Search Engines and Applications

Prof. Dik Lun Lee

Search Engines is Older Than you Thought

It has many names:

- Information retrieval (IR) – from 50's
- Document retrieval – from 60's
- Text retrieval – from 70's

and applications:

- Digital libraries
- Web search
- Vertical search (e.g., e-commerce)

What Kind of Data does IR Deal With?

- Unformatted or unstructured data (as opposed to relational database)
  - Textual data: papers, technical reports, newspaper articles
  - Completed untagged, plain-text data
- Semi-structured data
  - Web pages (HTML and XML files)
  - Email messages
- Non-textual data
  - images, graphics, video

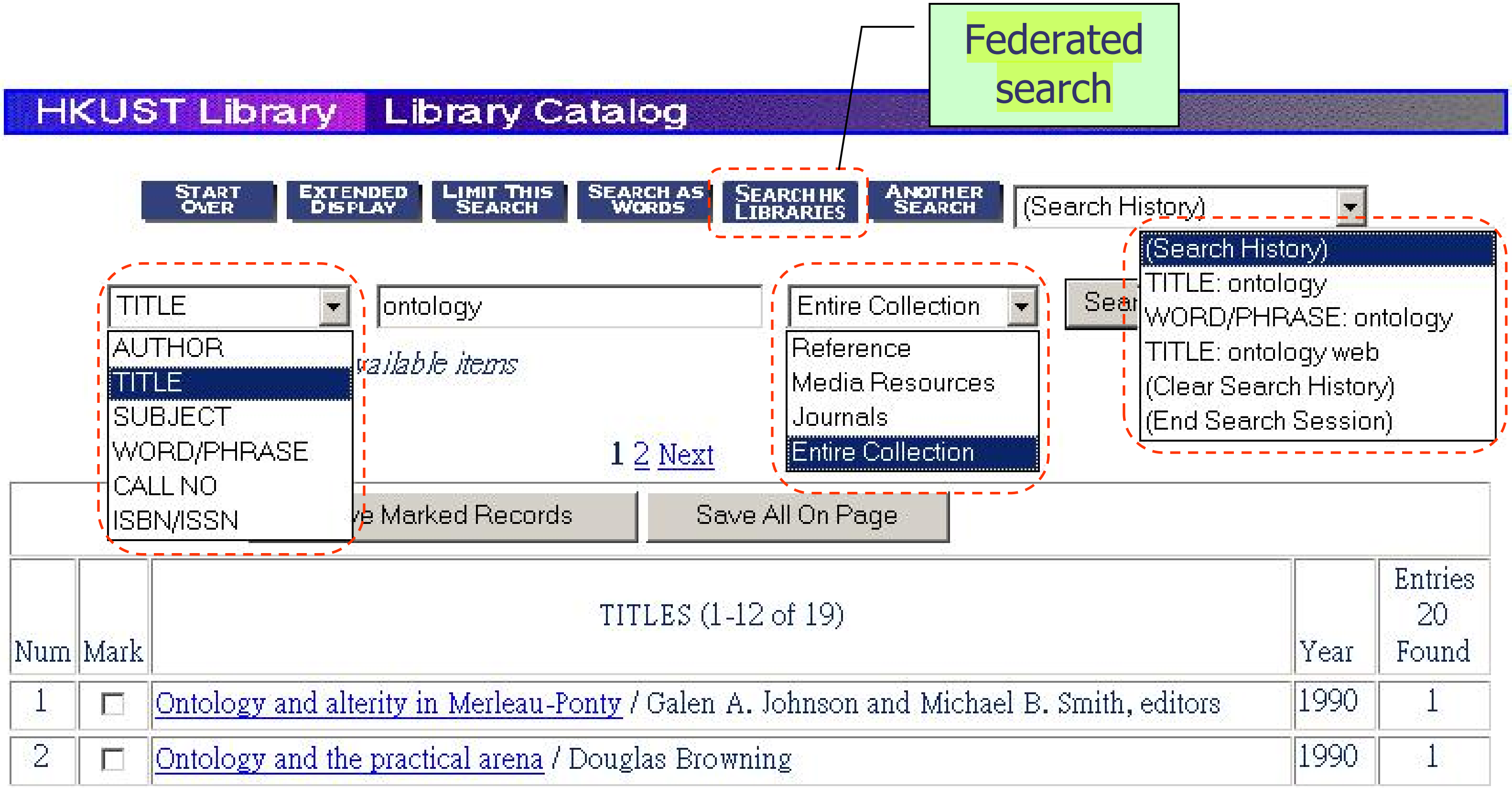
In this course, we study textual and web data

Examples of IR Systems :

- Search Engines are not just Google, Bing, Baidu (GBB)
  - These are global, web-scale search engines
- Most people used IR in some other ways, e.g.,
  - Library catalogue search; most library search systems support both structured and full text search
  - Amazon’s product search
  - Many others (Wikipedia search, …)

Library systems

- Books: http://ustlib.ust.hk/ (HKUST library)

