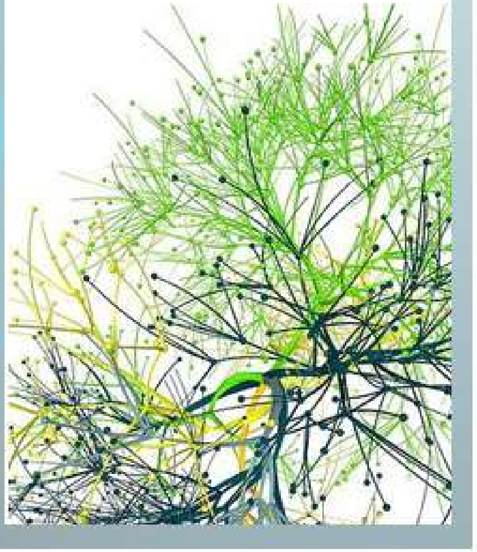
http Request receive

Link extract

URL Library

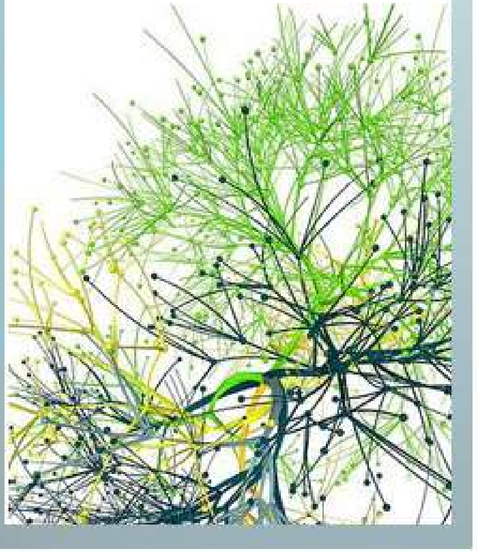
Original Page

Web Structure



The infrastructure of a web crawler

First, the web crawler reads one or more URL from the URL link library as initial input and performs domain name resolution
Then according to the domain name resolution results (IP) to visit the Web server, establish TCP connection, send requests, receive replies, store received data, and analyze extract link information (URL) into the URL connection library.
The crawler recursively performs this procedure until the URL link library is empty.



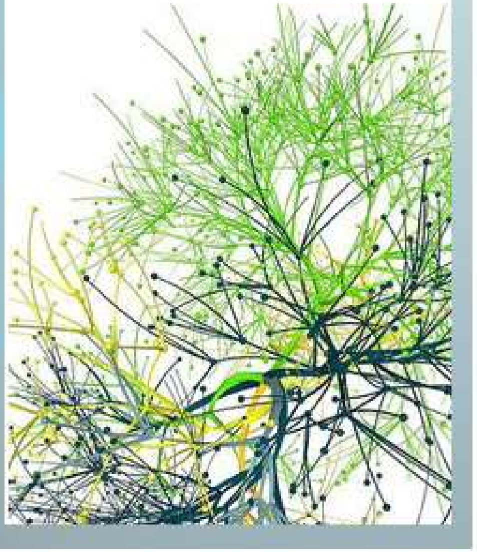
Information acquisition optimization

Network connection optimization strategy

- Persistent connection
- Multi-process concurrency design

Domain name system cache strategy: because the web crawler will call the domain name system frequently, the domain name system cache can improve the crawler performance.

- LRU (Least Recently Used) algorithm
- LFU (Lease Frequently Used) algorithm
- FIFO (first-in, first-out) algorithm



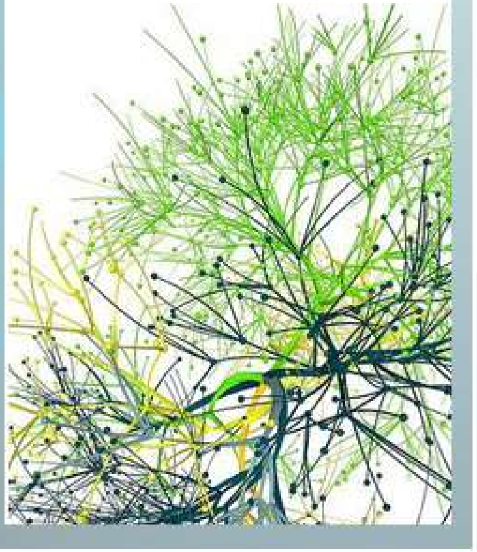
Web Scraping algorithm

Depth-first algorithm

When a Web collects page information, it starts with one or a set of predefined URL addresses, and then crawls the page based on the hyperlink depth in the page content until the search ends (no new URL are available).

