香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

# Search Engines and Applications for Web and Enterprise Data

Kenneth Wai-Ting LEUNG

Slides developed by Prof. Dik Lun LEE

# Search Engine was not Created by Google!

It has many names:

- Information retrieval (IR): dated back to 50’s as one of the major applications of computers
- Document retrieval: “Information” could mean many things; “document” refers to natural language texts organized in some predefined structures (books, reports, letters)
- Text retrieval: Texts are strings of characters with little or no structure; no images or videos

# Applications

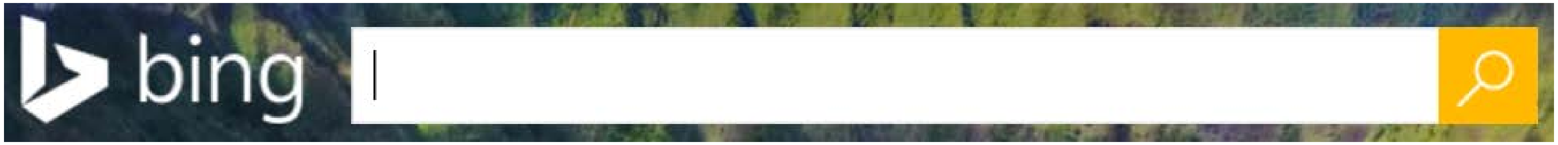
- Digital libraries: All materials in digital forms, accessible and searchable digitally
- Web search: Search anything accessible on the Web; include non-text content, although this course focuses on texts (HTML pages)
- Vertical search: Search in a particular domain, e.g., image, video, news, product (e-commerce) search
  - If we consider web search as "horizontal" search, vertical search focuses on a particular segment, topic or data type and provides better search functions for its focus compared to a general web search engine

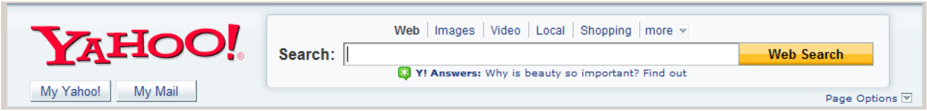
# Types of Data

- 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/multimedia data
  - images, graphics, video

# Examples of IR Systems :

- Examples of famous search engines are Google, Bing, Baidu (GBB), ...
  - Stand-alone search engines (i.e., interact directly with users and search is the only function they provide)





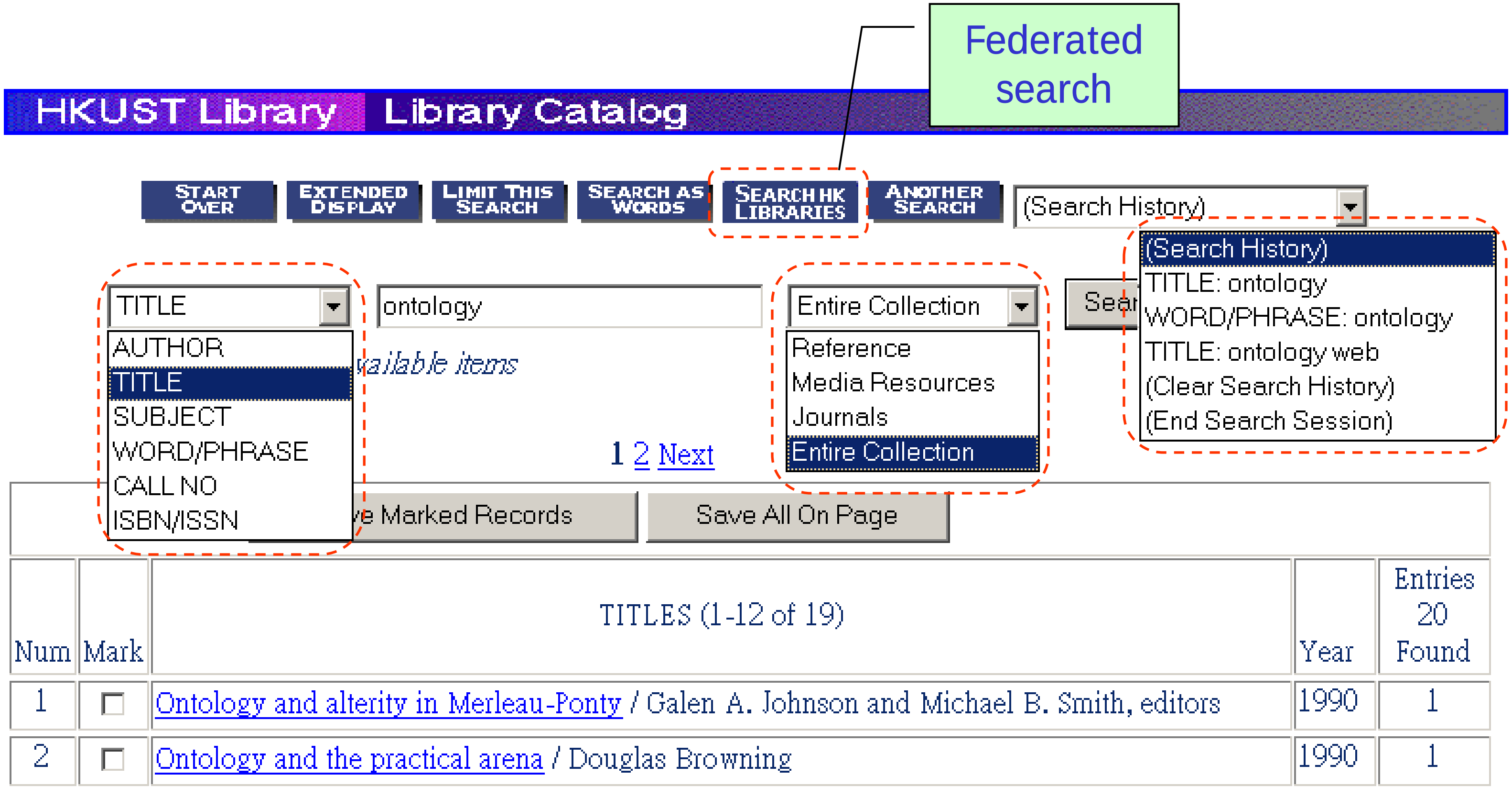
Y! Answers: Why is beauty so important? Find out

