



Searching the Web

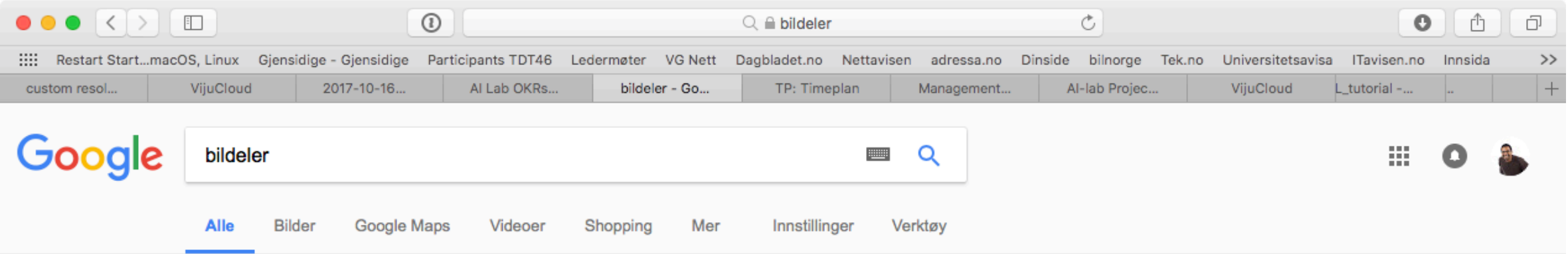
Uke 42 – Lecture 9

Brief (non-technical) history

- Early keyword-based engines
 - Altavista, Excite, Infoseek, Inktomi
- Sponsored search 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 save Inktomi
 - 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 followed suit, acquiring Overture (for paid placement) and Inktomi (for search)



Bestill billige bildeler - Rabatter fra 10% til 70% - autodeler.co.no

[Annonse] www.autodeler.co.no/Bildeler/Online 56 91 40 56
Over 1 000 000 Bildeler på nett. 2 års garanti. 100% høy kvalitet. Bestill nå!
Brands: BOSCH, HELLA, VEMO, ATE, BEHR, Febi Belstein, NK, SACHS, NGK, Valeo
Spesialtilbud: 19 % rabatt på alle produkter

Bildeler på nett? - Originale Bildeler Billig - eurodel.no

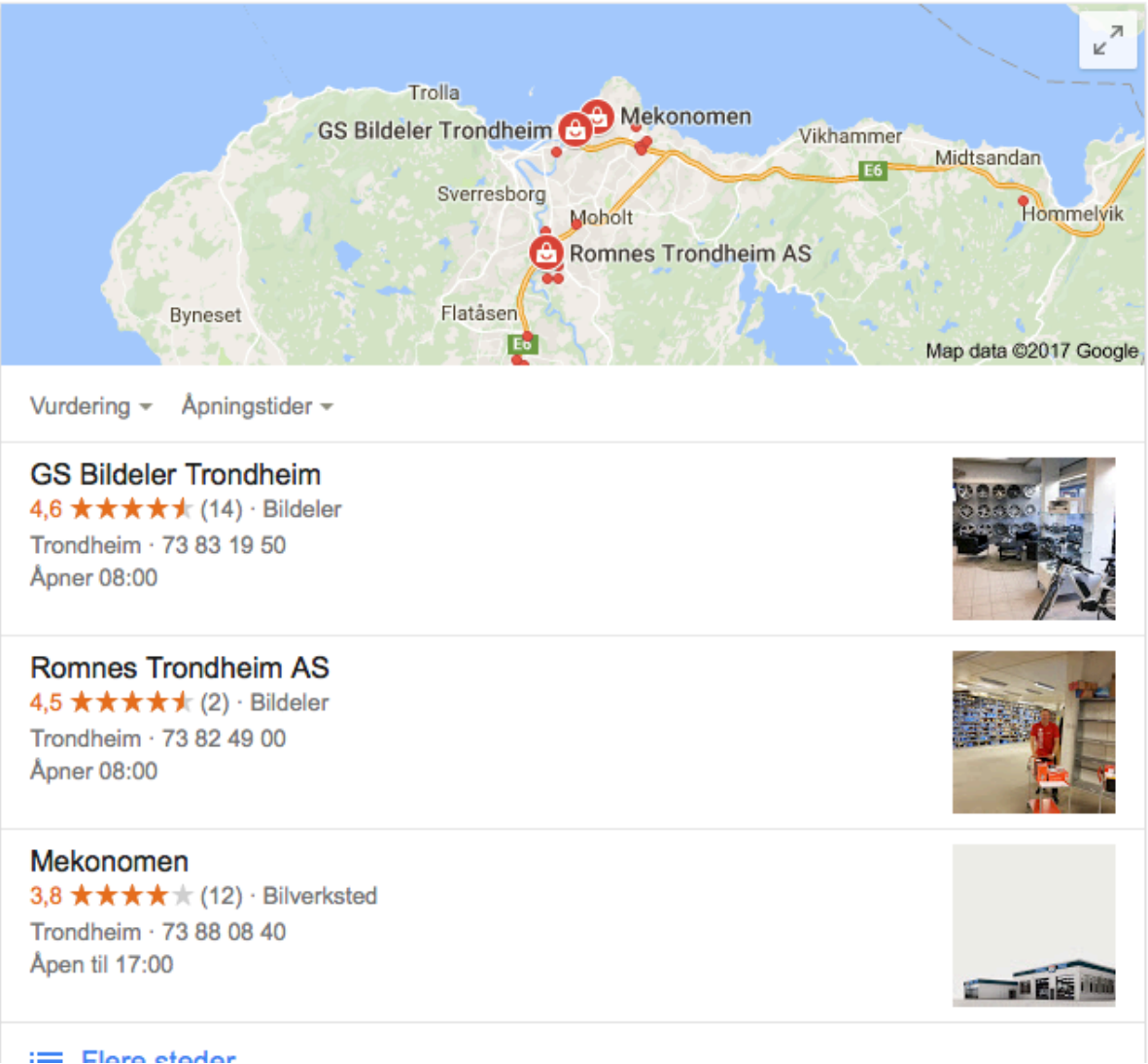
[Annonse] www.eurodel.no/Salg/Bildeler
4,5 ★★★★★ -vurdering for eurodel.no
Produsentens garanti på alle varer, 30 dagers åpent kjøp og original kvalitet.
Fasiliteter: Søk på bilnummeret, Søk på bilmerke, Søk etter OE nummer
Merkevarer: Volvo, Seat, Audi, Suzuki, VW, Honda, Peugeot, Mitsubishi

Bestill nye bildeler - Stort utvalg, rask levering - mister-auto.no

[Annonse] www.mister-auto.no/katalog/bildeler
3,6 ★★★★★ -vurdering for mister-auto.no
Skaff deg nye bildeler billig. Kjøp online med opp til 65% avslag!
Mobilvennlig webshop · 14 dagers angrerett · Autodeler til lavpris · Sikker betaling

Bildeler.no

<https://www.bildeler.no/>
Bildeler.no Audi VW Skoda. Et selskap i Komplet Group: Komplet.no MPX.no. TecDoc. Det er ikke tillatt å kopiere dataene, spesielt hele databasen. Det er ikke ...
[Velg bil](#) · [Kontakt oss](#) · [Om oss](#) · [Mellomaksel, oppheng](#)



