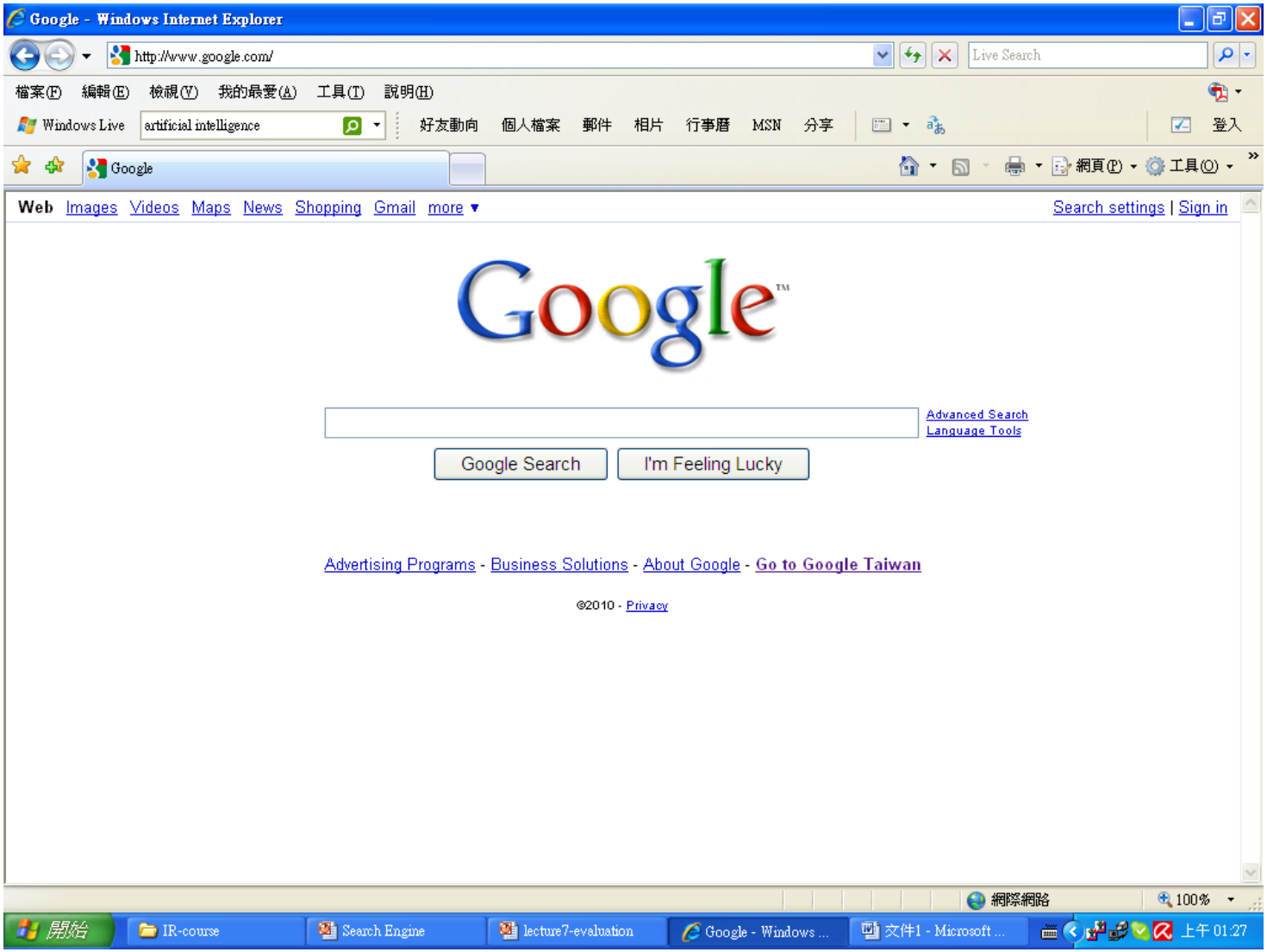
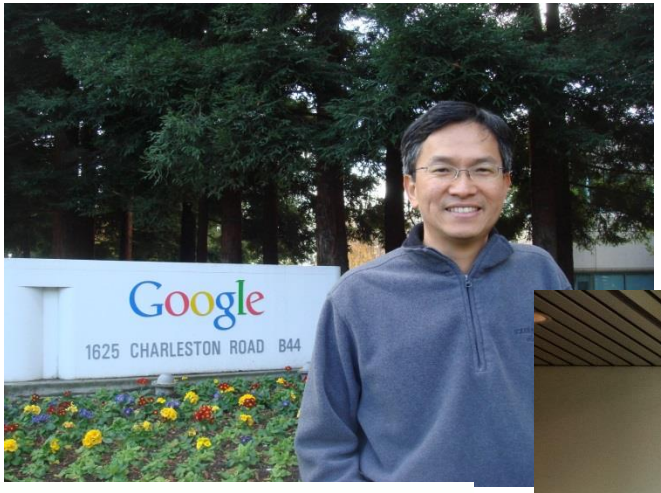
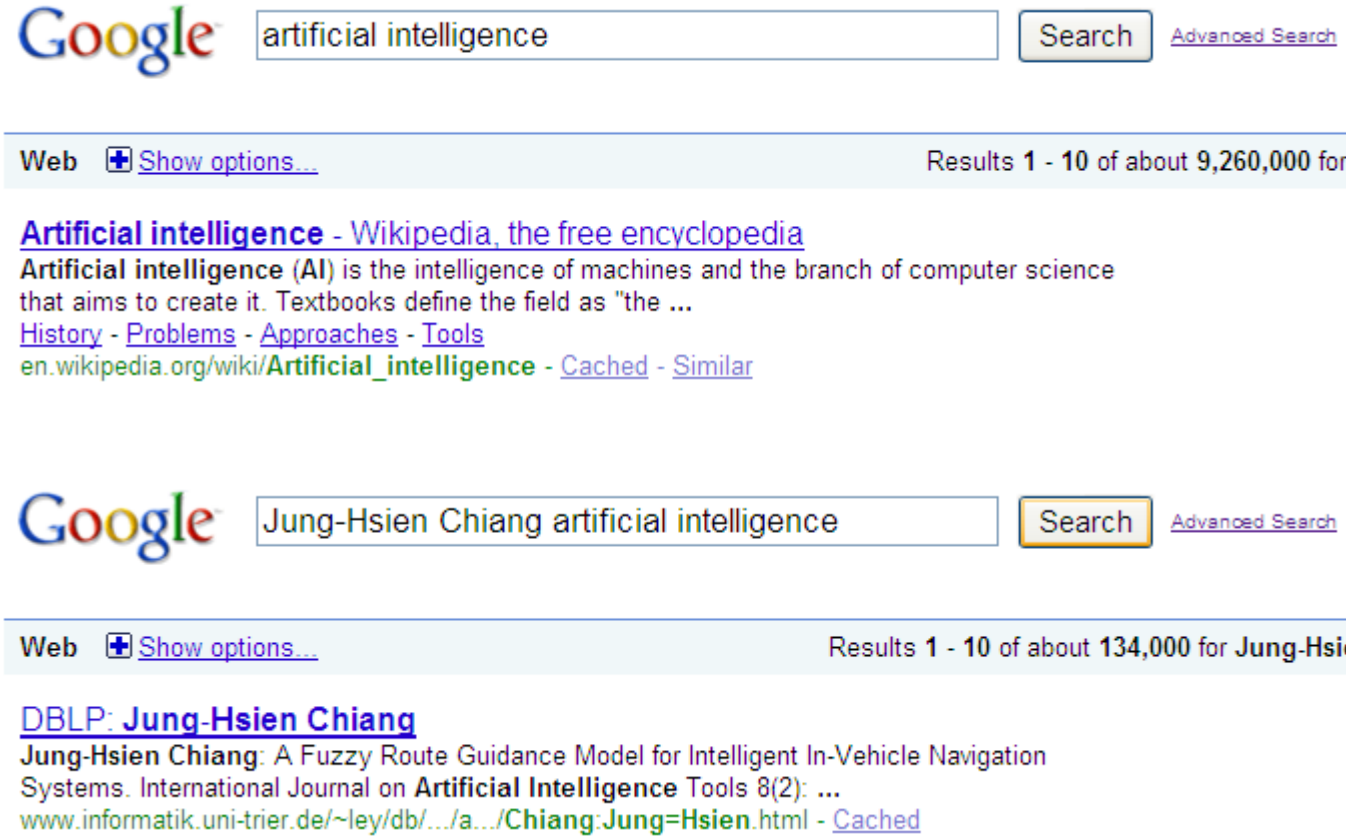
Google 與搜尋引擎的應用

