Web Search Basics

Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University

References:

1. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008. (Chapters 19 – 21 & associated slides)
2. Raymond J. Mooney’s teaching materials
3. Lan Huang. A Survey on Web Information Retrieval Technologies. Available at: <http://citeseer.nj.nec.com/336617.html>

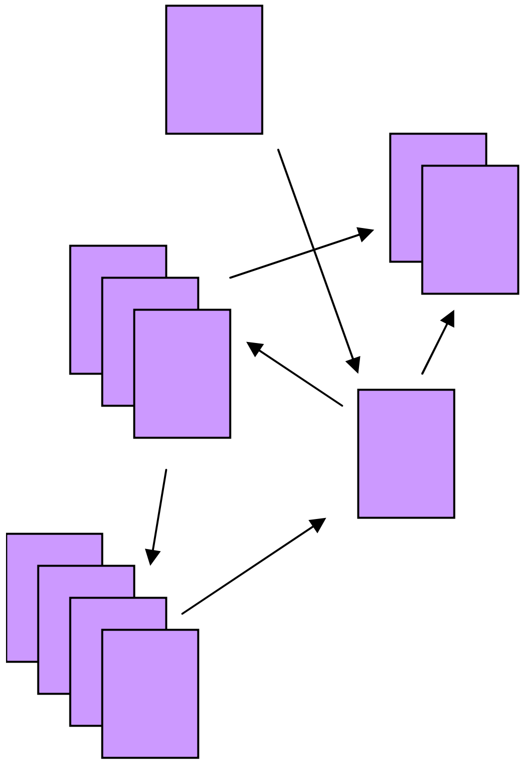
The World Wide Web (Web)

- Created in 1989 by Tim Berners-Lee at CERN (in Switzerland)

- An environment of accessing to interlinked and hypertext documents via the Internet
  - Client-server design for transfer text, images, videos, and other multimedia, encoded with html (hypertext markup language), via a protocol (http, hypertext transfer protocol)
    - The client side is usually a browser, a GUI environment, sending an http request to a web server (by specifying a URL, universal resource locator)
    - Asynchronous communication

$$\text{http://}\underbrace{\text{www.ntnu.edu.tw}}_{\text{domain}}\text{/infomation/contact.html}$$

Web Characteristics

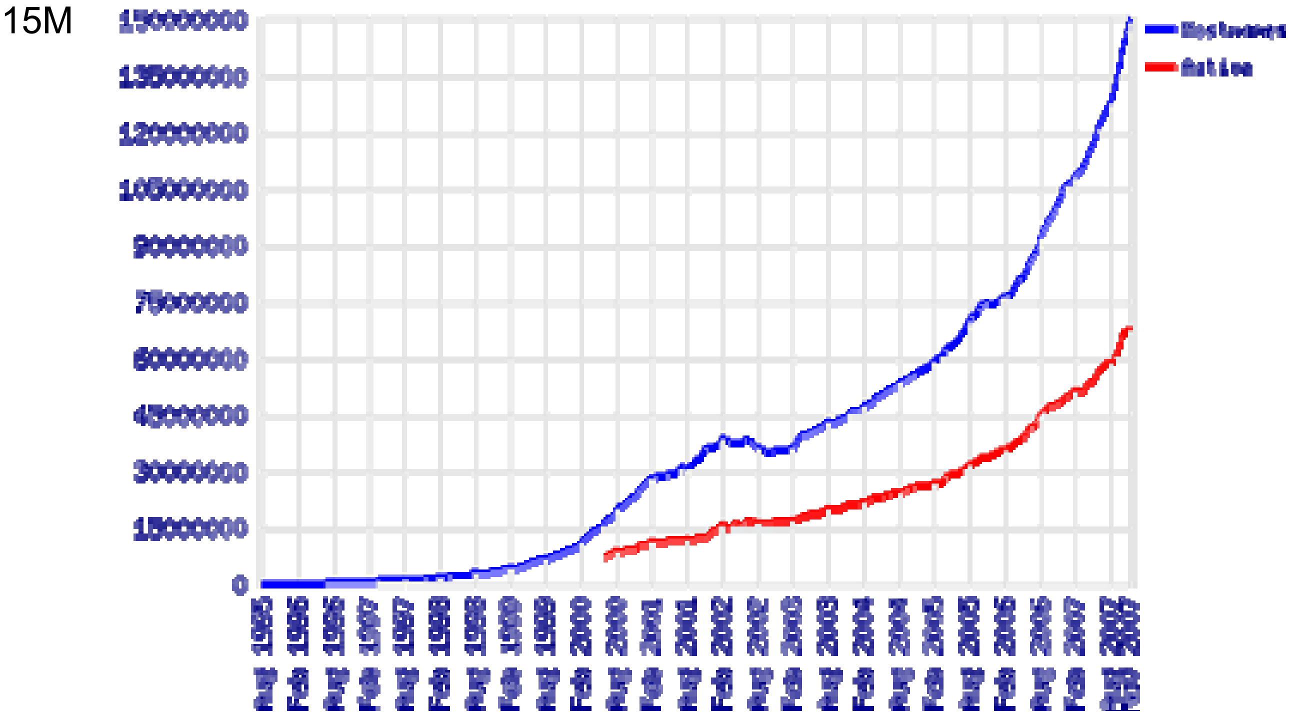


The Web

- No Control: democratization of creation and linking (publishing). Content includes truth, lies, obsolete information, contradictions
- Distributed Data: Documents spread over millions of different web servers…
- Heterogeneity: Unstructured (text, html, …), semi-structured (XML, annotated photos), structured (databases)…
- Variety of Languages: The types of languages used are more than 100
- Large Volume: Scale much larger than previous text corpora (slowed down from initial "volume doubling every few months" but still expanding)
- Volatile Data: content can be dynamically generated and removed
- …

Rapid Proliferation of Web Content

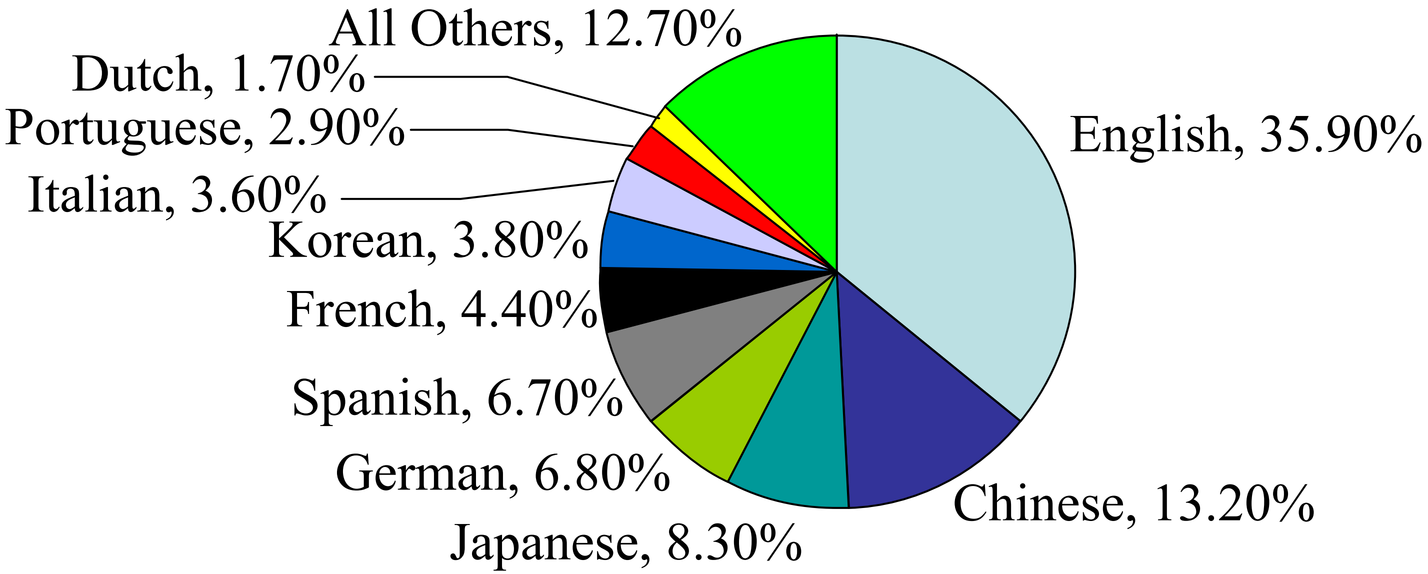
- Total Web Sites Across All Domains August 1995 - November 2007 (http://news.netcraft.com)



