

# Web Search Basics

Berlin Chen

Department of Computer Science & Information Engineering  
National Taiwan Normal University

## References:

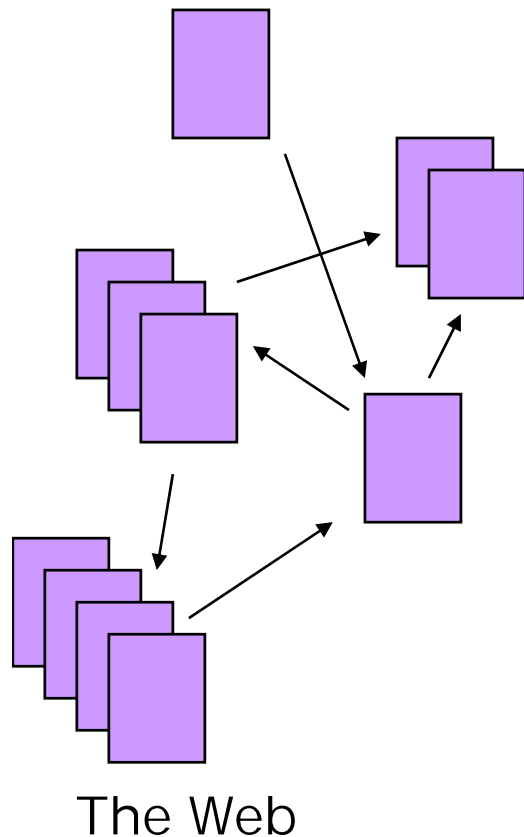
1. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008. (Chapters 19 – 21 & associated slides)
2. Raymond J. Mooney's teaching materials
3. Lan Huang. A Survey on Web Information Retrieval Technologies. Available at: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

<http://www.ntnu.edu.tw/infomation/contact.html>  
domain

# Web Characteristics

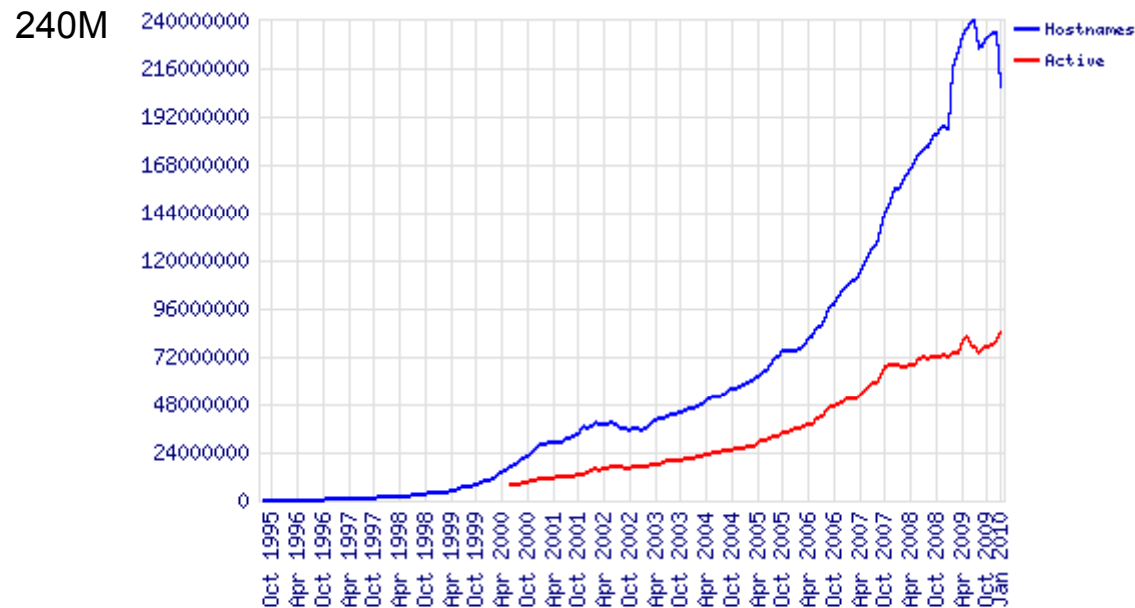


- **No Control**: democratization of creation and linking (publishing). Content includes truth, lies, obsolete information, contradictions
- **Distributed Data**: Documents spread over millions of different web servers...
- **Heterogeneity**: Unstructured (text, html, ...), semi-structured (XML, annotated 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 October 1995 - January 2010 (<http://news.netcraft.com>)