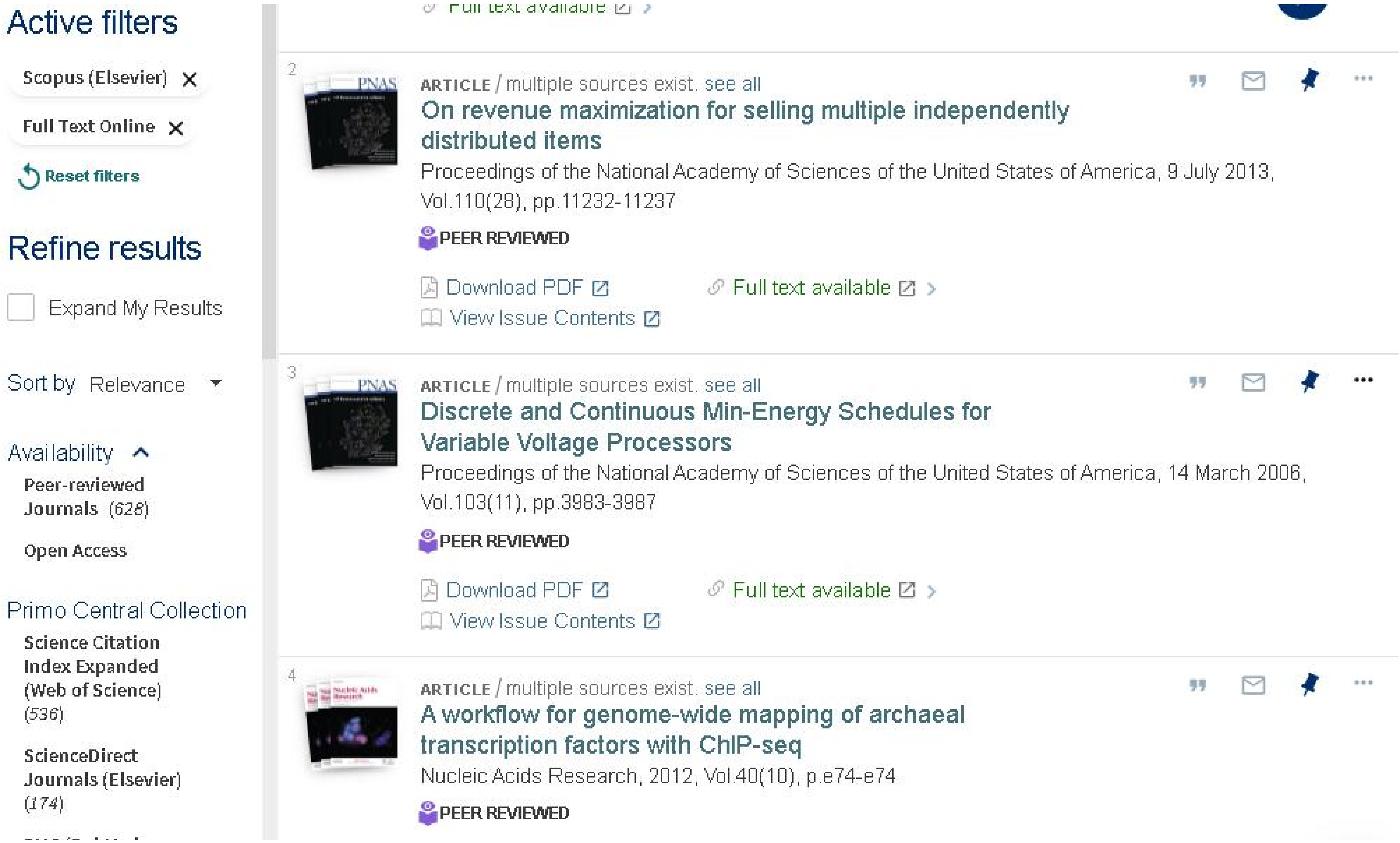


TITLE: ontology

Entire Collection

Num	Mark	TITLES (1-12 of 19)	Year	Entries 20 Found
1	□	Ontology and alterity in Merleau-Ponty / Galen A. Johnson and Michael B. Smith, editors	1990	1
2	□	Ontology and the practical arena / Douglas Browning	1990	1

Result Page has more Functions



Full text available

Active filters

Scopus (Elsevier) ✕

Full Text Online ✕

Reset filters

Refine results

□ Expand My Results

Sort by Relevance ▾

Availability ^

Peer-reviewed Journals (628)

Open Access

Primo Central Collection

Science Citation Index Expanded (Web of Science) (536)

ScienceDirect Journals (Elsevier) (174)

2

ARTICLE / multiple sources exist. see all

On revenue maximization for selling multiple independently distributed items

Proceedings of the National Academy of Sciences of the United States of America, 9 July 2013, Vol.110(28), pp.11232-11237

PEER REVIEWED

Download PDF

Full text available

View Issue Contents

3

ARTICLE / multiple sources exist. see all

Discrete and Continuous Min-Energy Schedules for Variable Voltage Processors

Proceedings of the National Academy of Sciences of the United States of America, 14 March 2006, Vol.103(11), pp.3983-3987