# Other Examples of IR Systems

- Most people used IR in some embedded ways
  - In Windows 10, search is one of the many functions of the operating system
  - Search is provided to users as a function or service offered by the application (e.g., in a library system) instead of a standalone search engine by itself)

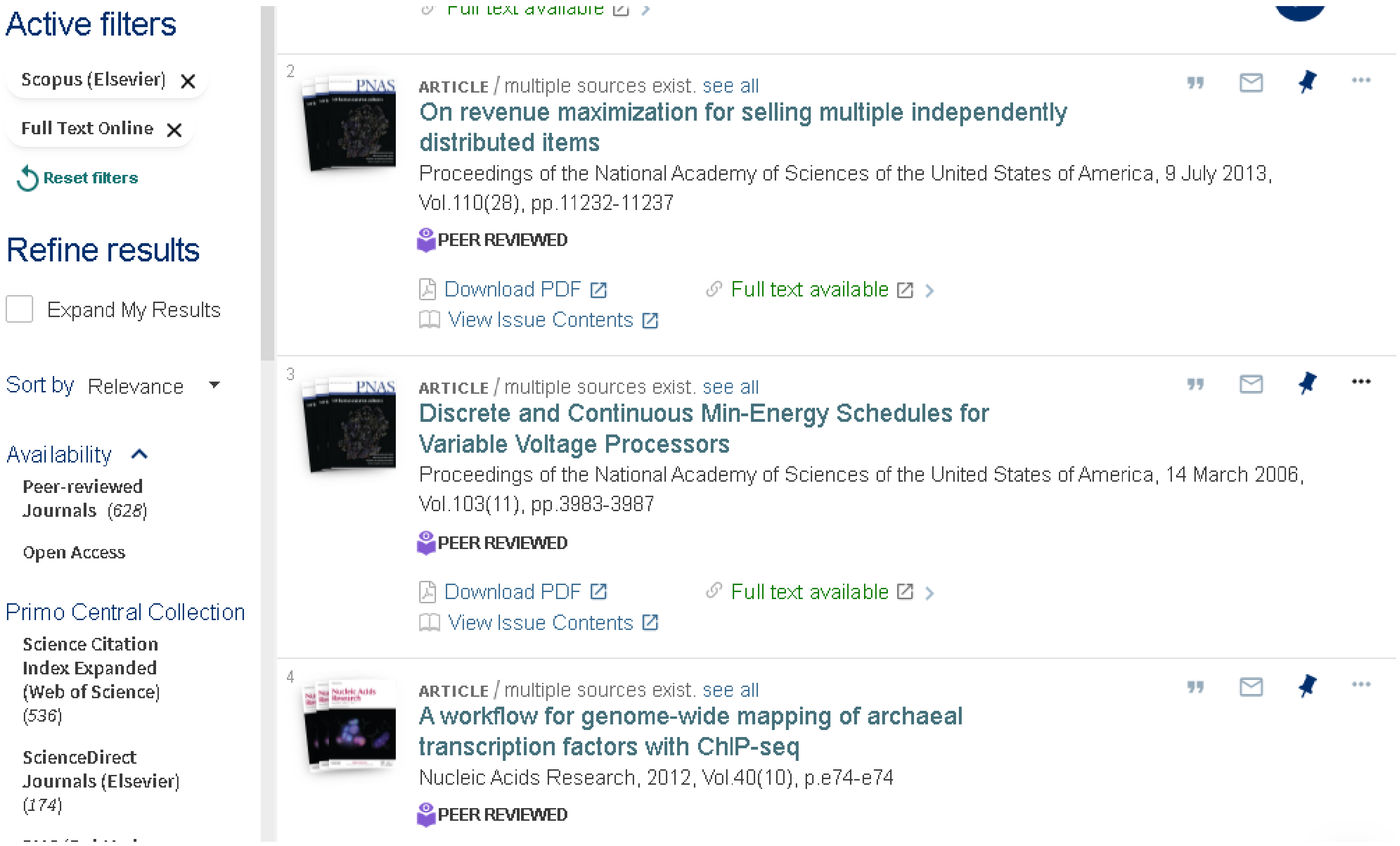
# Library systems

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



| 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



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)