Ads

Searching the Web



- Three forms of searching

1. **Specific queries** \Rightarrow encyclopaedia, libraries

- Exploit hyperlink structure

2. **Broad queries** \Rightarrow web directories

- Web directories: classify web documents by subjects

3. **Vague queries** \Rightarrow search engines

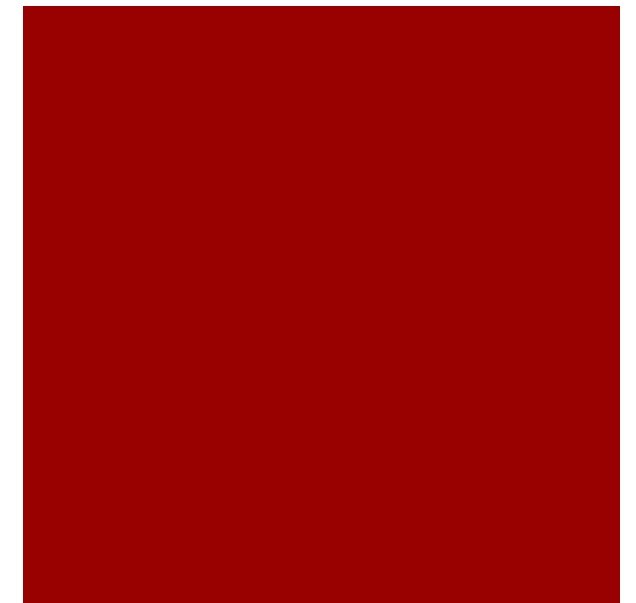
- index portions of web

Problem with the data

- **Distributed** data
- High percentage of **volatile data** – much is generated
- **Large** volume
 - June 2000 Google full-text index of 560 million URLs
- **Unstructured** data
 - gifs, pdf etc
- **Redundant** data
 - – mirrors (30% pages are near duplicates)
- **Quality** of data
 - false, poorly written, invalid, misspelt
- **Heterogeneous** data – media, formats, languages, alphabets

Search Engines

- Difference between standard IR systems
 - Only indices available, no text
- If insisting on text availability
 - **Keep local copy** of the Web pages
 - Too expensive
 - Access the remote Web pages
 - Too slow
- Architectures
 - **Centralized**
 - More popular
 - **Distributed**



alltheweb
• • • find it all • • •

Google

WebCrawler®

The Web's Top Search Engines Spun Together

Google

YAHOO!