PEER REVIEWED

Download PDF

Full text available

View Issue Contents

4

ARTICLE / multiple sources exist. see all

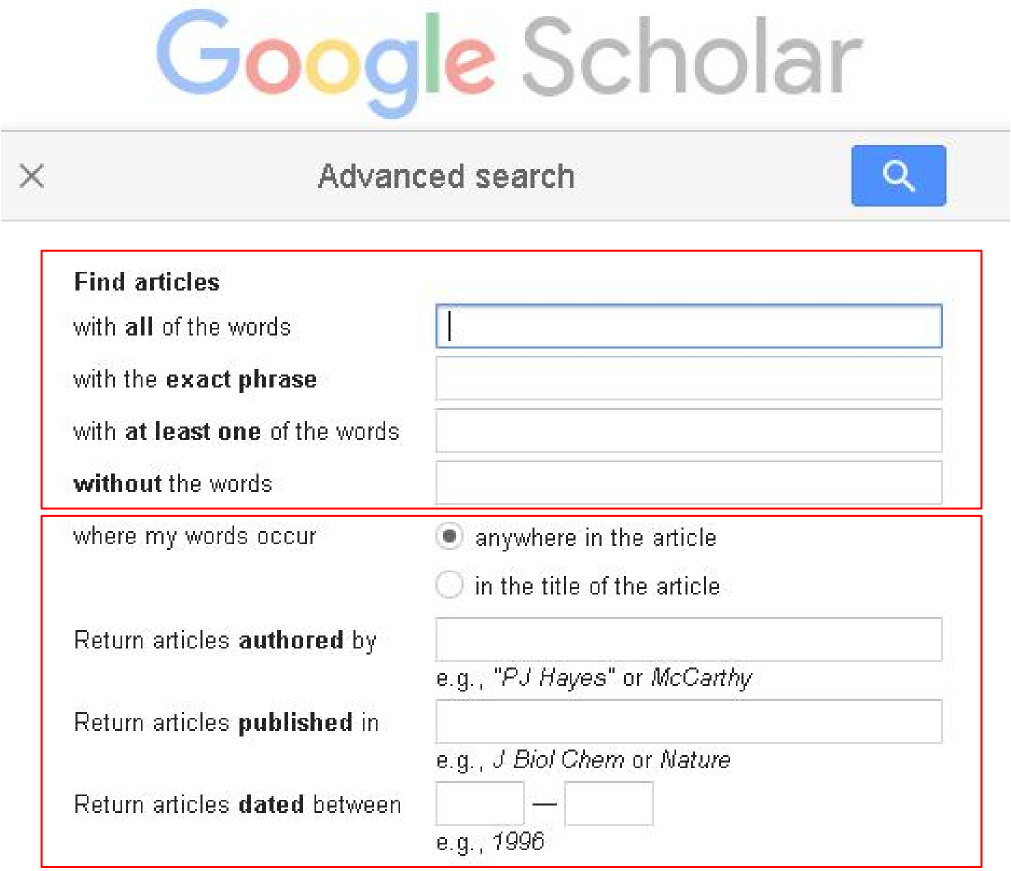
A workflow for genome-wide mapping of archaeal transcription factors with ChIP-seq

Nucleic Acids Research, 2012, Vol.40(10), p.e74-e74

PEER REVIEWED

- Unlike Google, libraries have more structured data (fields / facets)

How is it Compared to Google Scholar?



Google Scholar

Find articles

with **all** of the words

with the **exact phrase**

with **at least one** of the words

without the words

where my words occur ● anywhere in the article ○ in the title of the article

Return articles **authored** by

e.g., *"PJ Hayes"* or *McCarthy*

Return articles **published** in

e.g., *J Biol Chem* or *Nature*

Return articles **dated** between —

e.g., *1996*

Boolean conditions on keywords

Field search

Site Search

- A search engine for one site (or group of related sites)
- How is it different from GBB?
  - Data are more structured:
    - Data are grouped into "collections ‟, e.g., products, press releases, news, manuals, records dumped from database tables
    - Search can be applied to a subset of the collections
  - Query format:
    - Standard AND/OR, phrase, etc.
    - Search on fields: titles, authors, within date range, etc.
  - Result page: Grouped by document types, ranked by date or relevance, etc.
- Example: search on amazon.com; what search features are most useful to you that are available on GBB?

