

# Comp-8380: Information Retrieval

Jianguo Lu

September 11, 2023

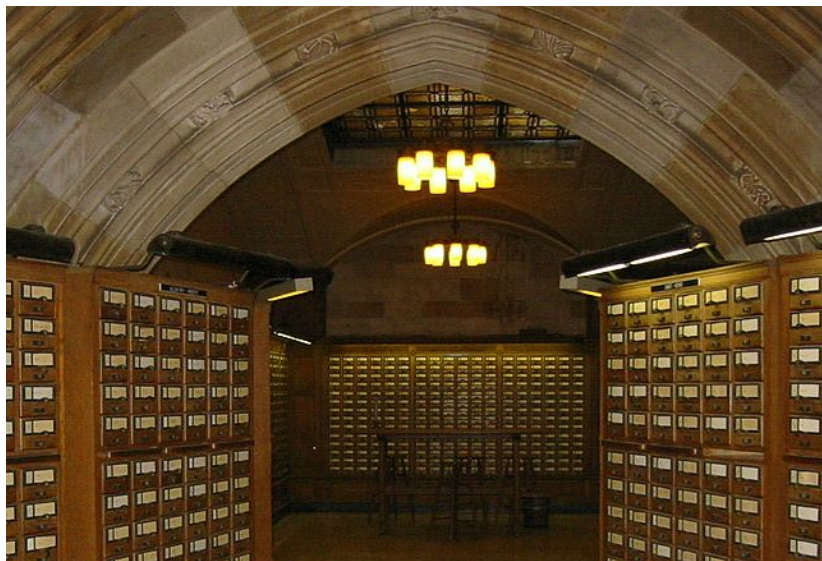
# Outline

- 1 what is IR
- 2 course schedule
- 3 grading scheme

# Outline

- 1 what is IR
- 2 course schedule
- 3 grading scheme

## IR not long time ago





## now IR is mostly about search engines



- there are many search engines ...

► [Study: Facebook use cuts productivity at work - Computerworld](#) 

[www.computerworld.com](http://www.computerworld.com) › Internet › Web 2.0 and Web Apps - Cached

Jul 22, 2009 – A Nucleus Research study found that **Facebook** work in the workplace is cutting employee **productivity**.

[Pulling the Plug on Facebook, Productivity/Time Management Article ...](#) 

[www.inc.com](http://www.inc.com) › Leadership and Managing › Human Resources - Cached

Pulling the Plug on **Facebook**, **Productivity/Time Management Article** - All that friending and superpoking wastes a lot of time at the office -- and could be ...

[Twitter and Facebook: The New Tools of Productivity or Distraction ...](#) 

[www.briansolis.com](http://www.briansolis.com)/.../twitter-and-facebook-the-new-tools-of-prod... - Cached

Mar 26, 2010 – RT **Twitter** and **Facebook**: Yools of **Productivity** or Distraction .... RT @PRSAcolo: **Twitter & Facebook**: New tools of **productivity** or ...

[Twitter, Facebook Can Improve Work Productivity | PCWorld Business ...](#) 

[www.pcworld.com](http://www.pcworld.com)/.../twitter\_facebook\_can\_improve\_work\_produc... - Cached

Apr 2, 2009 – Reach Older Users on **Facebook** and **Twitter** · The Web's Best **Productivity** Sites. According to a study by the Australian University, ...

[Is Facebook Killing Your Employees' Productivity? | WebProNews](#) 

[www.webpronews.com](http://www.webpronews.com)/is-facebook-killing-your-employees-produc... - Cached

Jul 21, 2009 – On the heels of a study indicating that social media can significantly impact a brand's bottom line positively, another one has come out ...

[Productivity Strategies | Facebook](#) 

[www.facebook.com](http://www.facebook.com)/beproductive - Cached

**Productivity Strategies** - To learn more about the Productive Today "Content Collaborative" faculty, click the "Info" tab or this direct link: | **Facebook**.

[Butt Out IT! Facebook "Productivity Loss" Is No Concern of Yours](#) 

[blogs.gartner.com](http://blogs.gartner.com)/.../butt-out-it-facebook-productivity-loss-is-no-co... - Cached

**Facebook "Productivity Loss" Is No Concern of Yours**, by Brian Prentice | November 23, 2008 | 10 Comments. Like my colleague Anthony Bradley, I also speak to ...

