

資訊檢索

邱志義

中央研究院資訊科學研究所

前言

- 資訊檢索是擷取、組織和利用資訊的重要技術
- 隨著數位典藏資料急遽成長，善用資訊檢索技術，資料才會成為有用的資訊

*Data is of no use
unless you can actually access it.*



Outline

- 資訊檢索簡介
- 檢索模型 - 布林模型與向量模型
- 效能評估

Outline

- 資訊檢索簡介
- 檢索模型 - 布林模型與向量模型
- 效能評估

Definition of Information Retrieval

- Information retrieval (IR) is **finding material** (usually documents) of an **unstructured** nature (usually text) that satisfies an **information need** from within **large collections** (usually stored on computers)

Definition of Information Retrieval

- Dealing with the *representation, storage, organization of, and access to* **information items**
 - Information items include text documents (often unstructured), Web pages (semi-structured), images, audios, videos, ...
- Converting information need to information items

Structured Data

- Structured data tends to refer to information in **tables**

Employee	Manager	Salary
Smith	Jones	50000
Chang	Smith	60000
Ivy	Smith	50000

- Typically allows numerical range and exact match (for text) of **Boolean queries**
 - *Salary < 60000 AND Manager = Smith*

Unstructure Data

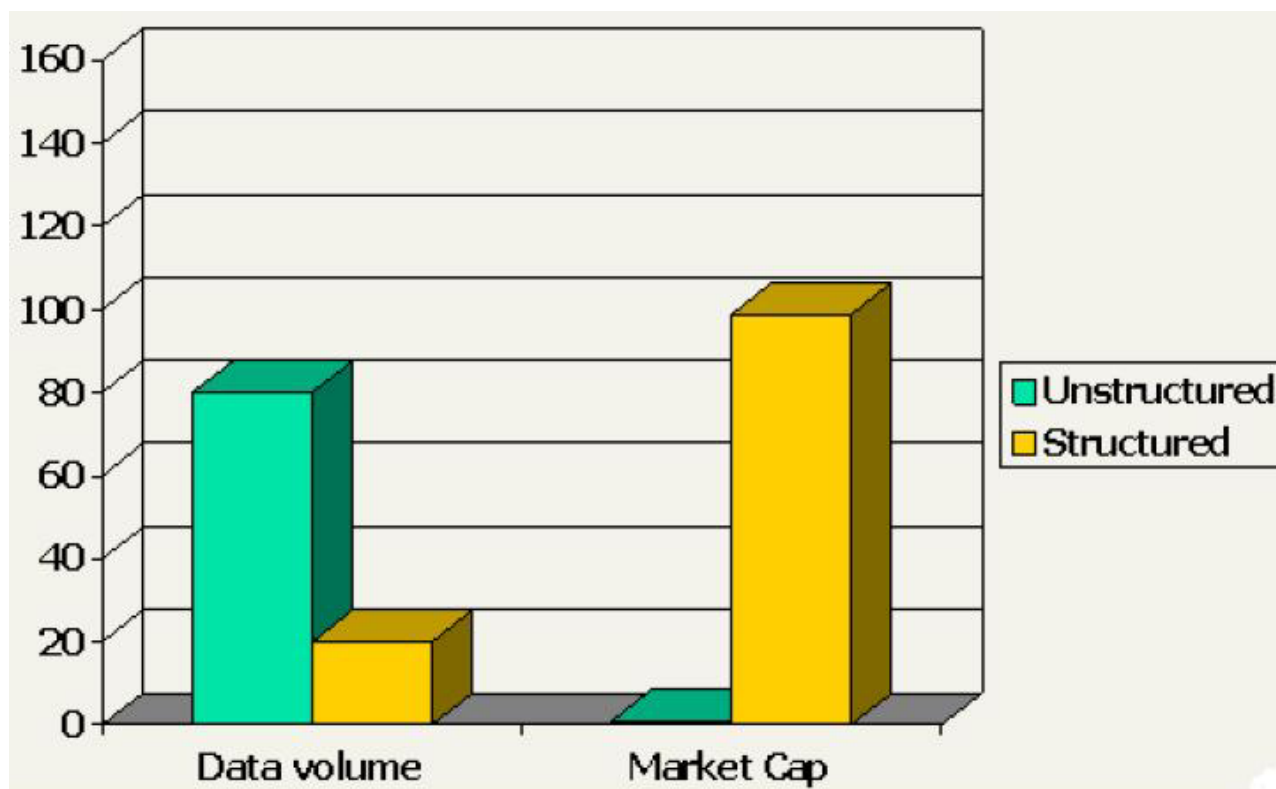
- Typically refers to free text
- Query formation
 - **Keyword queries** including operators
 - More sophisticated "**concept**" queries
 - find all web pages dealing with *drug abuse*

Database Retrieval vs. Information Retrieval

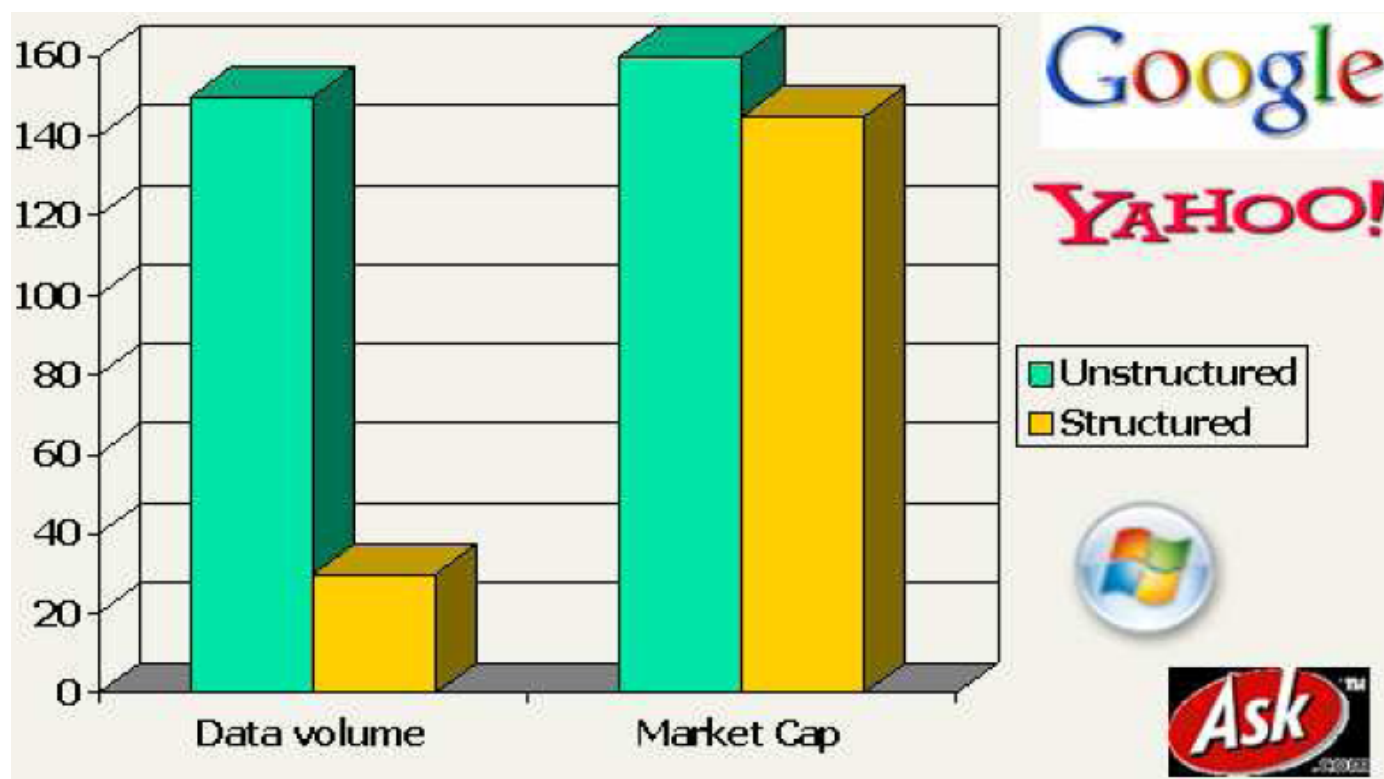
- Database System (DR Application)
 - Structured data
 - Exact search (no ranking)
 - Commercial success (Oracle, IBM DB2)

- Search Engine (IR Application)
 - Unstructured data
 - Approximate search (ranking)
 - Becoming dominant (Google, Yahoo!)

Unstructured vs. Structured data in 1996



Unstructured vs. Structured data in 2006



Need for IR

- With the advance of WWW - more than 100 billion web pages in March 2008 [wikipedia]
- Various needs for information:
 - Search for a specific information
 - Search an answer to a question
 - Search for information in a different language

Examples of IR Systems

- Database (library catalog)
 - Search by keywords, titles, authors, etc
- Text (Google, Yahoo)
 - Search by keywords. Some using queries in natural language
- Multimedia (QBIC, WebSeek, TinEye)
 - Search by audiovisual appearance (colors, texture, shapes...)
- Question answering systems (AskJeeves, Answerbus)
 - Search in (restricted) natural language
- Others:
 - Cross lingual information retrieval



Web · Images · News · Blogs & Feeds · Shopping · More »

who is the president in Taiwan

Search

[Advanced Search](#)

Web Search

Showing results 1-10 of 1



The Chief of State of [Taiwan](#) is President Chen Shui-bian, and the Head of State is Premier (President Of The Executive Yuan) Frank Hsieh

[World Factbook](#) | [Encyclopedia](#) | [BBC Profile](#) | [US Government Travel Info](#) | [Maps](#)

[Government Information Office, Republic of China \(Taiwan\)](#)

... comprehensive informatin about the ROC government on **Taiwan** includes announcements,official documents and AV presentations.

www.gio.gov.tw/ · [Save](#)

[Taiwan Security Research](#)

Taiwan's President, Mired in Scandals, Survives Recall Vote (Washington Post, June 28, 2006) **President** Chen Shui-bian easily survived a recall ...

www.taiwansecurity.org/ · [Cached](#) · [Save](#)

[R.O.C. President Lee Teng-hui -- A Profile](#)

... economy of **Taiwan** in its liberalization and internationalization efforts. Through the pragmatic diplomacy successfully promoted by ...

www.taipei.org/current/president/lee_p.htm · [Cached](#) · [Save](#)

[Taiwan \(Republic of China\)](#)

This site provides information about Republic of China on **Taiwan**, Visa, events, news, culture, Consulate Generals,

[Narrow Your Search](#)

[Government of Taiwan](#)

[Current President of Tai](#)

[Who Is the Prime Ministe
Taiwan](#)

[Who Is the Leader of Tai](#)

[Government Type of Tai](#)

[Taiwan President Late](#)

[Flag of Taiwan](#)

[Facts about Taiwan](#)

[Taiwan Info](#)

M

[Expand Your Search](#)

[President of Japan](#)

[President of China](#)

AnswerBus

where is Taiwan

Ask

Type in your question in English, French, Spanish, German, Italian or Portuguese.

Question:

where is Taiwan

Possible answers: [XML](#) [TXT](#)

- [Shi Xien is in northern and southern Taiwan, Hi-Lu is central and north central.](#)
- [Located across the Taiwan Strait from mainland China \(80 miles at the closest point\), Taiwan is a leading economic and trading center, with one of the largest ports in the world \(Kaohsiung\).](#)
- [About ATT | In the News | American Cultural Center | American Citizen Services | Visa Services | Services in Southern Taiwan | Agriculture Services | Commercial Services](#)
- [However, due to the Chinese Civil War between the Kuomintang \(KMT\) and the Chinese Communists, the 1951 San Francisco Peace Treaty between the Allies failed to name the recipient of Taiwan's sovereignty.](#)
- [Taiwan is a sub-tropical island, roughly 180 miles long, located less than 100 miles offshore of China's Fujian province.](#)



網頁 知識+ 圖片 影片 部落格 商家 新聞
商品 更多

礁溪的名產

搜尋

知識搜尋

礁溪的名產 搜尋結果約797個，以下為1 - 10個

搜尋特定分類：[餐飲情報](#) (32), [交通](#) (8), [旅遊](#) (722), [戶外活動](#) (4), [地方采風](#) (13)

搜尋結果排序方式：[相關性](#) | [日期](#)

1. [礁溪好康護照詳細內容](#)

這些是礁溪好康護照可使用的商家~~~ 宜蘭礁溪民宿·雅閣溫泉民宿~~~~~ 0918-525塹路18-4號 宜蘭礁溪名產·礁溪繁之鄉~~~~宜蘭縣礁溪鄉

分類：[台灣](#) 時間：2007/12/01

2. [請問一下宜蘭礁溪有蛇麼好玩的呢^](#)

那都方便。== 第一天你就先到【礁溪的（五峰旅瀑布）】晃一下，也不用時候也差不多要回家了。如果要買名產的話【宜蘭餅】【老元香】【老增壽

分類：[台灣](#) 時間：2005/12/27

3. [請問礁溪附近有什麼民產可以買!?](#)

有溫泉空心菜呀!跟一般外面賣的不一樣很大又很脆 如果不限定是礁溪的名產的話，在礁溪火車站附近有一家宜蘭餅 也很不錯吃，可以買來

分類：[其他](#) 時間：2005/10/05

4. [宜蘭縣有哪些好吃的東西、店面and觀光的名勝??](#)

一堆) 蘭陽技術學院後山看夜景 三星名產上將梨+蔥蒜 礁溪名產溫泉蕃茄~牛舌餅+蜜餞+鴨賞 礁溪火車站附近有一間賣包子

分類：[北部](#) 時間：2005/02/02

5. [宜蘭礁溪附近 有什麼好吃的?](#)

客棧。 礁溪除了以溫泉泡湯聞名，另有許多著名的風景名勝區；如果您想逛逛礁溪觀光商店街買名產或是嚐嚐當地美食小吃，以龍泉為中心，您將不必

分類：[飯店餐廳](#) 時間：2005/08/09

贊助網站

[一通電話禮盒到家 - 福堂餅行](#)

百年風味飄香，遠近馳名的地方名產，口感濃郁，料多味美，送禮自食兩相宜！

www.togotw.com/ft1899

[美食達人推薦的好吃xo干貝醬](#)

說到xo醬不能不提這家，真材實料，偌大的干貝，用料毫不手軟醬中之王。推薦給您

xo168.wemew.net

[龍情食品線上購物商城](#)

台灣名產、出國送禮、年節禮盒、訂婚喜餅、彌月、巧克力。

www.lung-ching.com.tw

[和美商行 - 名產](#)

古早味香腸，肉鬆，旗魚鬆純手工精選原料，傳承先民古早風味，佳節送禮自用兩相宜

www.homey888.com.tw

[礁溪 - 喆園豪雅民宿](#)

Cross-lingual Information Retrieval

LiveTrans

national palace museum

Source Language: Target Language: Fast Smart

Query/Translation	Relevant Pages	Relevant Images
national palace museum	<ul style="list-style-type: none"> * National Palace Museum [Gloss translation:] * TIT Museums: The National Palace Museum: 70 Years Young! [Gloss translation:] * Jades from the National Palace Museum [Gloss translation:] * National Palace Museum Exhibition [Gloss translation:] 	
國立故宮博物院	<ul style="list-style-type: none"> * 國立故宮博物院 [Gloss translation: national palace museum,] * 國立故宮博物院 預防性文物保存研習會 [Gloss translation: national palace museum to prevent cultural relic to conserve] * 國立故宮博物院院長 杜正勝 先生 [Gloss translation: national palace museum president sir] * 國立故宮博物院古文物及藝術品管理辦法 [Gloss translation: national palace museum cultural relic art to supervise means] 	









Automatic Translations: [國立故宮博物院](#); [故宮](#); [故宮博物院](#); [國立](#); [國立故宮博物館](#);
Dictionary Lookup:Unavailable!

