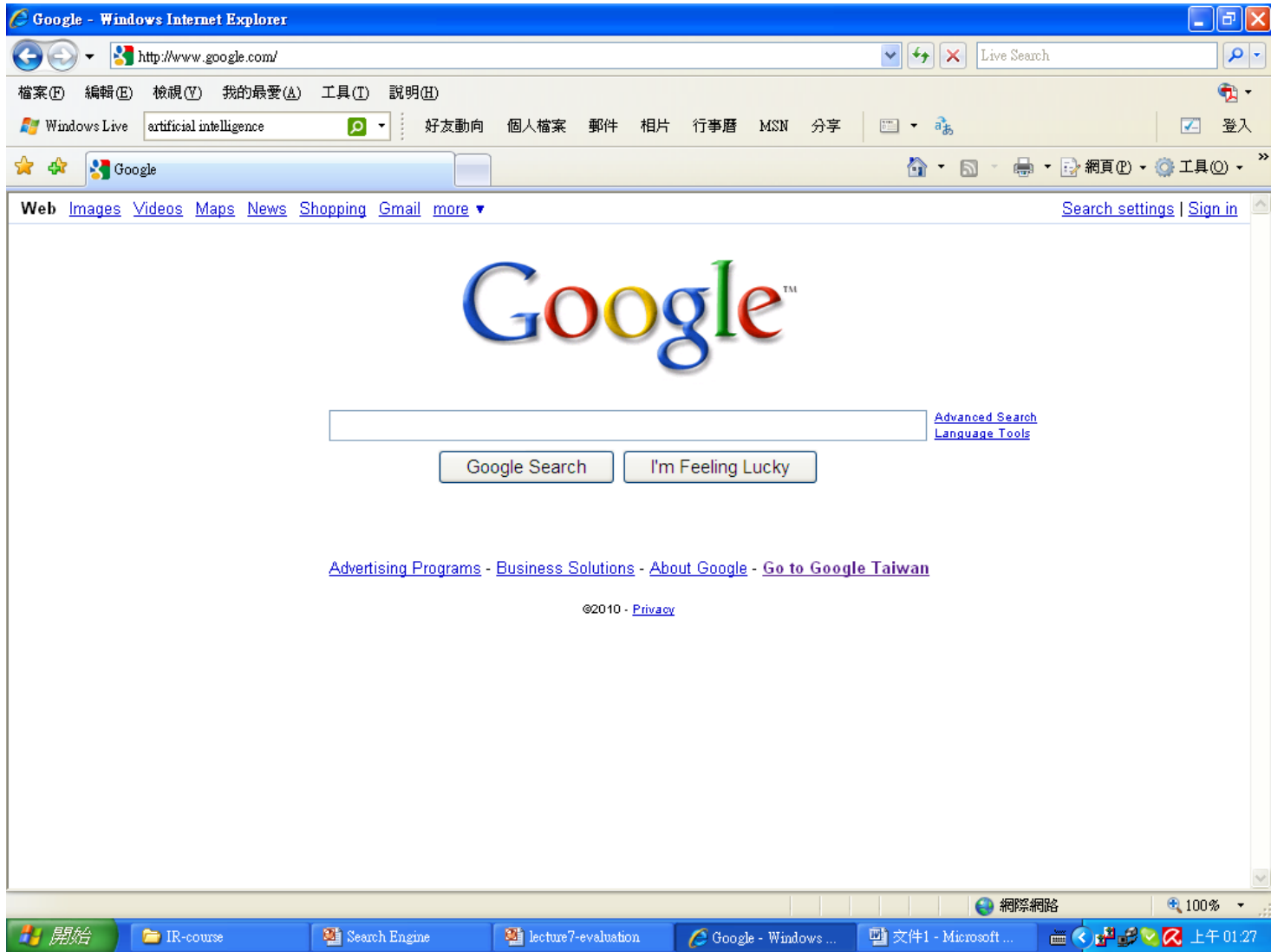
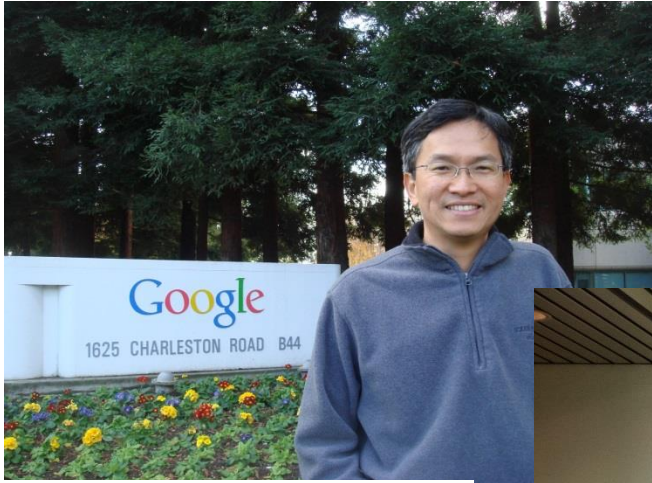




