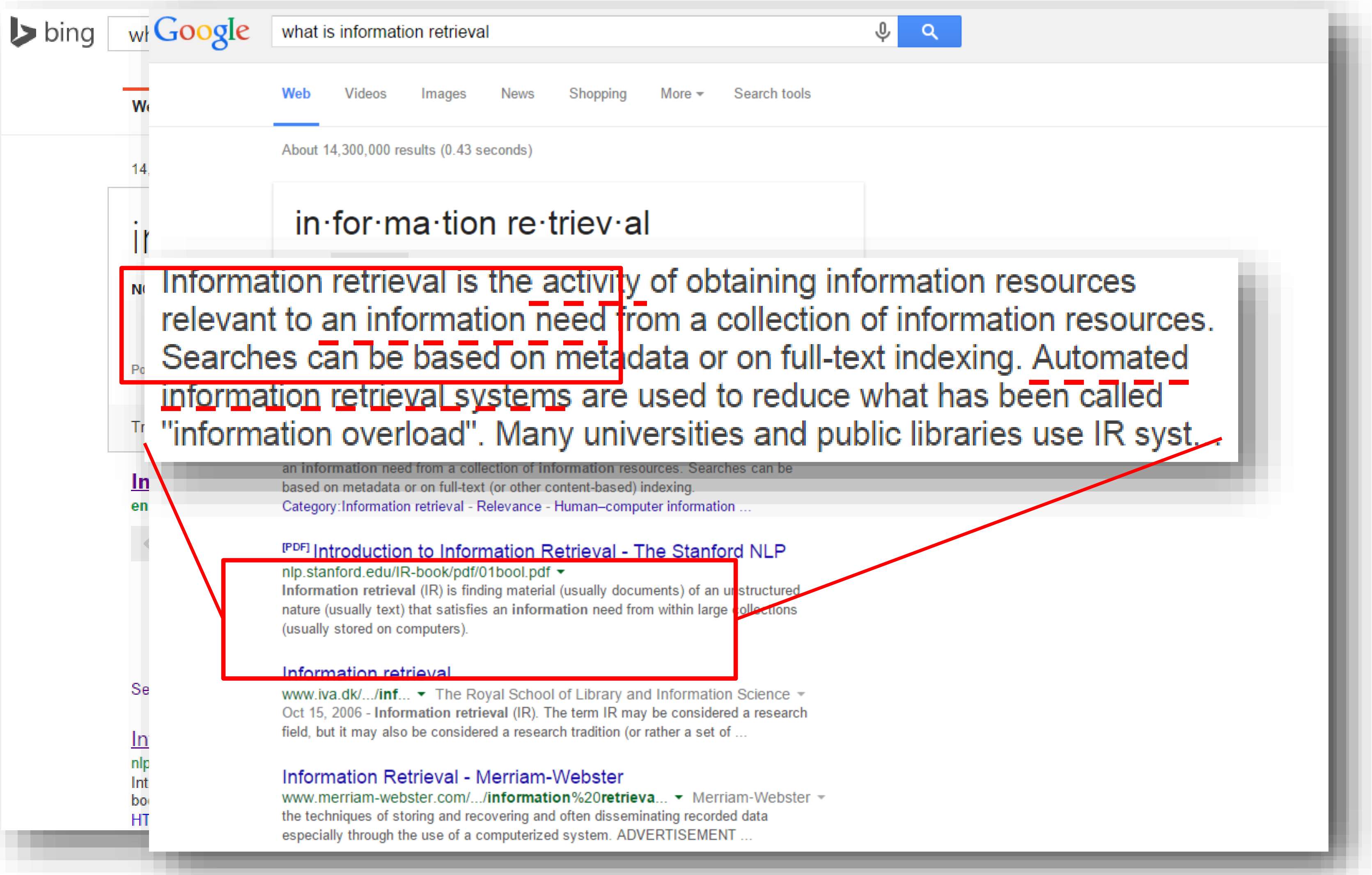
Introduction to Information Retrieval

Hongning Wang

CS@UVa

What is information retrieval?



what is information retrieval

About 14,300,000 results (0.43 seconds)

in·for·ma·tion re·triev·al

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...

an information need from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing.
Category:Information retrieval - Relevance - Human–computer information ...

[PDF] Introduction to Information Retrieval - The Stanford NLP
nlp.stanford.edu/IR-book/pdf/01bool.pdf
Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Information retrieval
www.iva.dk/.../inf... The Royal School of Library and Information Science
Oct 15, 2006 - Information retrieval (IR). The term IR may be considered a research field, but it may also be considered a research tradition (or rather a set of ...

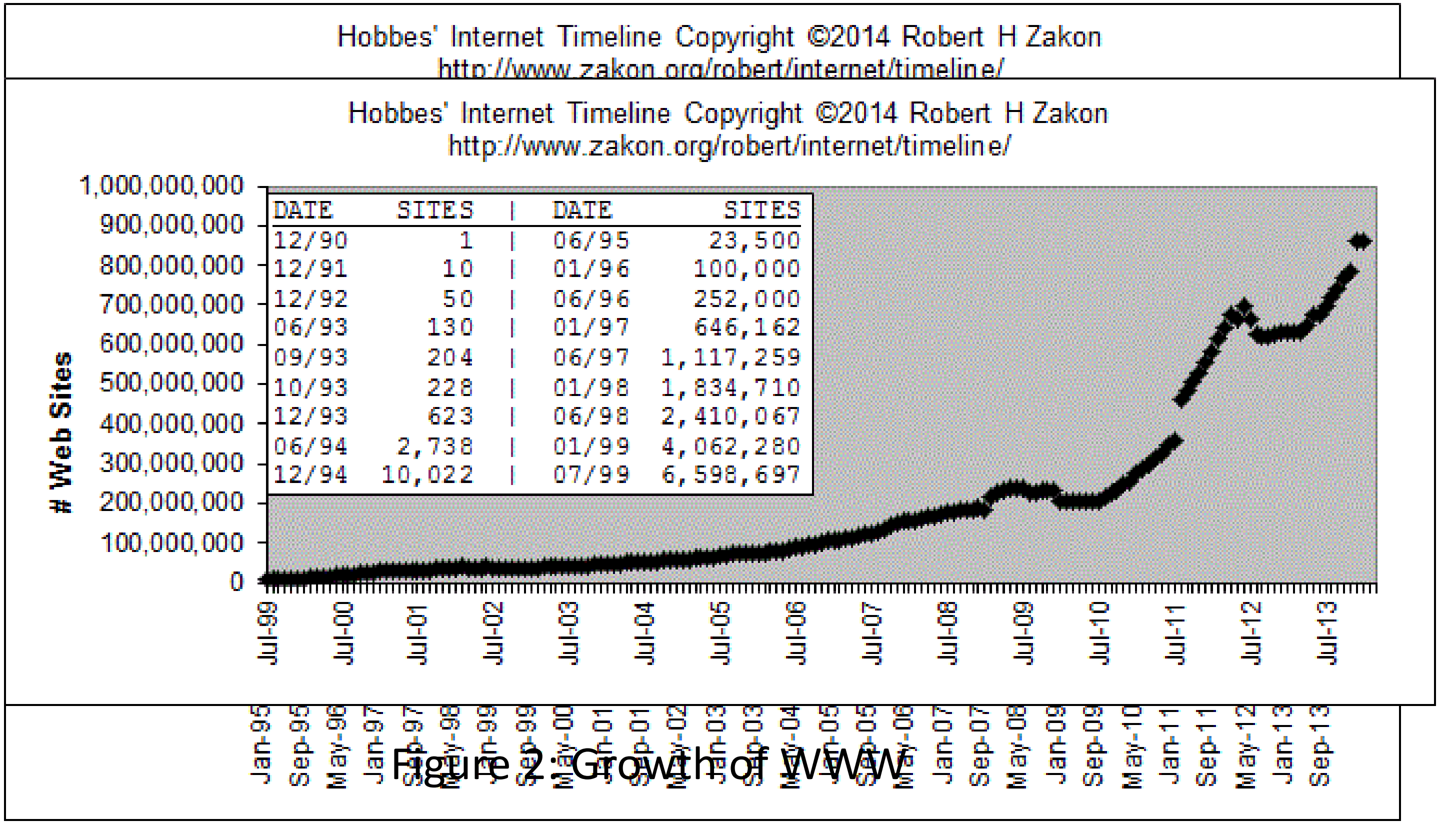
Information Retrieval - Merriam-Webster
www.merriam-webster.com/.../information%20retrieva... Merriam-Webster
the techniques of storing and recovering and often disseminating recorded data especially through the use of a computerized system. ADVERTISEMENT ...