  – A large fraction of growth in sites has come from the increasing number of blogging sites (in particular at Live Spaces, Blogger and MySpace) in the recent past

Internet Users by Languages

- End of 2004, total 801.4 millions

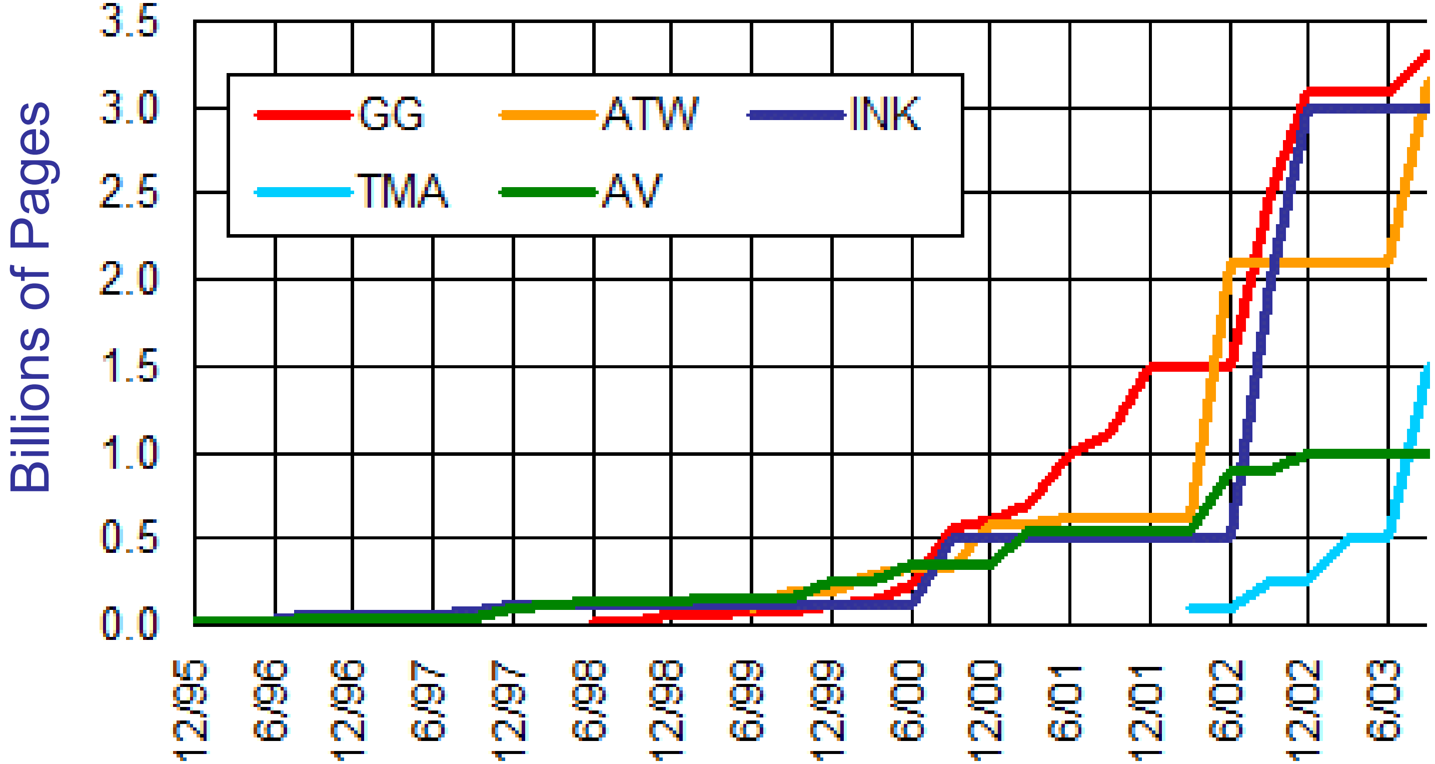


Access to Web Content

- Full-text index search engines
  - E.g., Google, Altavista, Excite, Infoseek, etc.
  - Keyword search supported by inverted indexes and ranking mechanisms

- Manual hierarchical taxonomies (directories) populated with web pages in categories
  - E.g., Yahoo!, Yam, etc.
  - Human editors assemble a large hierarchically structured directory of web pages
  - Users browse through trees of category labels