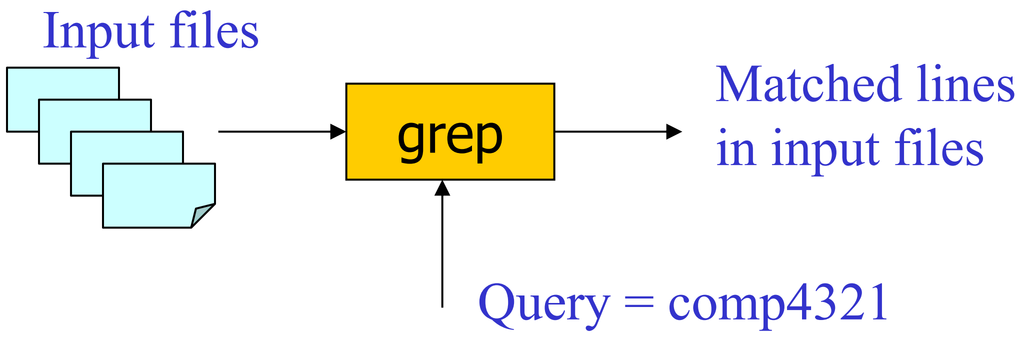
Embedded Search Engines on Devices

- Media and devices that come with a search engine
- A CD/DVD may contain a large amount of data (e.g., conference proceedings); a search engine embedded on it allows you to search the content immediately
  - E.g., Electronic encyclopaedia, product catalogues, corporate reports, etc.
- Search engines embedded on IOT devices
  - What in this world is going to generate the largest amount of data?
- Special requirements:
  - Tailored for the data and device
  - No user installation needed; built-in and executable
  - Provide adequate human/machine and machine/machine interfaces
  - Fast and resource sensitive (running on small devices)

How do you Search for Files on UNIX/LINUX?

- UNIX grep commands (grep, egrep, agrep, etc.)

```
$ grep comp4321 input-file1 input-file2 …
```



Input files

grep

Matched lines in input files

Query = comp4321

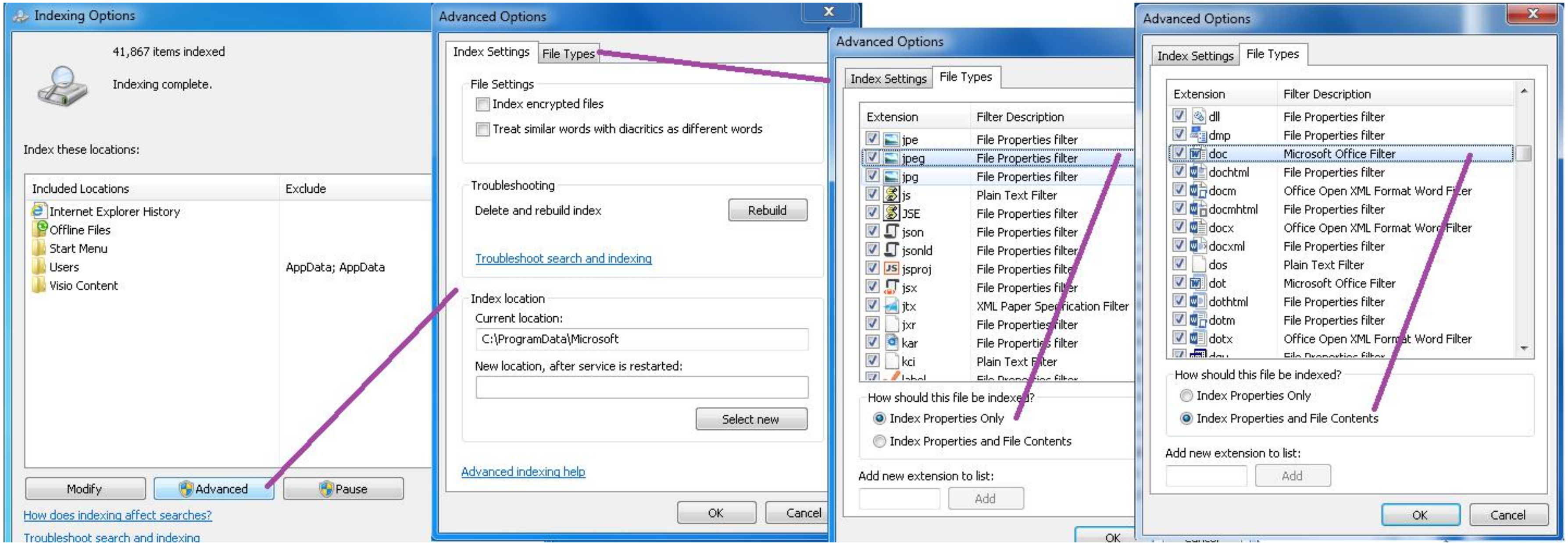
- man –k keyword
  - Search UNIX man pages

- Perform (regular expression) pattern matching

How do you Search for Files on Windows?

- Search for files: plain text, MS Office files, email, etc.
- Specify filenames, dates, file types, etc.
- Windows built-in search function

Index/Search on Windows 10



Indexing Options

41,867 items indexed

Indexing complete.

Index these locations:

Included Locations	Exclude
Internet Explorer History	
Offline Files	
Start Menu	
Users	AppData; AppData
Visio Content	

How does indexing affect searches?

