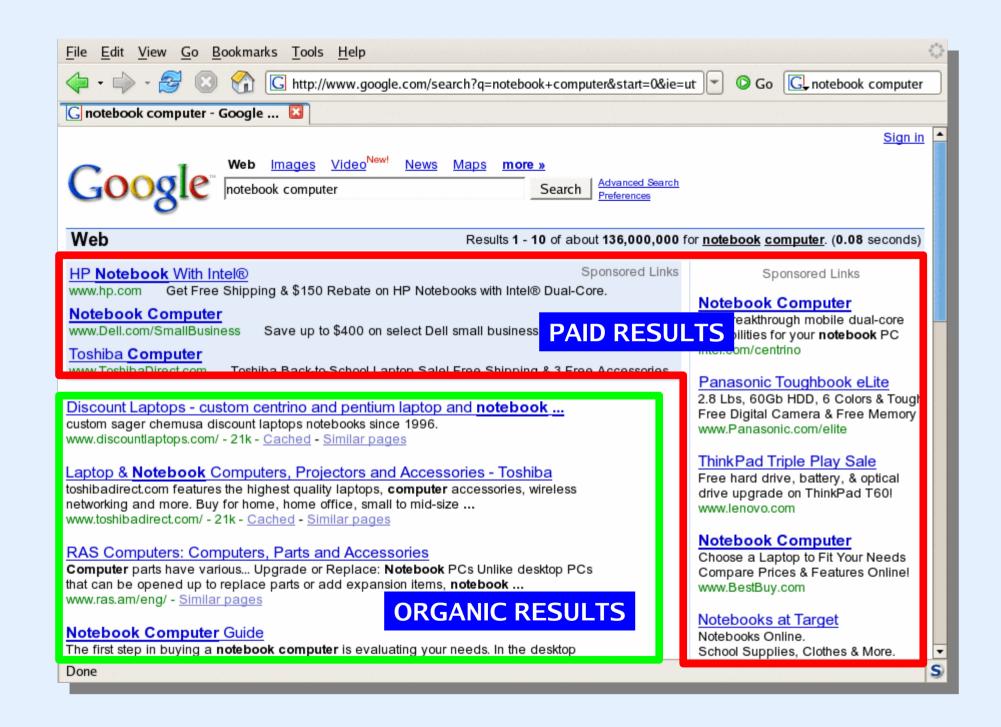
How Search Engines Work

- Today we show how a search engine works
 - What happens when a searcher enters keywords
 - What was performed well in advance
 - Also explain (briefly) how paid results are chosen
- If we have time, we will also talk about the size of the Web

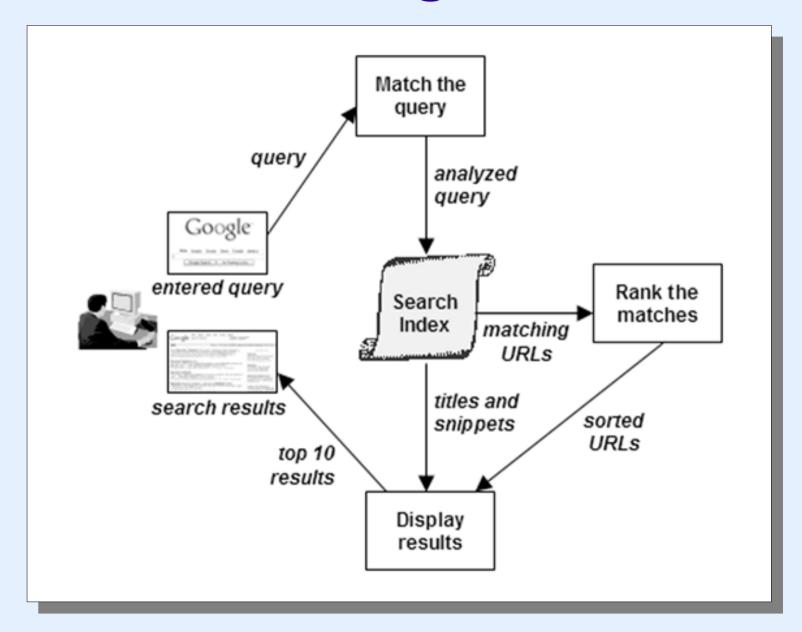
(If you *really* want to know how web search engines work, take my CSE345 WWW Search Engines course in the spring!)



