Technologies behind Internet Search Engine

Ming-Jer Lee
CTO
VisionNEXT Inc.

Type of Search Engine

- **Media**
  - Text
  - Image
  - Audio
  - Video
- **Scope**
  - General search engine
  - Domain specific topic
  - Language
- **Scale**
  - personal, content site, intranet, Internet
  - thousand, million or billion (documents, users, queries)
- **Structure**
  - non-structure, semi-structure, structure
- **User Interface**
  - Web-based, Standalone AP based, voice driven

Type of Internet Search Engine

- **Manual Index**
  - Yahoo index, Looksmart, Open Directory
- **Automatic index**
- **Metasearch**
- **Answer by human expert**
- **P2P**
- **…**

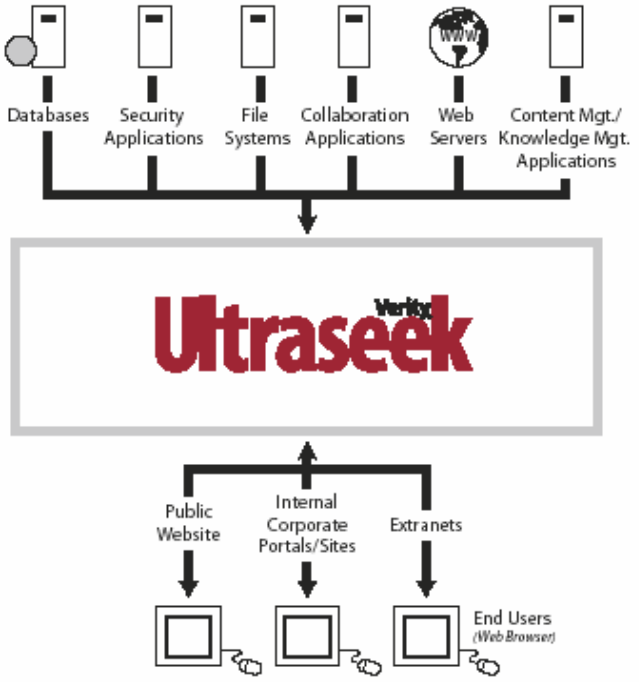
Search Engine in Business World

- Internet Search Engine(ISE)
  - Google, Openfind, VisionNEXT's eefind
- Enterprise Search Engine(ESE)
  - Verity, Convera, Virage, Tornado

	Discovery	Categorization	Information Search
ISE	**Spider follow links and Unstructured HTML, Office and PDF d ocuments**	**Manually categorization of web resources**	**- Keyword search - Boolean search - Search result ran king - Web page popula rity can be used as ISE weighting - User intension int erpretation**
ESE	**Structure and non-structured data in - File system - Database - Content management server - Collaboration server - Enterprise web site - Online news feed**	**Generally a mix of hum an input and automatic algorithms to maintain content category**	

image now. visionnext

ESE: Verity UltraSeek



Features

- Nature Language Search
- Application Integration
- Rapid Deployment
- Simple Administration
- Database Integration
- Security Integration
- Customized Interface
- Java API

image now.
visionnext

ESE: Convera RetrievalWare

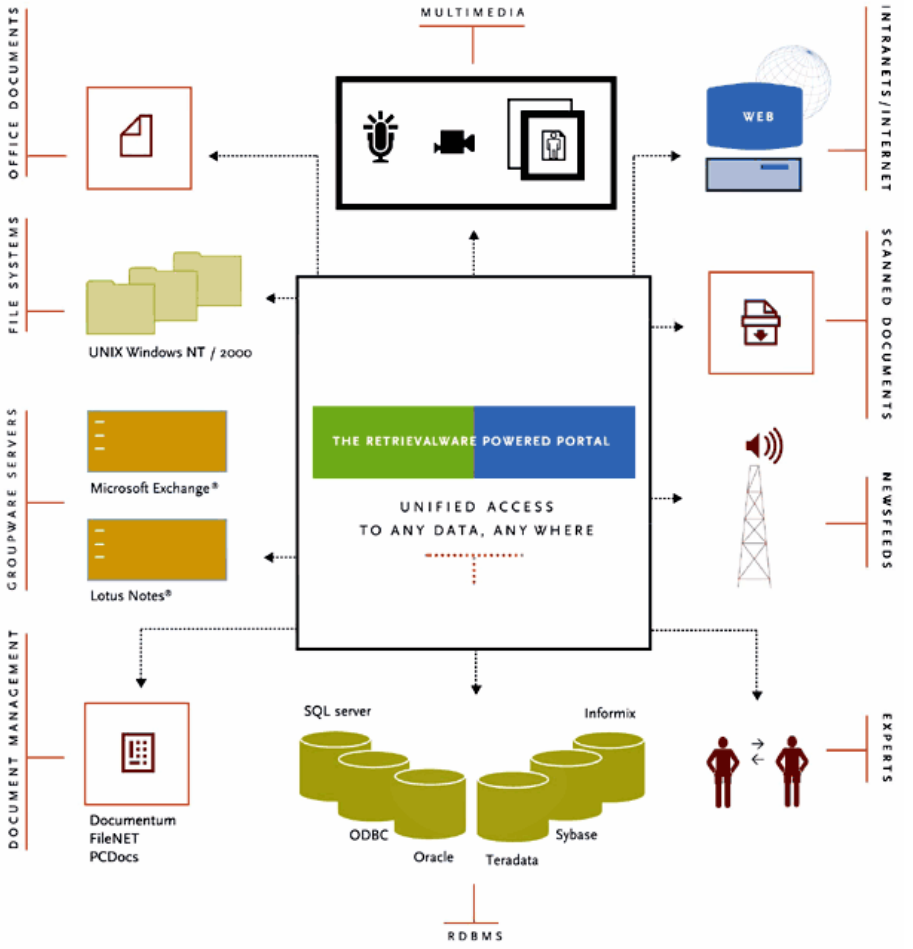
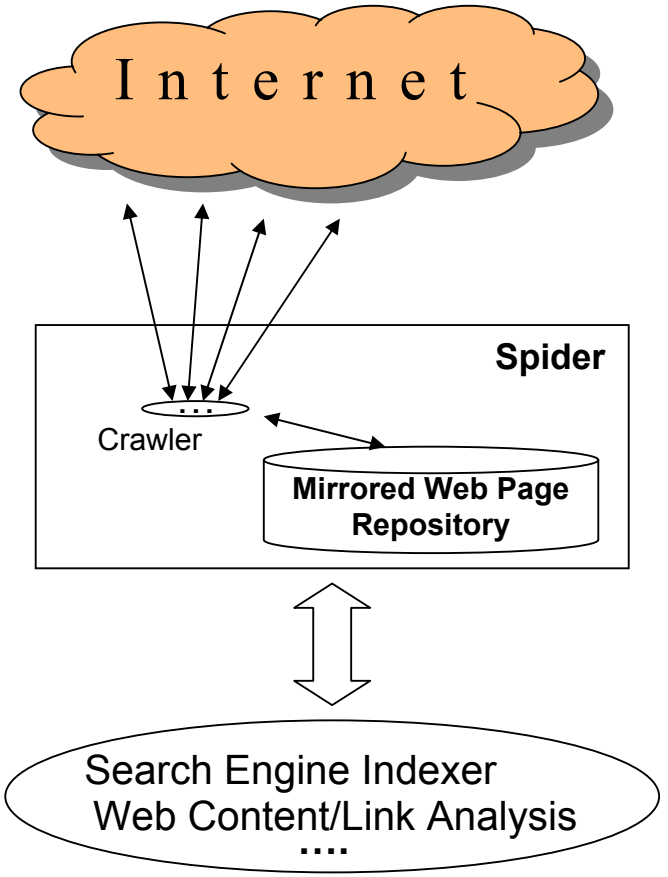