Growth of Web Pages Indexed



Google
Inktomi
AllTheWeb

Teoma
Altavista

SearchEngineWatch

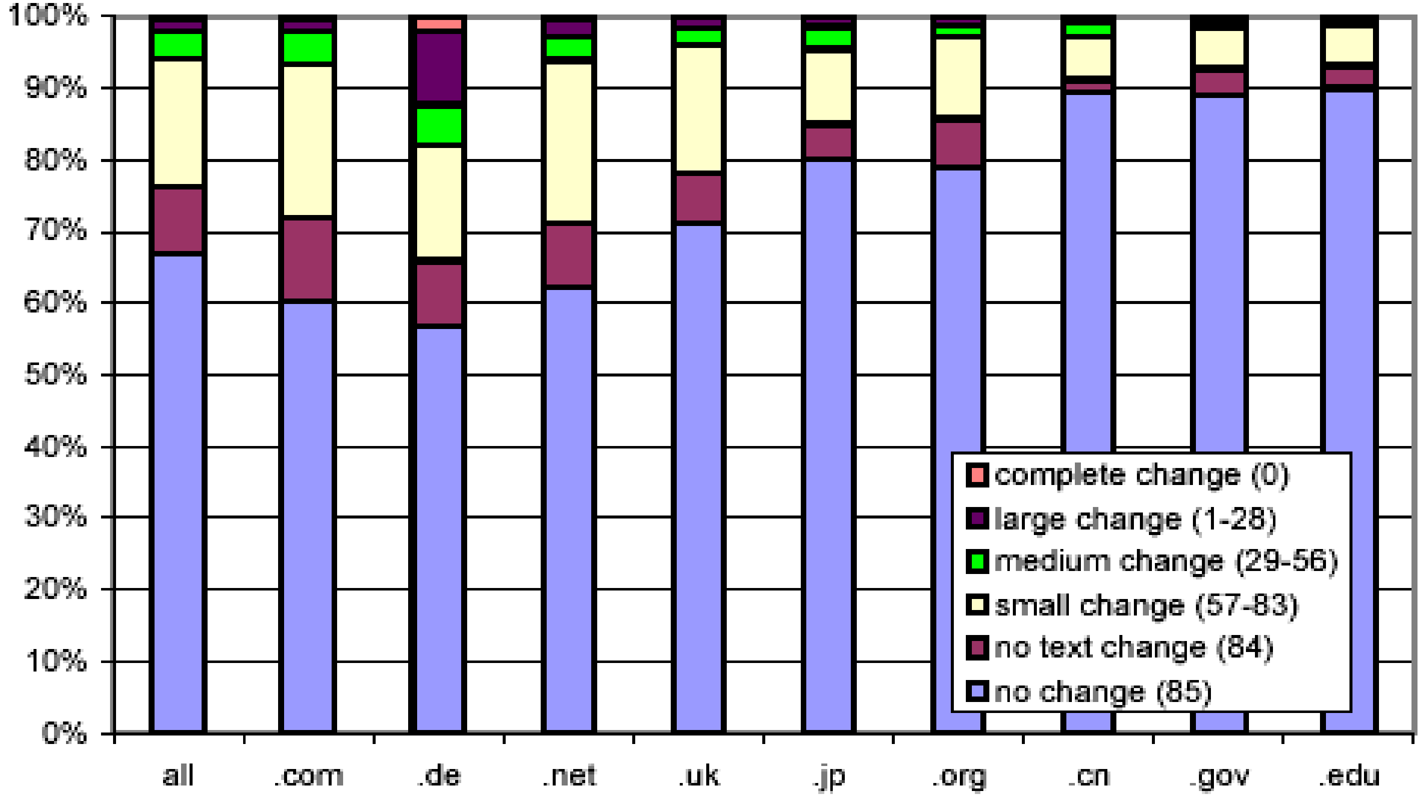
Link to Note from Jan 2004

Assuming 20KB per page,
1 billion pages is about 20 terabytes of data.

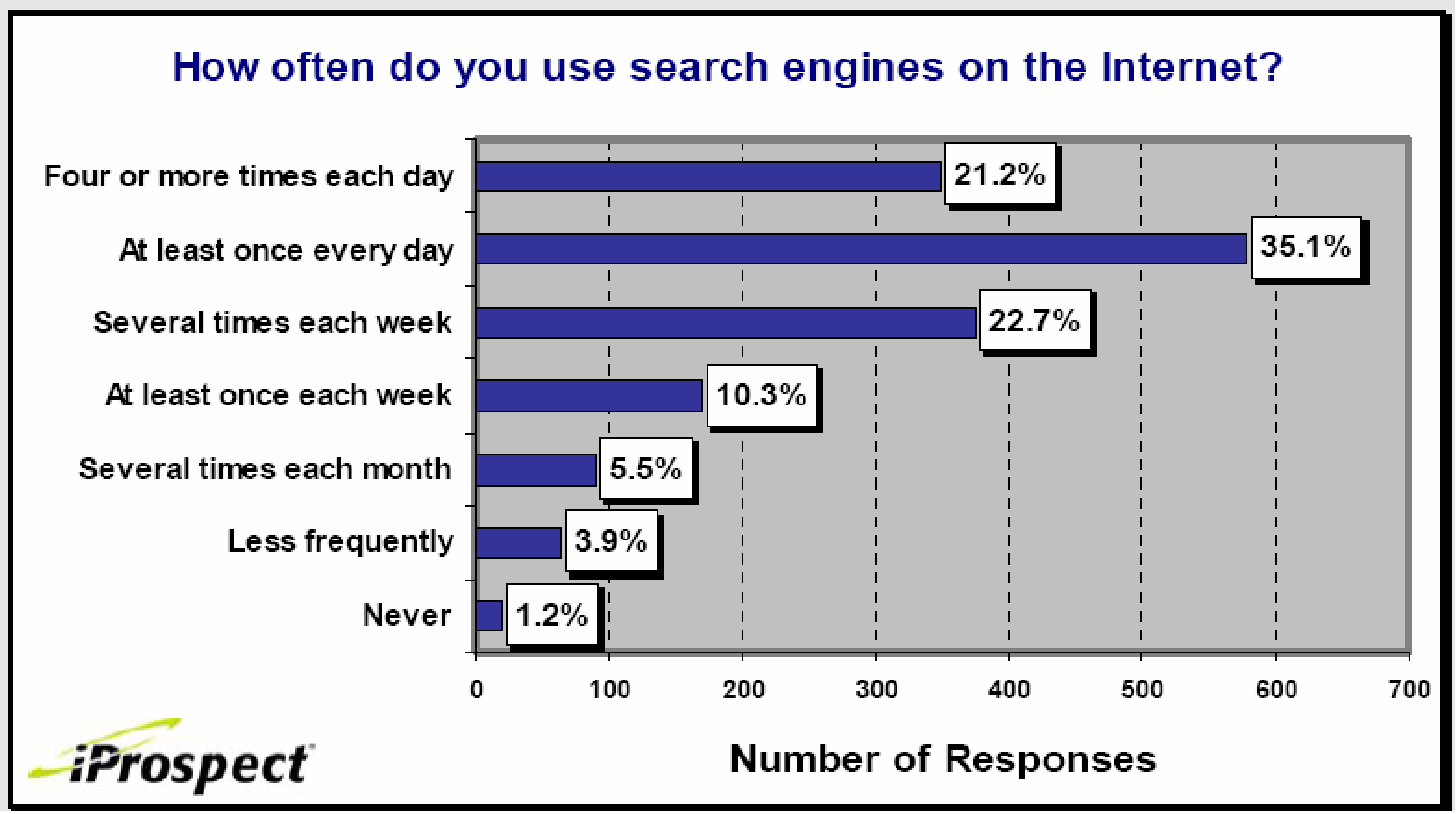
- This slide is adopted from Raymond J. Mooney's teaching materials

Rate of Change for Web Pages

- Fetterly et al. study (2002): several views of data, 150 million pages over 11 weekly crawls
    - Bucketed into 85 groups by extent of change



Frequency of Using Search Engines



http://www.iprospect.com

User Query Needs (1/4)

- User query roughly fall into three categories
  - Informational – want to learn about something
    - E.g., “Taroko”
  - Navigational – want to go to that page
    - E.g., “China Airlines”
  - Transactional – want to do something (web-mediated)
    - Purchasing a product, downloading a file or making a reservation

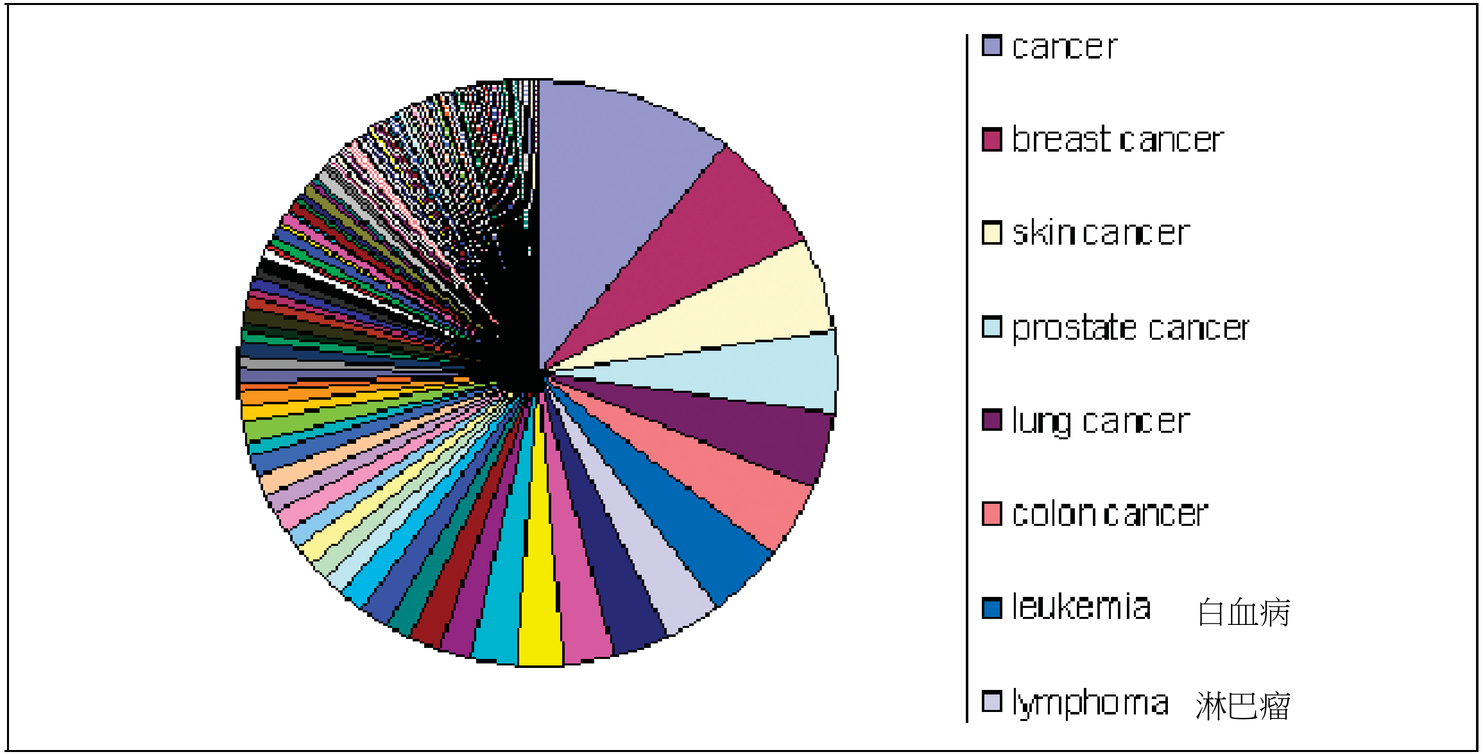
Discern which of these categories a query falls into can be challenging !