Building an index

- A search engine does not examine every page on the web when a user puts in a query
- The engine first builds an index
 - Custom database of all the words on all pages
 - Search engine also stores other information

Overview of organic search



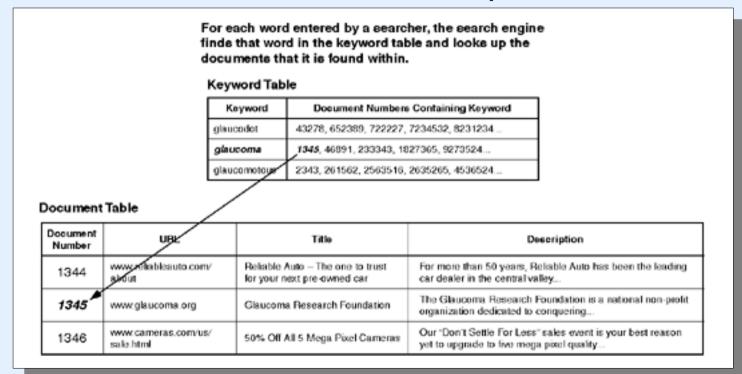
Matching the Search Query

- The search query is everything that the user types to get results
 - It is made up of one or more search terms, plus optional special characters
- Analyzing the Query
 - Expanding the query
 - Word variants: plural/singular, various verb forms
 - Spelling correction
 - Phrases, anti-phrases, and stop words
 - Word order
 - Search operators



Matching the Search Query

- Organic query matches
 - Find pages with each of the remaining query terms
 - Document IDs are listed in a term index
 - Document information is in a separate doc index



Matching the Search Query

- Paid placement matches
 - Similar to organic match, but using a separate database of ads
 - Uses similar processing to select which query terms to use
 - Advertisers choose which queries can match
 - Might require exact match, or allow broad matching
 - Simpler/faster because there are fewer ads to search through

Ranking Organic Matches

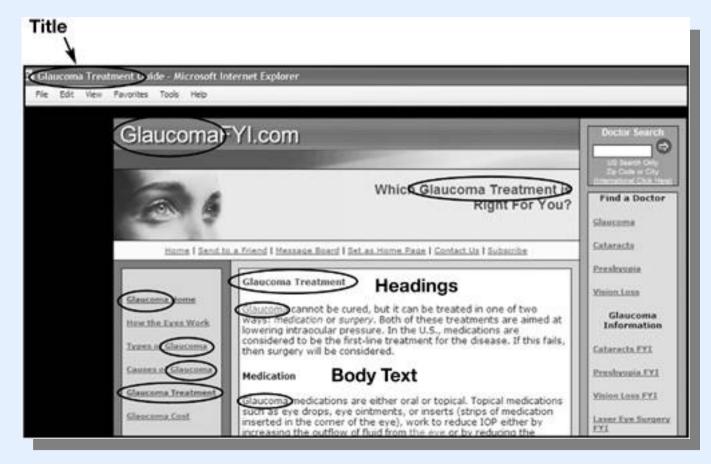
- This is a complex, active research area
 - Goal is to sort matching results from 'best' to 'worst'
 - Many factors contribute to different rankings in the various engines
 - Ranking functions are under continuous change
- Primary factors
 - Text analysis: keyword density and prominence
 - Link analysis: page and site authority estimates
 - Anchor text: terms used to describe page by others
 - Traffic analysis: which results get clicked on

Text Analysis: Keyword Density

- A.k.a. keyword weight
- Generally refers to the relative frequency of a term on the page
- Higher keyword density generally means that a document is more 'about' that keyword
- Natural text has a maximum reasonable density
 - The book cites a 7% density threshold
- Multi-term queries target keyword proximity
 - Pages with the same terms adjacent in same order would benefit most

