Temu-Kembali Informasi 2019

01: Pengantar Perkuliahan

Husni

Husni.trunojoyo.ac.id

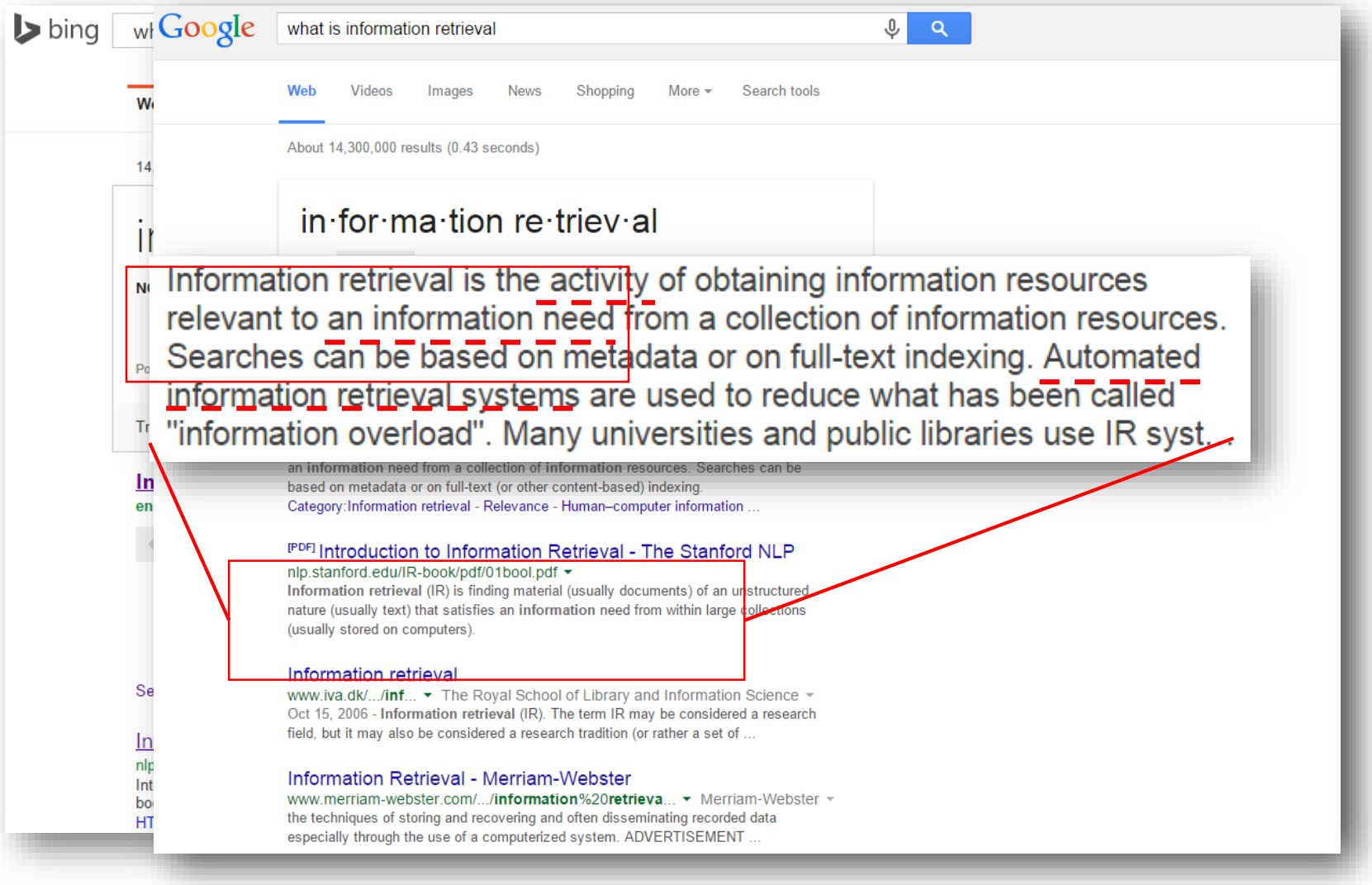
Outline

- Pengantar Information Retrieval
- Aturan Perkuliahan

Introduction to Information Retrieval

Husni

What is information retrieval?



Information Retrieval?

- Information retrieval (IR) adalah bidang yang konsen dengan perancangan, pengembangan, dan evaluasi sistem interaktif yang membantu pengguna memperoleh informasi.
- Diberikan suatu query dan corpus, temukan item yang relevan
  - **query**: ekspresi kebutuhan informasi dari pengguna
  - **corpus**: repository item-item yang dapat ditemu-kembalikan
  - **relevansi**: pemenuhan kebutuhan informasi pengguna
- Gerard Salton, 1968:
  - Information retrieval is a field concerned with the structure, analysis, organization, storage, and retrieval of information.

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

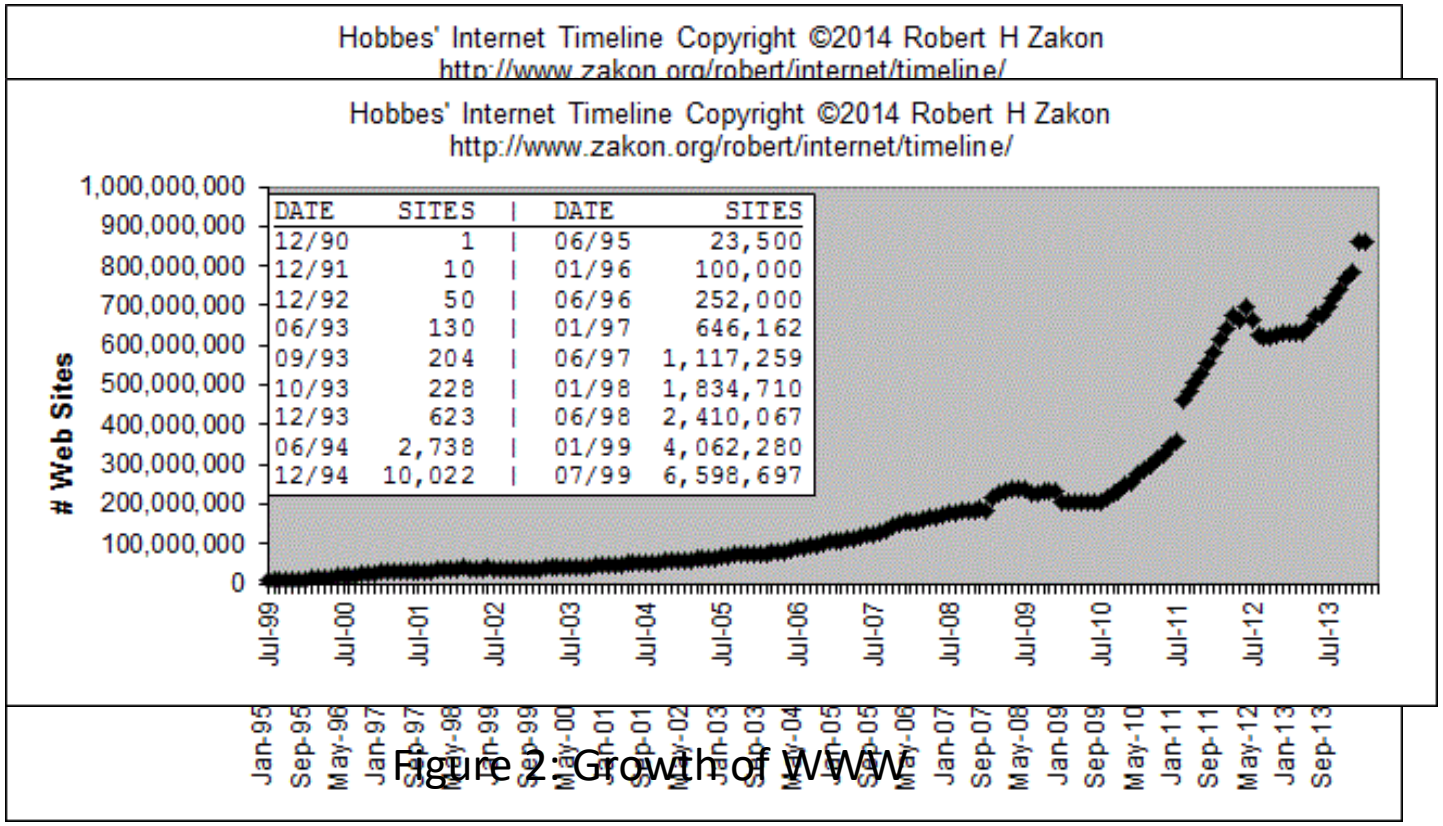


Figure 2: Growth of WWW

Figure 1: Growth of Internet

Why information retrieval

- Handling unstructured data
  - Structured data: database system is a good choice
  - Unstructured data is more dominant
    - Text in Web c
    - "*85 percent* c
    - Unknown ser