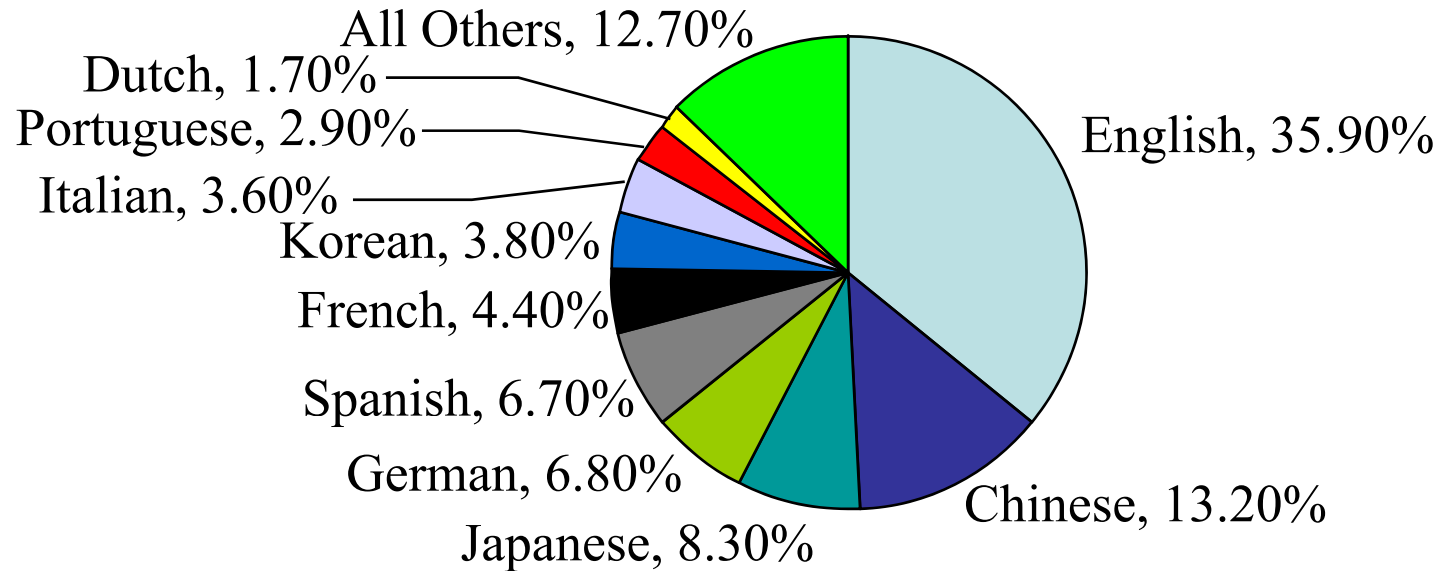
Total Sites Across All Domains August 1995 - January 2010



- 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

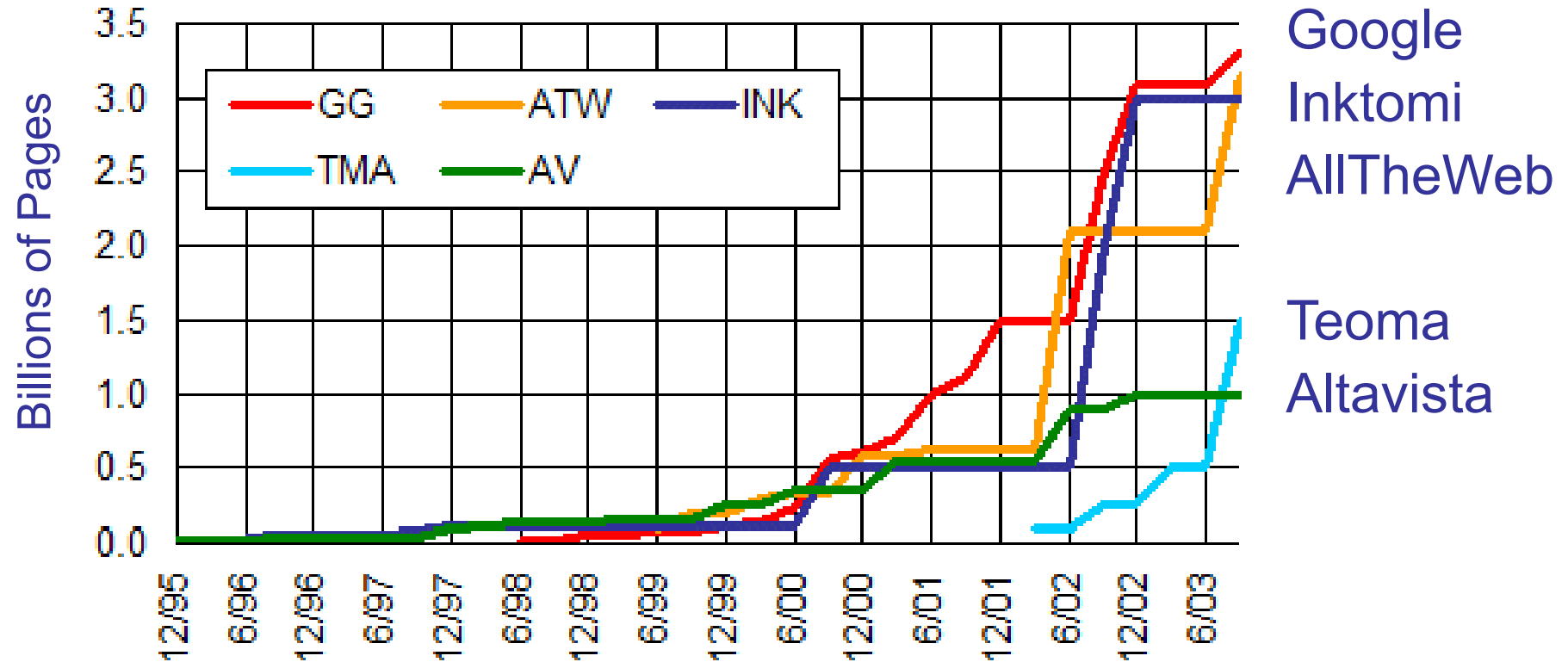
- End of 2004, total 801.4 millions



# 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 (i.e., portal sites)
  - E.g., Yahoo!, Yam, etc.
  - Human editors assemble a large hierarchically structured directory of web pages (entailing significant human effort!)
  - Users browse through trees of category labels ( $\geq 10,00$ )