# Google 與搜尋引擎的應用





# Search Examples



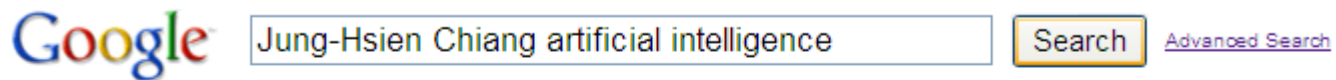
Web [+ Show options...](#) Results 1 - 10 of about 9,260,000 for ar

[Artificial intelligence](#) - [Wikipedia, the free encyclopedia](#)

**Artificial intelligence (AI)** is the intelligence of machines and the branch of computer science that aims to create it. Textbooks define the field as "the ...

[History](#) - [Problems](#) - [Approaches](#) - [Tools](#)

[en.wikipedia.org/wiki/Artificial\\_intelligence](http://en.wikipedia.org/wiki/Artificial_intelligence) - [Cached](#) - [Similar](#)

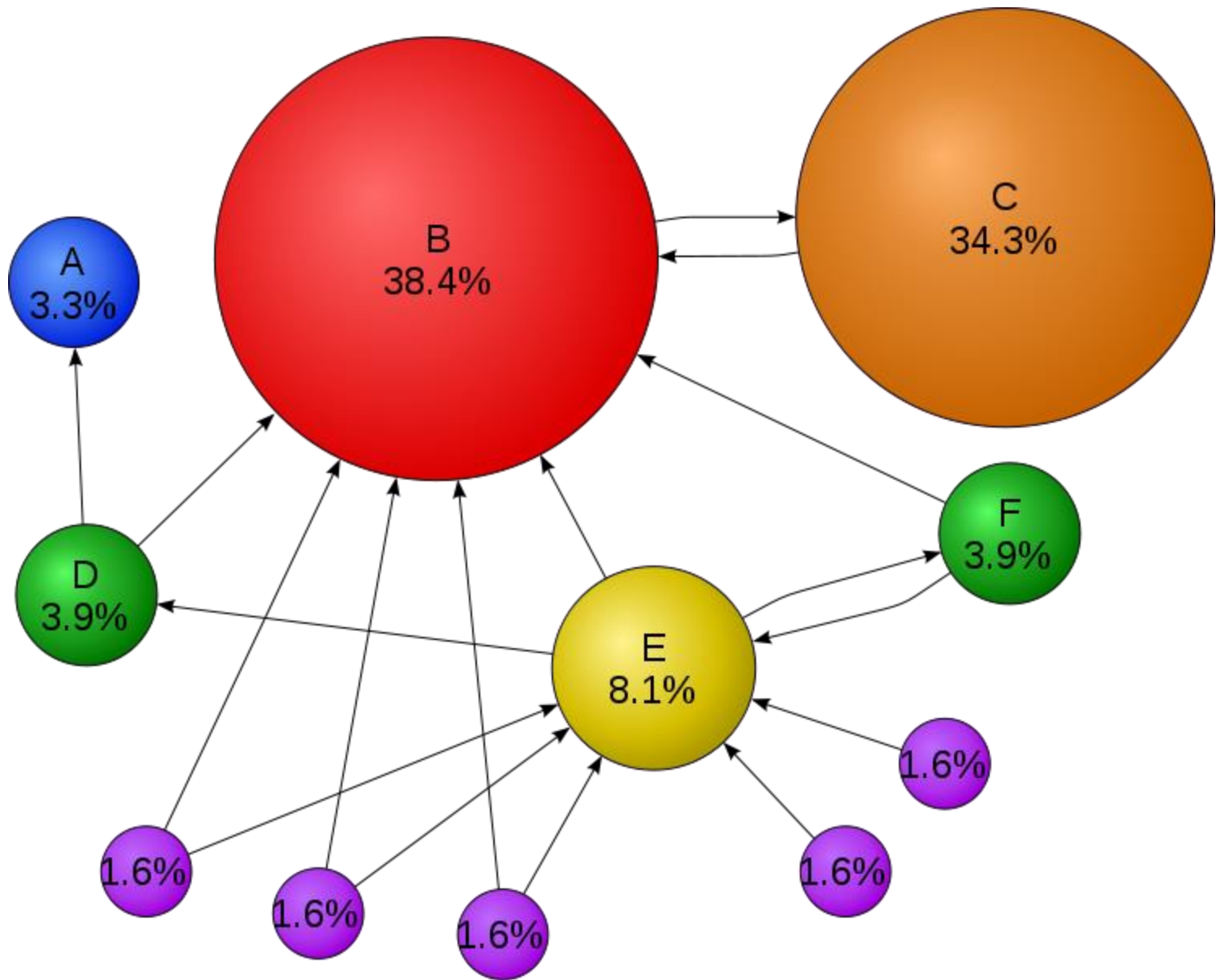


Web [+ Show options...](#) Results 1 - 10 of about 134,000 for Jung-Hsien

[DBLP: Jung-Hsien Chiang](#)

**Jung-Hsien Chiang:** A Fuzzy Route Guidance Model for Intelligent In-Vehicle Navigation Systems. *International Journal on Artificial Intelligence Tools* 8(2): ...