Image Search

[Web](#) [Images](#) [Maps](#) [News](#) [Shopping](#) [Gmail](#) [more](#) ▼ chihyi.chiu@gmail.com | [My Account](#) | [Sign out](#)

Google [Advanced Image Search](#)
[Preferences](#)
Moderate SafeSearch is on New! [Google Image Labeler](#)

Images Showing: Results **1 - 20** of about **215,000,000** for **university** [[definition](#)]. (0.25 seconds)

 <p>Ivan Franko National University of ... 491 x 322 - 25k - jpg www.franko.lviv.ua</p>	 <p>The University of St Andrews, ... 1500 x 915 - 2376k - png www.python.org [More from www.python.org]</p>	 <p>History of the University of Sydney 587 x 366 - 51k - jpg www.usyd.edu.au</p>	 <p>Pictures of the University of Glasgow 1200 x 798 - 312k - jpg www.gla.ac.uk [More from www.gla.ac.uk]</p>
 <p>Pictures of the University of Glasgow 1200 x 1793 - 766k - jpg www.gla.ac.uk</p>	 <p>University Library at St Andrews 477 x 320 - 86k - jpg www.st-andrews.ac.uk</p>	 <p>Events at the University of Hamburg 640 x 426 - 59k - jpg www.uni-hamburg.de</p>	 <p>View of The University of St Andrews 450 x 274 - 29k - jpg www.python.org</p>

Content-Based Image Retrieval



國立歷史博物館/師大/新視 提供

Fancy Browsing Interface for Image Search

<http://www.piclens.com/site/ie/>

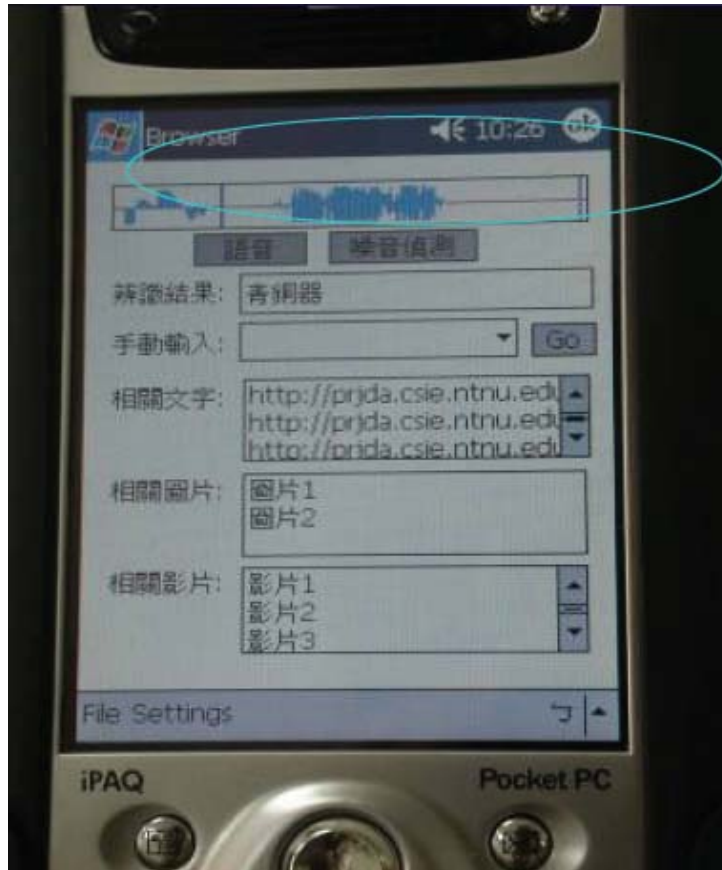


Find Near-duplicate Images

<http://tineye.com/>



Image Retrieval via Voice



國立歷史博物館/師大/新視 提供

Search Functions

- 結構化查詢
 - 特定結構文件: 網頁, 電子郵件, 新聞
 - allintitle: 李安 奧斯卡
- 關鍵詞查詢
 - 布林查詢: (A and B) or (C and D)
 - 同音(台灣, 臺灣), 同義(中華文化, 中國文化), 近似拼字
 - 術語推薦(台大: 台大醫院, 台大計算中心, 台大圖書館)
 - 相關回饋(relevance feedback)
 - 跨語言查詢: Sony, 新力, 索尼

Search Functions

- 範例式查詢 (Query by Example)
 - 相似網頁, 相似圖片
- 自然語言查詢 (Free Text Query)
 - 以較自然的方式來描述查詢主題, 如"世界盃足球舉辦地點"
(此處不強調語意理解)
- 問答 (Question Answering)
 - 人物、地點、組織、事物、時間、數字
 - 問題分析、文句擷取、答案抽取、答案排序 (強調語意理解)

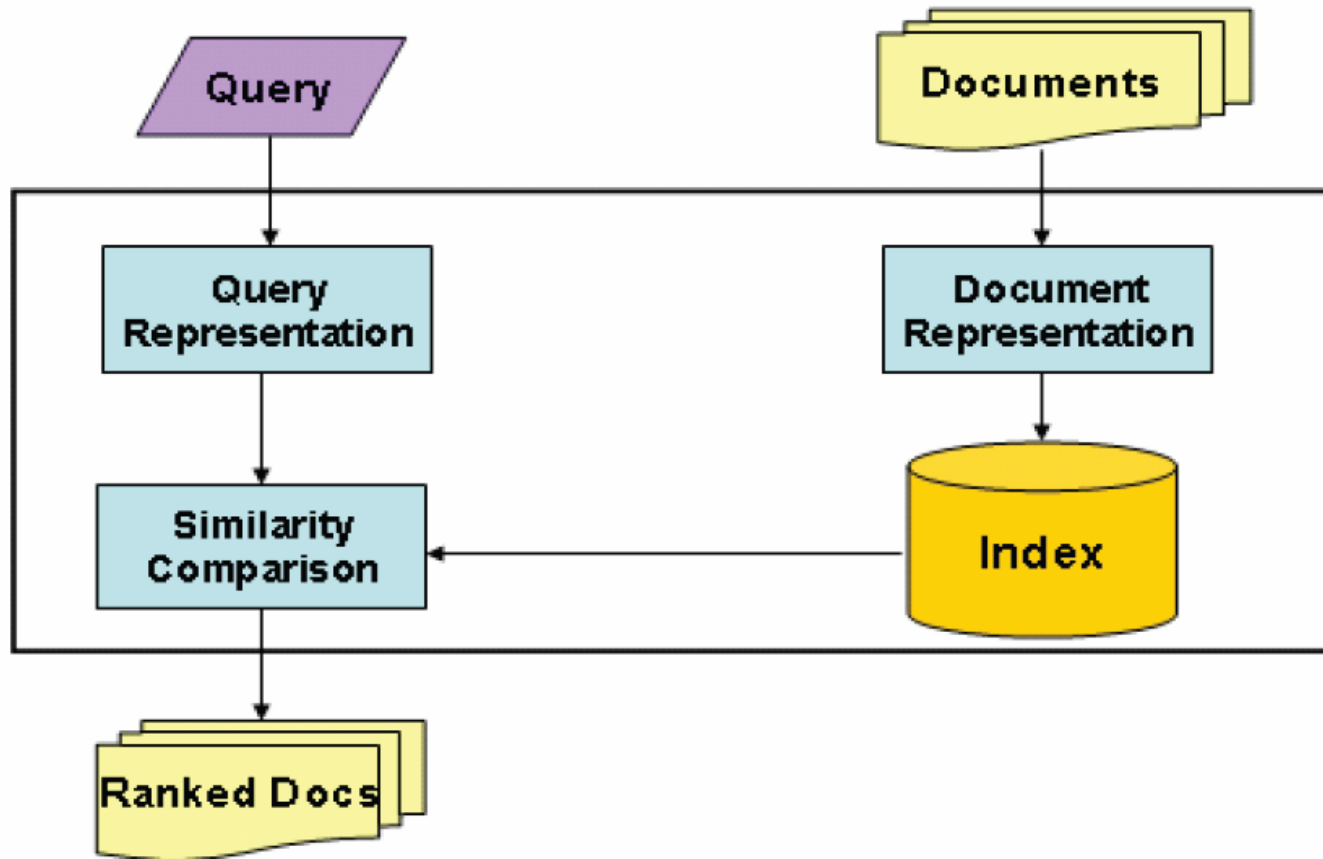
Result Presentation

- 資訊顯示處理
 - 相關性排序(Relevance Ranking)
 - 限制顯示筆數
 - 限制顯示資料的詳細程度(註解或摘要)
- 資訊摘要(Summarization)
- 資訊分類(Clustering)
- 提示處理(Highlight)
- 資訊傳遞(Delivery)
 - 資訊傳遞影響反應時間

Text IR System Architecture

The user task

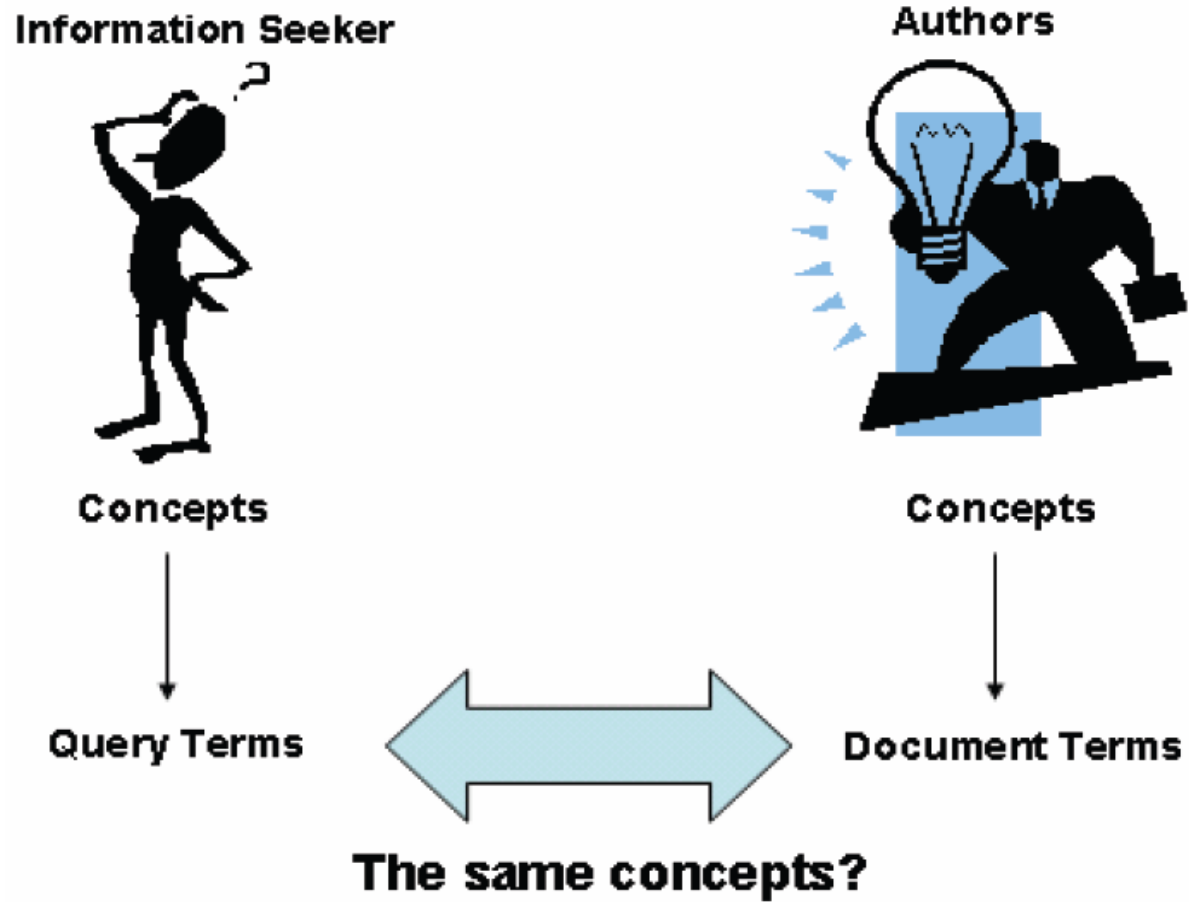
Logical view of document



Text IR System Architecture

- The user task
 - Convey the semantics of information need
 - Retrieving and browsing results
- Logical view of document
 - Full text representation and indexing
 - Build a set of index terms
 - Remove stopwords
 - Apply stemming
 - Identify phrases

The Problem



Information Needs and Queries

- **Information need:** the topic the user desires to know more
 - Dependent on prior knowledge, subjectivity, context
- **Query:** what the user conveys to the computer in an attempt to communicate the information need
 - Users don't express their information needs into queries well at first
- *User-training* plays a important role
 - In Google, general users use 1-3 word queries
 - In Westlaw, professional users use 10-12 words queries

Outline

- 資訊檢索簡介
- 檢索模型 - 布林模型與向量模型
- 效能評估

Boolean Model (布林模型)

- Weights assigned to terms are either “0” or “1”
 - '0' represents "absence": term **isn't** in the document
 - '1' represents "presence": term **is** in the document
- Expressing Querying with Boolean operators
 - AND, OR, NOT
- Returning all documents that satisfy the query

Boolean Operator

A \ B	0	1
0	0	1
1	1	1

A OR B

B	0	1
1	1	0

NOT B

A \ B	0	1
0	0	0
1	0	1

A AND B

A \ B	0	1
0	0	0
1	1	0

A NOT B
(= A AND NOT B)

Unstructured Data in 1650



Which plays of
Shakespeare contain the
words: **Brutus** and **Caesar**,
but not **Calpurnia**?

Term-document Incidence Matrix

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
ANTHONY	1	1	0	0	0	1	
BRUTUS	1	1	0	1	0	0	
CAESAR	1	1	0	1	1	1	
CALPURNIA	0	1	0	0	0	0	
CLEOPATRA	1	0	0	0	0	0	
MERCY	1	0	1	1	1	1	
WORSER	1	0	1	1	1	0	
...							

Incidence Vectors

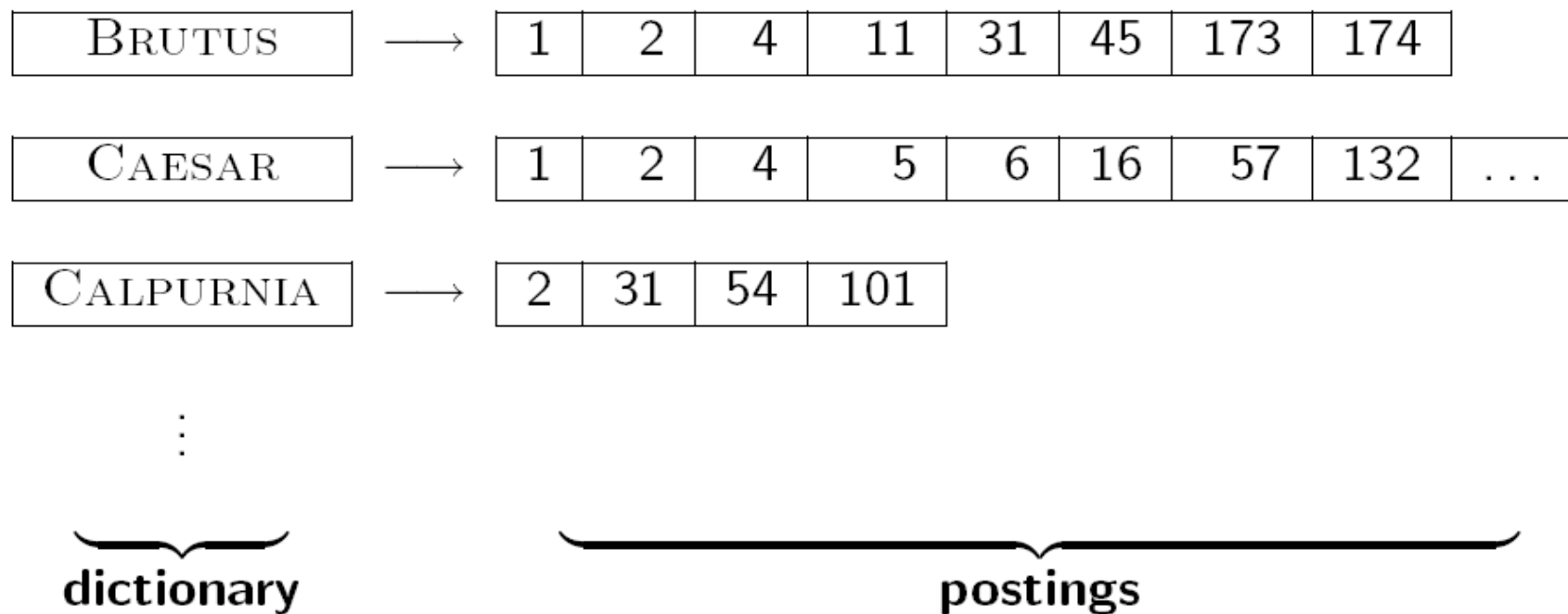
- To answer the query: **Brutus** and **Caesar** and not **Calpurnia**
 - Take the vectors for **Brutus**, **Caesar**, and **Calpurnia**
 - Complement the vector of **Calpurnia**
 - Do a (bitwise) and on the three vectors
 - 110100 and 110111 and 101111 = 100100

Can't Build the Incidence Matrix – too Large in a Big Collection

- Consider $N = 10^6$ documents, each with about 1000 terms
- On average 6 bytes per term, including spaces and punctuation \Rightarrow size of document collection is about 6 GB
- Assume there are $M = 500,000$ distinct terms in the collection
- $M = 500,000 \times 10^6 =$ half a trillion 0s and 1s

Inverted Index

- Actually the matrix is extremely sparse, we can only record the 1s for each term



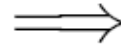
Inverted Index Construction

- Collect the documents to be indexed
 - Friends, Romans, countrymen So let it be with Caesar ...
- Tokenize the text
 - Friends Romans countrymen So ...
- Do linguistic preprocessing
 - friend roman countryman so ...
- Index the documents that each term occurs

Tokenization and Preprocessing

Doc 1. I did enact Julius Caesar: I was killed i' the Capitol; Brutus killed me.