image now.
visionnext

Technologies used in Internet Search Engine

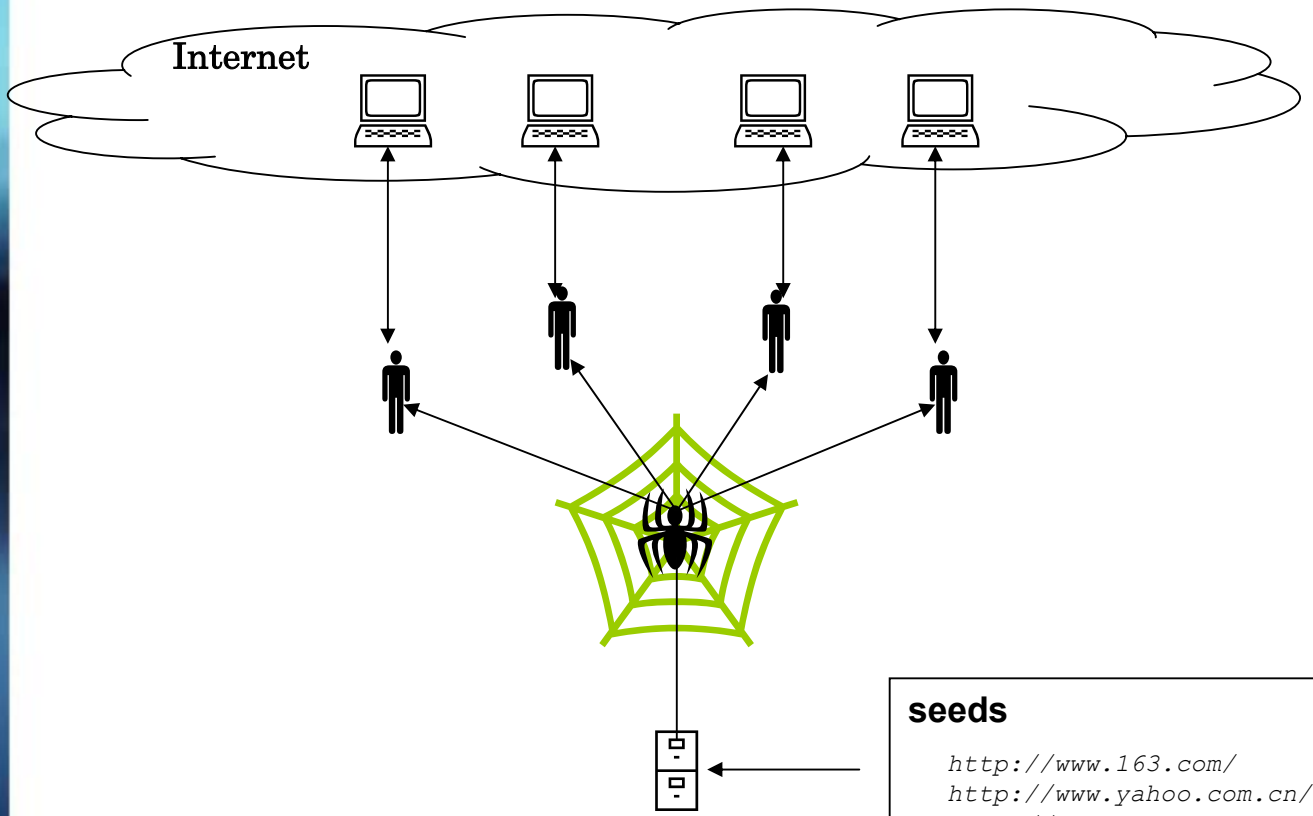
- **Information discovery**
  - Distributed system
  - Internet technology
    - Networking
    - DNS, IP
    - HTML
  - Storage system
  - Duplicate content detection
  - Information filtering
- **Index and Search**
  - Natural Language Processing
    - Spelling check
    - Term stemming
    - Thesaurus handling
  - Data structure for fast retrieval
    - Inverted file is industrial standard for text retrieval
    - Distributed index
    - Storage system design to minimize disk access
  - Cluster computing for scalable search
    - Google uses more than 15000 Linux PCs
    - Load balance issue
    - High availability issue
  - Multi-dimension index for multimedia content