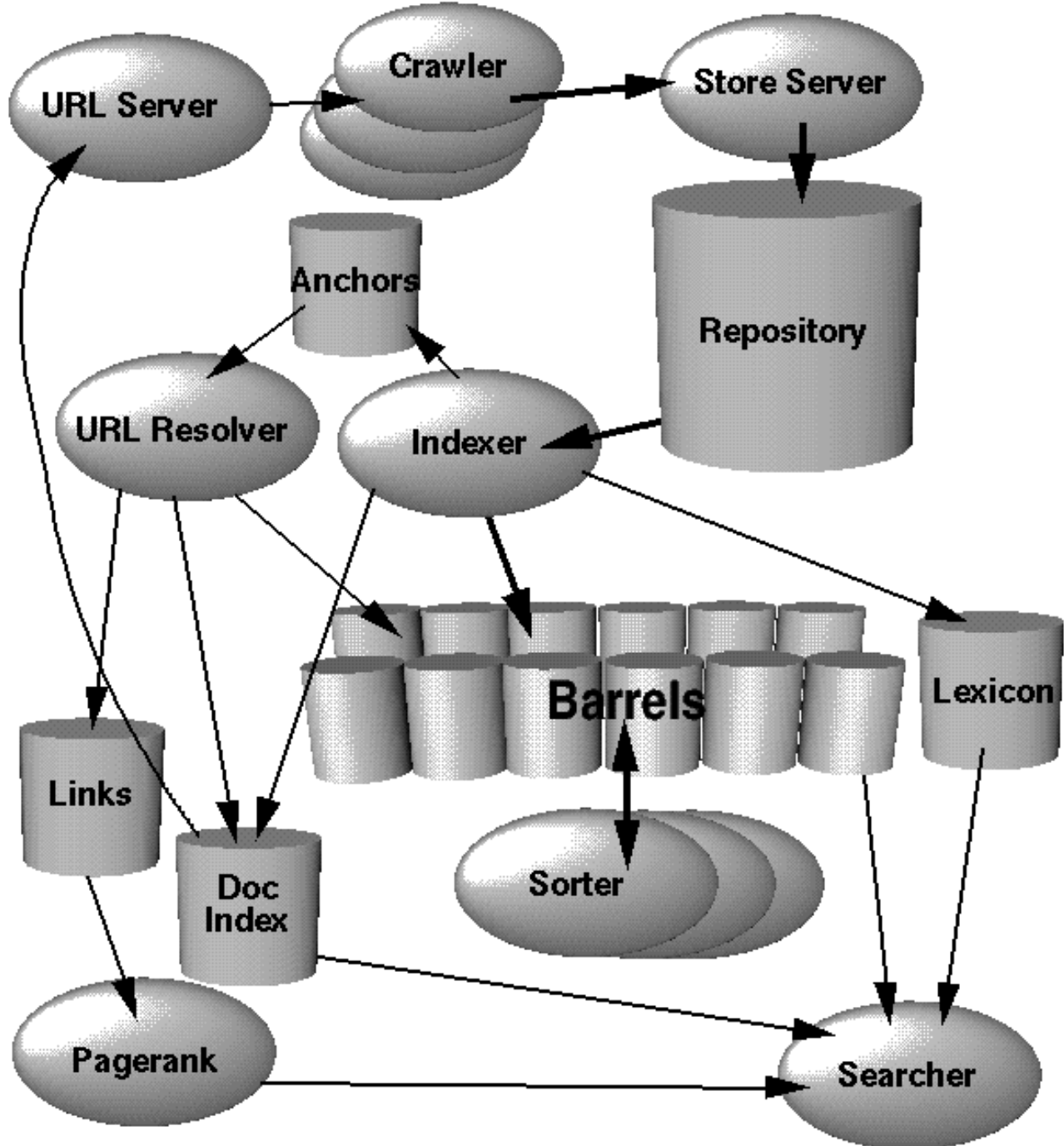
[www.informatik.uni-trier.de/~ley/db/.../a.../Chiang:Jung=Hsien.html](http://www.informatik.uni-trier.de/~ley/db/.../a.../Chiang:Jung=Hsien.html) - [Cached](#)



# System Features

1. **PageRank**: bringing order to the Web
2. Anchor Text
  - Associate the text of a link with both the page that the link is on and the page the link points to (**anchor propagation**).
  - Anchors often provide more accurate descriptions of web pages than the pages themselves.

