Acknowledgment: Kiduk Yang (2006); Manning, C.D., Raghavan, P. and Schütze, H., *Introduction to Information Retrieval, Cambridge University Press. 2008.*

Information Retrieval Systems

Information Retrieval on the Web:

Web, Characteristics, and Crawling

Weimao Ke

wk@drexel.edu

Outline

- Web Basics
- Overview of Web and IR on the Web
- Spam and search engine optimization
- Crawling

WEB BASICS

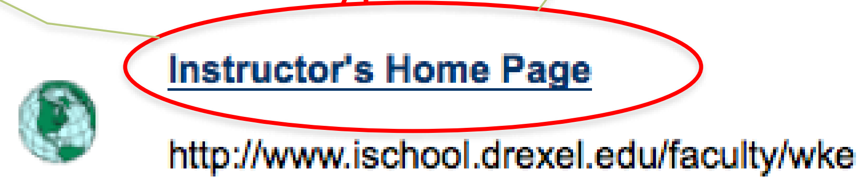
Basic terms on the Web

- URL: Uniform Resource Locator
  - The address of a Web page/document
- (X)HTML: Hypertext Markup Language
  - The language to author Web pages
- Hyperlink: a references to a document on the Web
  - A URL with anchor text (label)
- Anchor text: label of a hyperlink reference

URL: http://www.ischool.drexel.edu/faculty/wke

hyperlink

Anchor text: Instructor's Home Page



Instructor's Home Page

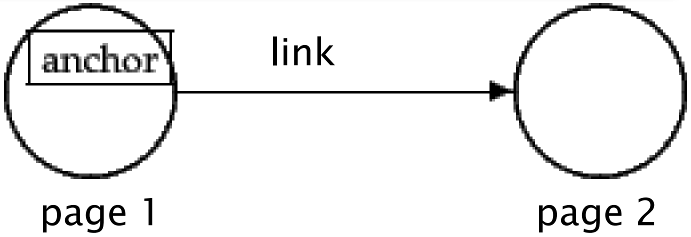
http://www.ischool.drexel.edu/faculty/wke

URL Example

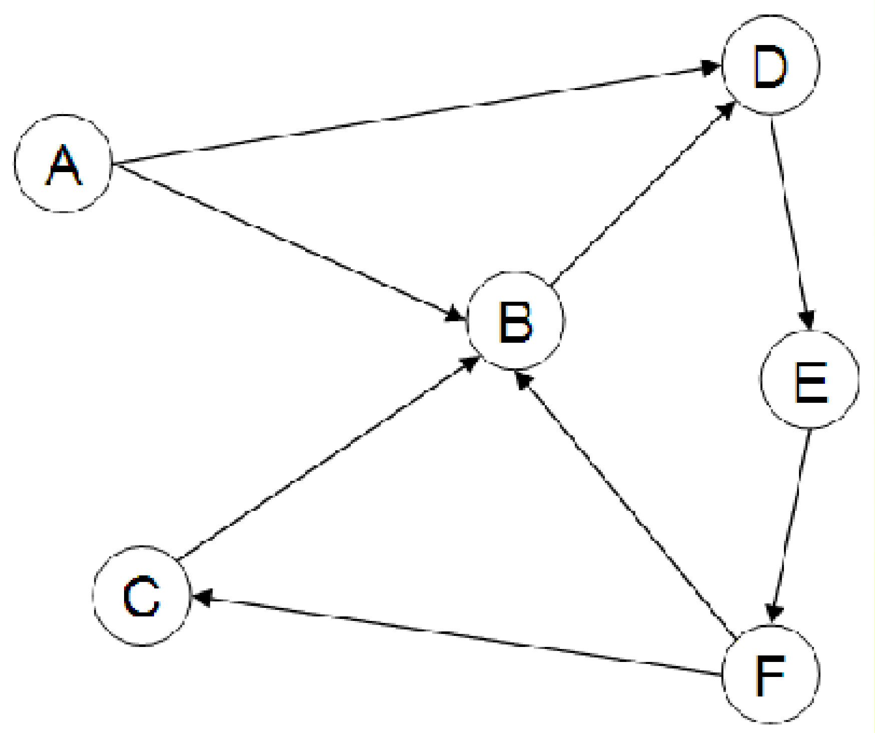
- Example: http://www.drexel.edu/about/history/brief.aspx
  - Protocol: HTTP, hypertext transfer protocol
  - Domain name: www.drexel.edu
  - Port: default 80
    - The above example is actually
    - http://www.drexel.edu:80/about/history/brief.aspx
  - Page: /about/history/brief.aspx
    - Directory/folder: "/about/history/"
    - File name: brief.aspx
- When you type a URL in the browser and hit Enter
  - Browser sends a request
    - of the page (brief.aspx)
    - to the server (www.drexel.edu) at port 80
    - using the indicated protocol (HTTP)
  - Servers sends HTML content back to browser
  - Browser displays the content

The Web as a graph

- A page is seen as a ***node***
- A hyperlink: an ***edge*** / ***arc***



- In-links and out-links
  - A has 2 out-links
  - Out-degree of A = 2
  - B has 3 in-links
  - In-degree of B = 3

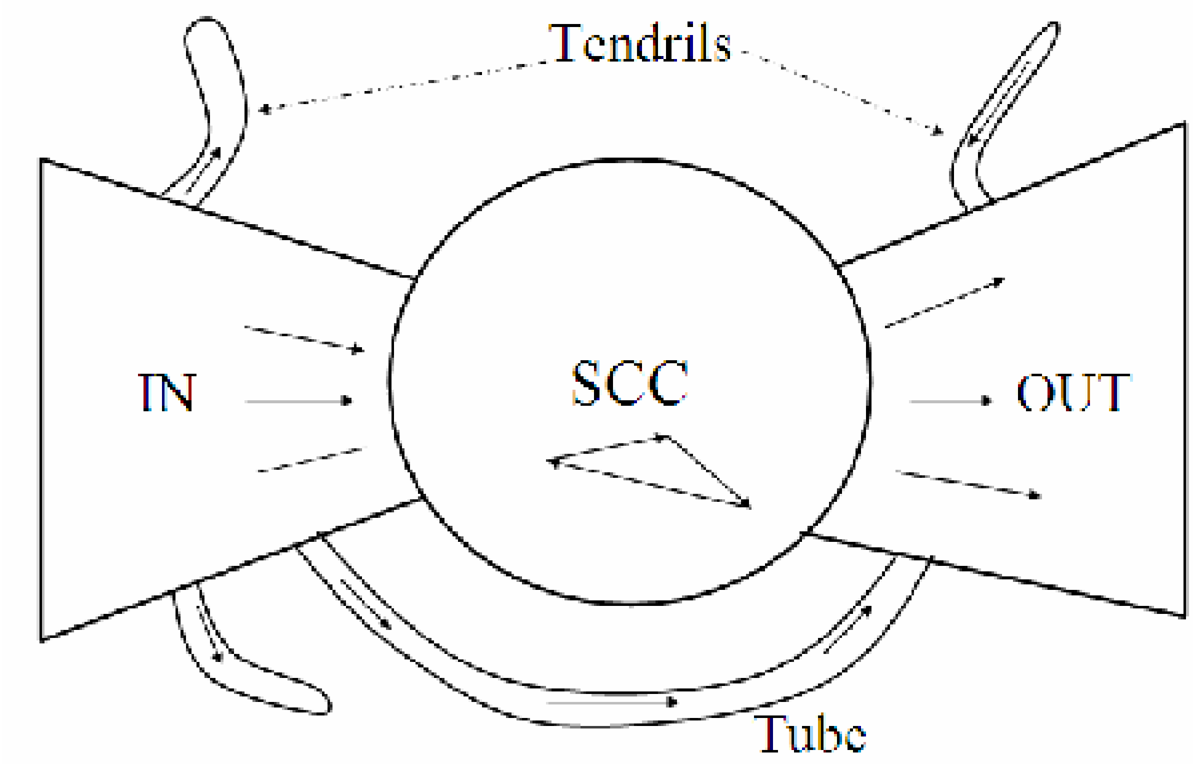


User, Content, and IR Systems on the Web

OVERVIEW

What does the Web look like?