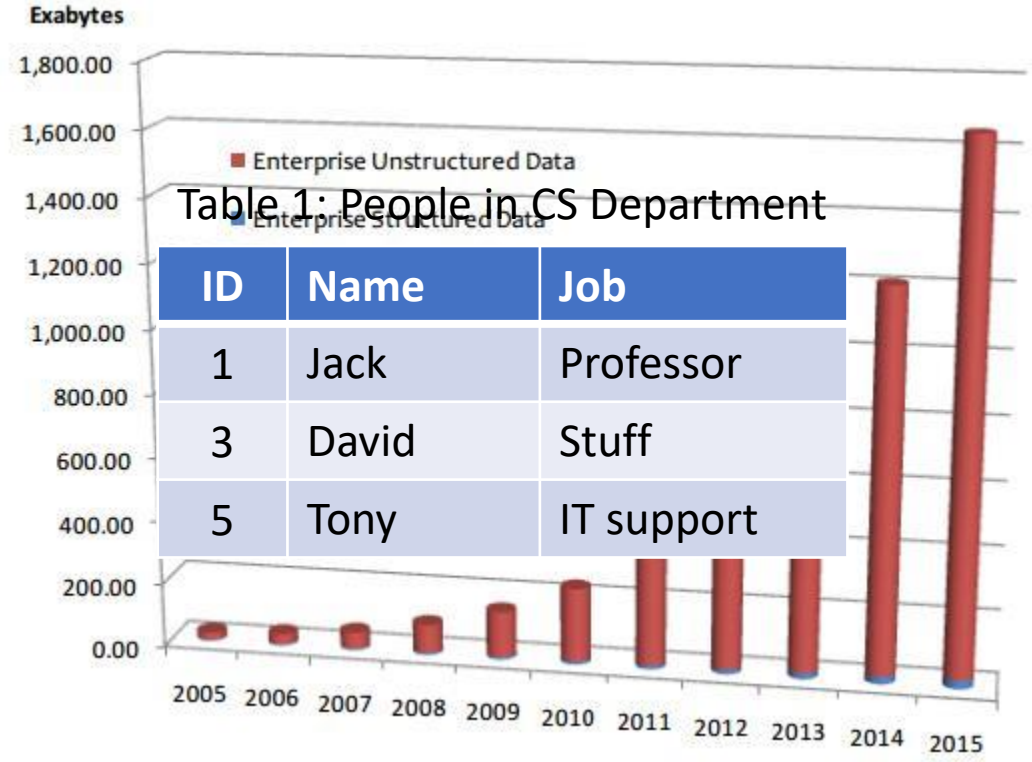


Table 1: People in CS Department

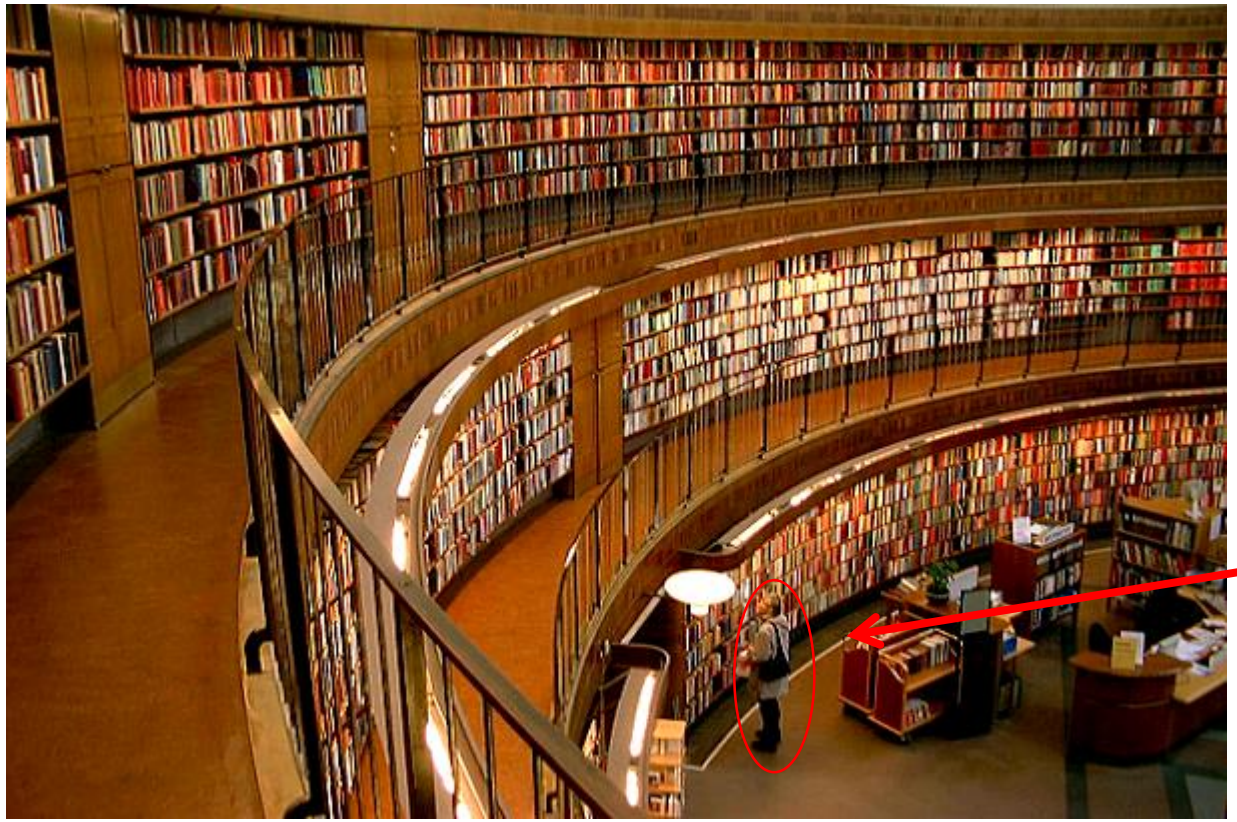
ID	Name	Job
1	Jack	Professor
3	David	Stuff
5	Tony	IT support

Total Enterprise Data Growth 2005-2015, IDC 2012

" - Merrill Lynch

Why information retrieval

- An essential tool to deal with information overload



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.*"
  - "*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.*"

-> WWW -> 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

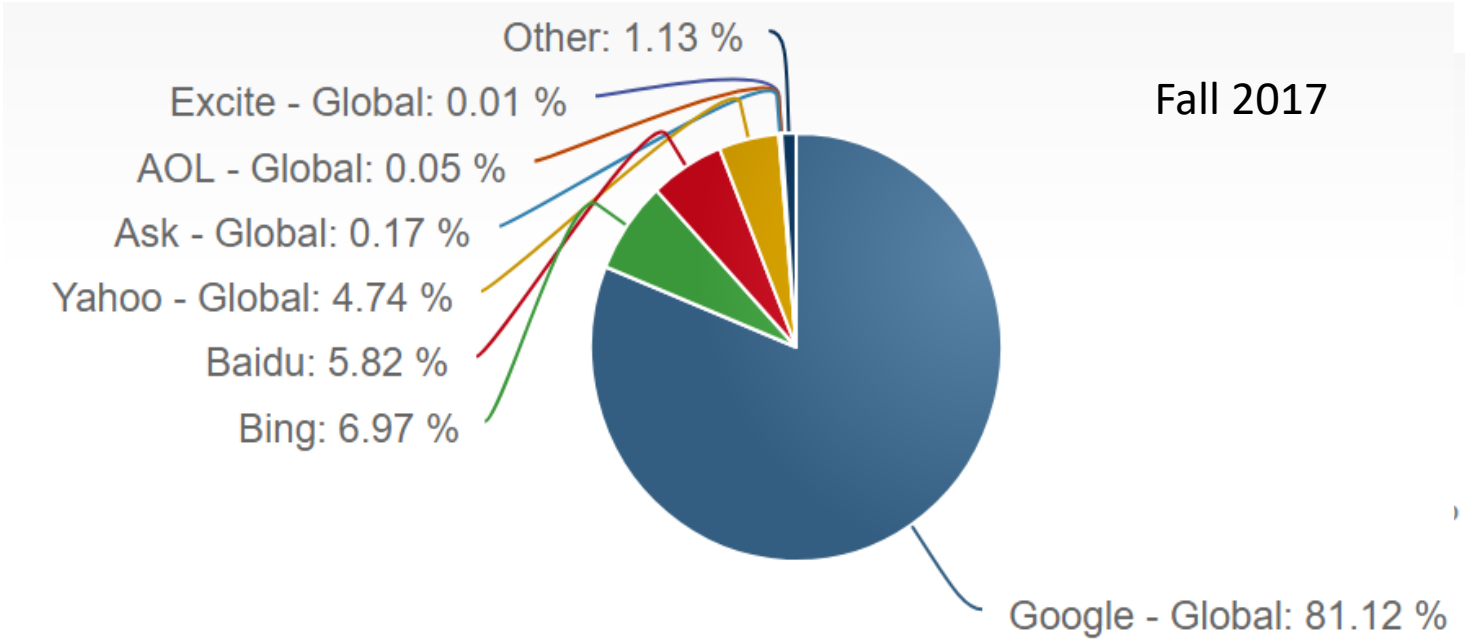
- Catalyst
  - Academia: Text Retrieval Conference (TREC) in 1992
    - *"Its purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies."*
    - *"… about one-third of the improvement 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*

