

Introduction to Information Retrieval

Hongning Wang

CS@UVa

What is information retrieval?

The image shows a Google search interface for the query "what is information retrieval". The search results page displays "About 14,300,000 results (0.43 seconds)". A large snippet of text is highlighted with a red box, containing the following text: "Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text indexing. Automated information retrieval systems are used to reduce what has been called 'information overload'. Many universities and public libraries use IR syst...". Below this snippet, a red box highlights a link to a PDF titled "Introduction to Information Retrieval - The Stanford NLP", with a red arrow pointing from the highlighted text in the snippet to this link. Other search results are partially visible, including one from "www.iva.dk" and one from "Merriam-Webster".

What is information retrieval?

- Apple's vision 35 years ago

[Knowledge Navigator](#)

Why information retrieval

- Information overload
 - “It refers to the difficulty a person can have understanding an issue and making decisions that can be caused by the presence of too much information.” - wiki



Why information retrieval

- Information overload

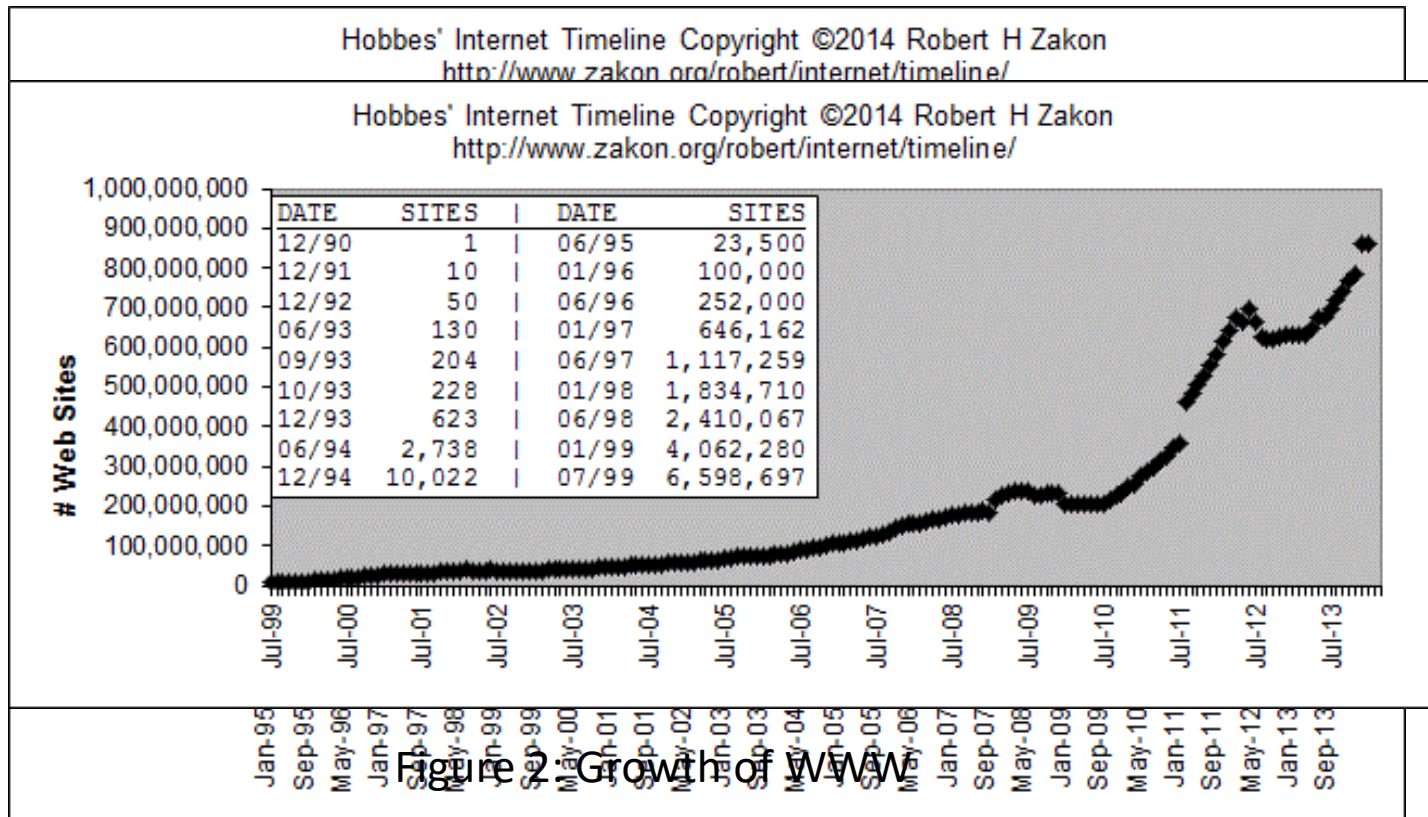


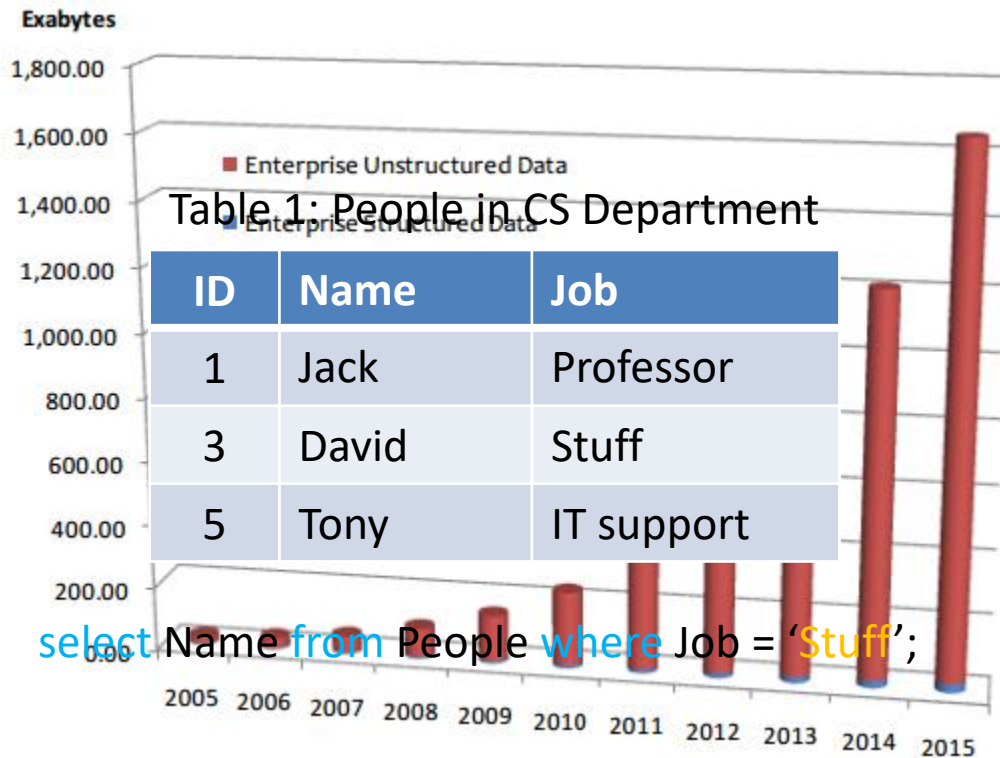
Figure 1: Growth of Internet

Why information retrieval

- Handling unstructured data
 - Structured data: database system is a good choice

– Unst

- Te
- “ξ
- UI
- U



lio, video...
as

Total Enterprise Data Growth 2005-2015, IDC 2012

Why information retrieval

- An essential tool to deal with information overload



You are here!