- A bowtie
  - Three major components: IN, SCC, OUT
  - Tubes and tendrils



Size of the Web

- Number of Web Sites:
  - 1991: August 6, Tim Berners-Lee published the Web
  - 1994: 700 - 12,000 web sites
  - 1991-1997: Explosive growth, at a rate of 850% per year.
  - 1998-2001: Rapid growth, at a rate of 150% per year.
  - 2002-2006: Maturing growth, at a rate of 25% per year.
  - 2007: over 100 million web sites
  - 2010: about 200 million web sites
  - …
  - 2020: 2 billion websites (400 million active)

http://www.useit.com/alertbox/web-growth.html
https://hostingtribunal.com/blog/how-many-websites/#gref

Size of the Web

- Number of Web Pages:
  - 1997: 200 million pages
  - 1998: 800 million pages
  - 2005: 11.5 billion pages
  - 2007: 22.5 billion pages
  - 2010: trillion+ pages
  - …
  - 2020: lost count, but Google indexed < 60 billion
- Dynamic pages
  - such as those on facebook, quora, twitter, etc.
- Deep web, dark web, …

What remains…

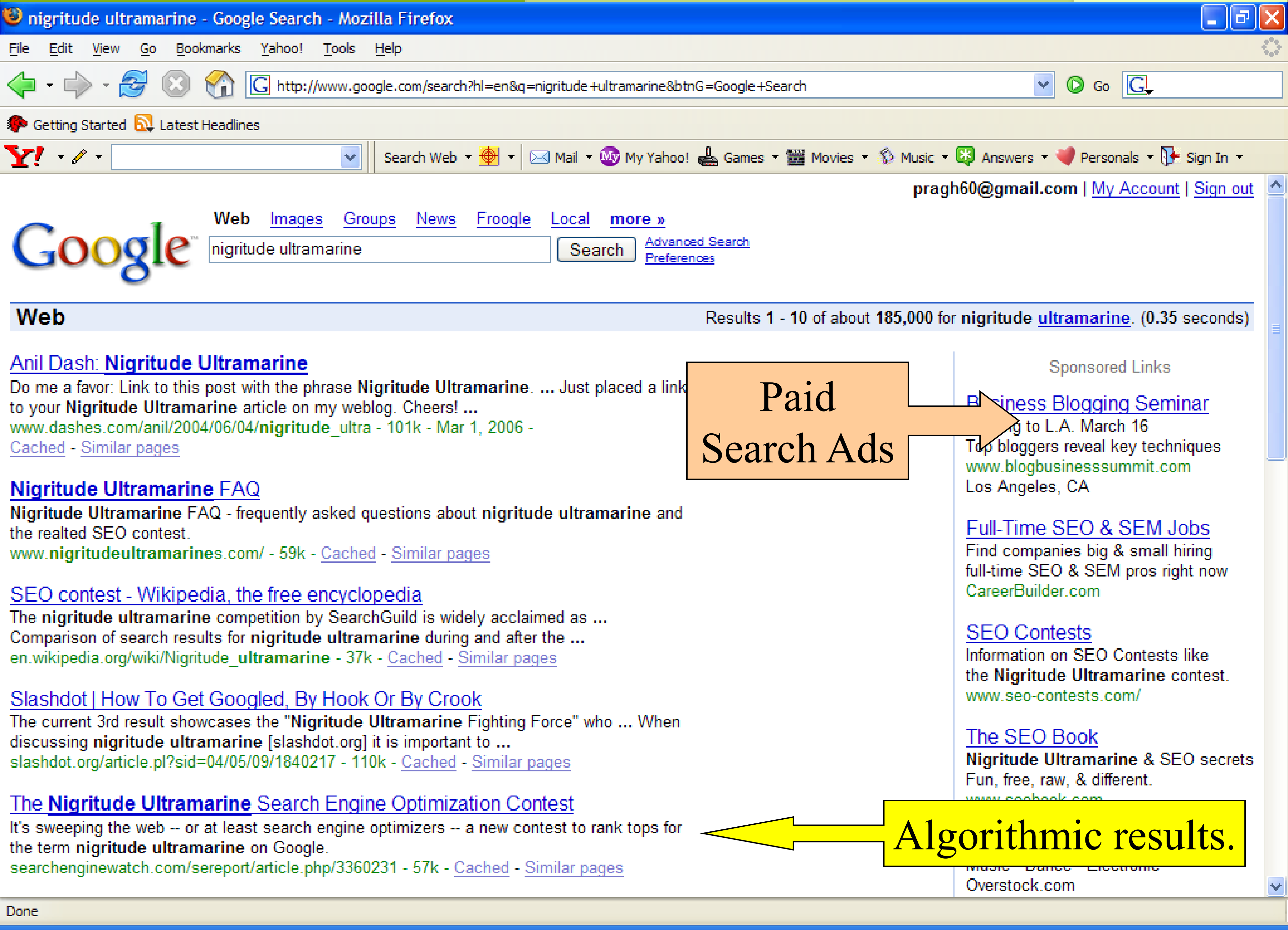
- Little change in terms of..
  - (X)HTML: standard markup language on the web
  - Hyperlinks: how pages connect to one another
  - “Small world”: 19-degree separation
    - Albert and Barabasi, 1999
  - Browser/Server architecture
    - Browser: the client (on your computer)
    - Web Server: the server (e.g., www.ischool.drexel.edu)

Brief (non-technical) history

- Early keyword-based engines ca. 1995-1997
  - Altavista, Excite, Infoseek, Inktomi, Lycos
- Paid search ranking: Goto (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 search "ads" to the side, independent of search results
  - Yahoo followed suit, acquiring Overture (for paid placement) and Inktomi (for search)
- 2005+: Google gains search share, dominating in Europe and very strong in North America
  - 2009: Yahoo! and Microsoft propose combined paid search offering


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**. ... Just placed a link 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.**nigritudeultramarine**s.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
...g 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/

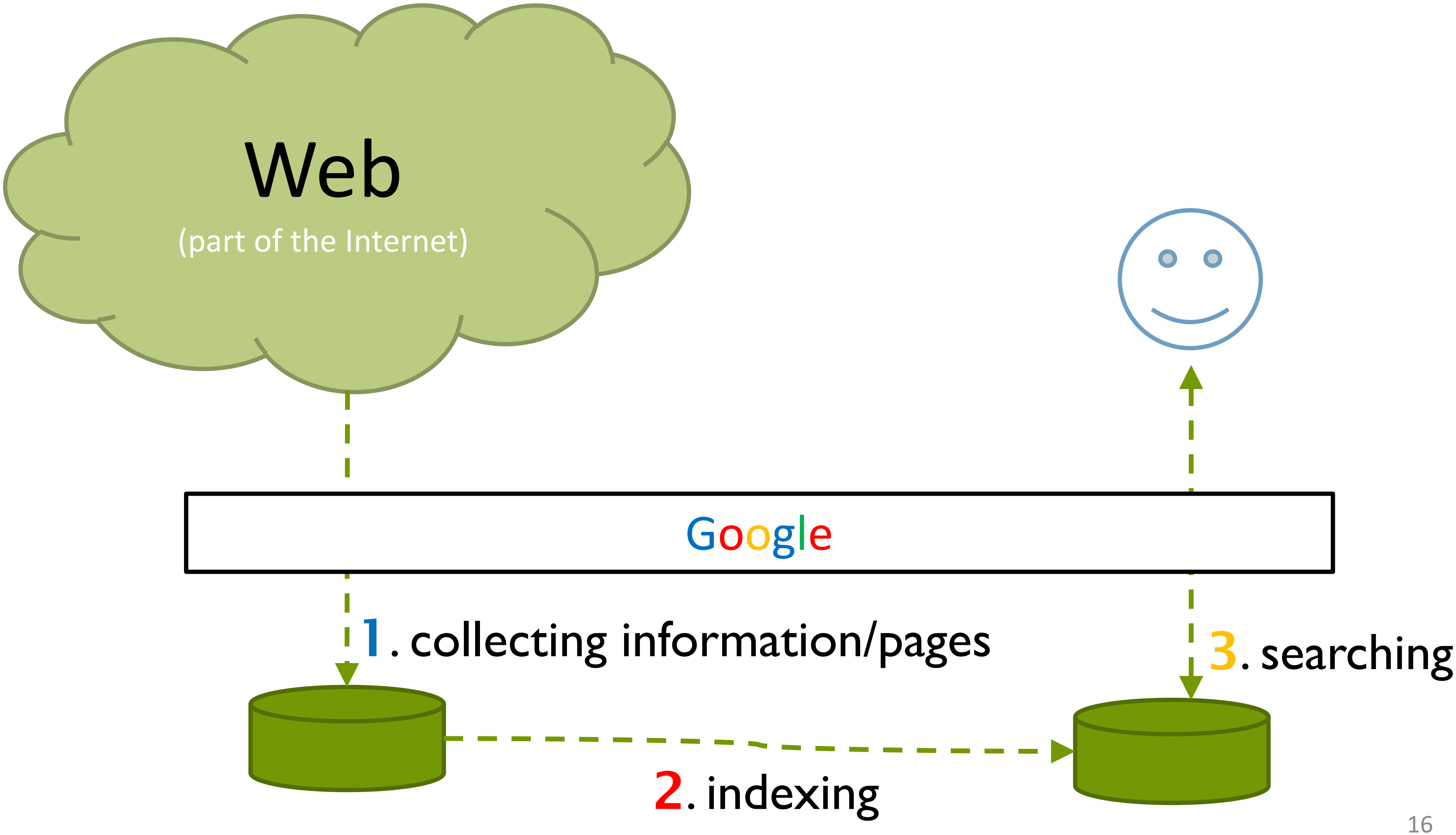
The SEO Book
Nigritude Ultramarine & SEO secrets
Fun, free, raw, & different.
www.seobook.com

Music - Dance - Electronic
Overstock.com

Paid Search Ads

Algorithmic results.

