Search Engines

COMP6218

Finding Information on the Web

- The Web is full of information
- How can information be found?
  - Browse strategies. Where is information stored? e.g. BBC News / Asia or ACM Digital Library/Web conference/2006
  - Search strategies. What does the information contain? A key phrase? Someone’s name?

Searching the Web

- Three forms of searching
  - Specific queries ⇒ encyclopaedia, libraries
    - Exploit hyperlink structure
  - Broad queries ⇒ web directories
    - Web directories: classify web documents by subjects
  - Vague queries ⇒ search engines
    - index portions of web

Problem with the data

- Distributed data
- High percentage of volatile data
- 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, mis-spelt
- Heterogeneous data – media, formats, languages, alphabets

Users and the Web

- How to specify a query?
- How to interpret answers?
  - Ranking
  - Relevance selection
  - Summary presentations
  - Large document presentation
- Main purpose: research, leisure, business, education
  - 80% do not modify query
  - 85% look first screen only
  - 64% queries are unique
  - 25% users use single keywords

Web search

- All queries answered without accessing texts – by indices alone
  - Local copies of web pages expensive (Google cache)
  - Remote page access unrealistic
- Links
  - Link topology, link popularity, link quality, who links
- Page structure
  - Words in heading > words in text etc
- Sites
  - Sub collections of documents, mirror site detection
- Names
- Presenting summaries
- Community identification
- Indexing Refresh rate
- Similarity engine
- Ranking scheme
- Caching and popularity measures

