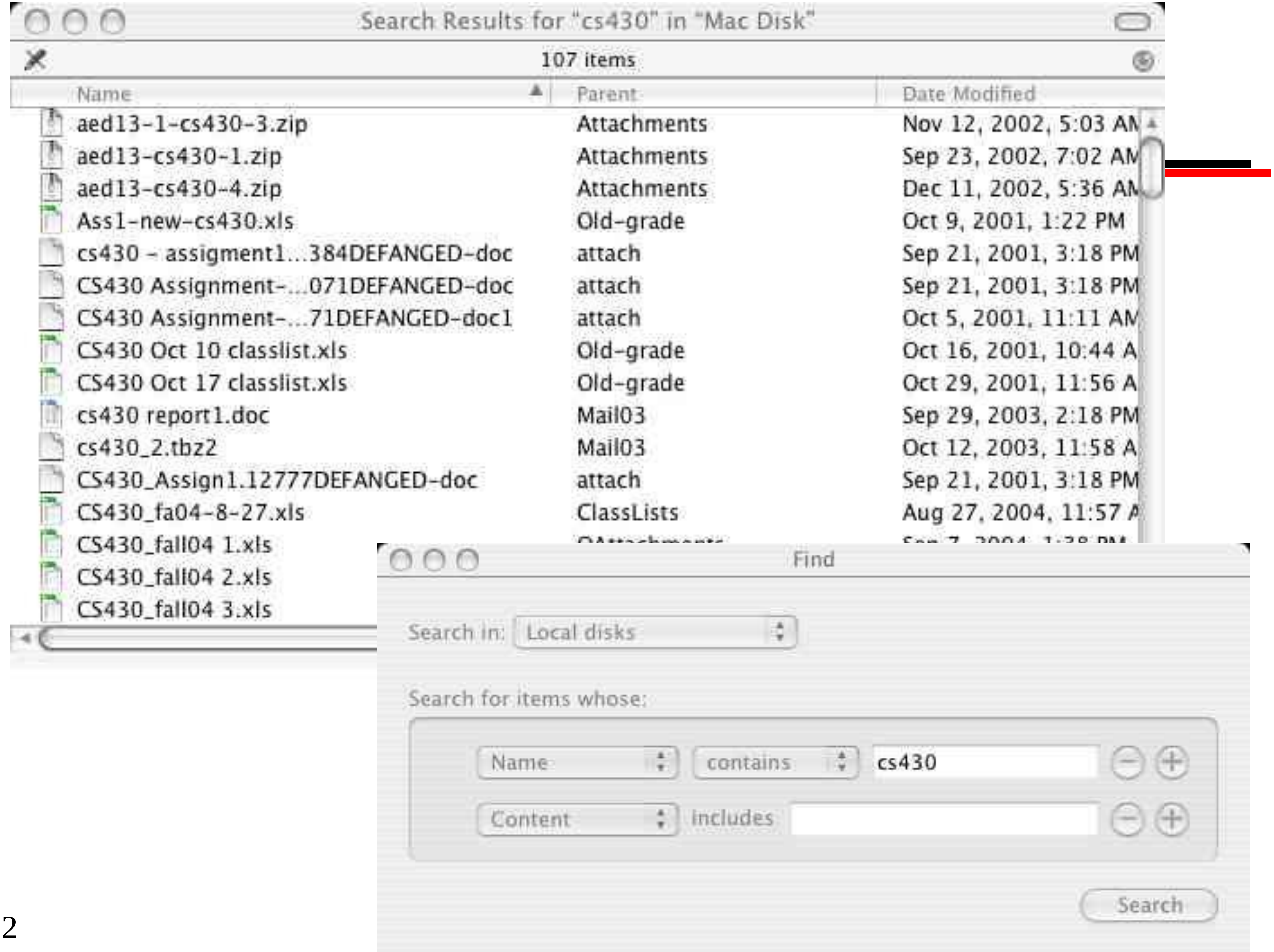
Information Retrieval



Search Results for "cs430" in "Mac Disk"