[Productivity Levels Plummet After Yale Student Makes Facebook Look ...](#) 


[www.batepost.com](http://www.batepost.com)/.../yale-student-makes-facebook-look-like-evil... - Cached

1 [Effective teaching practices using free Google services: conference tutorial](#)

[Paul Gestwicki](#), [Brian McNely](#)

October 2010 **Journal of Computing Sciences in Colleges** , Volume 26 Issue 1

**Publisher:** Consortium for Computing Sciences in Colleges

Full text available:  [Pdf](#) (22.76 KB)

**Bibliometrics:** Downloads (6 Weeks): 2, Downloads (12 Months): 48, Downloads (Overall): 48, Citation Count: 0


In this 90-minute tutorial, we will share our experiences using free Web services from Google in order to maximize teaching effectiveness. Participants will engage with these services as part of the tutorial. We have used and studied these technologies, ...

2 [Model-Based Engineering of Software: Three Productivity Perspectives](#)

[Shawn A. Bohner](#), [Sriram Mohan](#)

October 2009 **SEW '09: Proceedings of the 2009 33rd Annual IEEE Software Engineering Workshop**

**Publisher:** IEEE Computer Society

Full text available:  [Publisher Site](#)

**Bibliometrics:** Downloads (6 Weeks): n/a, Downloads (12 Months): n/a, Downloads (Overall): n/a, Citation Count: 0

Evolving software products is a tricky business, especially when the domain is complex and changing rapidly. Like other fields of engineering, software engineering productivity advances have come about largely through abstraction reuse, process, and ...

**Keywords:** Agent-Based Software Systems, Model-Driven Architecture, Model-Driven Development, Model-Based Software Development, Model-Based Software Engineering

3 [Absolute Beginner's Guide to Computer Basics, 5th edition](#)

[Michael Miller](#)

September 2009 **Absolute Beginner's Guide to Computer Basics, 5th edition**








**Publisher:** Que Publishing Company

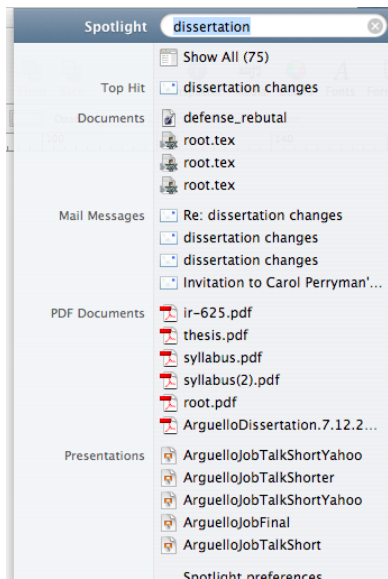
**Bibliometrics:** Downloads (6 Weeks): n/a, Downloads (12 Months): n/a, Downloads (Overall): n/a, Citation Count: 0

Everything casual users need to know to get the most out of their new Windows 7 PCs, software, and the Internet. The best-selling beginner's guide, now completely updated for Windows 7 and today's most popular Internet tools - including Facebook, craigslist, ...



## Places for **mexican food** near Chapel Hill, NC

- A** [Bandido's Mexican Cafe & Cantina](#)  - ★★★★★ 14 reviews - [Place page](#)  
[www.bandidoscafe.com](http://www.bandidoscafe.com) - 159 1/2 East Franklin Street, Chapel Hill - (919) 967-5048
- B** [Las Potrillos Mexican Restaurant](#)  - ★★★★★ 9 reviews - [Place page](#)  
[www.lospotrillos.net](http://www.lospotrillos.net) - 220 West Rosemary Street, Chapel Hill - (919) 932-4301
- C** [monterrey mexican restaurant](#)  - ★★★★★ 17 reviews - [Place page](#)  
[monterreychapelhill.com](http://monterreychapelhill.com) - 237 South Elliot Road, Chapel Hill - (919) 969-8750
- D** [Margaret's Cantina](#)  - ★★★★★ 19 reviews - [Place page](#)  
[www.margaretscantina.com](http://www.margaretscantina.com) - 1129 Weaver Dairy Road, Chapel Hill - (919) 942-4745
- E** [Qdoba Mexican Grill](#)  - ★★★★★ 19 reviews - [Place page](#)  
[www.qdoba.com](http://www.qdoba.com) - 100 West Franklin Street, Chapel Hill - (919) 929-8998
- F** [Cinco de Mayo](#)  - ★★★★★ 11 reviews - [Place page](#)  
[www.cincomayorestaurants.net](http://www.cincomayorestaurants.net) - 1502 East Franklin Street, Chapel Hill - (919) 929-6566
- G** [Chipotle Mexican Grill](#)  - ★★★★★ 15 reviews - [Place page](#)  
[www.chipotle.com](http://www.chipotle.com) - 301 W. Franklin St., Chapel Hill - (919) 942-2091





**neenjames** Neen James

Productivity tip: Follow ppl on Twitter that inspire, challenge and inform you - delete the clutter!