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



DATE	SITES	DATE	SITES
12/90	1	06/95	23,500
12/91	10	01/96	100,000
12/92	50	06/96	252,000
06/93	130	01/97	646,162
09/93	204	06/97	1,117,259
10/93	228	01/98	1,834,710
12/93	623	06/98	2,410,067
06/94	2,738	01/99	4,062,280
12/94	10,022	07/99	6,598,697

Figure 2: Growth of WWW

Figure 1: Growth of Internet

Why information retrieval

- Handling unstructured data
  - Structured data: database system is a good choice
  - Unst
    - Te lio, video…
    - “ as ui
    - U

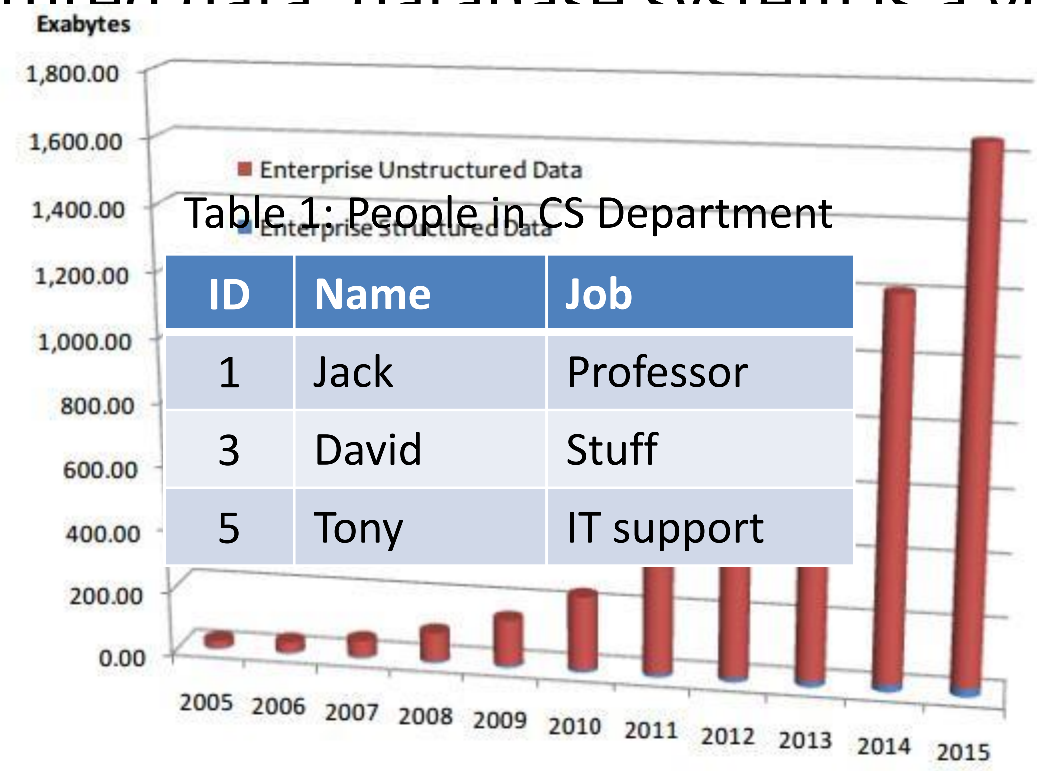


Table 1: People in CS Department

ID	Name	Job
1	Jack	Professor
3	David	Stuff
5	Tony	IT support

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

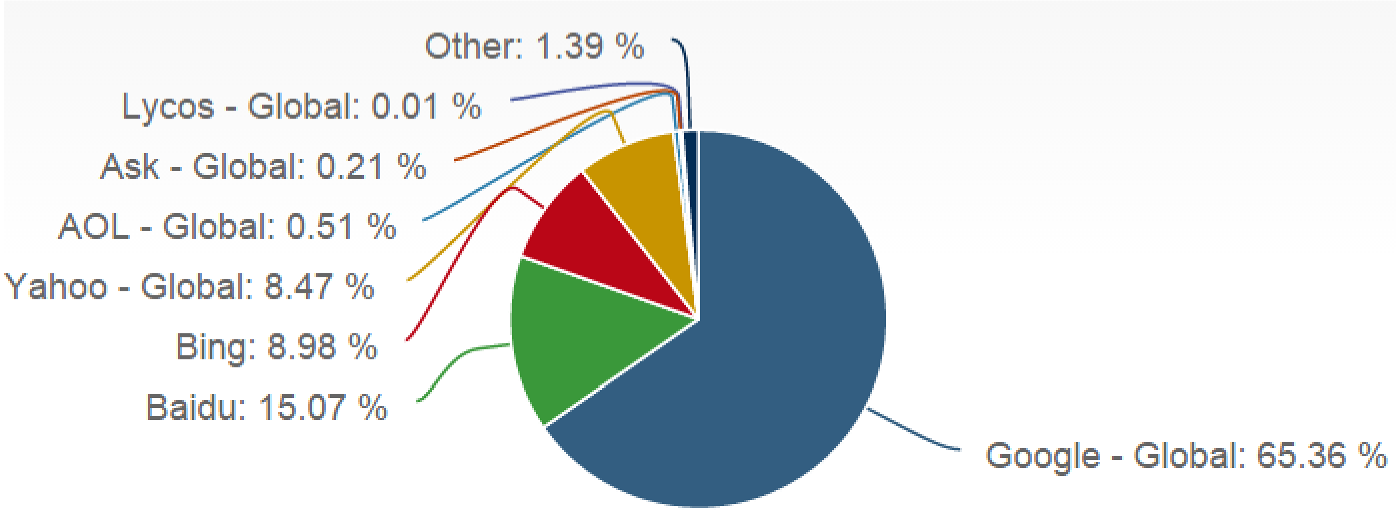
- Catalyst
  - Academia: Text Retrieval Conference (TREC) in 1992
    - “*Its purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale* *__evaluation__* *of text retrieval methodologies.*”
    - “*… about* *__one-third__* *of 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*