History of information retrieval

- Idea popularized in the pioneer article “***As We May Think***” by Vannevar Bush, 1945
 - “Wholly new forms of encyclopedias will appear, ready-made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified.” -> **WWW**
 - “A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility.” -> **Search engine**

Major research milestones

- Early days (late 1950s to 1960s): foundation of the field
 - Luhn's work on automatic indexing
 - Cleverdon's Cranfield evaluation methodology and index experiments
 - Salton's early work on SMART system and experiments
- 1970s-1980s: a large number of retrieval models
 - Vector space model
 - Probabilistic models
- 1990s: further development of retrieval models and new tasks
 - Language models
 - TREC evaluation
 - Web search
- 2000s-present: more applications, especially Web search and interactions with other fields
 - Learning to rank
 - Scalability (e.g., MapReduce)
 - Real-time search

History of information retrieval

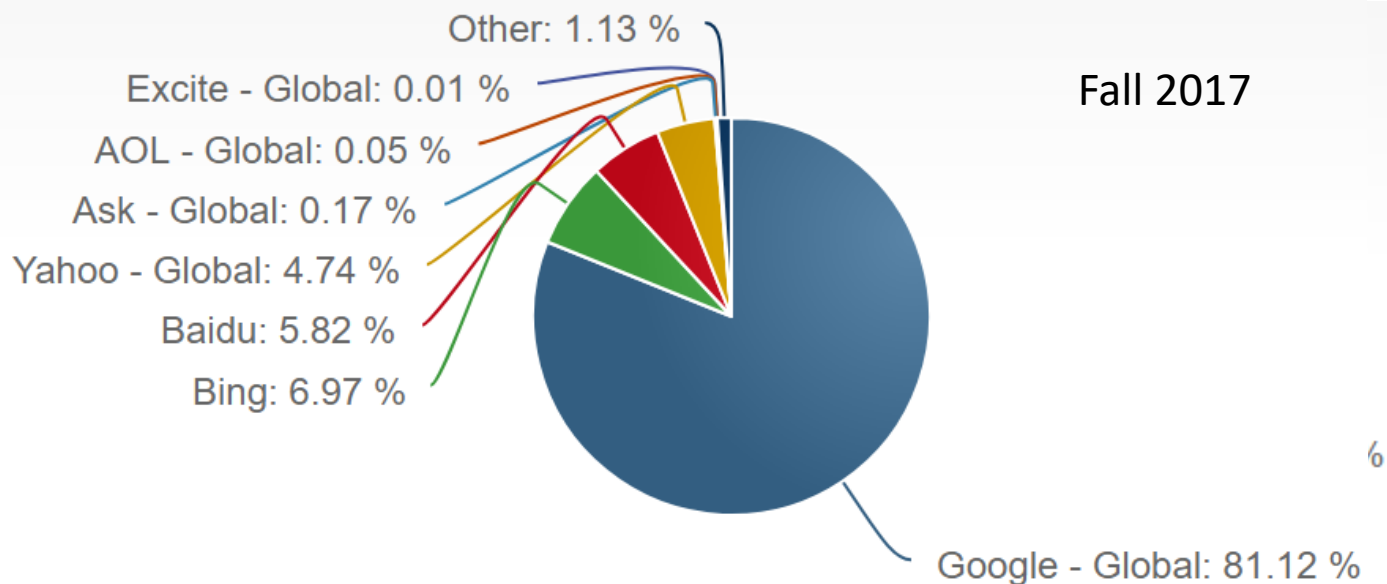
- Catalyst
 - Academia: Text Retrieval Conference (TREC) in 1992
 - *“Its purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies.”*
 - *“... about one-third of the improvement in web search engines from 1999 to 2009 is attributable to TREC. Those enhancements likely saved up to 3 billion hours of time using web search engines.”*
 - Till today, it is still a major test-bed for academic research in IR

History of information retrieval

- Catalyst
 - Industry: web search engines
 - WWW unleashed explosion of published information and drove the innovation of IR techniques
 - First web search engine: “*Oscar Nierstrasz at the University of Geneva wrote a series of Perl scripts that periodically mirrored these pages and rewrote them into a standard format.” Sept 2, 1993*”
 - Lycos (started at CMU) was launched and became a major commercial endeavor in 1994
 - Booming of search engine industry: *Magellan, Excite, Infoseek, Inktomi, Northern Light, AltaVista, Yahoo!, Google, and Bing*