Troubleshoot search and indexing

Advanced Options

File Settings

☐ Index encrypted files

☐ Treat similar words with diacritics as different words

Troubleshooting

Delete and rebuild index

Troubleshoot search and indexing

Index location

Current location:

C:\ProgramData\Microsoft

New location, after service is restarted:

Advanced indexing help

Advanced Options

Extension	Filter Description
☑ jpe	File Properties filter
☑ jpeg	File Properties filter
☑ jpg	File Properties filter
☑ js	Plain Text Filter
☑ JSE	File Properties filter
☑ json	File Properties filter
☑ jsonld	File Properties filter
☑ jsproj	File Properties filter
☑ jsx	File Properties filter
☑ jtx	XML Paper Specification Filter
☑ jxr	File Properties filter
☑ kar	File Properties filter
☑ kci	Plain Text Filter

How should this file be indexed?

● Index Properties Only

○ Index Properties and File Contents

Add new extension to list:

Advanced Options

Extension	Filter Description
☑ dll	File Properties filter
☑ dmp	File Properties filter
☑ doc	Microsoft Office Filter
☑ dochtml	File Properties filter
☑ docm	Office Open XML Format Word Filter
☑ docmhtml	File Properties filter
☑ docx	Office Open XML Format Word Filter
☑ docxml	File Properties filter
☑ dos	Plain Text Filter
☑ dot	Microsoft Office Filter
☑ dothtml	File Properties filter
☑ dotm	File Properties filter
☑ dotx	Office Open XML Format Word Filter

How should this file be indexed?

○ Index Properties Only

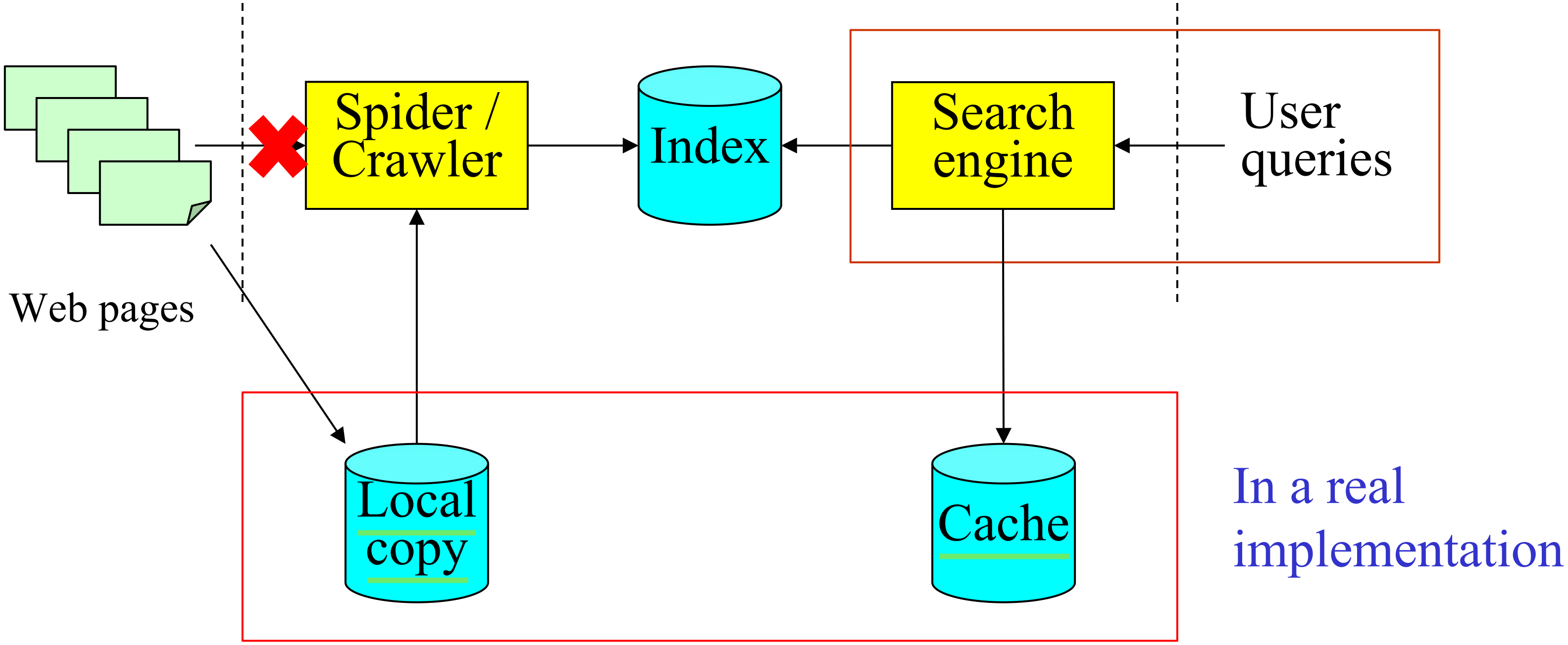
● Index Properties and File Contents

Add new extension to list:

- Windows 10 Index Option allows you to specify:
  - Folders to index
  - Index encrypted files or not
  - To index properties only or properties plus content for different file types
  - Rebuild index at any time