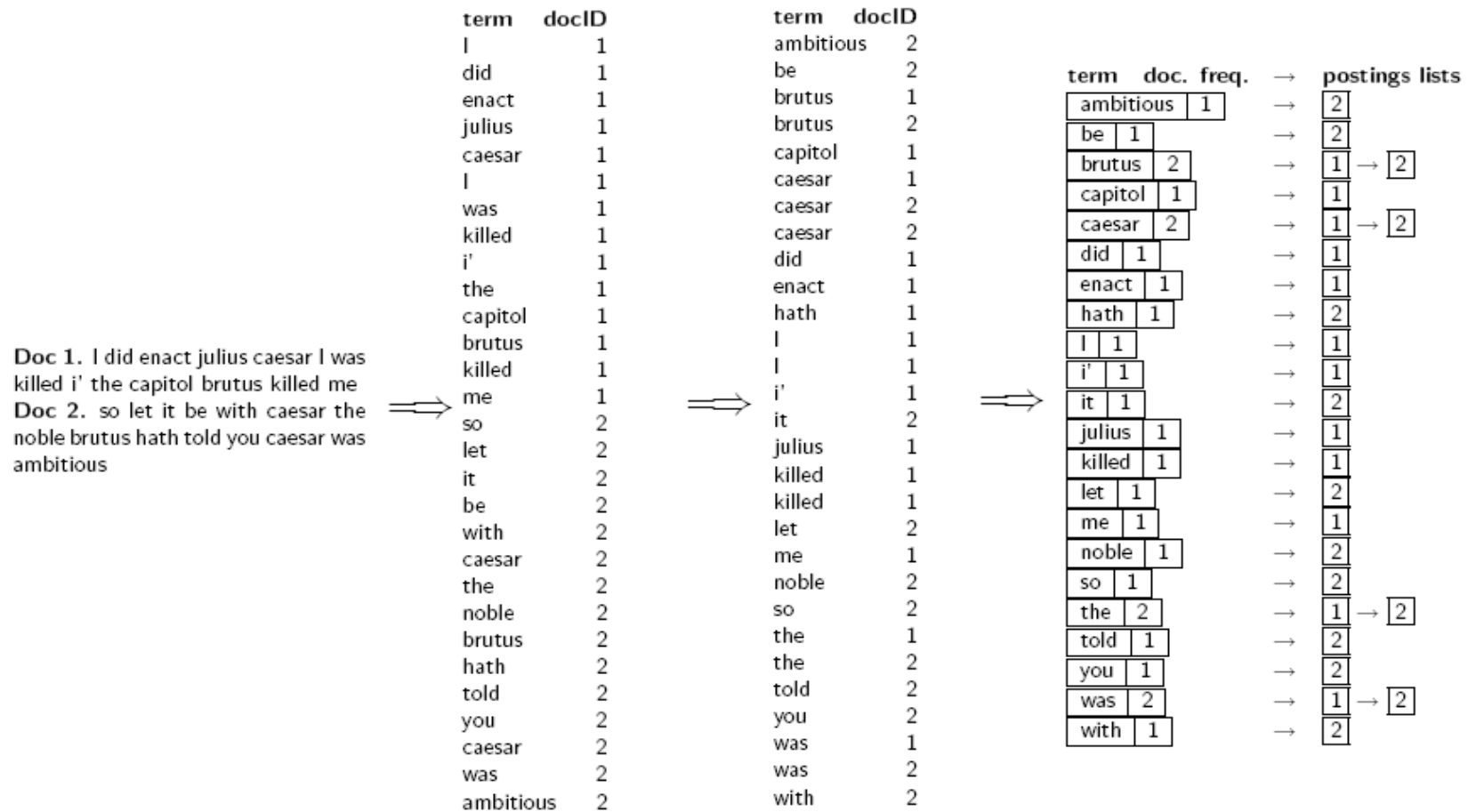
Doc 2. So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious:



Doc 1. I did enact julius caesar I was killed i' the capitol brutus killed me

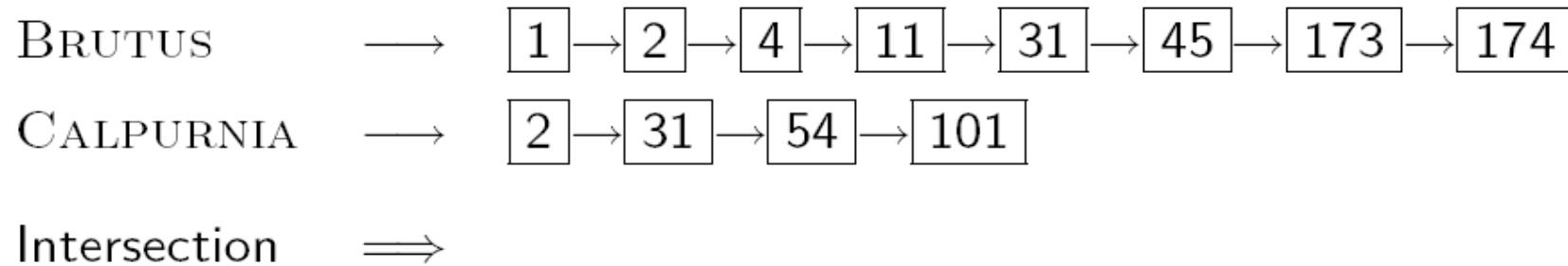
Doc 2. so let it be with caesar the noble brutus hath told you caesar was ambitious