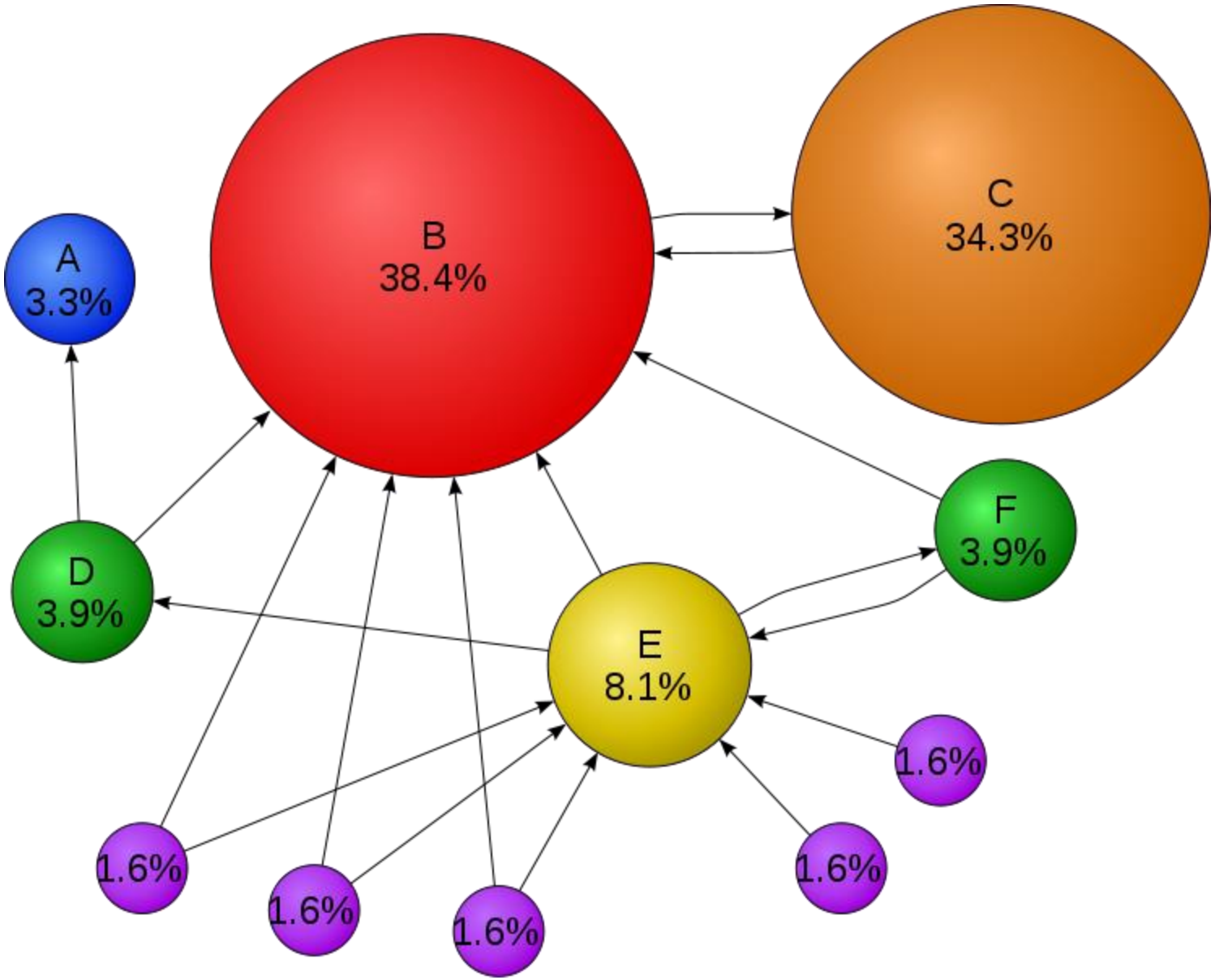

Google - Windows Internet Explorer
http://www.google.com/
Live Search
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
Windows Live
artificial intelligence
好友動向 個人檔案 郵件 相片 行事曆 MSN 分享
登入
Google
網頁(P) 工具(O)
Web Images Videos Maps News Shopping Gmail more ▼
Search settings | Sign in
Google
Advanced Search
Language Tools
Google Search
I'm Feeling Lucky
Advertising Programs - Business Solutions - About Google - Go to Google Taiwan
©2010 - Privacy
網際網路
100%
開始
IR-course
Search Engine
lecture7-evaluation
Google - Windows ...
文件1 - Microsoft ...
上午 01:27