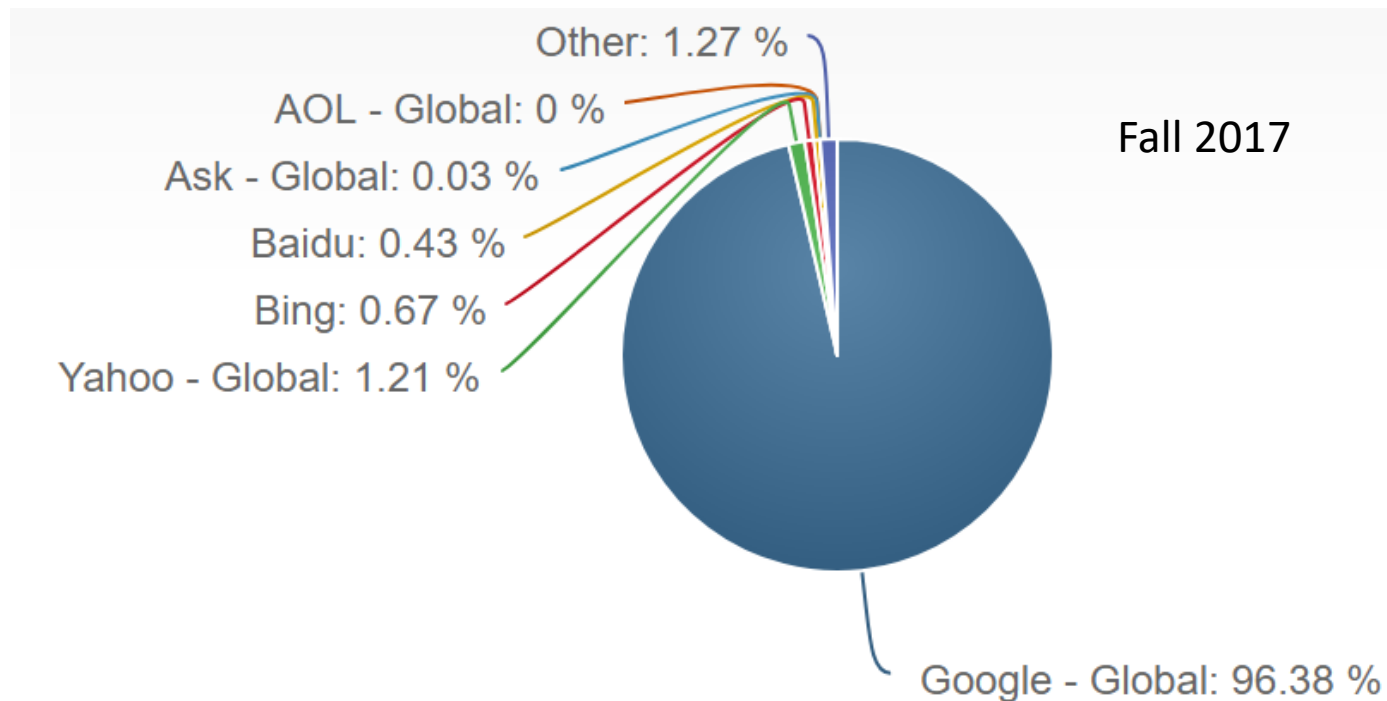
Major players in this game

- Global search engine market - desktop
 - By <http://marketshare.hitslink.com/search-engine-market-share.aspx>



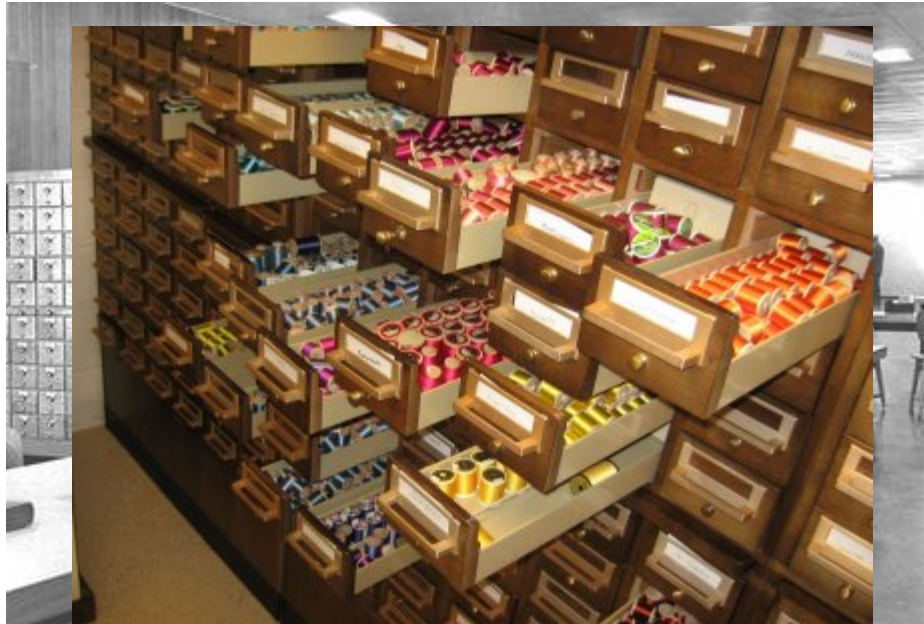
Major players in this game

- Global search engine market - mobile
 - By <http://marketshare.hitslink.com/search-engine-market-share.aspx>