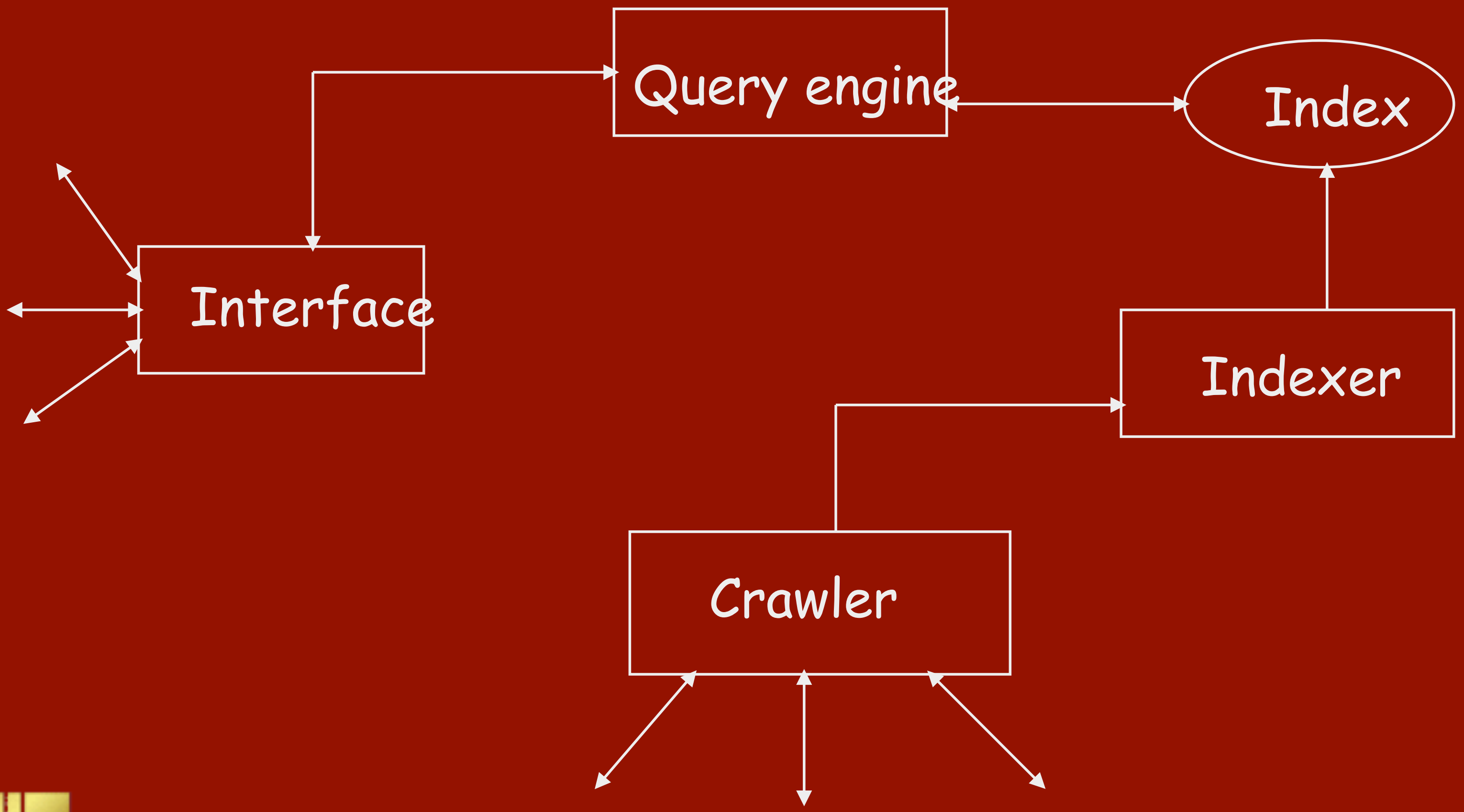
Spamming

- Most search engines have rules against
  - invisible text,
  - meta tag abuse,
  - heavy repetition
  - "domain spam”
    - overtly submission of "mirror“ sites in an attempt to dominate the listings for particular terms

Centralised architecture

- Crawler-indexer (most search engine)
- Crawler
  - Robot, spider, wanderer, walker, knowbot
  - Program that traverses web to send new or update pages to main server (where they are indexed)
  - Run on local server and send request to remote servers
- Centralised use of index to answer queries

Example (AltaVista)

