

Search Engines: Crawling the Web

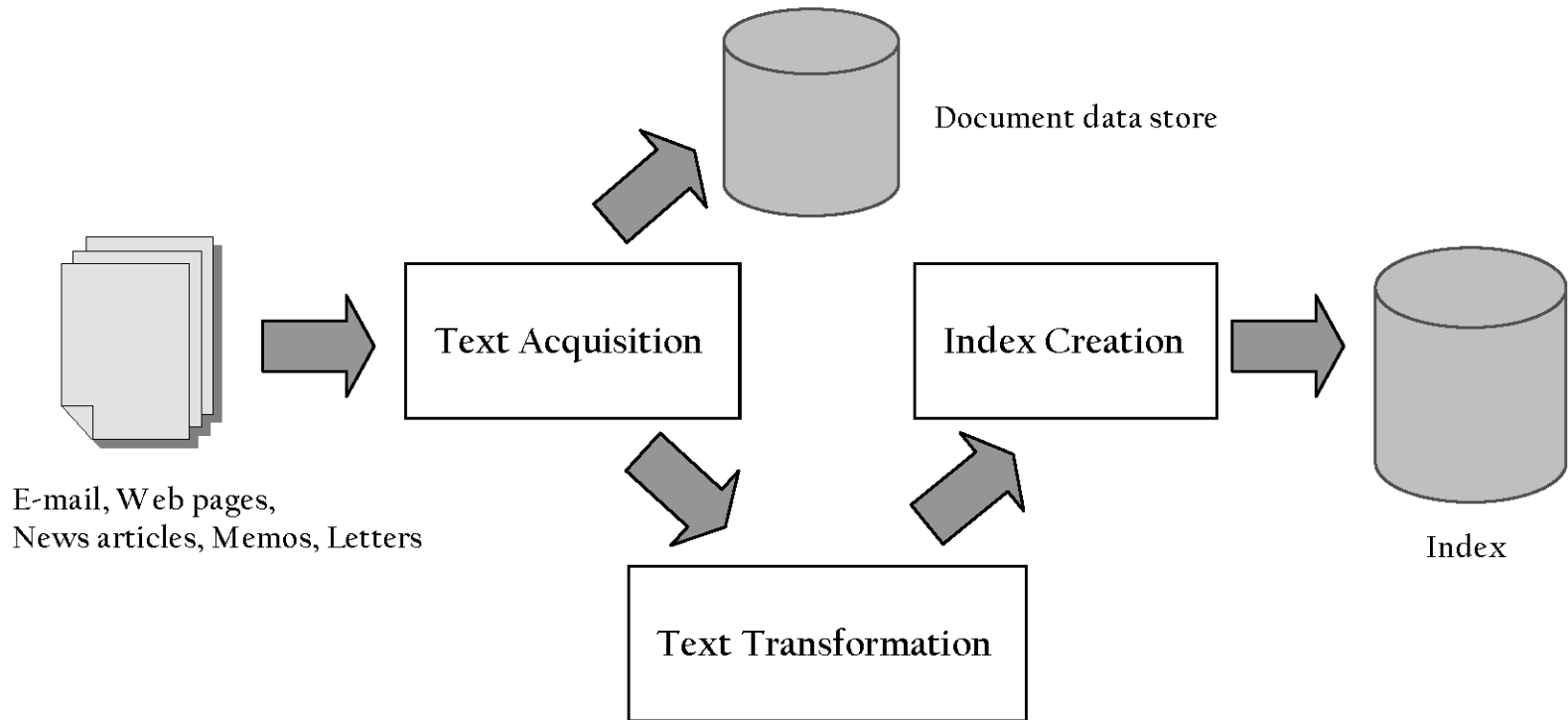
Elena Cabrio et Serena Villata
(equipe WIMMICS)



Search Engine Architecture

- A software architecture consists of software components, the interfaces provided by those components, and the relationships between them
 - describes a system at a particular level of abstraction
- Architecture of a search engine determined by 2 requirements
 - effectiveness (quality of results) and efficiency (response time and throughput)