107 items

Name	Parent	Date Modified
aed13-1-cs430-3.zip	Attachments	Nov 12, 2002, 5:03 AM
aed13-cs430-1.zip	Attachments	Sep 23, 2002, 7:02 AM
aed13-cs430-4.zip	Attachments	Dec 11, 2002, 5:36 AM
Ass1-new-cs430.xls	Old-grade	Oct 9, 2001, 1:22 PM
cs430 - assigment1...384DEFANGED-doc	attach	Sep 21, 2001, 3:18 PM
CS430 Assignment-...071DEFANGED-doc	attach	Sep 21, 2001, 3:18 PM
CS430 Assignment-...71DEFANGED-doc1	attach	Oct 5, 2001, 11:11 AM
CS430 Oct 10 classlist.xls	Old-grade	Oct 16, 2001, 10:44 A
CS430 Oct 17 classlist.xls	Old-grade	Oct 29, 2001, 11:56 A
cs430 report1.doc	Mail03	Sep 29, 2003, 2:18 PM
cs430_2.tbz2	Mail03	Oct 12, 2003, 11:58 A
CS430_Assign1.12777DEFANGED-doc	attach	Sep 21, 2001, 3:18 PM
CS430_fa04-8-27.xls	ClassLists	Aug 27, 2004, 11:57 A
CS430_fall04 1.xls	OAttachments	Sep 7, 2004, 1:38 PM
CS430_fall04 2.xls		
CS430_fall04 3.xls		

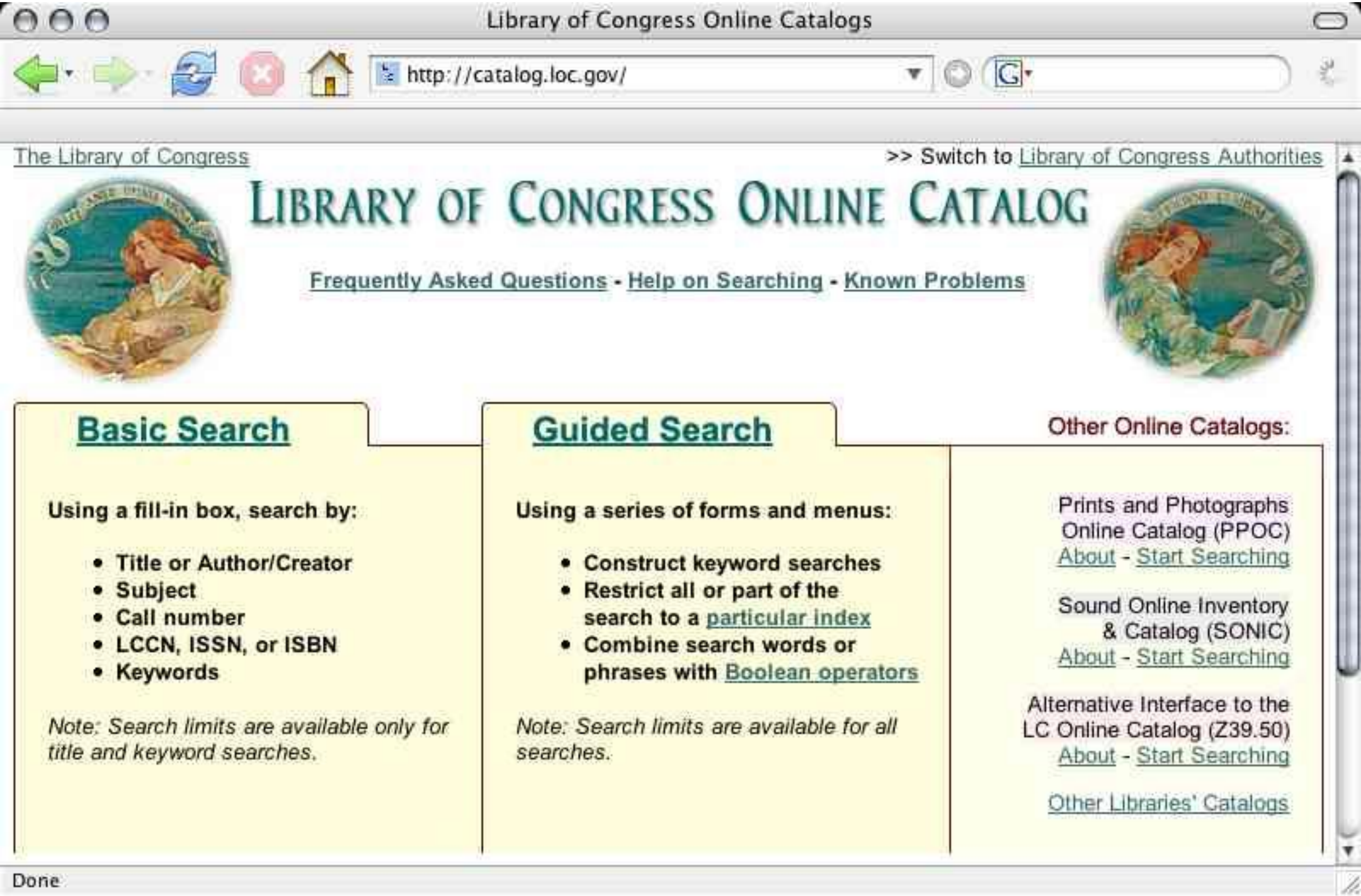
Find

Search in: Local disks

Search for items whose:

Name contains cs430

Content includes



The Library of Congress

>> Switch to Library of Congress Authorities

LIBRARY OF CONGRESS ONLINE CATALOG

Frequently Asked Questions - Help on Searching - Known Problems

Basic Search

Using a fill-in box, search by:

- **Title or Author/Creator**
- **Subject**
- **Call number**
- **LCCN, ISSN, or ISBN**
- **Keywords**

Note: Search limits are available only for title and keyword searches.

Guided Search

Using a series of forms and menus:

- **Construct keyword searches**
- **Restrict all or part of the search to a particular index**
- **Combine search words or phrases with Boolean operators**

Note: Search limits are available for all searches.

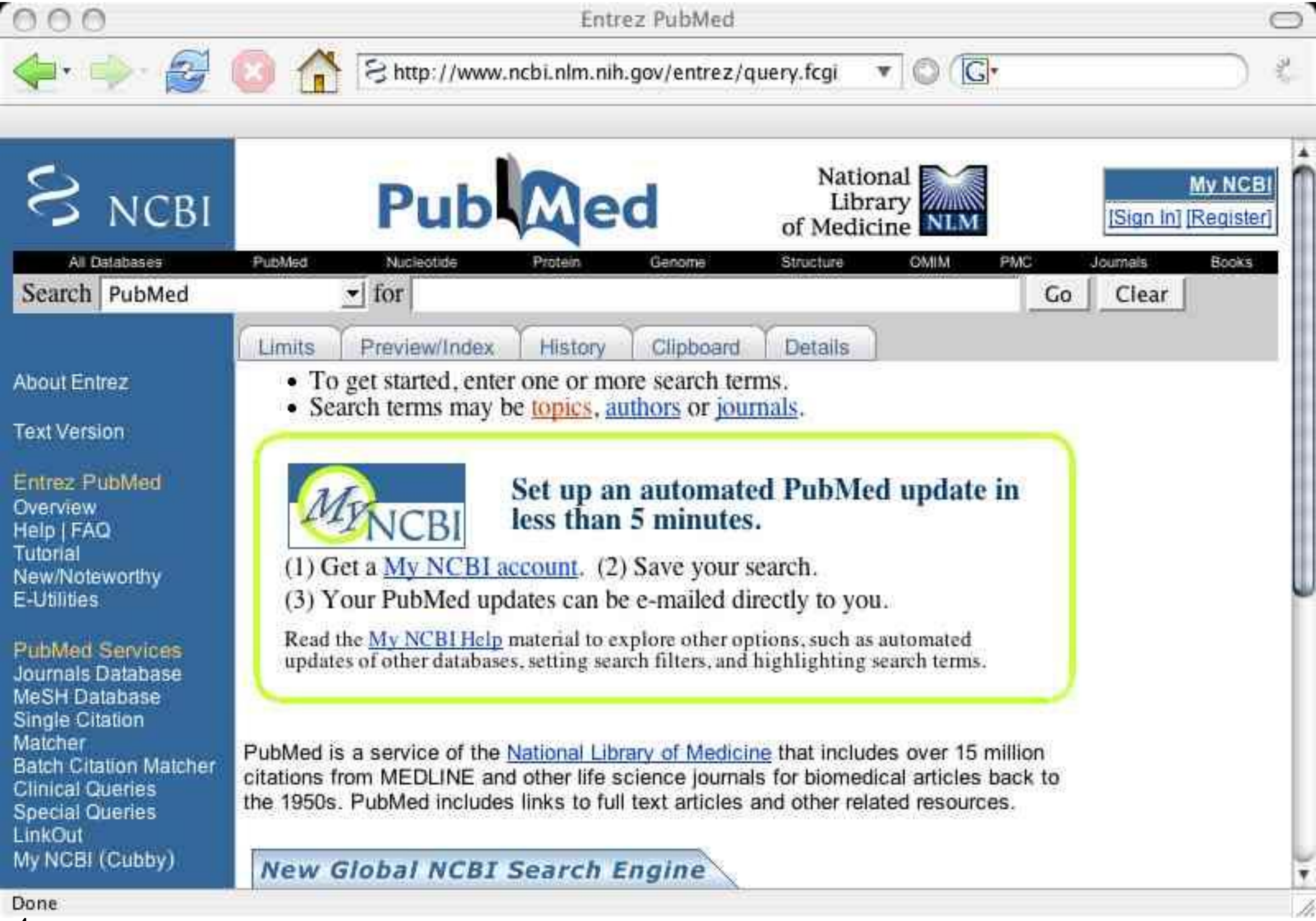
Other Online Catalogs:

Prints and Photographs Online Catalog (PPOC)
About - Start Searching

Sound Online Inventory & Catalog (SONIC)
About - Start Searching

Alternative Interface to the LC Online Catalog (Z39.50)
About - Start Searching

Other Libraries' Catalogs



- To get started, enter one or more search terms.
- Search terms may be topics, authors or journals.

Set up an automated PubMed update in less than 5 minutes.

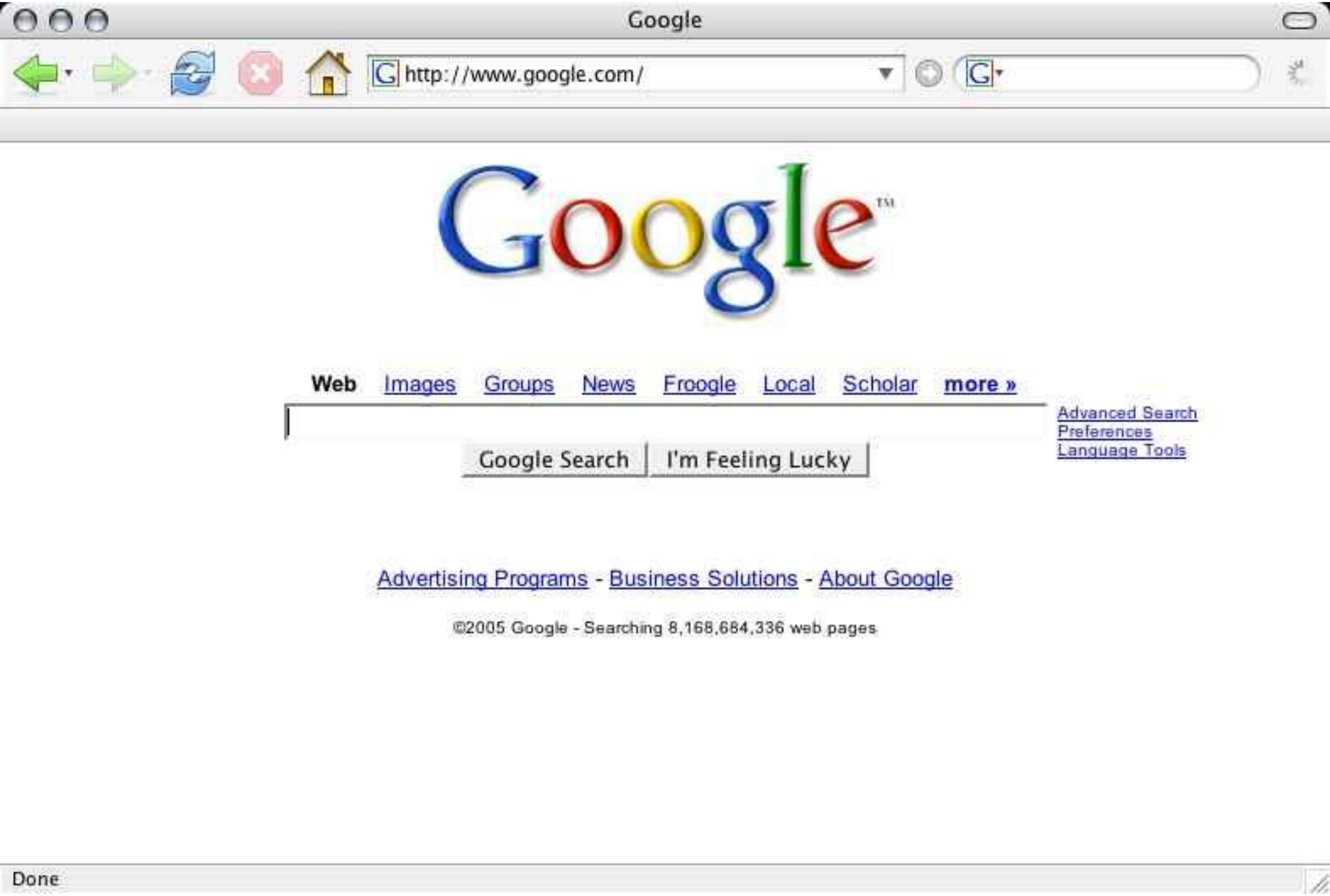
(1) Get a My NCBI account. (2) Save your search.