Full text available

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)

# Search in Different Applications

- Vertical Search: A search engine for one data of a focus area
  - Data could be maintained on multiple sites in the vertical search or aggregated from multiple external sites
    - E.g., Job search, News search, Movie search, ...
- Site Search: A search engine for one site (or group of related sites)
  - hsbc.com.hk, ust.hk, ...
- Custom Search: A search frontend to a (big) backend search engine to narrow search to a small set of websites
  - Ust.hk/search-engine?...

Both Google Site Search and Google Custom Search are Google products, but the idea is applicable to other search engines

- Enterprise Search: A search engine for a corporate intranet
  - Multiple types of data (databases, Office documents, emails, ...)
  - Different user roles (sales vs technical support vs CEO ...)
  - Security, security, security, ...

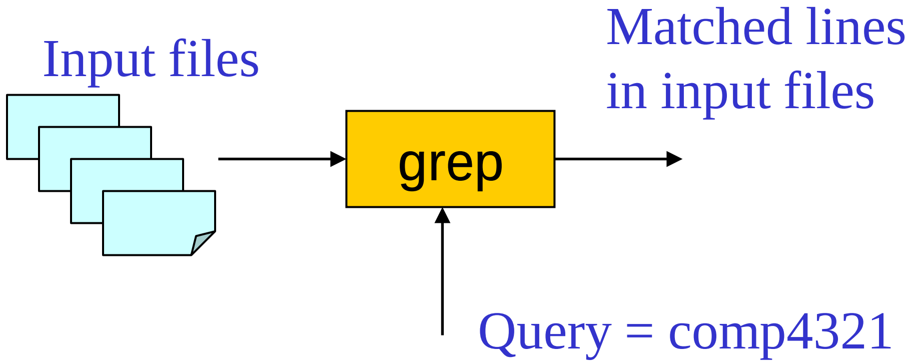
# Embedded Search Engines on Devices

- Search engines embedded on portable devices (mobile phones, USB thumb drives, CD ROMs)
- Search engines are tailored for the data on device
  - E.g., Electronic encyclopaedia, product catalogues, corporate reports, etc.
  - Don’t forget that a CD could hold 600 Mbytes of text!
- Special requirements:
  - No installation needed; built-in and executable
  - Provide adequate interface (e.g., web-based)
  - Fast and resource sensitive (running on small devices)

# File Search 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

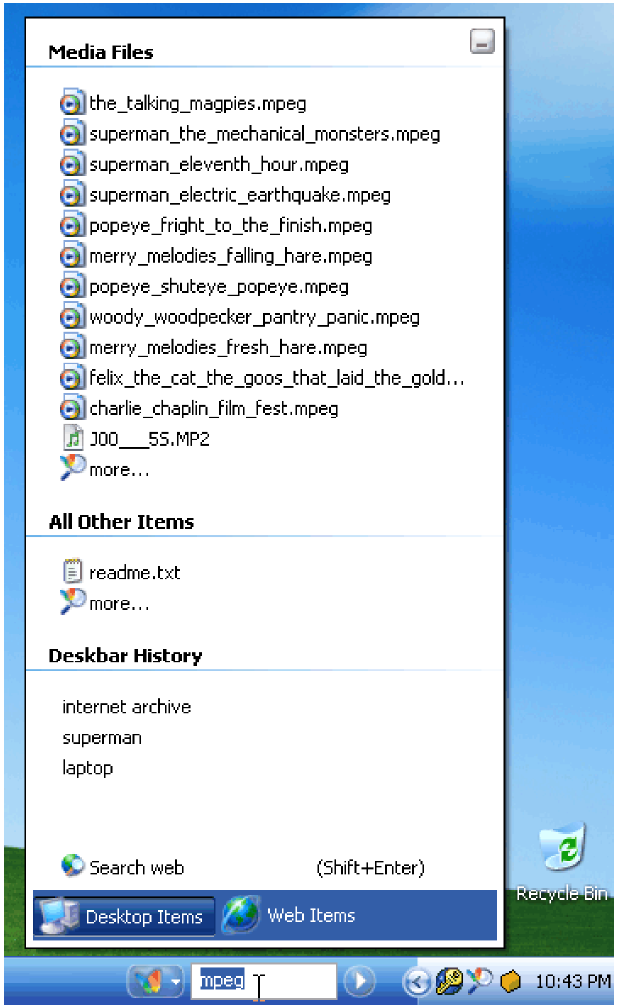
– man –k keyword
  - Search UNIX man pages

– These are simple "search engines" although search functions are extremely simple and primitive!

