Double-click on the key icons to read my *Power Point* notes. --BFB

Searching the Internet

How to find what you're looking for

by

Brenda F. Bell
ACGNJ

presented at the

25th Annual
Trenton Computer Festival
Edison, New Jersey

Speaker Notes

Searching Effectively

Three keys to effective searching

- Decide what you want to find out
- Choose one or more appropriate search sites
- Formulate a search strategy
  - Know enough about the subject to include related concepts
  - Use advanced-query logic

Speaker Notes

Library Terms

- collection
- catalog
- index
- acquisition
- retrieval
- scope
- coverage

Speaker Notes

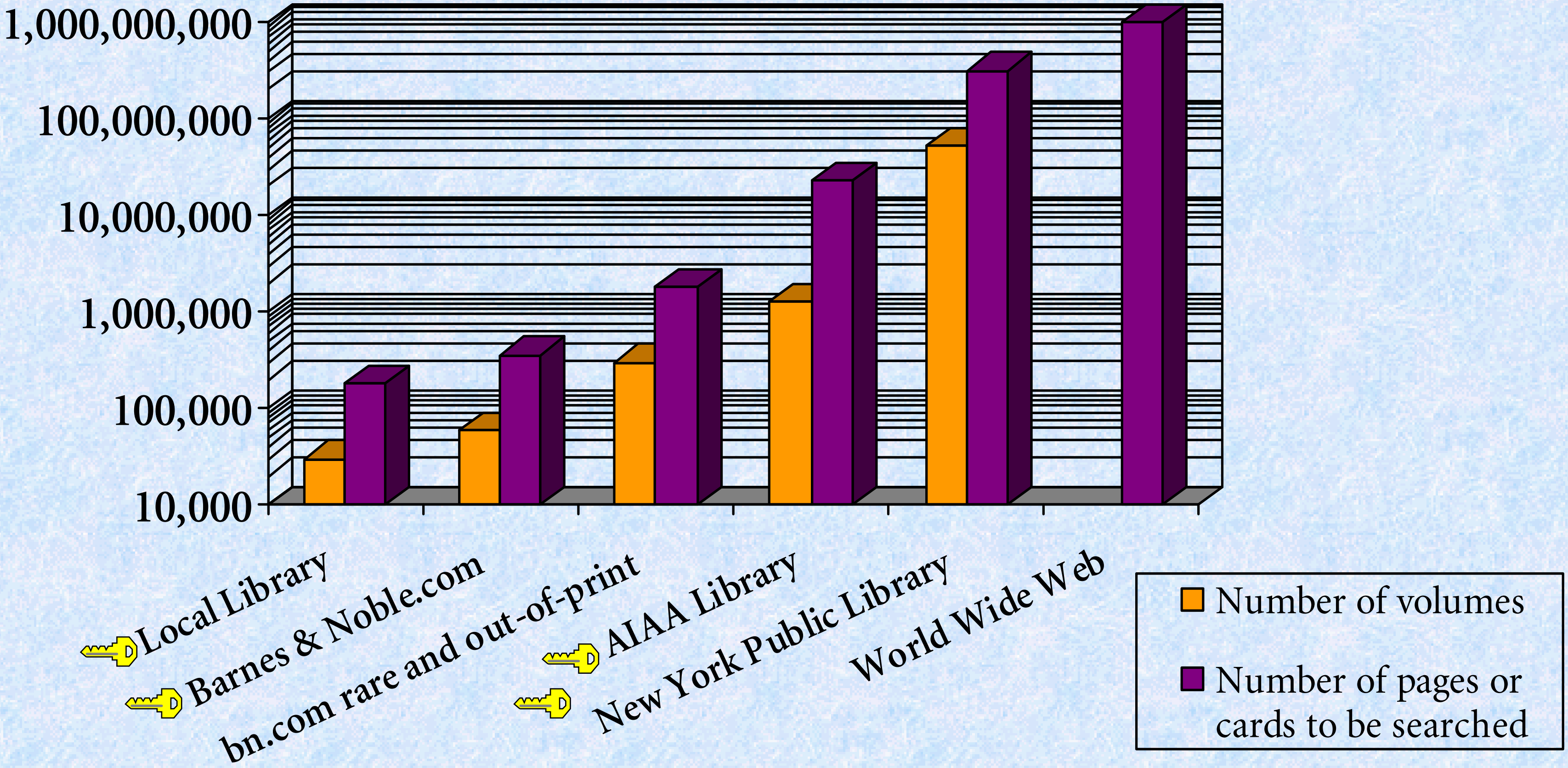
Problems in Searching the Internet

- **Sheer size**
  - over 1,000,000,000 pages and growing
- **Granularity (specificity)**
  - A lot of people with similar interests have sites with similar information. Which one (or several) have *exactly* what you want?
- **Dynamic collection**
  - Number, size, location, and linking of pages is constantly changing

Speaker Notes

Enormity of the Web Collection

(thousands of volumes)



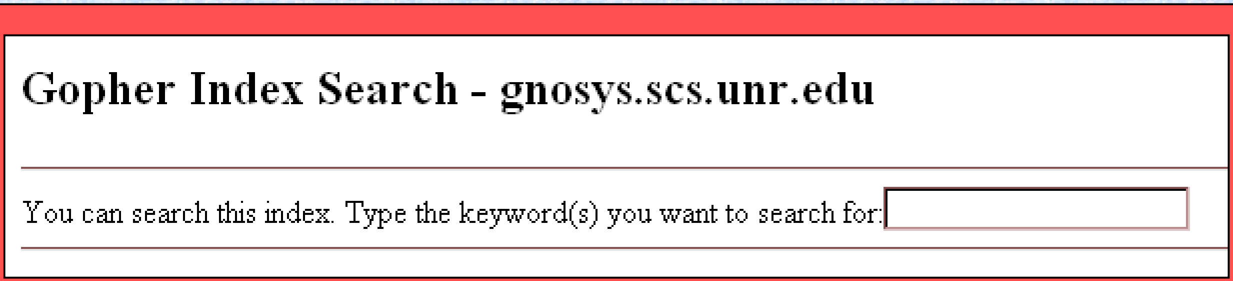
Speaker Notes

Internet search tools

- Historical tools
  - command-line interface
  - text-driven
  - menu-driven
- Web-based tools

Historical search tools

- **archie** — anonymous-ftp “archiver”
- **gopher** — menu-driven document servers
  - **veronica** — Very Easy Rodent-Oriented Netwide Index to Computerized Archives
  - **jughead** — Jonzy’s Universal Gopher Hierarchy Excavation and Display
- **WAIS** — Wide Area Information Server