# Growth of Web Pages Indexed



SearchEngineWatch

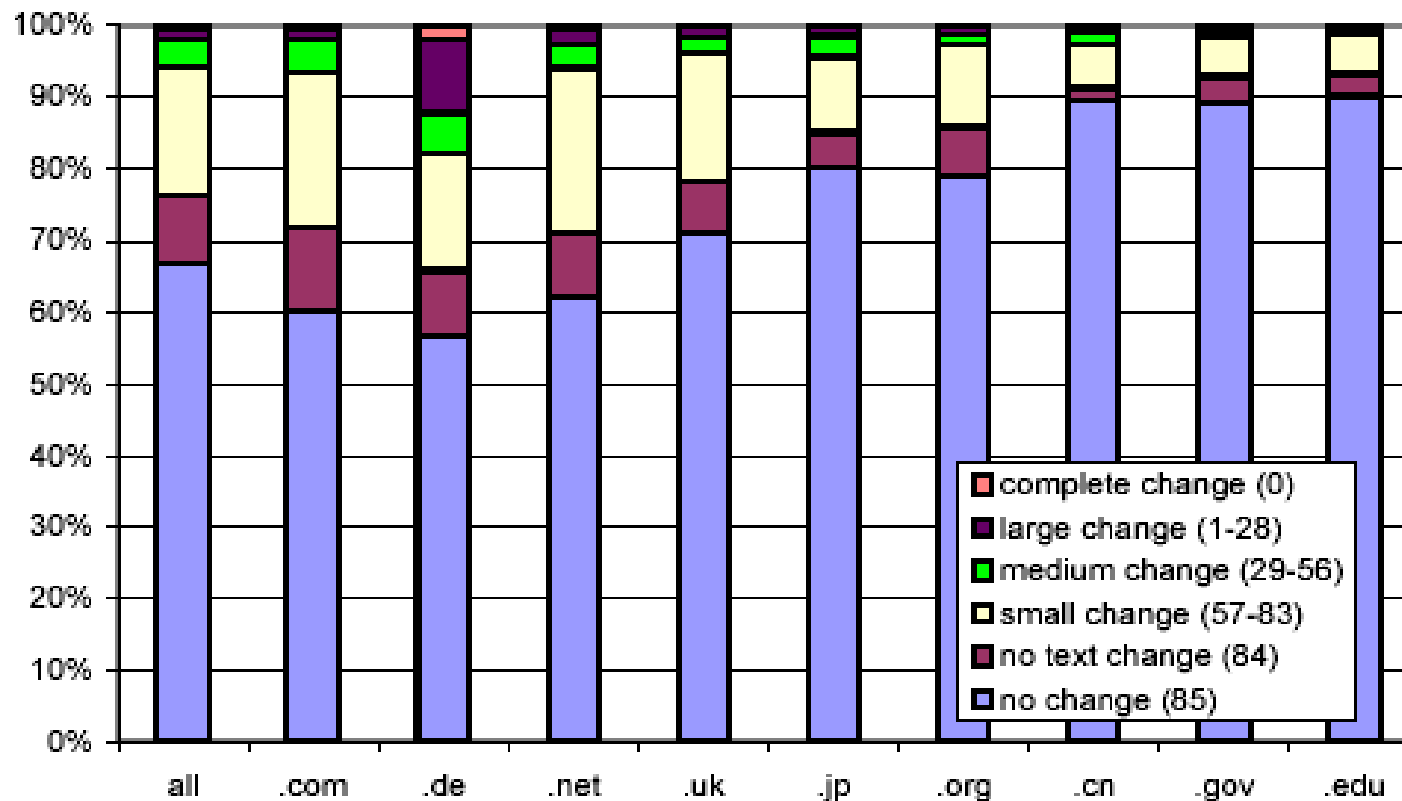
[Link to Note from Jan 2004](#)