Google
1625 CHARLESTON ROAD B44


Google
Google cloud print
http://goo.gl/sZHF7

Search Examples

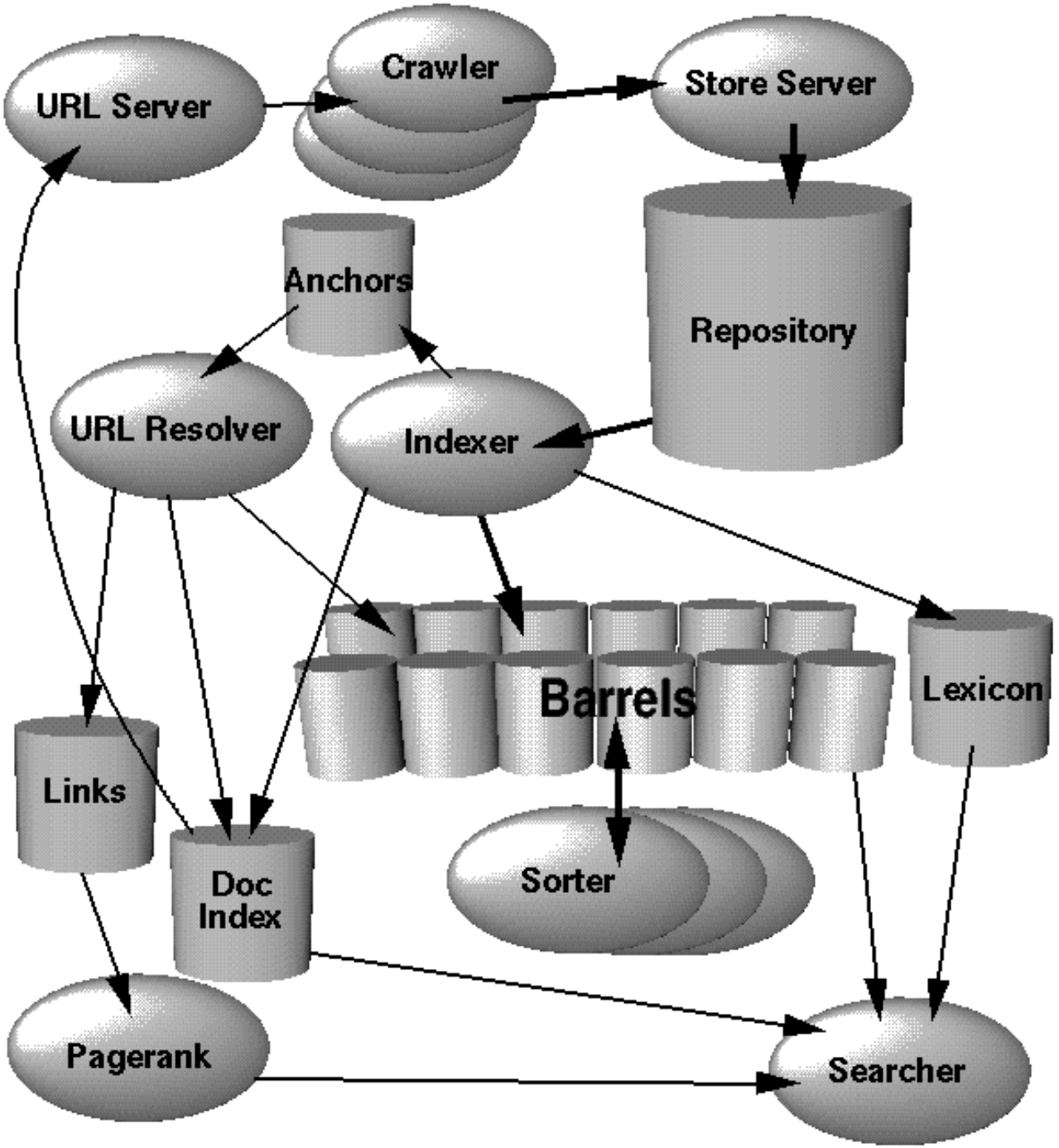



A
3.3%
B
38.4%
C
34.3%
D
3.9%
E
8.1%
F
3.9%
1.6%
1.6%
1.6%
1.6%
1.6%

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

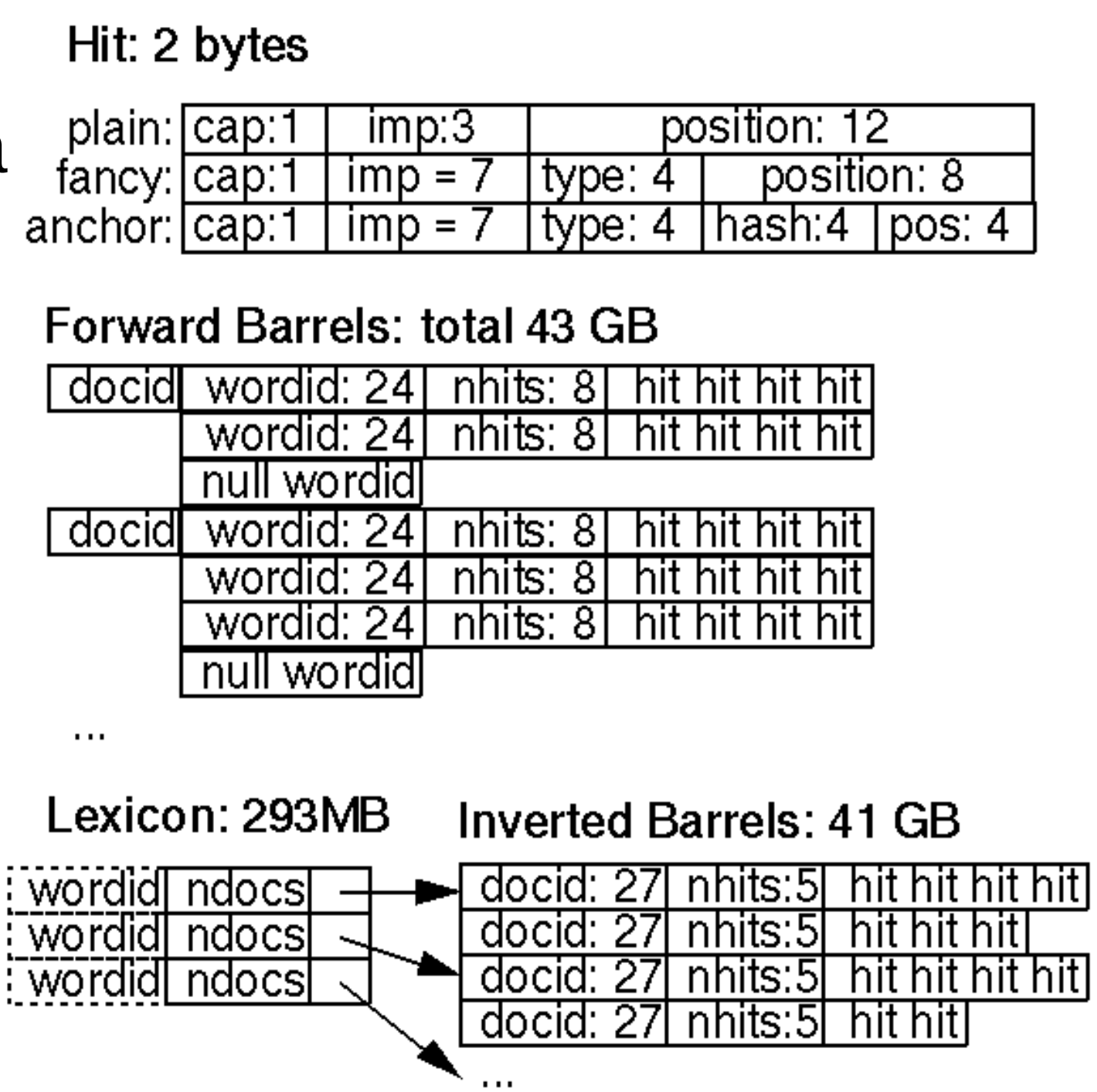


Major Data Structures
Hit Lists

- 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.