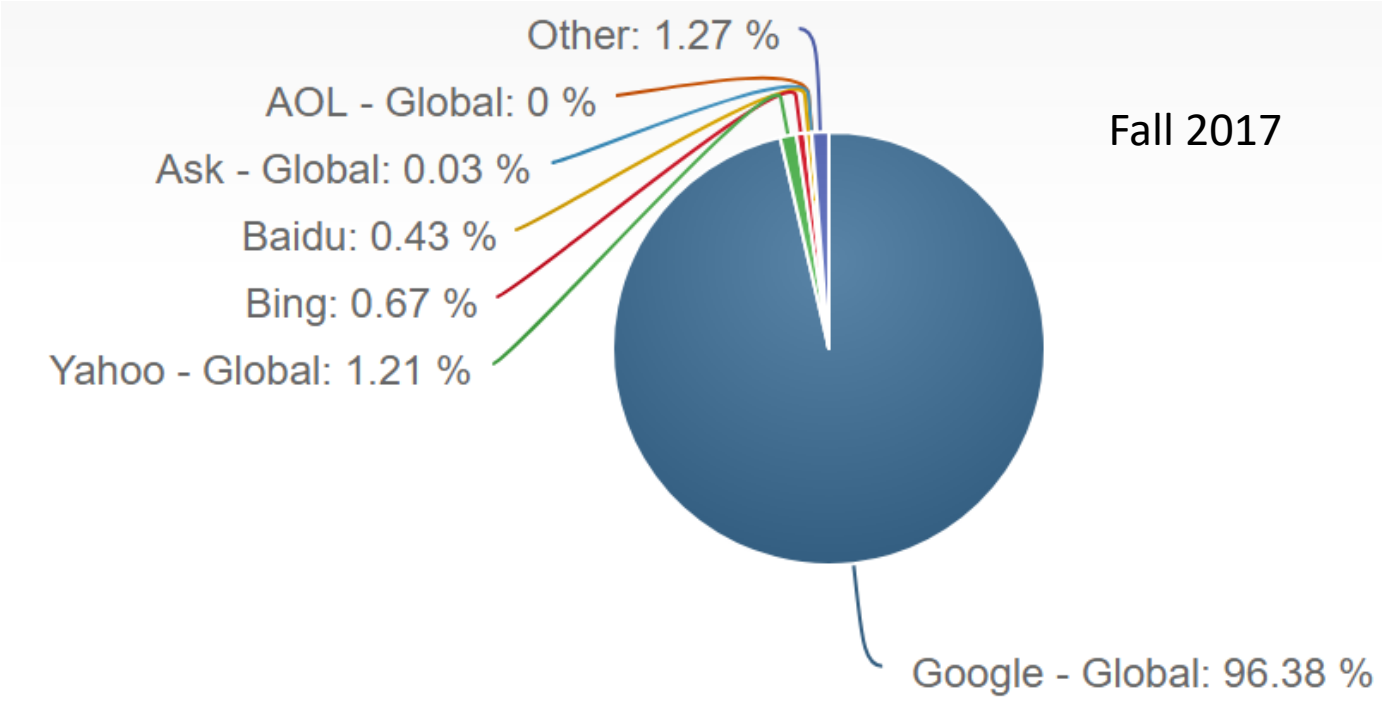
Major players in this game

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



Major players in this game

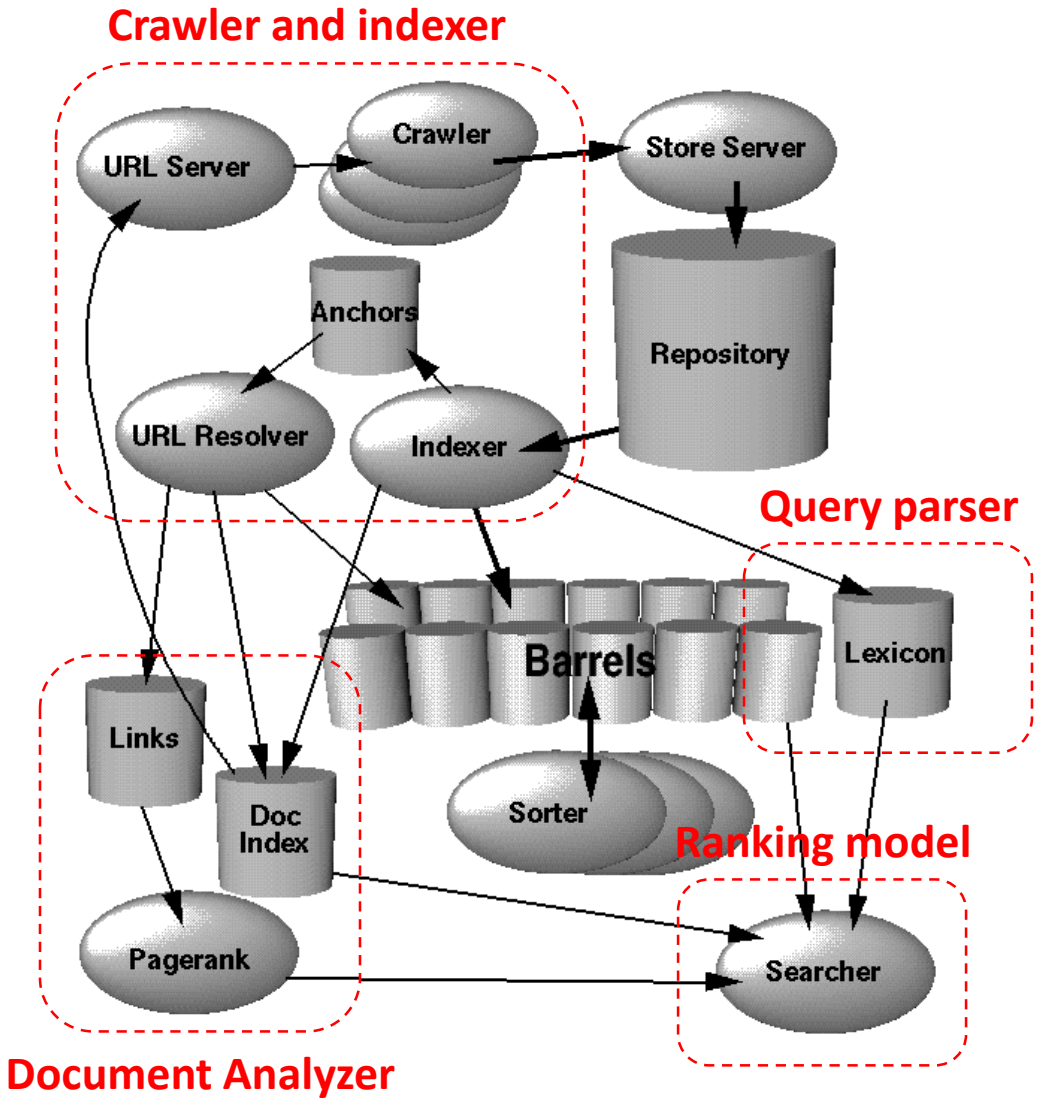
- Global search engine market - mobile
  - By http://marketshare.hitslink.com/search-engine-market-share.aspx