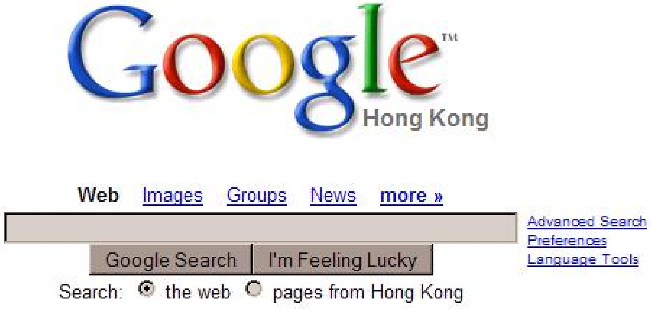
Web Search Engines (GBB: Google/Bing/Baidu)

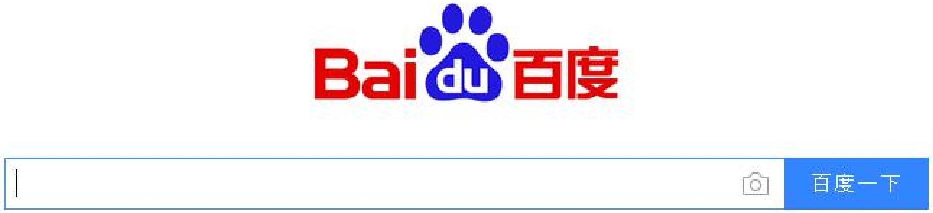
- World wide web search engines: we will cover them a lot
  - Most popular IR application nowadays, e.g., Google, Bing, Baidu
    - Other niche search engine DuckDuckGo, Yandex, etc.

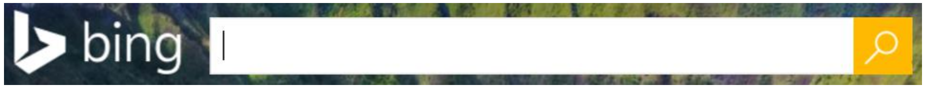


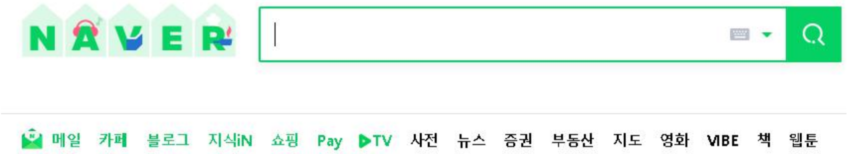
In a real implementation

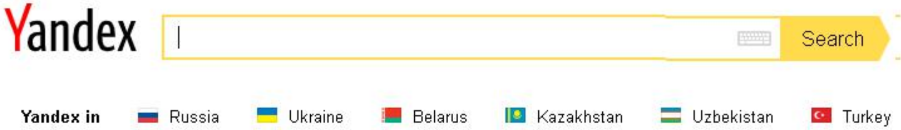
Google, Bing, Baidu











Why is IR Important?

- Most information available is in textual form and has no predefined format (e.g., emails and articles)
  - You may think businesses store data in structured databases, but >80% of business information is unstructured and mostly in text
- Integration of text retrieval capability in most relational database systems. SQL already supports limited search capability such as search based on regular expressions:
  - select * from Employee where Name like '%Lee%'
- Increasing number of online documentation systems (no more hardcopy!)
- Of course, the bloom of World Wide Web

Why is IR a Difficult Problem?

- The size of the web is doubling every year:
  - 50 million pages in November 1995
  - 320 million pages in December 1997
  - 800 million pages in February 1999
  - 1 billion pages in 2000
  - 3.5 billion in 2003 (openfind.com)
  - 8 billion in 2004 (google.com)
  - 20+ billion in 2005 (yahoo.com)
    - Google stopped releasing the size
  - 130 trillion in 2016
- Huge amount of data (e.g., WWW) dictates efficiency, effectiveness and user-friendliness

- Imagine spending "just 0.1 seconds on each page!
- Renders Natural Language Processing infeasible
- Google has an estimated 900,000 servers, and each query triggers >1000 servers (2011 data)

Why is IR a Difficult Problem? (Cont.)

- Unstructured data: difficult to capture semantics in documents. Compare:
  - "select * from Employee where Salary > 100,000"
  - "retrieve all news items about corporate takeover"

- Why is the second query more difficult to answer? The following query is even more difficult:
  - "retrieve all news items about corporate takeover involving an internet company"
  - Note: syntactic → semantic → real-world knowledge

- Documents have unrestricted subject domains
  - it is hard to predefine or pre-categorize the subject domains of documents

Why is IR a Difficult Problem? (Cont.)

- Diversified user base: expert to casual users
  - a system may be clumsy for an expert user but difficult to use for a casual user
  - a system may return information too general to be useful for an expert in the subject but too narrow for a general user

- Intention of information and user query is hard to capture
  - compare a README file and a user manual
  - compare a summary versus an in-depth report

One size cannot fit all!

Why is IR a Difficult Problem?