Generate Posting



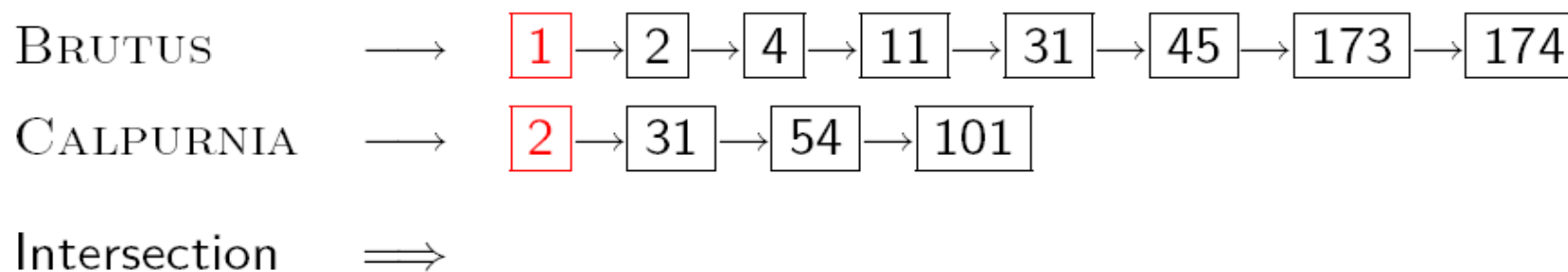
Processing Boolean Queries

- Consider the query: **Brutus AND Calpurnia**



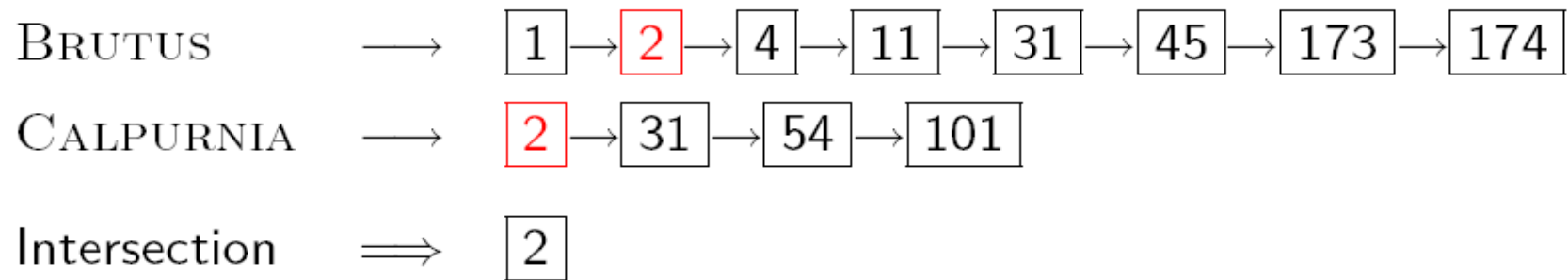
Processing Boolean Queries

- Consider the query: **Brutus AND Calpurnia**



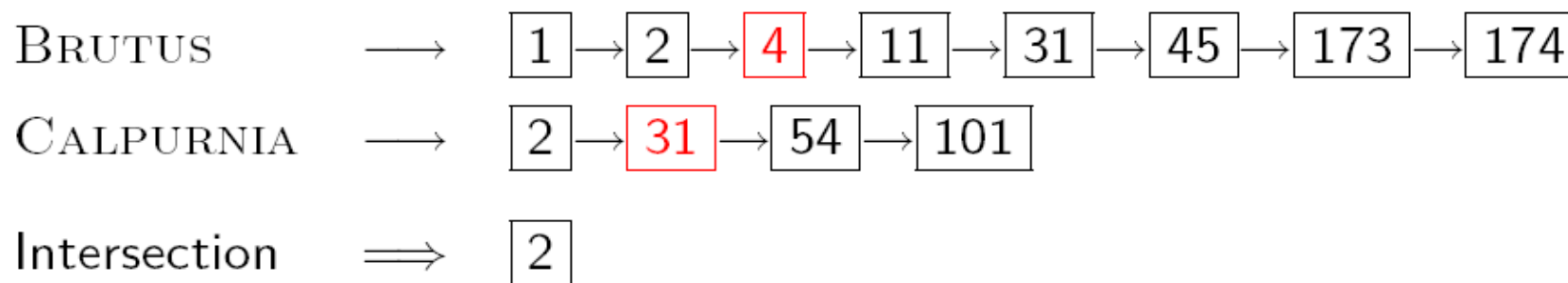
Processing Boolean Queries

- Consider the query: **Brutus AND Calpurnia**



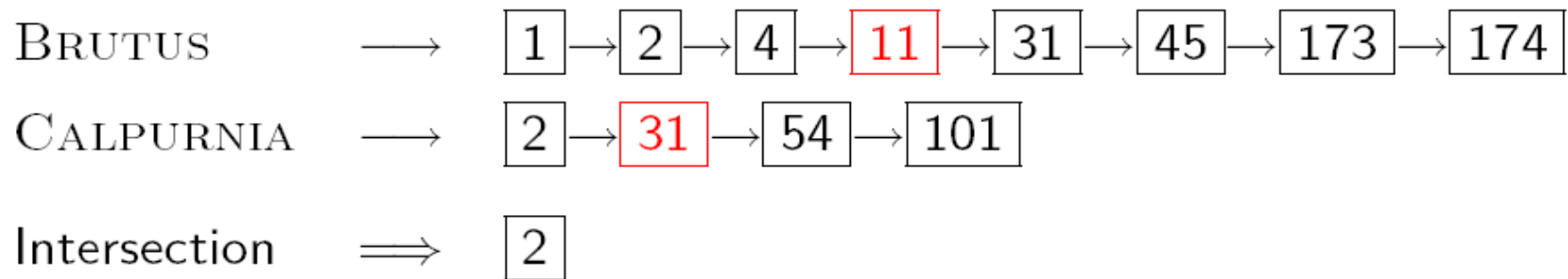
Processing Boolean Queries

- Consider the query: **Brutus AND Calpurnia**



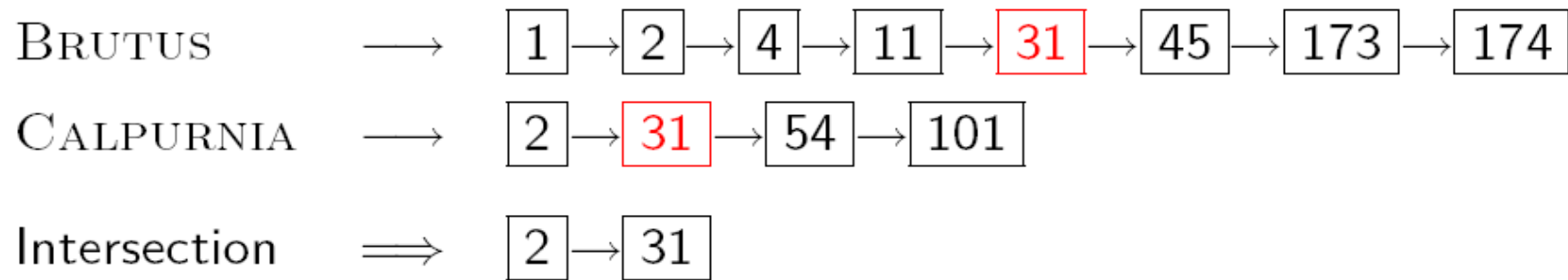
Processing Boolean Queries

- Consider the query: **Brutus AND Calpurnia**



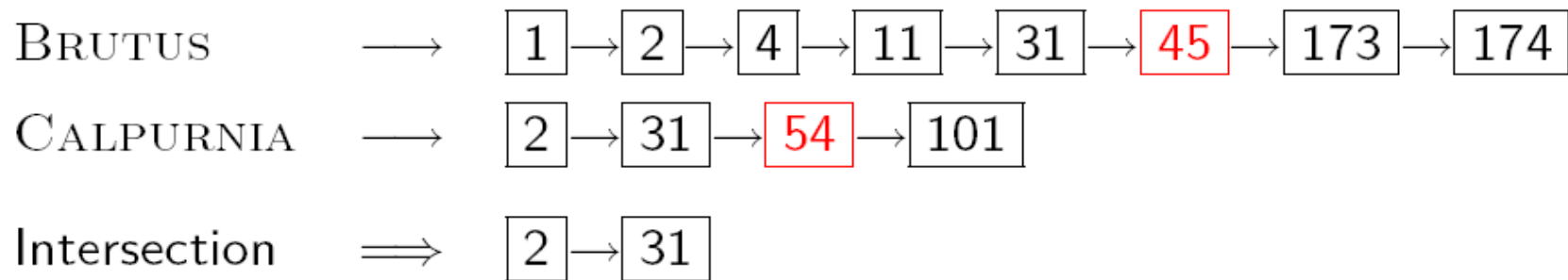
Processing Boolean Queries

- Consider the query: **Brutus AND Calpurnia**



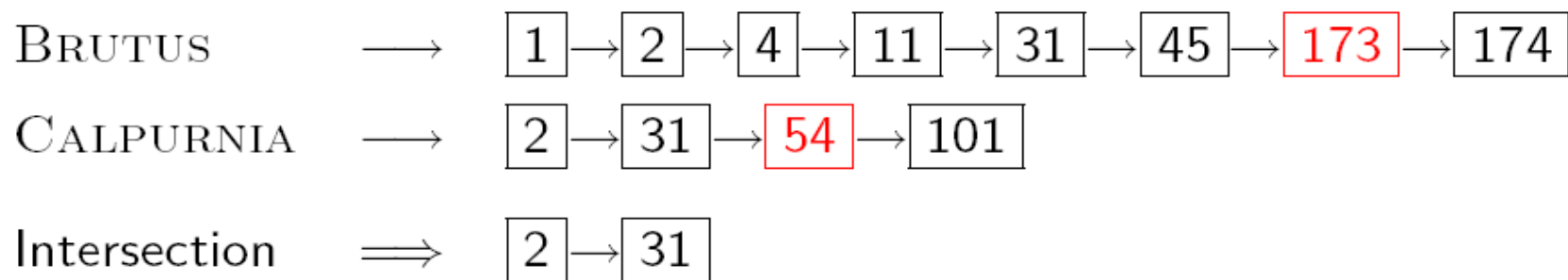
Processing Boolean Queries

- Consider the query: **Brutus AND Calpurnia**



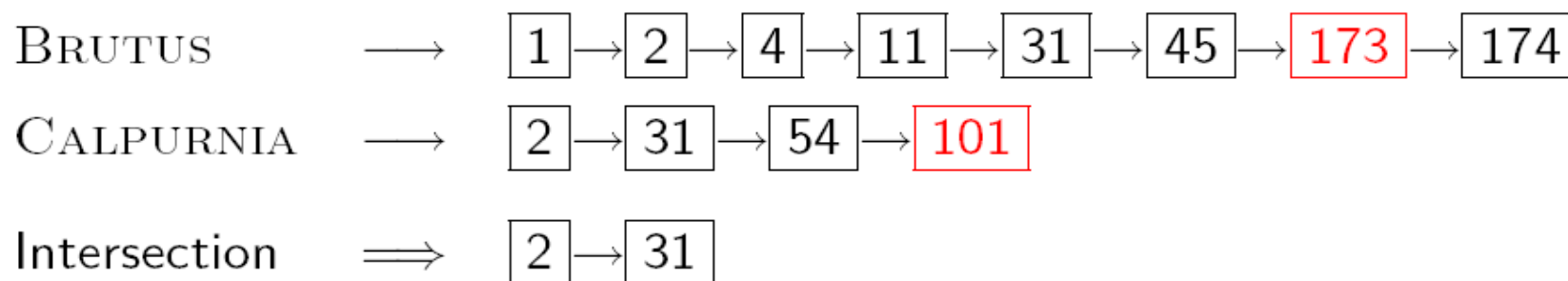
Processing Boolean Queries

- Consider the query: **Brutus AND Calpurnia**



Processing Boolean Queries

- Consider the query: **Brutus AND Calpurnia**



- This is linear in the length of the postings lists
- This only works if postings lists are sorted

Boolean Queries

- Primary commercial retrieval tool for 3 decades
- Many professional searchers (e.g., lawyers) still like Boolean queries – you know exactly what you are getting

Review of the Boolean Model

- Binary decision (relevant or non-relevant)
No partial match (compared to ranked retrieval)
- Good for expert users with precise understanding of their needs and the collection
- However, most users are not capable of writing Boolean queries (or they are, but they think it's too much work)

Review of the Boolean Model

- Retrieve too few or too many documents
 - Use AND operations tends to produce **high precision but low recall** searches
 - Use OR operators gives **low precision but high recall** searches
 - It is difficult or impossible to find a satisfactory middle ground

Review of the Boolean Model

- Query 1: "standard user dlink 650" → 200,000 hits
- Query 2: "standard user dlink 650 no card found" → 0 hits
- It takes a lot of skill to come up with a query that produces a manageable number of hits
- With a **ranked list** of documents it does not matter how large the retrieved set is

Ranked Retrieval

- We wish to return in order the documents most likely to be useful to the searcher
- Assign a score – say in $[0, 1]$ – to each document
- This score measures how well document and query "match"

Recall: Binary Incidence Matrix

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
ANTHONY	1	1	0	0	0	1	
BRUTUS	1	1	0	1	0	0	
CAESAR	1	1	0	1	1	1	
CALPURNIA	0	1	0	0	0	0	
CLEOPATRA	1	0	0	0	0	0	
MERCY	1	0	1	1	1	1	
WORSER	1	0	1	1	1	0	
...							

Each document is represented by a binary vector $\in \{0, 1\}^{|V|}$.

Term Frequency

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
ANTHONY	157	73	0	0	0	1	
BRUTUS	4	157	0	2	0	0	
CAESAR	232	227	0	2	1	0	
CALPURNIA	0	10	0	0	0	0	
CLEOPATRA	57	0	0	0	0	0	
MERCY	2	0	3	8	5	8	
WORSER	2	0	1	1	1	5	
...							

Each document is represented by a count vector $\in \mathbb{N}^{|V|}$.

Bag of Words Model

- We do not consider the order of words in a document
- "*John is quicker than Mary*" and "*Mary is quicker than John*" are represented the same way
- In a sense, this is a step back: The positional index was able to distinguish these two documents

Document Frequency

- The **document frequency** is the number of documents in the collection that the term occurs in
- Rare terms are more informative than frequent terms
 - 紅外線測量體感互動裝置 vs. Wii
 - 台灣雲豹棲息地 vs. 高山
- We use the **inverse document frequency** to factor this into computing the matching score

Collection Frequency vs. Document Frequency

- The collection Frequency of a term is the number of tokens of the term in the collection

Word	Collection frequency	Document frequency
INSURANCE	10440	3997
TRY	10422	8760

- Which word is a better search term (and should get a higher weight)?

tf-idf Weighting

- Product of its *tf* (term frequency) weight and its *idf* (inverse document frequency) weight

$$W = (1 + \log tf) \cdot \log \frac{N}{df} \quad \text{where } N \text{ is the number of documents}$$

- Best known weighting scheme in information retrieval
- Note: the “-” in tf-idf is a hyphen, not a minus sign!

Binary \rightarrow Count \rightarrow Weight Matrix

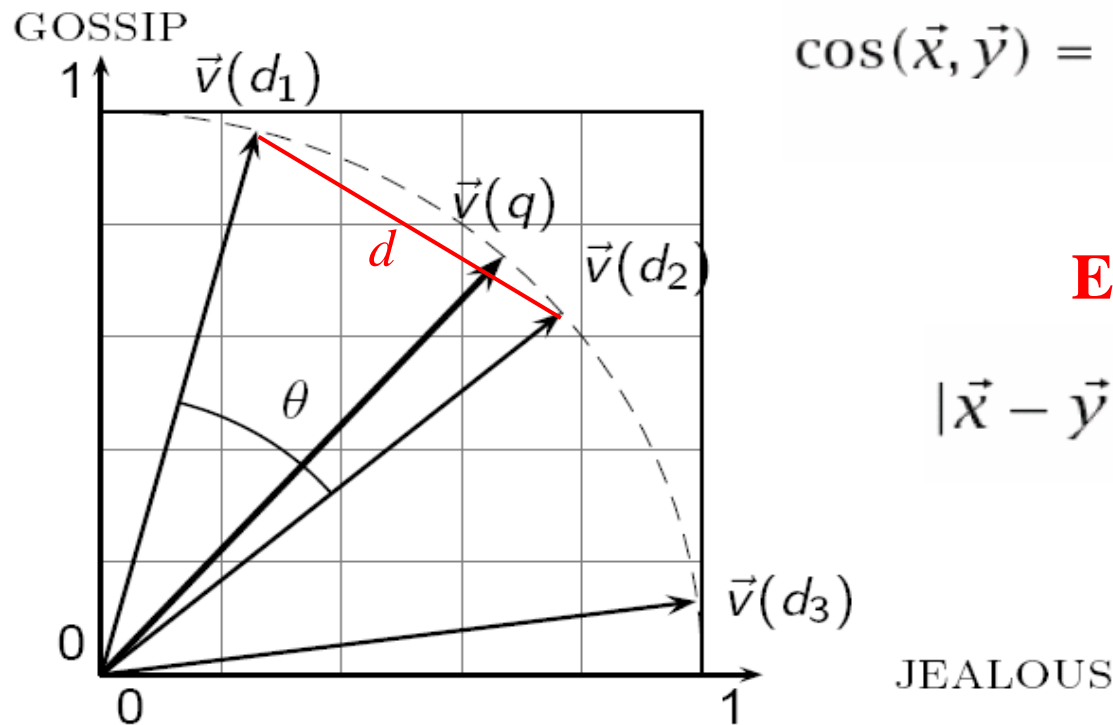
	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
ANTHONY	5.25	3.18	0.0	0.0	0.0	0.35	
BRUTUS	1.21	6.10	0.0	1.0	0.0	0.0	
CAESAR	8.59	2.54	0.0	1.51	0.25	0.0	
CALPURNIA	0.0	1.54	0.0	0.0	0.0	0.0	
CLEOPATRA	2.85	0.0	0.0	0.0	0.0	0.0	
MERCY	1.51	0.0	1.90	0.12	5.25	0.88	
WORSER	1.37	0.0	0.11	4.15	0.25	1.95	
...							

Each document is now represented by a **real-valued vector** of tf-idf weights

Vector Space Model

- Queries and documents represented as n -dimensional vectors in the space
 - Each dimension corresponds to a term
 - Documents are points or vectors in this space
- Rank documents according to their proximity to the query
 - proximity = similarity
 - proximity \approx negative distance

Vector Space Similarity



cosine similarity

$$\cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| |\vec{y}|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$$

Euclidean distance

$$|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Summary: Ranked Retrieval in the Vector Space

- Represent the query as a weighted tf-idf vector
- Represent each document as a weighted tf-idf vector
- Compute the cosine similarity or Euclidean distance between the query vector and each document vector
- Rank documents with respect to the query
- Return the top k (e.g., $k = 10$) documents to the user

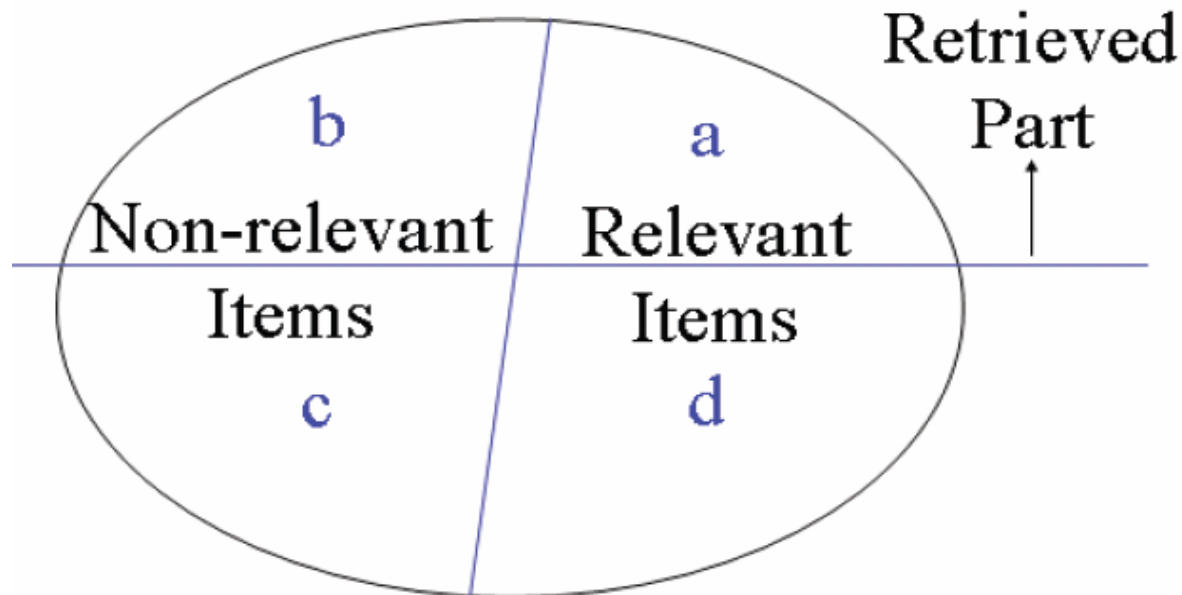
Outline

- 資訊檢索簡介
- 檢索模型 - 布林模型與向量模型
- 效能評估

Retrieval Effectiveness

- Recall (查全率)
 - 找到是正確的東西越多越好
- Precision (精確度)
 - 垃圾越少越好
- Ranking (排序)
 - 越需要的越前面越好
- Deadlink (死連結)
 - 死掉連結越少越好
- RefreshRate (更新率)
 - 更新率越高越好
- Response Time (反應速度)
 - 搜尋速度越快越好

Recall, Precision, and F-measure



$$Recall = \frac{a}{a + d}$$

$$Precision = \frac{a}{a + b}$$

$$F1 - measure = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

nDCG – Ranking Metric

- Normalized discounted cumulative gain

$$DCG_p = \sum_{i=1}^p \frac{2^{rel_i} - 1}{\log_2(1 + i)} \quad nDCG_p = \frac{DCG_p}{IDCG_p}$$

where i is the ranking position,
 rel_i is the relevance of i -th ranking position,
 $IDCG_p$ is the ideal DCG_p