User Query Needs (2/4)

- Ill-defined queries
  - Short
    - 2001: 2.54 terms avg, 80% < 3 words
    - 1998: 2.35 terms avg, 88% < 3 words
  - Imprecise terms
  - Suboptimal syntax
  - Low effort
- Specific behavior
  - 85% look over one result screen only (mostly above the fold)
  - 78% of queries are not modified (one query/session)
- Wide variance in
  - Needs
  - Expectations
  - Knowledge
  - Bandwidth

User Query Needs (3/4)

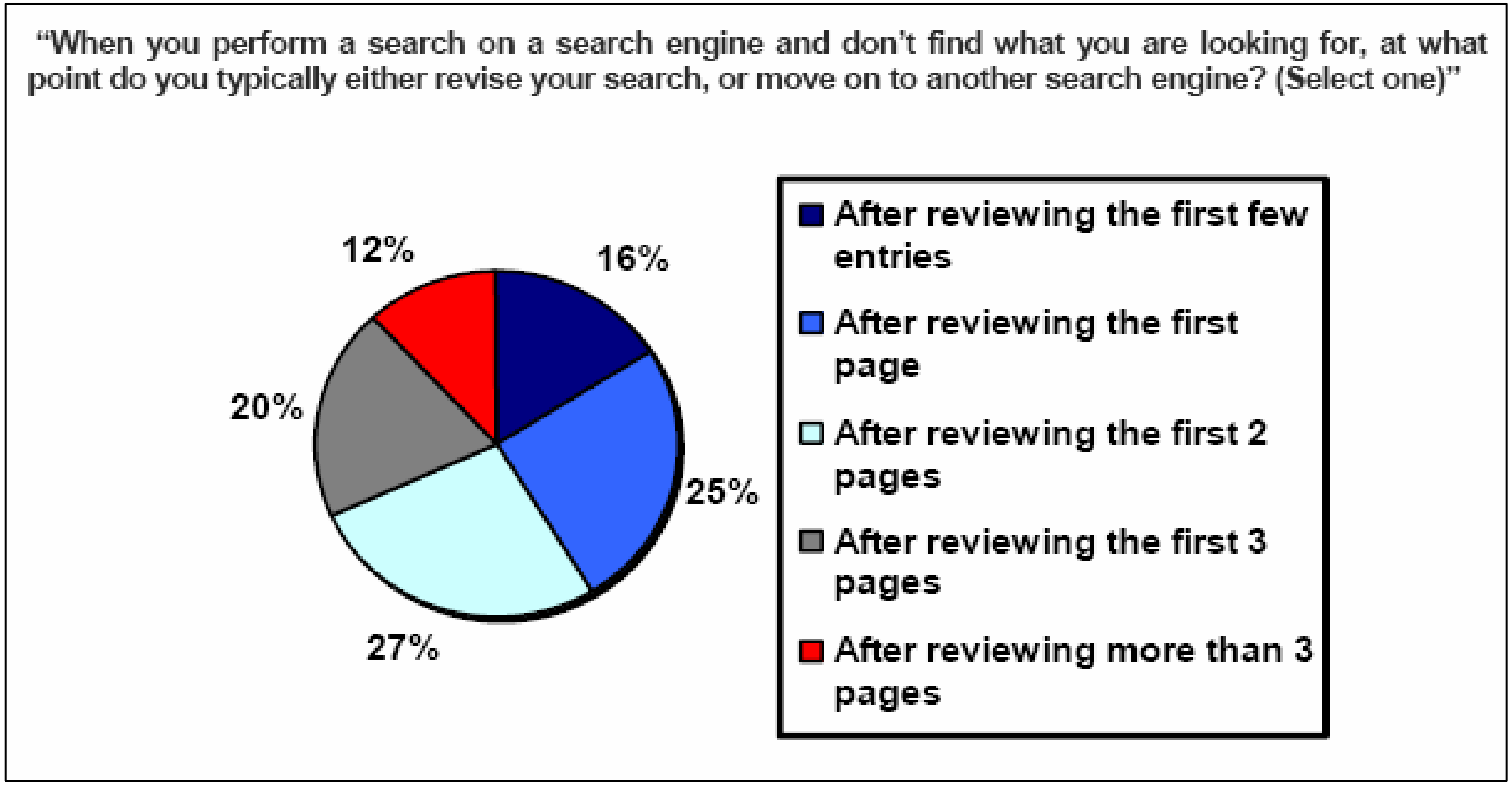
- Query Distribution



  - Power law: few popular broad queries, many rare specific queries

User Query Needs (4/4)

- How far do people look for results?