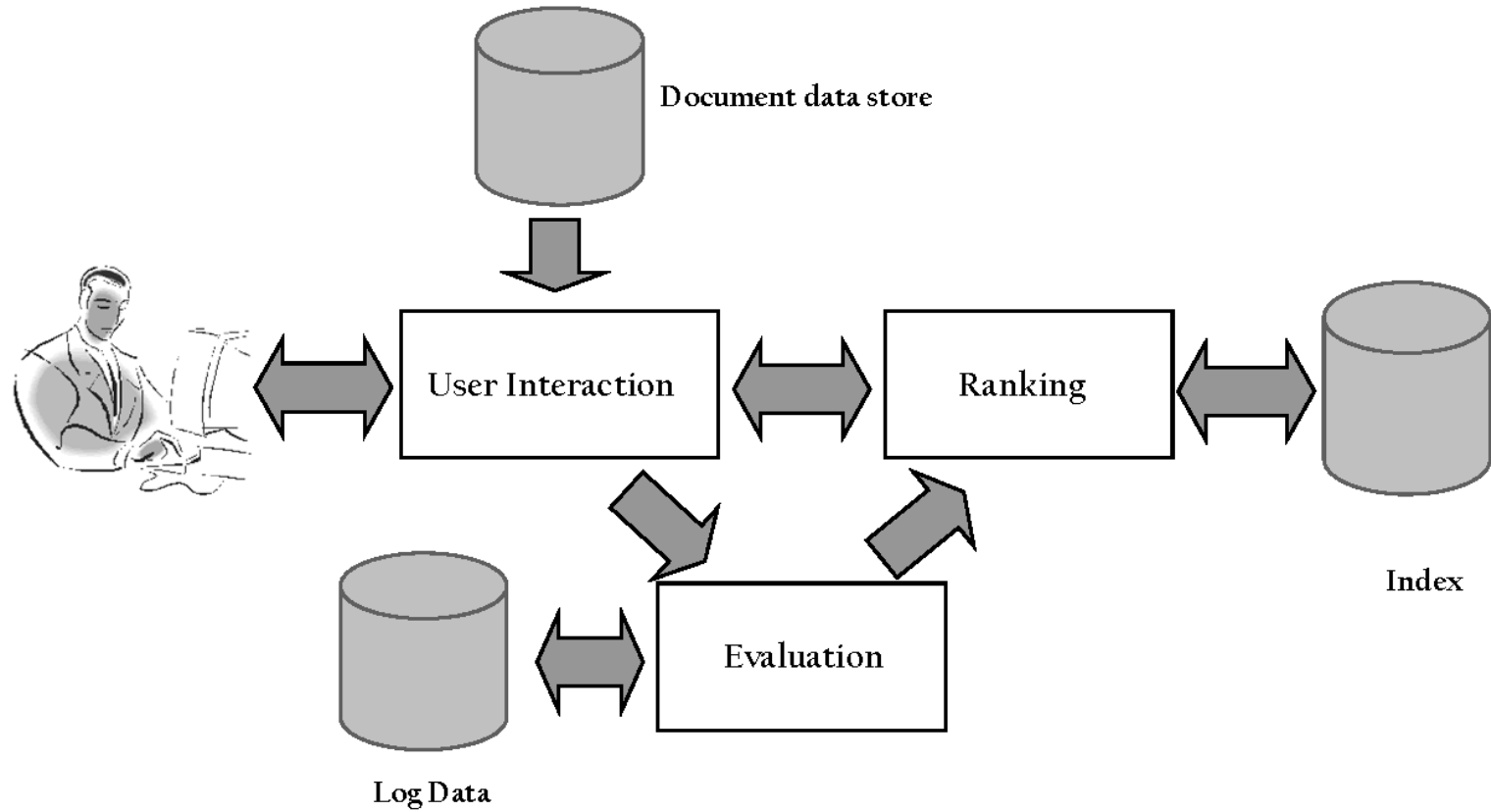
Indexing Process



Indexing Process

- Text acquisition
 - identifies and stores documents for indexing
- Text transformation
 - transforms documents into *index terms* or *features*
- Index creation
 - takes index terms and creates data structures (*indexes*) to support fast searching