Spider: Information Discovery on Inter net

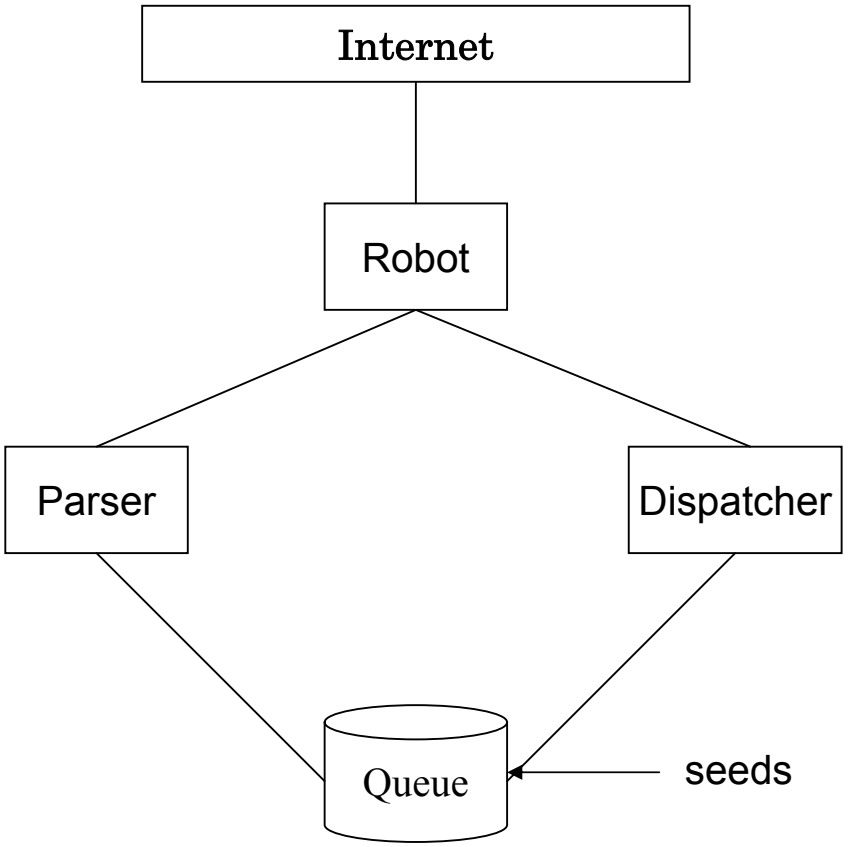


- **Affairs of a spider**
  - Crawl and explore the Web space
  - Maintain the fresh ness of the crawle d pages