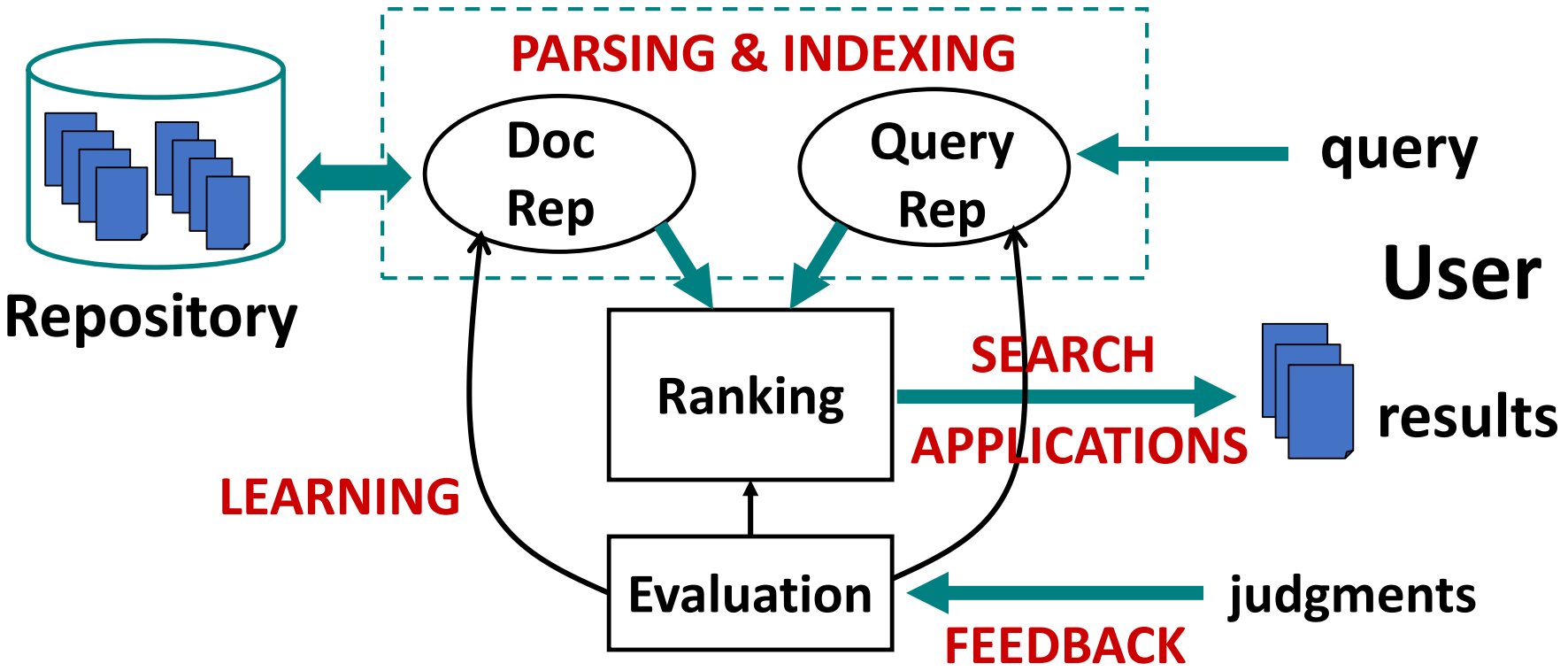
How to perform information retrieval

- Information retrieval when we did not have a computer

How to perform information retrieval



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.