Classic Web IR Model



Web

(part of the Internet)

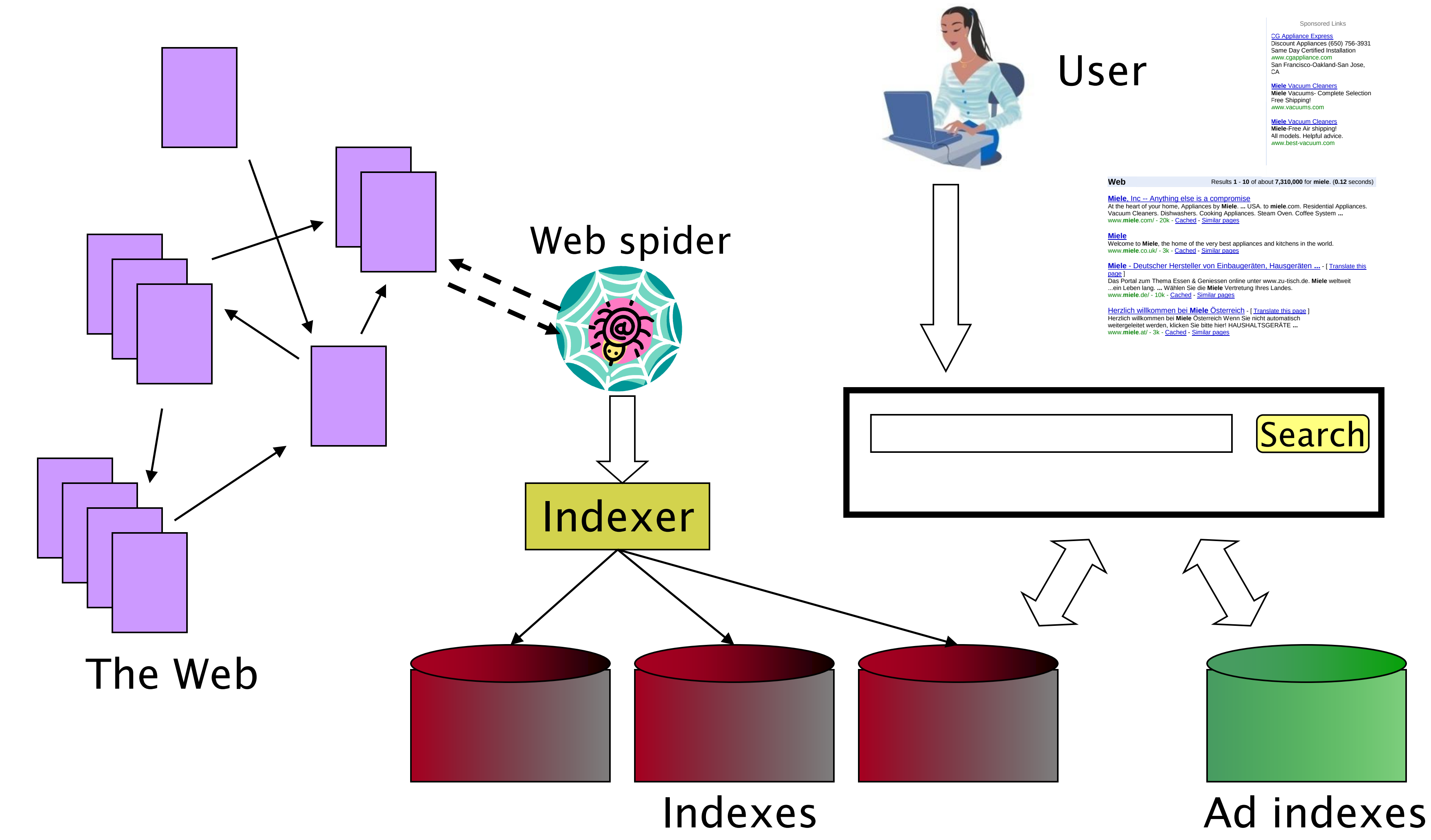
Google

1. collecting information/pages

2. indexing

3. searching

Another View of the Model



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

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

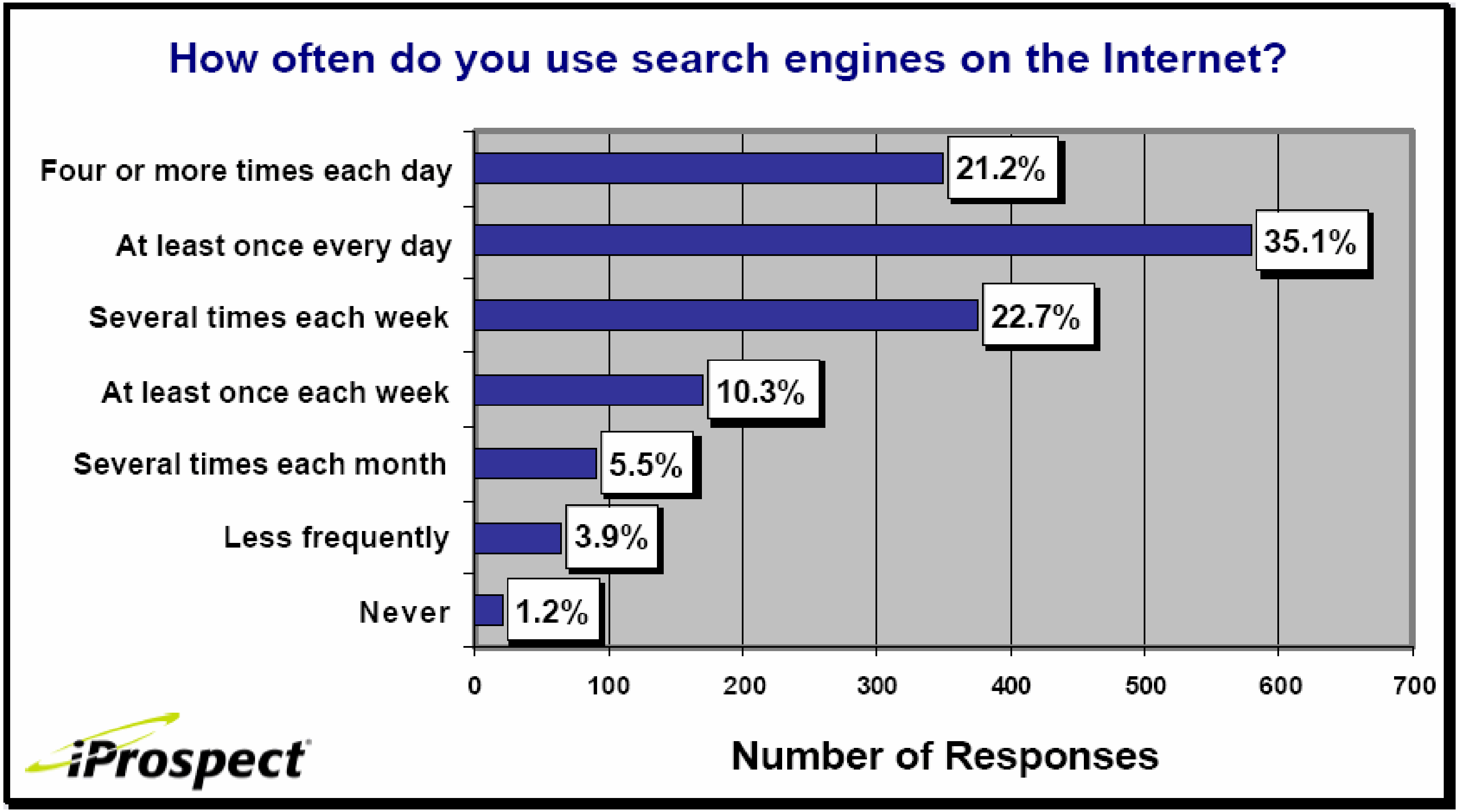
Miele, Inc -- Anything else is a compromise
At the heart of your home, Appliances by **Miele**. **...** USA. to **miele**.com. Residential Appliances. Vacuum Cleaners. Dishwashers. Cooking Appliances. Steam Oven. Coffee System **...**
www.**miele**.com/ - 20k - Cached - Similar pages

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

Search Engine Use



Think about these questions:

- Do web users who make the same query on a search engine always have the same information need?
- Do web pages that have the same keywords always have the same context?
- Is the content on a web page always the same? Have they ever changed? How often do they change?
- When you search, do you want to find the web pages with recently updated information or those that are created years ago and have never been changed.