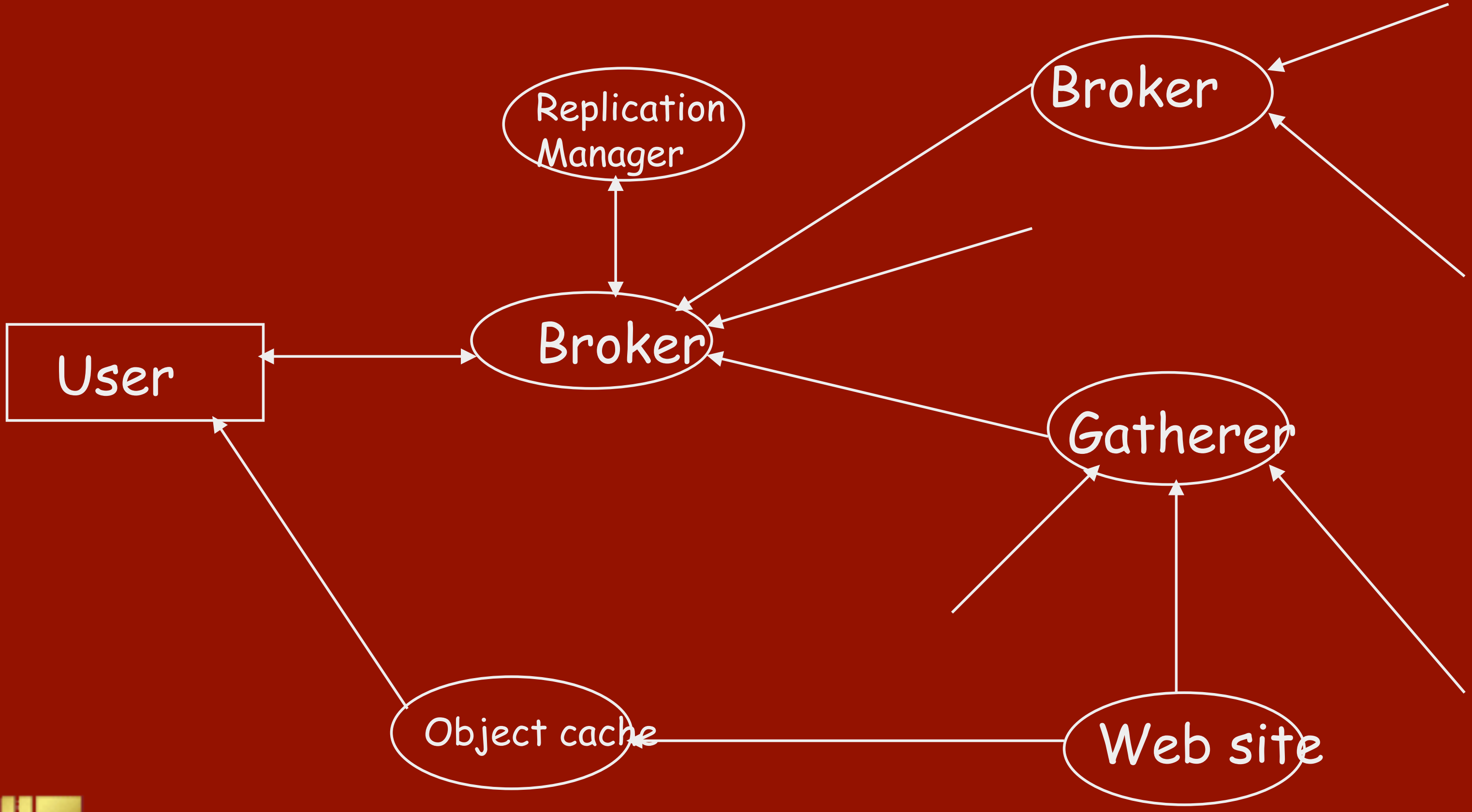


1998: 20 multi-processor machines, 130 GB of RAM, 500 GB disk space
2003: Estimated 10,000 servers (Google)
2008: Estimated 200,000 servers (Google)

Distributed architecture

- Harvest: harvest.transarc.com
- Gatherers:
  - Collect and extract indexing information from one or more web servers at periodic time
- Brokers
  - Provide indexing mechanism and query interface to data gathered
  - Retrieve information from gatherers or other brokers, updating incrementally their indices

Harvest architecture



Google Crawling

- Submission:
  - Add URL page (no need to do a "deep" submit)
  - Best way to ensure that your site is indexed is to build links.
- Crawling and Index Depth:
  - aims to refresh its index on a monthly basis,
  - if Google doesn't actually index a pages, it may still return it in a search because it makes extensive use of the text within hyperlinks.
  - This text is associated with the pages the link points at, and it makes it possible for Google to find matching pages even when these pages cannot themselves be indexed.

Ranking algorithms

- Variations of Boolean and vector space model
  - TF × IDF *term frequency x inverse document frequency*
- Hyperlinks between pages
  - pages pointed to by a retrieved page
  - pages that point to a retrieved page

  - Popularity: number of hyperlinks to a page
  - Relatedness: number of hyperlinks common in pages or pages referenced by same pages
    - PageRank (Google)
    - HITS (Clever)

Google Relevancy (1)

- Google ranks web pages based on the number, quality and content of *links pointing at them (citations)*.
- Number of Links
  - All things being equal, a page with more links pointing at it will do better than a page with few or no links to it.
- Link Quality
  - Numbers aren't everything. A single link from an important site might be worth more than many links from relatively unknown sites.

Google Relevancy (2)

- Link Content
  - The text in and around links relates to the page they point at. For a page to rank well for "travel," it would need to have many links that use the word travel in them or near them on the page. It also helps if the page itself is textually relevant for travel
- Ranking boosts on text styles
  - The appearance of terms in bold text, or in header text, or in a large font size is all taken into account. None of these are dominant factors, but they do figure into the overall equation.