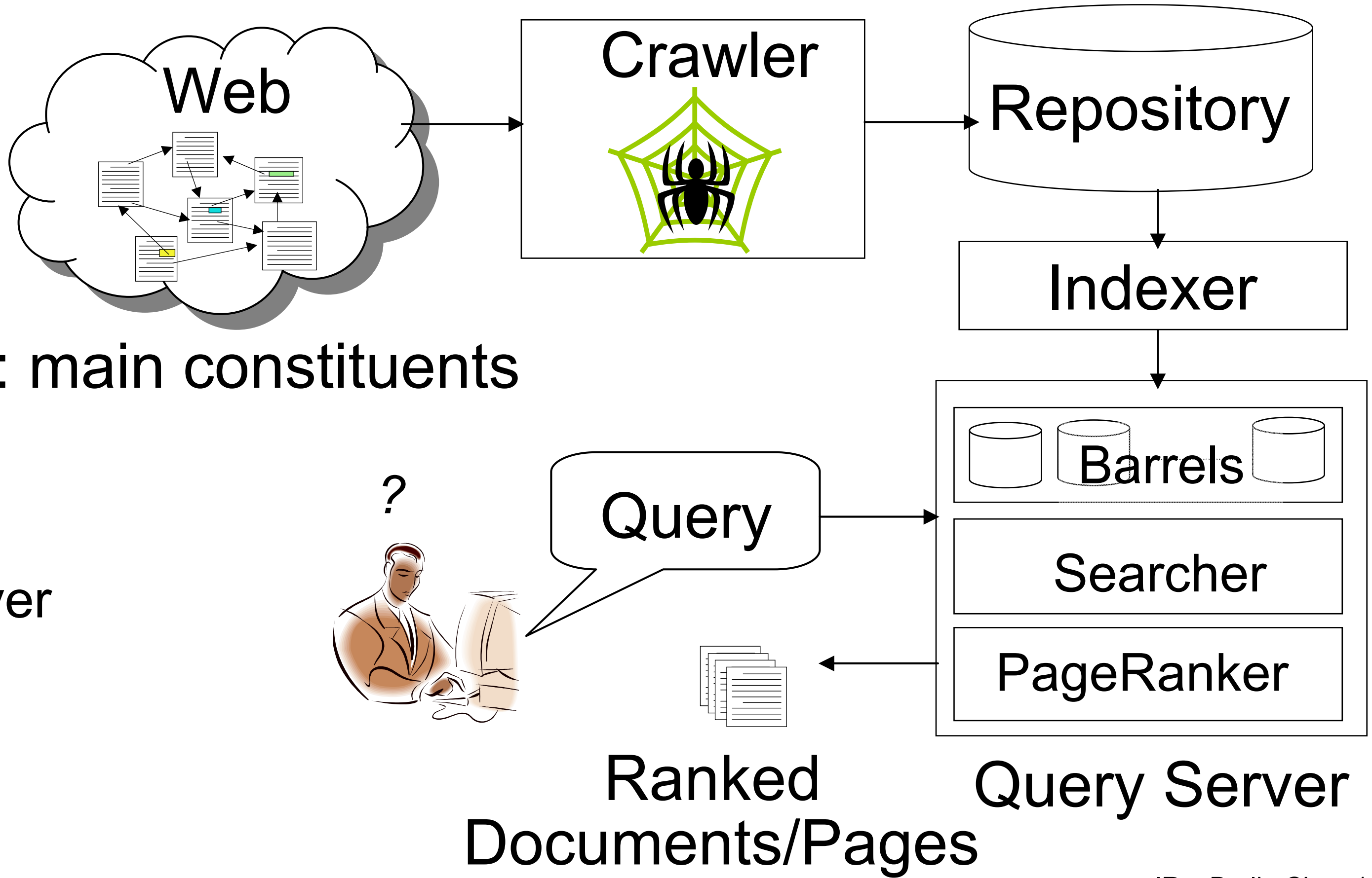


"When you perform a search on a search engine and don't find what you are looking for, at what point do you typically either revise your search, or move on to another search engine? (Select one)"

(Source: iprospect.com WhitePaper_2006_SearchEngineUserBehavior.pdf)

Web Search Engines (1/2)

- Goal
  - Return both high-relevance and high-quality (i.e., valuable) pages
    - Given the heterogeneity of the Web and the ill-formed queries
- Architecture: main constituents
  - Crawler
  - Indexer
  - Query Server

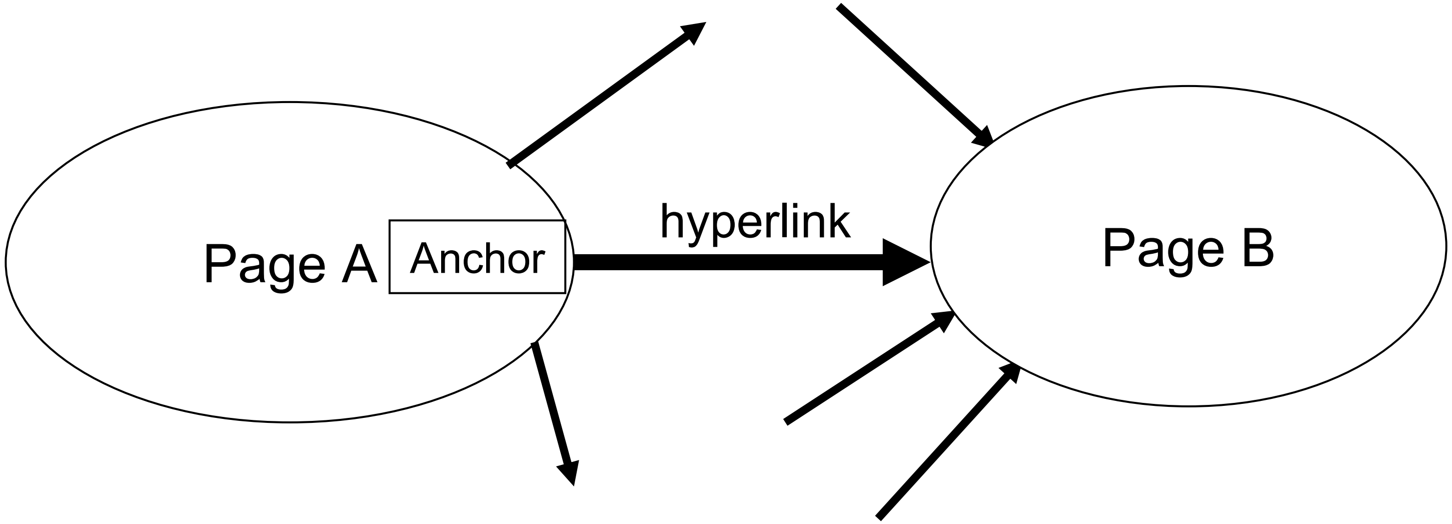


Web Search Engines (2/2)

- Crawler
  - Collect pages from the Web
  - Done by distributed crawlers
    - URL server sends lists of URL to be fetched by crawlers
    - Store server compresses and stores pages (full HTML texts) into a repository
      - Duplicate content detection
- Indexer
  - Process the retrieved pages/documents and represent them in efficient search data structures (inverted files)
- Query server
  - Accept the query from the user and return the result pages by consulting the search data structures

Hyperlink and Anchor Text (1/2)

- Web as a Directed Graph - Two intuitions
  - Hyperlinks from a web page as a form of conferral of authority
    - I.e., A hyperlink between pages denotes author perceived relevance (quality signal)