User Needs

- User information needs on the Web
  - **Informational** – want to learn about something (~40% / 65%)

    e.g., seeking information about "vector space"
  - **Navigational** – want to go to that page (~25% / 15%)

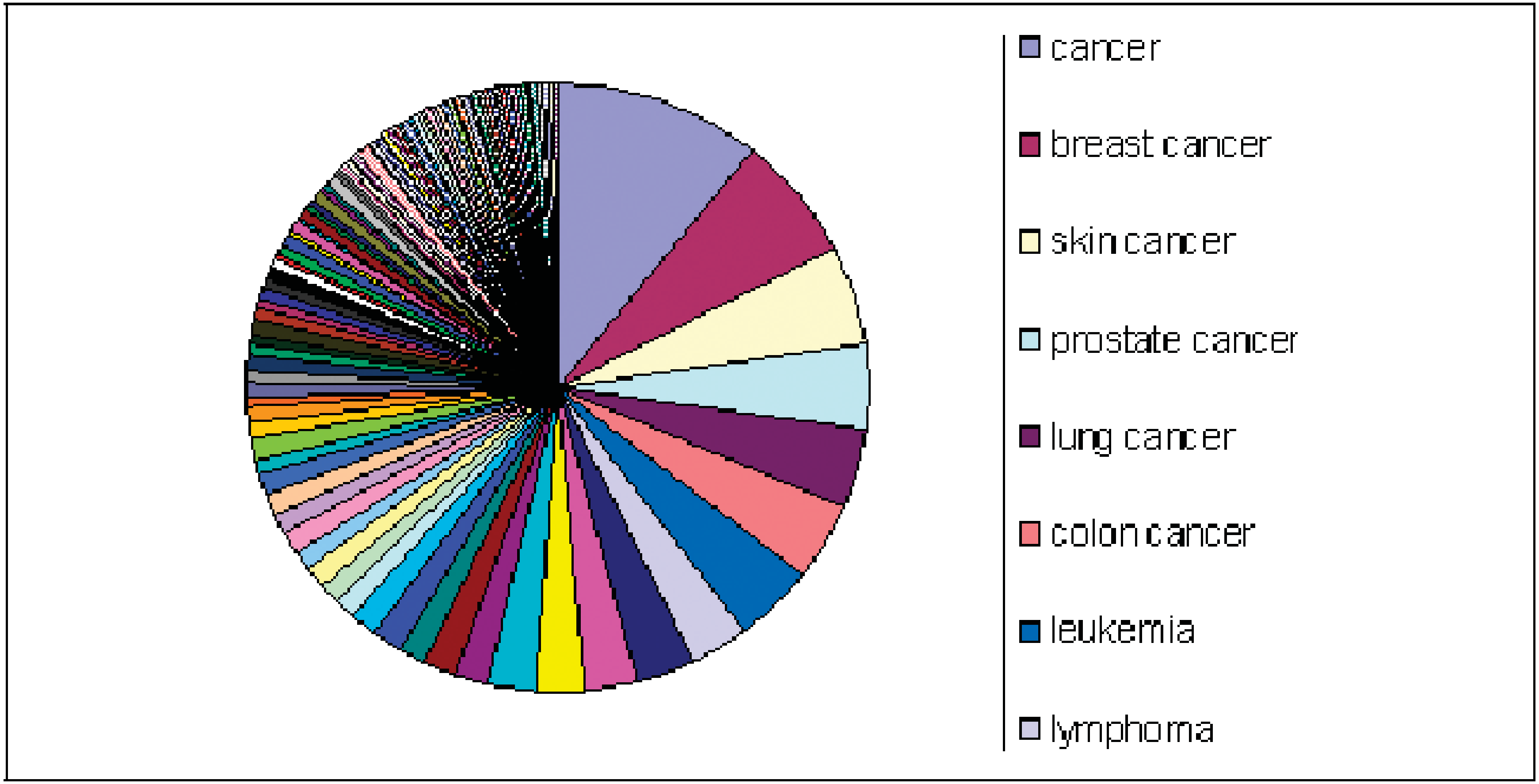
    e.g., to find the website of United Airlines
  - **Transactional** – want to do something (web-mediated) (~35% / 20%)
    - Access a service
    - Downloads
    - Shop

Seattle weather

Mars surface images

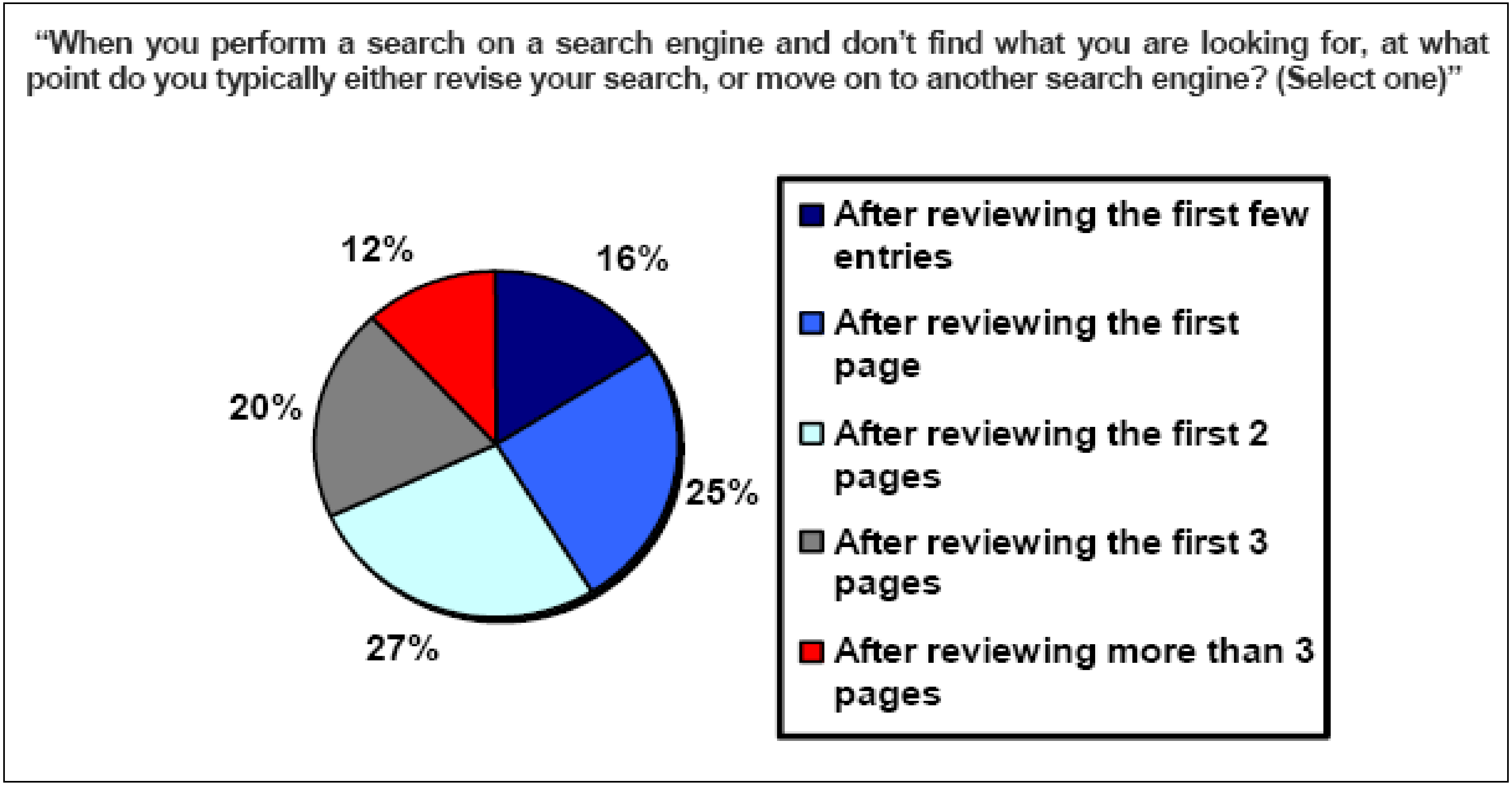
Canon S410

Query Complexity Distribution



Power law (like Zipf' s Law): few popular broad queries, many rare

How far do people look for results?