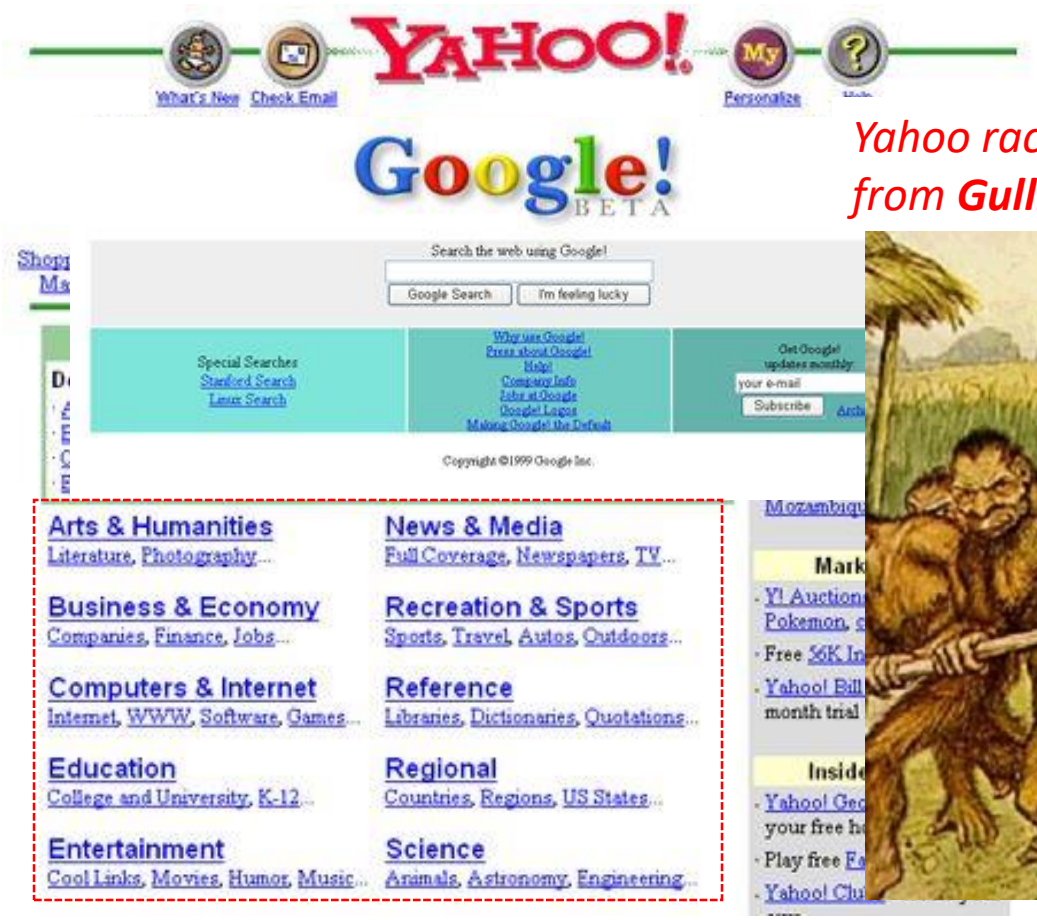
Core concepts in IR

- 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

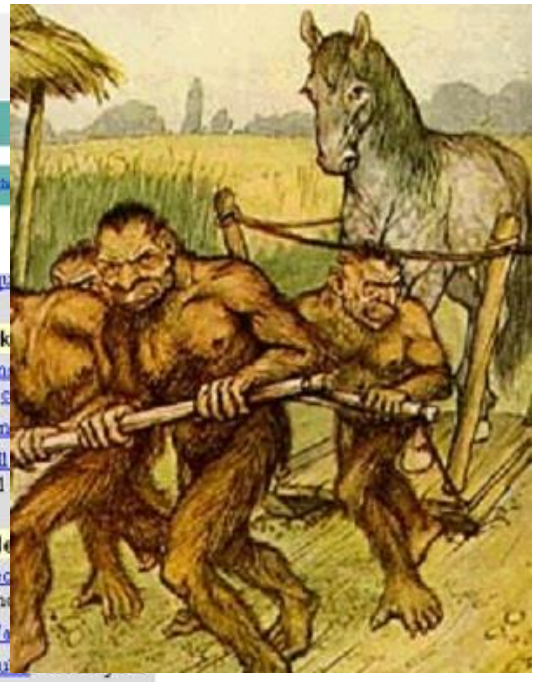
A glance of modern search engine

- In old times

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

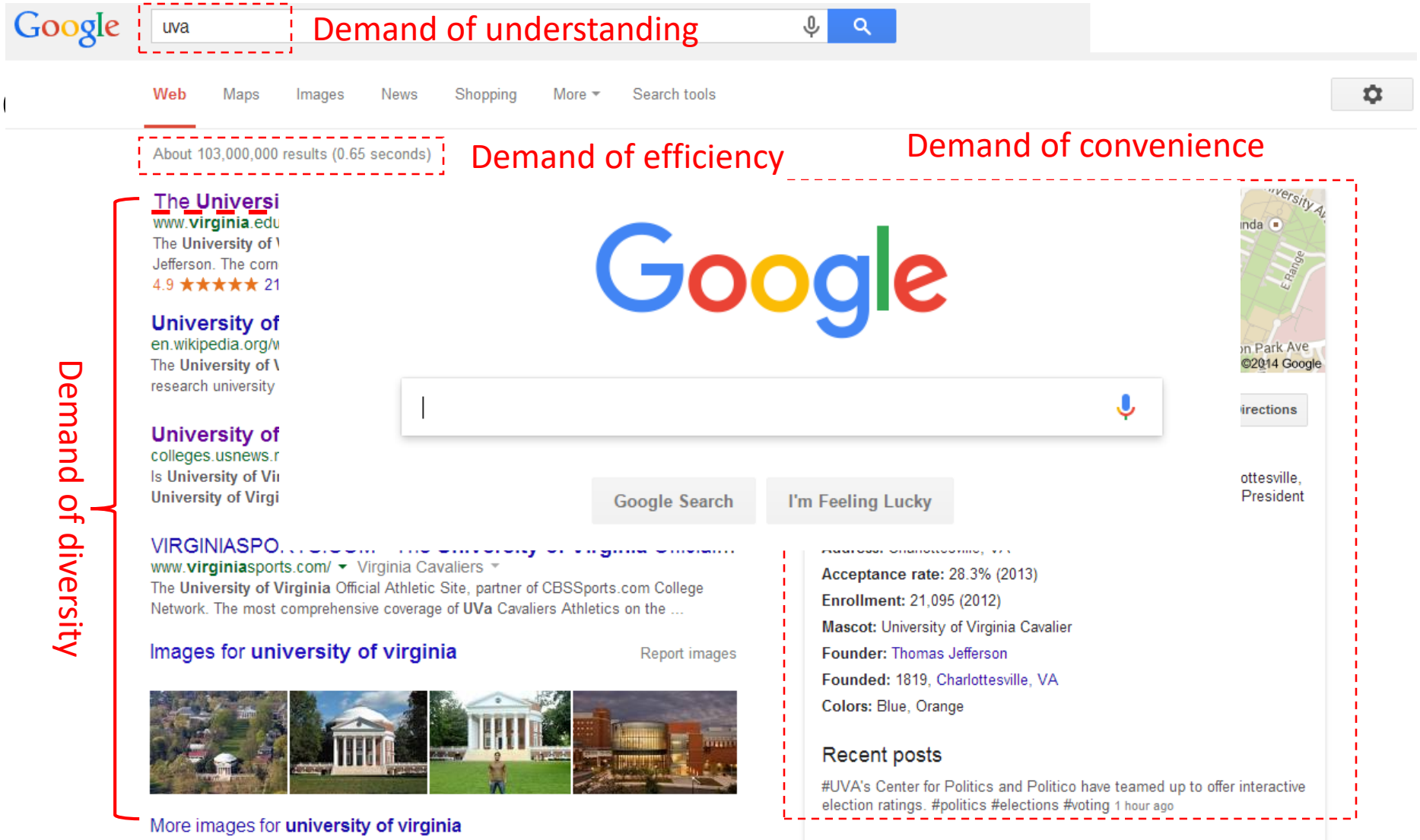


*Yahoo race of fictional beings from **Gulliver's Travels***



A glance of modern search engine

- Mo



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

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

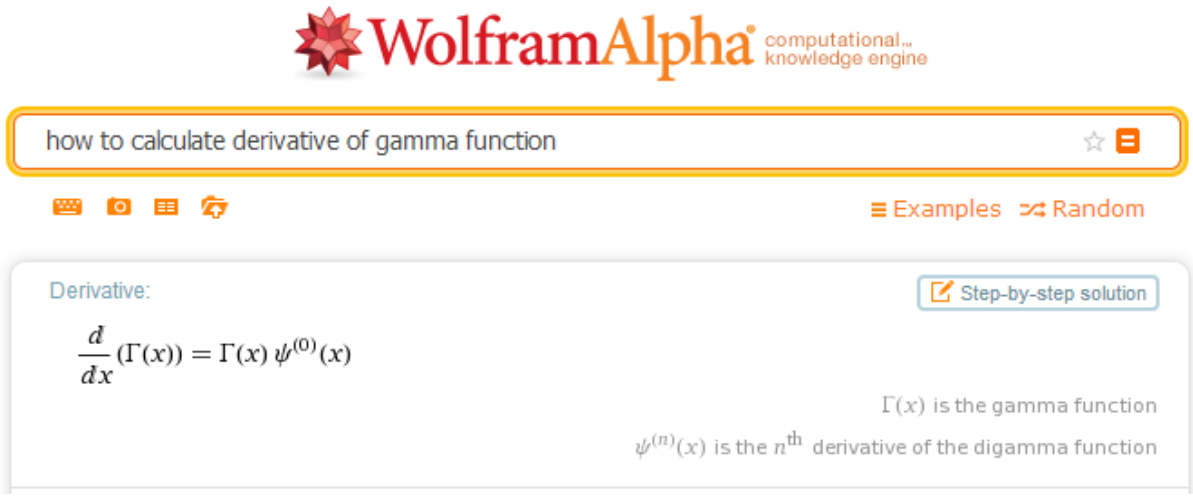
Google, Lagging Amazon, Races Across the Threshold Into the Home OCT. 2, 2016

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

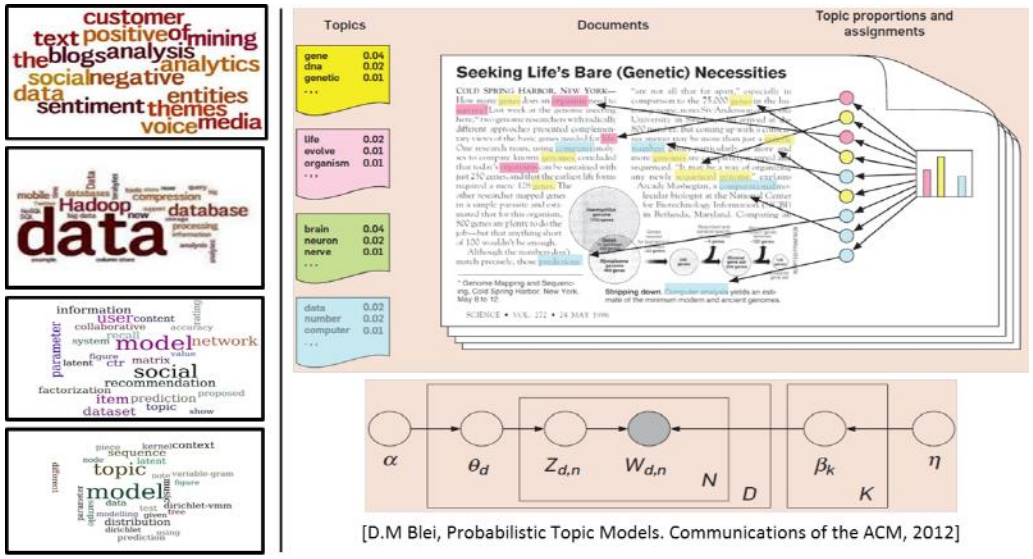
IR is not just about web search

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



IR is not just about web search

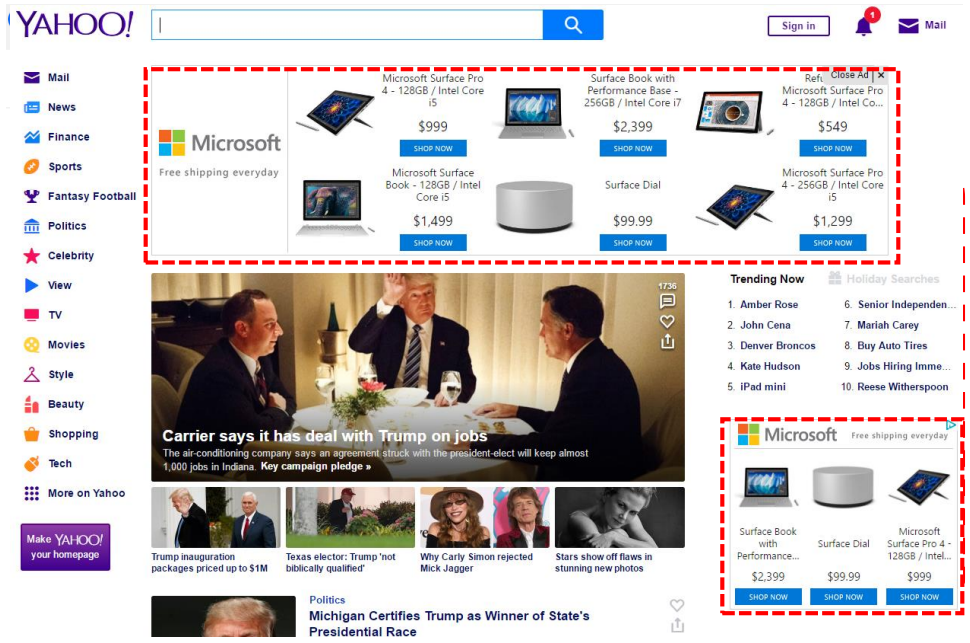
- Web search is just one important area of information retrieval, but not all
- Information retrieval also includes
  - Text mining