- **whois, X.500, netfind** — directory services



Gopher Index Search - gnosys.scs.unr.edu

You can search this index. Type the keyword(s) you want to search for:

Veronica interface to gopher at gopher://gnosys.scs.unr.edu:2347/7/ viewed in IE 4.0

Web-based Internet Search Tools

- free search sites
  - Yahoo, Infoseek, GoTo, Dogpile, Google
- subscription search sites
  - STNWeb, DialogWeb, CSA
- hybrid (paid and free) search sites
  - Northern Light Special Collection, Intelligence Online

Methods of Data Acquisition

- owner submission
  - the page or site author, webmaster, or publisher submits a number of Web pages or sites, or other documents, to a search site or a selection of sites
  - in some cases, this is similar to sending a review copy of your new book to the *New York Times Book Review*
- "spiders" and "crawlers"
  - computer programs that automatically search the Internet for new Web sites and add their URLs to a search site's existing catalog(s)
- solicitation
  - subscription databases and private libraries solicit contributions from authors and publishers

Types of Indexes

- **full-text index**
  - retrieves documents based an index of (almost) every word in every document in the collection
    - stop words like “and”, “the”, “it”, “or” are not indexed.
- **keyword retrieval**
  - retrieves documents based on a number of pre-defined key words, phrases, or concepts by which every document in the collection has been categorized
- **bibliographic index**
  - retrieves documents based on bibliographic information such as author, publisher, place of publication, and corporate sponsorship