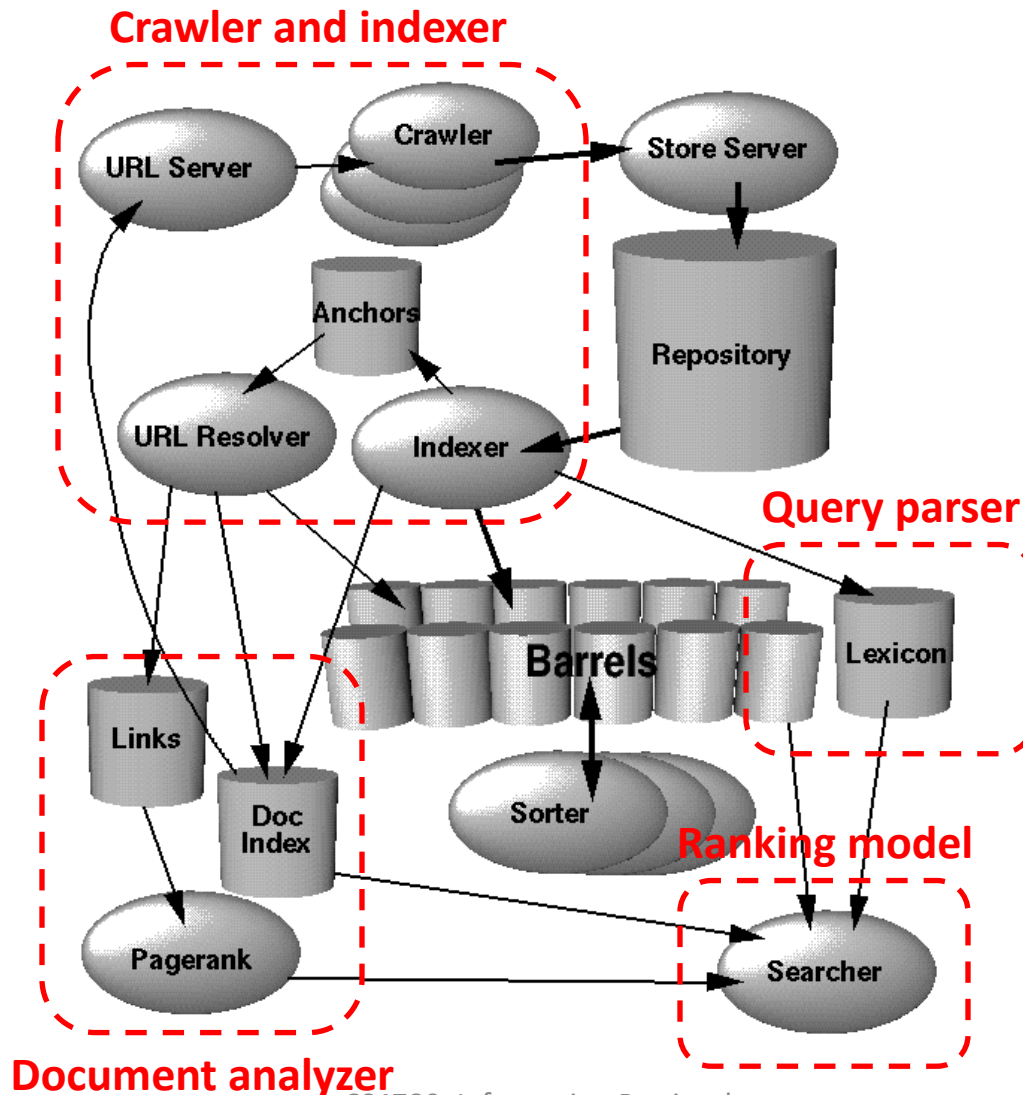


How to perform information retrieval

- Information retrieval when we did not have a computer

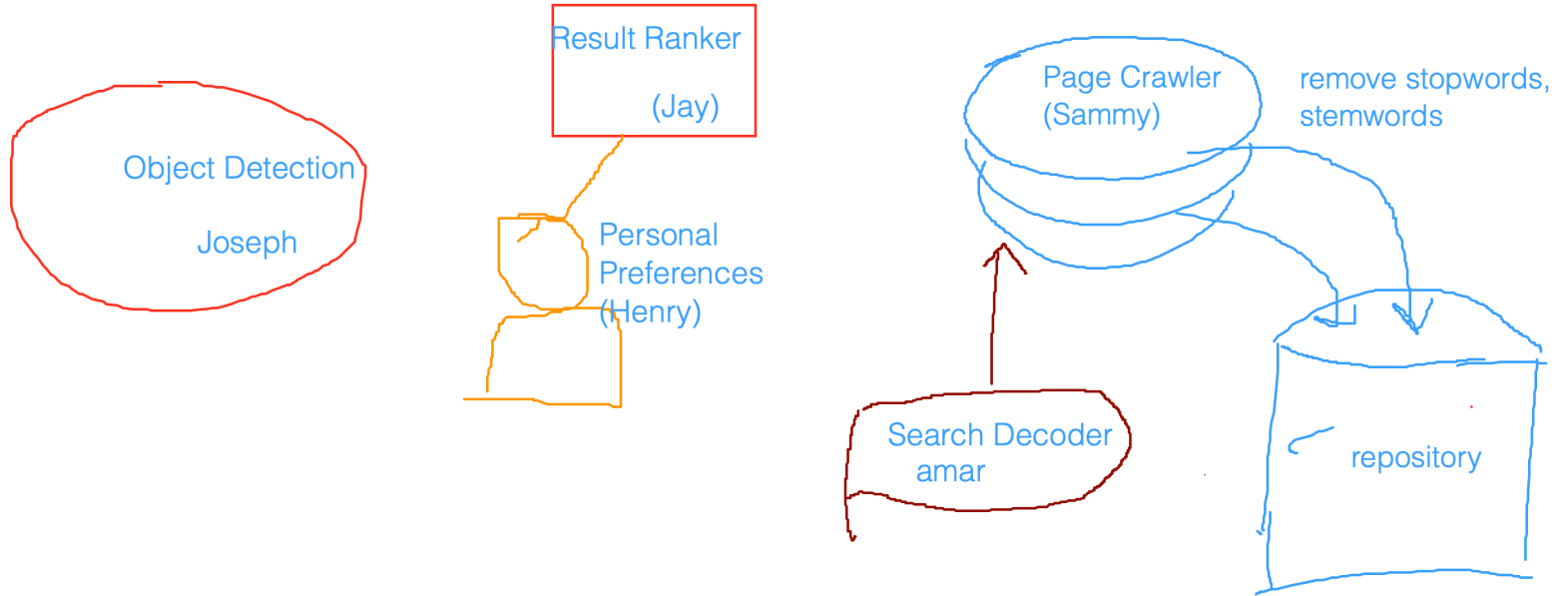


How to perform information retrieval

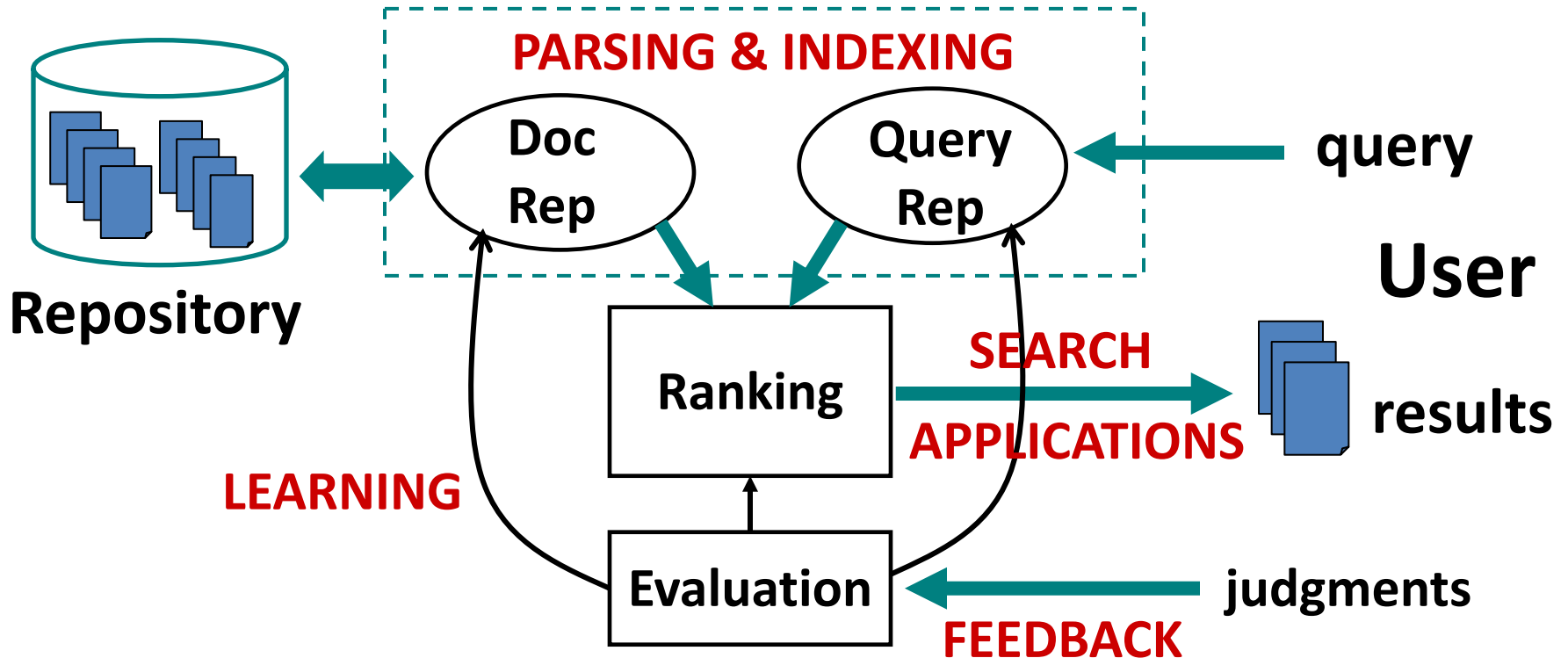


Crack into Google!