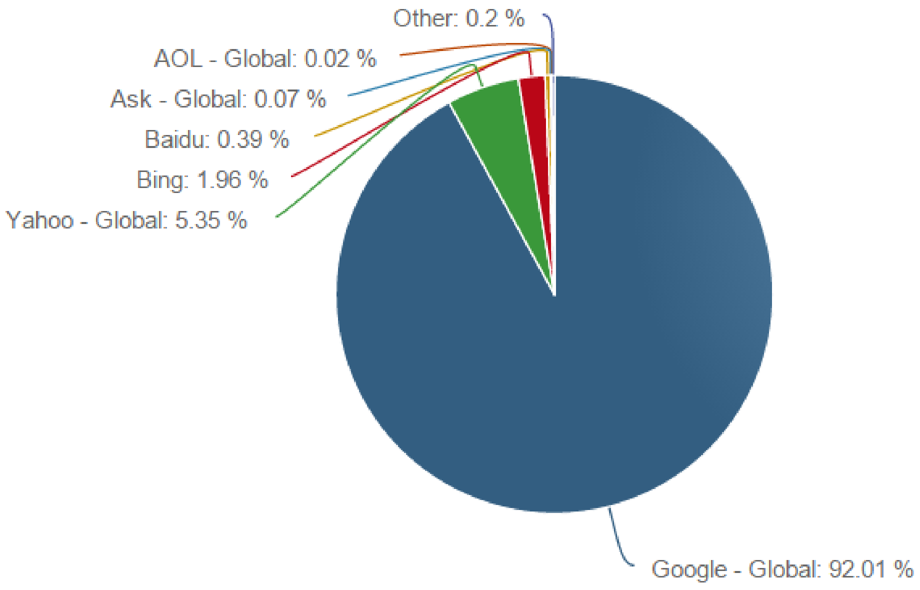
Major players in this game

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



Major players in this game

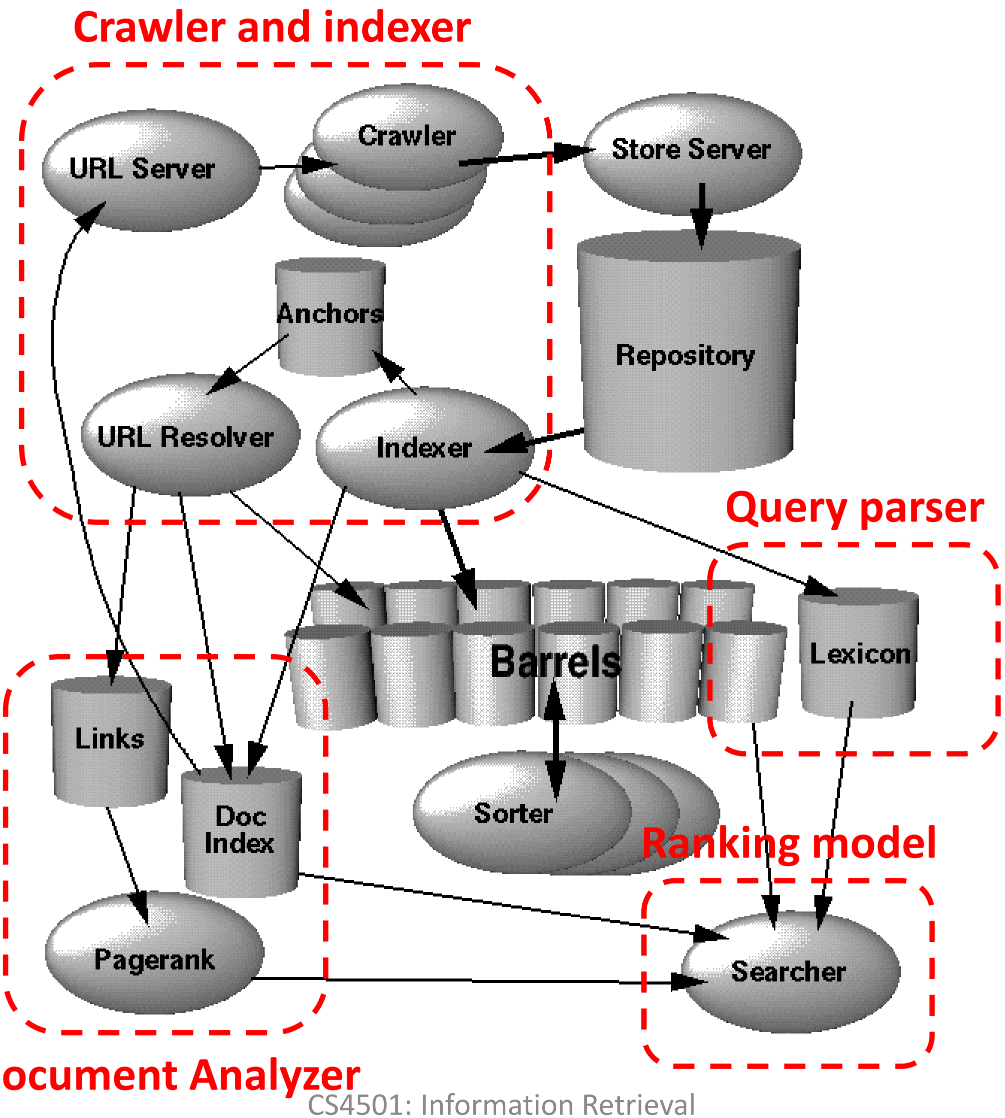
- Global search engine market - mobile
  - By http://marketshare.hitslink.com/search-engine-market-share.aspx