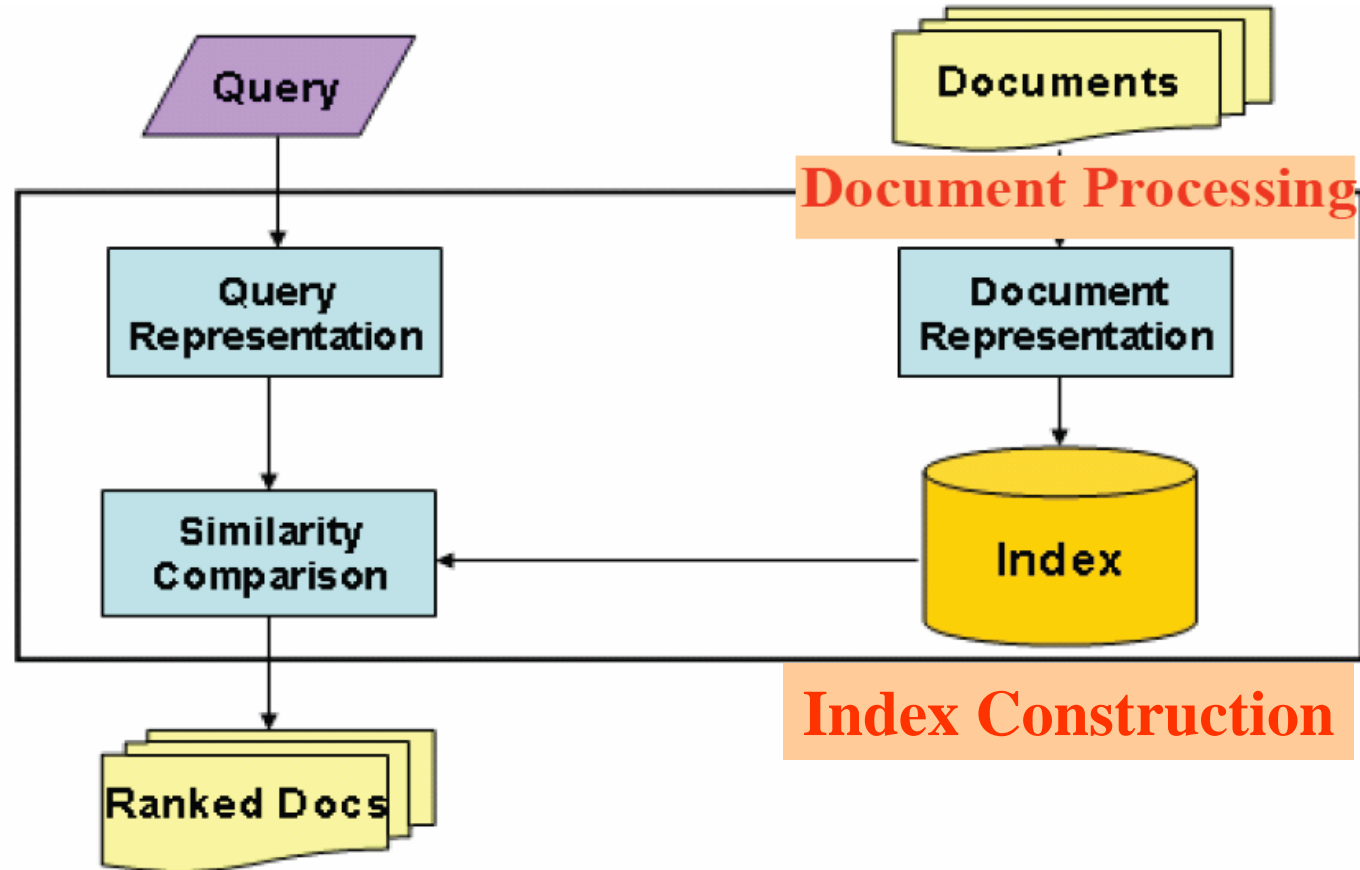
Advanced Topics

- 文件處理與索引建置
- 網路資訊檢索
- 網路廣告

Advanced Topics

- 文件處理與索引建置
- 網路資訊檢索
- 網路廣告

Document Processing for Indexing

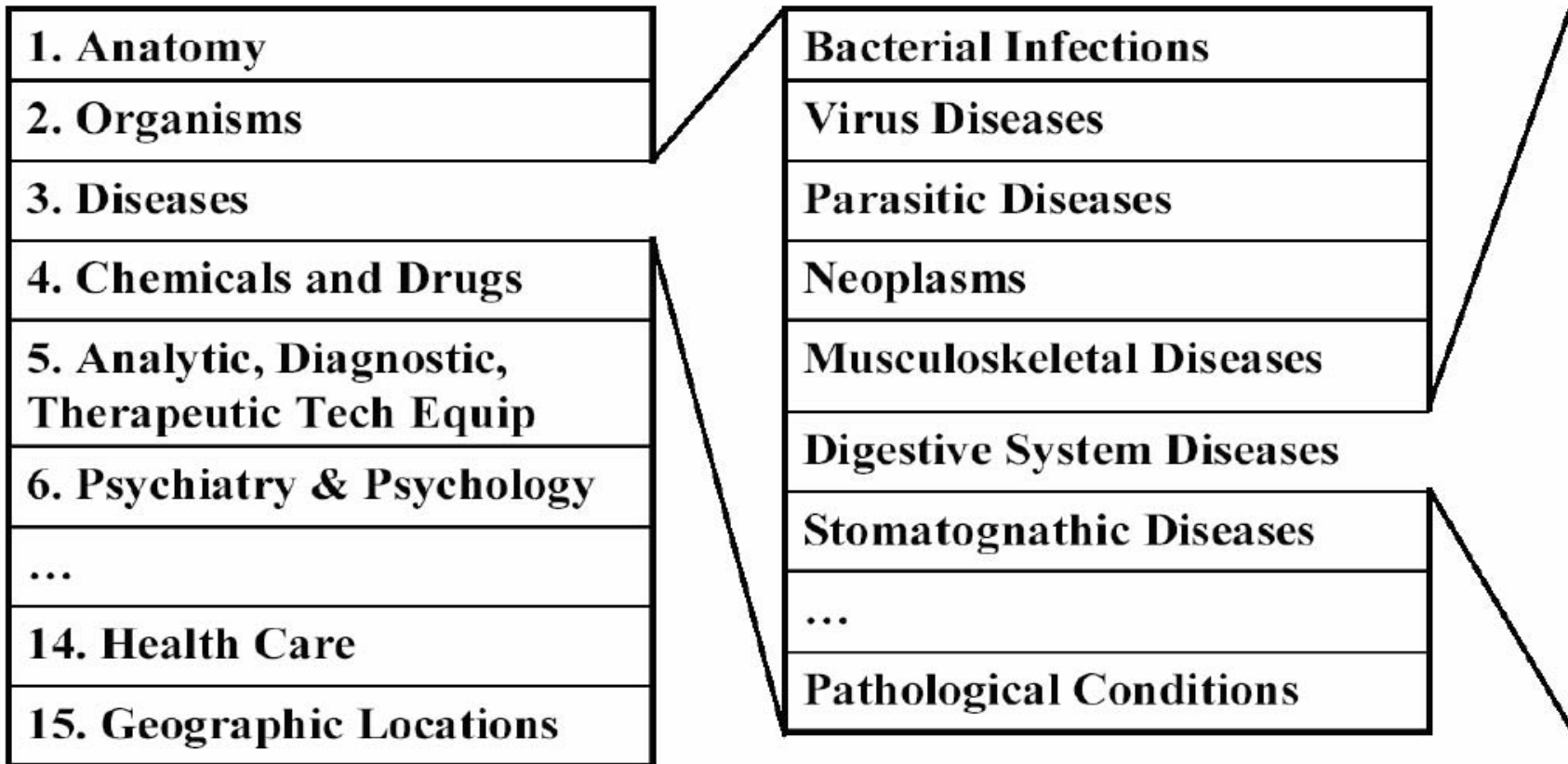


Manual Indexing

- Indexers decide which keywords to be indexed in a document based on **controlled vocabularies**
 - Examples: libraries, Medline, Yahoo
- Significant human costs, but no computational costs

Controlled Vocabularies

Example: Medical Subject Headings (MeSH)



Controlled Vocabulary Indexing

- There are many controlled vocabularies.
None is the best!
 - Library of Congress Subject Headings (LCSH)
 - Medical Subject Headings (MeSH)
 - LCSH is broad, while MeSH is detailed
- It solves the vocabulary mismatch problem. The domain ontology is explicit. Nice for browsing
- However, it is difficult and expensive to create, to use, and to maintain

Automatic Indexing

- Parse documents
- Scan for word tokens
- Stopword removal
- Word stemming
- Phrase recognition

Tokenization

- Design decisions
 - Numbers (510 B.C.)
 - Hyphenation (state-of-the-art, state of the art, B-52)
 - Capitalization
 - Punctuation (, ° 、 :)
 - Special characters
- Languages such as Chinese and Japanese need segmentation (部門聯絡, 飛越南太平洋, 土地公有政策)
- Record position information for proximity operators

Stopword Removal

- Stopwords are the words that can be discarded from a document representation
 - Function words: a, an, and, as, for, in, of, the, to, ...
 - About 400 words in English
- Removing stopwords makes some queries difficult to satisfy
 - Example: "*to be or not to be*"

Word Stemming

- **Stemming** is the process for reducing inflected (or sometimes derived) words to their stem – generally a written word form
- The stemming process is often called **conflation**
- Conflate terms manually is difficult and time consuming
- Automatic conflation using rules
 - Suffix stripping: "goes" \Rightarrow "go", "cats" \Rightarrow "cat"
 - Porter stemmer: "police", "policy" \Rightarrow "polic"

Phrase

- A sequence of related words carry a **more specific meaning** than the single words
 - e.g., "*home run*" vs. "*home*" and "*run*"
 - **Any other examples?**
- Two approaches for phrase recognition
 - Statistical approach
 - Part-of-speech tagging

Statistical Approach

- Consider all word bigrams
 - Example: "*hit a*", "*a home*", "*home run*", "*run yesterday*", ...
- Select by collection frequency or document frequency
 - Example: "*home run*" (54), "*run yesterday*" (1)
- If a pattern occurs often, it is probably a phrase
 - Counter example: "*announced yesterday*"

Part-of-Speech (POS) Tagging

- Assign POS tags
 - Usually with a probabilistic or rule-based part of speech tagger
 - Example: "... *hit*/**v** *a*/**art** *home*/**n** *run*/**n** ..."
 - 中文斷詞程式: <http://ckipsvr.iis.sinica.edu.tw/>
- Match phrases by POS patterns
 - n+v: "home run"
 - a+n: "white house"
 - a+n+v: "big home run" (Is this a good phrase?)

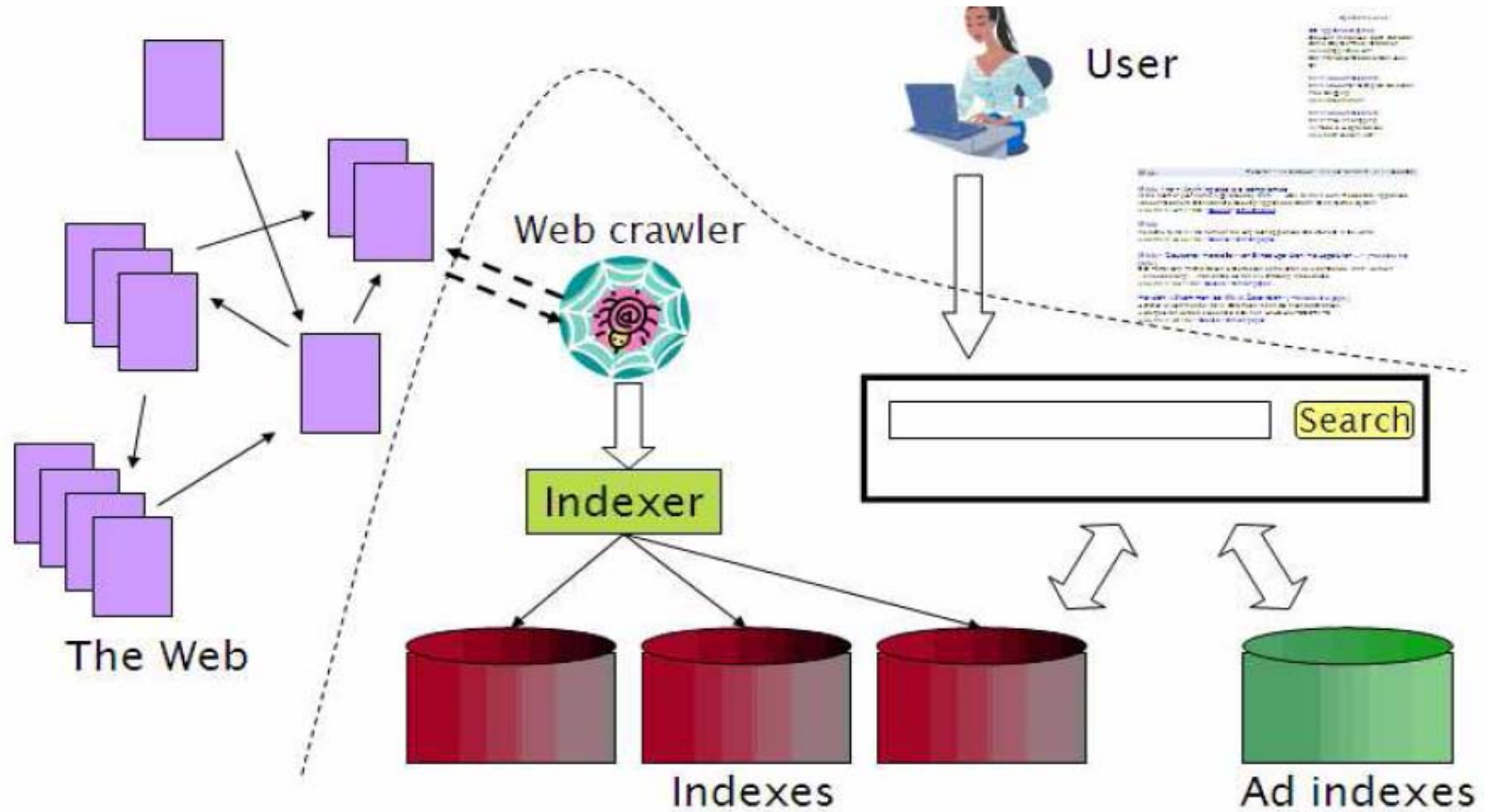
Index Granularity

- Words
 - Basic tokens
 - Word stems (詞幹)
- Phrases
 - Statistical recognition vs. part-of-speech
 - e.g., "information retrieval", "home run"
- Concepts
 - Manual or automatic recognition rules
 - e.g., "about medicine", "in the digital library society"

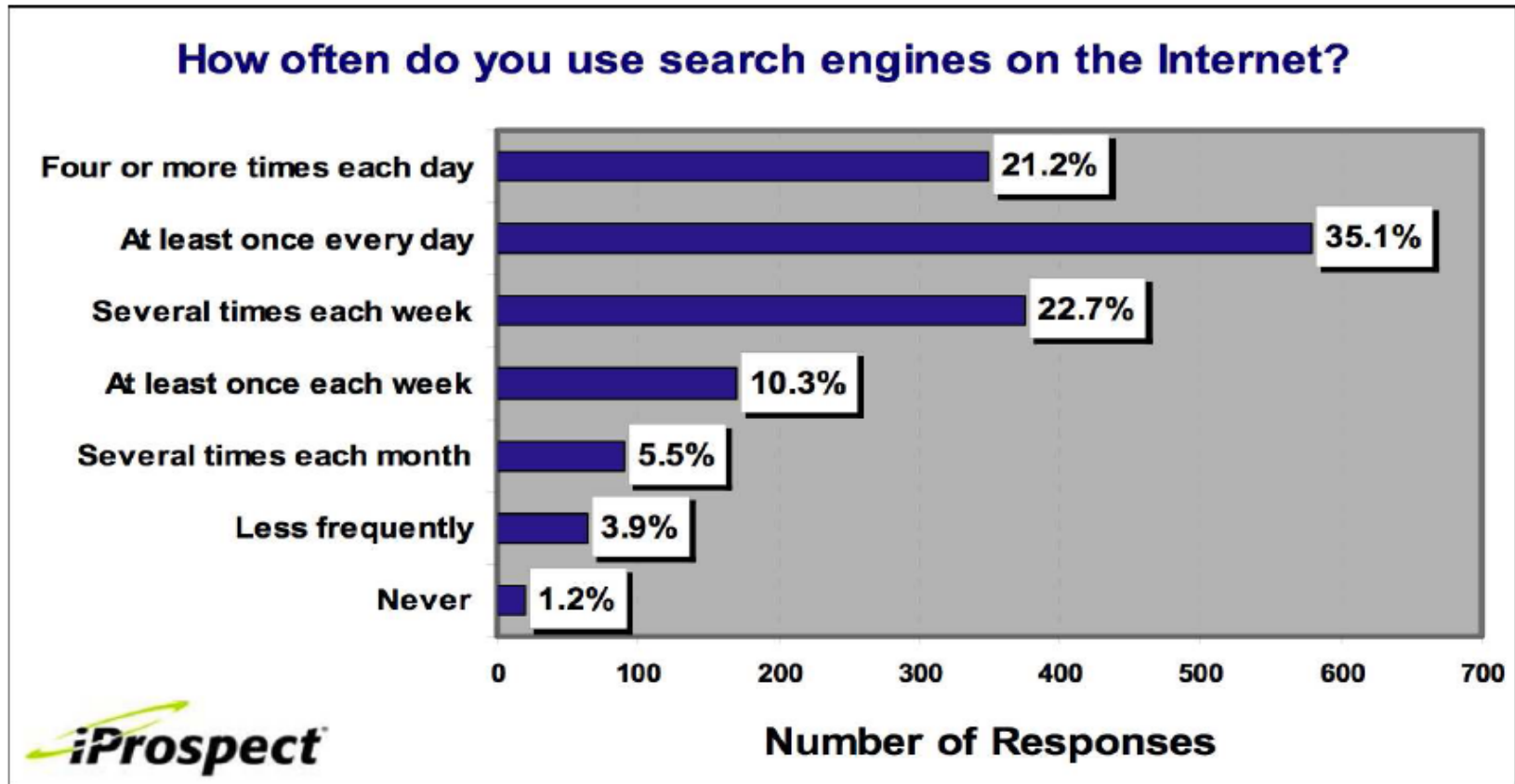
Advanced Topics

- 文件處理與索引建置
- 網路資訊檢索
- 網路廣告

Web Search Overview



Search Is a Top Activity on the Web



Without Search ...