# 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, Yahoo Desktop, Google Desktop, Windows Desktop, etc.

# Desktop Search Examples



**Media Files**

the_talking_magpies.mpeg
superman_the_mechanical_monsters.mpeg
superman_eleventh_hour.mpeg
superman_electric_earthquake.mpeg
popeye_fright_to_the_finish.mpeg
merry_melodies_falling_hare.mpeg
popeye_shuteye_popeye.mpeg
woody_woodpecker_pantry_panic.mpeg
merry_melodies_fresh_hare.mpeg
felix_the_cat_the_goos_that_laid_the_gold...
charlie_chaplin_film_fest.mpeg
J00___55.MP2
more...

**All Other Items**

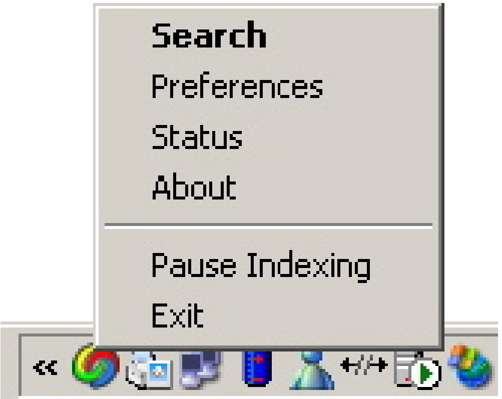
readme.txt
more...

**Deskbar History**

internet archive
superman
laptop

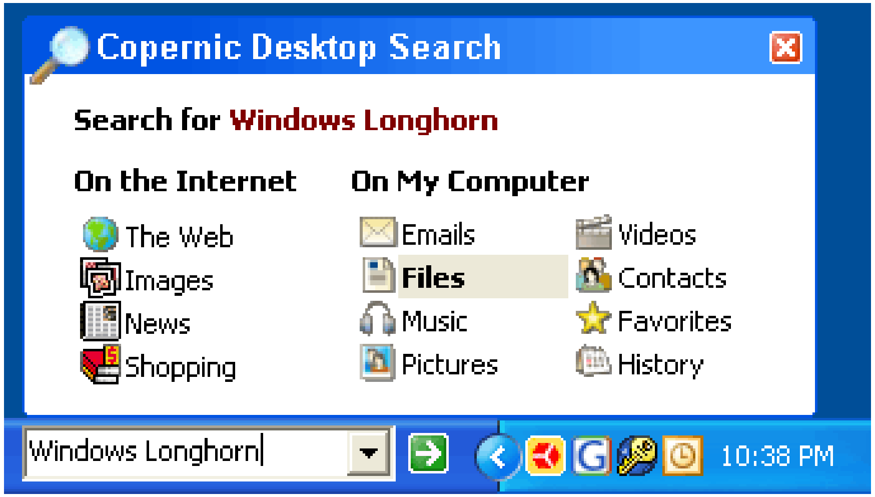
Search web (Shift+Enter)

mpeg



- Windows desktop search has been integrated into Windows
- Copernic is still available
- Google desktop has long been discontinued