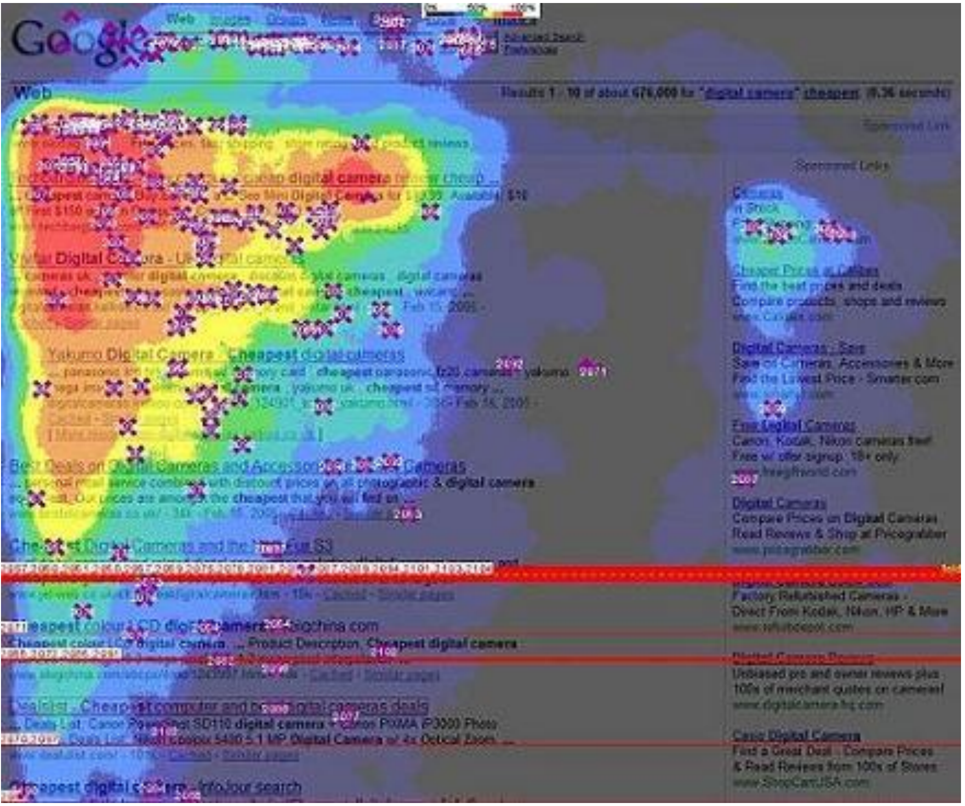


"When you perform a search on a search engine and don't find what you are looking for, at what point do you typically either revise your search, or move on to another search engine? (Select one)"

(Source: iprospect.com WhitePaper_2006_SearchEngineUserBehavior.pdf)

How do you look at search results?

- Eye tracking
- People pay attention to:
  - Top 2 – 3 items
  - Sponsored links


Google
Web
Results 1 - 10 of about 676,000 for "digital camera" cheapest. (0.36 seconds)
Sponsored Links

http://www.iqcontent.com/blog/2006/07/eyetracking-and-google-search-results/

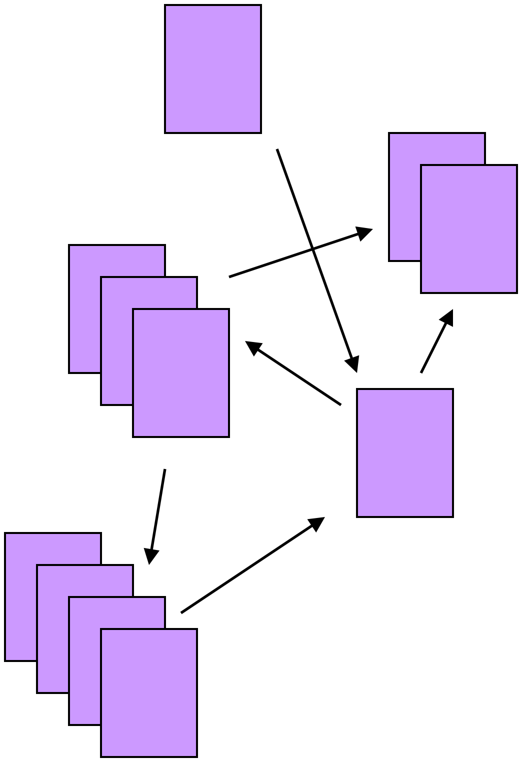
Users' empirical evaluation of results

- Quality of pages varies widely
  - Relevance is not enough
  - Other desirable qualities (non IR!!)
    - Content: Trustworthy, diverse, non-duplicated, well maintained
    - Web readability: display correctly & fast
    - No annoyances: pop-ups, etc
  - Can you find a page and guarantee its quality as well? Keyword does not tell you about the quality. Are there any other mechanisms that help in measuring the quality?
- Precision vs. recall
  - On the web, recall seldom matters
- What matters
  - Precision at 1? Precision above the fold?
  - Comprehensiveness – must be able to deal with obscure queries
    - Recall matters when the number of matches is very small
- User perceptions may be unscientific, but are significant over a large aggregate

Users' empirical evaluation of engines

- Relevance and validity of results
- UI – Simple, no clutter, error tolerant
- Trust – Results are objective
- Coverage of topics for polysemic queries
- Pre/Post process tools provided
  - Mitigate user errors (auto spell check, search assist,…)
  - Explicit: Search within results, more like this, refine ...
  - Anticipative: related searches
- Deal with idiosyncrasies
  - Web specific vocabulary
    - Impact on stemming, spell-check, etc
  - Web addresses typed in the search box

The Web document collection

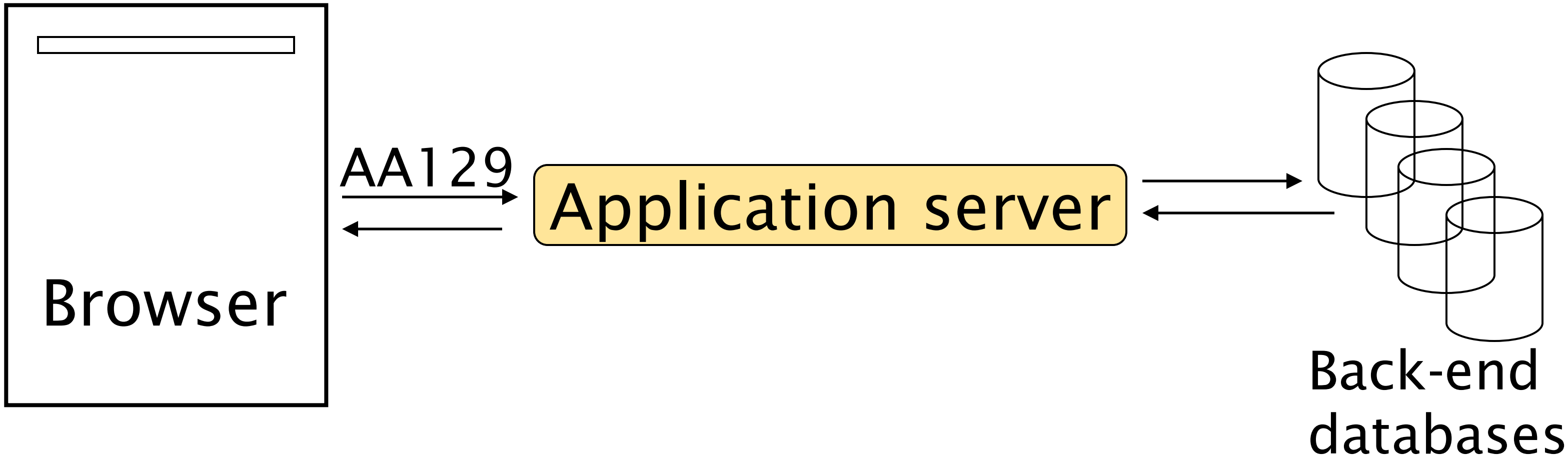


The Web

- 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 collections … but corporate records are catching up
- Growth – slowed down from initial “volume doubling every few months” but still expanding
- Content can be *dynamically generated*

The Web: Very 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



Can you index the dynamic content?