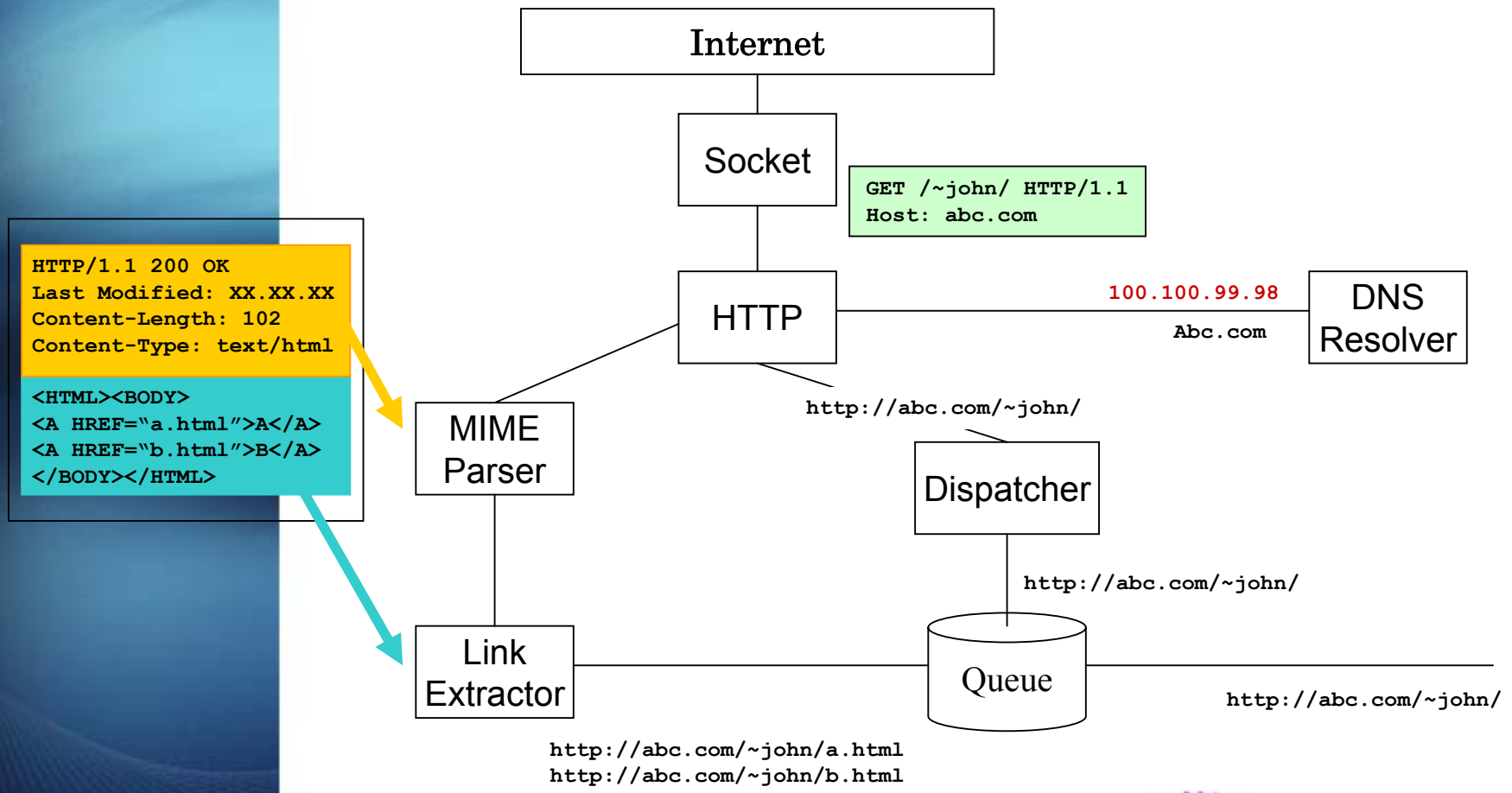
View of Crawling Process



One-Step Crawling Process



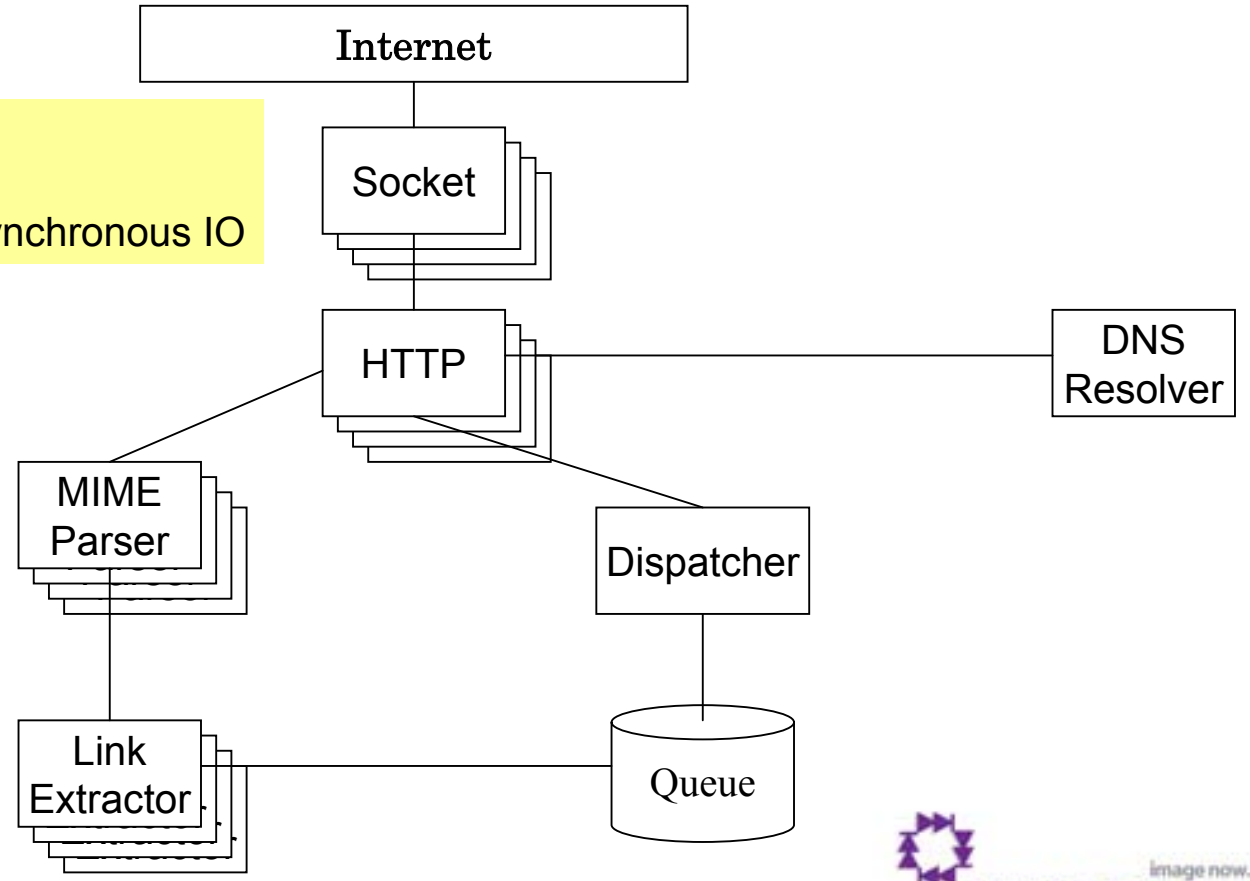
One-Step Crawling Process *(cont.)*



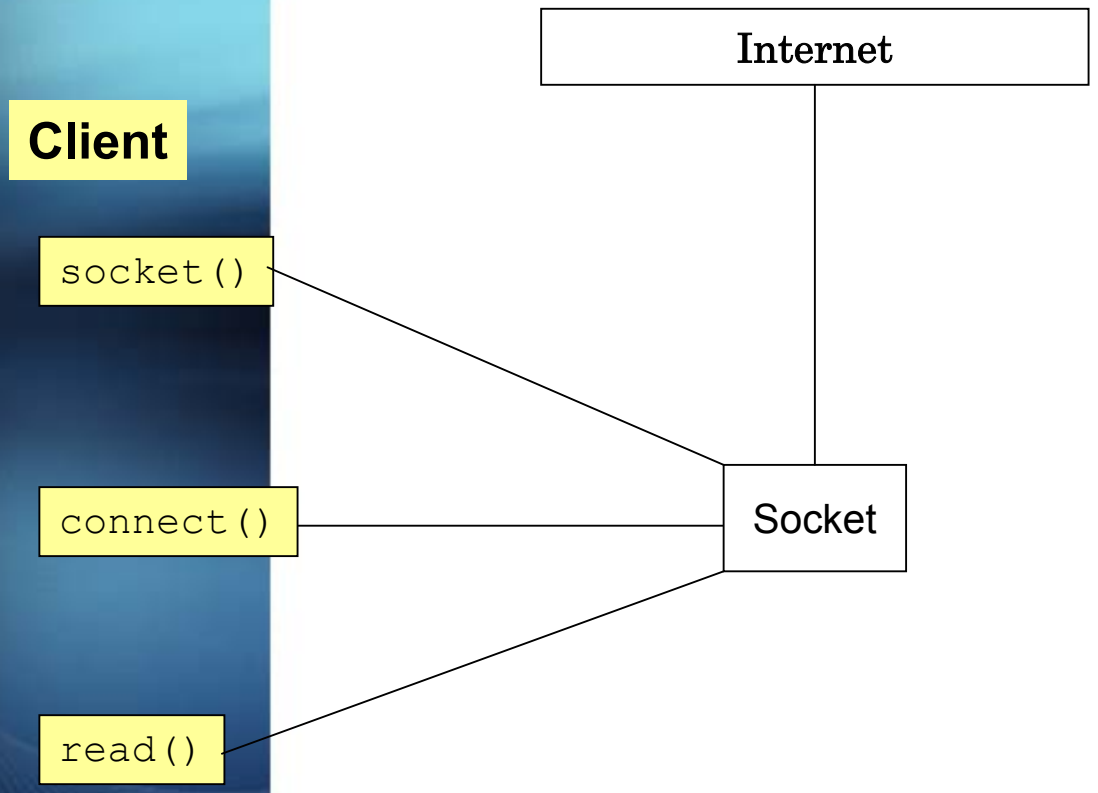
Parallel Crawling Process

Parallel Crawling