4 minutes ago



**mr\_Ostentatious** Jason Pitts

Took a day off from **twitter** to increase my **productivity** and ended up having a productive day!

1 hour ago



**adamwiebe** Adam Wiebe

Social media at work is here. Be wary of what is **and** is not productive. <http://nkd.in/DW3z8J>

3 hours ago



**ViggosDaddy** Gert van der Linde

A brief look: To tweet, or not to tweet? - How does **Twitter** affect our **productivity**, influence **and** how informe... <http://tinyurl.com/3wbz3m>

6 hours ago



**IncorrectMystic** Raghavender | raGz

**#productivity** day - So going be off **twitter** **and** other social networks till work is over :) bye tweeples for a while

6 hours ago



**Michael Jordan**

Page

12,856,455 people like this.

---



**Michael Jordan**

Carnegie Mellon

---



**Michael Jordan**

1 mutual friend

---



**Michael Jordan**

Page

215,268 people like this.

---



**Michael jordan**

Page

225,371 people like this.

---



**MICHAEL JORDAN**

Page

190,013 people like this.

---



**Michael Jordan**

Page

58,003 people like this.

## IR is more than web search

These days we frequently think first of web search, but there are many other cases:

- digital library search
- E-mail search, Searching your desktop and laptop computers
- Corporate knowledge bases, local business search, expert search
- Legal information retrieval, patent search
- news search
- image and video search
- (micro-)blog search
- product search, federated search
- social search, community Q&A, question-answering
- recommender systems
- opinion mining

## definition of information retrieval

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).

–from IIR book.

- Introduction to Information Retrieval, by C. Manning, P. Raghavan, and H. Schütze. Cambridge University Press
- book website <https://nlp.stanford.edu/IR-book/>

## definition of information retrieval

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).

–from IIR book.

- Introduction to Information Retrieval, by C. Manning, P. Raghavan, and H. Schütze. Cambridge University Press
- book website <https://nlp.stanford.edu/IR-book/>

## definition of information retrieval

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).

–from IIR book.

- Introduction to Information Retrieval, by C. Manning, P. Raghavan, and H. Schütze. Cambridge University Press
- book website <https://nlp.stanford.edu/IR-book/>



## definition of information retrieval

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).

–from IIR book.

- Introduction to Information Retrieval, by C. Manning, P. Raghavan, and H. Schütze. Cambridge University Press
- book website <https://nlp.stanford.edu/IR-book/>

## definition of information retrieval

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).

–from IIR book.

- Introduction to Information Retrieval, by C. Manning, P. Raghavan, and H. Schütze. Cambridge University Press
- book website <https://nlp.stanford.edu/IR-book/>

## definition of information retrieval

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).

–from IIR book.

- Introduction to Information Retrieval, by C. Manning, P. Raghavan, and H. Schütze. Cambridge University Press
- book website <https://nlp.stanford.edu/IR-book/>

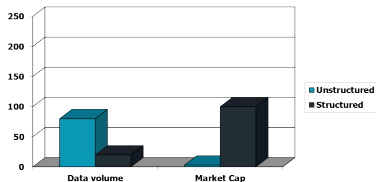
## definition of information retrieval

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).

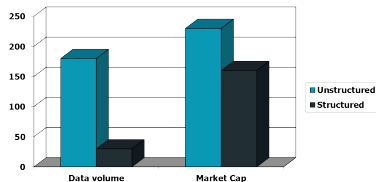
–from IIR book.

- Introduction to Information Retrieval, by C. Manning, P. Raghavan, and H. Schütze. Cambridge University Press
- book website <https://nlp.stanford.edu/IR-book/>

## Structured vs. unstructured data



in the 90's.



today

Information retrieval is finding material of an **unstructured** nature that satisfies an information need from within large collections

## other definitions

### Jaime Arguello

- Information retrieval (IR) is the science and practice of **designing, developing, and evaluating** systems that match information seekers with the information they seek.

### Gerard Salton, 1968:

- Information retrieval is a field concerned with the **structure, analysis, organization, storage, and retrieval** of information.