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

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

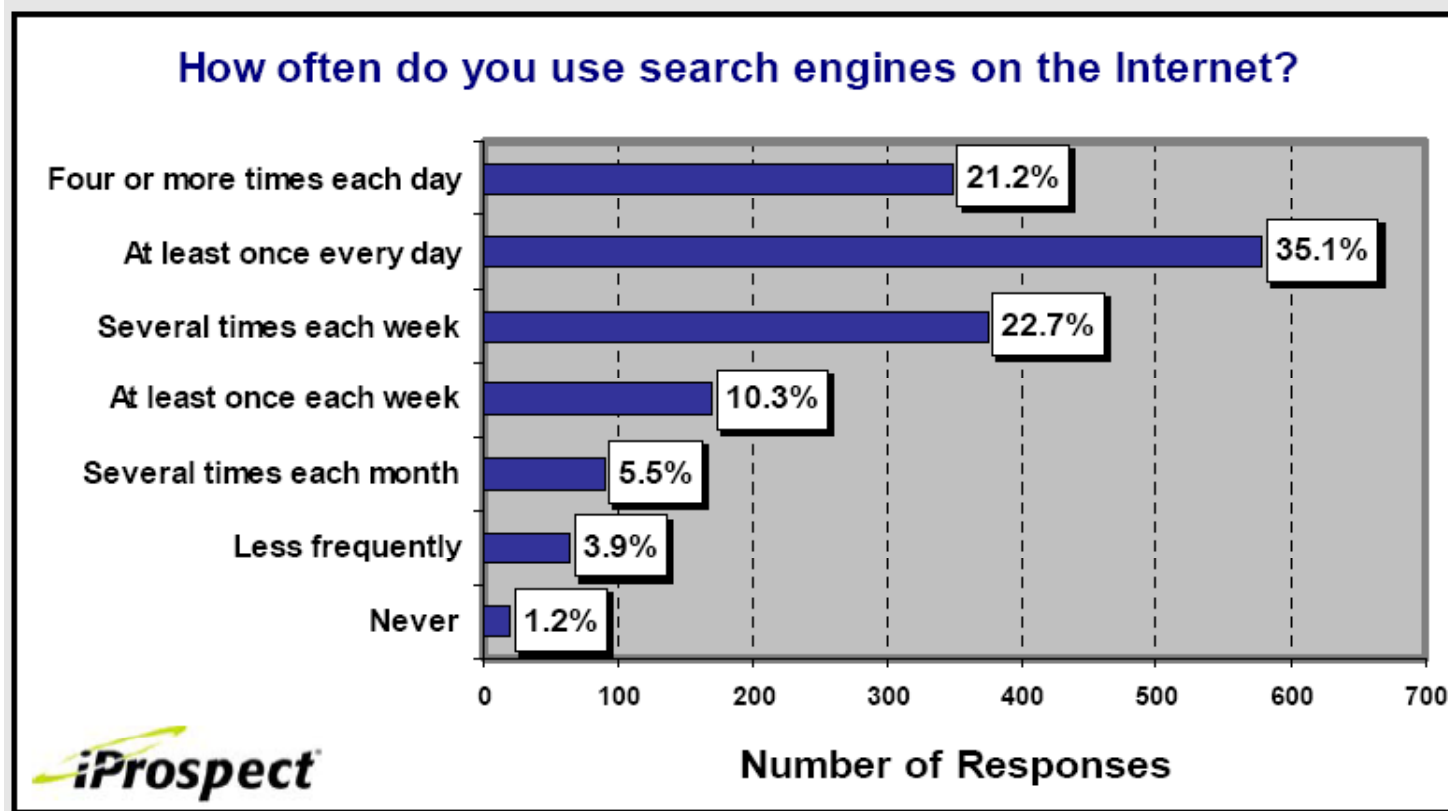
# Rate of Change for Web Pages

- 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

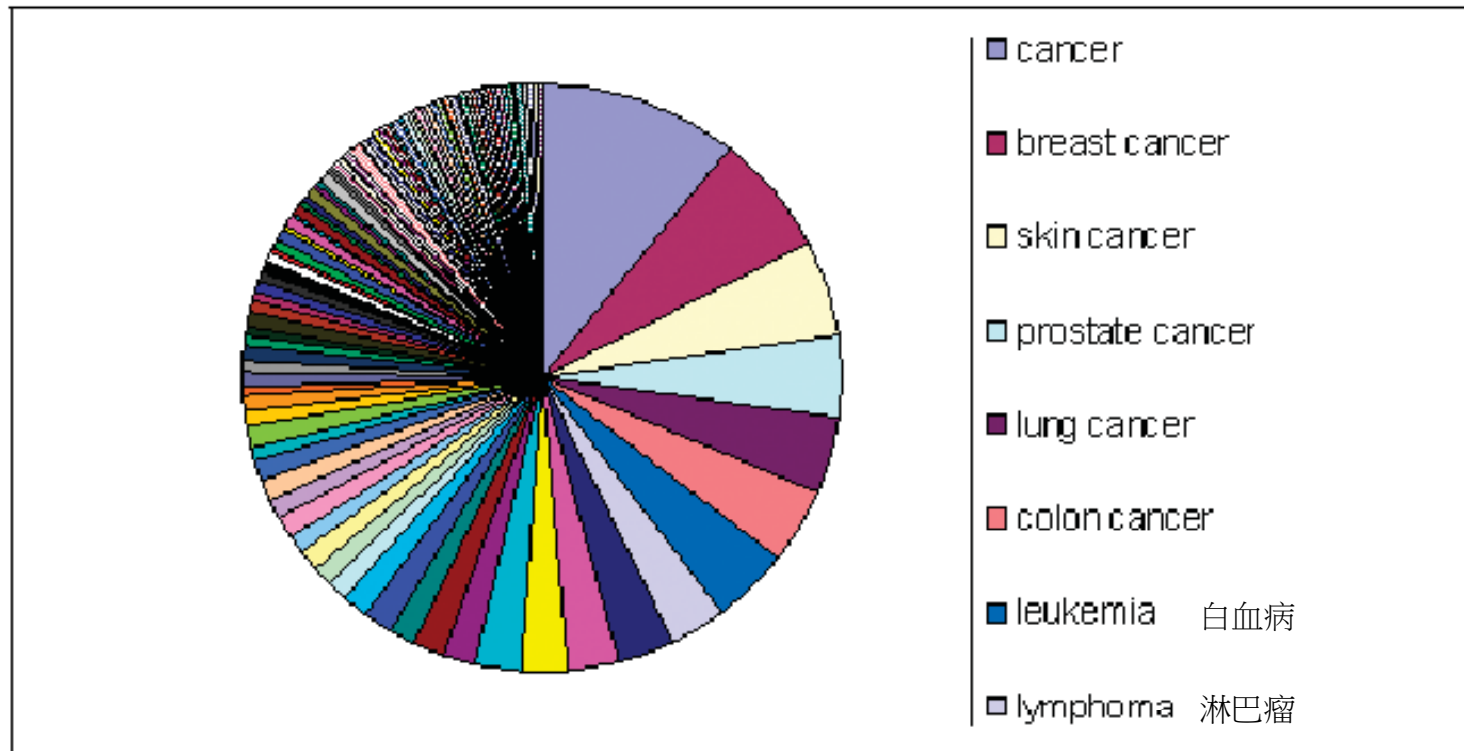
Discern which of these categories a query falls into can be challenging !