Search result

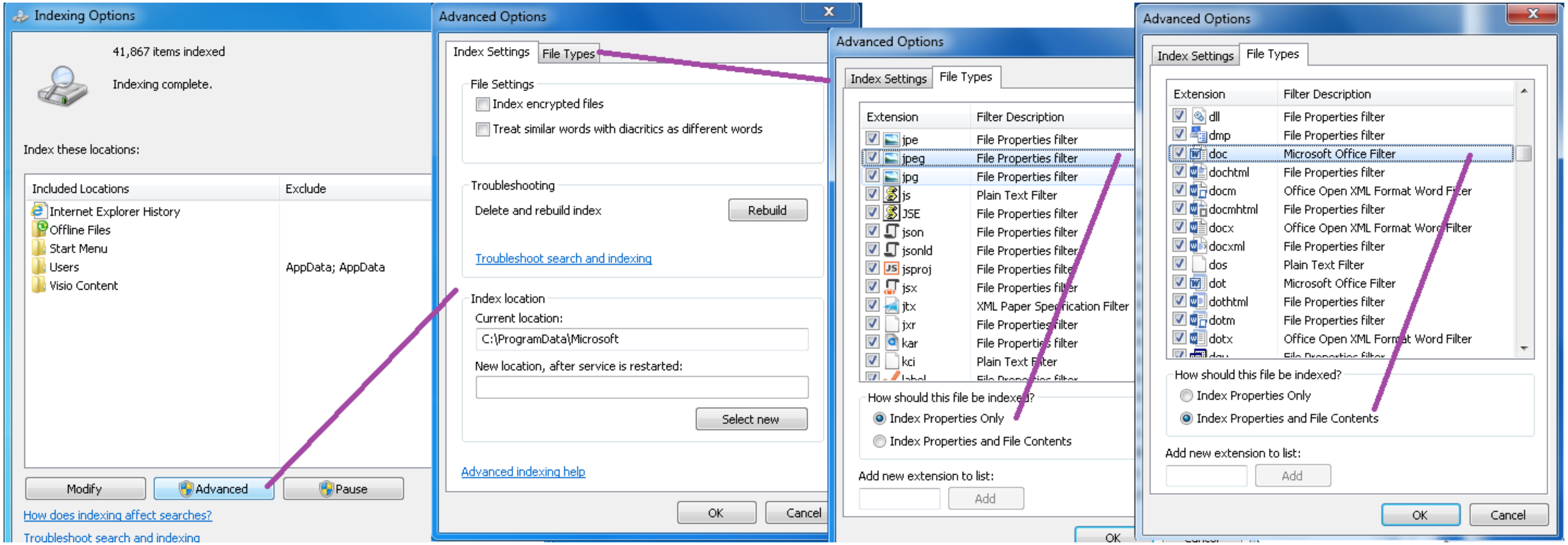


**Copernic Desktop Search**

**Search for Windows Longhorn**

Windows Longhorn

# 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 | |

Modify
Advanced
Pause
How does indexing affect searches?
Troubleshoot search and indexing

**Advanced Options** (Index Settings)

Index Settings
File Types

File Settings

☐ Index encrypted files

☐ Treat similar words with diacritics as different words

Troubleshooting

Delete and rebuild index

Index location

Current location: C:\ProgramData\Microsoft

New location, after service is restarted:

**Advanced Options** (File Types)

| 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 |
| ☑ label | File Properties filter |

How should this file be indexed?

● Index Properties Only

○ Index Properties and File Contents

Add new extension to list:

**Advanced Options** (File Types)

| 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

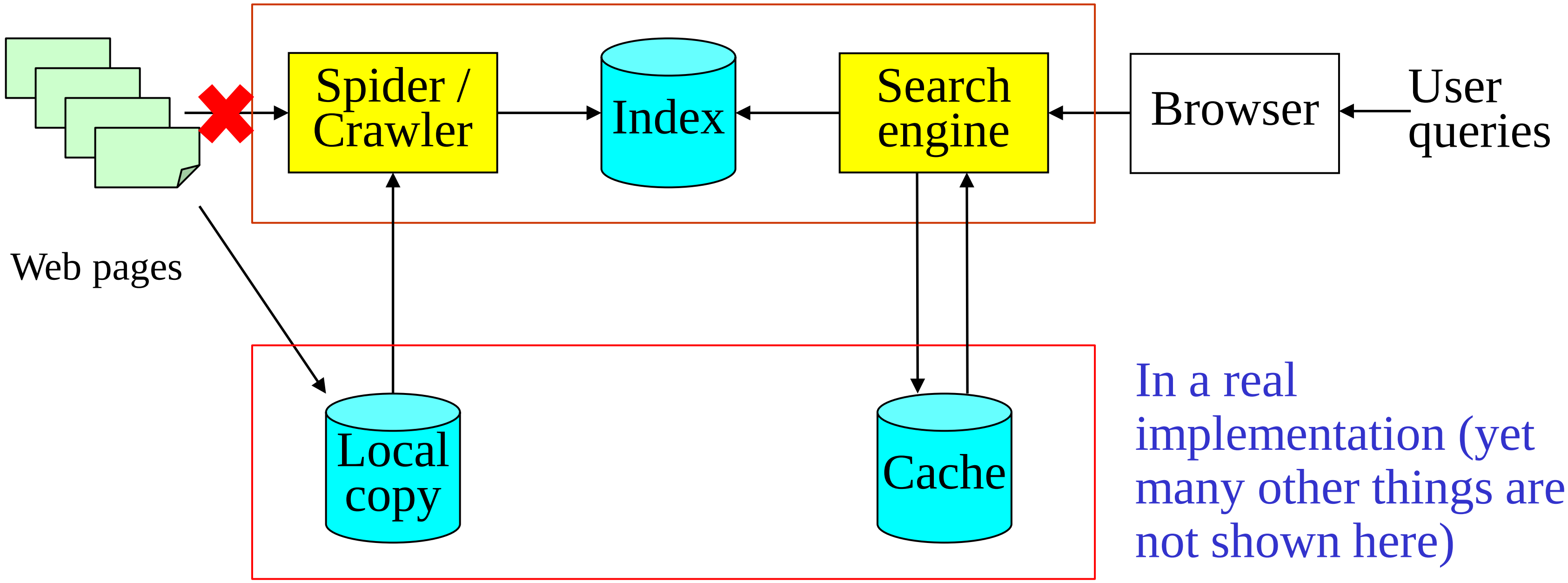
● Index Properties and File Contents

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)