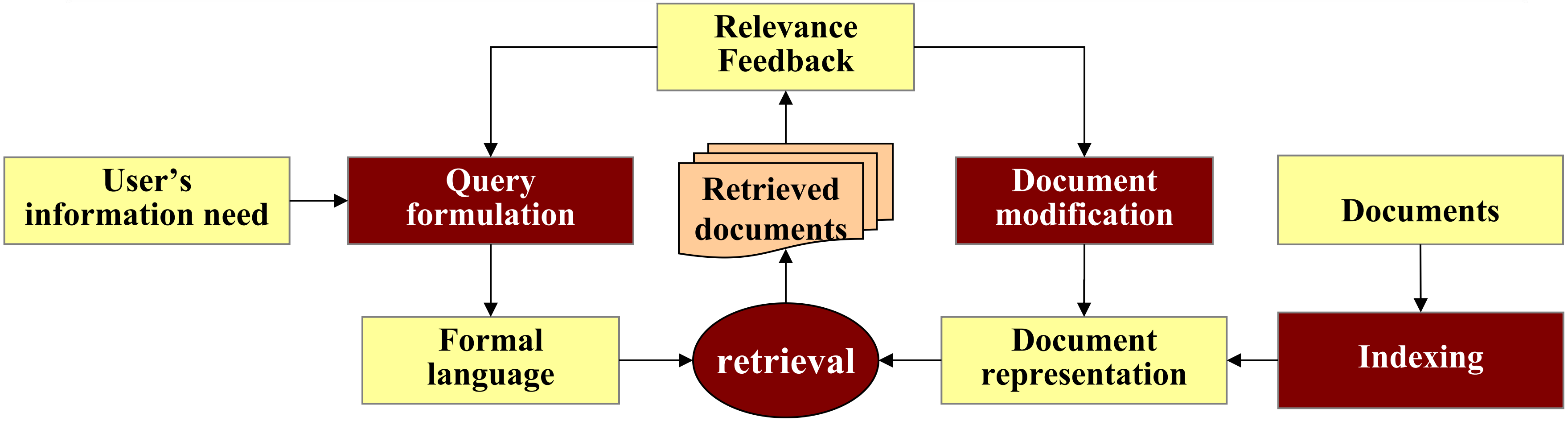
- Distributed and interlinked (e.g., Hypertext and WWW)
  - Where to start a search? Unlike in a centralize database, you have only one (or a few) database(s) to search.
  - How are the information related?

How fast

How good

- Efficiency vs. effectiveness
  - With a limited amount of resources, one can only improve efficiency and effectiveness to a certain degree. Moreover, improving efficiency often means degrading effectiveness, and vice versa.

Document Retrieval Model



- Document: a long string of characters contained in a single file
- Index: a list of important keywords from the documents, stored in some efficient file structure
- Query: Boolean (A and B or C), list of words, natural language
- Relevance feedback: try "similar pages" in Google

Evolution of Search Technologies

- Zeroth-generation search (1960 -)
  - Libraries, collections of electronic documents (legal documents, Lexis/Nexis, scientific databases)
  - Individual documents organized in folders or databases
  - Keyword-based search (looking for keywords)
  - Search on fields (title, author, date) in addition to search on full text body
  - Boolean (title="computer" AND body contains "IBM")
  - E.g., IBM Stairs
  - 0.5 generation: adding statistical to Boolean (e.g., how often does a keyword appear in a document and where?)

Evolution of Search Technologies (Cont.)

- First-generation search engines (web-based, 1993 -)
  - Statistical keyword match
    - traditional search methods applied to web
  - Add a spider / crawler
  - Earlier versions:
    - Altavista (started by Digital Equipment Corporation, then the 2nd largest computer company; sold to Yahoo!)
    - Infoseek (founded in 1994; Infoseek engineer Li Yanhong returned to China and founded Baidu; sold to Disney in 1998)
    - Lycos (started by CMU in 1994)
    - etc.

Evolution of Search Technologies (Cont.)