Text Analysis: Term Prominence

- Where do the query terms appear?
- Good places include:
 - Title
 - Headings
 - Start of body
- Terms in such places could get extra weight



Link Analysis: Estimating Authority

- A typical short query matches millions of pages
 - Many could even have the same textual (relevance) weight from keyword density and prominence
- Link analysis estimates the importance of each page, based on the link structure around it
- The more respected a site is, the more links point to it
- Some links are more important than others
 - A link from Yahoo (or the White House!) signifies much more than a link from geocities.com

Google's PageRank

- The best-known link analysis algorithm
 - Algorithm published in 1998
 - Very well-studied; improvements are still being made to it today
- The authoritativeness of a page grows if
 - More pages link to it
 - The pages that link to it increase their authority
- The original algorithm is not a significant component of Google's ranking approach today
 - Many have shown that it performs poorly now

Anchor Text

- What is a page about?
- Page builders often summarize a page (or the significant aspect of a page) in the anchor text (the text of a link)
 - These short descriptions look a lot like queries!
 - Can help determine value of link
 - A significant component for ranking today

Traffic Analysis

- Many engines will track which links you click on from a results page
- Such clicks can be considered "votes" for URLs
- Re-ordering based on clicks can improve ranking quality [Joachims et al., 2005]
- DirectHit search engine used click-throughs to generate top-10 results (purchased by Ask Jeeves in 2000)

Ranking Paid Placement

- Simplest approach: rank by highest bidder
 - Originally developed by Overture (a.k.a. goto.com)
 - Advertisers can change bids continuously, and can specify a particular budget
- Google's approach: rank by most valuable
 - Combination of bid and click-through rate
 - More relevant (clicked) ads move up in rank
 - Users find ads more useful

Notebook Computer

Get breakthrough mobile dual-core capabilities for your **notebook** PC intel.com/centrino

Panasonic Toughbook eLite

2.8 Lbs, 60Gb HDD, 6 Colors & Tough Free Digital Camera & Free Memory www.Panasonic.com/elite

ThinkPad Triple Play Sale

Free hard drive, battery, & optical drive upgrade on ThinkPad T60! www.lenovo.com

Notebook Computer

Choose a Laptop to Fit Your Needs Compare Prices & Features Online! www.BestBuy.com

Notebooks at Target

Notebooks Online. School Supplies, Clothes & More.

Displaying Search Results

- Once the set of results has been collected and ranked, the results page needs to be generated
- For first page, select top results (typically 10)
 - Look up title, URL for linking (and often display)
 - Generate snippet (portion of page text that illustrates query terms) or look up ad copy

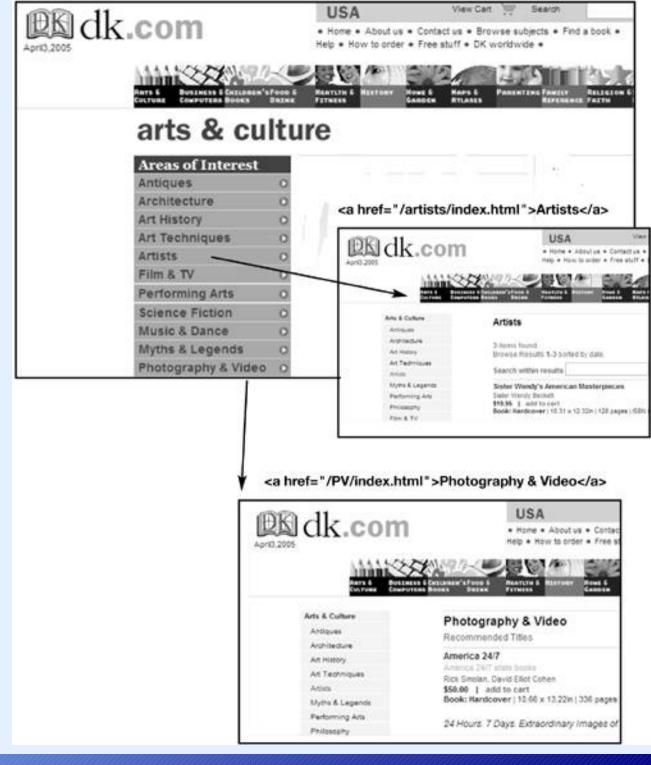
<u>Laptop & Notebook Computers, Projectors and Accessories - Toshiba</u> toshibadirect.com features the highest quality laptops, **computer** accessories, wireless networking and more. Buy for home, home office, small to mid-size ... www.toshibadirect.com/ - 21k - <u>Cached</u> - <u>Similar pages</u>