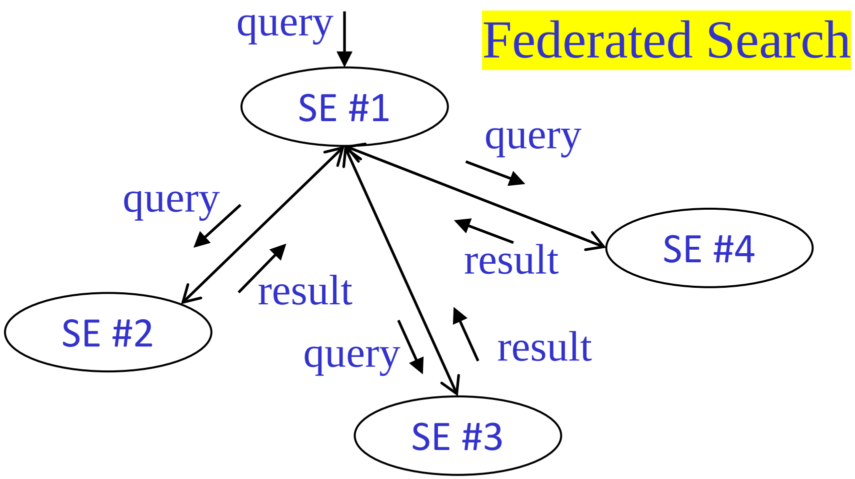
- World wide web search engines or Web Search
  - Most popular IR application nowadays, e.g., Google, Bing, Baidu
    - Other niche search engine DuckDuckGo, Yandex, etc.

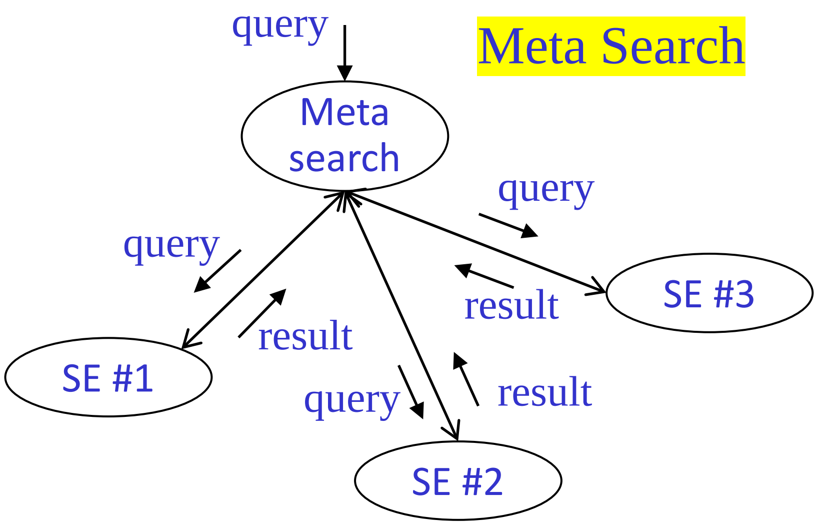


In a real implementation (yet many other things are not shown here)

# Federated and Meta Search

- Several search engines are used to return complete search results across all search engines
- A query will be dispatched to all search engines and search results are sent back to the originator, which will integrate the result





# Federated vs Meta Search

| Federated Search | Meta Search |
|---|---|
| Each node is a full-function SE for its own collection | Meta-search node passes queries and search results between users and underlying SEs; itself is not a SE |
| Search engines agree to join Federated Search | Agreement is not needed; unwilling SEs can block search requests |
| Agree to use the same standard for query/result representation and API (e.g., ANSI/ISO Z39.50 for libraries) | Query and results of underlying SEs can have different format; meta-search performs query transformation and data aggregation |
| SEs can collaborate to perform a search | Participating SEs do not collaborate with each other |
| E.g., HKALL (HK Academic Lib Link) | Dogpile.com |

# Differences from Web Search (GBB)

- Technologies for all these different forms of search are more or less the same, but in enterprise or product search
  - 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?

# Why is IR Important? Needed Everywhere!

- Most information available is in textual form and has no predefined format (e.g., emails and newsgroup 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 Difficult? Size!

- 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 you need to spend
“just one second more” on each page!
Renders Natural Language Processing methods infeasible