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

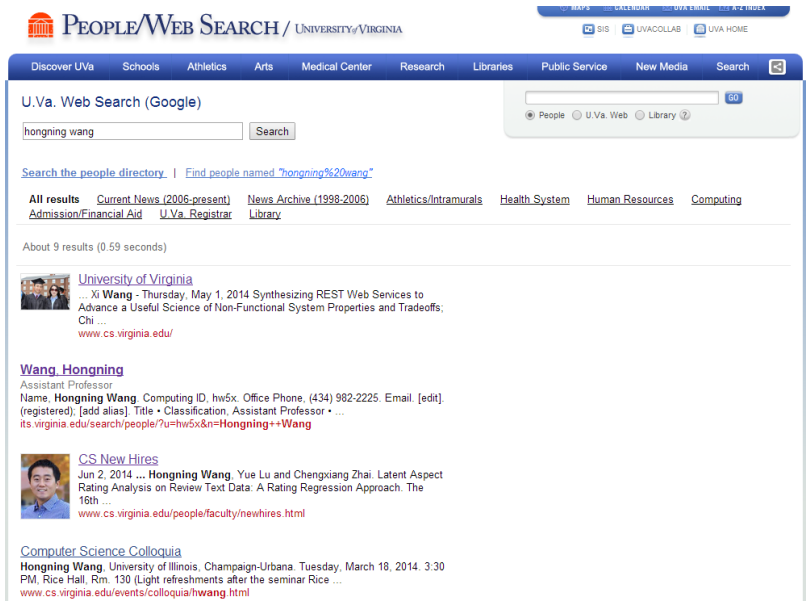
IR is not just about web search

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

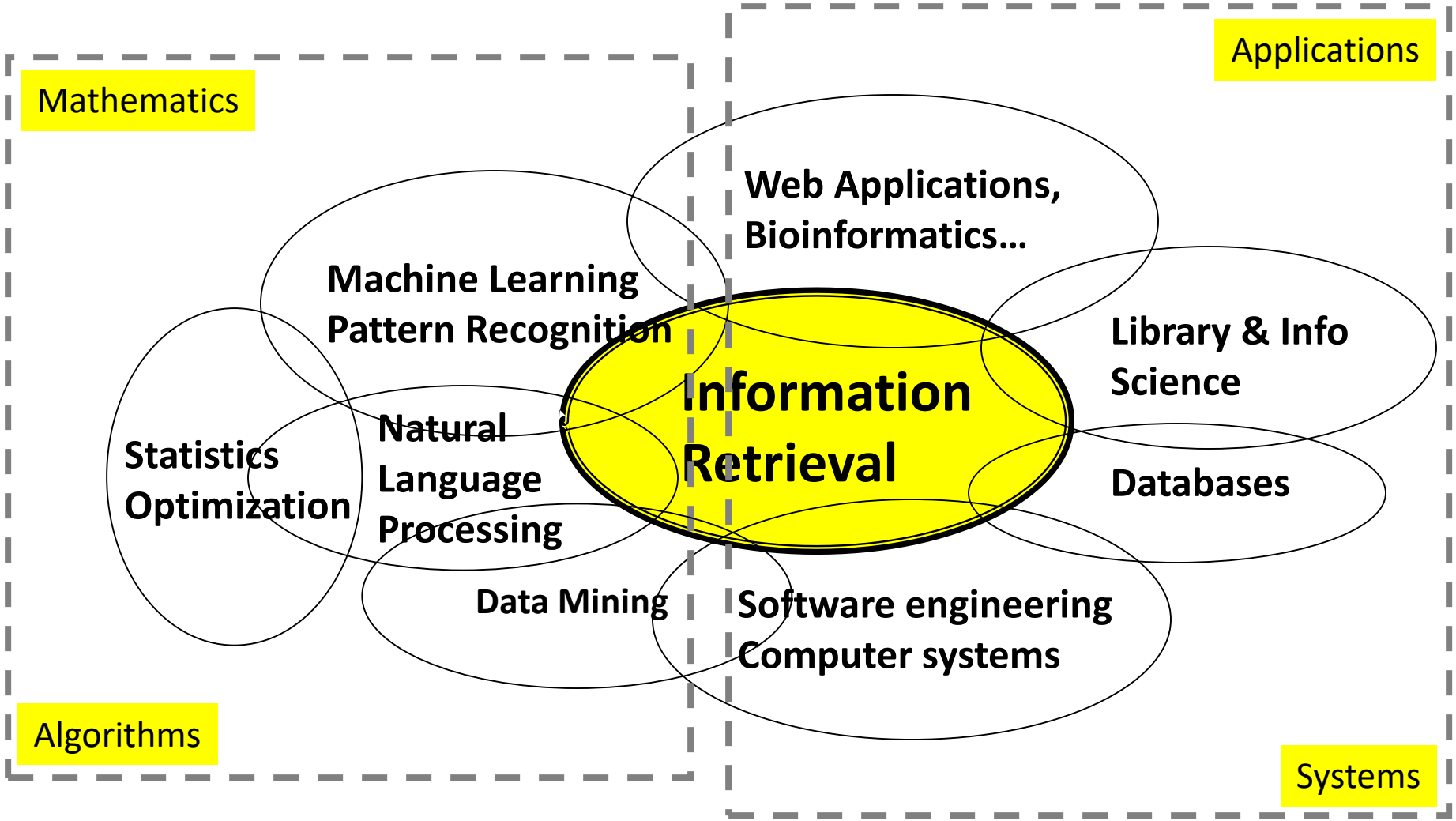


IR is not just about web search

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



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
  - Semi-structured representation: XML data; queries with structured information