RAS Computers: Computers, Parts and Accessories

Computer parts have various... Upgrade or Replace: **Notebook** PCs Unlike desktop PCs that can be opened up to replace parts or add expansion items, **notebook** ... www.ras.am/eng/ - <u>Similar pages</u>

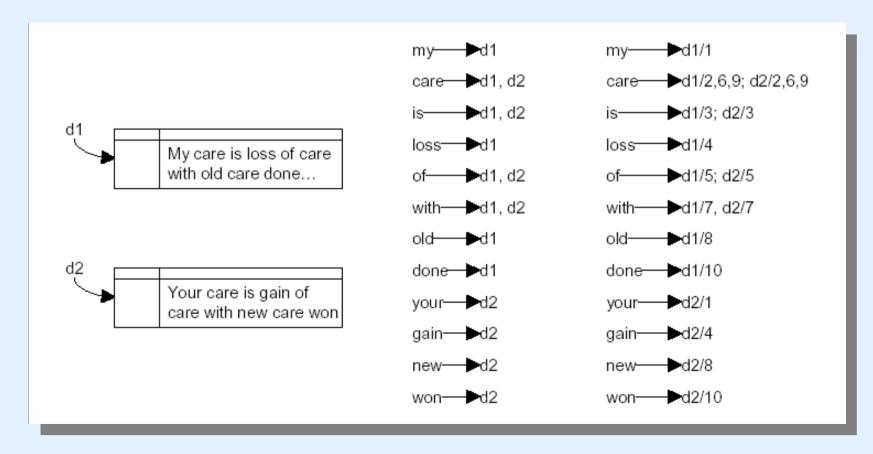
Collecting Material for the Organic Index

- Primarily using a crawler/spider
 - Given a seed
 list of links, visit
 each one and
 add any new
 URLs found to
 the list of links
 to visit



Building the Organic Index

- For each page retrieved, extract the text
 - For each term in the text, add the page's ID (and optionally, positions) to the list of docs for that term



Building the Organic Index

- For each page retrieved
 - Extract the links
 - Record anchor text for each link
 - Record Title and URL
- What to crawl?
 - Can't crawl all pages!
 - Need to re-crawl oft-changing pages
- Some engines allow trusted feeds (typically a form of paid inclusion) to get content indexed

Content Analysis

- Convert different types of documents
 - Use a single standard internal representation
 - Lots of file types: Word, PDF, PostScript, etc.
- Recognize language used

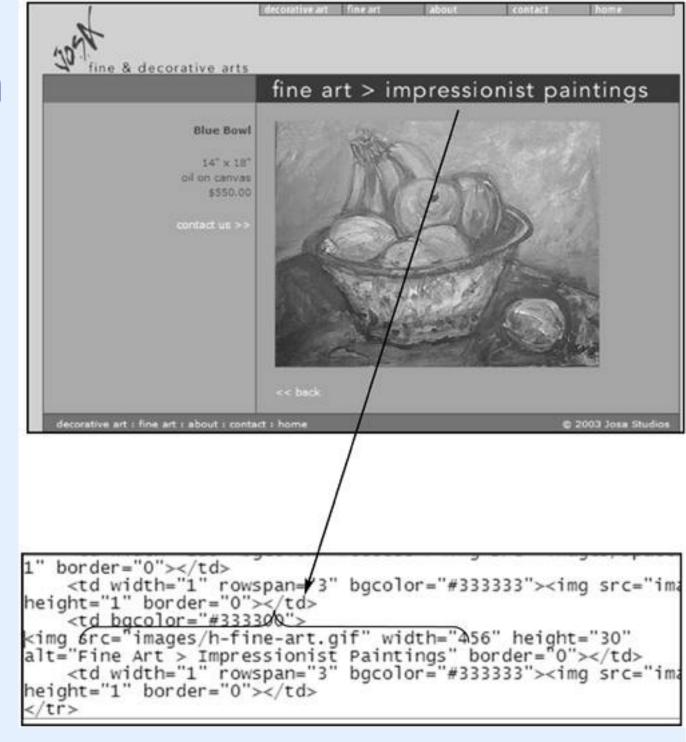
They also extract additional text from a page