bing

Ask



YAHOO!®

altavista™

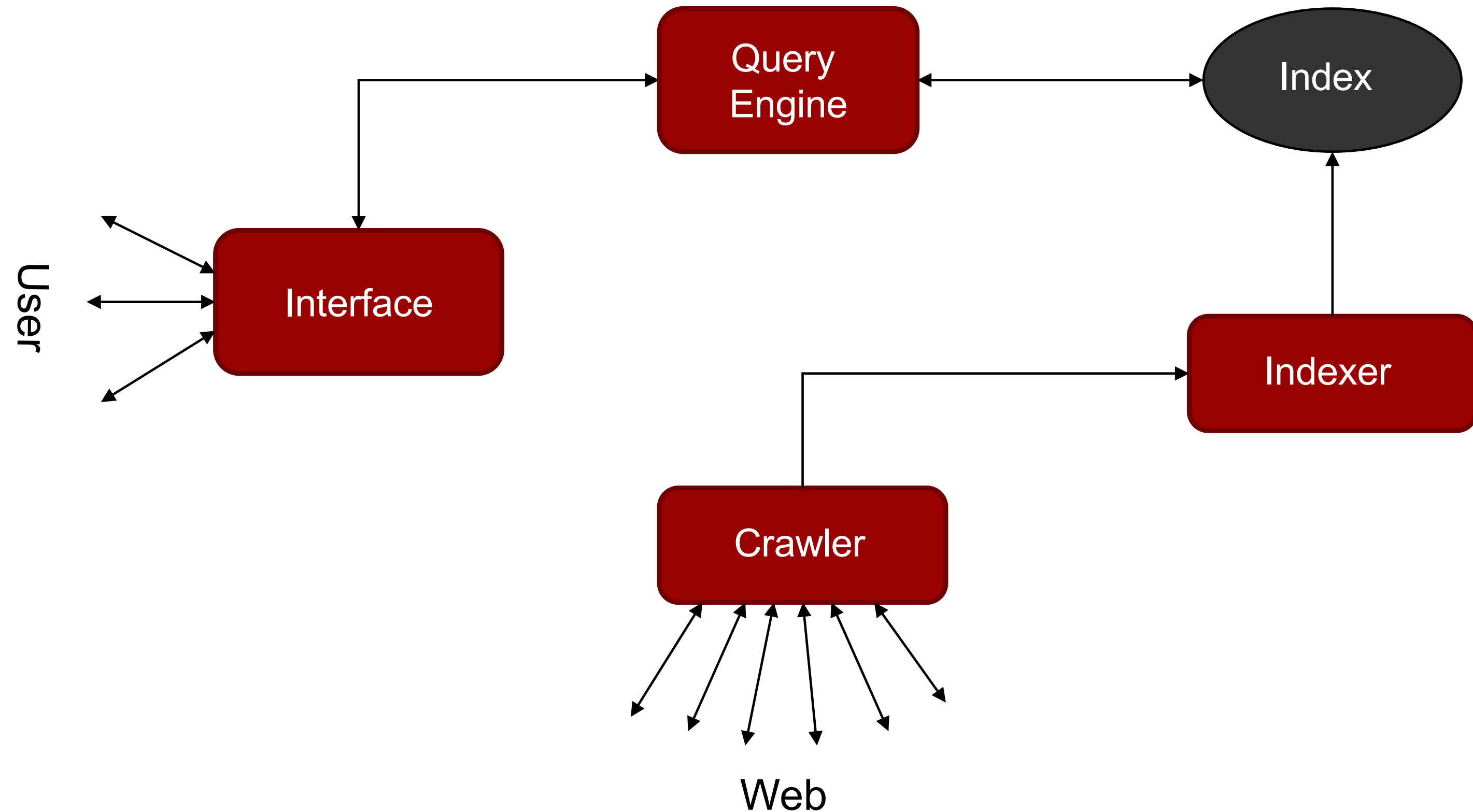
bing

Centralized Architecture

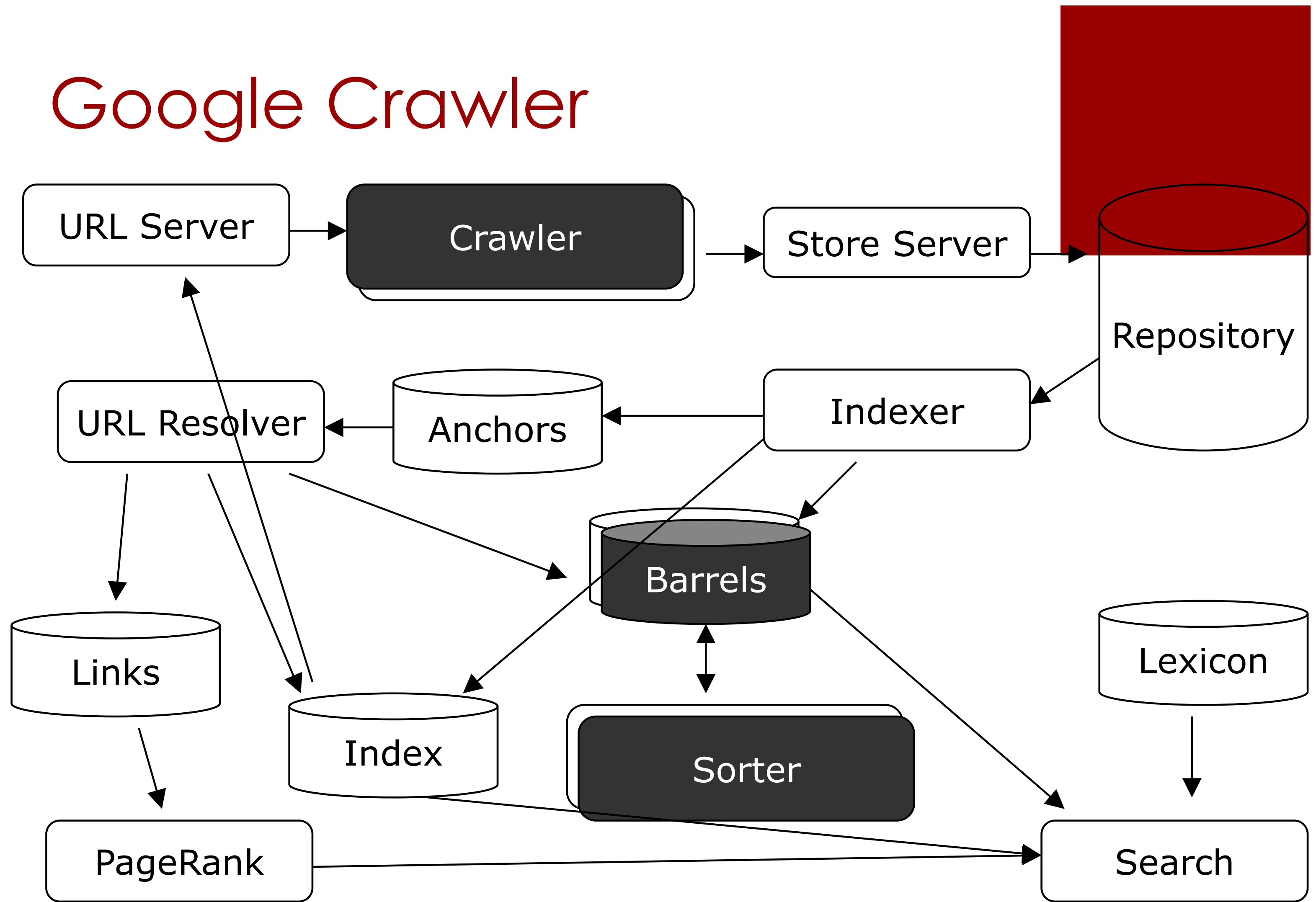
- Centralized crawler-indexer architecture
- Crawlers
 - Software agents that traverse the Web and send new or updated pages to the server
 - Aka, robots, spiders, wanderers, walkers, knowbots
 - Usually the code does not move to the remote machines
 - Just send the request to the remote machines



Centralized Architecture (2)



Google Crawler



Robots Exclusion Protocol

- <http://www.mydomain.com/robots.txt>

```
User-agent: *  
Disallow: /cgi-bin/  
Disallow: /tmp/  
Disallow: /~joe/
```

- HTML META Tag

```
<HTML>  
<HEAD>  
<META NAME="ROBOTS" CONTENT="NOINDEX, NOFOLLOW">  
</HEAD>  
...
```

Centralized Architecture (3)

- AltaVista architecture
 - Two parts: one for users, one for the indexing
 - 20 multiprocessor machines, 130 Gb of RAM, 500 Gb of disk
 - The query engine use most of these resources (70%)



Centralized Architecture (4)

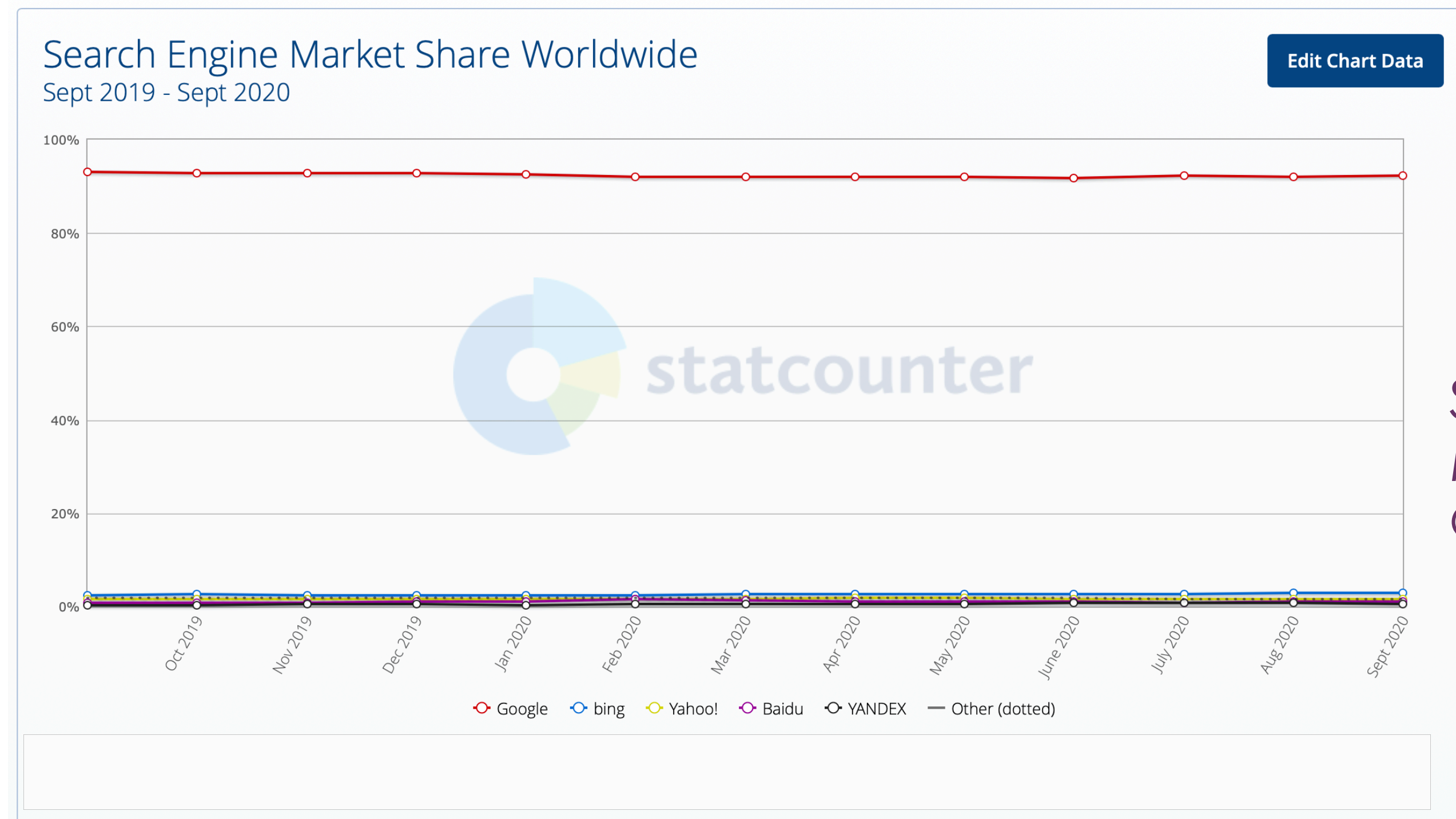
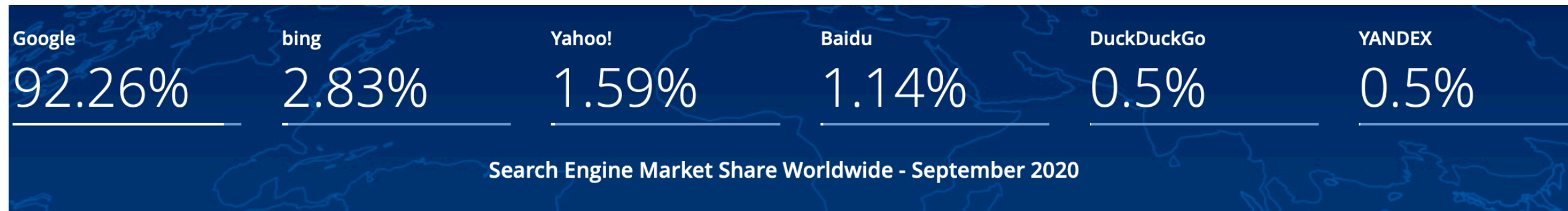
- Problem of the centralized architecture
 - Gathering of the data
 - Web is dynamic
 - Communication lines
 - High load at Web servers