How to perform information retrieval

- Information retrieval when we did not have a computer

How to perform information retrieval



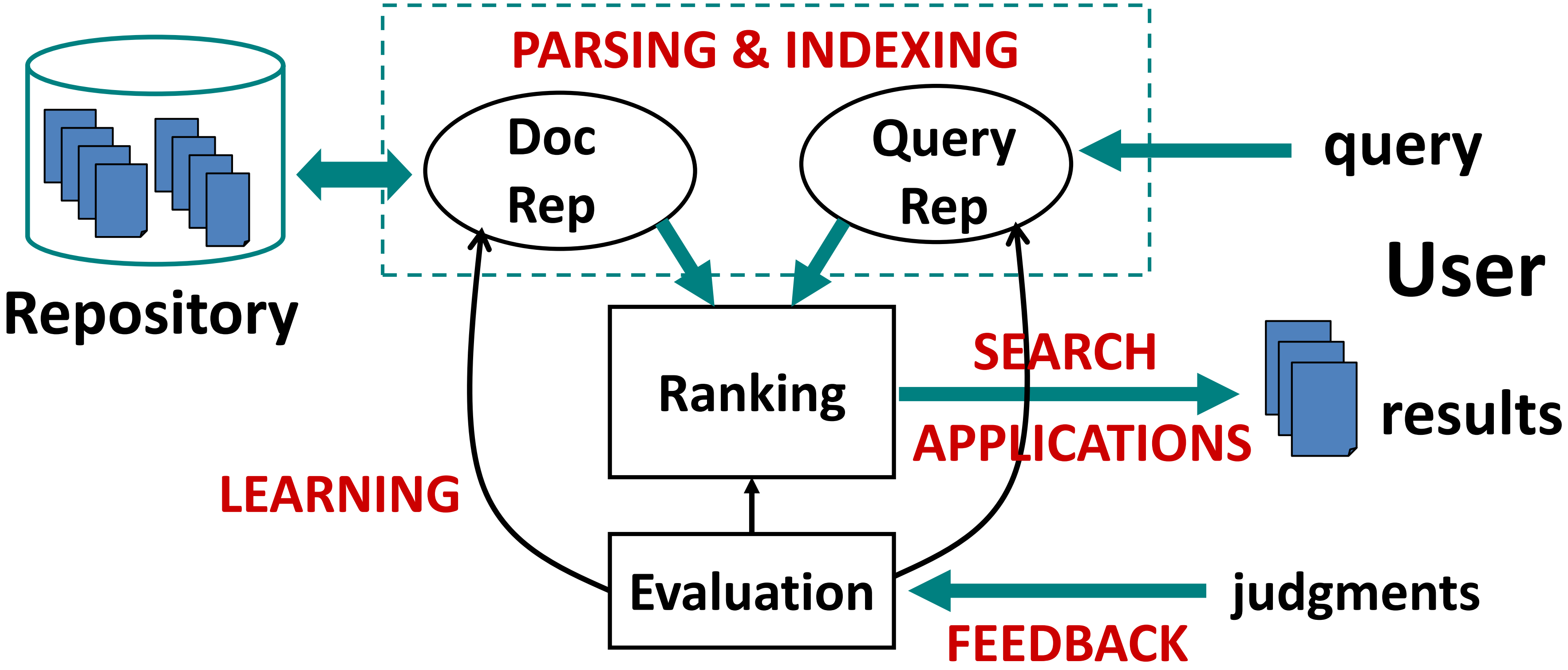
Crawler and indexer

Query parser

Ranking model

Document Analyzer

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.

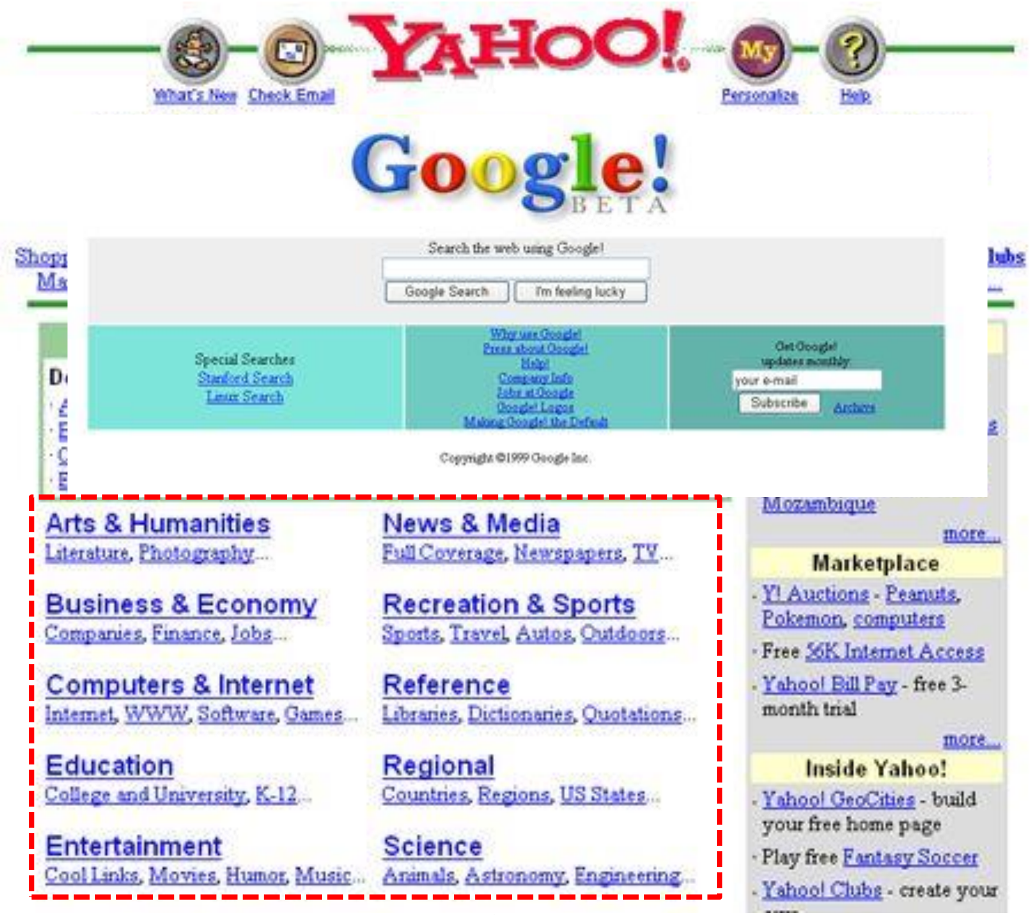
Core concepts in IR

- Query representation
  - Lexical gap: say v.s. said
  - Semantic gap: ranking model v.s. retrieval method
- Document representation
  - Specific 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

- In old times

*Yet Another **Hierarchical** Officious/Obstreperous/ Odiferous/Organized **Oracle***



Search the web using Google!

Google Search | I'm feeling lucky

Special Searches
Stanford Search
Linux Search

Why use Google!
Press about Google!
Help!
Company Info
Jobs at Google
Google! Logos
Making Google! the Default

Get Google! updates monthly:
your e-mail
Subscribe
Archive

Copyright ©1999 Google Inc.

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...