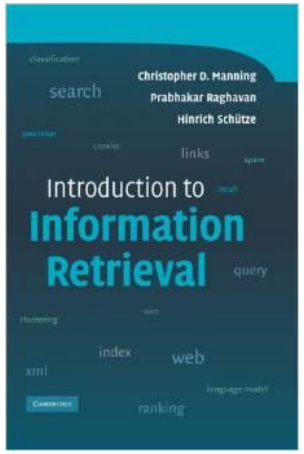
IR v.s. NLP

- 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

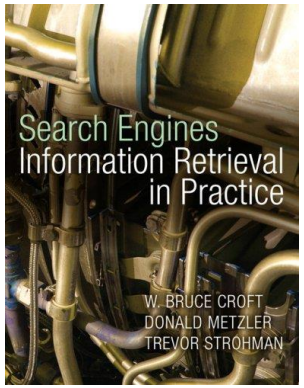
IR and NLP are getting closer

- 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

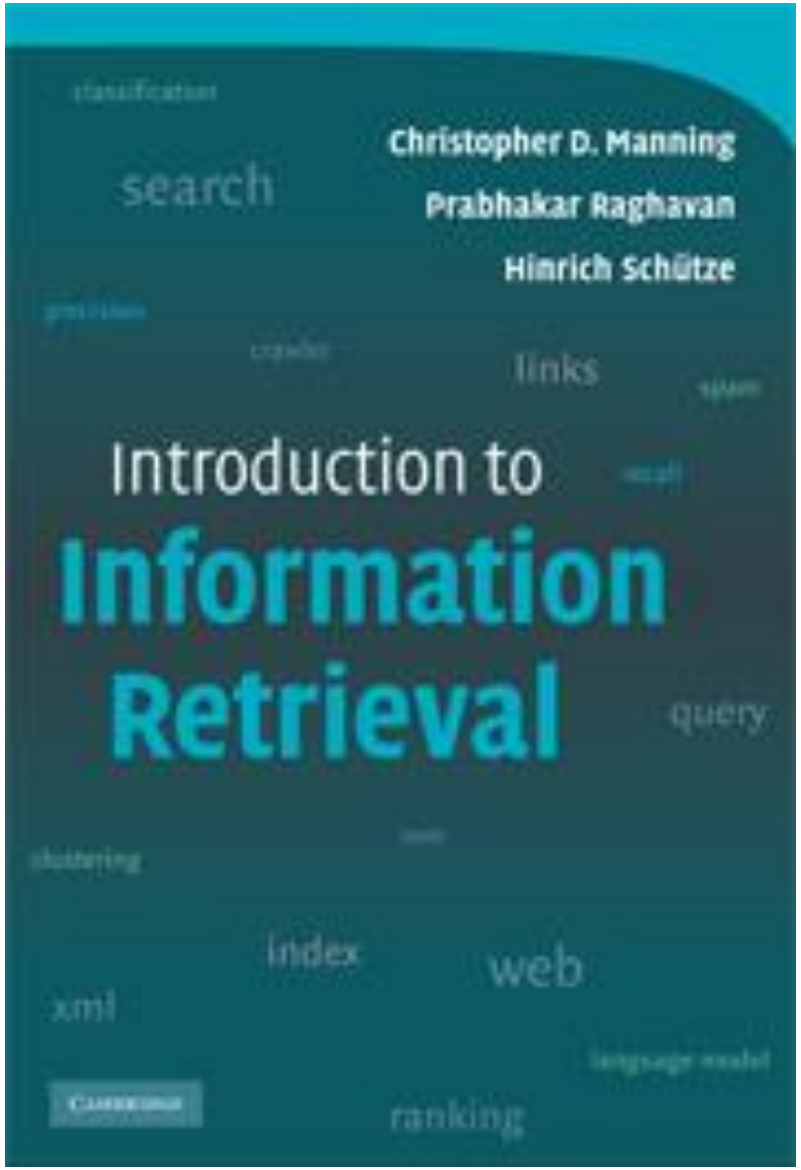
Text books



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

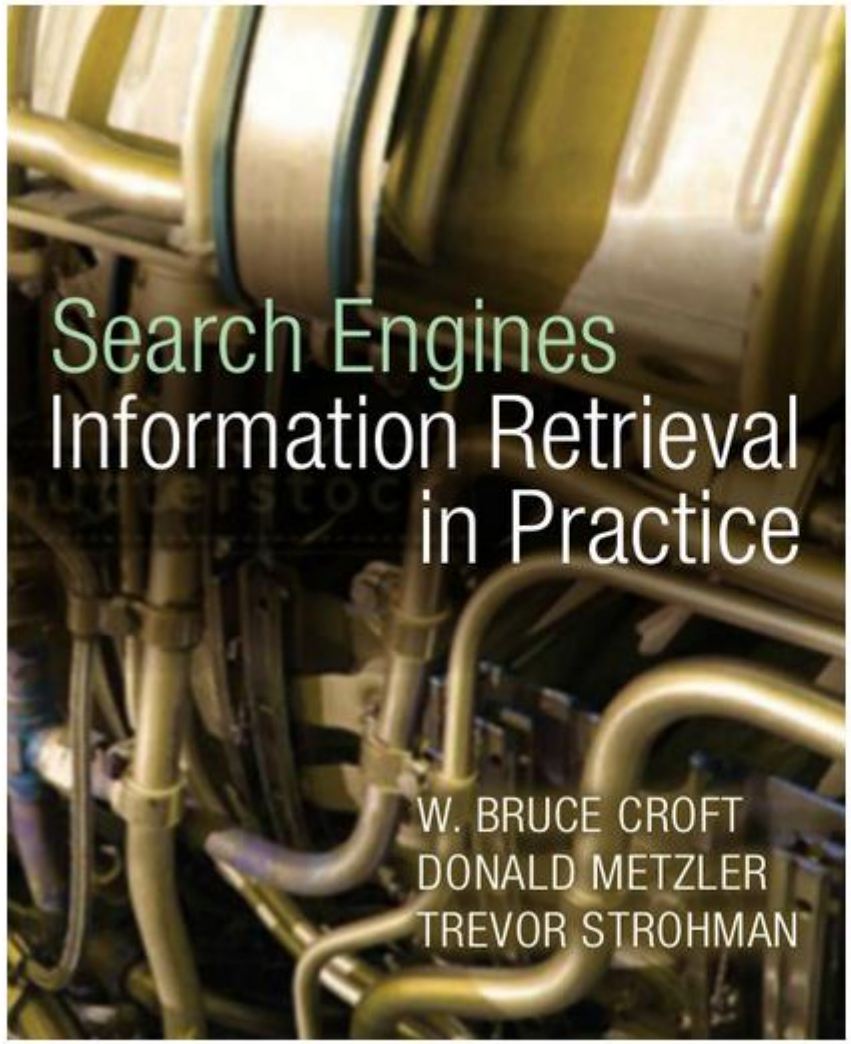


- ***Search Engines: Information Retrieval in Practice***. Bruce Croft, Donald Metzler, and Trevor Strohman, Pearson Education, 2009.



1. Boolean retrieval
2. The term vocabulary and postings lists
3. Dictionaries and tolerant retrieval
4. Index construction
5. Index compression
6. Scoring, term weighting and the vector space model
7. Computing scores in a complete search system
8. Evaluation in information retrieval
9. Relevance feedback and query expansion
10. XML retrieval
11. Probabilistic information retrieval
12. Language models for information retrieval
13. Text classification and Naive Bayes
14. Vector space classification
15. Support vector machines and machine learning on documents
16. Flat clustering
17. Hierarchical clustering
18. Matrix decompositions and latent semantic indexing
19. Web search basics
20. Web crawling and indexes
21. Link analysis

https://nlp.stanford.edu/IR-book/information-retrieval-book.html



BOOK DESCRIPTION

This book provides an overview of the important issues in information retrieval, and how those issues affect the design and implementation of search engines. Not every topic is covered at the same level of detail. The focus is on some of the most important alternatives to implementing search engine components and the information retrieval models underlying them. The target audience for the book is advanced undergraduates in computer science, although it is also a useful introduction for graduate students.

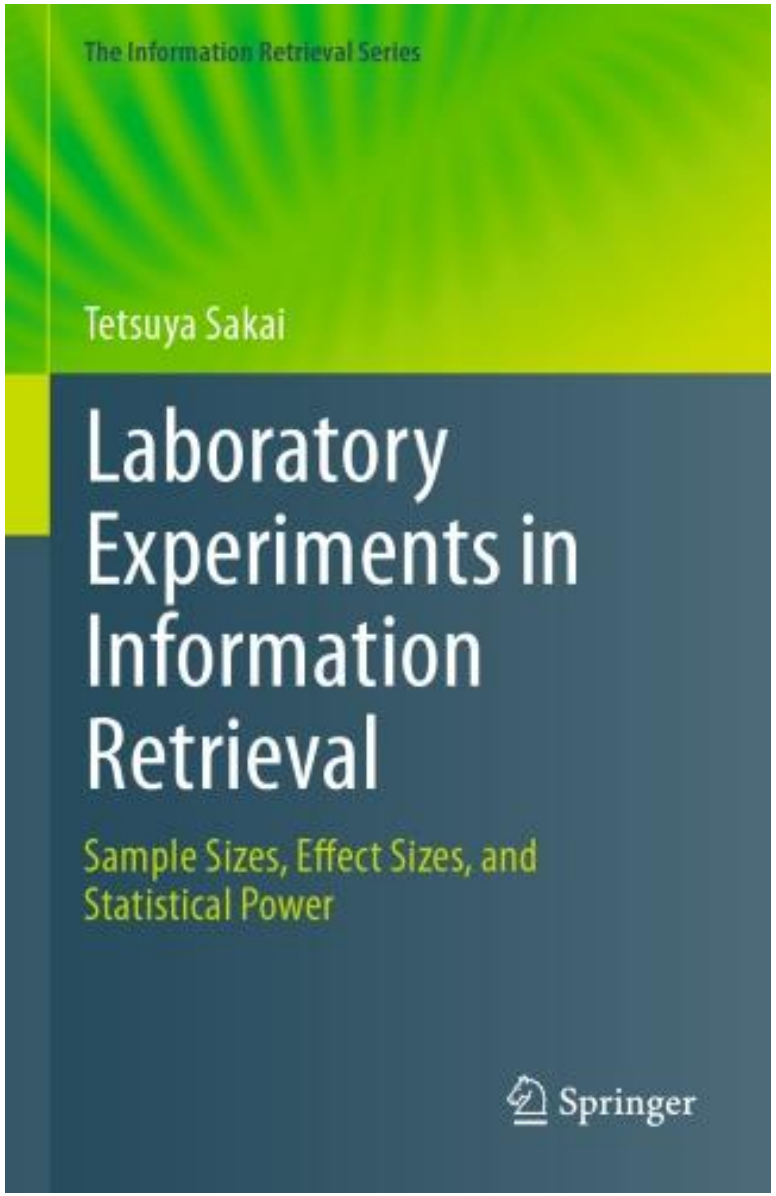
This version of the book is being made available for free download. It has been edited to correct the minor errors noted in the 5 years since the book's publication. The authors, meanwhile, are working on a second edition.

1. Search Engines and Information Retrieval
2. Architecture of a Search Engine
3. Crawls and Feeds
4. Processing Text
5. Ranking with Indexes
6. Queries and Interfaces
7. Retrieval Models
8. Evaluating Search Engines
9. Classification and Clustering
10. Social Search
11. Beyond Bag of Words

ciir.cs.umass.edu/downloads/SEIRiP.pdf

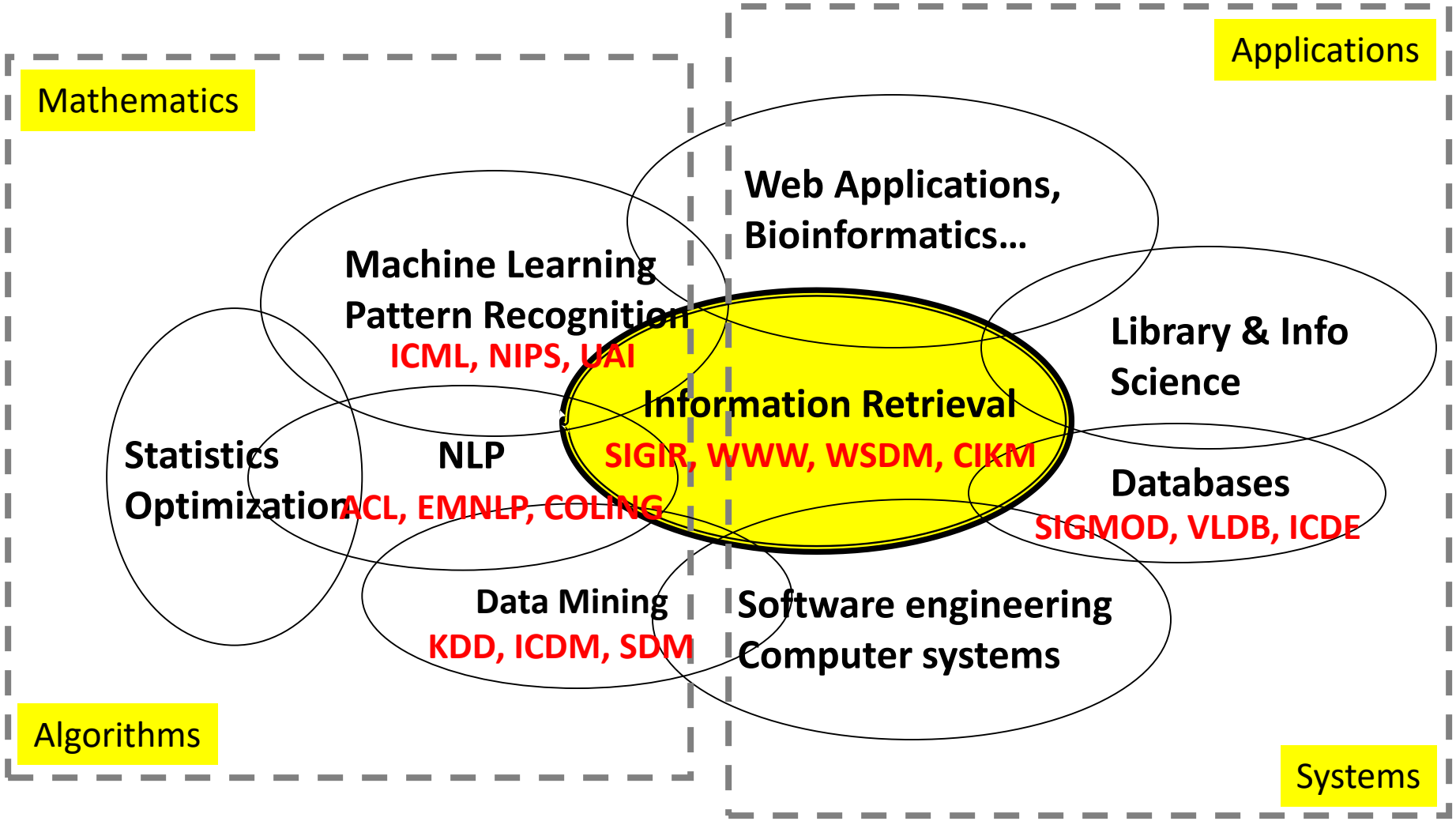
Referensi

1. R. Baeza-Yates, B. Ribeiro-Neto. Modern Information Retrieval: The Concepts and Technology behind Search, Pearson 2011.
2. S. Büttcher, C.L.A. Clarke, G.V. Cormack. Information Retrieval: Implementing and Evaluating Search Engines, MIT Press 2010.
3. ChengXiang Zhai, Sean Massung. Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining, ACM Press 2016.
4. Tetsuya Sakai. Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power, Springer 2018