(3) Your PubMed updates can be e-mailed directly to you.

Read the My NCBI Help material to explore other options, such as automated updates of other databases, setting search filters, and highlighting search terms.

PubMed is a service of the National Library of Medicine that includes over 15 million citations from MEDLINE and other life science journals for biomedical articles back to the 1950s. PubMed includes links to full text articles and other related resources.

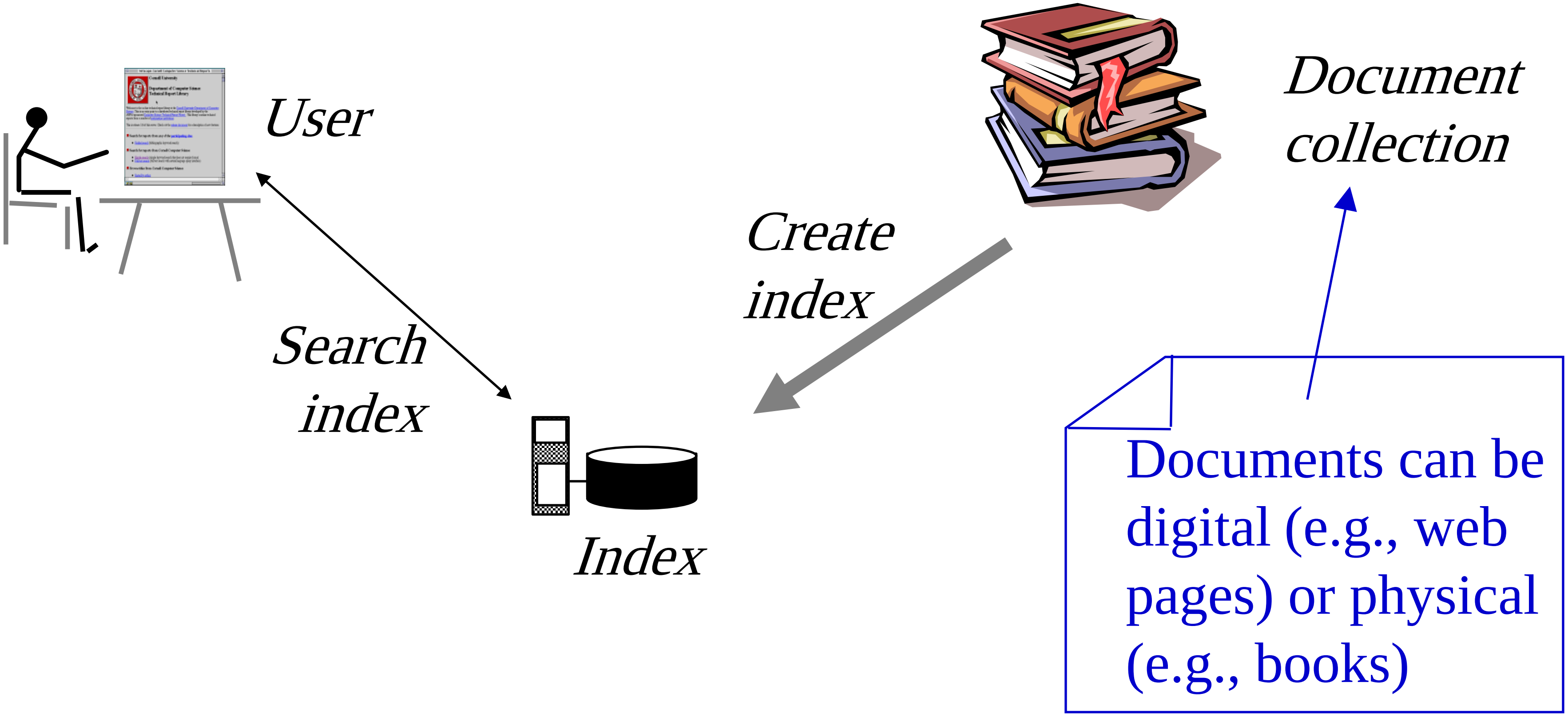
New Global NCBI Search Engine



©2005 Google - Searching 8,168,684,336 web pages

Indexes

Search systems rarely search document collections directly. Instead an **index** is built of the documents in the collection and the user searches the index.