- Multi-processes
- Multi-threads
- One Process with Asynchronous IO



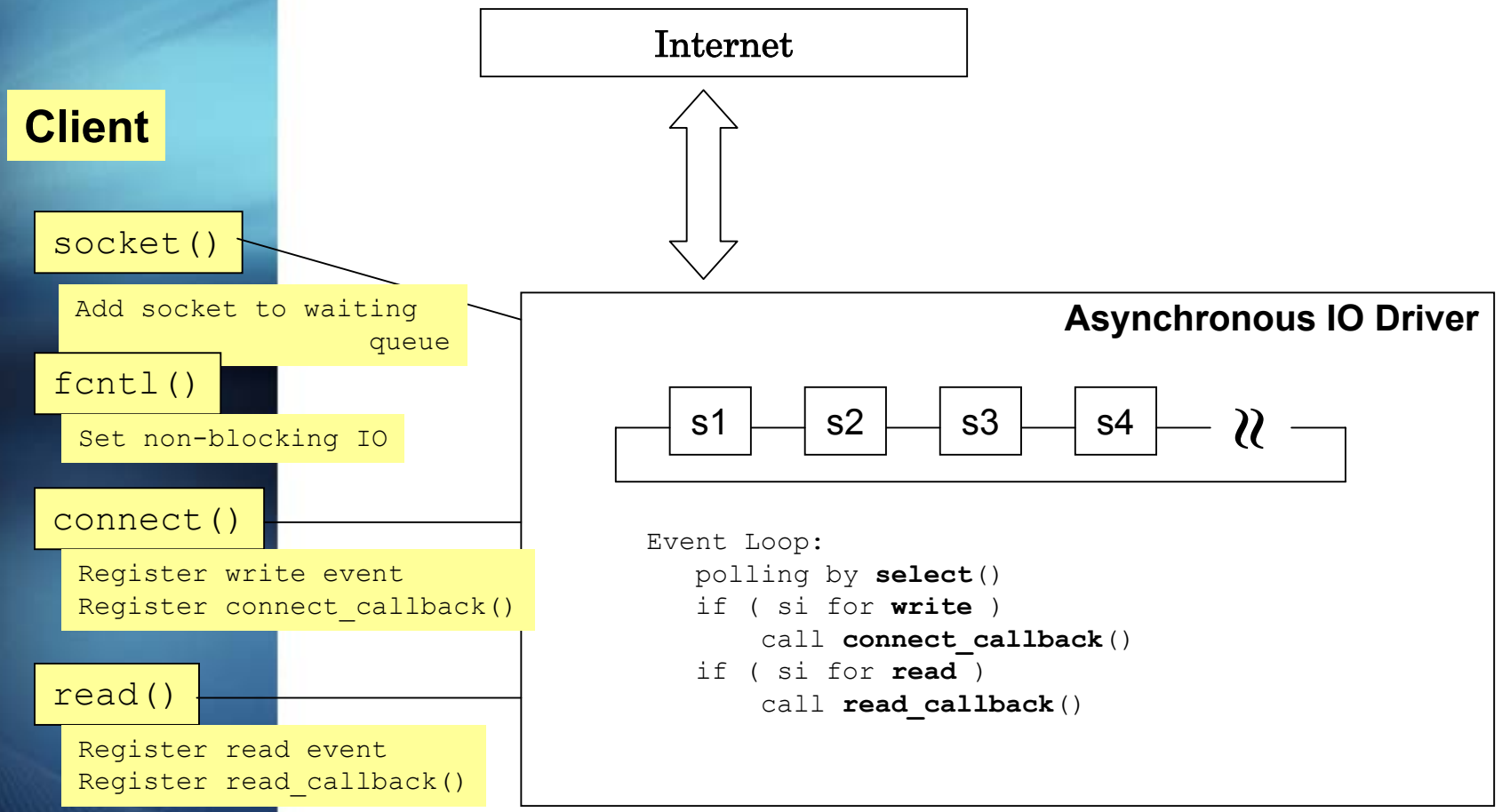
Socket Review



Default IO State:
Synchronous IO

Drawback:
The process is blocked on IO

IO-Driven Spider Infrastructure
(Asynchronous IO Driver)

