PURDUE UNIVERSITY | Department of Computer Science

CS47300: Web Information Search and Management

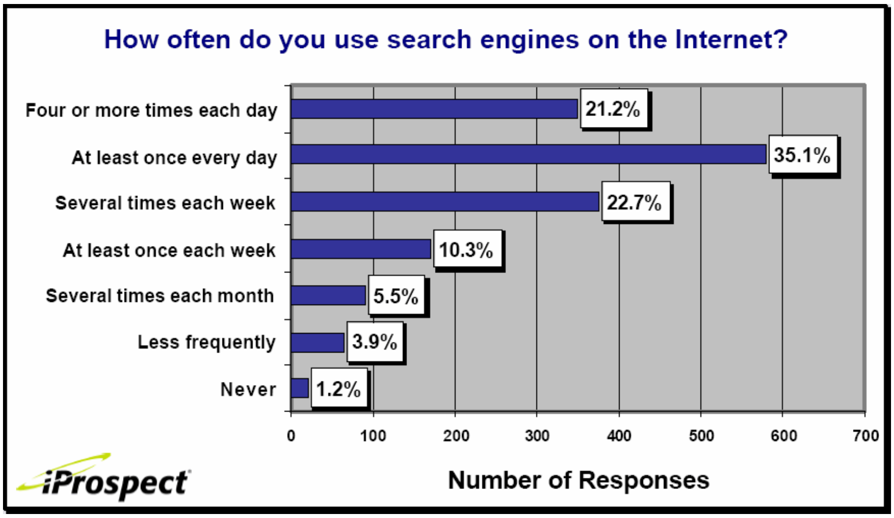
Web Search
Prof. Chris Clifton
14 September 2020
Some slides courtesy
Manning, Raghavan, and Schütze

Indiana Center for Database Systems ™

PURDUE UNIVERSITY
Department of Computer Science

Usage of Web Search

(iProspect Survey, 4/04, http://www.iprospect.com/premiumPDFs/iProspectSurveyComplete.pdf)



Without search engines the web wouldn’t scale

- No incentive in creating content unless it can be easily found – other finding methods haven’t kept pace (taxonomies, bookmarks, etc)
- The web is both a technology artifact and a social environment
  - “The Web has become the “new normal” in the American way of life; those who don’t go online constitute an ever-shrinking minority.” – [Pew Foundation report, January 2005]
- Search engines make aggregation of interest possible:
  - Create incentives for very specialized niche players
    - Economical – specialized stores, providers, etc
    - Social – narrow interests, specialized communities, etc

Without search engines the web wouldn’t scale

- The acceptance of search interaction makes “unlimited selection” stores possible:
  - Amazon, Netflix, etc
- Search has been the best mechanism for advertising on the web, a $15+ B industry.
  - Growing very fast but entire US advertising industry $250B – huge room to grow
  - Sponsored search marketing is about $10B
  - *2020: Statista estimates search ad revenue $159B, 200B by 2024*
  - *2019: Alphabet alone had advertising revenue $142B*

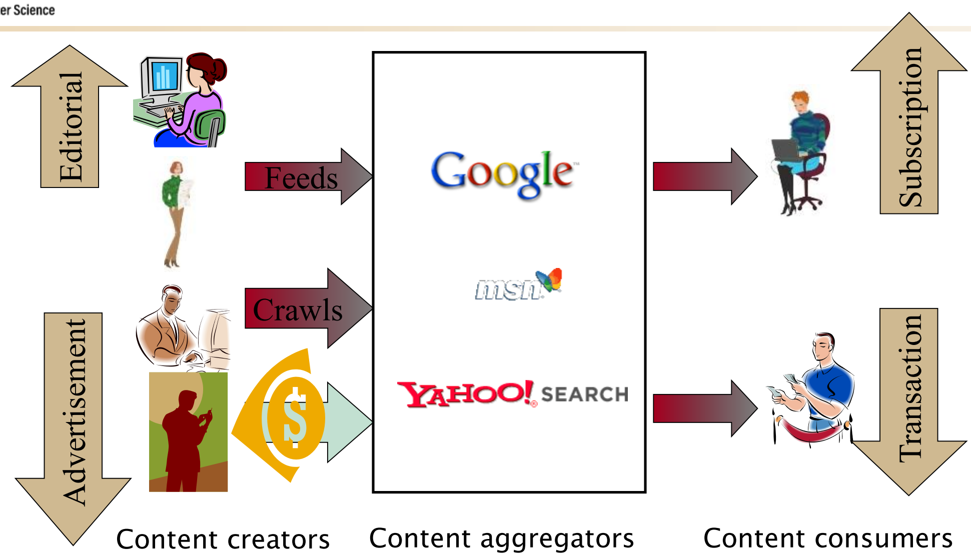
Classic IR

Relevance

- For each query Q and stored document D in a given corpus assume there exists relevance Score(Q, D)
  - Score is average over users U and contexts C
- Optimize Score(Q, D) as opposed to Score(Q, D, U, C)
- That is, usually:
  - Context ignored
  - Individuals ignored
  - Corpus predetermined