Centralized Architecture (5)

- The largest search engines (in 1998!)
 - AltaVista > HotBot > Northern Light > Excite
 - in terms of Web coverage
 - They cover 28-55% (or 14-34%) of all Web pages (300 million) in 1998
 - Some are using the same internal engine
 - HotBot, GoTo, Microsoft by Inktomi
 - Magellan by Excite's
 - Largest today: Google

Global Share: Web Search



Search Engine
Market Share
Oct. 2020

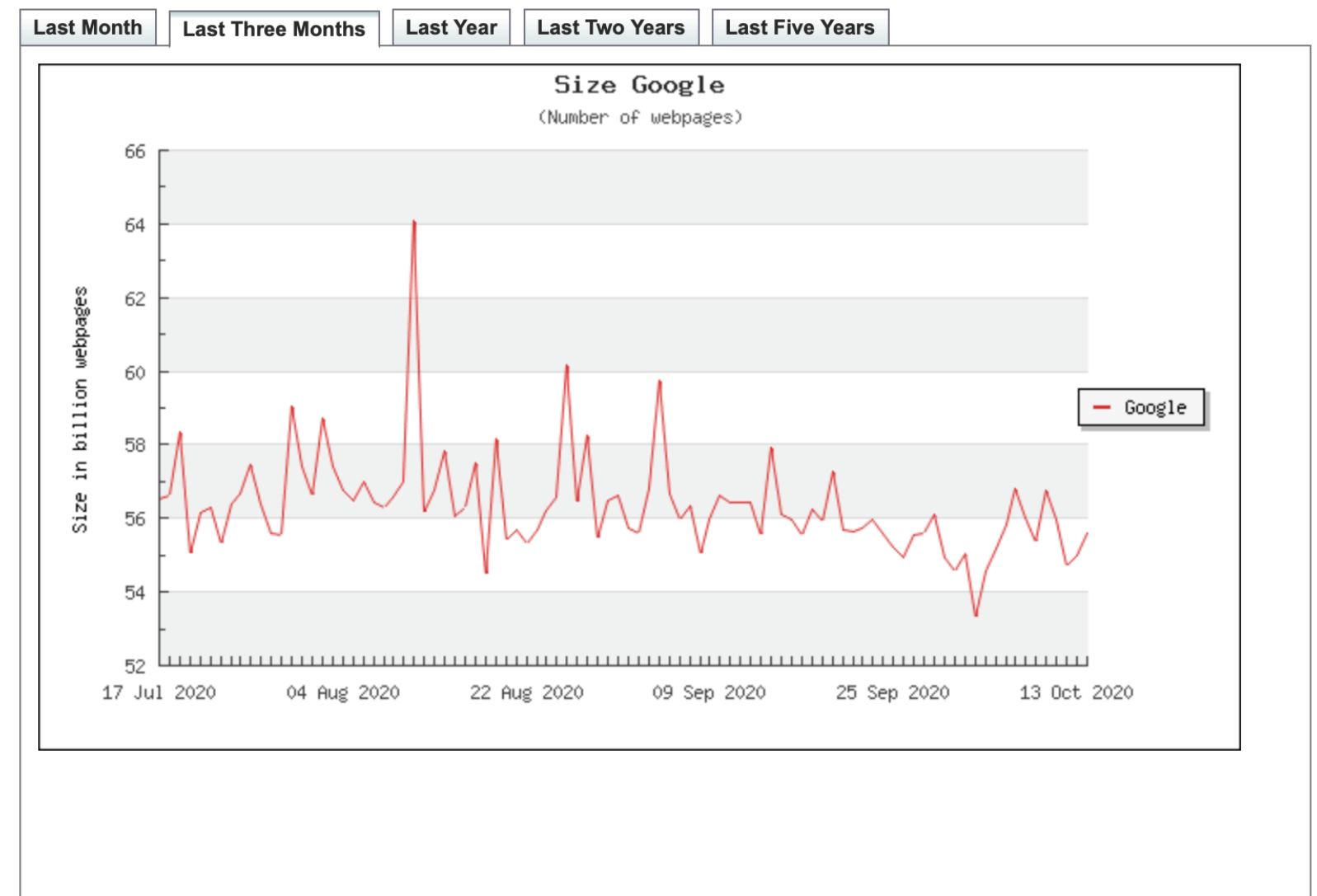
<http://gs.statcounter.com/search-engine-market-share>



- The best web engine:
 - comprehensive and relevant results
- Biggest index
 - > 5.5 billion pages visited and recorded
- Different kinds of index
 - smaller indexes containing a higher amount of the web's most popular pages, as determined by Google's link analysis system.
- Index refresh
 - Updated monthly/weekly
 - Daily for popular pages
- Serves queries from three data centres
 - two on West Coast of the US, one on East Coast.