Crack into Google!



How to perform information retrieval



We will cover:

- 1) Search engine architecture;
- 2) Retrieval models;
- 3) Retrieval evaluation;
- 4) Relevance feedback;
- 5) Link analysis;
- 6) Search applications.

Core concepts in IR

- Query representation
 - Lexical gap: say v.s. said
 - Semantic gap: ranking model v.s. retrieval method
- Document representation
 - Special data structure for efficient access
 - Lexical gap and semantic gap
- Retrieval model
 - Algorithms that find the most relevant documents for the given information need

A glance of modern search engine

Yet Another Hierarchical Official/Obstreperous/Odiferous/Organized Oracle

- In old times



Yahoo race of fictional beings from Gulliver's Travels



Arts & Humanities Literature, Photography...	News & Media Full Coverage, Newspapers, TV...
Business & Economy Companies, Finance, Jobs...	Recreation & Sports Sports, Travel, Autos, Outdoors...
Computers & Internet Internet, WWW, Software, Games...	Reference Libraries, Dictionaries, Quotations...
Education College and University, K-12...	Regional Countries, Regions, US States...
Entertainment Cool Links, Movies, Humor, Music...	Science Animals, Astronomy, Engineering...



A glance of modern search engine

Google | uva | Demand of understanding

Web Maps Images News Shopping More Search tools

About 103,000,000 results (0.65 seconds) | Demand of efficiency Demand of convenience

The Universi
www.virginia.edu
The University of \n Jefferson. The corn
4.9 ★★★★★ 21

University of
en.wikipedia.org/M
The University of \n research university

University of
colleges.usnews.r
Is University of Vir
University of Virgi

VIRGINIASPO...
www.virginiasp...
The University of Virginia Official Athletic Site, partner of CBSSports.com College Network. The most comprehensive coverage of UVA Cavaliers Athletics on the ...

Images for university of virginia Report images

Google

Google Search I'm Feeling Lucky

Directions
Charlottesville, President

Acceptance rate: 28.3% (2013)
Enrollment: 21,095 (2012)
Mascot: University of Virginia Cavalier
Founder: Thomas Jefferson
Founded: 1819, Charlottesville, VA
Colors: Blue, Orange

Recent posts
#UVA's Center for Politics and Politico have teamed up to offer interactive election ratings. #politics #elections #voting 1 hour ago

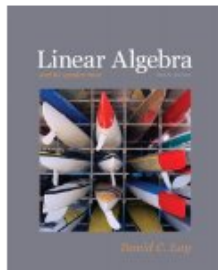
CS@UVA More images for university of virginia CS4780: Information Retrieval

Demand of diversity

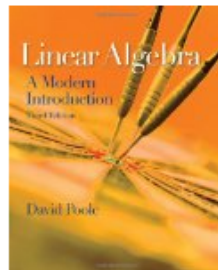
IR is not just about web search

- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
 - Recommendation

Recommended Based on Your Browsing History



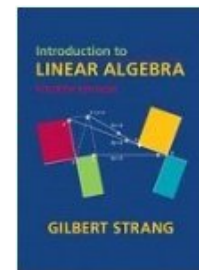
Linear Algebra and Its Applications...
› David C. Lay
Hardcover
★★★★☆ (84)
~~\$483.33~~ \$141.16



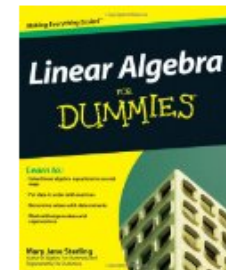
Linear Algebra: A Modern Introduction
› David Poole
Hardcover
★★★★☆ (41)
~~\$346.95~~ \$289.88



Linear Algebra
› G. E. Shilov
Paperback
★★★★☆ (34)
~~\$48.95~~ \$12.65



Introduction to Linear Algebra...
› Gilbert Strang
Hardcover
★★★★☆ (57)
~~\$87.50~~ \$83.13



Linear Algebra For Dummies
› Mary Jane Sterling
Paperback
★★★★☆ (29)
~~\$49.99~~ \$16.23

IR is not just about web search

SECTIONS

HOME SEARCH

The New York Times

TECHNOLOGY

Google and Walmart Partner With Eye on Amazon

By DAISUKE WAKABAYASHI and MICHAEL CORKERY AUG. 23, 2017



Walmart has also been trying to integrate its digital business with its vast network of more than 4,690 stores. It partnered with Google to take on Amazon, the heavyweight of online shopping.
Roger Kisby for The New York Times

RELATED COVERAGE



At Walmart Academy, Training Better Managers. But With a Better Future?
AUG. 8, 2017



Google, Lagging Amazon, Races Across the Threshold Into the Home
OCT. 2, 2016



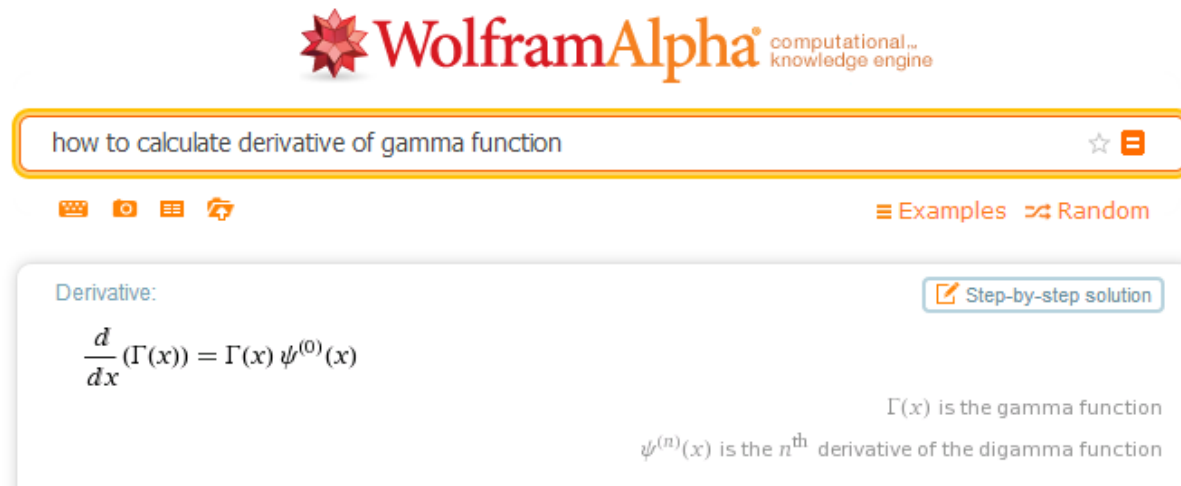
ECONOMIC TRENDS
The Amazon-Walmart Showdown That Explains the Modern Economy
JUNE 16, 2017