Bad assumptions in the web context

The coarse-level dynamics



Brief (non-technical) history

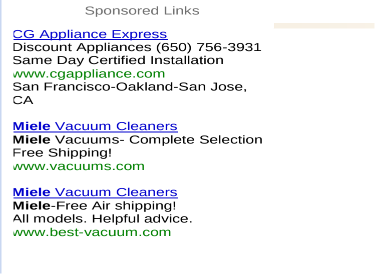
- Early keyword-based engines
  - – Altavista, Excite, Infoseek, Inktomi, ca. 1995-1997
- Paid placement ranking: Goto.com (morphed into Overture.com → Yahoo!)
  - – Your search ranking depended on how much you paid
  - – Auction for keywords: ***casino*** was expensive!

Brief (non-technical) history

- 1998+: Link-based ranking pioneered by Google
  - – Blew away all early engines: Great user experience in search of a business model
  - – Meanwhile Goto/Overture’s annual revenues were nearing $1 billion
- Result: Google added paid-placement “ads” to the side, independent of search results
  - – Yahoo follows suit, acquiring Overture (for paid placement) and Inktomi (for search)

Ads vs. search results

Google has maintained that ads (based on vendors bidding for keywords) do not affect vendors' rankings in search results



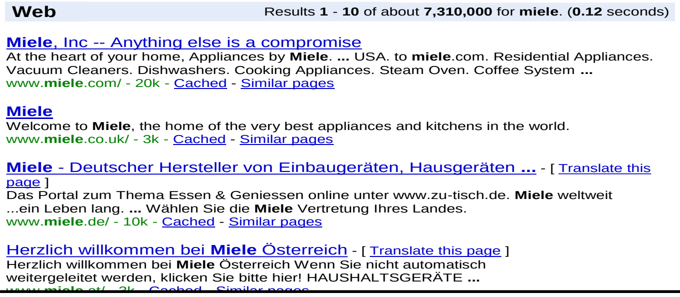
Sponsored Links

CG Appliance Express
Discount Appliances (650) 756-3931
Same Day Certified Installation
www.cgappliance.com
San Francisco-Oakland-San Jose, CA

Miele Vacuum Cleaners
Miele Vacuums- Complete Selection
Free Shipping!
www.vacuums.com

Miele Vacuum Cleaners
Miele-Free Air shipping!
All models. Helpful advice.
www.best-vacuum.com

Search = *miele*



Web Results **1** - **10** of about **7,310,000** for **miele**. (**0.12** seconds)

Miele, Inc -- Anything else is a compromise
At the heart of your home, Appliances by **Miele**. **...** USA. to **miele**.com. Residential Appliances. Vacuum Cleaners. Dishwashers. Cooking Appliances. Steam Oven. Coffee System **...**
www.**miele**.com/ - 20k - Cached - Similar pages

Miele
Welcome to **Miele**, the home of the very best appliances and kitchens in the world.
www.**miele**.co.uk/ - 3k - Cached - Similar pages