# 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)

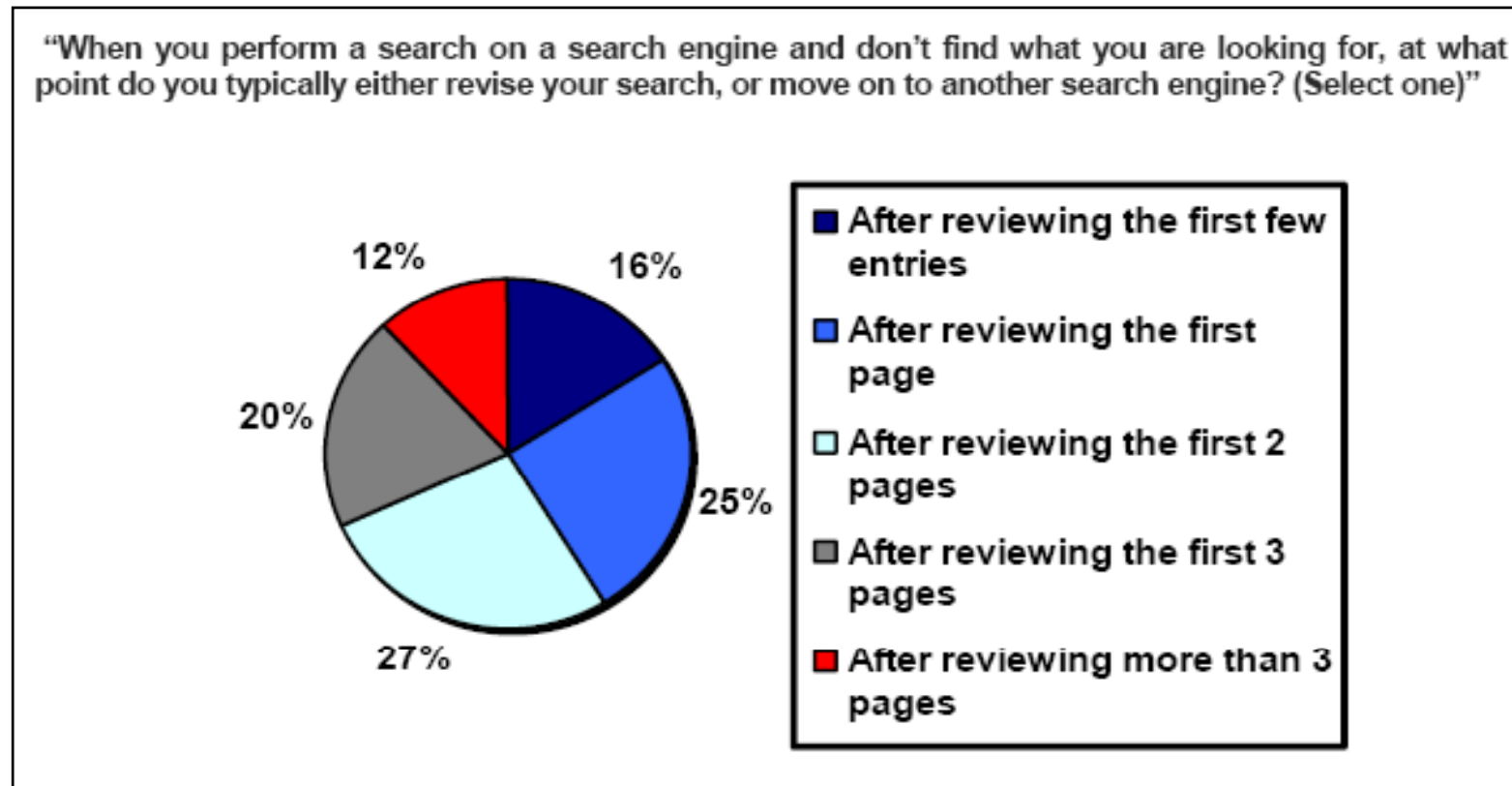
- Query Distribution



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