1. Preliminaries
2. t-Tests
3. Analysis of Variance
4. Multiple Comparison Procedures
5. The Correct Ways to Use Significance Tests
6. Topic Set Size Design Using Excel
7. Power Analysis Using R
8. Conclusions

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

Course Policy

Husni

Pop-up quiz

1. Let **a**=(1,2,3) and **b**=(2,3,-2), the inner product between **a** and **b** is

   (a) 0 (b) 1 (c) 2 (d) 3

2. Let A = $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$, what is A^{-1},

   (a) $\begin{pmatrix} -1 & -2 \\ -2 & -1 \end{pmatrix}$ (b) $\begin{pmatrix} -\frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} \end{pmatrix}$ (c) $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ (d) $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$

Pop-up quiz

3. What is the expectation of random variables drawn from Gaussian distribution N(0, 1),

   (a) 0 (b) 0.5 (c) 1 (d) 2

4. A biased coin with P(head)=0.2, in a sequence of 10 consecutive tossing, you have already got 9 tails, what is the probability you will have a head at the 10^{th} tossing,

   (a) 0 (b) 0.1 (c) 0.2 (d) $0.2*0.8^9$

Dosen Pengampu

- Husni, S.Kom., MT.
  - Graduated from University of Gadjah Mada (2001) and Bandung Institute of Technology (2010)
  - Research area
    - Web Information Retrieval (Text Mining)
    - Data Science (Analytics)
    - Internetworking Technology: Web Services
  - Industry experience
    - Internetworking
    - Software Development
    - IT Blueprint

Goal of this course

- Discuss fundamental problems in information retrieval
  - Building blocks of search engine systems
  - Wide coverage of important IR techniques
    - Personalized recommendation
    - Online advertising
- Get hands-on experience by developing practical systems/components
- Prepare students for doing cutting-edge research in information retrieval and related fields
  - Open the door to the amazing job opportunities in IT industry

Outcomes

- Example of letters from former students

Dear Pak Husni,
Thank you so much for teaching me Information Retrieval where I have benefitted the most this semester. I have got an internship position in Walmart Labs search team all because of the knowledge I leant from your class. Although you are strict on the grade, but after all I think it’s fair and still encourage me to learn better on IR.

Hi Professor,
My name is Handayani, and I just graduated from UTM in March.
I will start working full-time at Google starting next Monday and I just got my team assignment today. I will be working at Google‘s search ranking team. I still remembered the Information Retrieval class I took with you. That still remain one of my favorite IF classes at UTM! I'm sending this email just to let you know that you have a former student working on search engines.

Capaian Pembelajaran IR 2019

- **CLO1**: apply information retreival principles to locate relevant information in large collections of data
- **CLO2**: understand and deploy efficient techniques for the indexing of document objects that are to be retrieved
- **CLO3**: implement features of retrieval systems for web-based and other search tasks
- **CLO4**: analyse the performance of retrieval systems using test collections
- **CLO5**: make practical recommendations about deploying information retrieval systems in different search domains, including considerations for document management and querying

Topik Bahasan IR 2018

- Temu-Kembali Informasi & *Search Engine*
- Arsitektur *Search Engine*
- Web *Crawling* dan *Feeding*
- Pemrosesan Teks
- Pemeringkatan (Ranking) dengan Indeks
- Query dan Antarmuka
- Model-model Temu-Kembali
- Evaluasi terhadap *Search Engine*
- Klasifikasi dan Klasterisasi
- Pencarian Sosial
- Sistem Rekomendasi
- Sistem Rekomendasi bidang Pariwisata.

Syllabus IR 2019

Fokus pada Konsep *Retrieval*

- Pengantar Perkuliahan
- Bagian I: Arsitektur Search Engine
  - Topik 1.1: Arsitektur Dasar dari Search Engine
  - Topik 1.2: Web crawling dan Teknik Pemrosesan Teks
  - Topik 1.3: Inverted Index dan Pemrosesan Query
- Bagian II: Evaluasi Temu-Kembali
  - Topik 2.1: Evaluasi IR Klasik
  - Topik 2.2: Evaluasi IR Modern
- Bagian III: Model Temu-Kembali
  - Topik 3.1: Model Boolean dan Ruang Vektor

Syllabus IR 2019

- Bagian III: Model Temu-Kembali
  - *Topik 3.1: Model Boolean dan Ruang Vektor*
  - Topik 3.2: Prinsip Ranking Probabilistik
  - Topik 3.3: Model Bahasa
  - Topik 3.4: Pembelajaran Peringkat
- Bagian IV: Umpan-Balik Relevansi
  - Topik 4.1: Pemodelan Umpan-Balik
  - Topik 4.2: Pemodelan Umpan-Balik Implisit & Klik
- Bagian V: Analisis Tautan
  - Topik 5.1: Pagerank dan HITS

Character of this course

- Discussion oriented
  - This is how great ideas are created!
  - You are encouraged to express your thoughts, confusions, and suggestions
  - *Focusing on why, rather than how*

Prerequisites

- Programming skills – Important!
  - Basic data structures: CS 2150 or equivalent
  - **Python** or **Java** is required for machine problems
    - Most open source packages are written in Java and Java
  - Any language you choose for the rest of this course
- Math background
  - Probability
    - Discrete/continuous distributions, expectation, moments
  - Linear algebra
    - Vector, matrix, dot product
  - Optimization
    - Gradient-based methods

Structure of this course

- Six major topics will be covered by lectures
  - E.g., Search engine architecture, retrieval models, search evaluation, relevance feedback, and link analysis
- Latest development will be covered by paper reading assignments and presentations
  - E.g., mobile search, recommendation, personalization, **you name it!**

Grading policy

- Reading assignments (10%)
  - Peer evaluation, after each chapter
- Homework (35%)
  - Problem solving (~3)
- Midterm Exam (20%)
  - Check points of key concepts (in class, 75 minutes)
- Course project (35%)
  - In the exam week

No curving will be applied!

fairness will be guaranteed by the instructor

Grading policy

- Reading assignments (10%)
  - Peer evaluation, after each chapter
- Homework (35%)
  - Problem Solving (~3)
- Midterm Exam (20%)
  - Check points of key concepts (in class, 75 minutes)
- Course project (35%)
  - In the exam week
- Paper presentation (10%)
  - Bonus

No curving will be applied!

fairness will be guaranteed by the instructor

Reading assignments

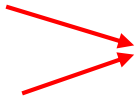
- Read the instructor selected papers after each chapter
- Open-ended essay questions
- Peer evaluation on course forum

Paper presentation

- Choose to present the most recent works in the area of information retrieval
- Peer evaluation
- It’s encouraged to select topics beyond the course content, so as to increase our topic coverage

Midterm exam

- In the second half of semester
- Covers all material we would have learnt by then
- In-class, 75 minutes
- Format
  - True/False question
  - Short answer questions
  - Short essay questions



Fact-based questions

Research-like open discussions

You design your midterm?

- After each chapter, based on your understanding, post one question related to the most important concept in that chapter on our course forum
- Read the others' posted questions and vote on them
- The top voted questions will be included in the midterm
- The authors of those top voted questions should provide the answer, and will get bonus points

Course project

- Topics
  - Implement algorithms in assigned research papers
  - Self-selected topics with permission from the instructor
- Team work
  - 4-5 students per group
- Evaluation
  - Two-page proposal (25%)
  - 15-minutes in-class presentation (40%)
  - Written report (35%)

Late policy

- Homework
  - Submit via Email (no extension)
  - Late penalty: 15%, two weeks after the due date; 50%, afterwards
- Course project
  - Final report is due before presentation (no extension)

Classroom participation

- HIGHLY APPRECIATED!
  - Helps me quickly remember your names
  - Reminds me what is still confusing
  - You can drive the lecture/discussion in this class!

Questions?

Thank you!