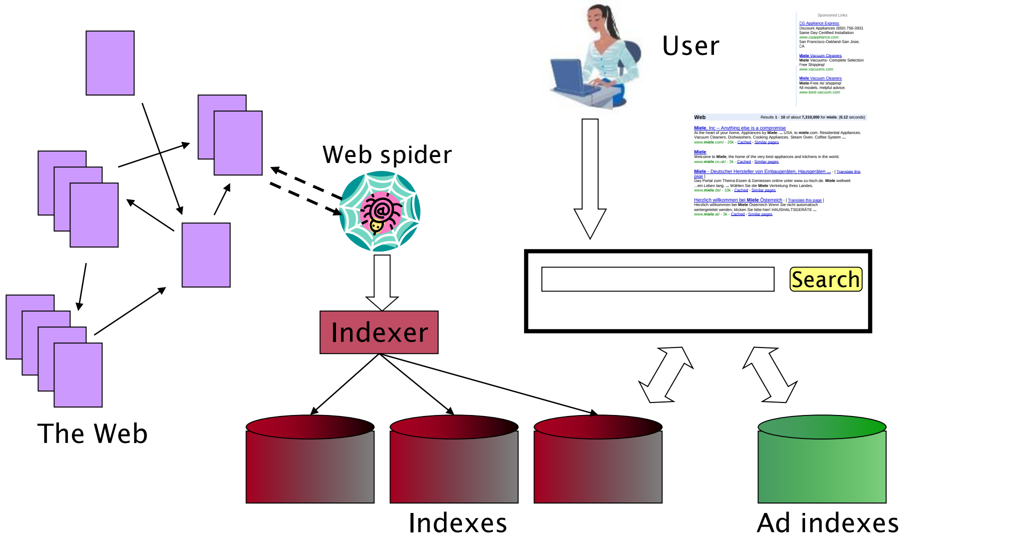
Miele - Deutscher Hersteller von Einbaugeräten, Hausgeräten **...** - [Translate this page]
Das Portal zum Thema Essen & Geniessen online unter www.zu-tisch.de. **Miele** weltweit ...ein Leben lang. **...** Wählen Sie die **Miele** Vertretung Ihres Landes.
www.**miele**.de/ - 10k - Cached - Similar pages

Herzlich willkommen bei **Miele** Österreich - [Translate this page]
Herzlich willkommen bei **Miele** Österreich Wenn Sie nicht automatisch weitergeleitet werden, klicken Sie bitte hier! HAUSHALTSGERÄTE **...**
www.**miele**.at/ - 3k - Cached - Similar pages

Ads vs. search results

- Other vendors (Yahoo, MSN) have made similar statements from time to time
  - Any of them can change anytime
- We will focus primarily on search results independent of paid placement ads
  - Although the latter is a fascinating technical subject in itself

Web search basics



User Needs

Need [Brod02, RL04]

- **Informational** – want to learn about something (~40% / 65%)
  `P53 Cancer`
- **Navigational** – want to go to that page (~25% / 15%)
  `United Airlines`
- **Transactional** – want to do something (web-mediated) (~35% / 20%)
  - Access a service
  - Downloads
  - Shop

  `Seattle weather`
  `Mars surface images`
  `Canon S410`
- Gray areas
  - Find a good hub
  - Exploratory search "see what's there"

  `Car rental Brasil`

PURDUE UNIVERSITY
Department of Computer Science

Web search users

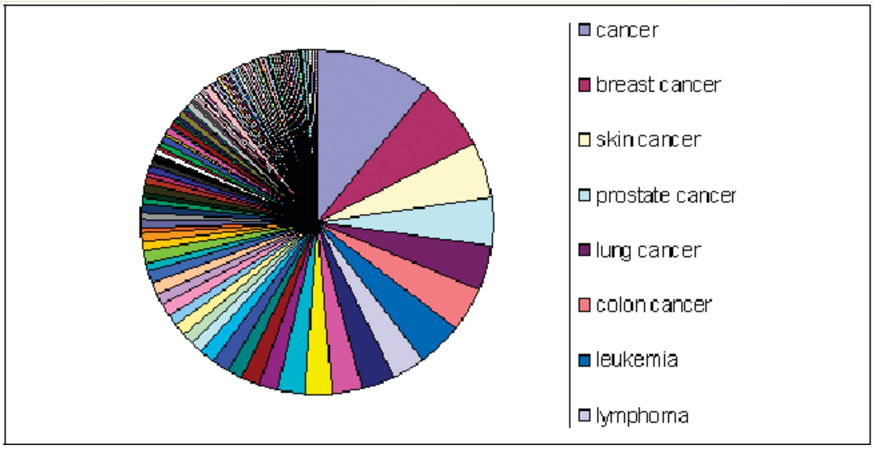
- Make ill defined queries
  - Short
    - AV 2001: 2.54 terms avg, 80% < 3 words)
    - AV 1998: 2.35 terms avg, 88% < 3 words [Silv98]
  - Imprecise terms
  - Sub-optimal syntax (most queries without operator)
  - Low effort
- Wide variance in
  - Needs
  - Expectations
  - Knowledge
  - Bandwidth

- Specific behavior
  - 85% look over one result screen only
  - 78% of queries are not modified (one query/session)
  - Follow links – "the scent of information"

  ...