Major Data Structures:

Forward and Reverse Indexes and the Lexicon

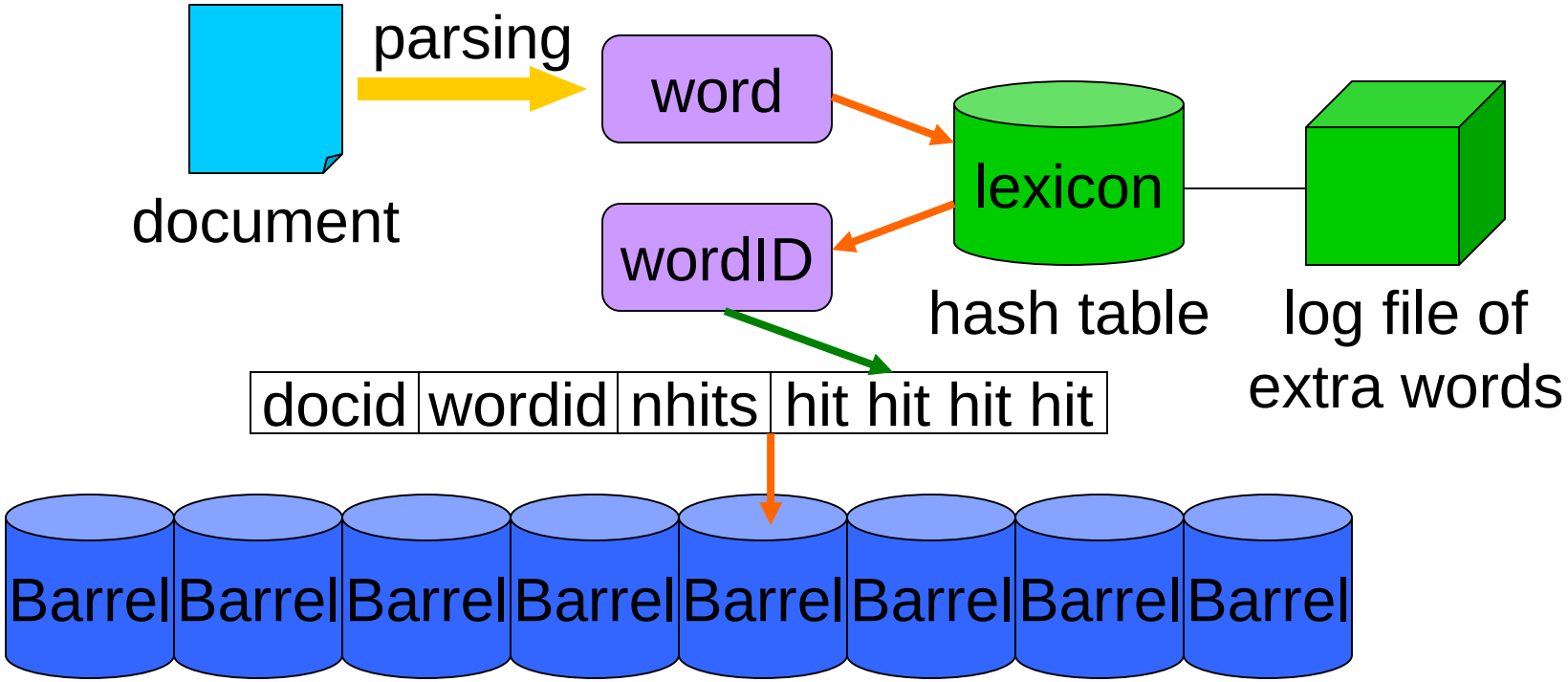


Crawling the Web

- 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

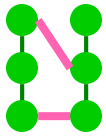
Indexing the Web
Forward Index

- Indexing documents into barrels



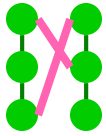
Indexing the Web
Inverted Index

- 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.