Walmart Rewrites Its E-Commerce Strategy With \$3.3 Billion Deal for Jet.com
AUG. 8, 2016

IR is not just about web search

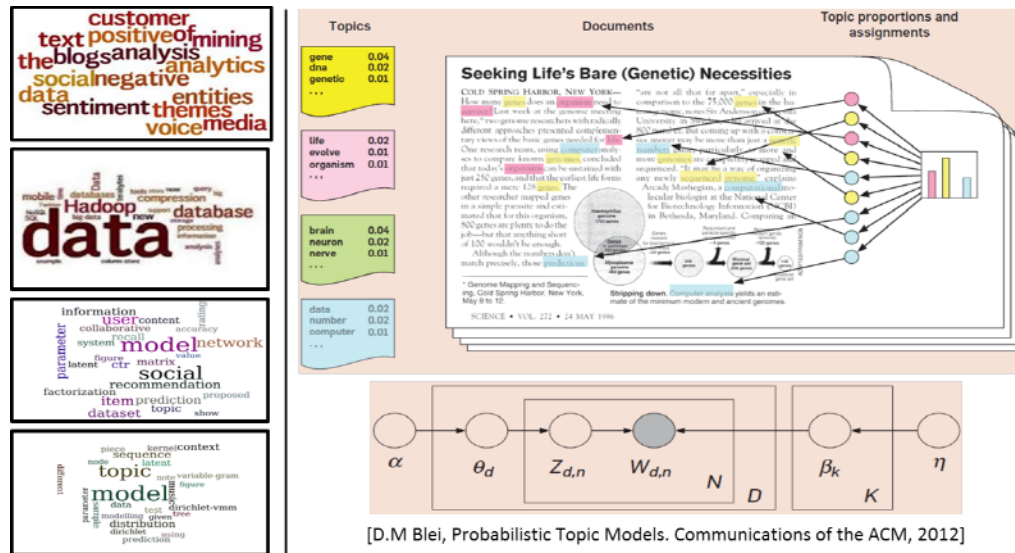
- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
 - Question answering



The image shows a screenshot of the WolframAlpha website. At the top, the WolframAlpha logo is displayed with the tagline 'computational... knowledge engine'. Below the logo is a search bar containing the text 'how to calculate derivative of gamma function'. To the right of the search bar are icons for a star and a menu. Below the search bar are several icons for different input methods (keyboard, voice, image, etc.) and links for 'Examples' and 'Random'. The main content area shows the result for the derivative of the gamma function. It starts with the text 'Derivative:' followed by the mathematical equation $\frac{d}{dx}(\Gamma(x)) = \Gamma(x) \psi^{(0)}(x)$. To the right of the equation is a button labeled 'Step-by-step solution'. Below the equation, there are two explanatory lines: ' $\Gamma(x)$ is the gamma function' and ' $\psi^{(n)}(x)$ is the n^{th} derivative of the digamma function'.

IR is not just about web search

- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
 - Text mining



[D.M Blei, Probabilistic Topic Models. Communications of the ACM, 2012]

IR is not just about web search

- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
 - Online advertising

The screenshot shows the Yahoo! homepage with a search bar at the top. Below the search bar, there are several advertisements for Microsoft products, including Surface Pro 4, Surface Book, and Surface Dial. A red dashed box highlights these ads. Below the ads, there is a news article titled "Carrier says it has deal with Trump on jobs" with a photo of Donald Trump and other people. To the right of the news article, there is a "Trending Now" section with a list of items. Below the trending section, there are more ads for Microsoft products, also highlighted with a red dashed box. The left sidebar contains various navigation links like Mail, News, Finance, Sports, etc.

Microsoft Free shipping everyday

Product	Price	Shop Now
Microsoft Surface Pro 4 - 128GB / Intel Core i5	\$999	SHOP NOW
Surface Book with Performance Base - 256GB / Intel Core i7	\$2,399	SHOP NOW
Microsoft Surface Pro 4 - 128GB / Intel Co...	\$549	SHOP NOW
Microsoft Surface Book - 128GB / Intel Core i5	\$1,499	SHOP NOW
Surface Dial	\$99.99	SHOP NOW
Microsoft Surface Pro 4 - 256GB / Intel Core i5	\$1,299	SHOP NOW

Carrier says it has deal with Trump on jobs
The air conditioning company says an agreement struck with the president-elect will keep almost 1,000 jobs in Indiana. [Key campaign pledge »](#)

Trending Now

1. Amber Rose
2. John Cena
3. Denver Broncos
4. Kate Hudson
5. iPad mini
6. Senior Independen...
7. Mariah Carey
8. Buy Auto Tires
9. Jobs Hiring Imme...
10. Reese Witherspoon

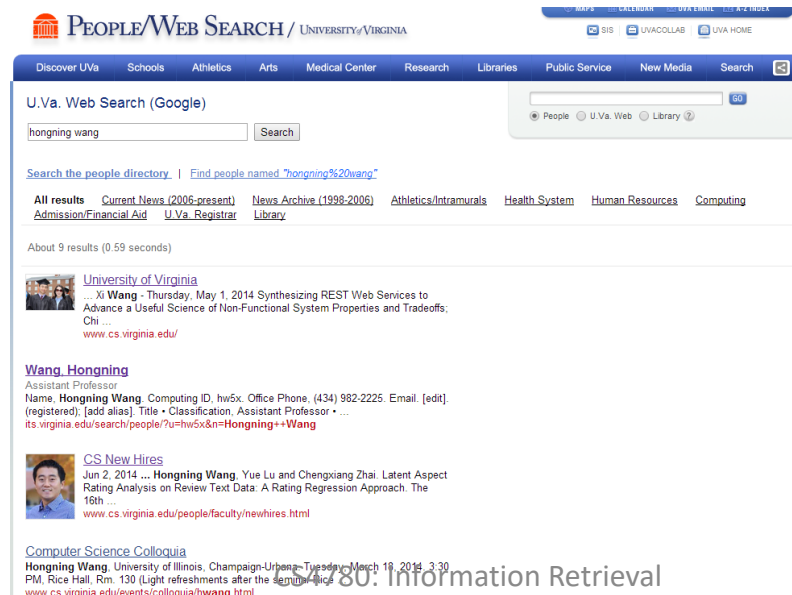
Microsoft Free shipping everyday

Product	Price	Shop Now
Surface Book with Performance...	\$2,399	SHOP NOW
Surface Dial	\$99.99	SHOP NOW
Microsoft Surface Pro 4 - 128GB / Intel...	\$999	SHOP NOW