User

Search index

Index

Create index

Document collection

Documents can be digital (e.g., web pages) or physical (e.g., books)

Documents

A textual **document** is a digital object consisting of a sequence of **words** and other symbols, e.g., punctuation.

The individual words and other groups of symbols used for retrieval are known as **tokens** or **terms**.

A textual document can be:

- **Free text**, also known as **unstructured** text, which is a continuous sequence of tokens.
- **Fielded text**, also known as **structured** text, in which the text is broken into sections that are distinguished by tags or other markup, e.g., a library catalog.

[methods of markup, e.g., XML. Partially structured text, e.g., web pages, is called **semi-structured** text.]

Documents

	Fielded text (structured)	Free text (unstructured)
Macintosh files	File names	
Library catalog	Catalog record	
Medical journals	Index record	Abstract or full article
Web search		Web page

Automatic indexing

The aim of **automatic indexing** is to build indexes and retrieve information without human intervention.

When the information that is being searched is **text**, methods of automatic indexing can be very effective.

Historical note

Much of the fundamental research in automatic indexing was carried out by **Gerald Salton**, Professor of Computer Science at Cornell, and his graduate students.

Information Retrieval from Large Collections

Information retrieval from very large collections relies on:

- Huge amounts of computer power carrying out fairly simple algorithms on vast amounts of data.

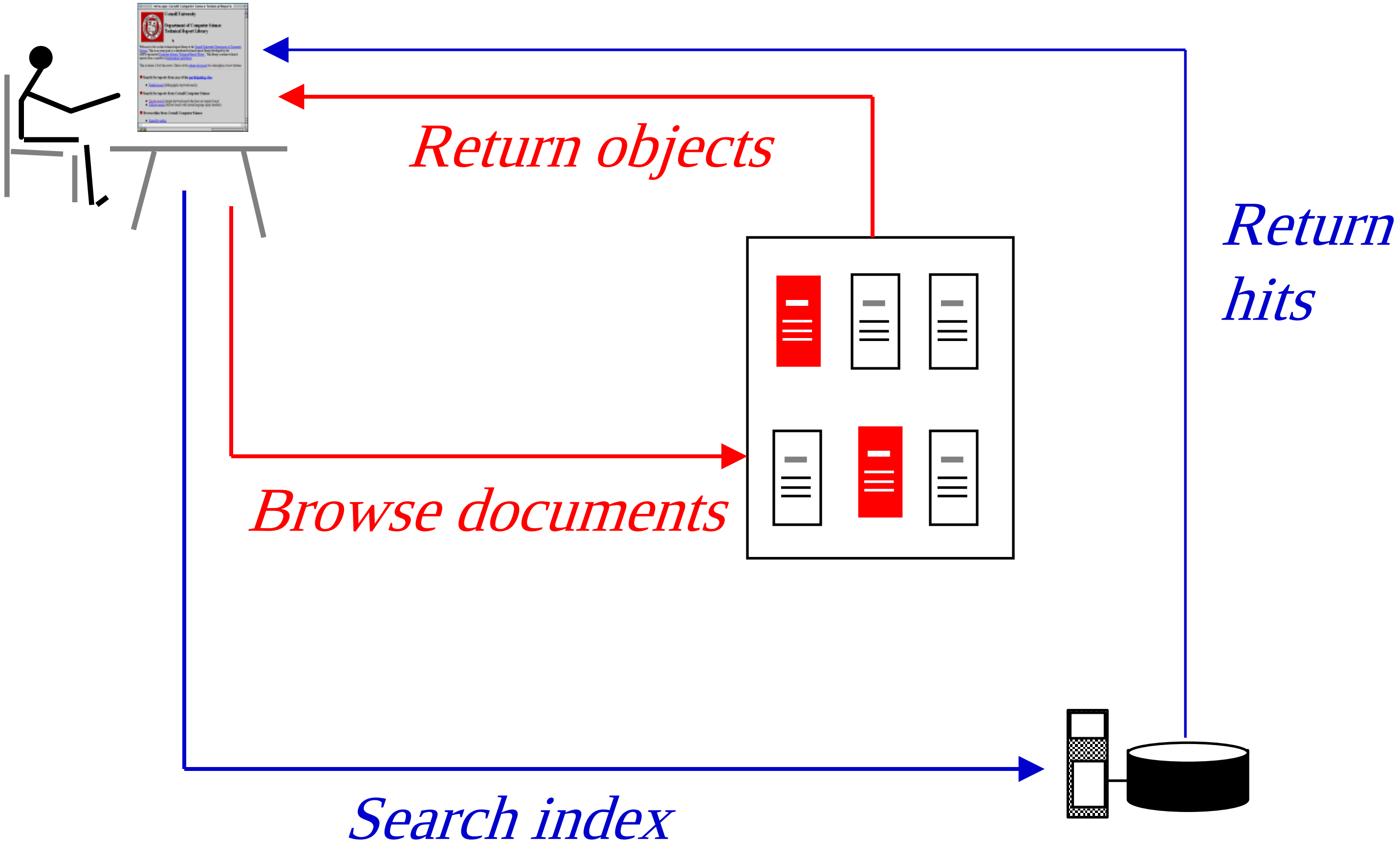
  High-performance computing

- The user's understanding of the information and the capabilities of the system.