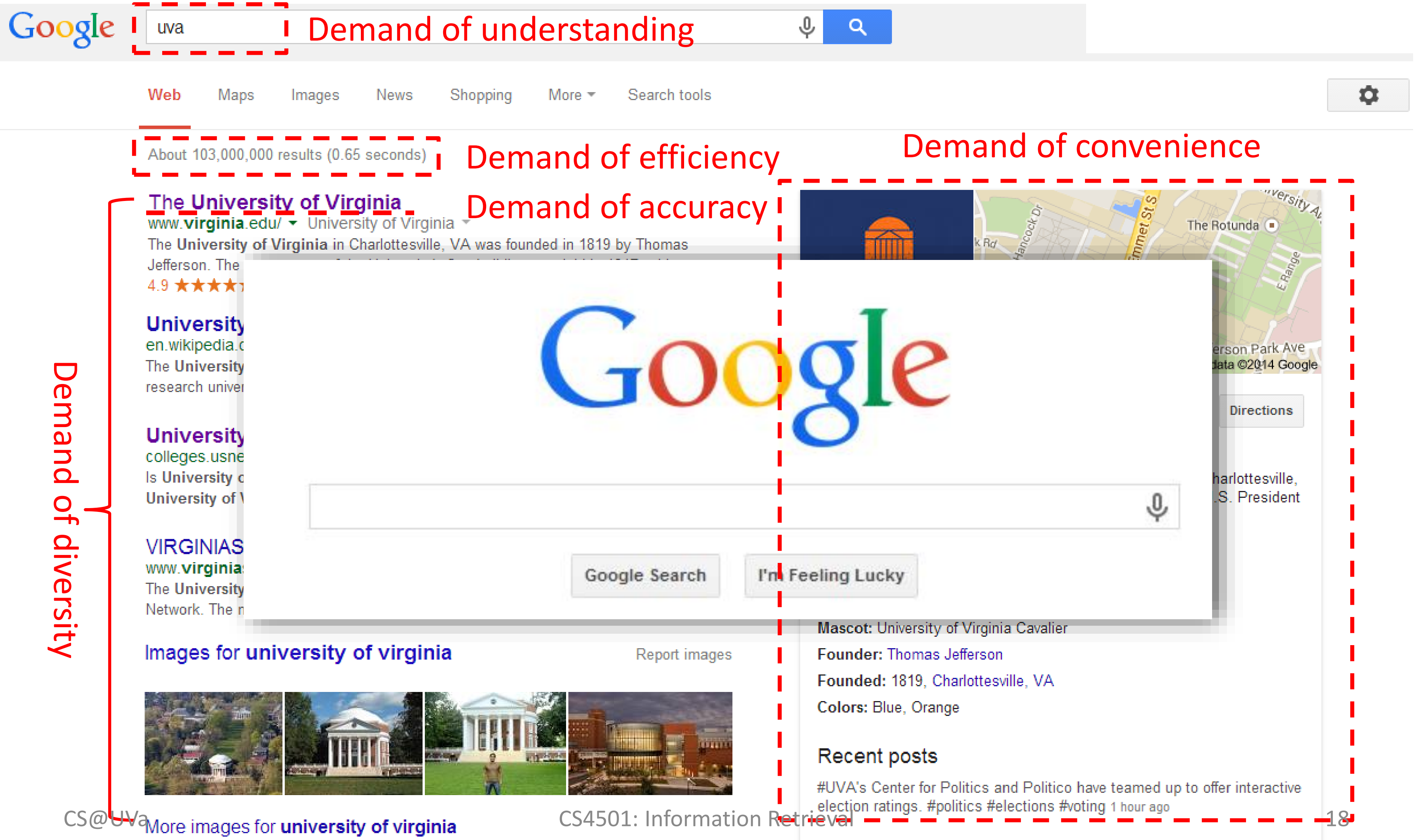
Mozambique
more...

Marketplace
- Y! Auctions - Peanuts, Pokemon, computers
- Free 56K Internet Access
- Yahoo! Bill Pay - free 3-month trial

more...

Inside Yahoo!
- Yahoo! GeoCities - build your free home page
- Play free Fantasy Soccer
- Yahoo! Clubs - create your

A glance of modern search engine



Demand of understanding

Demand of efficiency

Demand of accuracy

Demand of convenience

Demand of diversity

uva

Web Maps Images News Shopping More ▾ Search tools

About 103,000,000 results (0.65 seconds)

The University of Virginia
www.virginia.edu/ ▾ University of Virginia ▾
The University of Virginia in Charlottesville, VA was founded in 1819 by Thomas Jefferson. The
4.9 ★★★★

Mascot: University of Virginia Cavalier

Founder: Thomas Jefferson

Founded: 1819, Charlottesville, VA

Colors: Blue, Orange

Recent posts

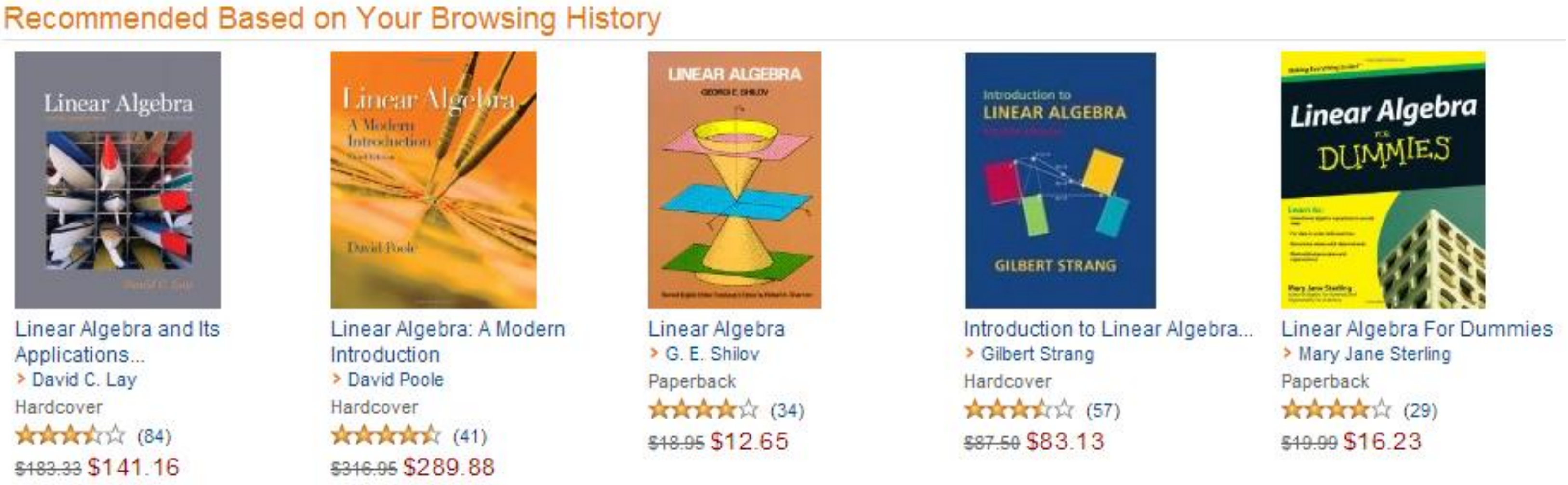
#UVA's Center for Politics and Politico have teamed up to offer interactive election ratings. #politics #elections #voting 1 hour ago

Images for university of virginia   Report images

More images for university of virginia

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
★★★½☆ (84)
~~$183.33~~ $141.16

Linear Algebra: A Modern Introduction
› David Poole
Hardcover
★★★★½ (41)
~~$316.95~~ $289.88