Query Process



Query Process

- User interaction
 - supports creation and refinement of query, display of results
- Ranking
 - uses query and indexes to generate ranked list of documents
- Evaluation
 - monitors and measures effectiveness and efficiency (primarily offline)

Details: Text Acquisition

- Crawler
 - Identifies and acquires documents for search engine
 - Many types – web, enterprise, desktop
 - Web crawlers follow *links* to find documents
 - Must efficiently find huge numbers of web pages (*coverage*) and keep them up-to-date (*freshness*)
 - Single site crawlers for *site search*
 - *Topical* or *focused* crawlers for vertical search
 - *Document* crawlers for enterprise and desktop search
 - Follow links and scan directories

Text Acquisition

- Feeds
 - Real-time streams of documents
 - e.g., web feeds for news, blogs, video, radio, tv
 - RSS is common standard
 - RSS “reader” can provide new XML documents to search engine
- Conversion
 - Convert variety of documents into a consistent text plus metadata format
 - e.g. HTML, XML, Word, PDF, etc. → XML
 - Convert text encoding for different languages
 - Using a Unicode standard like UTF-8

Text Acquisition

- Document data store
 - Stores text, metadata, and other related content for documents
 - Metadata is information about document such as type and creation date
 - Other content includes links, anchor text
 - Provides fast access to document contents for search engine components
 - e.g. result list generation
 - Could use relational database system
 - More typically, a simpler, more efficient storage system is used due to huge numbers of documents

Ranking

- Scoring
 - Calculates scores for documents using a ranking algorithm
 - Core component of search engine
 - Basic form of score is $\sum q_i d_i$
 - q_i and d_i are query and document term weights for term i
 - Many variations of ranking algorithms and retrieval models

Web Crawler

- Finds and downloads web pages automatically
 - provides the collection for searching
- Web is huge and constantly growing
- Web is not under the control of search engine providers
- Web pages are constantly changing
- Crawlers also used for other types of data

Retrieving Web Pages

- Every page has a unique *uniform resource locator* (URL)
- Web pages are stored on web servers that use HTTP to exchange information with client software
- e.g.,