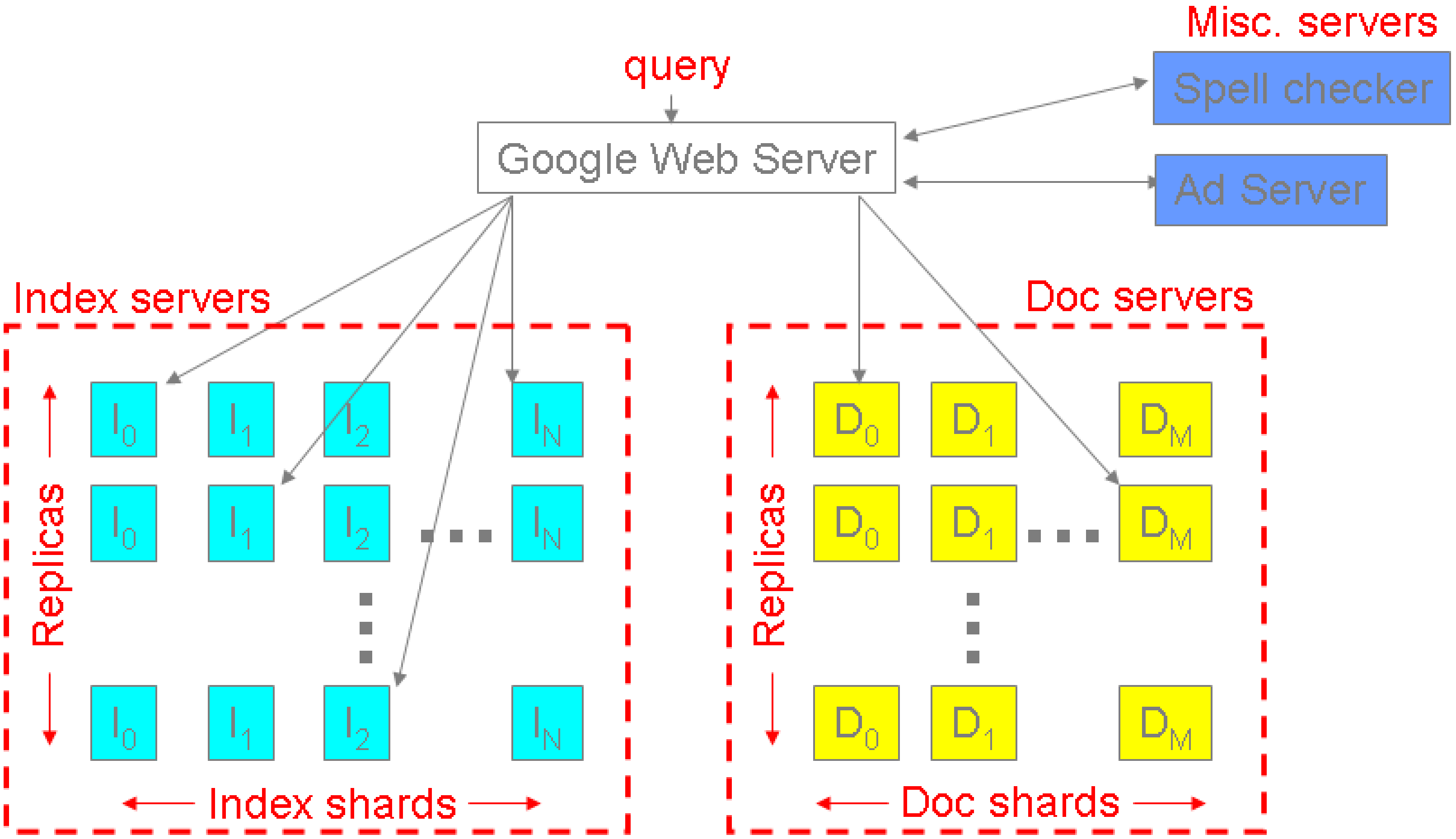
# Cloud Computing is Powerful: It can do what no PC can do

## Example: Google Search

- Is Google Search faster than search in Windows/Outlook/Word?
  - And Google Search must be much harder….
- How much storage does it take to store all of the web pages?
  - 100B pages * 10K per page = 1000T disk!
- Cloud computing has at its disposal
  - Essentially infinite amount of disk
  - Essentially infinite amount of computation
  - (Assuming they can be parallelized)

- Slide from a Google Presentation

# A Google Search Uses >1000 Machines Simultaneously



Elapsed time: 0.25s, machines involved: 1000+

# Why is IR Difficult? Semantics!

- 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? Diversity!

- 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!

# Indexing by Professionals (Librarians/Authors)

- High labor cost of trained human indexers
- Inconsistency in selecting index terms and judging relevance
  - thesauri created by two indexers in a given subject domain have only 60% of index terms in common
  - indexes obtained by two indexers from the same document with the same thesaurus have only 30% in common
  - documents obtained from two persons searching the same document set with the same question have only 40% in common
  - relevance judgments obtained by two users on the same set of documents and the same topic have only 60% in common
- Ref: Olson, Hope A., and Dietmar Wolfram. "Indexing consistency and its implications for information architecture: A pilot study." IA Summit (2006).

# Why is IR a Difficult Problem?

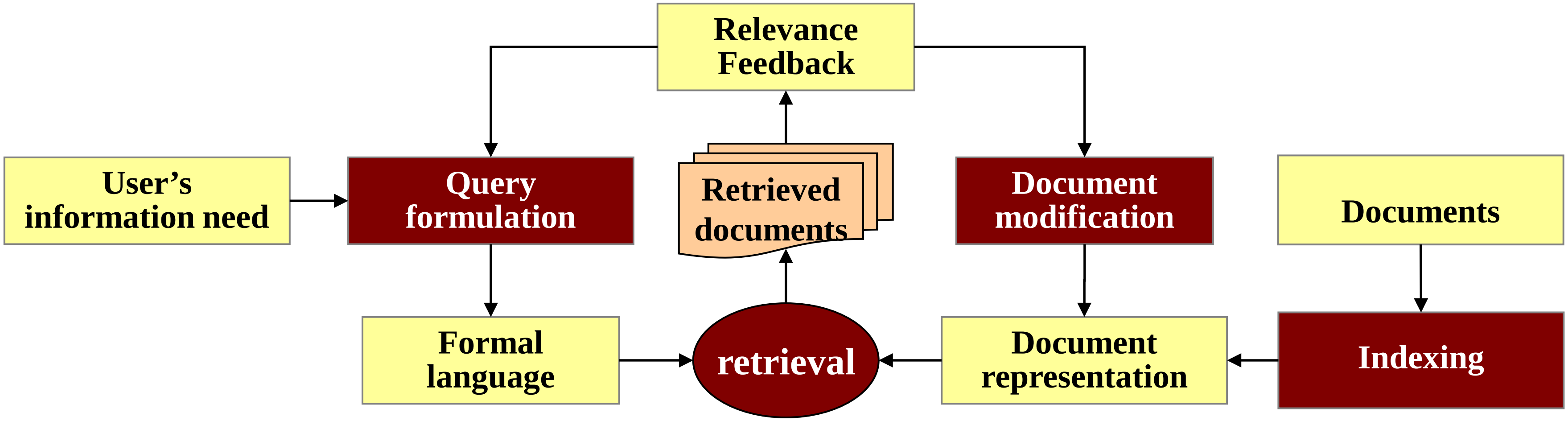
- Distributed and interlinked (e.g., Hypertext and WWW)
  - Where to start a search? Unlike in a centralize database, you have only one (or a few) database(s) to search.
  - How are the information related?