# Google Architecture



# Major Data Structures

## Hit Lists

- A **hit list** corresponds to a list of occurrences of a particular word in a particular document including **position**, **font** (relative), and **capitalization** information.
- Two types of hits:
  1. **Fancy hits**: hits occurring in a URL, title, anchor text, or meta tag.
  2. **Plain hits**: everything else.



Hit: 2 bytes

plain:	cap:1	imp:3	position: 12		
fancy:	cap:1	imp = 7	type: 4	position: 8	
anchor:	cap:1	imp = 7	type: 4	hash:4	pos: 4

# Major Data Structures:

Forward and Reverse Indexes and the Lexicon

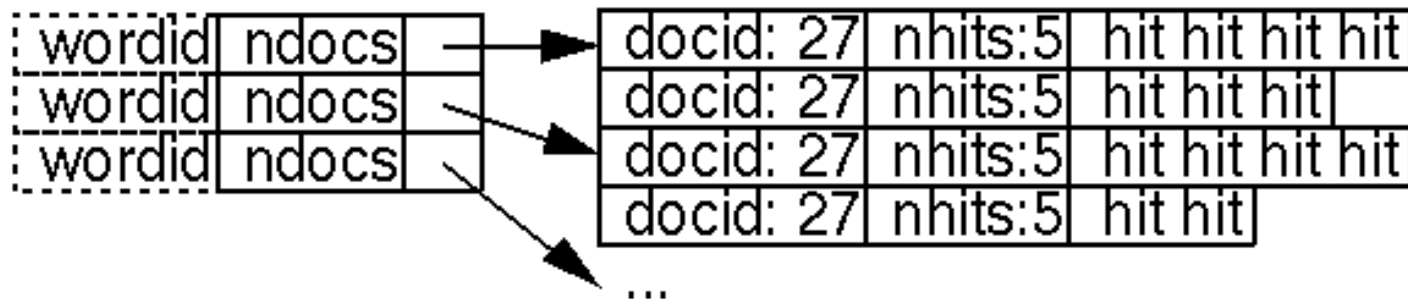
Forward Barrels: total 43 GB

docid	wordid: 24	nhits: 8	hit hit hit hit
	wordid: 24	nhits: 8	hit hit hit hit
	null wordid		
docid	wordid: 24	nhits: 8	hit hit hit hit
	wordid: 24	nhits: 8	hit hit hit hit
	wordid: 24	nhits: 8	hit hit hit hit
	null wordid		

...