Breadth first algorithm

To collect page information on the Web, start with one or a set of predefined URL addresses, and then crawl the page based on the hyperlink breadth of the page content, scraping the next layer of URL until the layer's URL is completely crawled, and returning at the end of the search.



Web Scraping algorithm

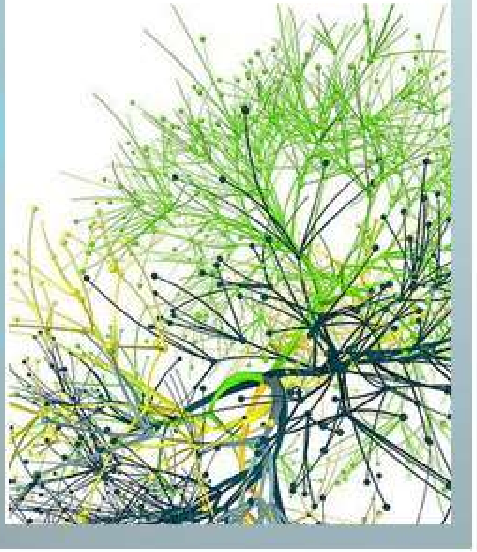
Content-based algorithm

Estimate link values based on keywords, similarity of topic documents, and Linked texts, and determine the algorithm for the corresponding search strategy. Link text is textual information that contains an explanation of the URL link and a summary of the content.

Algorithms based on HITS

Main idea: when crawling Web pages, use Authority/Hub crawling strategy. Authority is the number of times (in-degree value) the page has been referenced by other pages. Hub represents the number of times (out-of-degree value) pages refer to the page.

。



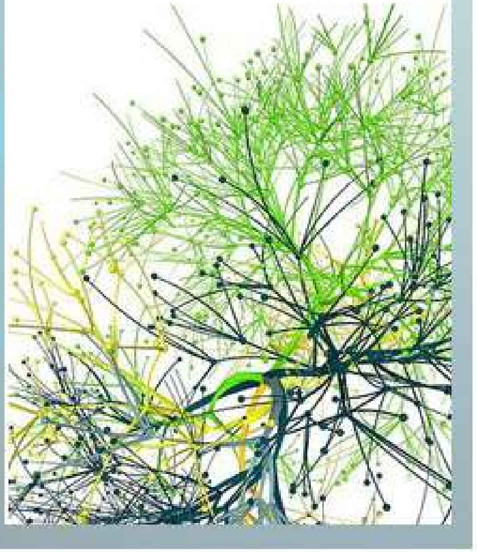
✓ Web Scraping algorithm

PageRank (Google's legendary technology)

Define PageRank: let's assume T1...Tn pages point to page A (the reference).Parameter d is a damping factor, whose value range belongs to (0,1), which is usually 0.85.C(A) is defined as the number of links to other pages of page A. The PageRank or PR(A) value of page A can be obtained by the following formula:

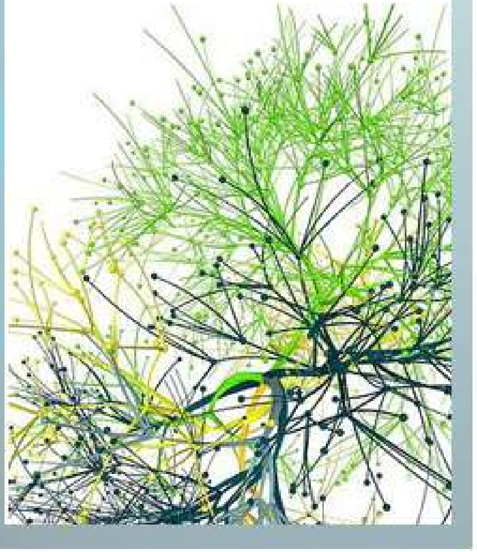
$$PR(A) = (1-d) + d(\frac{PR(T_1)}{C(T_1)} + ... + \frac{PR(T_n)}{C(T_n)})$$

Note: PageRank value is the probability distribution representation of Web pages, so the sum of PageRank values of all Web pages is 1.