Indexing for Retrieval

- indexing
  - manual, automated, machine-assisted
  - frequency, <META> tags, submitted abstracts
- general categories
- key words and phrases
  - free-text
  - controlled-vocabulary
  - may be based on full-text

Types of Search Sites

- intrasite
- dedicated search engine
- metasearch site
- portal
- special-interest portal
- vortal
- database portal

Choosing the Right Search Site(s)

	Intrasite	Dedicated search engine	Metasearch site	General interest portal	Special interest portal	Vortal	Database portal
General information		✓	✓	☆			
Highly-specific information	☆	✓	✓	✓	☆	☆	☆ (requires subscription)
Technical or industry-related information	☆			✓	☆	☆	☆ (requires subscription)

✓	possible to find
☆	best place to find

Sample Search —
SR-71 “Blackbird” reconnaissance aircraft

Potential Problems:

- confusion with songbirds
- confusion with businesses named “Blackbird ___________”
- confusion with “State Route 71”
- confusion with Honda CBR 1100 XX Super Blackbird motorcycle

First Results — “blackbird” and “SR-71 blackbird”

site	search string	number of “hits”	false hits in top 20 sites
Yahoo Site Matches	blackbird	30	18
	SR-71 blackbird	1	0/1 (link is dead)
Yahoo Web Page Matches	blackbird	18,733	16
	SR-71 blackbird	268	0
Lycos	blackbird	19,313	15
	SR-71 blackbird	1,830	0
Alta Vista	blackbird	42,600	16
	SR-71 blackbird	59,661	0
Google	blackbird	22,000	15
	SR-71 blackbird	3,139	0
Air Force Link	blackbird	35	6
	SR-71 blackbird	16	0
Web Crawler	blackbird	937	20
	SR-71 blackbird	14,168	14
Northern Light	blackbird	55,137	17
	SR-71 blackbird	4,463	0

First Results — “blackbird” and “SR-71 blackbird”

site	search string	number of “hits”	false hits in top 20 sites
Yahoo Site Matches	blackbird	30	18 (90%)
	SR-71 blackbird	1	0/1 (link is dead)
Yahoo Web Page Matches	blackbird	18,733	16 (80%)
	SR-71 blackbird	268	0
Lycos	blackbird	19,313	15 (75%)
	SR-71 blackbird	1,830	0
Alta Vista	blackbird	42,600	16 (80%)
	SR-71 blackbird	59,661	0
Google	blackbird	22,000	15 (75%)
	SR-71 blackbird	3,139	0
Air Force Link	blackbird	35	6 (30%)
	SR-71 blackbird	16	0
Web Crawler	blackbird	937	20 (100%)
	SR-71 blackbird	14,168	14
Northern Light	blackbird	55,137	17 (85%)
	SR-71 blackbird	4,463	0

First Results — "blackbird" and "SR-71 blackbird"

site	search string	number of "hits"	false hits in top 20 sites
Yahoo Site Matches	blackbird	30	18
	SR-71 blackbird	1	0/1 (link is dead)
Yahoo Web Page Matches	blackbird	18,733	16
	SR-71 blackbird	268	0
Lycos	blackbird	19,313	15
	SR-71 blackbird	1,830	0
Alta Vista	blackbird	42,600	16
	SR-71 blackbird	59,661	0
Google	blackbird	22,000	15
	SR-71 blackbird	3,139	0
Air Force Link	blackbird	35	6
	SR-71 blackbird	16	0
Web Crawler	blackbird	937	20
	SR-71 blackbird	14,168	14
Northern Light	blackbird	55,137	17
	SR-71 blackbird	4,463	0

Search strategies

- Boolean logic
  - AND, OR, NOT
- proximity
  - NEAR
- limited-vocabulary
  - broader, narrower, related terms
- string search
  - exact match
  - usually case-sensitive