  Human computer interaction

Machine-learning is widely used to fit parameters of the methods

Searching and Browsing: The Human in the Loop



Return objects

Return hits

Browse documents

Search index

Information Retrieval from Collections of Textual Documents

Major Categories of Methods

1. Ranking by **similarity to query** (vector space model)
2. Exact **matching** (Boolean)
3. Ranking of matches by **importance of documents** (PageRank)
4. Combinations of methods

   Example: Web search engines, such as Google and Yahoo, use a combination of methods, based on the first and third approaches, with the exact combination being chosen by machine learning.

Definitions

Information retrieval: Subfield of computer science that deals with automated retrieval of documents (especially text) based on their content and context.

Searching: Seeking for specific information within a body of information. The result of a search is a set of **hits**.

Browsing: Unstructured exploration of a body of information.

Linking: Moving from one item to another following links, such as citations, references, etc.

Definitions (continued)

Query: A string of text, describing the information that the user is seeking. Each word of the query is called a **search term**.

A query can be a single search term, a string of terms, a phrase in natural language, or a stylized expression using special symbols, e.g., a regular expression.

Full text searching: Methods that compare the query with every word in the text, without distinguishing the function of the various words.

Fielded searching: Methods that search on specific bibliographic or structural fields, such as author or title.

Definitions

Corpus: A collection of documents that are indexed and searched together.

Word list: The set of all terms that are used in the index for a given corpus (also known as a **vocabulary file**).

With full text searching, the word list is all the terms in the corpus, with **stop words** removed. Related terms may be combined by **stemming**.

C**ontrolled vocabulary**: A method of indexing where the word list is fixed. Terms from it are selected to describe each document.

Keywords: A name for the terms in the word list, particularly with controlled vocabulary.

Sorting and Ranking Hits

When a **user** submits a **query** to a **search system**, the system returns a set of **hits**. With a large collection of documents, the set of hits maybe very large.

The value to the user often depends on the order in which the hits are presented.

Three main methods:

- **Sorting** the hits, e.g., by date
- **Ranking** the hits by **similarity** between query and document
- **Ranking** the hits by the **importance** of the documents

Text Based Information Retrieval

Most **ranking** methods are based on the **vector space model.**

Most **matching** methods are based on **Boolean operators**.

Web search methods combine vector space model with ranking based on importance of documents.

Many practical systems combine features of several approaches.

In the basic form, all approaches treat **words** as **separate tokens** with minimal attempt to interpret them linguistically.

Word Frequency

Observation: Some words are more common than others.

Statistics: Most large collections of unstructured text documents have similar statistical characteristics. These statistics:

- influence the effectiveness and efficiency of data structures used to index documents
- many retrieval models rely on them

Word Frequency

Example

The following example is taken from:

Jamie Callan, *Characteristics of Text*, 1997

Sample of 19 million words

The next slide shows the 50 commonest words in rank order (*r*), with their frequency (*f*).