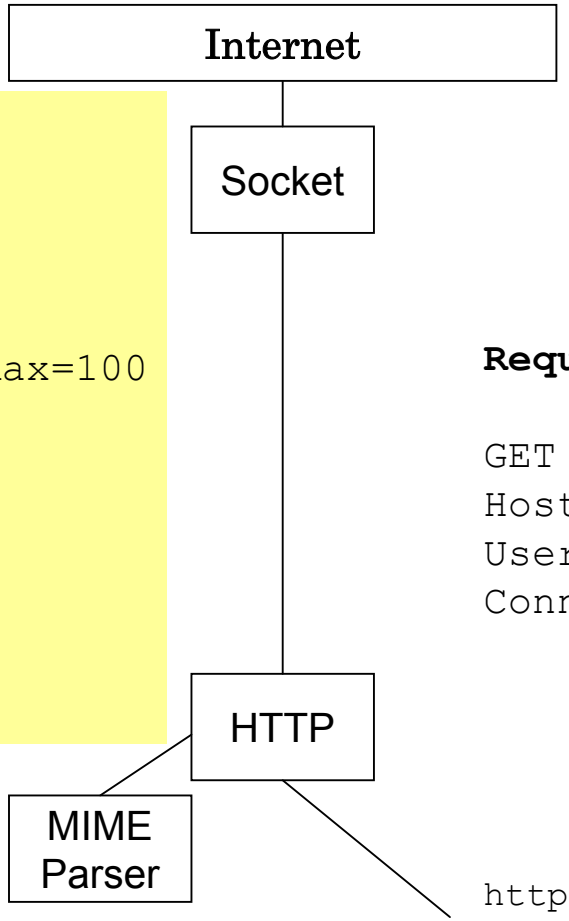


HTTP & MIME Header

Response:

```
HTTP/1.1 200 OK
Server: XXXX
Last-Modified: XXXX
Keep-Alive: timeout=15,max=100
Content-Length: 102
Content-Type: text/html

<HTML><BODY>
<A HREF="a.html">A</A>
<A HREF="b.html">B</A>
</BODY></HTML>
```



Request:

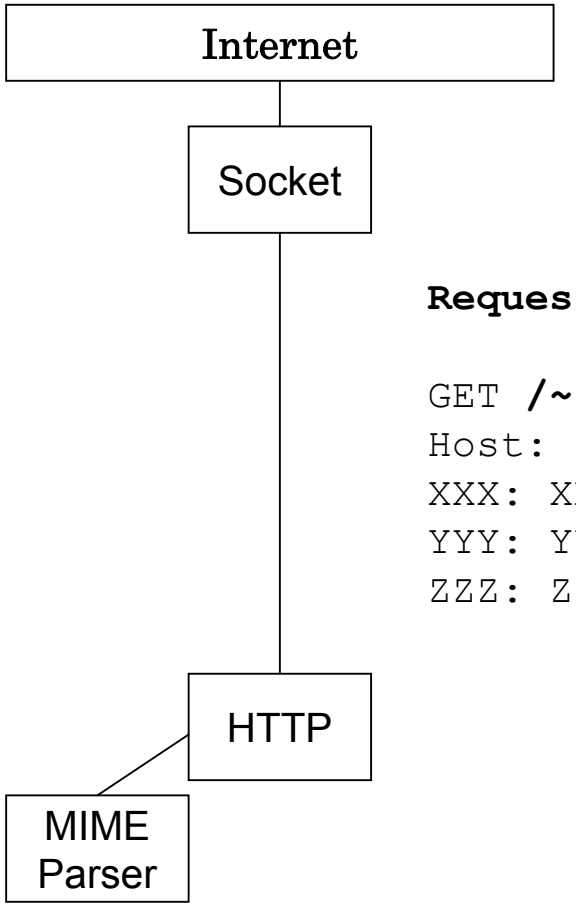
```
GET /~john/ HTTP/1.1
Host: abc.com
User-Agent: My Spider
Connection: Keep-Alive
```

`http://abc.com/~john/`

Redirection

Response:

```
HTTP/1.1 302 Found
Location: /~john/
```



Request:

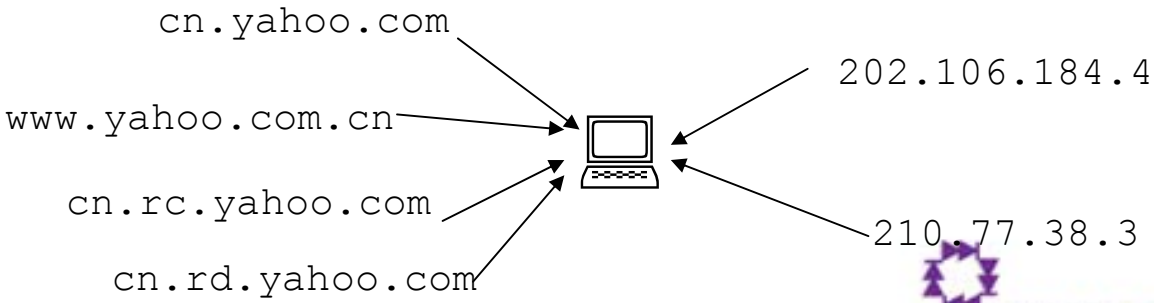
```
GET /~john/ HTTP/1.1 1
Host: abc.com
XXX: XXX
YYY: YYY
ZZZ: ZZZ
```

Link Extractor

- **Parse the HTML document and ext ract all the links that we are inter ested in**
- **Sample**
  - <A HREF="…">
  - <FRAME SRC="…">
  - <AREA HREF="…">
  - <META HTTP-EQUIV="refresh" CONTENT= "0; Url=/index.shtml">

Canonical Form of a URL

- **Canonical Form of a URL**
  - Normalization: A URL string is normalized by following steps:
    - Removal of the protocol prefix (http://) if p resent
    - Removal of :80 port number if present (Ho wever, non-standard port number are retai ned)
    - Conversion of the server name to lower cas e
- **Problem**