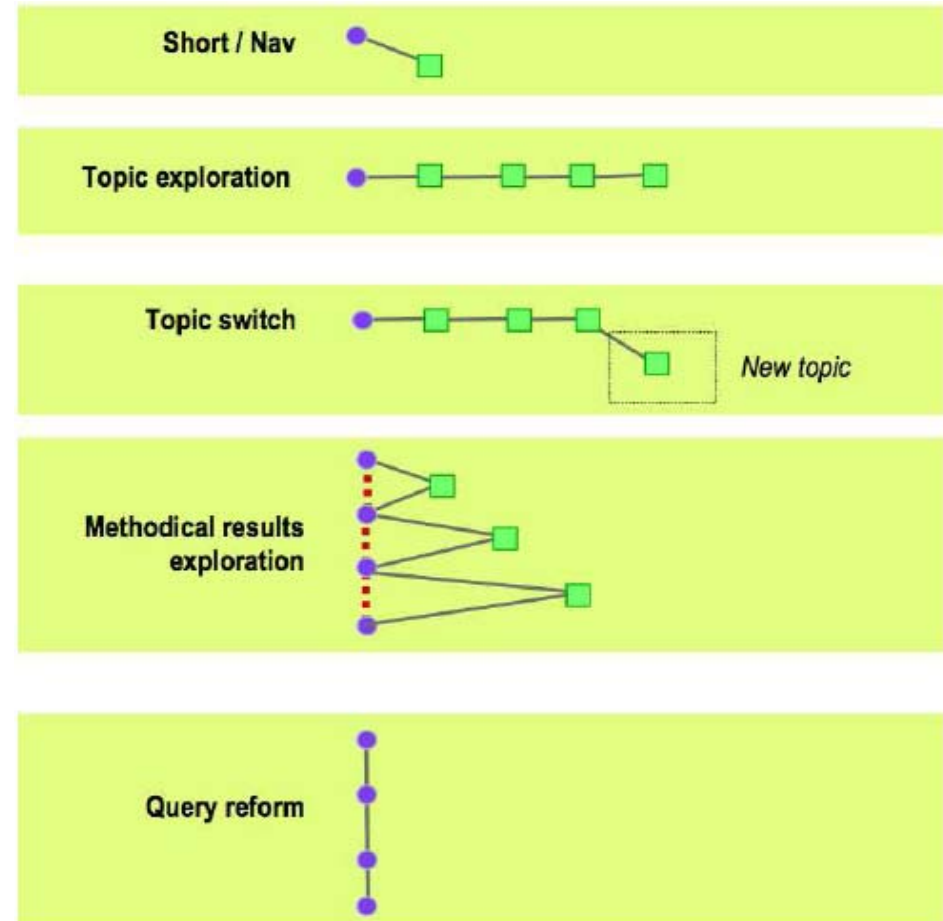
- Without search, content is hard to find in Web
- Without search, there is no incentive to create content
 - Why publish something if nobody will read it
 - Why publish something if I don't get ad revenue from it
- Somebody needs to pay for the Web
 - Servers, web infrastructure, content creation
 - A large part today is paid by search ads

Web IR: Differences from Traditional IR

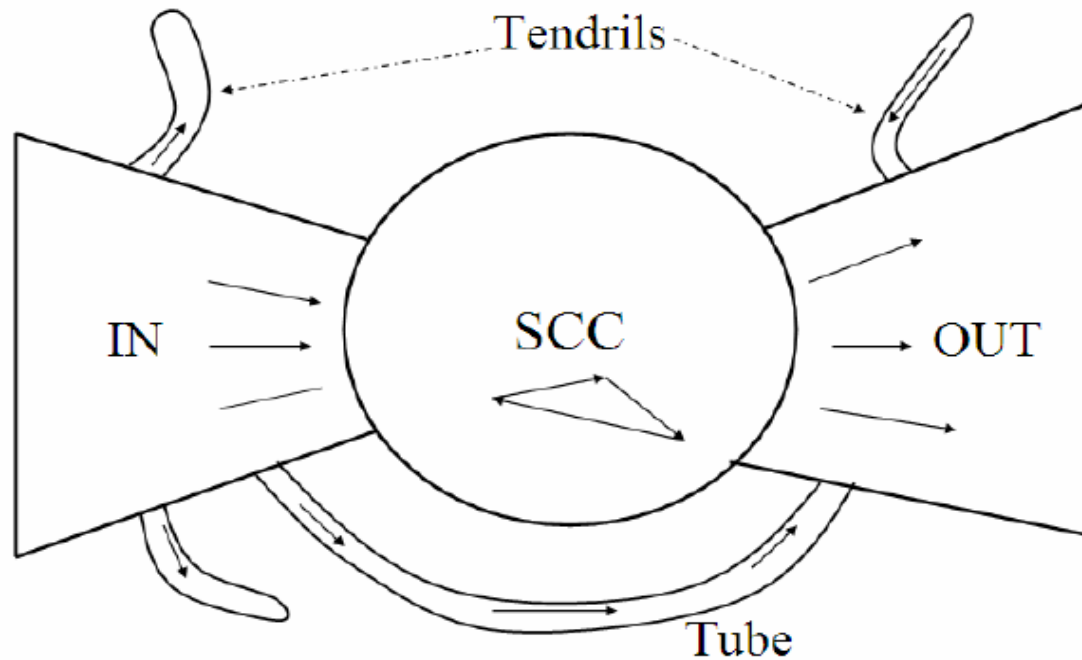
- **Links:** The web is a hyperlinked document collection
- **Queries:** Web queries are different, more varied and there are a lot of them. **How many?** 10^9 every day
- **Users:** Users are different, more varied and there are a lot of them. **How many?** 10^9
- **Documents:** Documents are different, more varied and a lot of them. **How many?** 10^{11} . **Indexed:** 10^{10}
- **Context:** Context is more important on the web than in many other IR applications.
- **Ads and spam**

Links

- Web search in most cases is interleaved with navigation... i.e., with following links



Bowtie Structure of the Web



- Strongly connected component (SCC) in the center
- Lots of pages that get linked to, but don't link (OUT)
- Lots of pages that link to other pages, but don't get linked to (IN)
- Tendrils, tubes, islands

Queries

- Most frequent queries on a large search engine on 2002.10.26

1	sex	16	crack	31	juegos	46	Caramail
2	(artifact)	17	games	32	nude	47	msn
3	(artifact)	18	pussy	33	music	48	jennifer lopez
4	porno	19	cracks	34	musica	49	tits
5	mp3	20	lolita	35	anal	50	free porn
6	Halloween	21	britney spears	36	free6	51	cheats
7	sexo	22	ebay	37	avril lavigne	52	yahoo.com
8	chat	23	sexe	38	www.hotmail.com	53	eminem
9	porn	24	Pamela Anderson	39	winzip	54	Christina Aguilera
10	yahoo	25	warez	40	fuck	55	incest
11	KaZaA	26	divx	41	wallpaper	56	letras de canciones
12	xxx	27	gay	42	hotmail.com	57	hardcore
13	Hentai	28	harry potter	43	postales	58	weather
14	lyrics	29	playboy	44	shakira	59	wallpapers
15	hotmail	30	lolas	45	traductor	60	lingerie

Query Log in Taiwan (AltaVista)

查詢語彙	查詢次數	比例
MP3	42561	1.95%
色情	24970	1.14%
情色	24363	1.12%
sex	20182	0.92%
模擬器	15071	0.69%
icq	13899	0.64%
同志	13622	0.62%
貼圖	12210	0.56%
桌面	12092	0.55%
桌面王	11680	0.53%
寫真集	11640	0.53%
蕃薯藤	10000	0.46%
情色文學	9817	0.45%
寫真	9530	0.44%
奇摩	9328	0.43%
bbs	8613	0.39%
kimo	8166	0.37%
104	7943	0.36%
小說	7456	0.34%
歌詞	7217	0.33%
註：總查詢次數：2,183,506		

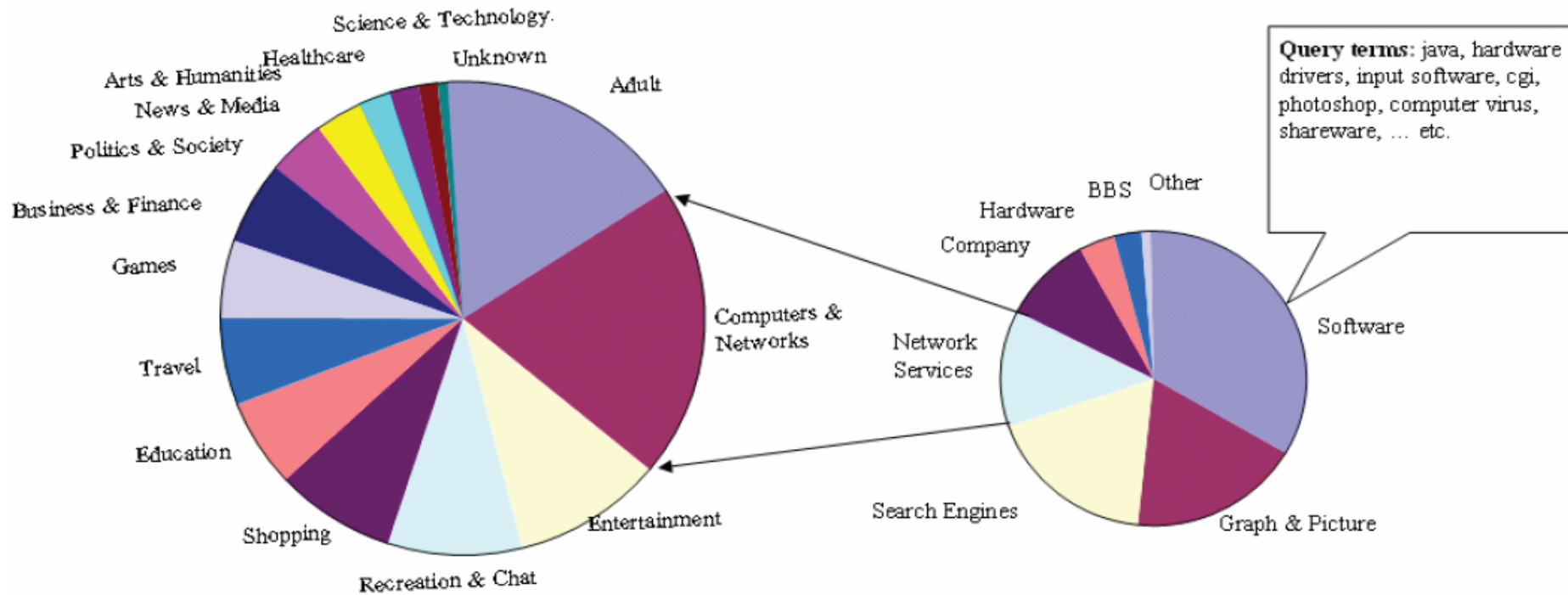
- More than 1/3 of these are queries for adult content
- Does this mean that most people are looking for adult content?
- A few very frequent queries, a large number of very rare queries

Image Query Log in Taiwan (PCHome)

PCHome 2002/01~2002/03

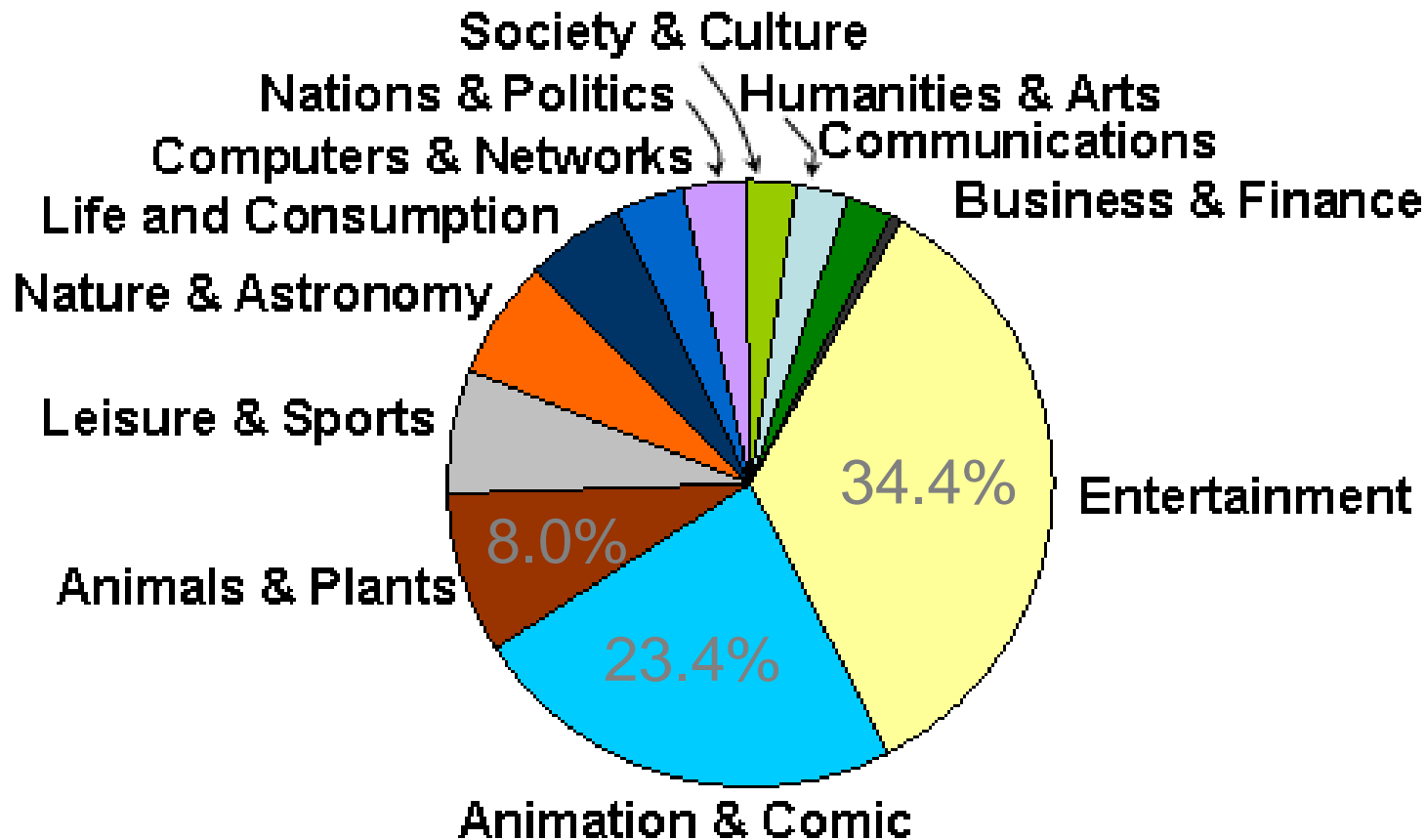
一月熱門查詢詞彙		二月熱門查詢詞彙		三月熱門查詢詞彙	
Qoo	40943	pucca	8079	交大水果妹妹	23608
2002 車展	16179	Qoo	4953	網路美女	16662
美女	8571	大頭狗	4655	水果妹	15485
金城武	6627	鹽水蜂炮	4529	交大水果妹	9277
S.H.E	4577	可愛的圖片	4177	薰衣草	8234
賤兔	3967	神隱少女	2819	蔡依林	5966
孫燕姿	3453	關穎珊	2696	賤兔	5659
怪獸電力公司	3334	哈姆太郎	2531	背景	5409
流氓兔	3113	李英愛	1703	荷莉貝莉	5119
宋慧喬	2564	後藤希美子	1352	丁文琪	4854
背景底圖	2395	許紹洋	1246	背景底圖	4610
西瓜熊	1923	怪獸電力公司	1144	丹佐華盛頓	3568
璩美鳳	1859	櫻花	1135	水果妹妹	3489
BMW	1581	賤兔	1105	美麗境界	3309
S.H.E	1576	松島菜菜子	1096	許慧欣	3244
周杰倫	1263			孫燕姿	2874
深田恭子	1249			周杰倫	2613
天心	1171			哈姆太郎	2594
喬丹	1148			大頭狗	2136

Types of Queries [Chuang2002]



Types of Image Queries

Clustering top 1000 distinct highest-frequent query terms from the image search engine in Taiwan (2002/7-9)



User Needs in Web Search

- **Informational user needs:** Find something about "*low hemoglobin*"
- Other user needs
 - **Navigational user needs:** I want to go to this web site: "*hotmail*", "*myspace*", "*United Airlines*"
 - **Transactional user needs:** I want to make a transaction to buy something: "*MacBook Air*", "*Acrobat Reader*"
 - **Chat with someone:** "*live soccer chat*"
- How can the search engine tell what the user intent for a particular query is?