Architecture: Indexing technology

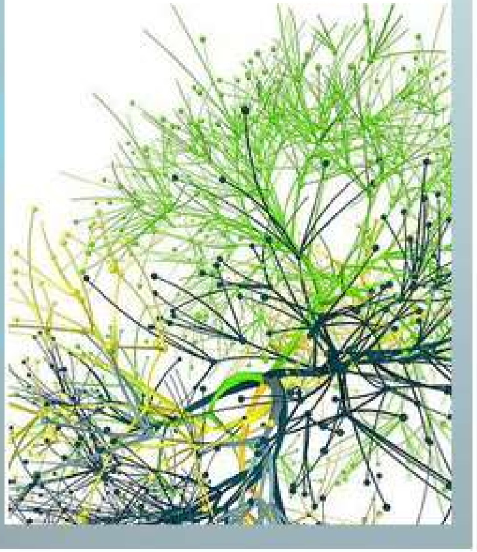
- Web crawlers capture back page information, need to put into the index database.
- The index establishment has a great impact on the search engine, excellent index can significantly improve the efficiency of the search engine system and the quality of search results.
- Text analysis technology is the support technology of establishing data index information.



✓ Indexing: pre-processing

When the Web search engine obtains the data information, it needs to preprocess the data first, such as cutting sentences into meaningful words. Due to the particularity of Chinese language, it is a difficult technical problem how to segment vocabulary reasonably.

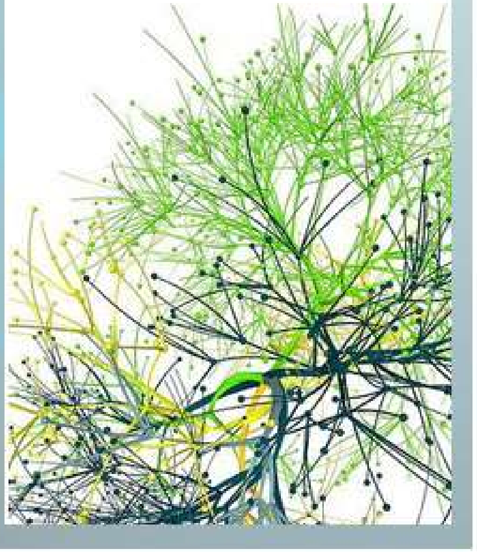
Chinese word participle is completely different from English word participle. In Chinese, only characters/sentences/paragraphs have obvious separators, but words do not have formal separators.



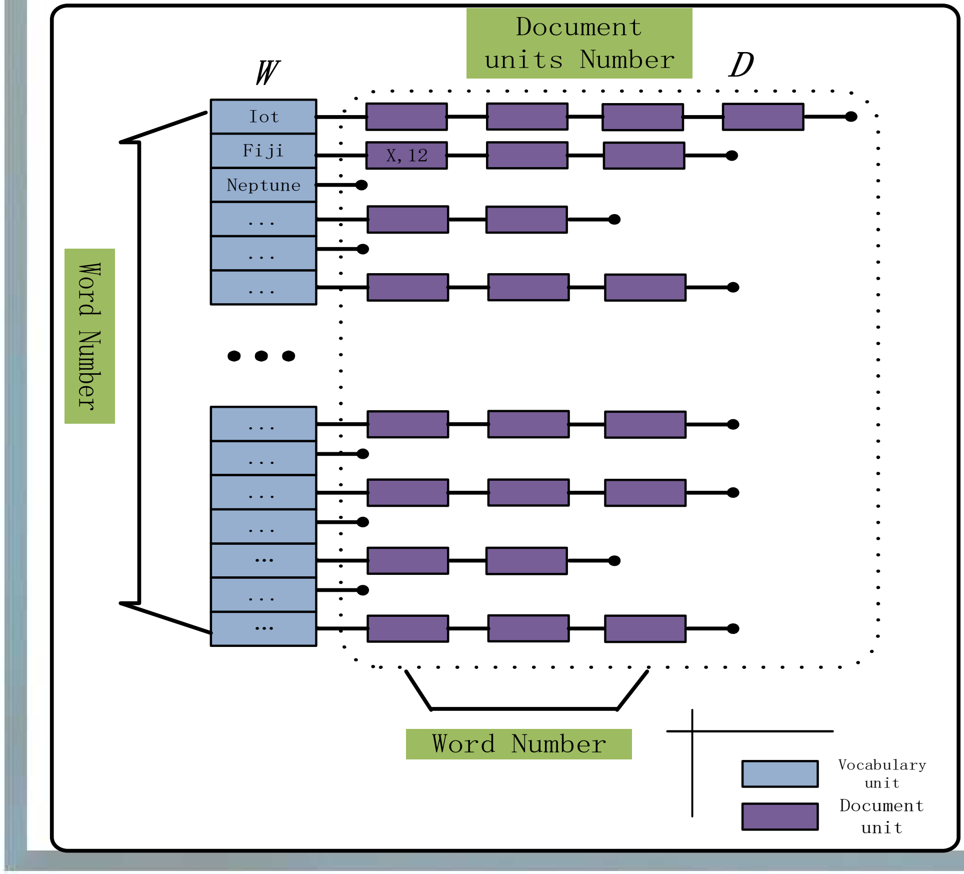
✓ Index creation: inverted file model

Inverted files are the data structures that correspond to a collection of words W and a collection of documents D.