# User Query Needs (4/4)

- How far do people look for results?

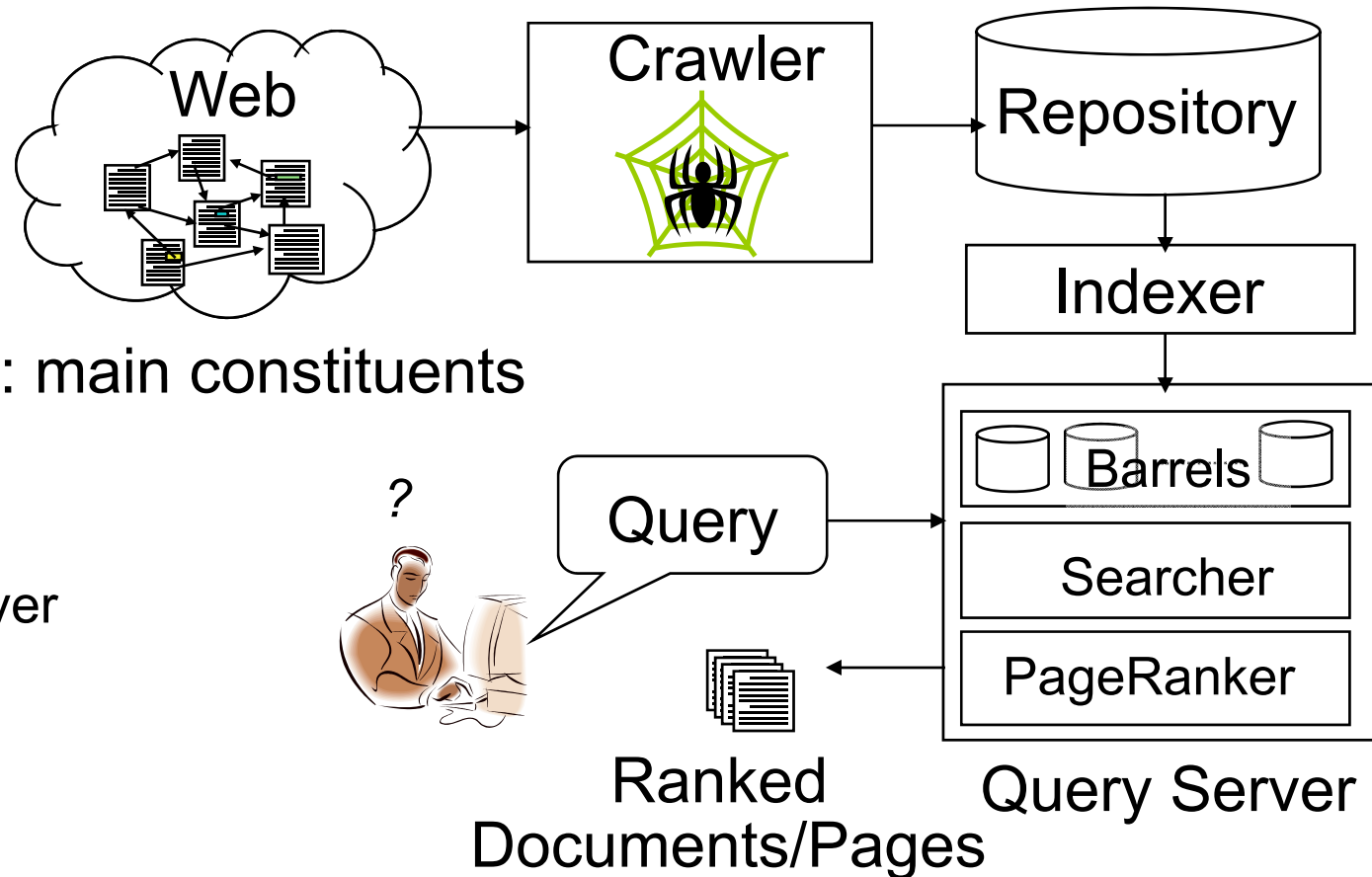


(Source: [iprospect.com](http://iprospect.com) WhitePaper\_2006\_SearchEngineUserBehavior.pdf)