Lexicon: 293MB

Inverted Barrels: 41 GB



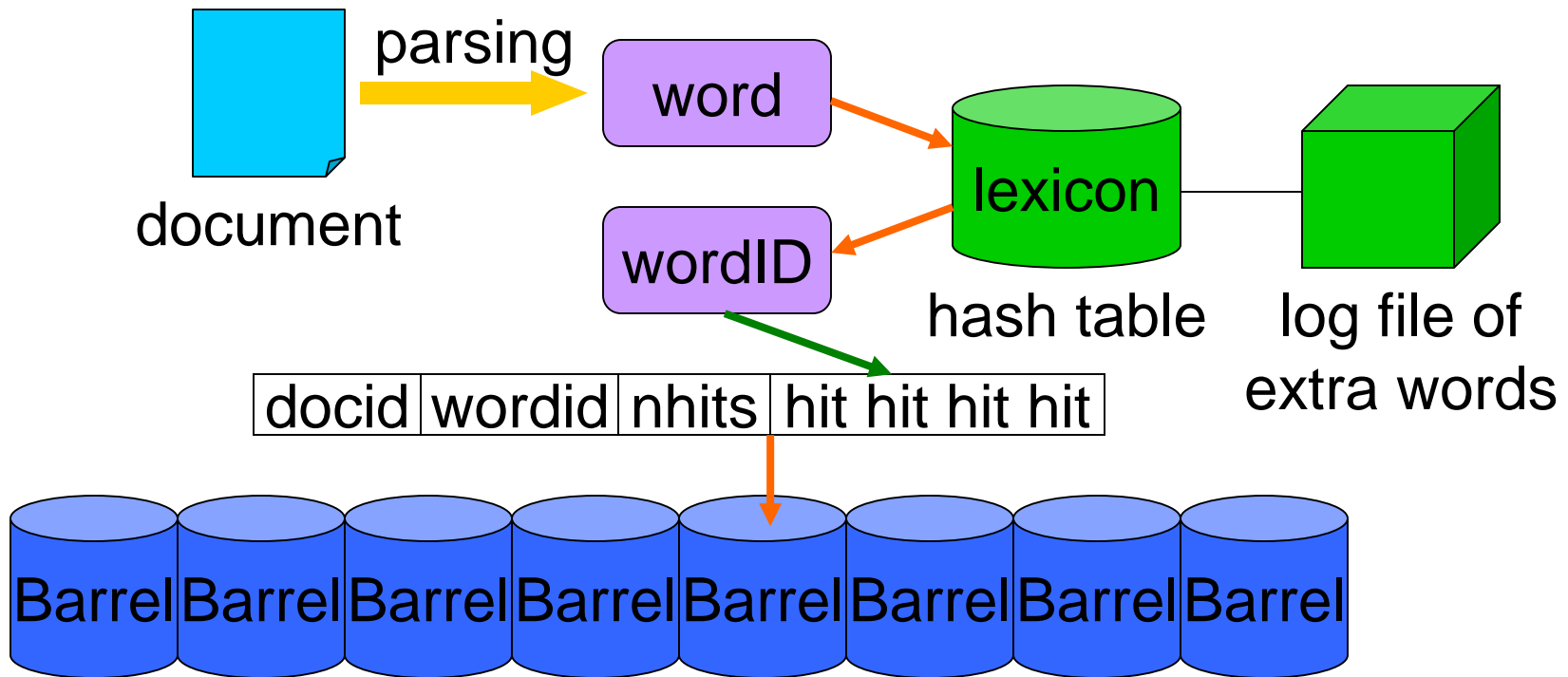
# Crawling the Web

- Google has a fast distributed crawling system.
- A single URLserver serves lists of URLs to a number of crawlers.
- Each crawler keeps roughly 300 connections open at once.
- Connection states: looking up DNS, connecting to host, sending request, and receiving response.
- DNS cache

# Indexing the Web

## Forward Index

- Indexing documents into barrels

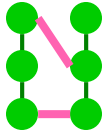


# Indexing the Web

## Inverted Index

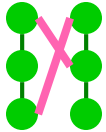
- Issue: in what order the docID's should appear in the doclist

1. Sorted by docID



This allows for quick merging of different doclists for multiple word queries.

2. Sorted by a ranking of the occurrence of the word in each document



This makes answering one word queries trivial and makes it likely that the answers to multiple word queries are near the start.

# Indexing the Web

## Inverted Index

- A compromise between these options, keeping two sets of inverted barrels
  1. One set for hit lists which include **title** or **anchor** hits
  2. Another set for all hit lists.
- Check the first set of barrels first and if there are not enough matches within those barrels, check the larger ones.

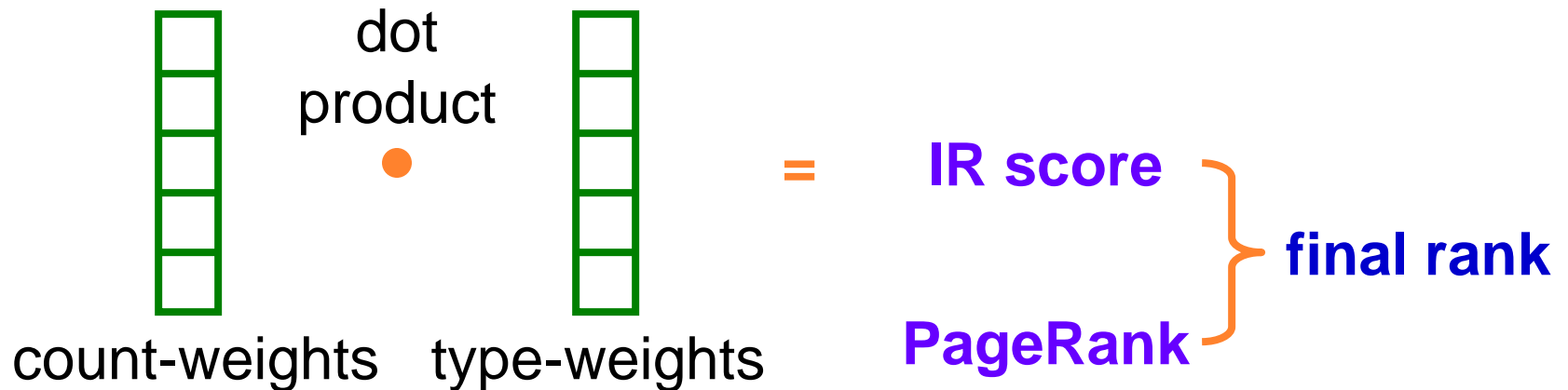
# Searching

1. Parse the query.
2. Convert words into wordIDs.
3. Seek to the start of the doclist in the short barrel for every word.
4. Scan through the doclists until there is a document that matches all the search terms.
5. Compute the rank of that document for the query.
6. If we are in the short barrels and at the end of any doclist, seek to the start of the doclist in the full barrel for every word and go to step 4.
7. If we are not at the end of any doclist go to step 4.
8. Sort the documents that have matched by rank and return the top k.