What search engines

(and sight-impaired users)

don't see

- They cannot read images (even text in images)
- Often they do not read Flash content or JavaScript



What search engines can see

- Image names
- Image alt text

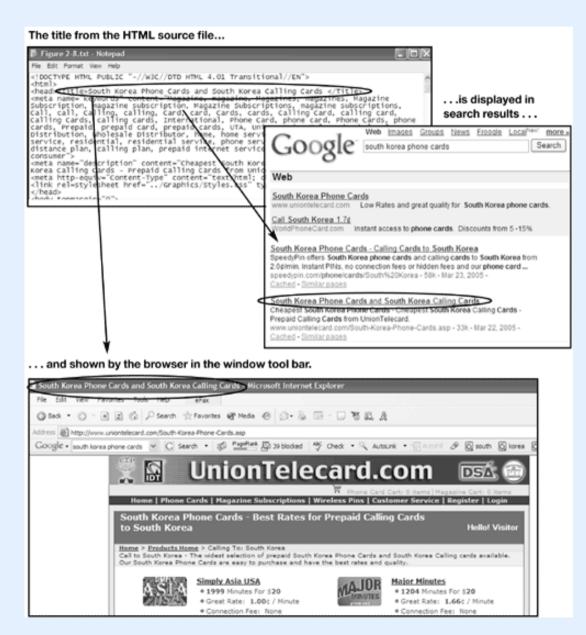


Zebra T-shirt

- Sizes Adult (S-3XL) and Children are Youth (XS-L)
- Prices Adult \$14, Children \$10 and (2XL-3XL) \$16
- <u>Colors</u> White, Ash, Nat, Yellow, Rose, Spruce, Sky, Blue
- See how they would look on different T-shirt colors
- See other Wild Animals shirts available at our site
- See the List Over 400 designs available online
- See us FRAMED and Java Interactive

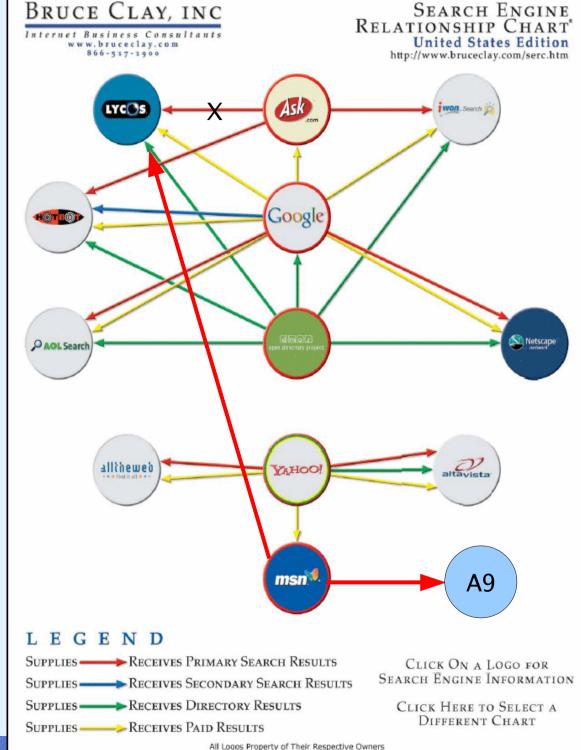
What search engines can see

- Image names
- Image alt text
- Meta text
 - Title
 - Description
 - Keywords
 - (often ignored)
 - Other directives
- URL text



Search Engine Relationships

- Business relationships have changed significantly over the past five years or so.
- See the Search Engine Relationship Chart as it can also show connections over time.
- There are more players than shown (such as Gigablast, Snap.com) and lots of international engines.