# 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

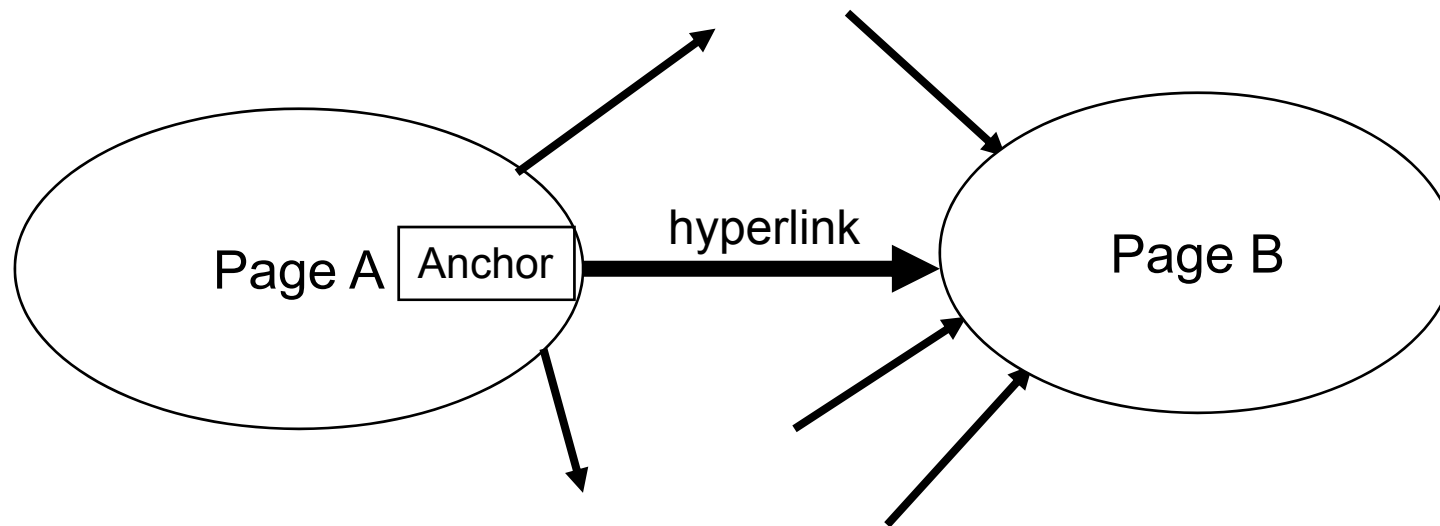
- Crawler
- Indexer
- Query Server

# 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/ posting 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)



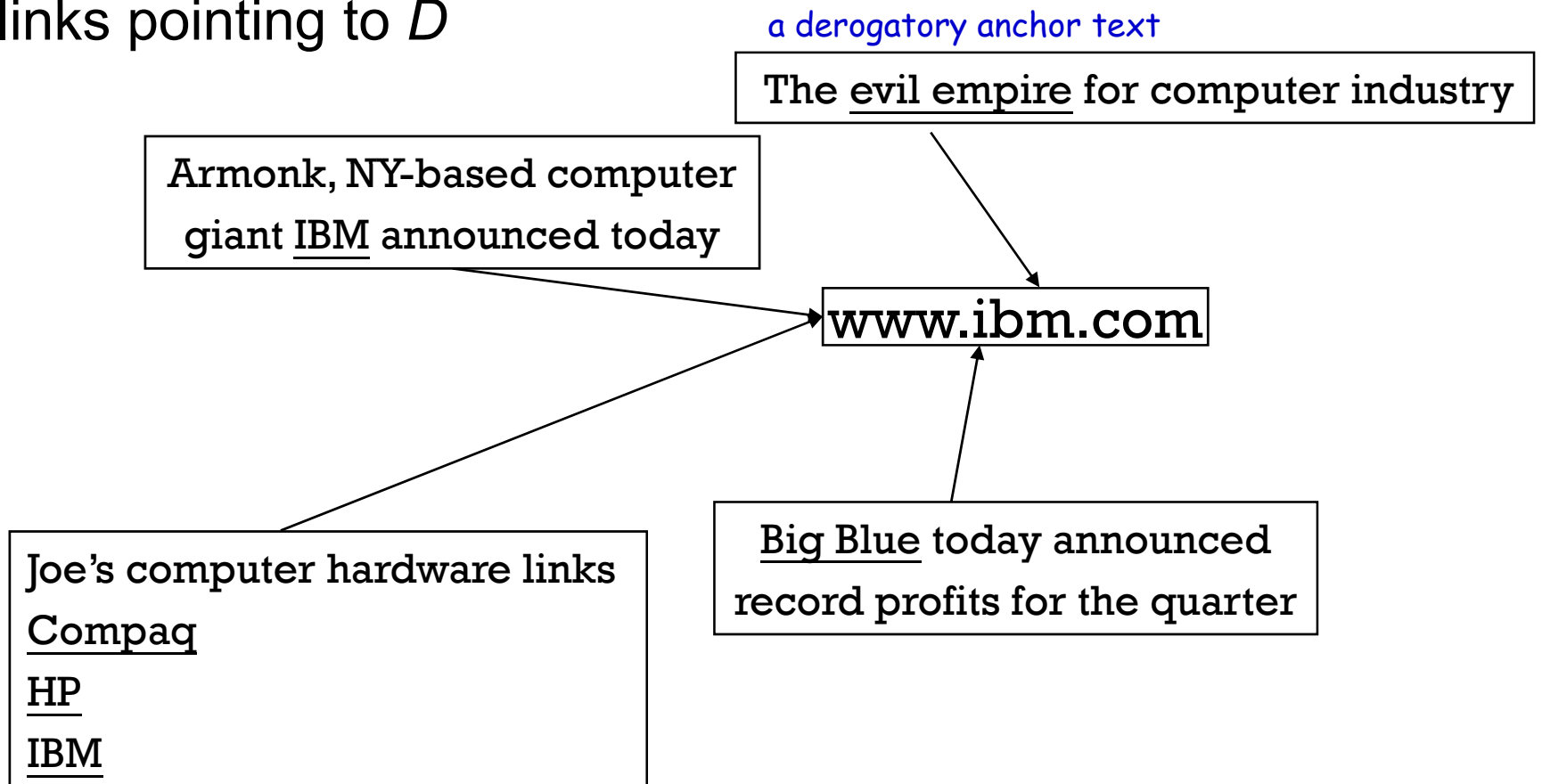
- The **anchor (text)** of the hyperlink describes the target page (textual context)
  - A short summary of the target page