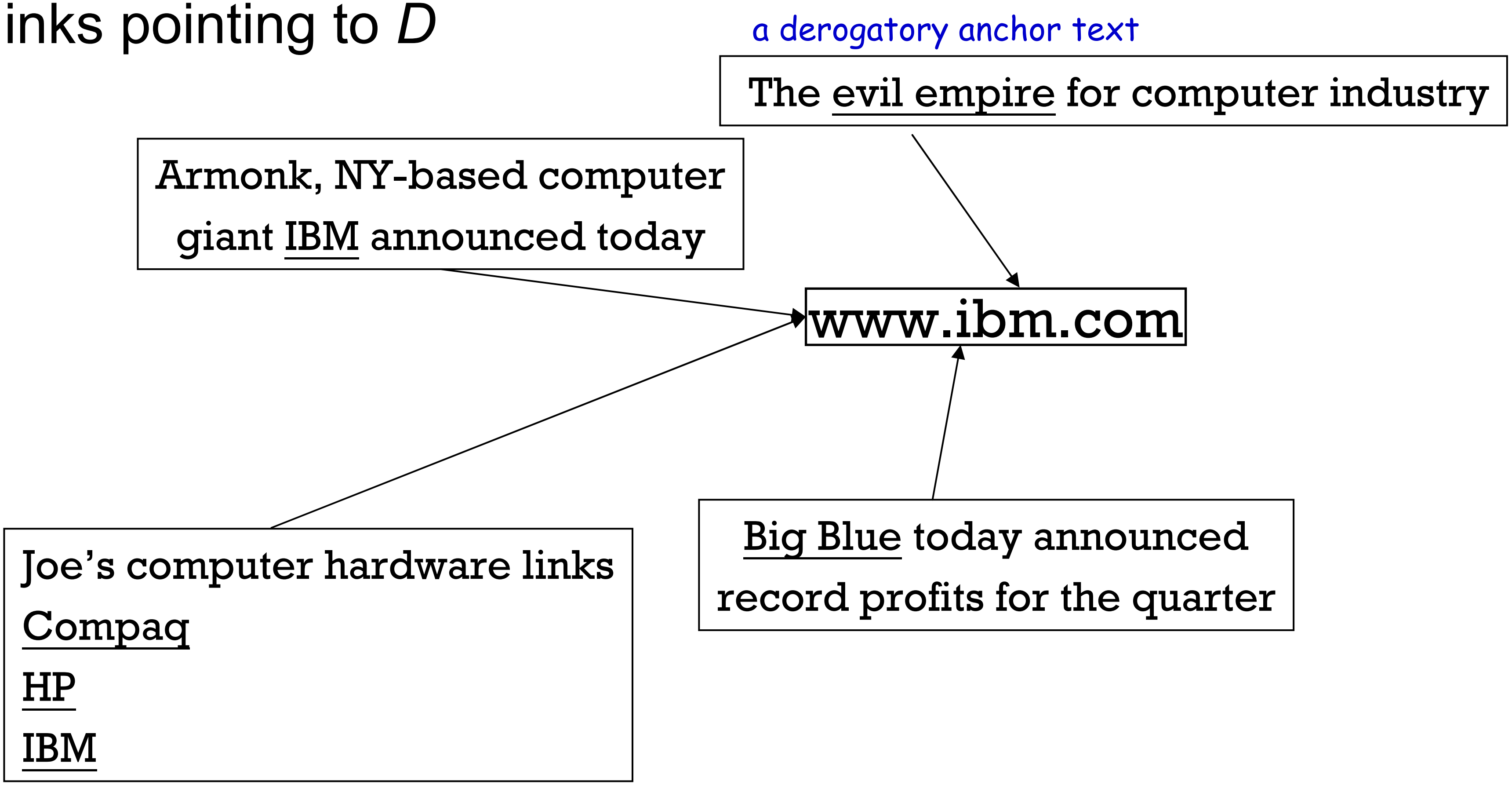


  - The anchor (text) of the hyperlink describes the target page (textual context)
    - A short summary of the target page

<a href="http://www.acm.org./jacm/"> Journal of the ACM </a>

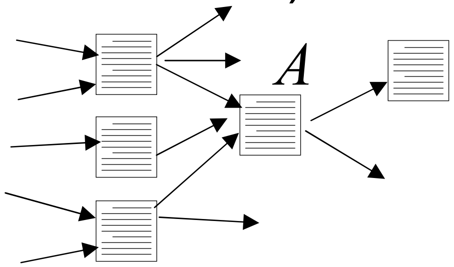
Hyperlink and Anchor Text (2/2)

- When indexing a document *D*, include anchor text from links pointing to *D*


a derogatory anchor text
The evil empire for computer industry
Armonk, NY-based computer giant IBM announced today
www.ibm.com
Joe's computer hardware links
Compaq
HP
IBM
Big Blue today announced record profits for the quarter

PageRank Algorithm

- Proposed by L. Page and Brain, 1998
- Notations
  - A page *A* has pages $T_1 \ldots T_n$ which point to it (citations)
  - *d* range from 0~1, a damping factor (Google sets to be 0.85)
  - C(*A*)：Number of links going out of page A



- PageRank of a page *A*

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

  - PageRank of each page is randomly assigned at the initial iteration and its value tends to be saturated through iterations

- A page with a high PageRank value
  - Many pages pointing to it
  - Or, there are some pages that point to it and have high PageRank values

Business Models for Web Search (1/3)

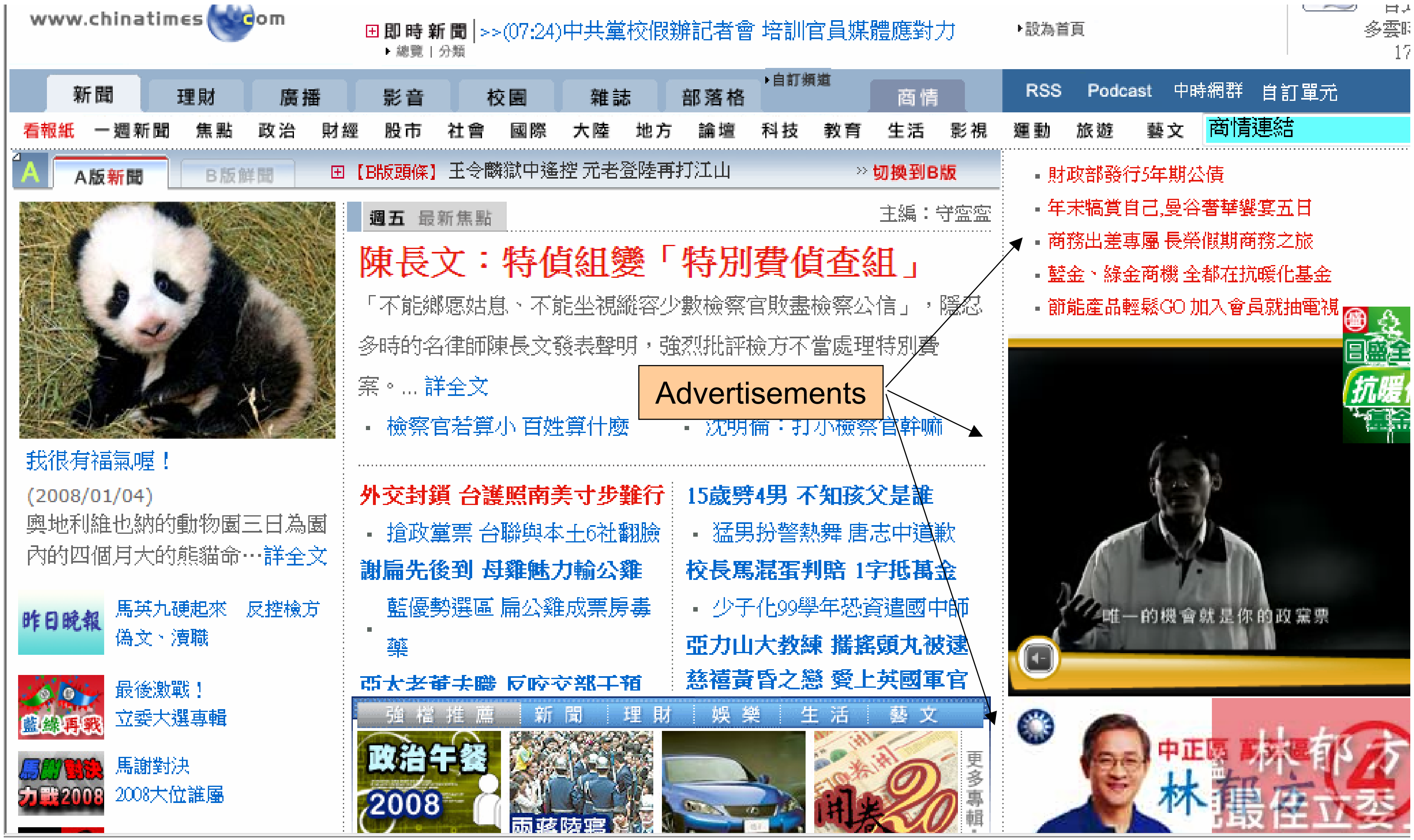
- Advertisers pay for banner ads (advertisements) on the site that do not depend on a user's query
  - CPM: Cost Per Mille (thousand impressions). Pay for each ad display
  - CPC: Cost Per Click. Pay only when user clicks on ad
  - CTR: Click Through Rate. Fraction of ad impressions that result in clicks throughs. CPC = CPM / (CTR * 1000)
  - CPA: Cost Per Action (Acquisition). Pay only when user actually makes a purchase on target site