The establishment of inverted file index is the core work of the establishment of index database.



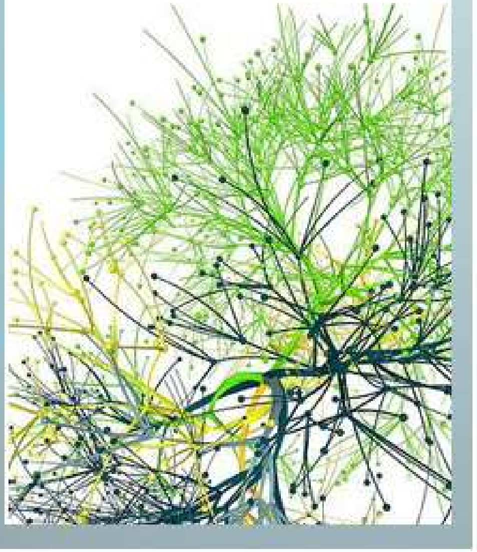
Index creation: inverted file model



Index module architecture

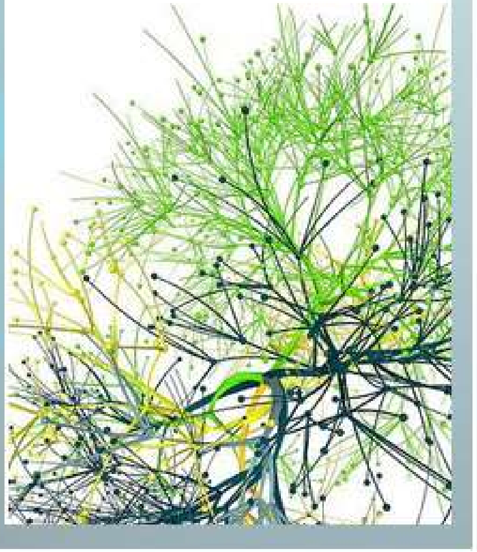
Architecture: search services

- The search service is the final step in the Web search engine workflow, which expands the search based on the query keywords submitted by the user and returns the matching results to the user.
- The quality of search service directly affects the user satisfaction of Web search engine.



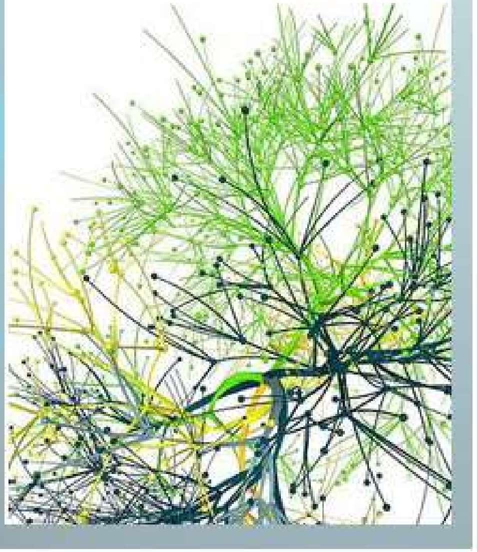
Search services: Results display

- Accept user input and submit user search request.
- Reasonably present to users based on the list of search results.
- Under the premise of protecting privacy, record the detailed information of users' usage behavior, so as to improve the satisfaction of the next service.