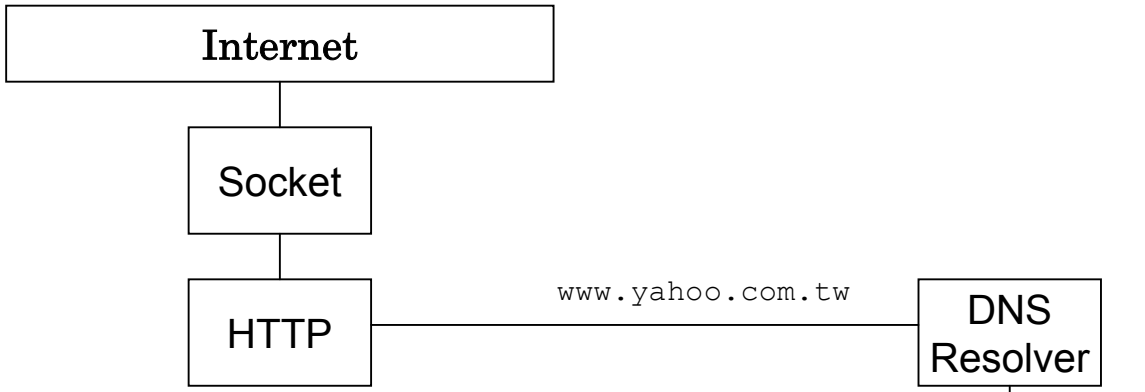


DNS Lookup

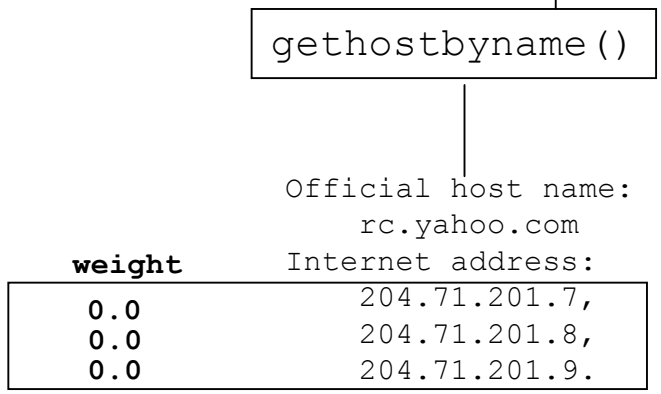
Contact the Domain Name Service (DNS) to resolve the host name into its IP address

- **Problem:**
  - DNS resolution is a well-documented bottleneck of m ost web crawlers
  - Most system DNS lookup implementation is synchron ized
- **Strategies:**
  - Keep a local host-to-IP cache to decrease the overhe ad by the default DNS lookup routine (e.g. *gethostby name*)
  - Or, implement a non-synchronized DNS resolver

DNS Lookup *(cont.)*



Special deal with **multi-homed** host:
- *choose the fastest IP to connect*



Update weight:

1. $w(n+1) = \alpha\, w(n) + (1-\alpha)\, \delta t$
2. $w(n+1) = w(n) * decay$

δt: connection time

α: sensitivity factor

Filter

Prevent reloading visited documents or downloading unnecessary ones

- **Problem:**
  - Host name alias

    i.e., multiple host correspond to the same IP
  - Alternative paths on the same host

    i.e., symbolic links
  - Replication across different hosts

    e.g., site mirroring
  - Non-indexed documents such as images, *.zip, *.mp3, etc.

Filter *(cont.)*

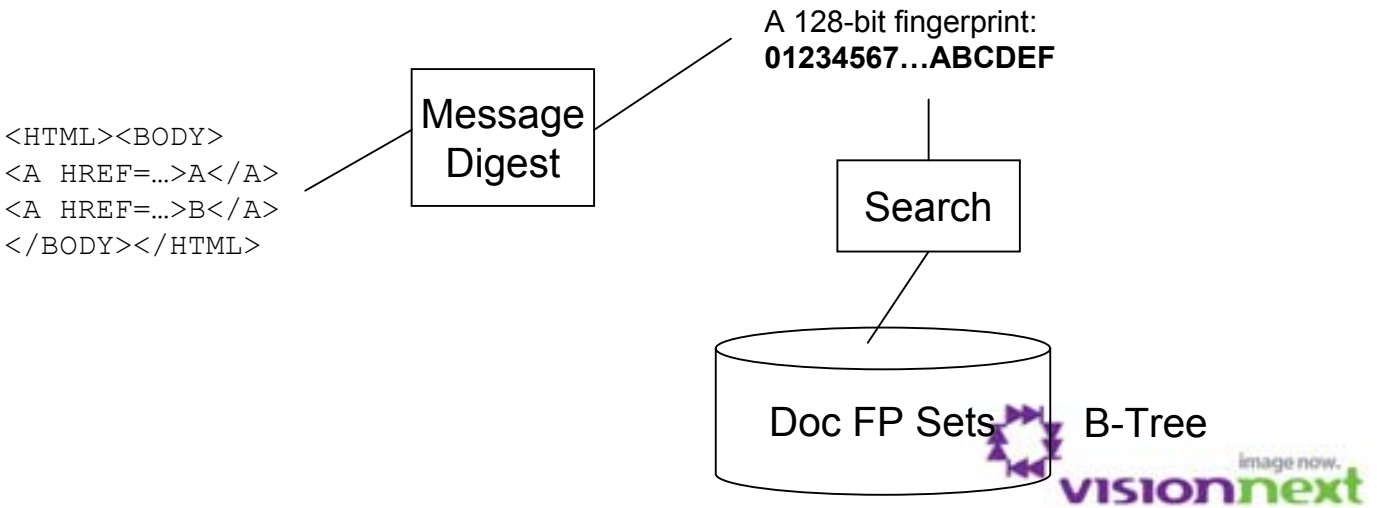
- **Strategies:**
  - URL constraints
    - Specify some regular expression rules for d omain, ip, prefix, protocol type, file suffix, e tc.

      e.g., `exclude="\.mp3$|\.jpg$|\.gif$"`
      `include="htm$|html$|/[^/\.]*$"`
      `dn-cst="\.cn$"`
      `ip-cst="2`
      `;211.100.0.0:0.0.127.255`
      `;61.128.0.0:0.3.255.255"`
  - URL-seen test
    - Check whether a URL has been fetched:

Filter *(cont.)*

- Content-seen test
  - Check whether a document has been fetche d
  - Represent a document as a fixed-size finger print (e.g., MD5) and perform a fast search on the document fingerprint set to measure the document resemblance

e.g.,



Robots Exclusion

To be polite in the crawling process

- **Strategies:**
  - Follow the Robots Exclusion Protocol

    e.g.,

    http://www.example.com/robots.txt

    ```
    # robots.txt for http://www.example.com/
    User-agent: *
    Disallow: /privacy/
    Disallow: /personal.html
    ```