# The Ranking System

## 1. Single word query

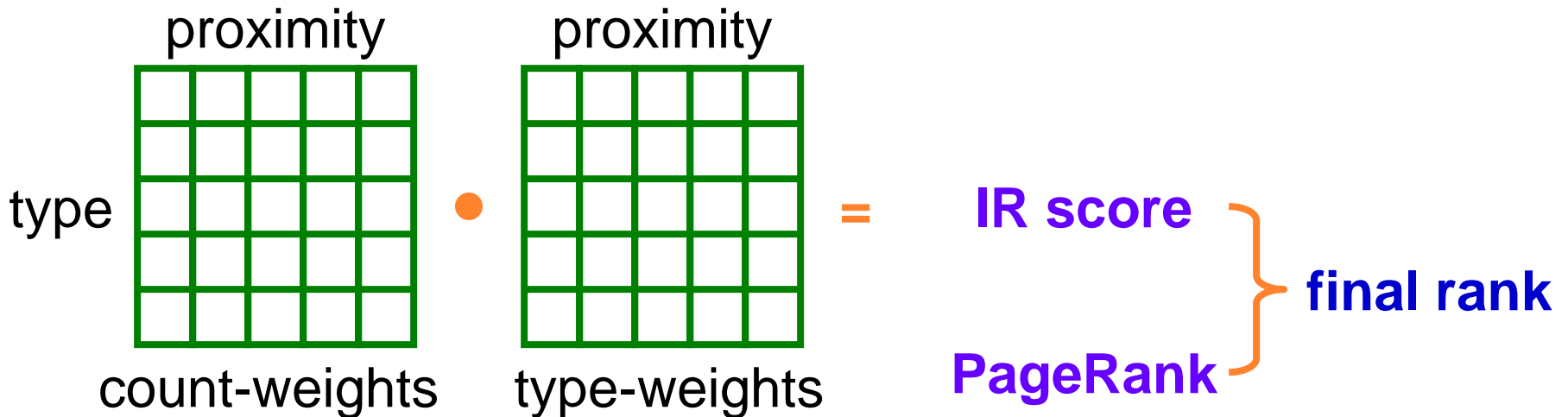
Consider each hit to be one of several different types (title, anchor, URL, plain text large font, plain text small font, ...).