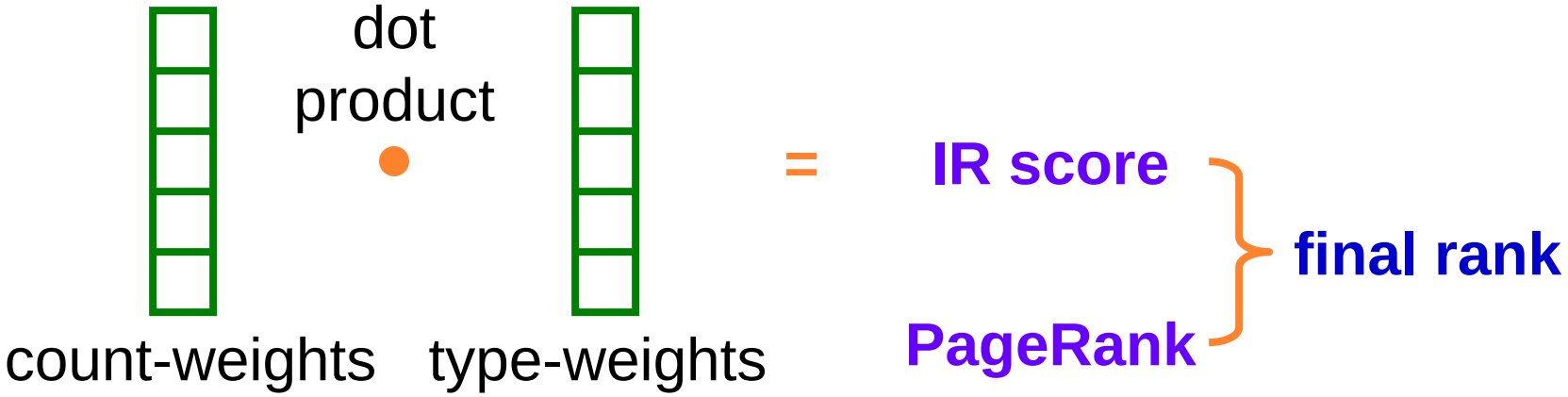
Searching

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

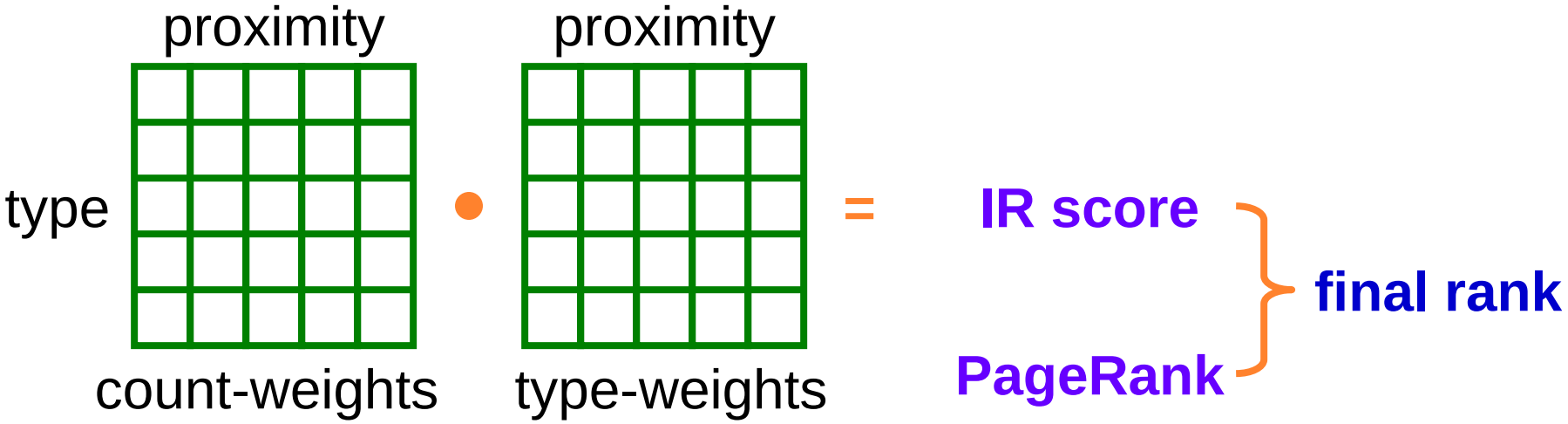
   Consider each hit to be one of several different types (title, anchor, URL, plain text large font, plain text small font, …).