**The size of the World Wide Web:
Estimated size of Google's index**



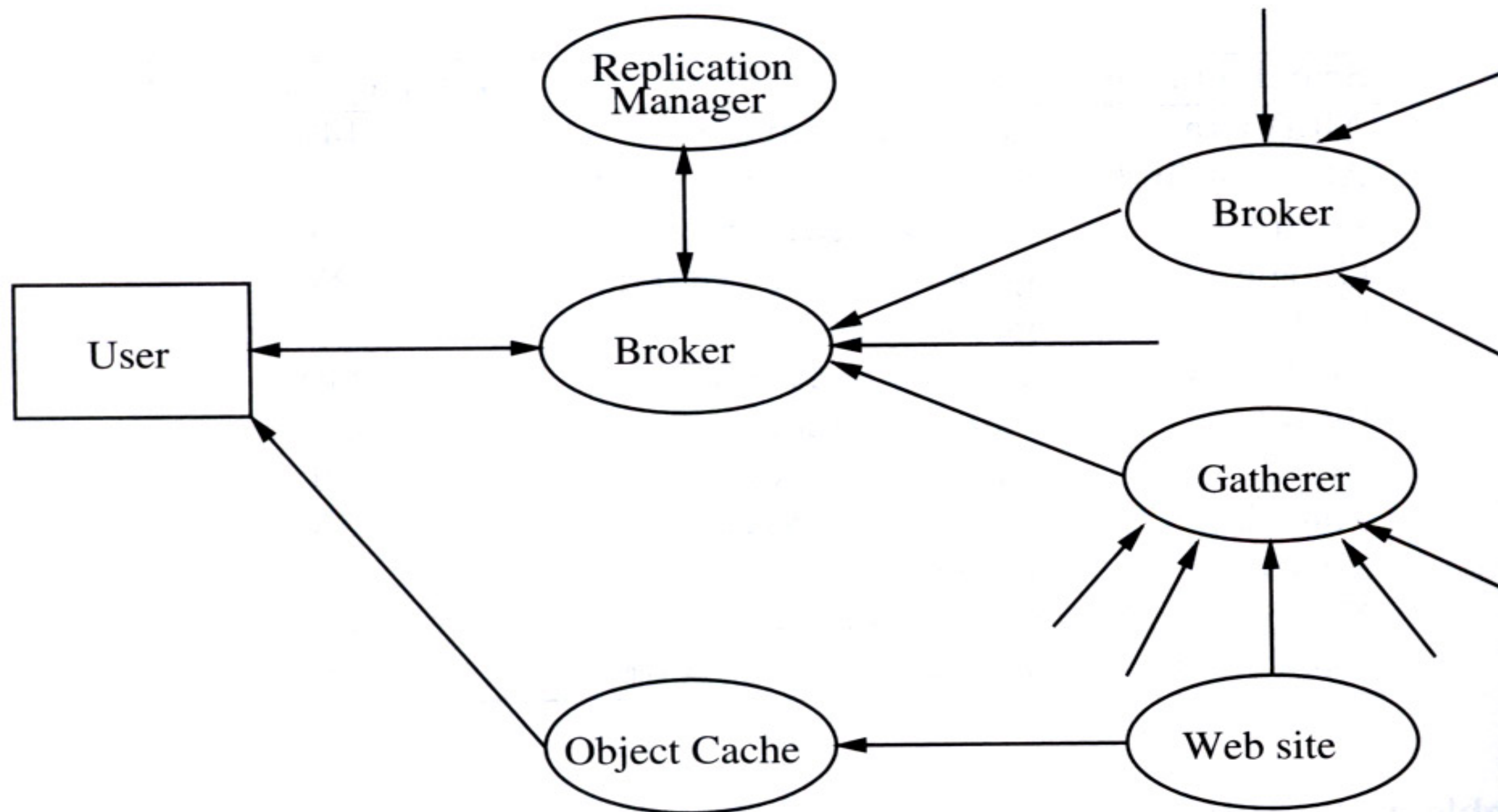
Centralized Architecture (6)

- Other search engines
 - Ask Jeeves! (<http://www.ask.com/>)
 - Simulate an interview
 - DirectHit (<http://www.directhit.com/>)
 - Ranks the answer Web pages in order of popularity
 - Search Broker (<http://webglimpse.org/sb/>)
 - Specific topics
 - No longer maintained...
 - DejaNews (www.dejanews.com)
 - Searches USENET archives
 - Bought by Google – became groups.google.com



Distributed Architecture

- Distributed architecture to gather and distribute data
 - E.g., Harvest
- Advantage
 - More efficient
- Disadvantage
 - Requires coordination of several Web servers
- Problems of crawler-indexer architecture
 - Web servers receive requests from different crawlers, increasing their load
 - Crawlers retrieve entire objects, but most of the content is discarded
 - Information is gathered independently by each crawler
 - No coordination



Distributed Architecture (3)



- Two elements in Harvest
 - **Gatherers**
 - Collect and extract indexing information from the Web servers
 - Gathering times are periodic
 - **Brokers**
 - Retrieve information from gatherers or other brokers
 - Update incrementally their indices
- Gatherers and brokers can communicate in flexible ways
 - From a gatherer to multiple brokers
 - Brokers to brokers

Distributed Architecture (4)



- Other features of Harvest
 - Topic specific brokers in Harvest
 - Can avoid the generic indices
 - E.g., size of the vocabulary
 - Registration brokers
 - Allow other brokers to register information about gatherers and brokers
 - Object caches
 - Reduce the network and server load

Distributed Architecture (5)



- About Harvest
 - Many Harvest applications
 - CIA, NASA, ...
 - Public domain:
 - <http://harvest.sourceforge.net/>
 - Commercial versions
 - Netscape's Catalog Server
 - Network Appliances' cache

User Interfaces

- Two things
 - Query interface
 - Answer interface
- Query Interface
 - The same word sequence in the query is regarded differently in different search engines
 - AltaVista: OR
 - Google: AND
 - Logical views of the text are different in different search engines
 - Stopwords, stemming, case sensitive

User Interfaces (2)

- Complex queries
 - AltaVista
 - Boolean query
 - Range query
 - HotBot
 - Boolean (i.e., all, any)
 - Must contain, must not contain
 - Published date range
 - Page associated media (image, audio, VRML, ...)
 - Location/domain (.edu, .com)
 - Page depth in a Web site
 - Word stemming
- NorthernLight
 - Words in title
 - Words in URL
 - Published date range
 - Web source
 - Journals, news, personal
 - Languages/Countries
 - subjects

User interface (3)

The screenshot shows the AlltheWeb.com Advanced Web search interface in a web browser window. The browser's address bar displays the URL `http://www.alltheweb.com/advanced?advanced=1`. The page features a navigation bar with links to Web, News, Pictures, Video, Audio, and FTP files. Below this, a blue banner reads "Advanced Search - Use the following filters to execute a more accurate search".

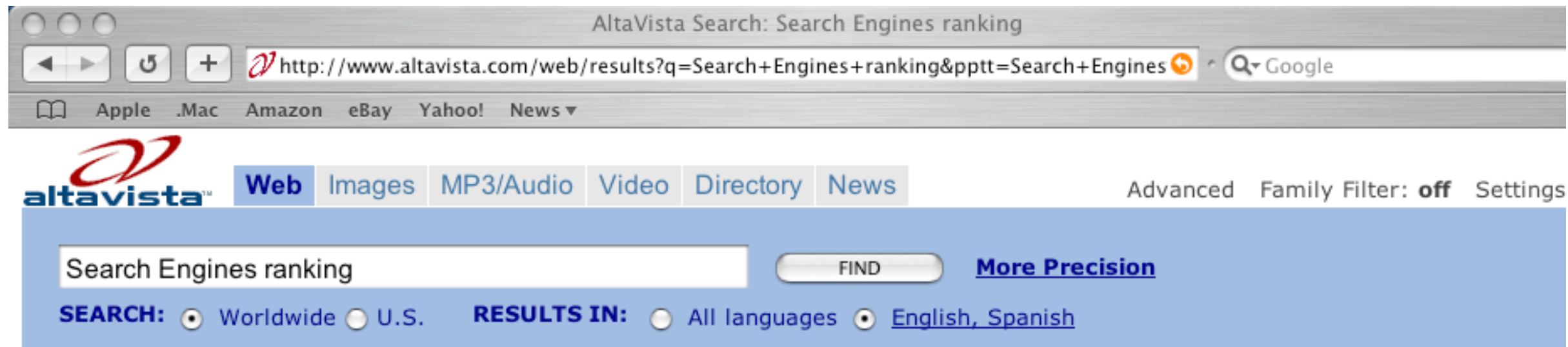
The main search area is titled "First select a type of search" and includes a "Query language guide" link. It offers two search methods: "Search for -" with a dropdown menu set to "all of the words" and a text input field, and "Boolean -" with instructions to "Create a boolean query using the operators **and**, **or**, **andnot** and **rank**." and a text input field with a "see examples" link. A "Search" button is located below these options.