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



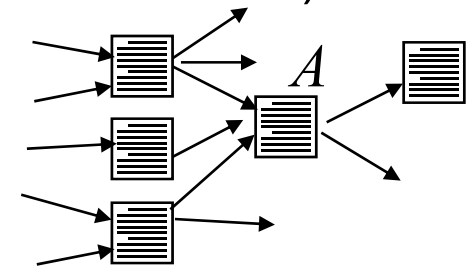
# Hyperlink and Anchor Text (2/2)

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



# PageRank Algorithm

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



- PageRank of a page  $A$

$$PR(A) = \underbrace{(1 - d)}_{\text{contributed by "teleport" operation}} + d \left( \frac{PR(T_1)}{C(T_1)} + \dots + \frac{PR(T_n)}{C(T_n)} \right)$$

- 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

# Spam

- Span (in the context of web search) is the manipulation of web page content for the purpose of appearing high up in search results for selected keywords
  - “paid inclusion” (or search engine marketing, SEM) vs. “search engine optimizers (SEOs)”
- Spam has sprung up a research subarea of the so-called “adversarial information retrieval”
- “link analysis” – the exploitation of the link structure of the Web – somehow can help to mitigate the problems caused by spam

# 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)

The screenshot shows the www.chinatimes.com news portal. At the top, there is a navigation bar with categories like 新聞, 理財, 廣播, 影音, 校園, 雜誌, 部落格, 商情, RSS, Podcast, 中時網群, and 自訂單元. Below this is a secondary navigation bar with sub-categories like 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, and 商情連結. The main content area is divided into several sections. On the left, there is a large image of a panda with the headline "我很有福氣喔!". Below it, there are several smaller news items with headlines like "昨日晚報" and "最後激戰!". In the center, there is a large article titled "陳長文：特偵組變「特別費偵查組」" with a sub-headline "「不能鄉愿姑息、不能坐視縱容少數檢察官敗盡檢察公信」". To the right of this article, there is a box labeled "Advertisements" with arrows pointing to various ad spots. Below the main article, there are several smaller news items with headlines like "外交封鎖 台護照南美寸步難行" and "15歲劈4男 不知孩父是誰". At the bottom, there is a banner for "政治午餐 2008" and another for "中正區 林郁方 最佳立委".

www.chinatimes.com

即時新聞 | >>(07:24)中共黨校假辦記者會 培訓官員媒體應對力

設為首頁

多雲 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版新聞 【B版頭條】 王令麟獄中遙控 元老登陸再打江山 >> 切換到B版

週五 最新焦點 主編：守靈室

**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  
藍優勢選區 扁公雞成票房毒藥  
亞太老番夫職 反時交部干預

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

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

昨日晚報 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 馬謝對決 2008六位誰屬

強檔推薦 新聞 理財 娛樂 生活 藝文

政治午餐 2008 兩蔣陵寢

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

唯一的機會就是你的政黨票

中正區 林郁方 最佳立委

# Business Models for Web Search (3/3)

- Bid keywords (search engine)

The screenshot shows a Mozilla Firefox browser window with the Google search engine. The search query is "nigritude ultramarine". The results are divided into two main sections:

- Algorithmic results (left):** These are the organic search results. The top result is from Anil Dash's blog, "Nigritude Ultramarine". Other results include a FAQ page, a Wikipedia entry about an SEO contest, and a Slashdot article about getting googled.
- Advertisements (right):** These are sponsored links. The top ad is for a "Business Blogging Seminar" in Los Angeles. Other ads include "Full-Time SEO & SEM Jobs" and "The SEO Book".

Annotations on the image:

- An orange box labeled "Advertisements" has an arrow pointing to the sponsored links section.
- A yellow box labeled "Algorithmic results" has an arrow pointing to the organic search results section.