Context

- What can we do to guess user intent?
- Guess user intent independent of context:
 - Precomputed "typing" of queries
- Better: Guess user intent based on context:
 - Geographic context
 - Context of user in this session (e.g., previous query)
 - Context provided by personal profile (Yahoo/MSN do this, Google claims it doesn't)

Guess by "Typing" of Queries

- Calculation: 5+4
- Unit conversion: 1 kg in pounds
- Currency conversion: 1 euro in kronor
- Tracking number: 8167 2278 6764
- Flight info: LH 454
- Area code: 650
- Map: 動物園導覽
- Stock price: msft
- Albums/movies etc: 海角七號

The Spatial Context: Geo-search

- Three relevant locations
 - Server (nytimes.com → New York)
 - User (located in Palo Alto)
 - Web page content (nytimes.com article about Albania)
- Locating the user
 - IP address
 - Information provided by user (e.g., in user profile)
 - Mobile phone and GPS
- **Geo-tagging:** parse text and identify the coordinates of the geographic entities (**Important NLP problem!**)
 - East Palo Alto CA → Latitude: 37.47 N, Longitude: 122.14 W

Use Context to Modify Query Results

- **Result restriction:** Don't consider inappropriate results
 - For user on google.fr ... only show .fr results
- **Ranking modulation:** use a rough generic ranking, rerank based on personal context
- **Contextualization/personalization** is an area of search with a lot of potential for improvement

Users

- Use short queries (average < 3)
- Rarely use operators
- Don't want to spend a lot of time on composing a query
- Only look at the first couple of results
- Want a simple UI, or like a information portal?
- Extreme variability in terms of user needs, user expectations, experience, knowledge, . . .
 - Industrial/developing world, English/Chinese, old/young, rich/poor, differences in culture and class

Term Length per Query (Taiwan)

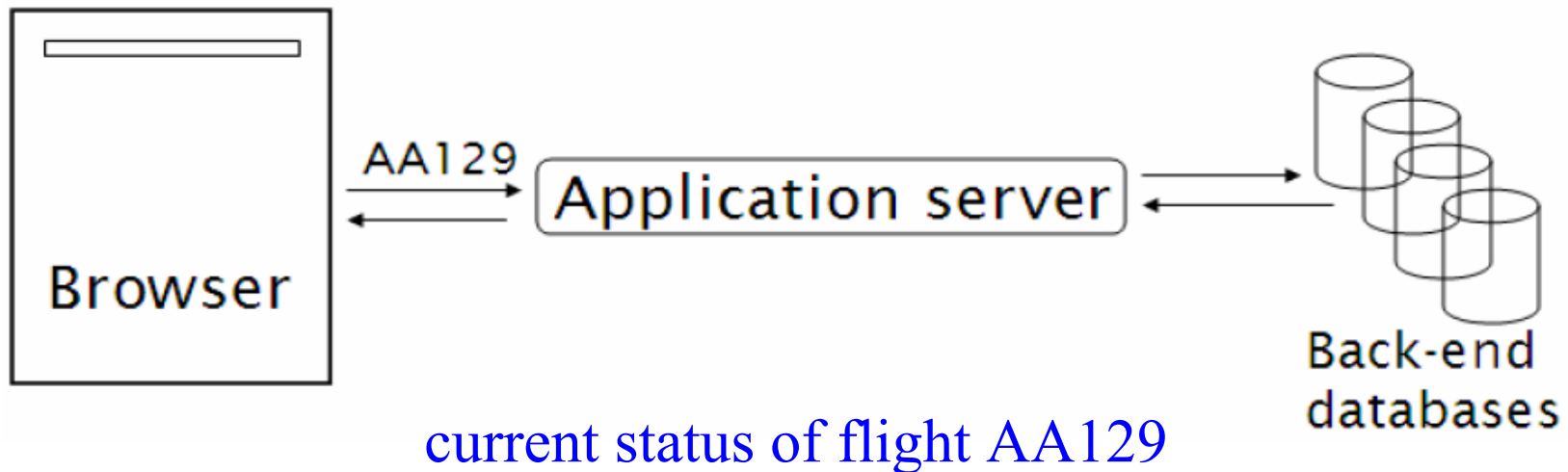
	In Chinese	In English	All
Dreamer	3.18 characters	1.22 words	6.31 bytes
GAIS	3.55 characters	1.10 words	7.26 bytes

How Do Users Evaluate Search Engines

- Classic IR relevance can also be used for web IR
- Trust, duplicate elimination, readability, loads fast, no pop-ups, ... (any others?)
- On the web, precision is more important than recall
 - Precision at 1, precision at 10, precision on the first 2-3 pages
 - But there is a subset of queries where recall matters

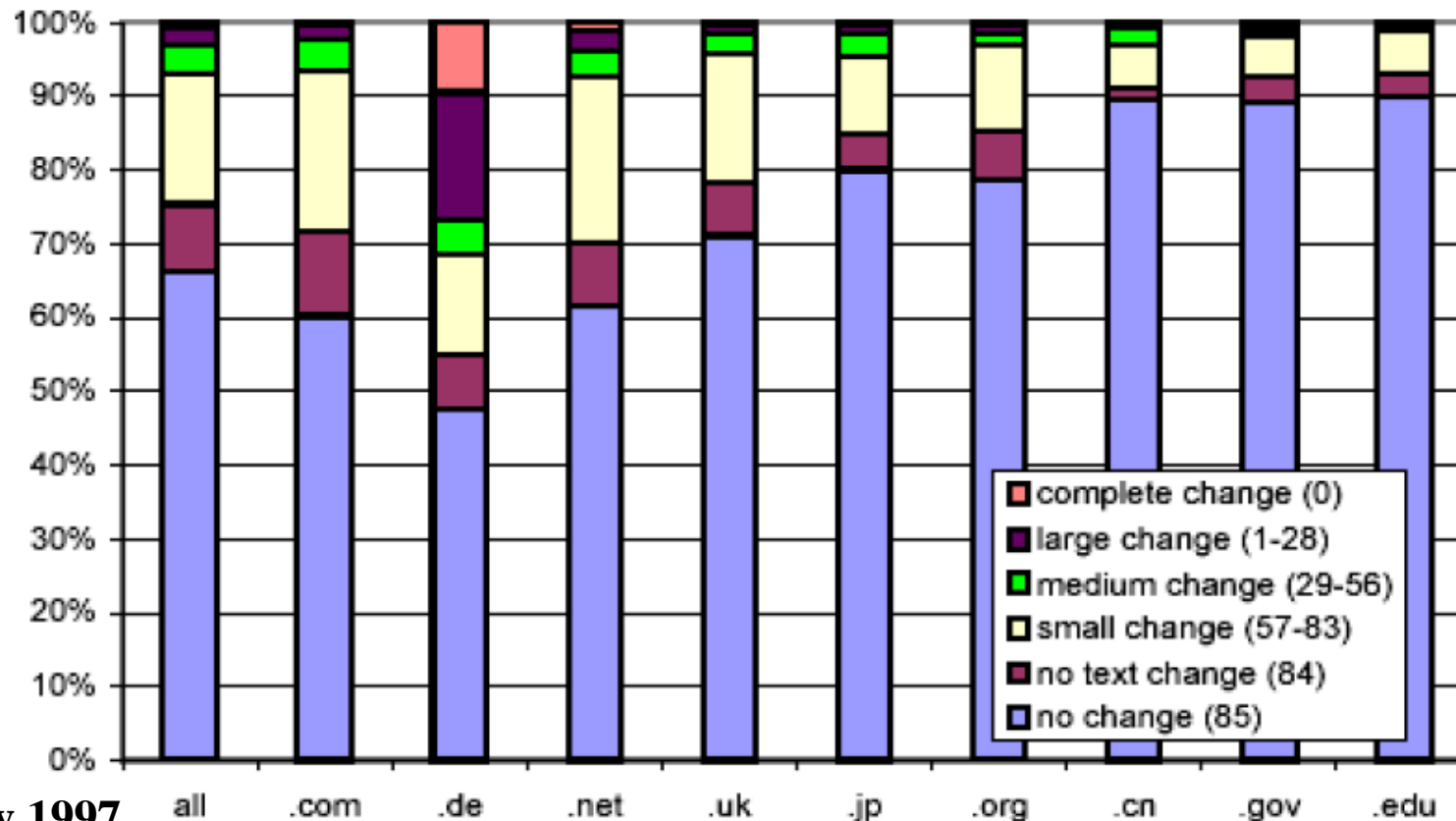
Documents

- Some documents are **generated dynamically** from scratch when the user requests them – usually from underlying data in a database



Dynamic Content

- Most (truly) dynamic content is ignored by web spiders. It's too much to index it all



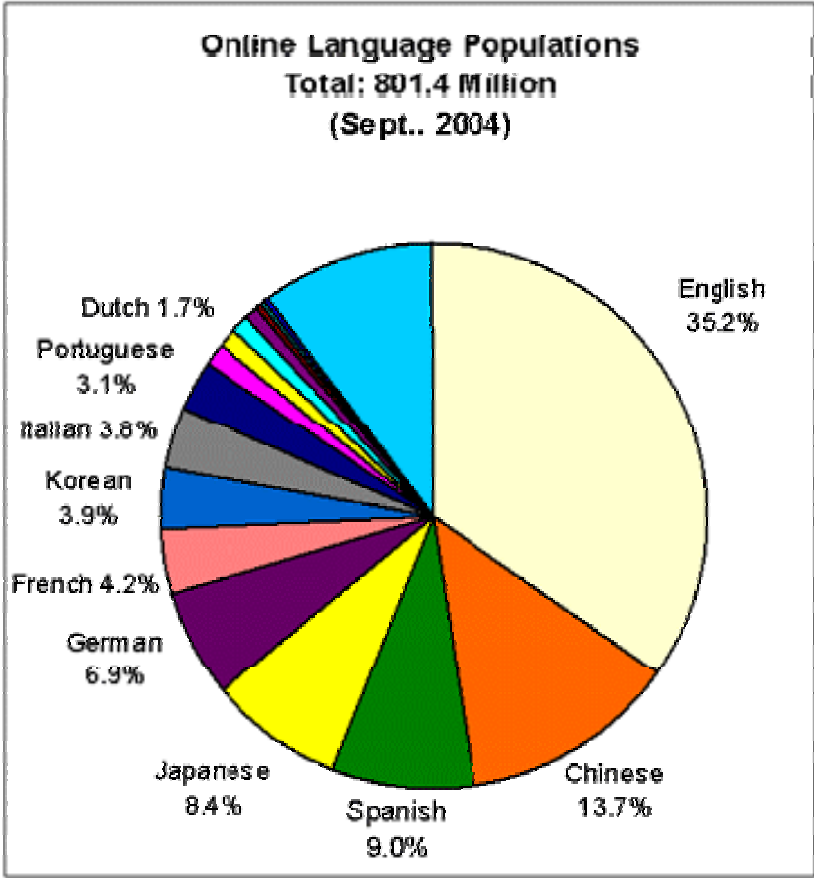
Fetterly 1997

2009/1/8

Multilinguality

- Documents in a large number of languages
- Queries in a large number of languages
- First cut: Don't return English results for a Chinese query
- However: Frequent mismatches query/document languages
- Translation is important

Online Language Population



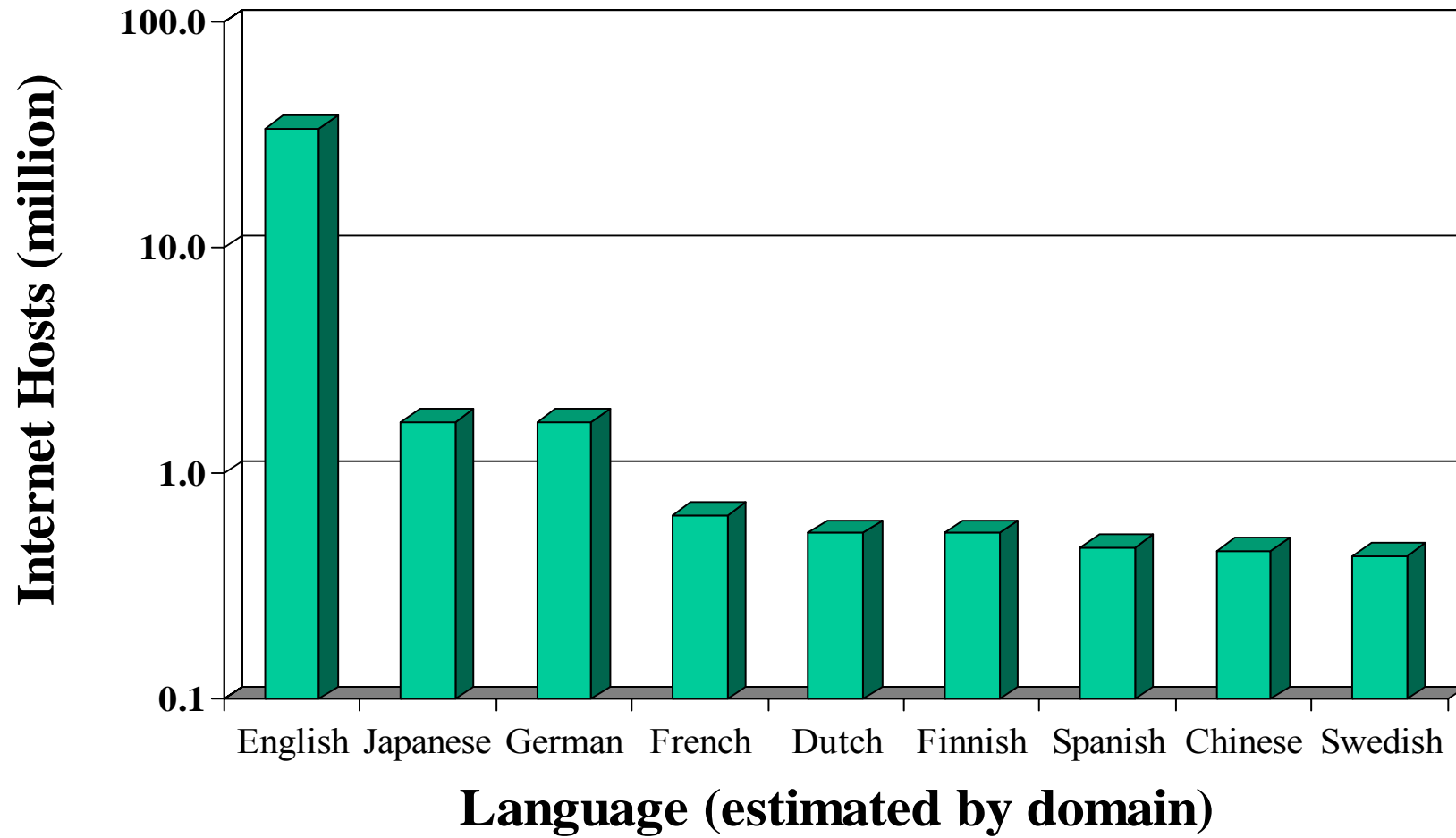
Global Reach

Top Ten Languages Used in the Web

Internet World Stats (Jun. 30, 2006)