# The Ranking System

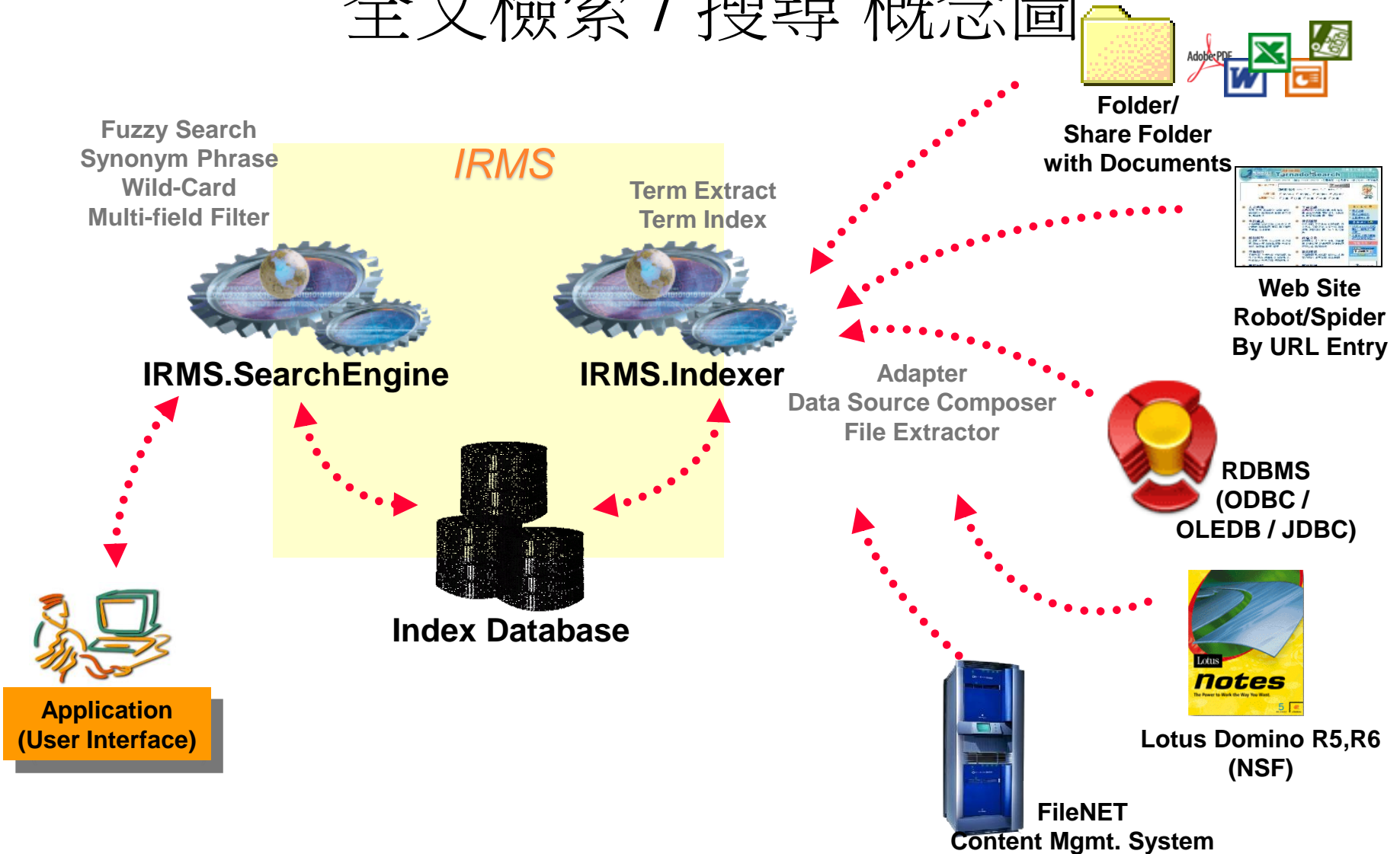
## 2. Multi-word query

Hits occurring close together in a document are weighted higher than hits occurring far apart.





# 全文檢索 / 搜尋 概念圖



# 搜尋引擎的定位與價值

- 資訊整合檢索
  - 將異質的資訊源進行整合的索引，以提供跨資訊源的搜尋服務。
- 非結構性文件之索引與搜尋
  - 針對非結構性文件、結構與半結構性的資訊可以透過檢索的技術來提供快速的搜尋。
- 搜尋效能與功能的滿足
  - 透過資訊索引相關的演算法，以達到快速搜尋以及特殊搜尋功能的滿足，彌補傳統循序搜尋於資料量大時的搜尋速度限制。

# 多國語言查詢與搜尋結果呈現

**IRMS 整合搜尋結果**

查詢關鍵字: 화제의 | するなど熱戦 | 年度營運計劃 | 創新哲學

搜尋範圍: 多國語言資料(9)

查詢時間: 6.78

符合筆數: 4

每頁顯示筆數: 15

- 바른언론 바른통신 인터넷 연합뉴스>뉴스>속보 (화제 의)(1)**  
(부산=연합뉴스) 특별취재단= 국내의 한 요트 선수가 16년만에 같은 자리에서 아시아게임 금메달을 따내 화제가 되고 있다. **화제의** 주인공은 이동우(해운대구청)와 짝을 이뤄 요트 남자 420급에 출전한 박종우(29). 지난 86년 서울아시아게임 때 15세 이하 종목인 옴니미스트에 출전해 금메달을 땀던 박종우가 이번 부산아시아게임에서 다시 금메달을 목에 건 것. 86년 당시 요트는 부산에서 열렸기 때문에 박종우는 16년만에 같은 장소에서 2번째 금메달을 따내  
[資料大小]35K [資料時間]2002/10/09 15:59:28 [資料來源]多國語言資料  
[資料類型]  
[資料連結路徑]C:\IRMS1\_01\EO\_SAMPLE\MultiLingual\KR\1559-sid213-07.htm
- 山陽新聞地域ニュース・倉敷・総社圏版 (2002年10月9日掲載) (するなど熱戦)(1)**  
「第8回トシビアンカップベタンク大会」(清音村ベタンク協会主催、山陽新聞常盤販売所共催)が6日、清音村三因のふるさとふれあい広場で開かれた。清音村、倉敷市から24チーム48人が出場。4チームずつ6ブロックに分かれた予選リーグの後、決勝トーナメントを行った。ベタンクはビュットと呼ばれる小さな球をめぐって鉄のボールを投げ合い、ビュットに、近い近づけるかを競うフランス生まれのスポーツ。選手らは真剣な表情で狙いを定め、相手のボールをまじき飛ばし、最後に大逆転 **するなど熱戦**を展開した。同村では、200  
[資料大小]18K [資料時間]2002/10/09 15:54:28 [資料來源]多國語言資料  
[資料類型]  
[資料連結路徑]C:\IRMS1\_01\EO\_SAMPLE\MultiLingual\JP\1554-sid212-09.htm
- “世纪之交的哲学”国际学术研讨会在北京召开 (创新哲学)(1)**  
中共中央政治局委员、中国社会科学院院长李铁映出席了研讨会的开幕式，并发表了题为“把握时代，**创新哲学**”的重要学术讲演。  
[資料大小]14K [資料時間]2002/10/09 14:43:34 [資料來源]多國語言資料  
[資料類型]  
[資料連結路徑]C:\IRMS1\_01\EO\_SAMPLE\MultiLingual\GB\1443-sid210-16.htm
- Untitled (年度營運計劃)(1)**  
能夠有這麼全面性、開創性的員工福利，是因為宏碁董事長施振榮的「福利事業化」概念。施振榮要求，人力資源處和福委會，必須把員工福利當做一項「事業 (business)」在經營，把所有同事當作「客戶」來看待。因此，華蒼幼兒中心、健身中心等單位的負責人，每年都要像各部門主管一樣，提出「**年度營運計劃**」，並且要為是否達到目標負責。在強大的營運壓力下，各福利事業無不卯足勁，以推陳出新的服務來滿足宏碁員工的需求。