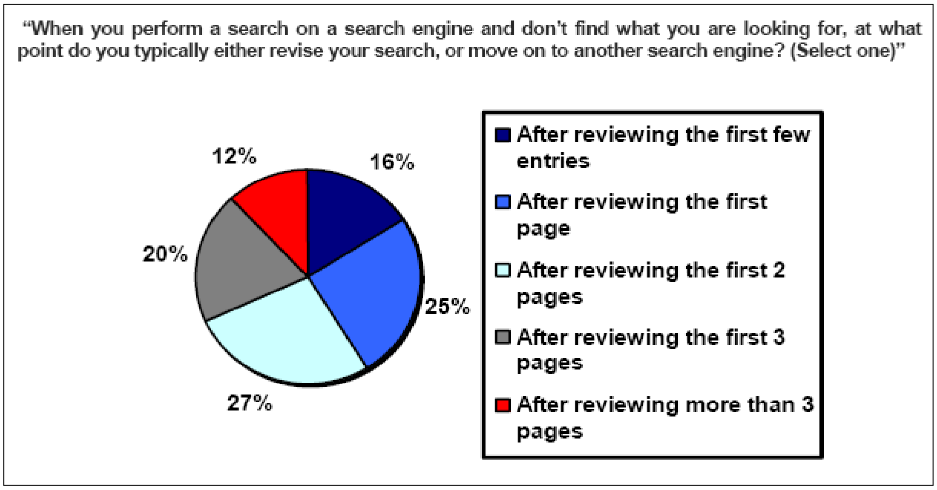
PURDUE UNIVERSITY
Department of Computer Science

Query Distribution



Power law: few popular (typically broad) queries,
many rare (typically more specific) queries

How far do people look for results?



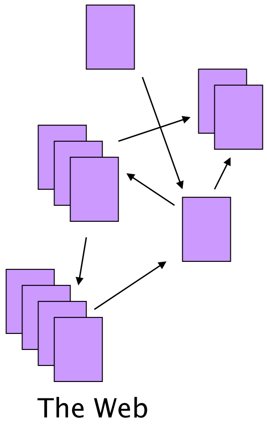
"When you perform a search on a search engine and don't find what you are looking for, at what point do you typically either revise your search, or move on to another search engine? (Select one)"

(Source: iprospect.com WhitePaper_2006_SearchEngineUserBehavior.pdf)

Users' empirical evaluation of results

- Quality of pages varies widely
  - Relevance is not enough
  - Other desirable qualities (non IR!!)
    - Content: Trustworthy, new info, non-duplicates, well maintained,
    - Web readability: display correctly & fast
    - No annoyances: pop-ups, etc
- Precision vs. recall
  - On the web, recall seldom matters

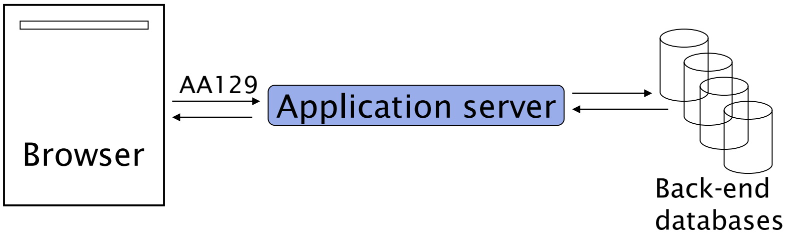
The Web corpus



- No design/co-ordination
- Distributed content creation, linking, democratization of publishing
- Content includes truth, lies, obsolete information, contradictions …
- Unstructured (text, html, …), semi-structured (XML, annotated photos), structured (Databases)…
- Scale much larger than previous text corpora … but corporate records are catching up.
- Content can be dynamically generated

The Web: Dynamic content

- A page without a static html version
  - E.g., current status of flight AA129
  - Current availability of rooms at a hotel
- Usually, assembled at the time of a request from a browser
  - Typically, URL has a '?' character in it



Dynamic content

- Most dynamic content is ignored by web spiders
  - Many reasons including malicious spider traps
- Some dynamic content (news stories from subscriptions) are sometimes delivered as dynamic content
  - Application-specific spidering
- Spiders commonly view web pages just as Lynx (a text browser) would
- Note: even “static” pages are typically assembled on the fly (e.g., headers are common)

Other characteristics of the Web

- Significant duplication
  - Syntactic – 30%-40% (near) duplicates [Brod97, Shiv99b, etc.]
  - Semantic – ???
- High linkage
  - More than 8 links/page in the average
- Complex graph topology
  - Not a small world; bow-tie structure [Brod00]
- Spam
  - Billions of pages