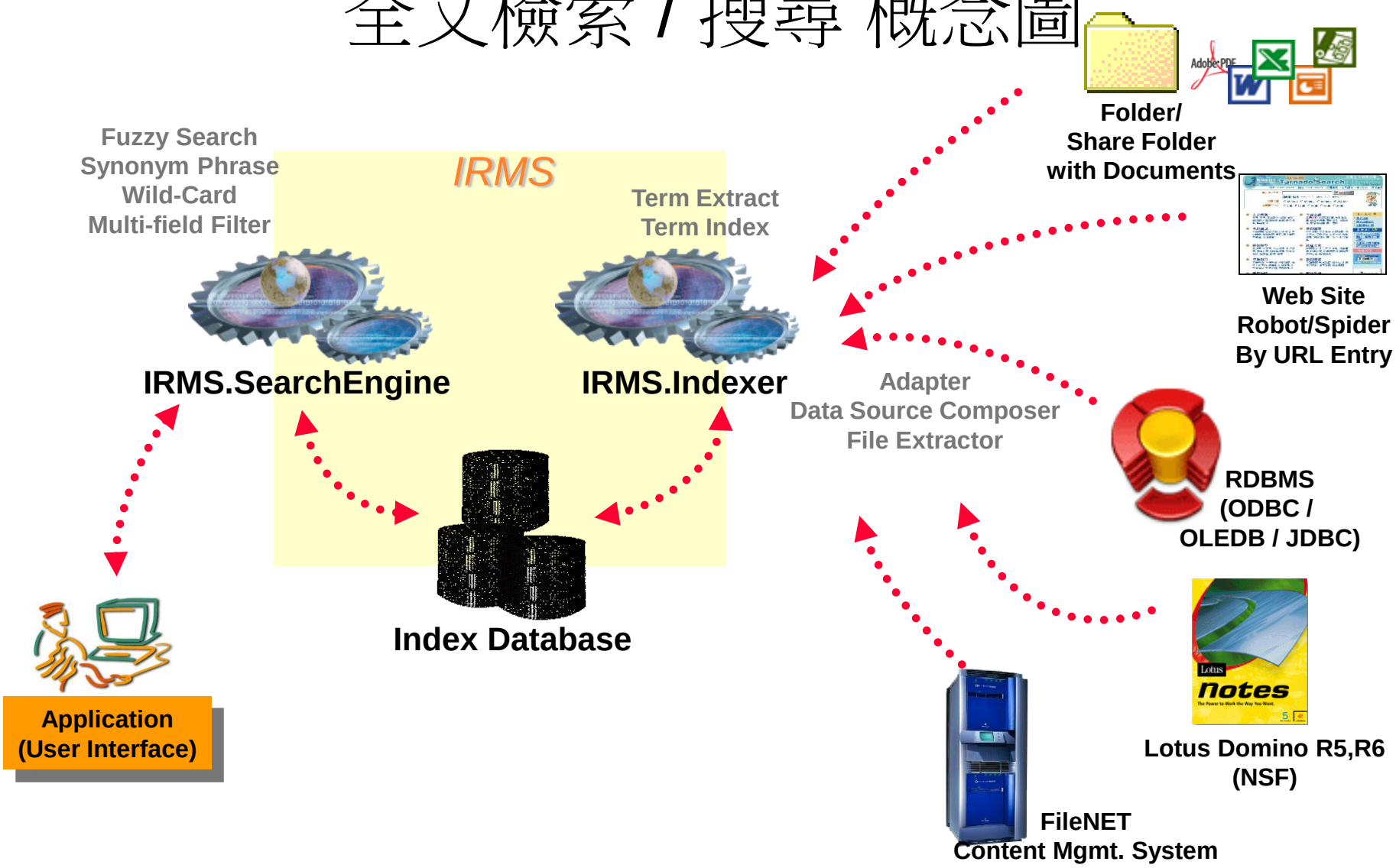
The Ranking System

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


Fuzzy Search
Synonym Phrase
Wild-Card
Multi-field Filter
IRMS
Term Extract
Term Index
IRMS.SearchEngine
IRMS.Indexer
Index Database
Application
(User Interface)
Folder/
Share Folder
with Documents
Web Site
Robot/Spider
By URL Entry
Adapter
Data Source Composer
File Extractor
RDBMS
(ODBC /
OLEDB / JDBC)
Lotus Domino R5,R6
(NSF)
FileNET
Content Mgmt. System