Search service: web snapshot

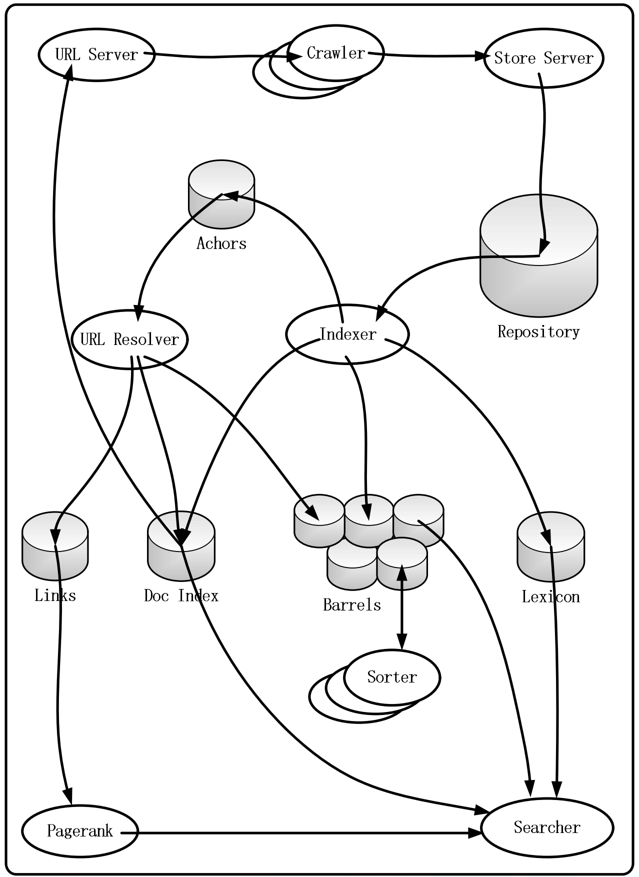
- Data on the Web is changing all the time, so there is always the possibility that the retrieved page information no longer exists.
- In order to improve the service quality, Web search engines need to take a snapshot of the searched page information, so that users can view the page through the snapshot function if the original page information fails.



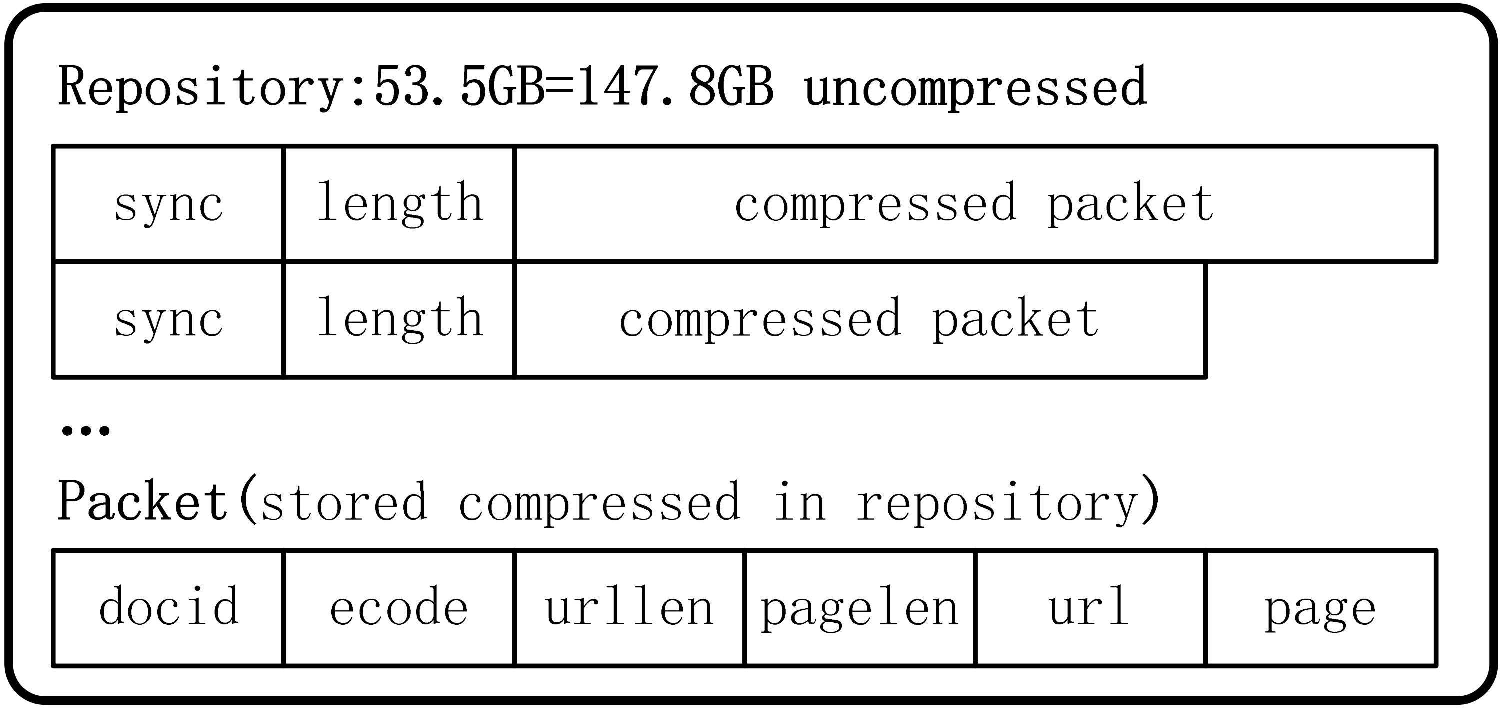
Case analysis of a google-like Web search engine *