SPAM and Search Engine Optimization

SPAM

The trouble with paid search ads …

- It costs money. What's the alternative?
- *Search Engine Optimization:*
  - "Tuning" your web page to rank highly in the algorithmic search results for select keywords
  - Alternative to paying for placement
  - Thus, intrinsically a marketing function
- Performed by companies, webmasters and consultants ("Search engine optimizers") for their clients
- Some perfectly legitimate, some very shady

Search engine optimization (Spam)

- Motives
  - Commercial, political, religious, lobbies
  - Promotion funded by advertising budget
- Operators
  - Contractors (Search Engine Optimizers) for lobbies, companies
  - Web masters
  - Hosting services
- Forums
  - E.g., Web master world (www.webmasterworld.com)
    - Search engine specific tricks
    - Discussions about academic papers ☺

Simplest forms

- First generation engines relied heavily on *tf/idf*
  - The top-ranked pages for the query **`maui resort`** were the ones containing the most **`maui`**’s and **`resort`**’s
- SEOs responded with dense repetitions of chosen terms
  - e.g., **`maui resort maui resort maui resort`**
  - Often, the repetitions would be in the same color as the background of the web page
    - Repeated terms got indexed by crawlers
    - But not visible to humans on browsers

Pure word density cannot be trusted as an IR signal

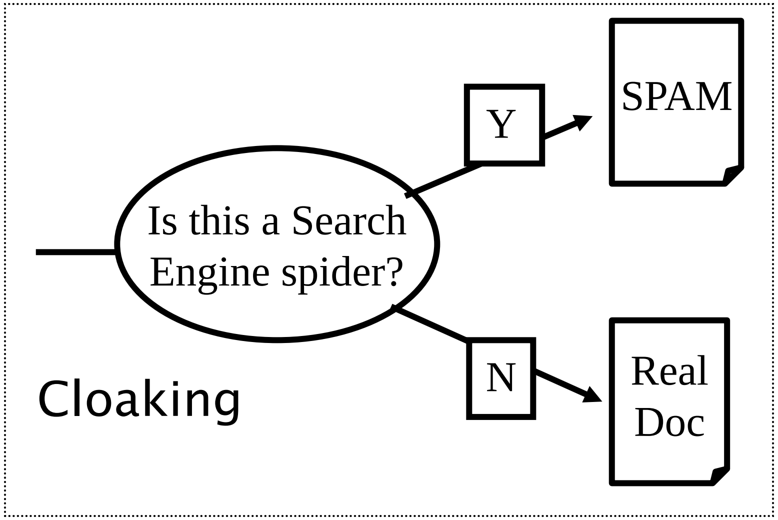
Variants of keyword stuffing

- Misleading meta-tags, excessive repetition
- Hidden text with colors, style sheet tricks, etc.

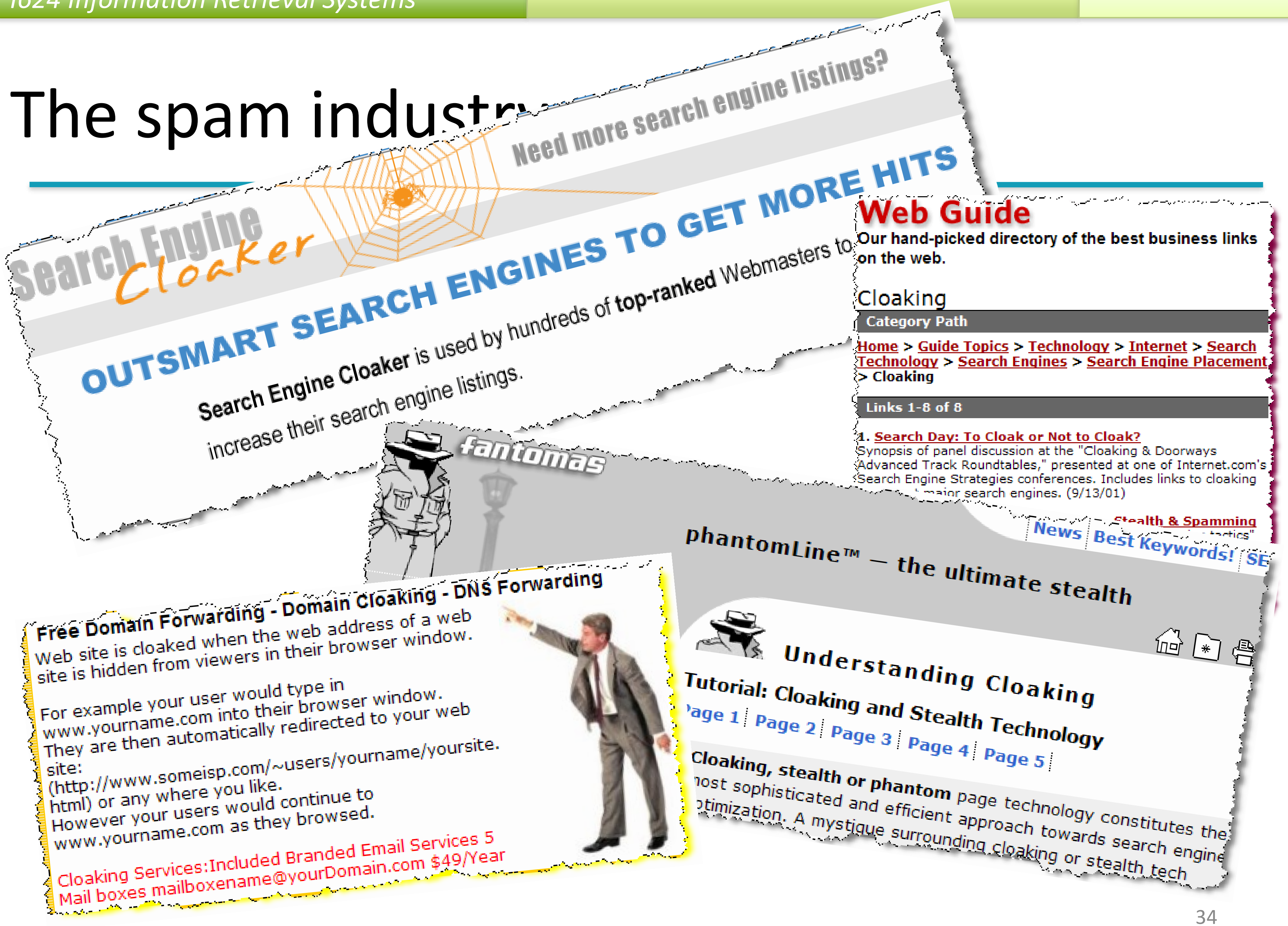
Meta-Tags =
“… London hotels, hotel, holiday inn, hilton, discount, booking, reservation, sex, mp3, britney spears, viagra, …”

Cloaking

- Serve fake content to search engine spider
- DNS cloaking: Switch IP address. Impersonate



The spam industry



Need more search engine listings?

OUTSMART SEARCH ENGINES TO GET MORE HITS

Search Engine Cloaker is used by hundreds of **top-ranked** Webmasters to increase their search engine listings.

Web Guide

Our hand-picked directory of the best business links on the web.

Cloaking

Category Path

Home > Guide Topics > Technology > Internet > Search Technology > Search Engines > Search Engine Placement > Cloaking

Links 1-8 of 8

1. Search Day: To Cloak or Not to Cloak?
Synopsis of panel discussion at the "Cloaking & Doorways Advanced Track Roundtables," presented at one of Internet.com's Search Engine Strategies conferences. Includes links to cloaking ... major search engines. (9/13/01)

Stealth & Spamming ... tactics"

phantomLine™ — the ultimate stealth

Understanding Cloaking

Tutorial: Cloaking and Stealth Technology

Page 1 | Page 2 | Page 3 | Page 4 | Page 5 |

Cloaking, stealth or phantom page technology constitutes the most sophisticated and efficient approach towards search engine optimization. A mystique surrounding cloaking or stealth tech

Free Domain Forwarding - Domain Cloaking - DNS Forwarding

Web site is cloaked when the web address of a web site is hidden from viewers in their browser window.