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。

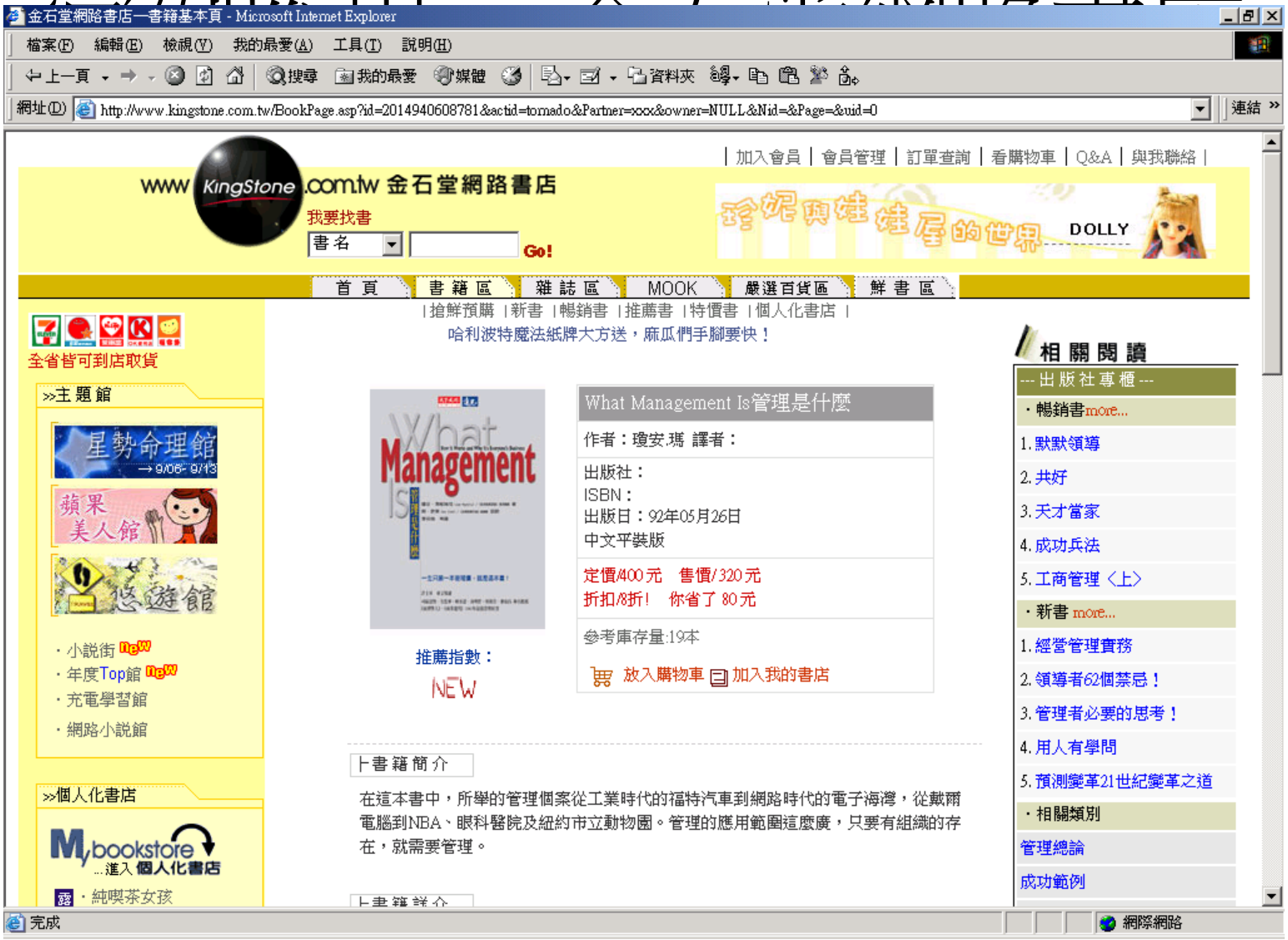
多國語言查詢與搜尋結果呈現

Korean

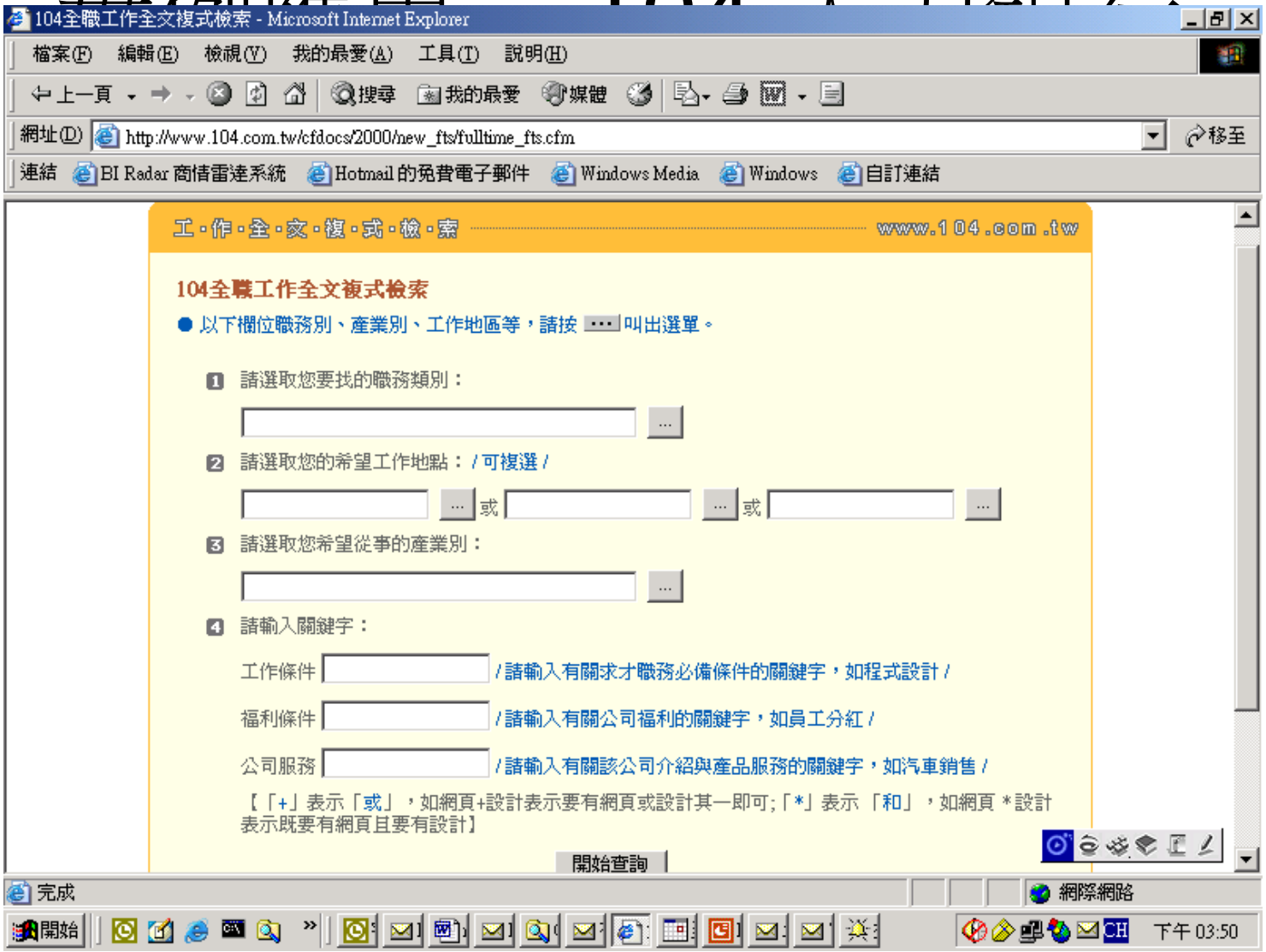
Japanese

Simplified Chinese

Traditional Chinese


金石堂網路書店—書籍基本頁 - 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 資料夾
網址(D) http://www.kingstone.com.tw/BookPage.asp?id=20149406087 81&actid=tornado&Partner=xxx&owner=NULL&Nid=&Page=&uid=0
連結
加入會員 | 會員管理 | 訂單查詢 | 看購物車 | Q&A | 與我聯絡 |
www KingStone .com.tw 金石堂網路書店
我要找書
書名
Go!
珍妮與娃娃屋的世界 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
星勢命理館 9/06~9/13
蘋果美人館
悠遊館
・小說街 NEW
・年度Top館 NEW
・充電學習館
・網路小說館
>>個人化書店
Mybookstore ...進入 個人化書店
・純喫茶女孩
What Management Is管理是什麼
作者：瓊安.瑪 譯者：
出版社：
ISBN：
出版日：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
放入購物車 加入我的書店
推薦指數：
NEW
書籍簡介
在這本書中，所舉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。
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
完成
網際網路

案例應用 — 104人力銀行



實例應用－高雄捷運局


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
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
搜尋
搜尋類型
分類查詢 進階查詢
全文檢索字串
工程
每頁顯示筆數
5 10 15 20 25 筆
文件狀態
全部 現行文件 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
升冪排列 降冪排列
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