- Advertisers bid for "keywords". Ads for highest bidders displayed when user query contains a purchased keyword
  - PPC: Pay Per Click. CPC for bid word ads (e.g. Google AdWords)

- This slide is adopted from Raymond J. Mooney's teaching materials

Business Models for Web Search (2/3)

- Paid banner ads (news portal)



⊞ 即時新聞 | >>(07:24)中共黨校假辦記者會 培訓官員媒體應對力
‣ 總覽 | 分類

‣ 設為首頁

多雲時
17

新聞 理財 廣播 影音 校園 雜誌 部落格 ‣ 自訂頻道 商情 RSS Podcast 中時網群 自訂單元

看報紙 一週新聞 焦點 政治 財經 股市 社會 國際 大陸 地方 論壇 科技 教育 生活 影視 運動 旅遊 藝文 商情連結

A A版新聞 B版鮮聞 ⊞【B版頭條】王令麟獄中遙控 元老登陸再打江山 » 切換到B版

我很有福氣喔！

(2008/01/04)

奧地利維也納的動物園三日為園內的四個月大的熊貓命…詳全文

馬英九硬起來　反控檢方偽文、瀆職

最後激戰！立委大選專輯

馬謝對決 2008大位誰屬

週五 最新焦點　主編：守盜盜

陳長文：特偵組變「特別費偵查組」

「不能鄉愿姑息、不能坐視縱容少數檢察官敗盡檢察公信」，隱忍多時的名律師陳長文發表聲明，強烈批評檢方不當處理特別費案。... 詳全文

- 檢察官若算小 百姓算什麼
- 沈明倫：打小檢察官幹嘛

外交封鎖 台護照南美寸步難行

- 搶政黨票 台聯與本土6社翻臉

謝扁先後到 母雞魅力輸公雞

- 藍優勢選區 扁公雞成票房毒藥

亞太老董去職 反嗆交部干預

15歲劈4男 不知孩父是誰

- 猛男扮警熱舞 唐志中道歉

校長罵混蛋判賠 1字抵萬金

- 少子化99學年恐資遣國中師

亞力山大教練 攜搖頭丸被逮

慈禧黃昏之戀 愛上英國軍官

強檔推薦 | 新聞 | 理財 | 娛樂 | 生活 | 藝文

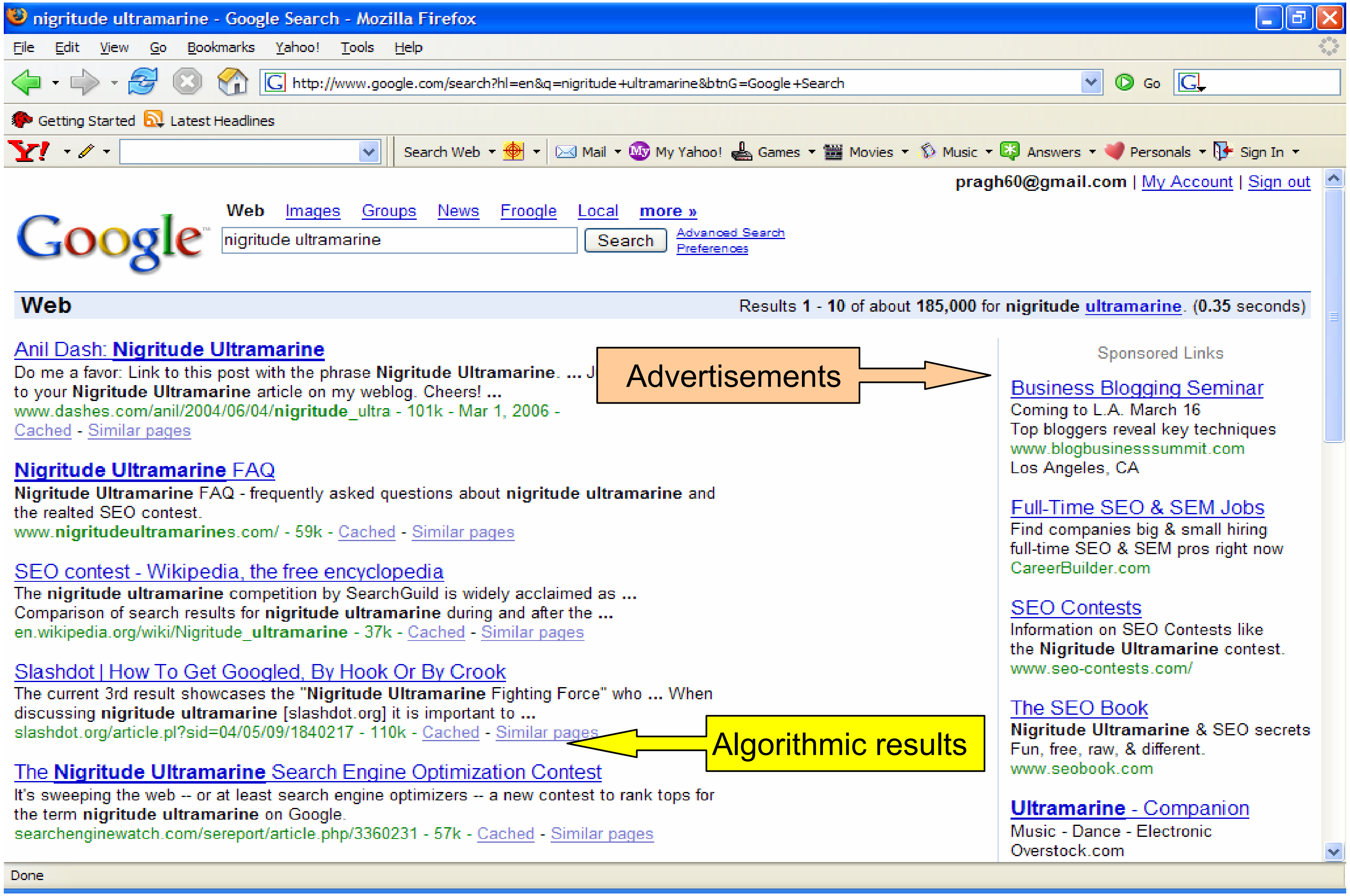
更多專輯

- 財政部發行5年期公債
- 年末犒賞自己,曼谷奢華饗宴五日
- 商務出差專屬 長榮假期商務之旅
- 藍金、綠金商機 全都在抗暖化基金
- 節能產品輕鬆GO 加入會員就抽電視

Advertisements

Business Models for Web Search (3/3)

- Bid keywords (search engine)



nigritude ultramarine - Google Search - Mozilla Firefox
File Edit View Go Bookmarks Yahoo! Tools Help
http://www.google.com/search?hl=en&q=nigritude+ultramarine&btnG=Google+Search
Go
Getting Started
Latest Headlines
Search Web
Mail
My Yahoo!
Games
Movies
Music
Answers
Personals
Sign In
pragh60@gmail.com | My Account | Sign out
Google
Web Images Groups News Froogle Local more »
nigritude ultramarine
Search
Advanced Search
Preferences
Done

Web Results **1** - **10** of about **185,000** for **nigritude ultramarine**. (**0.35** seconds)

Anil Dash: **Nigritude Ultramarine**
Do me a favor: Link to this post with the phrase **Nigritude Ultramarine**. **...** J
to your **Nigritude Ultramarine** article on my weblog. Cheers! **...**
www.dashes.com/anil/2004/06/04/**nigritude**_ultra - 101k - Mar 1, 2006 -
Cached - Similar pages