Politics
Michigan Certifies Trump as Winner of State's Presidential Race

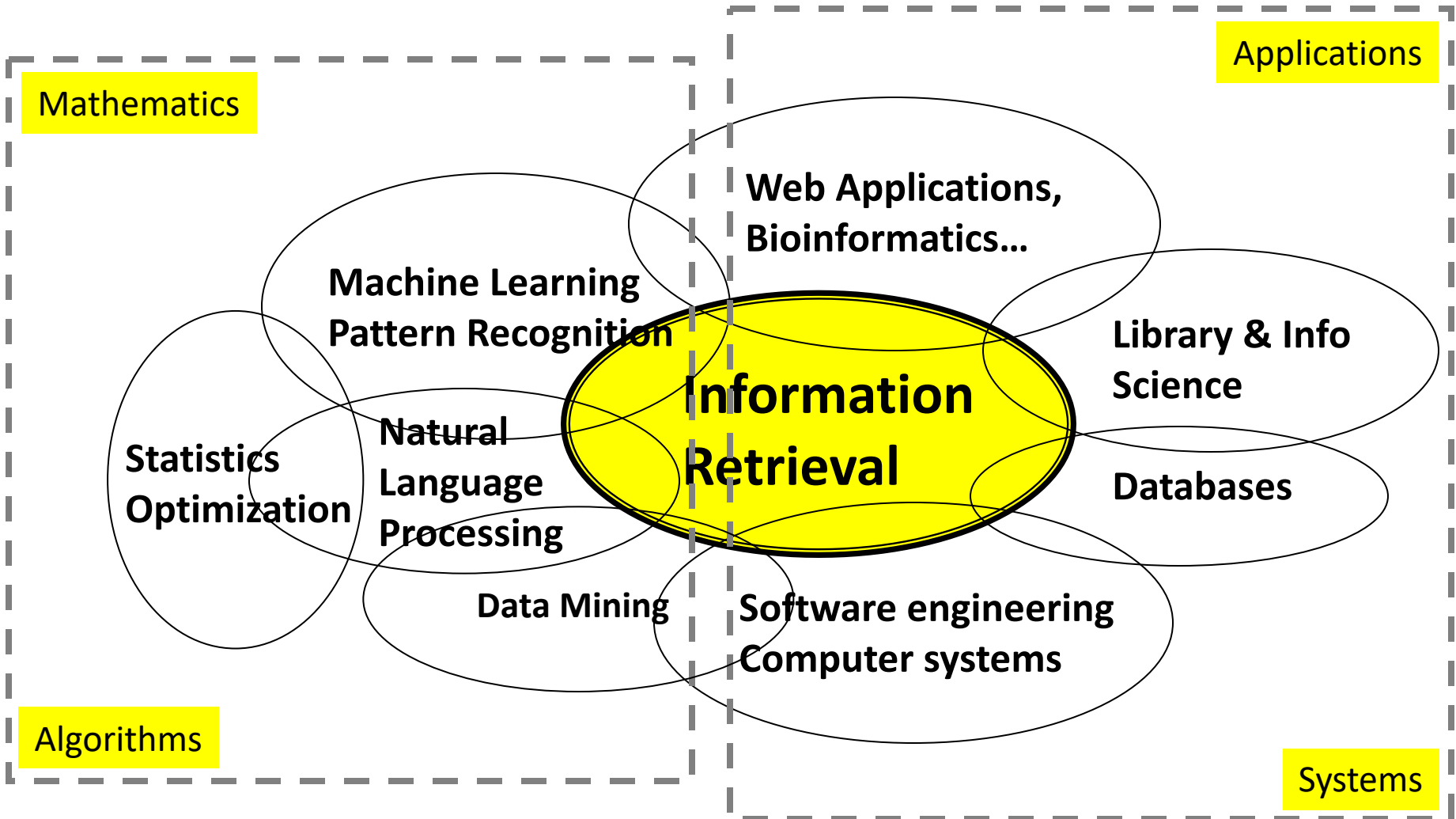
IR is not just about web search

- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
 - Enterprise search: web search + desktop search



The screenshot displays the 'PEOPLE/WEB SEARCH / UNIVERSITY OF VIRGINIA' interface. At the top, there are navigation links for 'Discover UVA', 'Schools', 'Athletics', 'Arts', 'Medical Center', 'Research', 'Libraries', 'Public Service', 'New Media', and 'Search'. Below this is a search bar with the text 'U.Va. Web Search (Google)' and a search button. The search term 'hongning wang' is entered in the search bar. Below the search bar, there are several tabs: 'All results', 'Current News (2006-present)', 'News Archive (1998-2006)', 'Athletics/Intramurals', 'Health System', 'Human Resources', and 'Computing'. The search results show 'About 9 results (0.59 seconds)'. The first result is for the 'University of Virginia' website, dated Thursday, May 1, 2014, with the title 'Synthesizing REST Web Services to Advance a Useful Science of Non-Functional System Properties and Tradeoffs'. The second result is for 'Wang, Hongning', an Assistant Professor, with a link to his profile page. The third result is for 'CS New Hires', dated Jun 2, 2014, with the title 'Hongning Wang, Yue Lu and Chengxiang Zhai. Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach. The 16th ...'. The fourth result is for 'Computer Science Colloquia', dated March 18, 2015, with the title 'Hongning Wang, University of Illinois, Champaign-Urbana. Tuesday, March 18, 2015, 3:30 PM, Rice Hall, Rm. 130 (Light refreshments after the seminar)'. The search results are displayed in a clean, organized layout with clear headings and links.