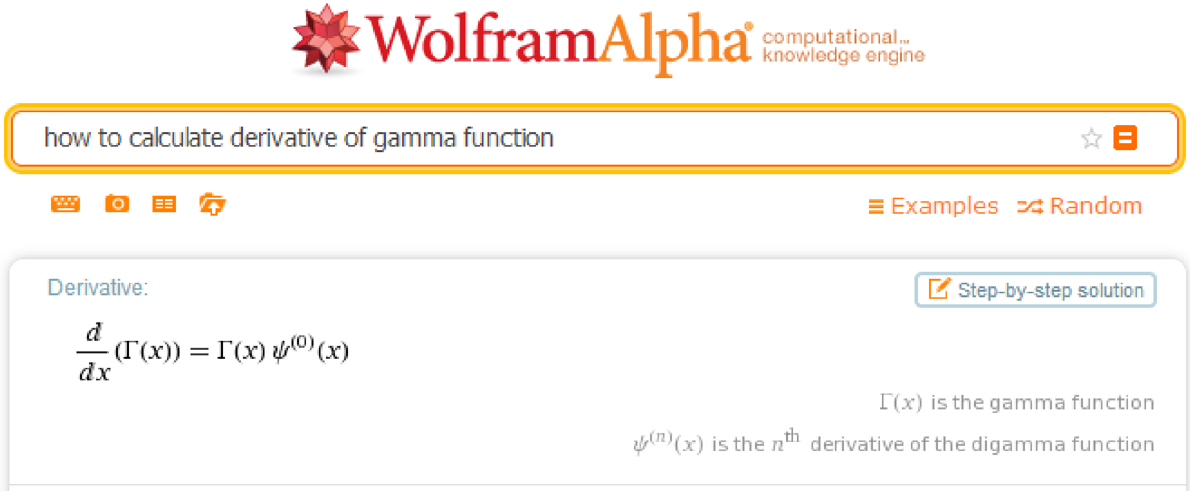
Linear Algebra
› G. E. Shilov
Paperback
★★★★☆ (34)
~~$18.95~~ $12.65

Introduction to Linear Algebra...
› Gilbert Strang
Hardcover
★★★½☆ (57)
~~$87.50~~ $83.13

Linear Algebra For Dummies
› Mary Jane Sterling
Paperback
★★★★☆ (29)
~~$19.99~~ $16.23

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



how to calculate derivative of gamma function

Derivative:

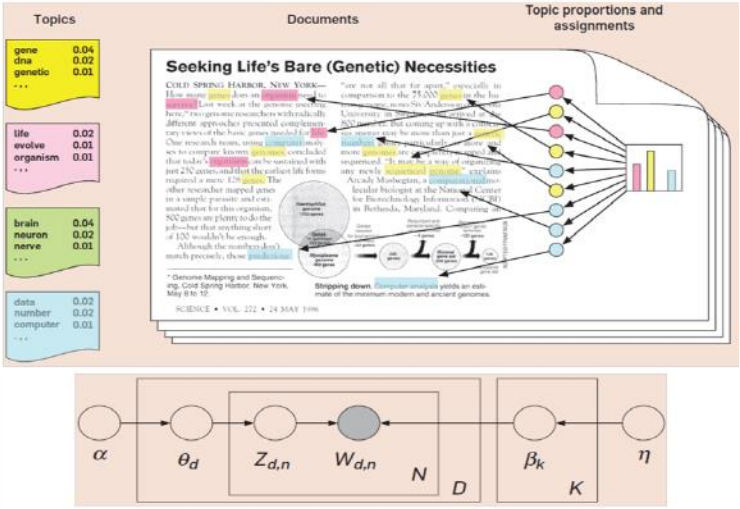
$$\frac{d}{dx}(\Gamma(x)) = \Gamma(x)\,\psi^{(0)}(x)$$

$\Gamma(x)$ is the gamma function

$\psi^{(n)}(x)$ is the n^{th} derivative of the digamma function

IR is not just about web search

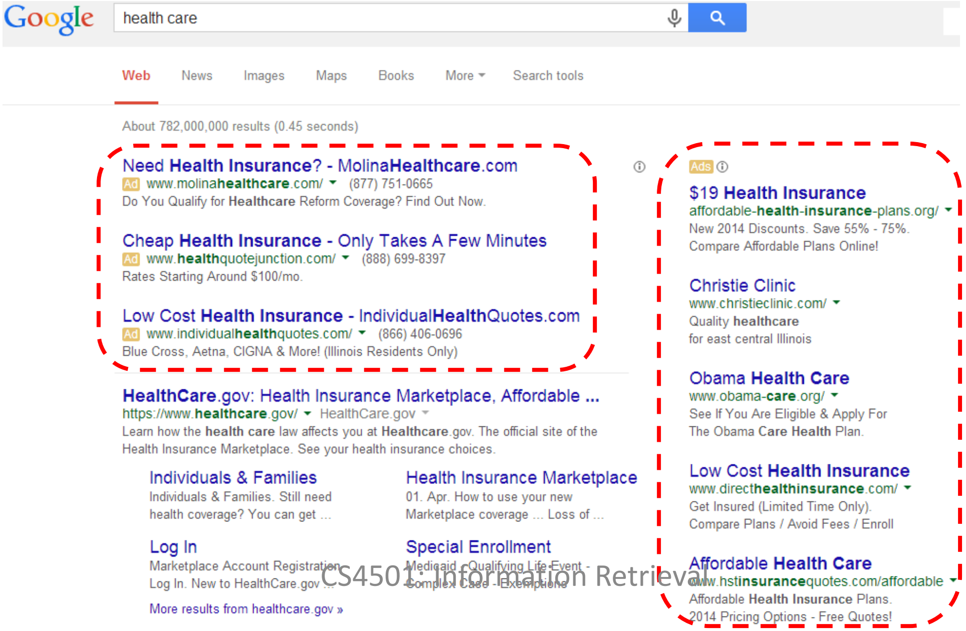
- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
  - Text mining



[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



Google health care

Web News Images Maps Books More Search tools

About 782,000,000 results (0.45 seconds)

Need **Health Insurance**? - Molina**Healthcare**.com
Ad www.molina**healthcare**.com/ (877) 751-0665
Do You Qualify for **Healthcare** Reform Coverage? Find Out Now.

Cheap **Health Insurance** - Only Takes A Few Minutes
Ad www.**health**quotejunction.com/ (888) 699-8397
Rates Starting Around $100/mo.

Low Cost **Health Insurance** - Individual**Health**Quotes.com
Ad www.individual**health**quotes.com/ (866) 406-0696
Blue Cross, Aetna, CIGNA & More! (Illinois Residents Only)

HealthCare.gov: Health Insurance Marketplace, Affordable ...
https://www.**healthcare**.gov/ HealthCare.gov
Learn how the **health care** law affects you at **Healthcare**.gov. The official site of the Health Insurance Marketplace. See your health insurance choices.

Individuals & Families
Individuals & Families. Still need health coverage? You can get ...

Health Insurance Marketplace
01. Apr. How to use your new Marketplace coverage ... Loss of ...

Log In
Marketplace Account Registration Log In. New to HealthCare.gov ...

Special Enrollment
Medicaid · Qualifying Life Event - Complex Case - Exemptions

More results from healthcare.gov »

Ads

$19 Health Insurance
affordable-**health-insurance**-plans.org/
New 2014 Discounts. Save 55% - 75%.
Compare Affordable Plans Online!