Nigritude Ultramarine FAQ
Nigritude Ultramarine FAQ - frequently asked questions about **nigritude ultramarine** and
the realted SEO contest.
www.**nigritudeultramarines**.com/ - 59k - Cached - Similar pages

SEO contest - Wikipedia, the free encyclopedia
The **nigritude ultramarine** competition by SearchGuild is widely acclaimed as **...**
Comparison of search results for **nigritude ultramarine** during and after the **...**
en.wikipedia.org/wiki/Nigritude_**ultramarine** - 37k - Cached - Similar pages

Slashdot | How To Get Googled, By Hook Or By Crook
The current 3rd result showcases the "**Nigritude Ultramarine** Fighting Force" who **...** When
discussing **nigritude ultramarine** [slashdot.org] it is important to **...**
slashdot.org/article.pl?sid=04/05/09/1840217 - 110k - Cached - Similar pages

The **Nigritude Ultramarine** Search Engine Optimization Contest
It's sweeping the web -- or at least search engine optimizers -- a new contest to rank tops for
the term **nigritude ultramarine** on Google.
searchenginewatch.com/sereport/article.php/3360231 - 57k - Cached - Similar pages

Sponsored Links

Business Blogging Seminar
Coming to L.A. March 16
Top bloggers reveal key techniques
www.blogbusinesssummit.com
Los Angeles, CA

Full-Time SEO & SEM Jobs
Find companies big & small hiring
full-time SEO & SEM pros right now
CareerBuilder.com

SEO Contests
Information on SEO Contests like
the **Nigritude Ultramarine** contest.
www.seo-contests.com/

The SEO Book
Nigritude Ultramarine & SEO secrets
Fun, free, raw, & different.
www.seobook.com

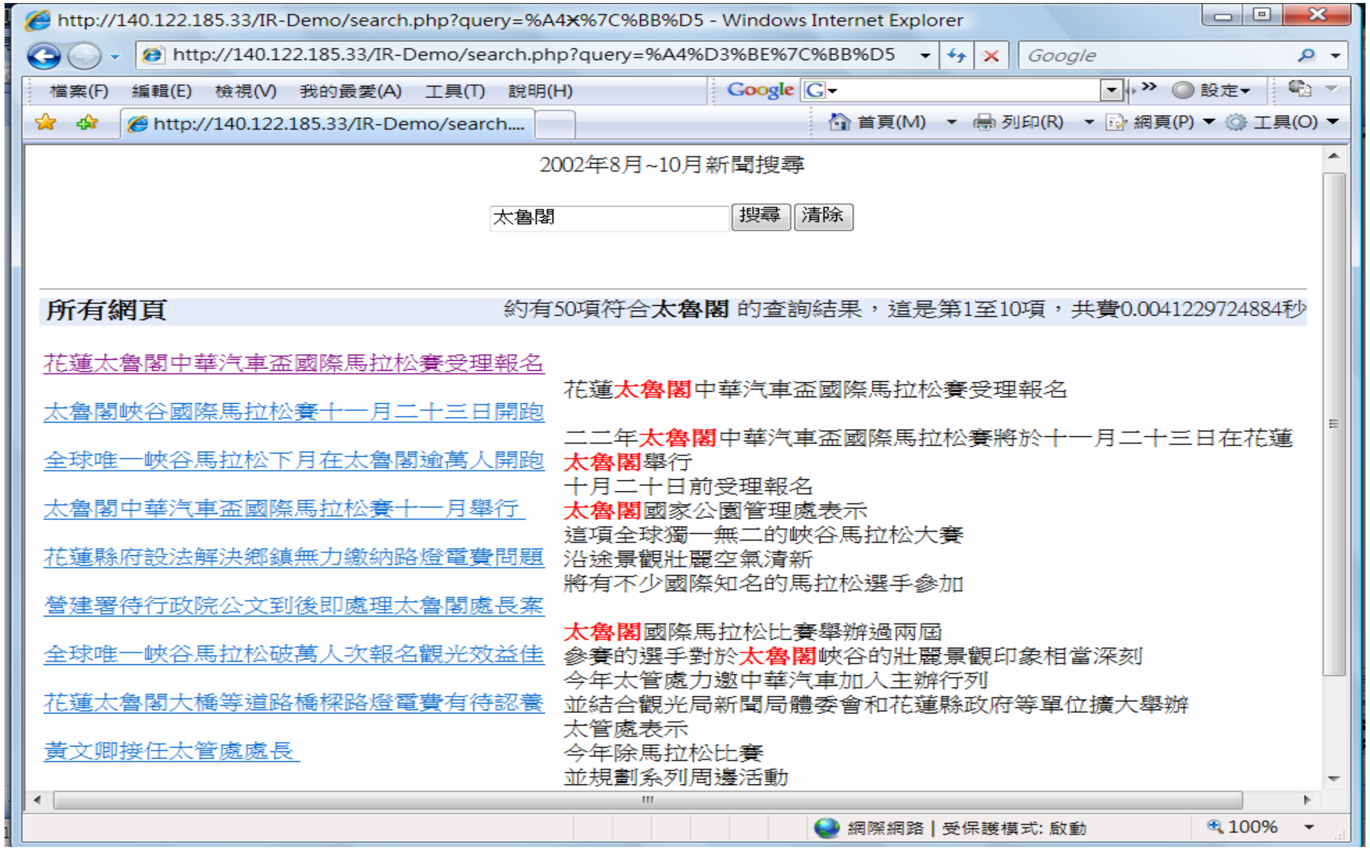
Ultramarine - Companion
Music - Dance - Electronic
Overstock.com

Advertisements

Algorithmic results

Final Project Description

- Reference site: http://140.122.185.33/IR-Demo/


http://140.122.185.33/IR-Demo/search.php?query=%A4x%7C%BB%D5 - Windows Internet Explorer
http://140.122.185.33/IR-Demo/search.php?query=%A4%D3%BE%7C%BB%D5
Google
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
Google G
設定
http://140.122.185.33/IR-Demo/search....
首頁(M) 列印(R) 網頁(P) 工具(O)
網際網路 | 受保護模式: 啟動
100%

2002年8月~10月新聞搜尋

太魯閣 搜尋 清除

所有網頁 約有50項符合**太魯閣** 的查詢結果，這是第1至10項，共費0.0041229724884秒

花蓮太魯閣中華汽車盃國際馬拉松賽受理報名
太魯閣峽谷國際馬拉松賽十一月二十三日開跑
全球唯一峽谷馬拉松下月在太魯閣逾萬人開跑
太魯閣中華汽車盃國際馬拉松賽十一月舉行
花蓮縣府設法解決鄉鎮無力繳納路燈電費問題
營建署待行政院公文到後即處理太魯閣處長案
全球唯一峽谷馬拉松破萬人次報名觀光效益佳
花蓮太魯閣大橋等道路橋樑路燈電費有待認養
黃文卿接任太管處處長

花蓮**太魯閣**中華汽車盃國際馬拉松賽受理報名

二二年**太魯閣**中華汽車盃國際馬拉松賽將於十一月二十三日在花蓮
太魯閣舉行
十月二十日前受理報名
太魯閣國家公園管理處表示
這項全球獨一無二的峽谷馬拉松大賽
沿途景觀壯麗空氣清新
將有不少國際知名的馬拉松選手參加

太魯閣國際馬拉松比賽舉辦過兩屆
參賽的選手對於**太魯閣**峽谷的壯麗景觀印象相當深刻
今年太管處力邀中華汽車加入主辦行列
並結合觀光局新聞局體委會和花蓮縣政府等單位擴大舉辦
太管處表示
今年除馬拉松比賽
並規劃系列周邊活動

- Contact TA for details of Corpus and Internet/Web Application Programs
- Project Due: 25 Jan. 2008