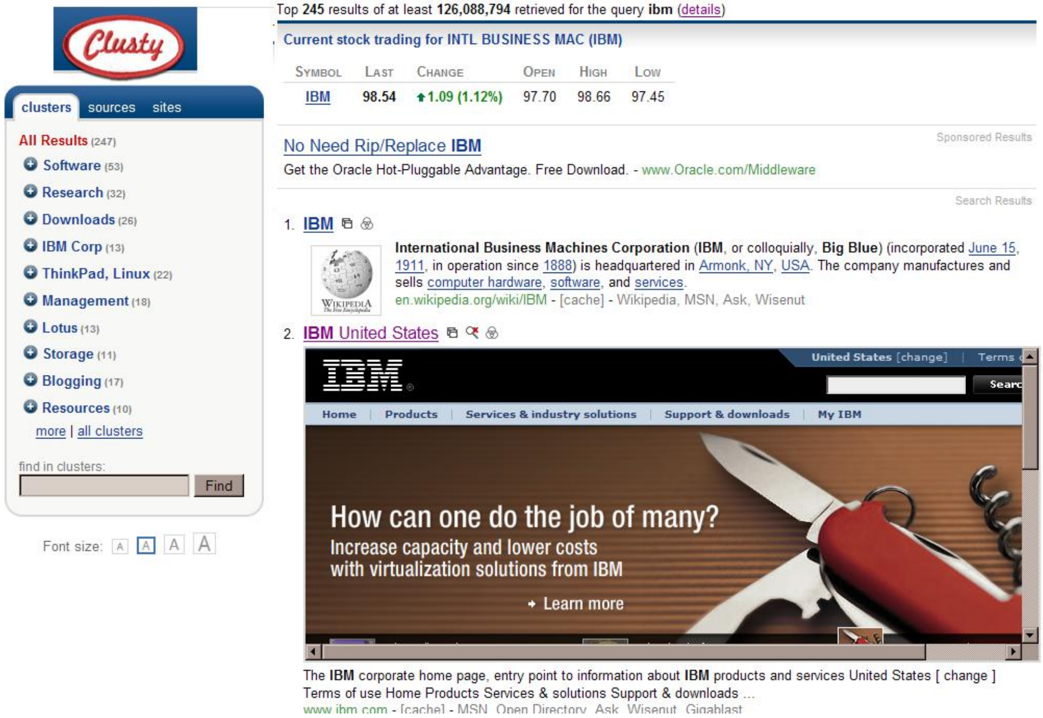
- Second-generation search engines (1997 -)
  - In addition to keyword matching, relying heavily on link analysis (thus capitalizing the special property of web)
  - Google, Fast (sold to Microsoft), etc. etc.

Evolution of Search Technologies (Cont.)

- Third-generation search engines (2001-)
  - Incorporate advanced search features, e.g., automatic categorization

Challengers:

- Teoma (acquired by ask.com)
- Wisenut (acquired by Looksmart)
- Vivisimo (own clusty.com; started by CMU in 2000; acquired by IBM)
- Powerset (acquired by Microsoft in 2008 at allegedly US$ 100m)
- Companies that you will start!



All Results (247)
- Software (53)
- Research (32)
- Downloads (26)
- IBM Corp (13)
- ThinkPad, Linux (22)
- Management (18)
- Lotus (13)
- Storage (11)
- Blogging (17)
- Resources (10)

more | all clusters

find in clusters: Find

Font size: A A A A

Top **245** results of at least **126,088,794** retrieved for the query **ibm** (details)

Current stock trading for INTL BUSINESS MAC (IBM)

SYMBOL	LAST	CHANGE	OPEN	HIGH	LOW
IBM	**98.54**	↑1.09 (1.12%)	97.70	98.66	97.45

Sponsored Results

No Need Rip/Replace **IBM**

Get the Oracle Hot-Pluggable Advantage. Free Download. - www.Oracle.com/Middleware

Search Results

1. IBM

   International Business Machines Corporation (**IBM**, or colloquially, **Big Blue**) (incorporated June 15, 1911, in operation since 1888) is headquartered in Armonk, NY, USA. The company manufactures and sells computer hardware, software, and services.
   en.wikipedia.org/wiki/IBM - [cache] - Wikipedia, MSN, Ask, Wisenut

2. **IBM** United States

WIKIPEDIA
The Free Encyclopedia
IBM
United States [change] | Terms
Searc
Home | Products | Services & industry solutions | Support & downloads | My IBM
How can one do the job of many?
Increase capacity and lower costs
with virtualization solutions from IBM
Learn more

The **IBM** corporate home page, entry point to information about **IBM** products and services United States [change] Terms of use Home Products Services & solutions Support & downloads ...
www.ibm.com - [cache] - MSN, Open Directory, Ask, Wisenut, Gigablast

The Search Industry (and our Job Market)

- Enterprise search
  - Companies deploy their own search engines to enhance the productivity of knowledge workers
  - Endeca (Oracle), Autonomy (Micro Focus), Lucene/Elasticsearch, Microsoft SharePoint/Fast, and Google/Azure Cloud Search, …
- Classified and local search
  - Yellow/White page directories, recruitment and travel web sties; ad placement is the largest source of revenue (Craigslist, Openrice, …)
- Search marketing
  - Companies offering search engine optimization (SEO) services to help websites ranking their pages high in search results

Take Home Messages

- Search engine is rooted in “information retrieval” used by academics
- IR existed even before computers were invented (e.g., manual catalogs in libraries)
- Search engine does NOT just mean web search (Google.com and Bing.com), it includes intranet and enterprise search engines
- Search engine could search structured information (as in library systems)
- Search engine is difficult primarily because it has to “understand” what the user wants through a few query keywords and the semantic content of the pages
- Scaling up/out is also important

Exercise: Identify Differences between Web Search and Structured Data?

	Product search (e.g., amazon.com)	Public web search (e.g., google.com)
Types of content	Mostly structured data (authors, titles, etc.) and some unstructured data (reviews)	Most unstructured data (web page content) and some structured data (last modified date, filetype, etc.)
Query functions	???	???
Search result refinement	???	???