For example your user would type in www.yourname.com into their browser window. They are then automatically redirected to your web site: (http://www.someisp.com/~users/yourname/yoursite.html) or any where you like. However your users would continue to www.yourname.com as they browsed.

Cloaking Services:Included Branded Email Services 5 Mail boxes mailboxename@yourDomain.com $49/Year

More spam techniques

- **Doorway pages**
  - Pages optimized for a single keyword that re-direct to the real target page
- **Link spamming**
  - Mutual admiration societies, hidden links, awards
  - *Domain flooding:* numerous domains that point or re-direct to a target page
- **Robots**
  - Fake query stream – rank checking programs
    - "Curve-fit" ranking programs of search engines
  - Millions of submissions via Add-Url

The war against spam

- Quality signals - Prefer authoritative pages based on:
  - Votes from authors (linkage signals)
  - Votes from users (usage signals)
- Policing of URL submissions
  - Anti robot test
- Limits on meta-keywords
- Robust link analysis
  - Ignore statistically implausible linkage (or text)
  - Use link analysis to detect spammers (guilt by association)
- Spam recognition by machine learning
  - Training set based on known spam
- Family friendly filters
  - Linguistic analysis, general classification techniques, etc.
  - For images: flesh tone detectors, source text analysis, etc.
- Editorial intervention
  - Blacklists
  - Top queries audited
  - Complaints addressed
  - Suspect pattern detection

More on spam

- Web search engines have policies on SEO practices they tolerate/block
  - http://help.yahoo.com/help/us/ysearch/index.html
  - http://www.google.com/intl/en/webmasters/
- Adversarial IR: the unending (technical) battle between SEO's and web search engines
- Research http://airweb.cse.lehigh.edu/

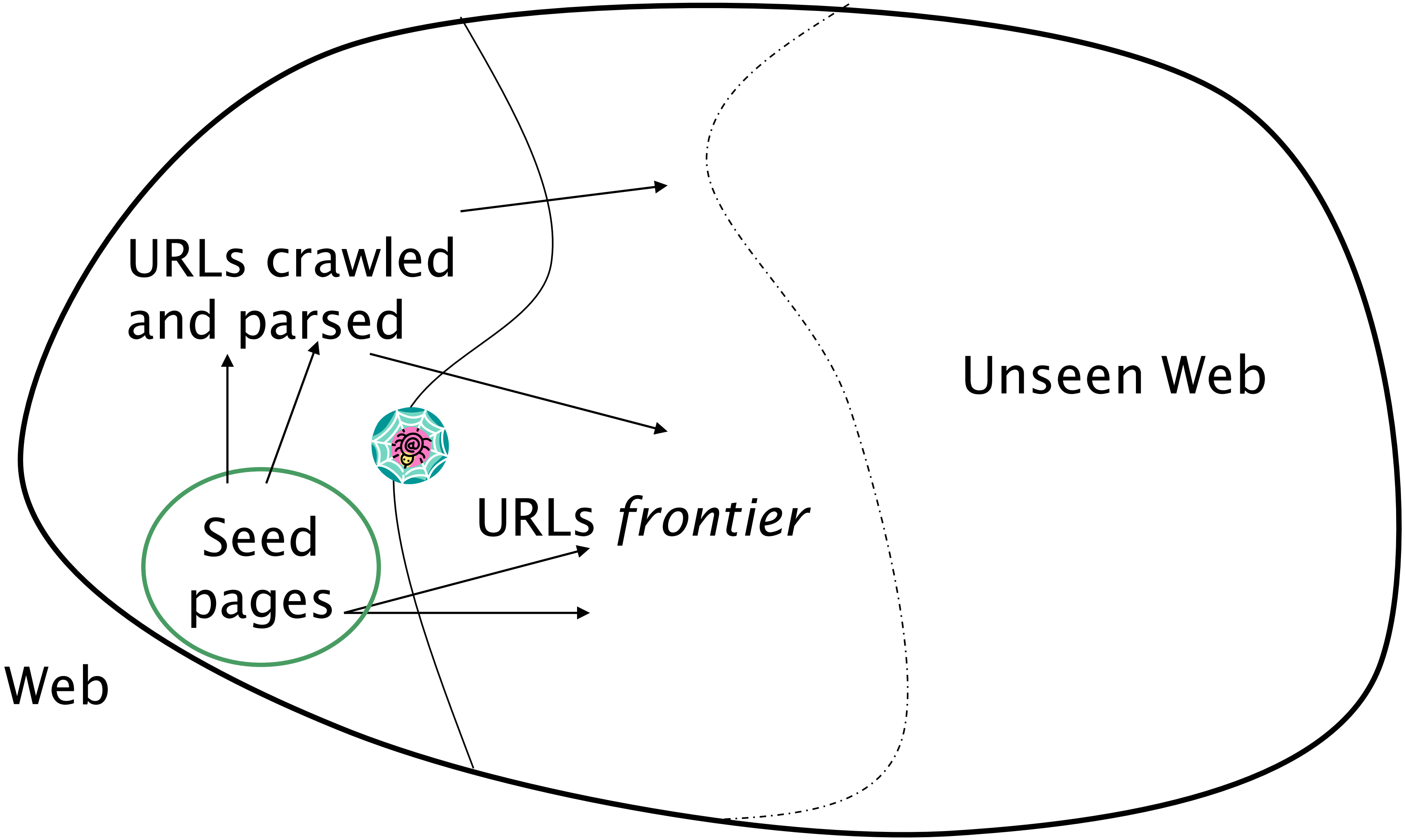
Information Collection

CRAWLING

Basic crawler operation

- Begin with known “seed” URLs
- Fetch and parse them
  - Extract URLs they point to
  - Place the extracted URLs on a queue
- Fetch each URL on the queue and repeat

Crawling picture



Simple picture – complications

- Web crawling isn't feasible with one machine
  - All of the above steps distributed
- Malicious pages
  - Spam pages
  - Spider traps – some dynamically generated
- Even non-malicious pages pose challenges
  - Latency/bandwidth to remote servers vary
  - Webmasters' stipulations
    - How "deep" should you crawl a site's URL hierarchy?
  - Site mirrors and duplicate pages
- Politeness – don't hit a server too often

What any crawler *must* do

- Be Polite: Respect implicit and explicit politeness considerations
  - Only crawl allowed pages
  - Respect *robots.txt* (more on this shortly)
- Be Robust: Be immune to spider traps and other malicious behavior from web servers