http://www.cs.umass.edu/csinfo/people.html

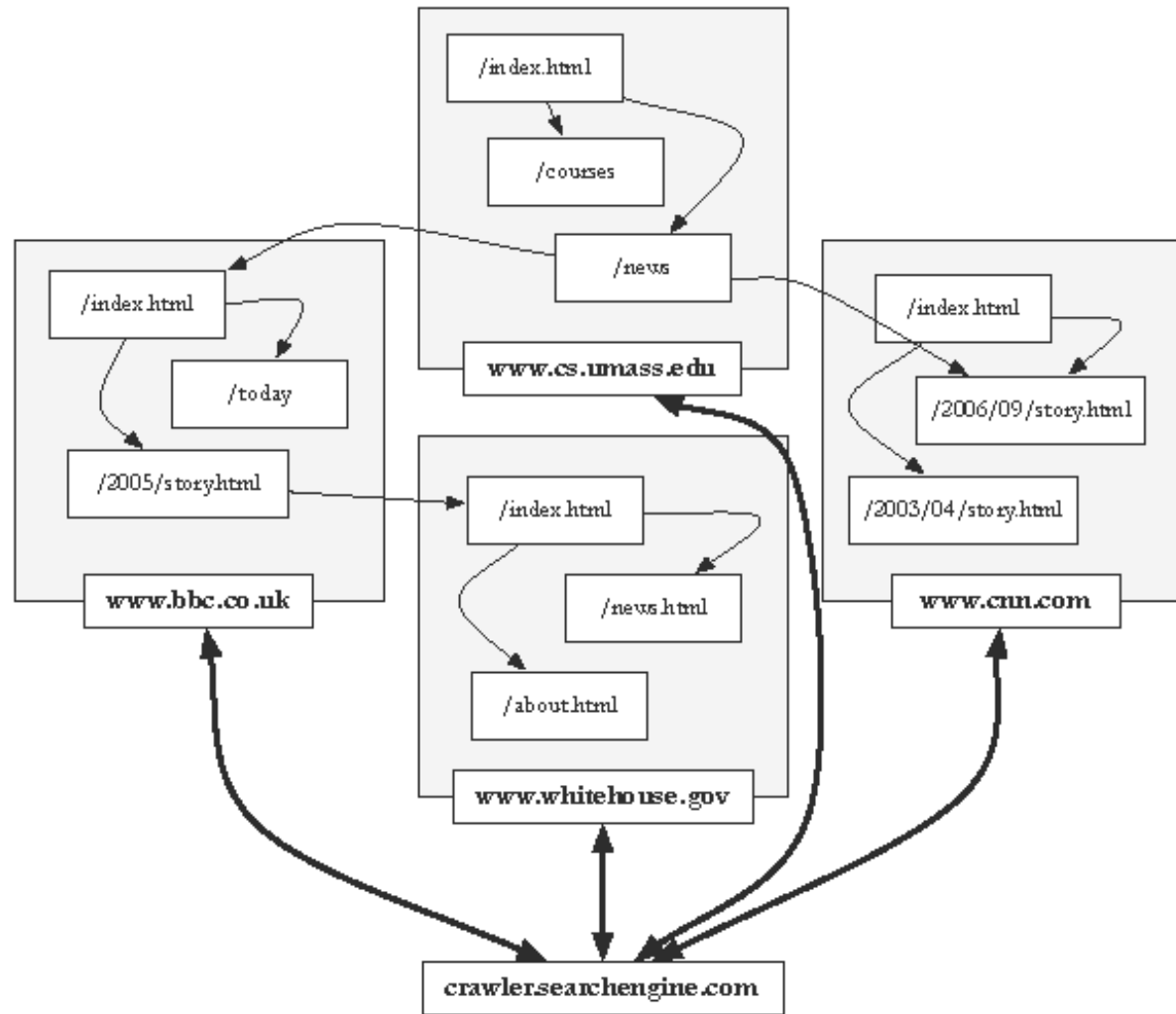
The diagram illustrates the structure of the URL `http://www.cs.umass.edu/csinfo/people.html`. It is broken down into three parts, each with a double-headed arrow pointing to its corresponding component in the URL above:

- http** is labeled as the **scheme**.
- www.cs.umass.edu** is labeled as the **hostname**.
- /csinfo/people.html** is labeled as the **resource**.

Retrieving Web Pages

- Web crawler client program connects to a *domain name system* (DNS) server
- DNS server translates the hostname into an *internet protocol* (IP) address
- Crawler then attempts to connect to server host using specific *port*
- After connection, crawler sends an HTTP request to the web server to request a page
 - usually a GET request

Crawling the Web



Web Crawler

- Starts with a set of *seeds*, which are a set of URLs given to it as parameters
- Seeds are added to a URL request queue
- Crawler starts fetching pages from the request queue
- Downloaded pages are parsed to find link tags that might contain other useful URLs to fetch
- New URLs added to the crawler's request queue, or *frontier*
- Continue until no more new URLs or disk full

Web Crawling

- Web crawlers spend a lot of time waiting for responses to requests
- To reduce this inefficiency, web crawlers use threads and fetch hundreds of pages at once
- Crawlers could potentially flood sites with requests for pages
- To avoid this problem, web crawlers use *politeness policies*
 - e.g., delay between requests to same web server

Controlling Crawling

- Even crawling a site slowly will anger some web server administrators, who object to any copying of their data
- Robots.txt file can be used to control crawlers

```
User-agent: *  
Disallow: /private/  
Disallow: /confidential/  
Disallow: /other/  
Allow: /other/public/
```

```
User-agent: FavoredCrawler  
Disallow:
```

```
Sitemap: http://mysite.com/sitemap.xml.gz
```

Simple Crawler Thread

```
procedure CRAWLERTHREAD(frontier)
  while not frontier.done() do
    website ← frontier.nextSite()
    url ← website.nextURL()
    if website.permitsCrawl(url) then
      text ← retrieveURL(url)
      storeDocument(url, text)
      for each url in parse(text) do
        frontier.addURL(url)
      end for
    end if
    frontier.releaseSite(website)
  end while
end procedure
```