How fast

How good

- Efficiency vs. effectiveness
  - With limited resources, one can only improve efficiency and effectiveness to a certain degree.
  - 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 (mostly vector space model, which we will learn next) applied to web
  - Add a spider / crawler to download web pages
  - 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.)

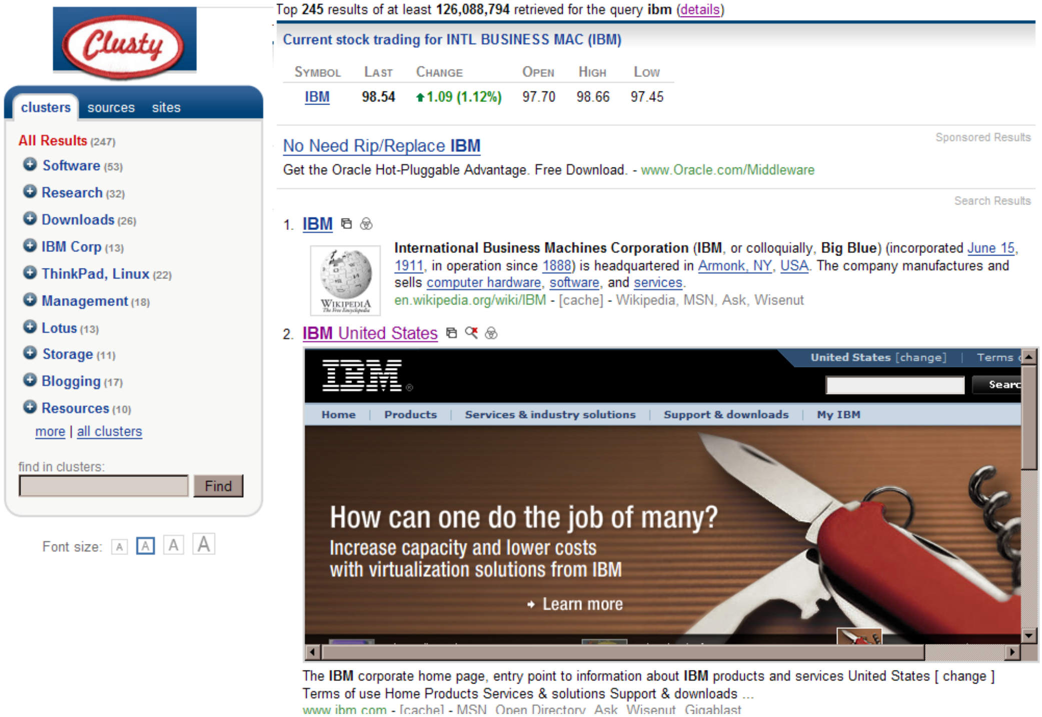
- Second-generation search engines (1997 - )
  - In addition to keyword matching, relying heavily on link analysis (thus capitalizing on the special property of web)
  - Using links to measure the quality of web page, thus fundamentally expand the dimension of ranking
  - Google, Fast (sold to Microsoft), 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 o
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)

- **GBB:** Global web search engines attract billions of searches every day; advertisement is the major source of revenue; technological competitiveness is a must (winner takes all!)
- **Enterprise search:** Companies deploy their own search engines to enhance productivity; vendors include Endeca (Oracle), Microsoft (SharePoint), and Google (Site/Custom Search)
- **Various vertical search:** Business directories, recruitment and travel web sties; advertisement is the largest source of revenue
- **Search engine marketing (SEM):** Marketing via search engine ad placements
- **Search engine optimization (SEO):** Companies helping websites to rank high in GBB

# 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, manual keyword extraction)
- 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); how is structured information represented in HTML?
- Search is difficult because it has to “understand” what the user wants through a few query keywords and retrieve 10 best pages out of billions of pages based on the semantic content of the pages
- In addition to sophistication of search, scaling up remains important
- High quality ranking at sub-second speed => Great **U**ser e**X**perience