Google-like search engine architecture

- The server URL
- Web page crawler
- Storage server
- URL interpreter
- collator
- Page Rank
- searcher



Case analysis of a google-like Web search engine *



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

The structure of Google's data warehouse

Case analysis of a google-like Web search engine *

Query assessment process

1. Parsed Query
2. Turn the word wordID
3. Start with the bucket document list for each word
4. Scan the document list until a document matches all the search terms
5. Calculate the score for the query corresponding to this document
6. If you reach the end of the bucket's document list, look up the full barrel document list for each word and jump to step 4
7. If you don't reach the end of any document list, skip to step 4
8. Sort the matching documents by score, and return the top k

Content

12.1 Introduction to search engines

12.2 Search engine architecture

12.3 Internet of things search engine

What new features should search engines have under the background of Internet of things?

12.3 Internet of things search engine

New thinking of search engine in Internet of things era

- Consider the relationship between search engine and objects from the perspective of intelligent objects, actively identify objects and extract useful information.
- The multi-modal information utilization from the user's perspective makes the query results more accurate, more intelligent and more customized.

Conclusion

Review

This chapter introduces the development of search engine, discusses the architecture of search engine (information collection, index technology, search service), and puts forward the new thinking of search engine under the background of Internet of things.

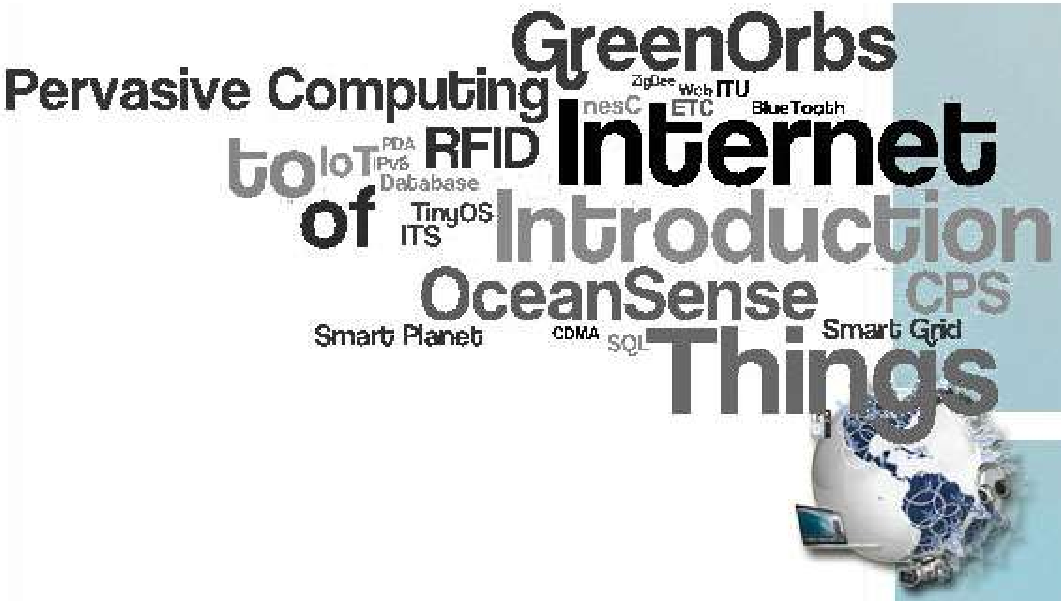
Key Points

- Grasp the role of three modules of Web search engine (Web crawler module, index module, search module).
- Understand the three problems that search engines need to solve (response time, keyword search, search result ordering).

Conclusion

Key Points

- Master the basic process of search engine information collection and the basic structure and workflow of web crawler program.
- Understand ways to optimize web crawlers. Grasp the characteristics and process of common web page crawling algorithms.
- Understand the difficulties of index technology preprocessing, understand the inverted file model.
- Illustrate the architecture of the Google Web search engine.



Thank you!

Internet of Things