Freshness

- Web pages are constantly being added, deleted, and modified
- Web crawler must continually revisit pages it has already crawled to see if they have changed in order to maintain the *freshness* of the document collection
 - *stale* copies no longer reflect the real contents of the web pages

Freshness

- HTTP protocol has a special request type called HEAD that makes it easy to check for page changes
 - returns information about page, not page itself

```
Client request: HEAD /csinfo/people.html HTTP/1.1
                Host: www.cs.umass.edu
```

```
                HTTP/1.1 200 OK
                Date: Thu, 03 Apr 2008 05:17:54 GMT
                Server: Apache/2.0.52 (CentOS)
                Last-Modified: Fri, 04 Jan 2008 15:28:39 GMT
```

```
Server response: ETag: "239c33-2576-2a2837c0"
                Accept-Ranges: bytes
                Content-Length: 9590
                Connection: close
                Content-Type: text/html; charset=ISO-8859-1
```

Freshness

- Not possible to constantly check all pages
 - must check important pages and pages that change frequently
- Freshness is the proportion of pages that are fresh
- Optimizing for this metric can lead to bad decisions, such as not crawling popular sites
- *Age* is a better metric

Focused Crawling

- Attempts to download only those pages that are about a particular topic
 - used by *vertical search* applications
- Rely on the fact that pages about a topic tend to have links to other pages on the same topic
 - popular pages for a topic are typically used as seeds
- Crawler uses *text classifier* to decide whether a page is on topic

Deep Web

- Sites that are difficult for a crawler to find are collectively referred to as the *deep* (or *hidden*) *Web*
 - much larger than conventional Web
- Three broad categories:
 - private sites
 - no incoming links, or may require log in with a valid account
 - form results
 - sites that can be reached only after entering some data into a form
 - scripted pages
 - pages that use JavaScript, Flash, or another client-side language to generate links

Sitemaps

- Sitemaps contain lists of URLs and data about those URLs, such as modification time and modification frequency
- Generated by web server administrators
- Tells crawler about pages it might not otherwise find
- Gives crawler a hint about when to check a page for changes

Sitemap Example

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://www.company.com/</loc>
    <lastmod>2008-01-15</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.7</priority>
  </url>
  <url>
    <loc>http://www.company.com/items?item=truck</loc>
    <changefreq>weekly</changefreq>
  </url>
  <url>
    <loc>http://www.company.com/items?item=bicycle</loc>
    <changefreq>daily</changefreq>
  </url>
</urlset>
```

Distributed Crawling

- Three reasons to use multiple computers for crawling
 - Helps to put the crawler closer to the sites it crawls
 - Reduces the number of sites the crawler has to remember
 - Reduces computing resources required
- Distributed crawler uses a hash function to assign URLs to crawling computers
 - hash function should be computed on the host part of each URL

Document Feeds

- Many documents are *published*
 - created at a fixed time and rarely updated again
 - e.g., news articles, blog posts, press releases, email
- Published documents from a single source can be ordered in a sequence called a *document feed*
 - new documents found by examining the end of the feed

Document Feeds

- Two types:
 - A *push feed* alerts the subscriber to new documents
 - A *pull feed* requires the subscriber to check periodically for new documents
- Most common format for pull feeds is called *RSS*
 - Really Simple Syndication, RDF Site Summary, Rich Site Summary, or ...

RSS Example

```
<?xml version="1.0"?>
<rss version="2.0">
  <channel>
    <title>Search Engine News</title>
    <link>http://www.search-engine-news.org/</link>
    <description>News about search engines.</description>
    <language>en-us</language>
    <pubDate>Tue, 19 Jun 2008 05:17:00 GMT</pubDate>
    <ttl>60</ttl>

    <item>
      <title>Upcoming SIGIR Conference</title>
      <link>http://www.sigir.org/conference</link>
      <description>The annual SIGIR conference is coming!
        Mark your calendars and check for cheap
        flights.</description>
      <pubDate>Tue, 05 Jun 2008 09:50:11 GMT</pubDate>
      <guid>http://search-engine-news.org#500</guid>
    </item>
```

RSS Example

...