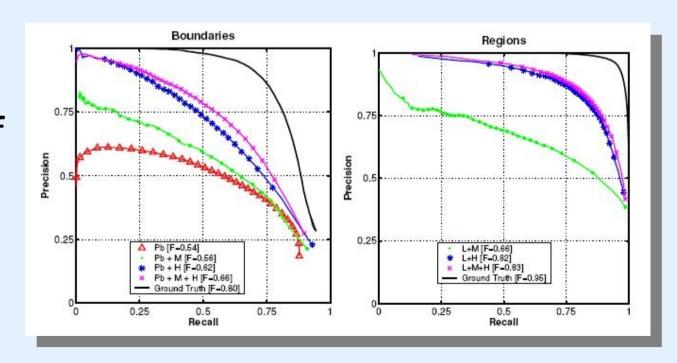


Evaluating Organic Search Results

- Precision: fraction of search results that are correct (relevant) to a query
- Recall: fraction of all correct (relevant) answers included in a set of search results

Evaluating Organic Search Results

- Precision: fraction of search results that are correct (relevant) to a query
- Recall: fraction of all correct (relevant) answers included in a set of search results
- Improving one usually results in worsening of the other



Evaluating Organic Search Results

- Precision: fraction of search results that are correct (relevant) to a query
- Recall: fraction of all correct (relevant) answers included in a set of search results
- Improving one usually results in worsening of the other
- In web search, neither can be measured exactly!
 - Still useful to think about how a change will affect performance



Depends!



Depends!



- What if I turn on a laptop that can produce links to an infinite number of pages?
 - Proposed by Andrei Broder who has studied this

 Perhaps you mean the size of the index used by web search engines?



- Perhaps you mean the size of the index used by web search engines?
 - This is a recurring debate
 - In 2005, Google was reporting 8B+ pages indexed
 - Yahoo then announced it had indexed almost 20B
 - Google declared Yahoo as counting differently
 - Google no longer reports its index size
 - and regularly underreports the number of machines it uses

- Perhaps you mean the size of the index used by web search engines?
 - This is a recurring debate
 - In 2005, Google was reporting 8B+ pages indexed
 - Yahoo then announced it had indexed almost 20B
 - Google declared Yahoo as counting differently
 - Google no longer reports its index size
 - and regularly underreports the number of machines it uses
- Estimates of intersection size in 1995 of top 4 indexes was only about 2.7B (different crawls!)

- Perhaps you mean the size of the index used by web search engines?
 - This is a recurring debate
 - In 2005, Google was reporting 8B+ pages indexed
 - Yahoo then announced it had indexed almost 20B
 - Google declared Yahoo as counting differently
 - Google no longer reports its index size
 - and regularly underreports the number of machines it uses
- Estimates of intersection size in 1995 of top 4 indexes was only about 2.7B (different crawls!)
- What about pages not indexed by the engines?

- How large is the indexable web?
 - That is, ignoring the pages that require passwords, links within flash content, or forms to be filled in (search boxes, registration, etc.)
 - Recent estimate is > 11.5B [Gulli & Signorini, 2005]
 - Fairly close in time to Yahoo's 20B claim

- How large is the indexable web?
 - That is, ignoring the pages that require passwords, links within flash content, or forms to be filled in (search boxes, registration, etc.)
 - Recent estimate is > 11.5B [Gulli & Signorini, 2005]
 - Fairly close in time to Yahoo's 20B claim
- The hidden web (the rest) is 2-500 times larger!
 - Again, just reported estimates...

So it is impossible to know the size of the Web!