Christie Clinic
www.christieclinic.com/
Quality **healthcare**
for east central Illinois

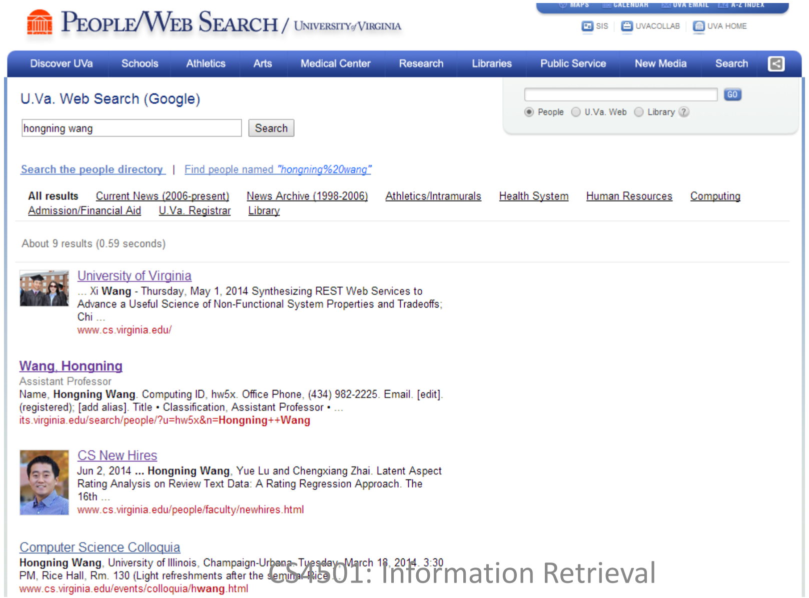
Obama **Health Care**
www.obama-**care**.org/
See If You Are Eligible & Apply For
The Obama **Care Health** Plan.

Low Cost **Health Insurance**
www.direct**healthinsurance**.com/
Get Insured (Limited Time Only).
Compare Plans / Avoid Fees / Enroll

Affordable **Health Care**
www.hst**insurance**quotes.com/affordable
Affordable **Health Insurance** Plans.
2014 Pricing Options - Free Quotes!

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



U.Va. Web Search (Google)

hongning wang

Search the people directory | Find people named *"hongning%20wang"*

All results Current News (2006-present) News Archive (1998-2006) Athletics/Intramurals Health System Human Resources Computing Admission/Financial Aid U.Va. Registrar Library

About 9 results (0.59 seconds)

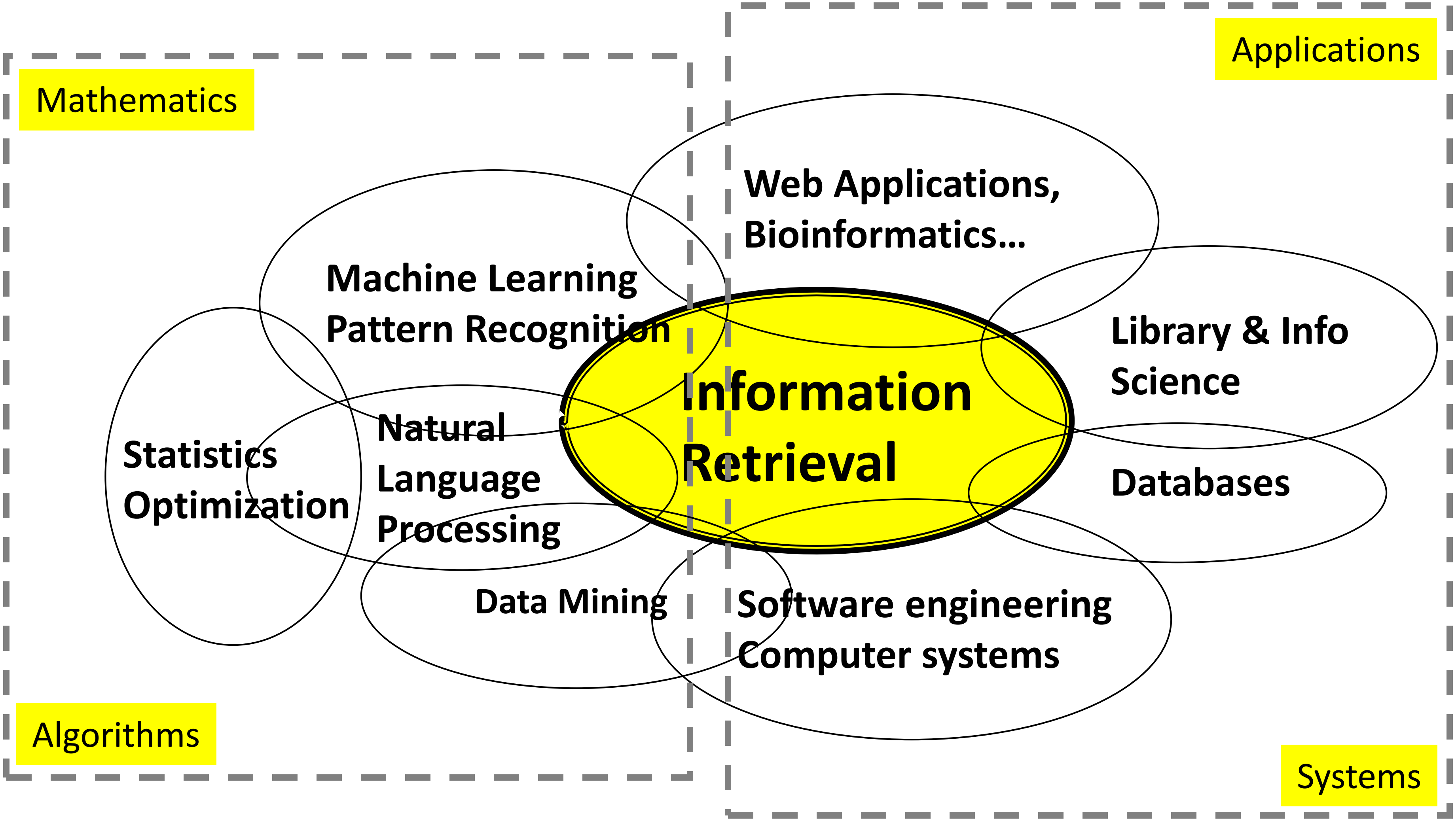
University of Virginia
... Xi **Wang** - Thursday, May 1, 2014 Synthesizing REST Web Services to Advance a Useful Science of Non-Functional System Properties and Tradeoffs; Chi ...
www.cs.virginia.edu/

Wang, Hongning
Assistant Professor
Name, **Hongning Wang**. Computing ID, hw5x. Office Phone, (434) 982-2225. Email. [edit]. (registered); [add alias]. Title • Classification, Assistant Professor • ...
its.virginia.edu/search/people/?u=hw5x&n=**Hongning++Wang**