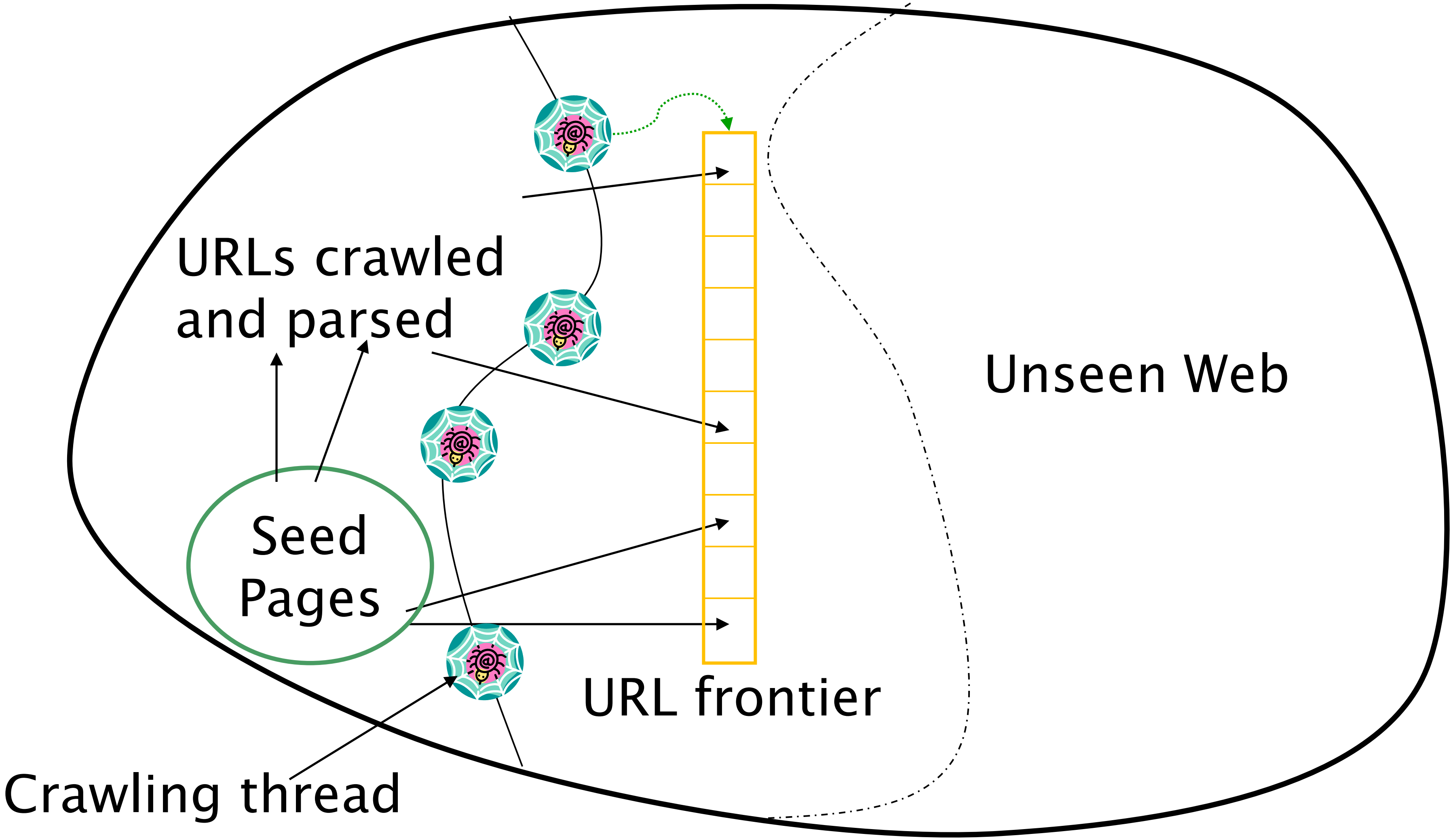
What any crawler *should* do

- Be capable of distributed operation: designed to run on multiple distributed machines
- Be scalable: designed to increase the crawl rate by adding more machines
- Performance/efficiency: permit full use of available processing and network resources

What any crawler *should* do

- Fetch pages of "higher quality" first
- Continuous operation: Continue fetching fresh copies of a previously fetched page
- Extensible: Adapt to new data formats, protocols

Updated crawling picture



URL frontier

- Can include multiple pages from the same host
- Must avoid trying to fetch them all at the same time
- Must try to keep all crawling threads busy

Explicit and implicit politeness

- Explicit politeness: specifications from webmasters on what portions of site can be crawled
  - robots.txt
- Implicit politeness: even with no specification, avoid hitting any site too often

Robots.txt

- Protocol for giving spiders (“robots”) limited access to a website, originally from 1994
  - www.robotstxt.org/wc/norobots.html
- Website announces its request on what can(not) be crawled
  - For a URL, create a file `URL/robots.txt`
  - This file specifies access restrictions

Robots.txt Example

- The following rule in robots.txt indicates:

  “No robot should visit any URL starting with ***/yoursite/temp/***, except the robot called ***searchengine***. ”

```
User-agent: *
Disallow: /yoursite/temp/

User-agent: searchengine
Disallow:
```

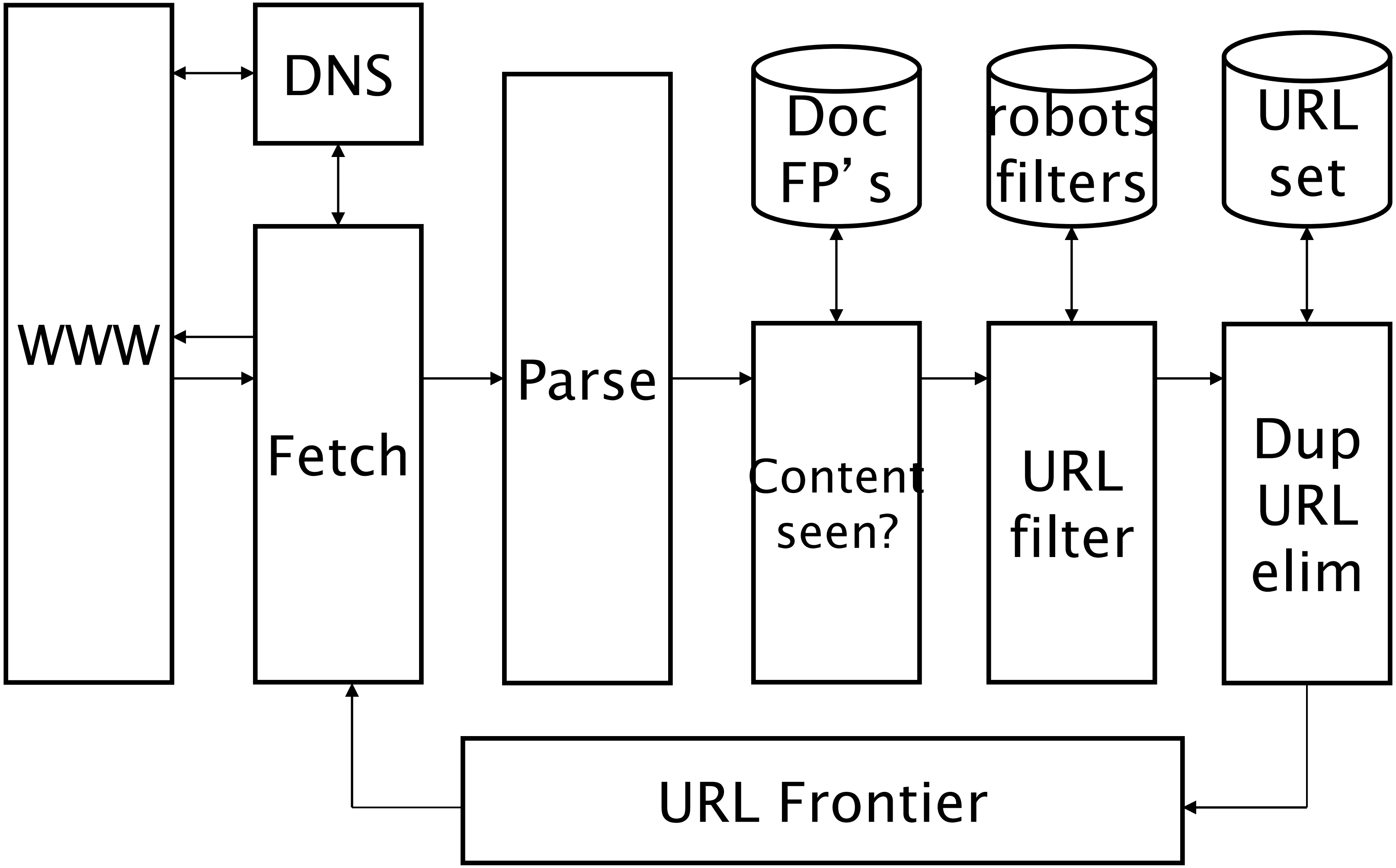
Processing steps in crawling

- Pick a URL from the frontier
- Fetch the document at the URL
- Parse the URL
  - Extract links from it to other docs (URLs)
- Check if URL has content already seen
  - If not, add to indexes
- For each extracted URL
  - Ensure it passes certain URL filter tests
  - Check if it is already in the frontier (duplicate URL elimination)

Which one?

e.g., only crawl .edu, obey robots.txt, etc.

Basic crawl architecture



DNS (Domain Name Server)

- A lookup service on the internet
  - Given a URL, retrieve its IP address
  - Service provided by a distributed set of servers – thus, lookup latencies can be high (even seconds)

Content seen?

- Duplication is widespread on the web
- If the page just fetched is already in the index, do not further process it
- This is verified using document fingerprints or shingles

URL frontier: two main considerations

- Politeness: do not hit a web server too frequently
- Freshness: crawl some pages more often than others
  - e.g., pages (such as News sites) whose content changes often

These goals may conflict each other.

Resources

- IIR Chapters 19, 20