  - Obey the Robots Meta Information

    e.g.,

    ```
    <META NAME="robots" CONTENT="nofollow,noindex">
    ```

Other Spider Issues

- **Dispatching URLs**
  - Prevent overloading one particular web server
    - Only one robot is responsible to one server at one time
- **Recovering from failures**
- **Keeping the network bandwidth in good use**
  - Keep as much connections (roughly several hundreds) as possible at the same time
- **Cache strategies**
  - Keep in-memory caches for those steps with high locality
    - DNS lookup, URL-seen test are but Content-seen test is not

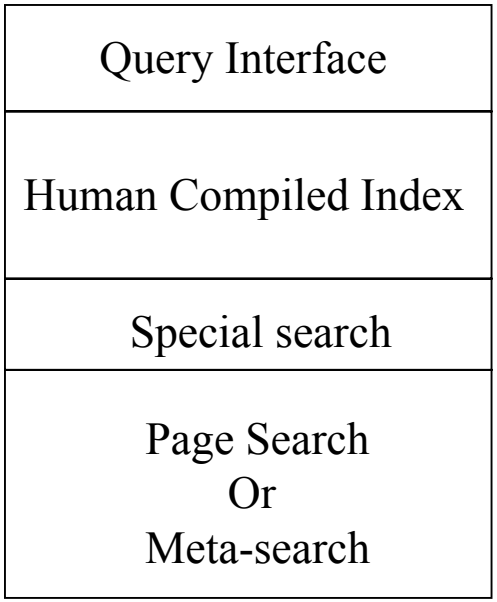
Human Index for the Internet

- **High precision, low recall**
- **Subject directory tree(Yahoo!)**
- **Expert guide (about.com)**
- **Q&A search (ask.com)**
- 你問我答 **(ExpertCentral.com)**

Automatic Indexing for the Internet

- **Low precision, high recall**
- **Full-text index/scan(excite, lycos, infoseek)**
- **Large scale indexing(Alta Vista)**
- **Popularity based indexing (Direct Hit)**
- **Search result clustering (Northern Light)**
- **Link Analysis(Google)**

A traditional Internet Search Engine



Ask Jeeves, Oingo, SimpliFind…

Yahoo!, Looksmart, Open Directory

Realnames, DirectHit, “shopping search”, Ad

Alta Vista, Google, Northern Light…

Search engine 愈來愈專業分工化；新search engine 往往專注於其重點特色的發展，其他的component 則與他人合作,以求快速發展

Challenge of Internet Search Engine

- 蒐集資料的困難性: 資料隨時可能更新, 變動
- 大量資料處理困難. **Indexing, programming** 與系統管理的複雜度均很高
- 資料品質良莠不齊, 搜索引擎需在大量資料中找的快, 更需找的好
- **Internet Search Engine** 往往是人潮聚集之處, 欲提供快速精確的服務, 需要較高的成本在**Hardware** 上, 此外, **Network Bandwidth** 需求極大, 所需成本上升
- 用戶的查詢通常很短, 提供的資訊量不足, 較難提供使用者好的查詢結果

Index

- **Manual Index**
- **Index system design Consideration**
  - Index Speed
  - Retrieval Speed
  - Incremental indexing
  - Index size
  - Compression
  - Distributed index
- **Determine object importance**
  - Hyperlink citation
  - Manual selection
  - Term weighting
- **Replicate of indexes to increase index availability**
- **Full-text index**
  - Inverted file
- **Database index**
  - B-Tree

Search

- Cluster computing
- Pre-compute everything at index time as possible
- Cache mechanism
- Top 400 results are enough for users instead of fin ding out more than 1 million results
- Minimize internal network communication
- Minimize disk access time in search time
- Minimize the web page size

Operational cost

- **Management of server farm**
  - System failure handling
  - Power management
- **Spider operation**
- **Index & search server switching operation**

Future Search Engine Improvement

- **Search algorithm side**
  - Term weighting
  - Document importance weighting
  - Utilize user feedback information
  - Term disambiguation
- **Content side**
  - Annotate content with keywords
  - Machine understandable metadata, Ontolgy
  - Semantic Web
- **User side**
  - Clustering/categorization
  - Natural language query

Some new topics

- **Question Answering**
- **Wireless Search**
- **Cross-Language Search**
- **Multimedia Search**
- **Recommendation**
- **Information Filtering**

Multimedia Search

- **Speech Retrieval**
- **Web Image Retrieval**
- **Content-based Image Retrieval**
- **Video Retrieval**
- **Music Retrieval**

網路圖片搜尋


PC home Online 網路家庭-圖片搜尋 - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛 記錄
網址(D) http://image.pchome.com.tw/cgi-bin/searchpic.cgi?type=button&query=%B0%EA%A4%FD%A5%F8%C3Z&Submit=%B7j%B4M&
移至
PC home > 圖片搜尋
找圖 國王企鵝 搜尋
全部 只找大圖（200X200以上）
POWERED BY
VisionNEXT
PC home圖片搜尋為您找到35 張 符合 國王企鵝 的圖 頁次：1/4
相關查詢 企鵝 無尾熊
國王企鵝 King Penguins,peuguin
(768 X 512)
圖片資訊..
國王企鵝 King Penguins,peuguin
(768 X 512)
圖片資訊..
國王企鵝家族~
(385 X 403)
圖片資訊..
國王企鵝 King Penguins,peuguin
國王企鵝 King Penguins,peuguin
國王企鵝~
開始
disse...
NTU...
收件...
3.5 ...
chien...
PC h...
CH
下午 01:17

以圖查圖


eefind 網際圖片搜尋 - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛 記錄
網址(D) http://www.want2.com.tw/cgi-bin/cbir.cgi?query=t146481.gif
移至
排列方式 4 行 4 列 重排 共 100 張圖 7頁 回網圖
頁次 : 1 / 7
1 2 3 4 5 6 7
史努比
繁 68x102 3Kb
繁 94x105 3Kb
繁 110x149 3Kb
繁 108x138 5Kb
繁 148x161 4Kb
繁 230x138 6Kb
繁 97x79 3Kb
繁 132x120 4Kb
開始
disse...
NTU...
收件...
3.5 ...
chien...
eefin...
CH
下午 01:12

Discussion