TOP TEN LANGUAGES IN THE INTERNET	Internet Users, by Language	% of all Internet Users	World Population 2006 Estimate for Language	Internet Penetration by Language	Internet Growth for Language (2000 - 2005)
English	312,757,646	30.6 %	1,125,664,397	27.8 %	128.0 %
Chinese	132,301,513	13.0 %	1,340,767,863	9.9 %	309.6 %
Japanese	86,300,000	8.5 %	128,389,000	67.2 %	83.3 %
Spanish	80,593,698	7.9 %	429,293,261	18.8 %	229.2 %
German	56,853,104	5.6 %	95,982,043	59.2 %	106.0 %
French	40,974,004	4.0 %	381,193,149	10.7 %	235.9 %
Korean	33,900,000	3.3 %	73,945,860	45.8 %	78.0 %
Portuguese	32,372,000	3.2 %	230,846,275	14.0 %	327.3 %
Italian	28,870,000	2.8 %	59,115,261	48.8 %	118.7 %
Russian	23,700,000	2.3 %	143,682,757	16.5 %	664.5 %
TOP TEN LANGUAGES	828,621,965	81.0 %	4,008,879,867	20.7 %	156.0 %
Rest of World Languages	194,241,342	19.0 %	2,490,817,193	7.8 %	421.6 %
WORLD TOTAL	1,022,863,307	100.0 %	6,499,697,060	15.7 %	183.4 %

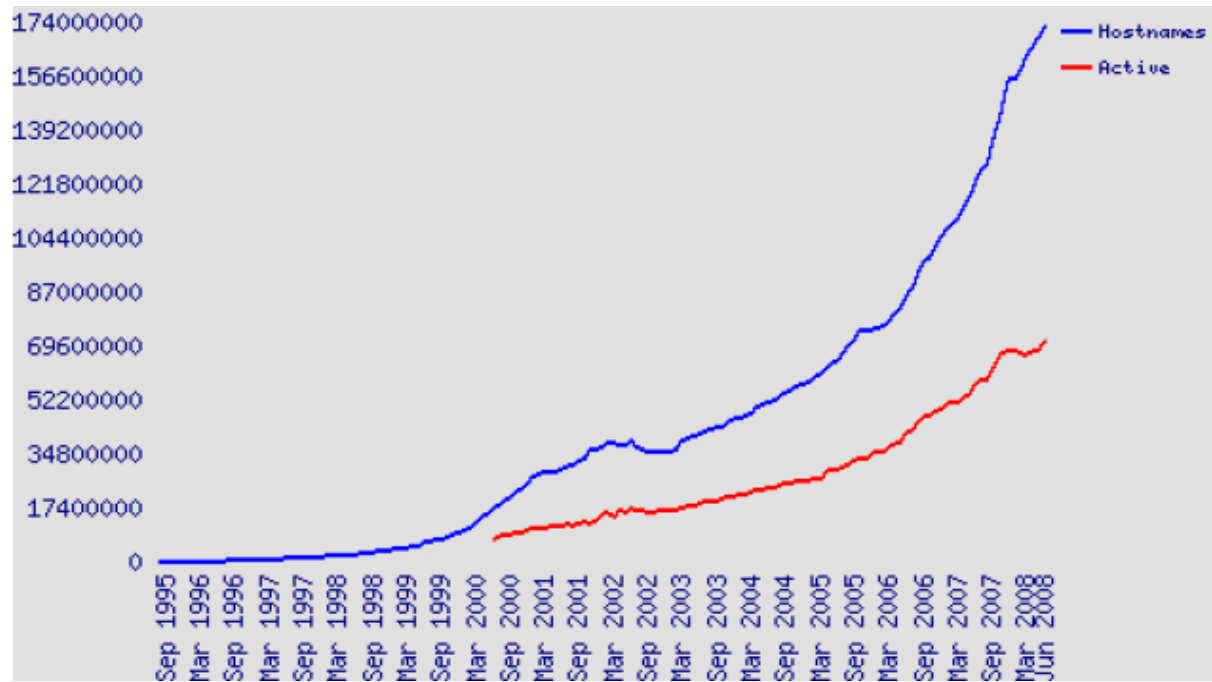
Document Languages



Duplicate Document

- Significant duplication: 30%-40% duplicates in some studies
- Duplicates in the search results were common in the early days of the web
- Today's search engines eliminate duplicates very effectively

Size



The Web keeps growing

Who Cares about the Web Size

- Search engine designers
 - How many pages do I need to be able to handle?
- Crawler designers
 - which policy will crawl close to N pages?
- Users
 - They may switch to the search engine that has the best coverage of the web
- Media

How to Estimate the Size of the Web

- OR-query of frequent words in a number of languages
- But page counts of google search results are only rough estimates

Number of Chinese Web Pages

所有網頁 圖片 新聞 網上論壇 雜誌搜尋 Gmail 更多 ▼ chihyi.chiu@gmail.com | [我的帳戶](#) | [登出](#)

Google 的 [進階搜尋](#) | [使用偏好](#)

搜尋： 所有網頁 中文網頁 繁體中文網頁 台灣的網頁

所有網頁 約有4,930,000,000項符合的查詢結果，以下是第 1-10項，共費0.22 秒。

[史萊姆的第一個家](#)
軟體搜集, 下載, 教學. 每日更新, 軟體種類眾多.
[www.slime.com.tw/](#) - 2k - [頁庫存檔](#) - [類似網頁](#) - [加入筆記本](#)

[歷史上的今天](#) - 簡 - [[轉為繁體網頁](#)]
可以了解歷史的這一天發生的事件, 含查詢系統。
[www1.wst.net.cn/scripts/flex/TodayOnHistory/](#) - 7k - [頁庫存檔](#) - [類似網頁](#) - [加入筆記本](#)

[老徐-徐靜蕾-新浪BLOG](#) - 簡 - [[轉為繁體網頁](#)]
前幾日差點丟了, 又說起有人吃狗肉什麼的, 嚇得本博趕緊不由分說當即驅車30 ... 後天早上5點半, 最高溫度零下8度的嚴寒, 本老博要與之奮戰, 據說有人很崩潰, 老博做人 ...
[blog.sina.com.cn/xujinglei](#) - 71k - [頁庫存檔](#) - [類似網頁](#) - [加入筆記本](#)

[音樂的遐思](#)
香港電台網上廣播站e-Learning爲了推廣古典音樂, 於2003年9月30日推出《音樂的遐思》網頁, 邀得著名作家李歐梵教授設計了十個單元, 由淺入深, 跟網友介紹他最喜歡的 ...
[www.rthk.org.hk/elearning/musicfantasy/](#) - 4k - [頁庫存檔](#) - [類似網頁](#) - [加入筆記本](#)

[我的E政府](#)
本網頁使用script可是您的瀏覽器並不支援, 使用的script並沒有影響您閱讀本站網頁的 ... 主計處資料顯示, 女性雇主及自營人數逐年成長, 請問您覺得女性想自行當老闆的 ...
[www.gov.tw/](#) - 43k - [頁庫存檔](#) - [類似網頁](#) - [加入筆記本](#)

[UrMap你的地圖網](#)
提供給您最快速、最方便的地圖搜尋引擎, 只要輸入地址、路名或關鍵字就可以查到您所要的位置, 還可以看到全台灣的最新衛星影像。
[www.urmap.com/](#) - 11k - [頁庫存檔](#) - [類似網頁](#) - [加入筆記本](#)

[鳥哥的Linux私房菜](#)
鳥哥的Linux私房菜, VBird's Linux Technical Documents.
[linux.vbird.org/](#) - 1k - [頁庫存檔](#) - [類似網頁](#) - [加入筆記本](#)

4,930,000,000 pages

Advanced Topics

- 文件處理與索引建置
- 網路資訊檢索
- 網路廣告

First Generation of Search Ads: Goto (1996)



- No separation of ads/docs. Just one result!
- Buddy Blake bid the maximum (\$0.38) for this search.
- He paid \$0.38 to Goto every time somebody clicked on the link.

Ranking of Advertisers in Search Results

Web Images Maps News Shopping Gmail more Sign In

Google Search [Advanced Search](#) [Preferences](#)

Web Results 1 - 10 of about 807,000 for discount broker [definition]. (0.12 seconds)

Discount Broker Reviews
Information on online **discount brokers** emphasizing rates, charges, and customer comments and complaints.
www.broker-reviews.us/ - 94k - [Cached](#) - [Similar pages](#)

Discount Broker Rankings (2008 Broker Survey) at SmartMoney.com
Discount Brokers. Rank/ **Brokerage/** Minimum to Open Account, Comments, Standard Commis- sion*, Reduced Commission, Account Fee Per Year (How to Avoid), Avg. ...
www.smartmoney.com/brokers/index.cfm?story=2004-discount-table - 121k - [Cached](#) - [Similar pages](#)

Stock Brokers | Discount Brokers | Online Brokers
Most Recommended. Top 5 **Brokers** headlines. 10. Don't Pay Your **Broker** for Free Funds May 15 at 3:39 PM. 5. Don't **Discount** the Discounters Apr 18 at 2:41 PM ...
www.fool.com/investing/brokers/index.aspx - 44k - [Cached](#) - [Similar pages](#)

Discount Broker
Discount Broker - Definition of **Discount Broker** on Investopedia - A stockbroker who carries out buy and sell orders at a reduced commission compared to a ...
www.investopedia.com/terms/d/discountbroker.asp - 31k - [Cached](#) - [Similar pages](#)

Discount Brokerage and Online Trading for Smart Stock Market ...
Online stock **broker** SogoTrade offers the best in **discount brokerage** investing. Get stock market quotes from this internet stock trading company.
www.sogotrade.com/ - 39k - [Cached](#) - [Similar pages](#)

15 questions to ask discount brokers - MSN Money
Jan 11, 2004 ... If you're not big on hand-holding when it comes to investing, a **discount broker** can be an economical way to go. Just be sure to ask these ...
moneycentral.msn.com/content/Investing/Startinvesting/P66171.asp - 34k - [Cached](#) - [Similar pages](#)

Sponsored Links

Rated #1 Online Broker
No Minimums. No Inactivity Fee
Transfer to Firsttrade for Free!
www.firsttrade.com

Discount Broker
Commission free trades for 30 days.
No maintenance fees. Sign up now.
TDAMERITRADE.com

TradeKing - Online Broker
\$4.95 per Trade, Market or Limit
SmartMoney Top **Discount Broker** 2007
www.TradeKing.com

Scottrade Brokerage
\$7 Trades, No Share Limit. In-Depth Research. Start Trading Online Now!
www.Scottrade.com

Stock trades \$1.99-\$3
100 free trades, up to \$100 back for transfer costs, \$500 minimum
www.sogotrade.com

\$3.95 Online Stock Trades
Market/Limit Orders, No Share Limit and No Inactivity Fees
www.Marsco.com

INGDIRECT | ShareBuilder

SogoTrade appears in search results.

SogoTrade appears in ads.

Do search engines rank advertisers higher than non-advertisers?

All major search engines claim no.

Do Ads Influence Editorial Content?

- Similar problem at newspapers/TV channels
- A newspaper is reluctant to publish harsh criticism of its major advertisers
- No known case of this happening with search engines yet?

How Are Ads Placed?

- Advertisers bid for keywords
- Open system: Anybody can participate and bid on keywords
- Advertisers are only charged when somebody clicks on your ad
- **How does the advertiser determine its bid price?**
 - Basis is a **second price auction**, but with twists
 - Squeeze an additional fraction of a cent from each ad means billions of additional revenue for the search engine.

Keywords with High Bids

According to <http://www.cwire.org/highest-paying-search-terms/>

\$69.1	mesothelioma treatment options	\$35.86	pennsylvania medical malpractice attorney
\$65.85	personal injury lawyer michigan	\$35.86	medical malpractice ohio
\$62.59	student loans consolidation	\$35.71	automobile insurance quote
\$61.44	car accident attorney los angeles	\$35.4	loan consolidating
\$59.44	online car insurance quotes	\$35.34	commercial insurance quote
\$59.39	arizona dui lawyer	\$35.33	tax attorney
\$57.87	michigan car accident attorney	\$35.15	home equity loans
\$56.59	free auto insurance quote	\$34.81	instant auto insurance quotes
\$53.17	personal injury lawyers los angeles	\$34.8	home equity loan rates
\$52.31	free online auto insurance quote	\$34.79	home owners insurance quotes
\$50.4	accident attorney michigan	\$34.71	home equity line
\$50.35	michigan auto accident attorney	\$34.53	compensation solicitors
\$49.25	accident helpline	\$34.38	automobile insurance quotes
\$47.74	automobile accident lawyers	\$34.37	term insurance quotes
\$47.49	dui defense attorneys	\$34.26	instant car insurance quotes
\$46.44	asbestos cancer	\$34.02	auto insurance online quote
\$46.34	arizona dui	\$33.49	new york criminal attorney
\$45.8	business liability insurance quote	\$33.45	secured loan
\$43.86	loan consolidation	\$33.44	equity lines
\$42.98	student loan consolidation	\$33.41	criminal lawyer new york
\$40.7	dui defense lawyers	\$33.36	refinance mortgage
\$40.1	home equity line of credit	\$33.12	equity loan rates
\$39.81	life insurance quotes	\$33.07	manhattan mini storage
\$39.78	criminal lawyers new york	\$32.46	equity line
\$39.32	loan federal consolidation	\$32.45	home equity credit
\$39.23	refinancing	\$32.02	loan consolidate
\$38.72	equity line of credit	\$31.98	secured loan consolidation
\$37.96	lasik eye surgery new york city	\$31.93	laser hair removal new york city
\$37	2nd mortgage	\$31.51	home equity rates
\$35.9	free car insurance quote	\$31.37	free credit report com

Ad Ranking

- First cut: according to bid price
 - Bad idea: open to abuse
 - Example: query "*accident*" → ad "*buy a new car*"
- Instead: rank based on bid price and relevance
- Key measure of ad relevance: **clickthrough rate**
- Result: A non-relevant ad will be ranked low
 - Hope to achieve win-win-win long-term
- Other ranking factors: **location, time of day, quality, and loading speed of landing page**

A Win-Win-Win?

- The **search engine company** gets revenue every time somebody clicks on an ad
- The **user** only clicks on an ad if they are interested in the ad
 - Search engines punish misleading and nonrelevant ads
 - As a result, users are often satisfied with what they find after clicking on an ad
- The **advertiser** finds new customers in a cost-effective way

A Win-Win-Win?

- Why is web search potentially more attractive for advertisers than TV spots, newspaper ads or radio spots?
- Someone who just searched for "Saturn Aura Sport Sedan" is infinitely more likely to buy one than a random person watching TV
- Most importantly, the advertiser only pays if the customer took an action indicating interest (i.e., clicking on the ad)

Actually It's often not a Win-win-win

- Example: keyword arbitrage
 - Buy a keyword at Google
 - Then redirect traffic to a third party that is paying much more than you had to pay to Google
 - This rarely makes sense for the user
- Ad spammers keep inventing new tricks
- The search engines need time to catch up with them

Who Own a Search Term?

- Example: geico
 - During part of 2005: The search term “geico” on Google was bought by competitors.
 - Geico lost this case in the United States
- Currently in the courts: Louis Vuitton case in Europe

Thanks for Your Attention!

- Any question?
- Available resource
 - 數位典藏技術導論 <http://ebook.iis.sinica.edu.tw>
 - Introduction to Information Retrieval
<http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html> (Some slide contents are excerpted from this book)