Korean

Japanese

Simplified Chinese

Traditional Chinese

www **KingStone** .com.tw 金石堂網路書店

我要找書

書名

[加入會員](#) | [會員管理](#) | [訂單查詢](#) | [看購物車](#) | [Q&A](#) | [與我聯絡](#)

玲妮與娃娃屋的世界 DOLLY 

[首頁](#) | [書籍區](#) | [雜誌區](#) | [MOOK](#) | [嚴選百貨區](#) | [鮮書區](#)

| [搶鮮預購](#) | [新書](#) | [暢銷書](#) | [推薦書](#) | [特價書](#) | [個人化書店](#) |  
哈利波特魔法紙牌大方送，麻瓜們手腳要快！



全省皆可到店取貨

>>主題館



- 小說街 **NgW**
- 年度Top館 **NgW**
- 充電學習館
- 網路小說館

>>個人化書店



露 · 純喫茶女孩



推薦指數：  
**NEW**

What Management Is 管理是什麼

作者：瓊安·瑪 譯者：

出版社：

ISBN：

出版日期：92年05月26日

中文平裝版

定價400元 售價/320元  
折扣/8折! 你省了 80元

參考庫存量:19本

書籍簡介

在這本書中，所學的管理個案從工業時代的福特汽車到網路時代的電子海灣，從戴爾電腦到NBA、眼科醫院及紐約市立動物園。管理的應用範圍這麼廣，只要有組織的存在，就需要管理。

書籍詳介

相關閱讀

--- 出版社專櫃 ---

· 暢銷書 more...

1. 默默領導
2. 共好
3. 天才當家
4. 成功兵法
5. 工商管理 <上>

· 新書 more...

1. 經營管理實務
2. 領導者62個禁忌!
3. 管理者必要的思考!
4. 用人有學問
5. 預測變革21世紀變革之道

· 相關類別

[管理總論](#)

[成功範例](#)

# 104 人力銀行

104全職工作全文複式檢索 - 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 媒體

網址(D) http://www.104.com.tw/cfdocs/2000/new\_fts/fulltime\_fts.cfm 移至

連結 BI Radar 商情雷達系統 Hotmail的免費電子郵件 Windows Media Windows 自訂連結

工·作·全·文·複·式·檢·索 www.104.com.tw

### 104全職工作全文複式檢索

● 以下欄位職務別、產業別、工作地區等，請按 ... 叫出選單。

- 1 請選取您要找的職務類別：  
 ...
- 2 請選取您的希望工作地點： / 可複選 /  
 ... 或  ... 或  ...
- 3 請選取您希望從事的產業別：  
 ...
- 4 請輸入關鍵字：  
工作條件  / 請輸入有關求才職務必備條件的關鍵字，如程式設計 /  
福利條件  / 請輸入有關公司福利的關鍵字，如員工分紅 /  
公司服務  / 請輸入有關該公司介紹與產品服務的關鍵字，如汽車銷售 /

【「+」表示「或」，如網頁+設計表示要有網頁或設計其一即可；「\*」表示「和」，如網頁 \* 設計表示既要有網頁且要有設計】

開始查詢

完成 網際網路

開始 下午 03:50

# 工程規範管理 系統操作手冊

工程規範管理 - 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 媒體

網址(D) http://127.0.0.1/KMRT/ERuleDoc.nsf 移至 連結 »

## 工程規範管理

**搜尋**

搜尋類型	<input type="radio"/> 分類查詢 <input checked="" type="radio"/> 進階查詢
全文檢索字串	工程
每頁顯示筆數	<input type="radio"/> 5 <input type="radio"/> 10 <input checked="" type="radio"/> 15 <input type="radio"/> 20 <input type="radio"/> 25 筆
文件狀態	<input checked="" type="radio"/> 全部 <input type="radio"/> 現行文件 <input type="radio"/> 歷史文件
文件名稱	
核准/頒訂日期	~
文件製訂單位	
文件核准/頒訂單位	
本局主辦單位	
核准/頒訂日期排序	<input checked="" type="radio"/> 升羈排列 <input type="radio"/> 降羈排列

開始搜尋

正在開啓網頁 http://127.0.0.1/KMRT/ERuleDoc.nsf/Search?OpenForm&Seq=2... 網際網路

**Anonymous**

- 文件管理
  - 目錄依編號
  - 目錄依階層
  - 全部文件依編號
  - 全部文件依階層
  - 草稿
  - 被駁回
  - 待審核
  - 現行文件
- 查詢檢索
- 報表列印
- 使用者說明
- 系統管理
  - 設定關鍵字
  - 資料備存
- 回局內首頁
- 登入
- 登出
- 變更密碼