CS New Hires
Jun 2, 2014 **... Hongning Wang**, Yue Lu and Chengxiang Zhai. Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach. The 16th ...
www.cs.virginia.edu/people/faculty/newhires.html

Computer Science Colloquia
Hongning Wang, University of Illinois, Champaign-Urbana. Tuesday, March 18, 2014. 3:30 PM, Rice Hall, Rm. 130 (Light refreshments after the seminar-Rice ...
www.cs.virginia.edu/events/colloquia/h**wang**.html

Related Areas



IR v.s. DBs

- Information Retrieval:
  - Unstructured data
  - Semantics of object 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
  - 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

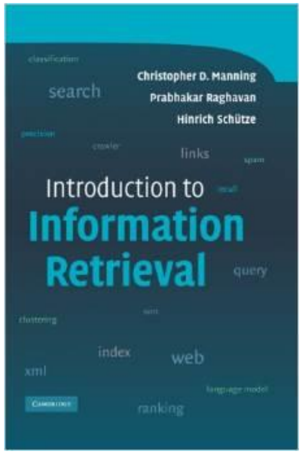
- Natural language processing
  - Cognitive, symbolic and computational approaches
  - Semantic (deep) understanding of language
  - (often times) small scale problems

IR and NLP are getting closer

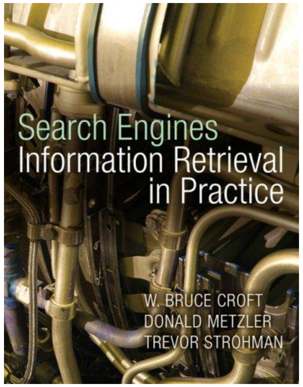
- IR => NLP
  - Larger data collections
  - Scalable/robust NLP techniques, e.g., translation models

- NLP => IR
  - Deep analysis of text documents and queries
  - Information extraction for structured IR tasks

Text books

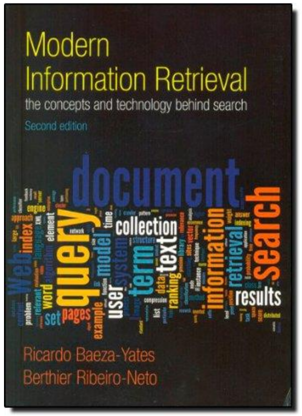


- ***Introduction to Information Retrieval***. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.

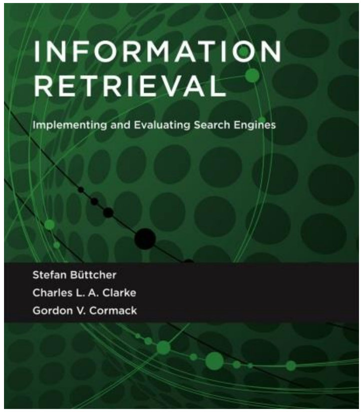


- ***Search Engines: Information Retrieval in Practice***. Bruce Croft, Donald Metzler, and Trevor Strohman, Pearson Education, 2009.

Text books

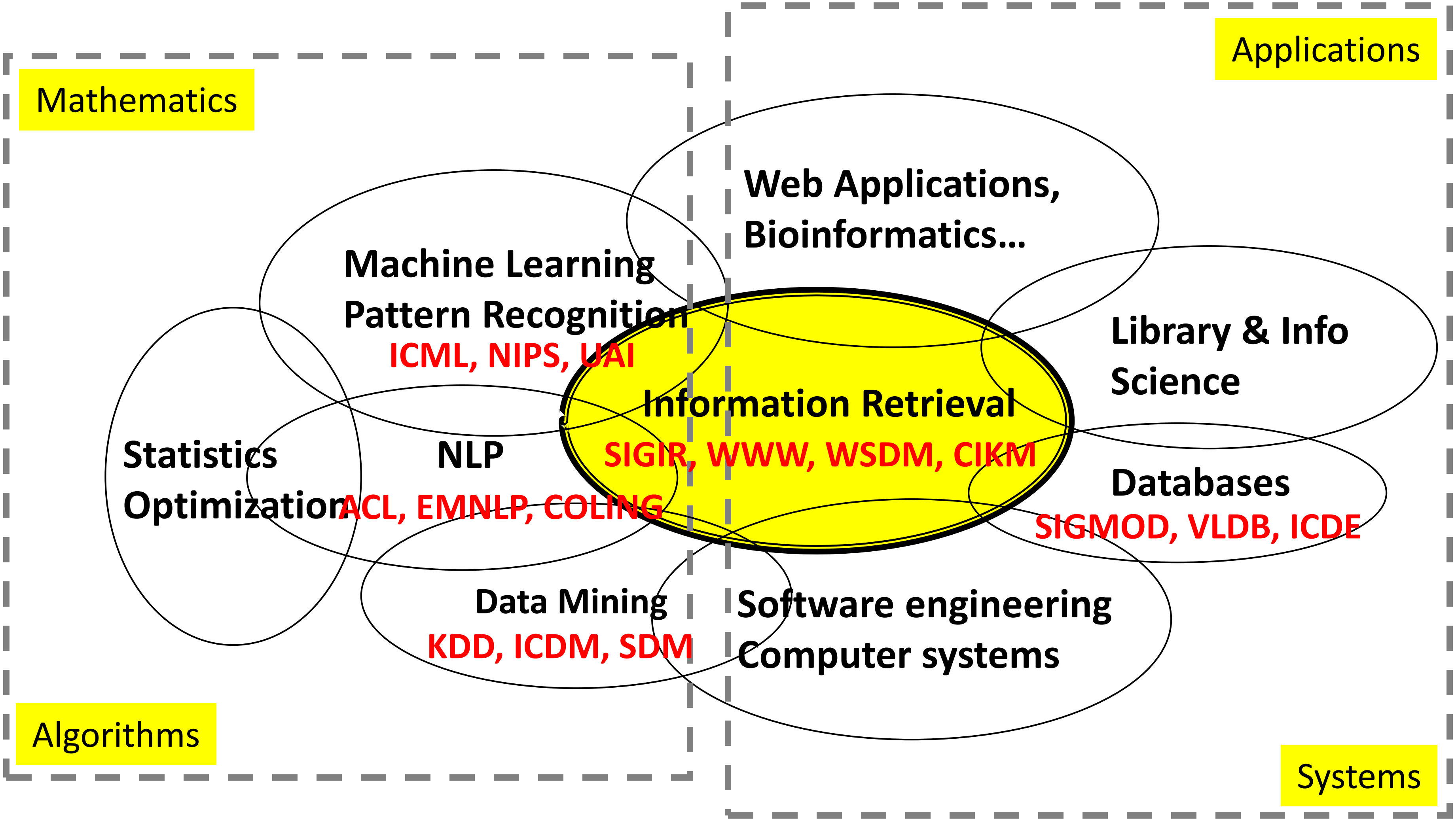


- ***Modern Information Retrieval.*** Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Addison-Wesley, 2011.



- ***Information Retrieval:*** Implementing and Evaluating Search Engines. Stefan Buttcher, Charlie Clarke, Gordon Cormack, MIT Press, 2010.

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