```
<item>
  <title>New Search Engine Textbook</title>
  <link>http://www.cs.umass.edu/search-book</link>
  <description>A new textbook about search engines
    will be published soon.</description>
  <pubDate>Tue, 05 Jun 2008 09:33:01 GMT</pubDate>
  <guid>http://search-engine-news.org#499</guid>
</item>
</channel>
</rss>
```

Noise Example

CNN.com Member Center Sign In Register International Edition

SEARCH THE WEB RSS FEED Search

Home Page World U.S. Weather Business Sports Analysis Politics Law Technology Science & Space Health Entertainment Offbeat Travel Education Special Reports Video Autos I-Reports

IMPACT & WORLD TAKE ACTION SERVICES E-mails RSS Podcasts Mobile CNN Pipeline SEARCH WEB CHANNEL Search

SCIENCE & SPACE

Aquarium plays whale shark matchmaker

Two females flown 8,000 miles for double date in Atlanta

Monday, June 5, 2006, Posted: 5:28 p.m. EDT (21:28 GMT)

ATLANTA, Georgia (CNN) -- Ralph and Norton, meet Alice and Trixie.

The Georgia Aquarium's two male whale sharks got some female companionship on Saturday, when they were joined by two females transported to Atlanta from Taipei, Taiwan.

Researchers are hoping the sharks will mate.

The females -- 11 feet and 14 feet long -- were flown more than 8,000 miles by UPS, which reconfigured a company B-747 freighter with advanced marine life support systems to carry them. (Watch what it took to get the sharks together -- 1:55)

The pilot said they treated the massive fish like first-class passengers.

"As we were doing the descent, we asked to start down a little sooner to make a nice shallow descent, to not make things too uncomfortable back there for the whale sharks," UPS pilot Capt. Bob Crum said.

The plane's center of balance was carefully planned, according to a statement from the aquarium, and veterinarians accompanied the sharks.


The delivery company also brought the two males to Atlanta, where researchers can study the whale sharks' behavior, breeding and development.

The whale sharks -- named after the main characters in the 1950s sitcom "The Honeymooners" -- were delivered to the aquarium in special transportation containers.

The Georgia Aquarium, which opened in November, is the world's largest aquarium. It was a \$250 million gift to Georgia from Bernie Marcus, co-founder of The Home Depot and his wife, Bill, through the Marcus Foundation.

It is the only aquarium outside of Asia to showcase whale sharks, which are the largest fish on Earth.

The aquarium's 6.2-million gallon "Ocean Voyager" tank can hold up to six whale sharks, but it's room for the whale sharks to start a family.



Alice the whale shark swims into the Ocean Voyager tank at the Georgia Aquarium for the first time.

Image: [REDACTED]

YOUR E-MAIL ALERTS

Atlanta (Georgia)

Taiwan

ACTIVATE or Create Your Own

Manage Alerts | What is This?

Subscribe to Time for \$1.00

SPACE

Section Page | Video

Astronauts prepare for third spacewalk

- Astronomers vie to make biggest telescope
- NASA to beam Beatles song to North Star
- U.S. plans for falling satellite

TOP STORIES

Home Page | Video | Most Popular

- Russians choose Putin's successor
- Iran's president makes landmark visit to Iraq
- Israel PM: Attacks on militants go on
- Cable arrested in abandoned baby case

International Edition Languages CNN TV CNN International Headline News Transcripts Advertise with Us About Us

SEARCH THE WEB RSS FEED Search

© 2007 Cable News Network. A Time Warner Company. All Rights Reserved. Terms under which this service is provided to you. Read our privacy guidelines Contact Us Site Map

External sites open in new window; not endorsed by CNN.com. See (Privacy) Pay service with live and archived video. Learn more

Download audio news Add RSS headlines

Content block

Finding Content Blocks

- Other approaches use DOM structure and visual (layout) features

For instance, you can use: XPath/jQuery to inspect the DOM and retrieve the desired HTML tags

Example (Chrome):
`$x('//div[@class="listing"]')`