Related Areas



IR v.s. DBs

- Information Retrieval:
 - Unstructured data
 - Semantics of objects are subjective
 - Simple keyword queries
 - Relevance-drive retrieval
 - Effectiveness is primary issue, though efficiency is also important
- Database Systems:
 - Structured data
 - Semantics of each object are well defined
 - Structured query languages (e.g., SQL)
 - Exact retrieval
 - Emphasis on efficiency

IR and DBs are getting closer

- IR => DBs

- Approximate search is available in DBs
- Eg. in MySQL

```
mysql> SELECT * FROM articles  
-> WHERE MATCH (title,body)  
AGAINST ('database');
```

- DBs => IR

- Use information extraction to convert unstructured data to structured data, e.g., knowledge base
- Semi-structured representation: XML data; queries with structured information

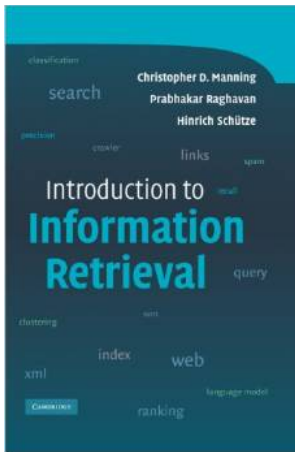
IR v.s. NLP

- Information retrieval
 - Computational approaches
 - Statistical (shallow) understanding of language
 - Handle large scale problems
- Natural language processing
 - Cognitive, symbolic and computational approaches
 - Semantic (deep) understanding of language
 - (often times) small scale problems

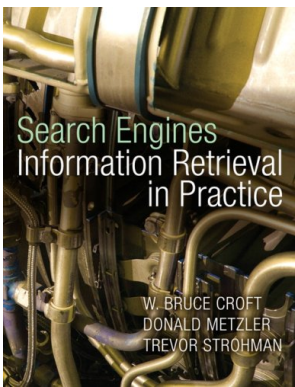
IR and NLP are getting closer

- IR => NLP
 - Larger data collections
 - Scalable/robust NLP techniques, e.g., translation models
- NLP => IR
 - Deep analysis of text documents and queries
 - Information extraction for structured IR tasks
 - Natural language based QA systems

Text books

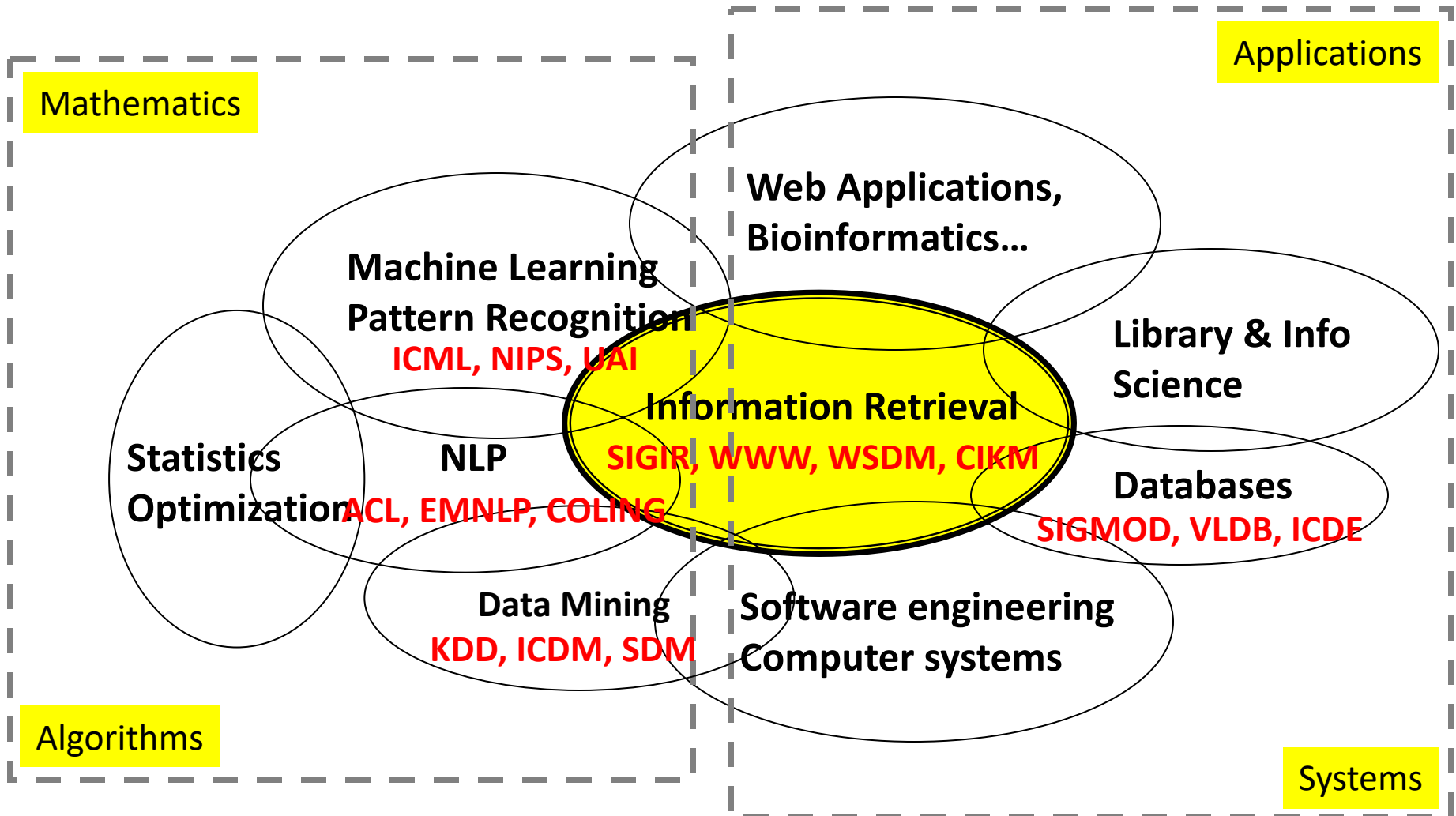


- ***Introduction to Information Retrieval.*** Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.



- ***Search Engines: Information Retrieval in Practice.*** Bruce Croft, Donald Metzler, and Trevor Strohman, Pearson Education, 2009.

What to read?



- Find more on course website for resource

IR in future

- Mobile search
 - Desktop search + location? Not exactly!!
- Interactive retrieval
 - Machine collaborates with human for information access
- Personal assistant
 - Proactive information retrieval
 - [Knowledge navigator](#)
- And many more
 - You name it!

What you should know

- IR originates from library science for handling unstructured data
- IR has many important application areas, e.g., web search, recommendation, and question answering
- IR is a highly interdisciplinary area with DBs, NLP, ML, HCI

Today's reading

- *Bush, Vannevar. "As we may think." The atlantic monthly 176, no.1 (1945): 101-108.*
- Introduction to Information Retrieval
 - Chapter 1: Boolean Retrieval