Below the search options, a section titled "Use the following to include and exclude additional criteria" contains several filter sections:

- Language -** Find results written in: Preferred (dropdown), Unicode (UTF-8) (dropdown).
- Word Filters -** Includes three rows of filters: "Must include", "Should include", and "Must not include". Each row has a text input field and a dropdown menu set to "in the text". A "+ Add a filter" link is at the bottom.
- Domain Filters -** Filter results from specific domains (com, gov, dell.com, etc.). It includes fields for "Include results from" and "Exclude results from", a section for "Find results from a specific geographic region", and a dropdown menu for "Only find results from" set to "Any".
- IP-address Filters -** Only find results from the following IP-address(es) and/or range(s). It includes a text input field.
- Media Types -** Find results that contain the following media types. It is divided into two columns, each with "Include - Exclude" headers. The first column lists Images, Audio (midi, wav, au), Video (mov, qt, avi), and RealVideo & RealAudio. The second column lists Macromedia Flash, Java applets, JavaScript, and VBScript. Each item has an "Include" checkbox and an "Exclude" checkbox.

The browser's status bar at the bottom shows the "File format" dropdown menu.

User interface (4)



Sponsored Matches [About](#)

[Top Search Engine Ranking](#)

SEO Inc. has placed more Web sites in high **search** engine rankings than virtually any other SEO company on the planet.

www.seoinc.com

[Rank Top 3 from \\$99 a Month Guaranteed](#)

Get top three rank on for your Web site on Yahoo, Netscape, Altavista, Lycos, Excite and 13 other top **search engines**, from \$99 a month with no set-up or hidden fees.

1-2-3webposition.com

[Top Rank on Yahoo! or Google \\$39.95 Week](#)

Top three rank on Yahoo! and 13 others, or first page on Google, AOL and two others, \$39.95 flat weekly fee - no setup or other fees, no long-term commitment. Guaranteed rankings.

www.hiposition.com

[Professional PPC Bid Management Service](#)

Increase Web site traffic, save time and money. Manage bids on multiple PPC **search engines**, track visitors and conversions, and review Web analytics in one interface. 14-day free trial.

www.gotoast.com

AltaVista found 253,653 results [About](#)

[SUBMIT EXPRESS search engine optimization free url submission ranking](#)

... URL to 40 major **search engines** for FREE ! Search ... and give you **ranking** advice before submission. Once submitted we will email you a report of which **search engines**

Refine your search with AltaVista Prisma

[?]

Click a term to focus your search. Click >> to replace your search.

[[Go Back](#)]

[Engine Optimization](#) >>

[Major Search Engines](#) >>

[Search Engine Optimization](#) >>

[Search Engine Placement](#) >>

[Search Engine Positioning](#) >>

[Search Engine Ranking](#) >>

[Search Engine Submission](#) >>

[Top Search Engines](#) >>

[Meta Tags](#) >>

[Pay Per Click](#) >>

[Web Site Promotion](#) >>

[Website Promotion](#) >>

User interface (6)

Avansert søk

Finn sider ...

med alle disse ordene:

med nøyaktig dette ordet eller uttrykket:

med noen av disse ordene:

uten noen av disse ordene:

med tall fra:

til

Gjør dette direkte i søkefeltet:

Skriv inn de viktige ordene: golden retriever valp

Sett eksakte ord i anførselstegn: "rat terrier"

Skriv inn OR mellom alle ordene du vil bruke: miniatyr OR standard

Sett inn et minustegn rett før ord du ikke vil bruke:
-gnager, - "Jack Russell"

Sett to punktumer mellom tallene, og legg til en måleenhet:
10..35 lb, \$300..\$500, 2010..2011

Deretter kan du avgrense resultatene dine etter ...

språk:

alle språk

Finn sider på det språket du velger.

område:

en hvilken som helst region

Finn sider som er publisert i et bestemt område.

siste oppdatering:

når som helst

Finn sider som er oppdatert innenfor tiden du oppgir.

nettsted eller domene:

Søk på et bestemt nettsted (som f.eks. wikipedia.org), eller avgrens søket til et domene som for eksempel .edu, .org eller .gov

ord som vises:

hvor som helst på siden

Søk etter ord på hele siden, i sidens tittel eller i nettadressen, eller søk etter linker til siden du leter etter.

[Sikkert Søk:](#)

Vis de mest relevante resultatene

Gi beskjed til [Sikkert Søk](#) om seksuelt eksplisitt innhold skal filtreres.

filtype:

alle formater

Finn sider i det formatet du foretrekker.

[bruksrettigheter:](#)

ikke filtrert etter lisens

Finn sider du står fritt til å bruke selv.

Avansert søk

Du kan også

[Finn sider som ligner, eller som viser til, en nettadresse](#)

[Søk i sider du har besøkt](#)

[Bruk operatører i søkefeltet](#)

[Tilpass søkeinnstillingene](#)



User Interfaces (6)

- Answer Interfaces
 - Usually top ten ranked Web pages
 - Associated with URL, size, date, part of content
 - Typically ordered by relevance
 - Some order by URL, or date
 - Option to find similar documents for each document
 - Can expand the query

Ranking

- Use variations of Boolean or vector model
- Difficulties
 - Ranking just with indices not with text
 - Ranking algorithms are usually proprietary
 - Impossible to measure recall



Ranking (2)

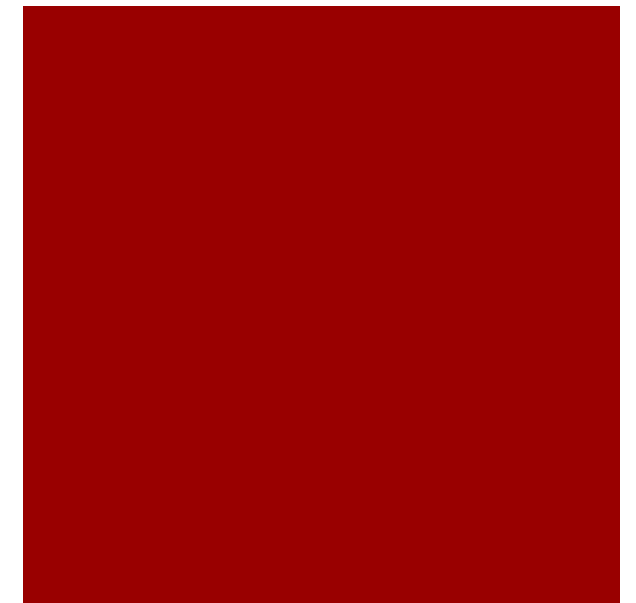
- Three ranking algorithms by [Yuwono and Lee]
 - **Boolean spread**
 - Based on (extended) Boolean model extended to pages in/out linked by the answer set
 - **Vector spread**
 - Same with vector model (tf-idf)
 - **Most-cited**
 - Based on the terms in pages pointing to the answer set
- Vector model yielded the best
 - average precision was 70%

Ranking (3)

- Ranking using hyperlink information
 - The number of incoming hyperlinks to a page represent the popularity or quality of the page
 - Three approaches
 - WebQuery
 - HITS
 - PageRank
- **WebQuery [Li]**
 - First take an answer set to a query
 - Rank the page by analyzing how the pages are connected
 - Extend the answer set by adding pages that are highly connected to the original set.