“Advanced Search” Syntax

site	AND	OR	NOT	string
Yahoo	“matches on all words (AND)” in search options screen	“matches on any word (OR)” in search options screen	-___	“___”
Lycos	“all the words (AND match)” in advanced search screen	“any words (OR match)” in advanced search screen	-___	“exact phrase (quoted query)” in advanced search screen
Alta Vista	AND in advanced search screen	OR in advanced search screen	NOT in advanced search screen	“___”
Google	(default)	(does not support)	-___	“___”
Air Force Link (uses Verity search engine)	AND	OR	-___	“___”
Web Crawler	AND	OR	NOT	“___”
Northern Light	AND	OR	NOT	“___”

More Specifiers

site	required	proximity	wildcard	parentheses
Yahoo	+___		___*	(no)
Lycos	+___			
Alta Vista	+___	NEAR (within 10 words)	___*	
Google	(default)		(none)	
Air Force Link	+___	<NEAR> (may not be implemented)		??
Web Crawler	+___		___*	yes
Northern Light	+___		___*___ (any length string) % (single character)	nesting

Still more features...

Search by

- URL, title, HTML tag
  (“zone” search)
- date
- language of document
- natural language

Retrieve data sorted by

- relevance (number of hits within a page)
  - ACCRUE (“fuzzy OR”)
- date (newest first)
- other options may be available on a given site

Results of complex search for info on SR-71's jet fuel

site	search string	number of "hits"	false hits in top 20 sites
Yahoo	SR-71 and "jet fuel"	30	18 (28/30)
Lycos	ALL THE WORDS: SR-71 blackbird "jet fuel"	50	13 (36/47)
Alta Vista	"SR-71" AND blackbird AND "jet fuel"	30	27 (23/30 total)
Google	SR-71 blackbird jet fuel SR-71 blackbird jet fuel JP-7	246 12	16/20 0
Air Force Link	fuel AND blackbird; fuel AND SR-71	13 8	15/16 (total unique hits)
Web Crawler	SR-71 AND blackbird AND jet fuel	41	20 (38/38 unique hits)
Northern Light	"SR-71" and "jet fuel"	57	12 (45/57)

Types of false hits from complex search

- **General SR-71 sites**
  - photos, history, politics
- **What's the plane doing now?**
  - LASRE, NASA tests
- **SR-71 is used as a comparison point**
  - NOVA Online
- **Area 51** ("the UFO place")
- **Dead links**
- **passing mention** (excerpts from Joe Weber's novel, *Prime Target*)

Why so many false hits?

- **Improper search strategy**
- **Pay for position**
- **<META> tags have misleading information**
- **“invisible” text**
  - **outside of tags, same color as page bkgd, etc.**
- **bad links**
  - **database not updated**
    - site owner didn’t submit changes
    - crawler didn’t visit site recently

What’s being done to limit deliberate “false hits”?

- “family filters” limit access to porn sites
  - some might overfilter, denying access to sites on breast cancer, for example
- site review before acceptance
- ignoring <META> tags
- blacklisting of sites that use inappropriate <META> tags
- full-text indexing makes it harder to ignore porn-related terms

Looking towards the future

- More information will be available online
  - e-publishing
  - e-commerce
- better query processing
  - context-sensitive filtering
  - pattern-matching
  - neural nets
  - inference engines
  - fuzzy logic
- more for-pay indexes; fewer free search sites
  - database industry shakeout: online v. traditional
  - intellectual property laws

Recap

We have explored:

- the history of Internet search technology
- search theory and syntax
- search industry vocabulary
- types of search sites
- why searches sometimes go wrong
- where Internet searching might go from here

Finale

To search the Internet effectively, ask yourself:

- **What do I want to know?**
  - define your search parameters
- **Where am I likely to find it?**
  - choose the most appropriate search sites
- **How do I ask for it?**
  - design a good search strategy
  - use advanced search techniques where possible