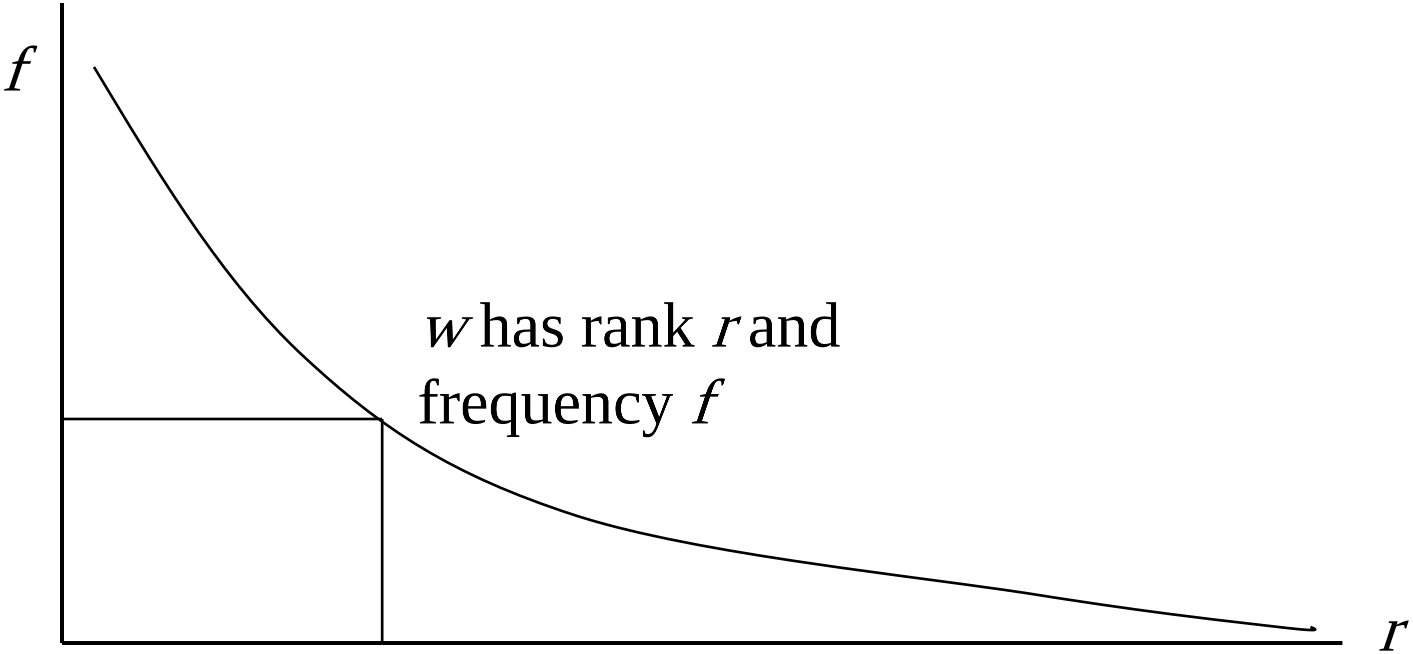
	f		***f***		***f***
the	1,130,021	from	96,900	or	54,958
of	547,311	he	94,585	about	53,713
to	516,635	million	93,515	market	52,110
a	464,736	year	90,104	they	51,359
in	390,819	its	86,774	this	50,933
and	387,703	be	85,588	would	50,828
that	204,351	was	83,398	you	49,281
for	199,340	company	83,070	which	48,273
is	152,483	an	76,974	bank	47,940
said	148,302	has	74,405	stock	47,401
it	134,323	are	74,097	trade	47,310
on	121,173	have	73,132	his	47,116
by	118,863	but	71,887	more	46,244
as	109,135	will	71,494	who	42,142
at	101,779	say	66,807	one	41,635
mr	101,679	new	64,456	their	40,910
with	101,210	share	63,925		

Rank Frequency Distribution

For all the words in a collection of documents, for each word w

f is the frequency that w appears

r is rank of w in order of frequency. (The most commonly occurring word has rank 1, etc.)



Rank Frequency Example

The next slide shows the words in Callan's data normalized. In this example:

r is the rank of word w in the sample.

f is the frequency of word w in the sample.

n is the total number of word occurrences in the sample.

	1000r(f/n)		*1000r(f/n)*		*1000r(f/n)*
the	59	from	92	or	101
of	58	he	95	about	102
to	82	million	98	market	101
a	98	year	100	they	103
in	103	its	100	this	105
and	122	be	104	would	107
that	75	was	105	you	106
for	84	company	109	which	107
is	72	an	105	bank	109
said	78	has	106	stock	110
it	78	are	109	trade	112
on	77	have	112	his	114
by	81	but	114	more	114
as	80	will	117	who	106
at	80	say	113	one	107
mr	86	new	112	their	108
with	91	share	114		

Zipf's Law

If the words in a collection are ranked, *r*, by their frequency, *f*, they roughly fit the relation:

$$r * (f/n) = c$$

Where *n* is the number of word occurrences in the collection, 19 million in the example.

Different collections have different constants *c*.

In English text, *c* tends to be about 0.1.

Methods that Build on Zipf's Law

Stop lists: Ignore the most frequent words (upper cut-off). *Used by almost all systems.*

Significant words: Ignore the most frequent and least frequent words (upper and lower cut-off). *Rarely used.*

Term weighting: Give differing weights to terms based on their frequency, with most frequent words weighed less. *Used by almost all ranking methods.*