WebQuery [Li]

- First take an answer set to a query
- Rank the page by analyzing how the pages are connected
- Extend the answer set by adding pages that are highly connected to the original set.



HITS [Kleinberg]

- S : set of pages that are one link adjacent (in/out) **with the answer set**
- **Authorities**: pages that have high incoming links (in S)
- **Hubs**: pages that have high outgoing links (in S)
- Hub and authority value of a page p
 - $H(p) = \sum_{u \in S \mid p \rightarrow u} A(u)$
 - $A(p) = \sum_{v \in S \mid v \rightarrow p} H(v)$
- Consider the link weight and page score
 - Link weight:
 - similarity between the surrounding content and the query
 - Page score
 - similarity between the page content and the query
- Result: better precision

PageRank

- Developed and part of Google
- q : probability that a user randomly jump to the page
- $1 - q$: probability that a user visit the page following a link
- $L(a)$: number of outgoing links of page a
- p_1, \dots, p_n : pages pointing page a
- Page rank

$$PR(a) = \frac{q}{T} + (1 - q) \sum_{i=1}^n \frac{PR(p_i)}{L(p_i)}$$

- T total # pages on the web graph, q is typically 0.15
- $PR(p_i)$ is normalized by $L(p_i)$

Ranking (cont'd)

- How to make a page highly ranked by search engines
 - Include informative titles, headings, meta fields
 - Include good links
 - Don't repeat keywords
 - Some search engines penalize
 - Select terms that directly represent the subject



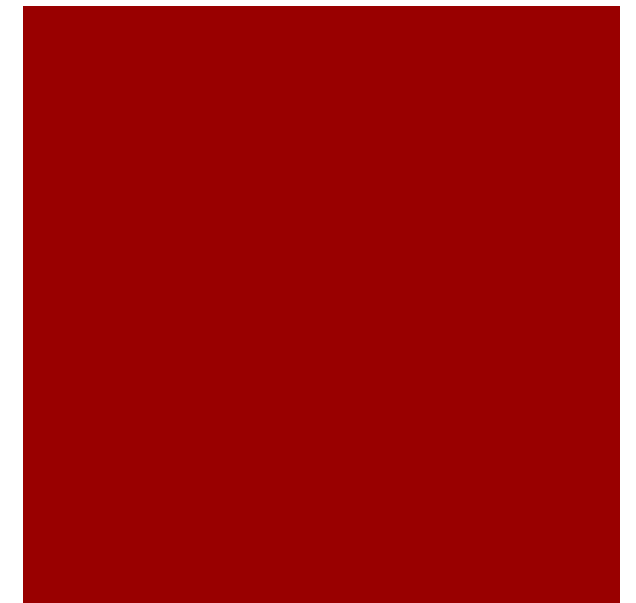
Crawling the Web

- Start with a set of URLs (or popular URLs)
 - Extract other URLs recursively in breadth-first or depth-first
 - Can pre-determine the depth of a web site
 - Allow users to submit top Web site to be added to the initial set
 - If multiple crawlers are used, difficult to coordinate not to visit the same pages
 - Solution: partition the web by countries or internet names, assign robots by the partition

Crawling the Web (2)

- How frequently the index is updated
 - How old is an index to a Web page
 - Varies a lot: One day to two months
 - Stars in the sky view
 - Percentage of invalid links: 2-9%
 - Usually put the index date for each page in the answer set
 - Some search engine learn the change frequency of a page and visit accordingly
 - Visit the popular pages (high incoming links) more frequently
 - State-of-art performance: traverse 10 million Web pages a day

Crawling the Web (3)



- The order of traversing
 - Breadth-first
 - Good for web site that are structured by related topics
 - Coverage is wide but shallow
 - A web server is bombarded with many rapid requests
 - Depth-first
 - Narrow but deep traversal
- Better pages first
 - e.g., using PageRank

Crawling the Web (4)

- High network traffic problem caused by robots
 - A set of guidelines have been developed
 - A special file is place at the root of each Web server
 - Indicate the restrictions at that site
 - Some pages that should not be indexed
- Pages that crawlers cannot index
 - HTML with frames, image maps
 - Dynamically generated pages, password protected pages
 - Flash-based pages

Indices

- Use variants of inverted file
 - Elements in the occurrence list are pages
- Size of the index
 - a short description of a Web page
 - Creation date, size, title, first lines
 - e.g., size of the description = 500 bytes (URL, the description) * 100 million pages = 50 Gb
 - State of art: index size is 30% of the text
 - 150 Gb for 100 million pages

Indices (2)

- Main memory
 - Usually keep the whole answer set although the initial is few
- Full inversion
 - Occurrences are word positions in a document
 - Costly but can easily support phrase and proximity queries
- Using the fixed size logical blocks
 - Reduce the size of the pointers
 - # of blocks < # pages
 - Reduce the number of pointers
 - Infrequent words tend to cluster in the same block
 - Used by Glimpse (Harvest)

Web Directories

- Web coverage provided by directories
 - Less than 1% of all Web pages
 - Answers sets are usually much more relevant
- Examples of Web directory
 - Yahoo!: the oldest. 750,000 Web pages classified
 - eBLAST, LookSmart, Magellan, NewHoo
 - Some also provide the search engine

Web Directories - Yahoo

The screenshot shows the Yahoo! Directory homepage in a web browser window. The browser's address bar displays 'http://dir.yahoo.com/'. The page features a search bar with a dropdown menu showing 'the Web' and 'the directory' (selected). Below the search bar, there are several category links: Business & Economy, Computers & Internet, News & Media, Entertainment, Recreation & Sports, Health, Government, Regional, Society & Culture, Education, Arts & Humanities, Science, Social Science, and Reference. A 'NEW ADDITIONS' section is also visible. On the right side, there are sections for 'YAHOO! PICKS' and 'BUZZ MOVERS'. The page is powered by HP, as indicated by the logo.

Yahoo! Directory

http://dir.yahoo.com/

Apple .Mac Amazon eBay Yahoo! News

YAHOO! directory

Directory Home

Search ☐ the Web ☒ the directory Search

Advanced Search | Suggest a Site

Business & Economy
[B2B](#), [Finance](#), [Shopping](#), [Jobs](#)...

Computers & Internet
[Internet](#), [WWW](#), [Software](#), [Games](#)...

News & Media
[Newspapers](#), [TV](#), [Radio](#)...

Entertainment
[Movies](#), [Humor](#), [Music](#)...

Recreation & Sports
[Sports](#), [Travel](#), [Autos](#), [Outdoors](#)...

Health
[Diseases](#), [Drugs](#), [Fitness](#)...

Government
[Elections](#), [Military](#), [Law](#), [Taxes](#)...

Regional
[Countries](#), [Regions](#), [U.S. States](#)...

Society & Culture
[People](#), [Environment](#), [Religion](#)...

Education
[College and University](#), [K-12](#)...

Arts & Humanities
[Photography](#), [History](#), [Literature](#)...

Science
[Animals](#), [Astronomy](#), [Engineering](#)...

Social Science
[Languages](#), [Archaeology](#), [Psychology](#)...

Reference
[Phone Numbers](#), [Dictionaries](#), [Quotations](#)...

YAHOO! PICKS
[Taking the Long View](#) - wide and wonderful.
[read review](#) - [more picks](#)

BUZZ MOVERS
1 [Timothy Treadwell](#)
2 [Wyatt Earp](#)
3 [Carla Gugino](#)
4 [Atlantic Monthly](#)
5 [Maria Shriver](#)
[more](#)

ASK YAHOO!
[Why do people hang pairs of shoes from power lines?](#)
[more](#)

NEW ADDITIONS
[Thu Oct 9](#) - [Wed Oct 8](#) - [Tue Oct 7](#) - [Mon Oct 6](#) - [Sun Oct 5](#) - [Sat Oct 4](#) - [Fri Oct 3](#)

Yahoo! Directory

YAHOO!
DIRECTORY

Owner	Yahoo!
Launched	January 1994; 23 years ago
Current status	Defunct as of December 26th 2014

Web Directories (2)

- Categories in Web directories
 - Hierarchical taxonomies that classify human knowledge
 - E.g., Yahoo!
- Pages are submitted to the Web directory
 - Then reviewed, accepted, classified
- Taxonomy as a acyclic graph
 - Instead of tree
 - There are cross reference

Web Directories (3)

- Advantages of Web directories
 - The answer is useful in most cases
- Disadvantages
 - The classification is not specialized enough
 - Not all Web pages are classified
- Automatic classification
 - Using clustering
 - Use NLP to extract all relevant terms
- Manual classification
 - By a limited number of people

Combining Searching and Browsing



- Browsing
 - Following hyperlinks
- Searching
 - Running search engine
- WebGlimpse's approach
 - Attach a small search box to the bottom of every Web page
 - Can cache the neighbors

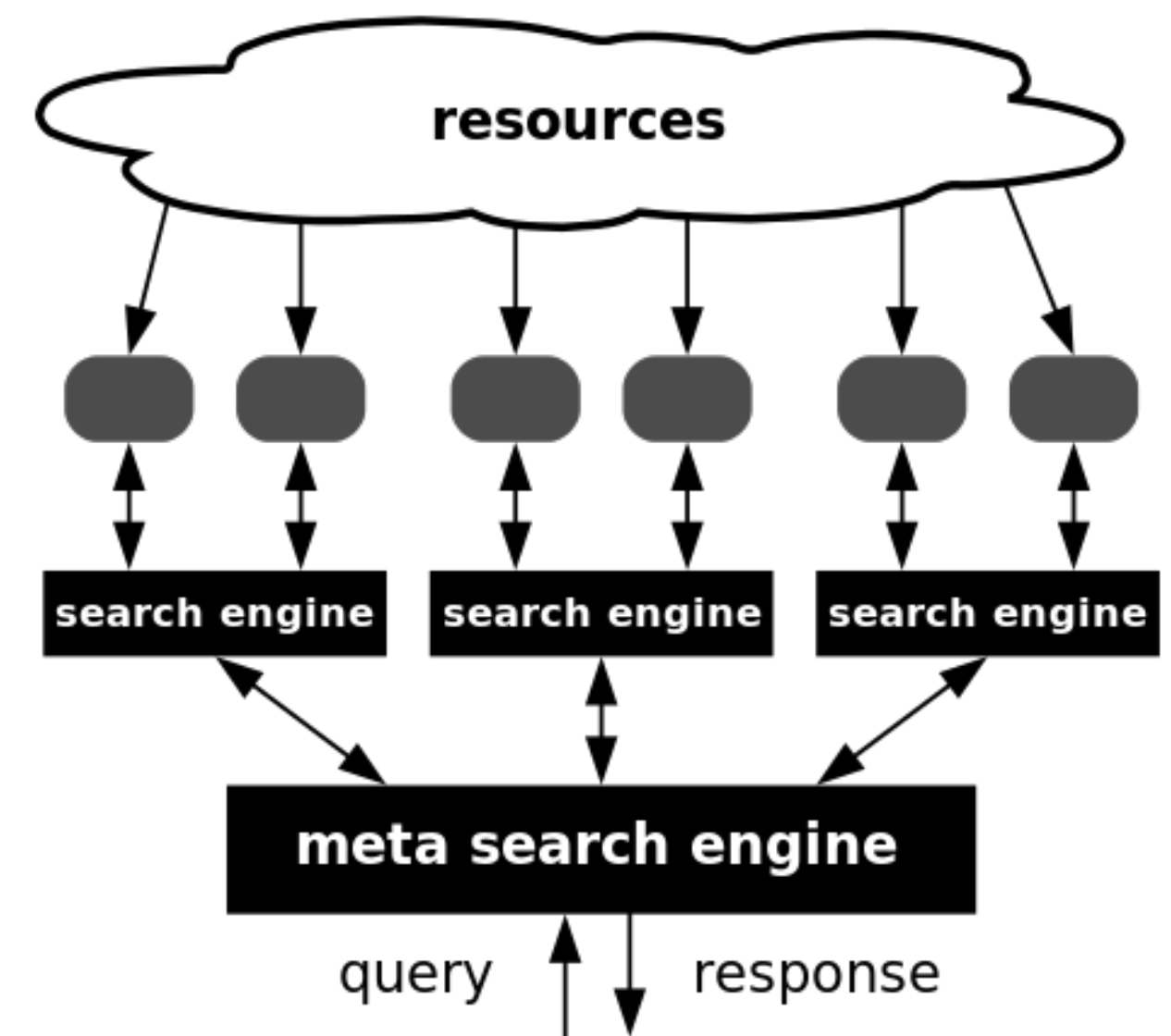
Combining Searching and Browsing (2)



- Helpful tools
 - Add-ons to browsers
 - Alexa: current site's popularity, speed of access, overall quality. Suggest related sites
 - WebTaxi
 - Visualization tools for Web
 - Microsoft's SiteAnalyst, Dynamic Diagrams' MAPA, IBM's Mapuccino, ...
 - Visualization tools for large answers

Meta-searchers

- What is meta-searchers
 - Send a query to several search engines, Web directories, and other databases
 - Collect answers and unify them
- Examples
 - Metacrawler, SavvySearch
- Advantages
 - Many sources
 - User has the single interface



User Problems

- User problems
 - Novice users don't know how to start and get better answer
 - Don't know the logical view of text used by the system
 - e.g., case sensitive or not
 - Typos or variations of words
 - About 10-20% of matches are lost
 - Foreign names: 50% of matches are lost
 - Boolean logic
 - Use AND or OR differently
 - e.g., when choosing between two things, user mean 'exclusive or'

User Problems (2)

- Search engine problems
 - Slow; the answer is too large, not relevant, not always up to date
- Analysis on user query logs [1998]
 - Main purposes of query are research, leisure, business, and education
 - Most users don't care about advertising
 - 25% of users use single word in the query
 - Average query length: 2.35
 - 80% of users do not modify query
 - 85% of users look at only the first screen