# The search task

Given a query and a corpus, find relevant items

- query: user's expression of their information need
- corpus: a repository of retrievable items
- relevance: satisfaction of the user's information need

## Corpus: definition from Webster

- a : all the writings or works of a particular kind or on a particular subject; especially : the complete works of an author
- b : a collection or body of knowledge or evidence; especially : a collection of recorded utterances used as a basis for the descriptive analysis of a language

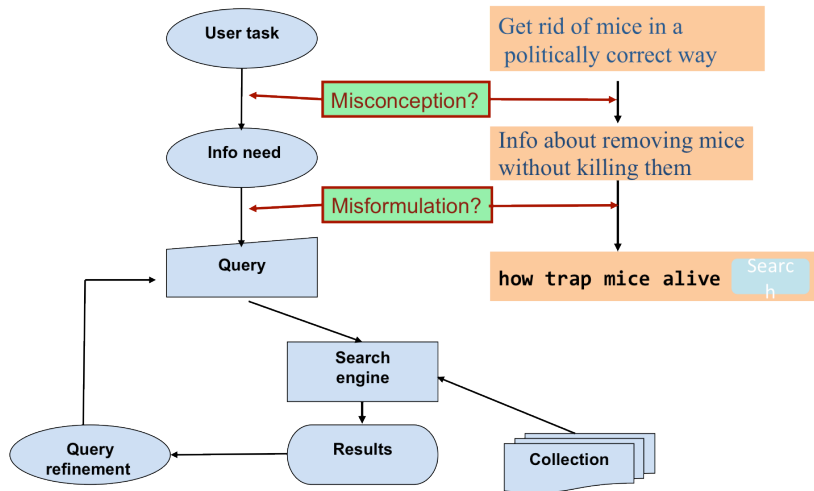
# Why is IR fascinating?

## Information retrieval is an uncertain process

- Query
  - users don't know what they want
  - users don't know how to convey what they want
  - computers can't elicit information like a librarian
  - computers can't understand natural language text well
- Relevance
  - the search engine can only guess what is relevant
  - the search engine can only guess if a user is satisfied
  - over time, we can only guess how users adjust their short- and long-term behavior



## classic search model

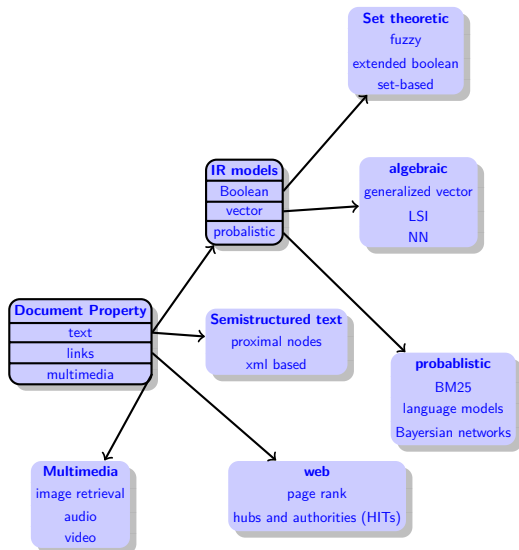


- A query is an impoverished description of the user's information need
- Highly ambiguous to anyone other than the user

## Retrieval Model

- A formal method that predicts the degree of relevance of a document to a query

# taxonomy of IR models



## Boolean Retrieval Model

- The user describes their information need using boolean constraints (e.g., AND, OR, and AND NOT)
- The burden is on the user to formulate a good boolean query

## Example

- Which plays of Shakespeare contain the words Brutus AND Caesar but NOT Calpurnia
- One choice: use grep command in unix.
  - grep all of Shakespeare's plays for Brutus and Caesar,
  - strip out lines containing Calpurnia
- Why is that not the answer?
  - Slow (for large corpora)
  - NOT Calpurnia is non-trivial
  - Other operations (e.g., find the word Romans near countrymen) not feasible
  - Ranked retrieval (best documents to return)

so we need to index the text

## Example

- Which plays of Shakespeare contain the words Brutus AND Caesar but NOT Calpurnia
- One choice: use grep command in unix.
  - grep all of Shakespeare's plays for Brutus and Caesar,
  - strip out lines containing Calpurnia
- Why is that not the answer?
  - Slow (for large corpora)
  - NOT Calpurnia is non-trivial
  - Other operations (e.g., find the word Romans near countrymen) not feasible
  - Ranked retrieval (best documents to return)

so we need to index the text

## Example

- Which plays of Shakespeare contain the words Brutus AND Caesar but NOT Calpurnia
- One choice: use grep command in unix.
  - grep all of Shakespeare's plays for Brutus and Caesar,
  - strip out lines containing Calpurnia
- Why is that not the answer?
  - Slow (for large corpora)
  - NOT Calpurnia is non-trivial
  - Other operations (e.g., find the word Romans near countrymen) not feasible
  - Ranked retrieval (best documents to return)

so we need to index the text

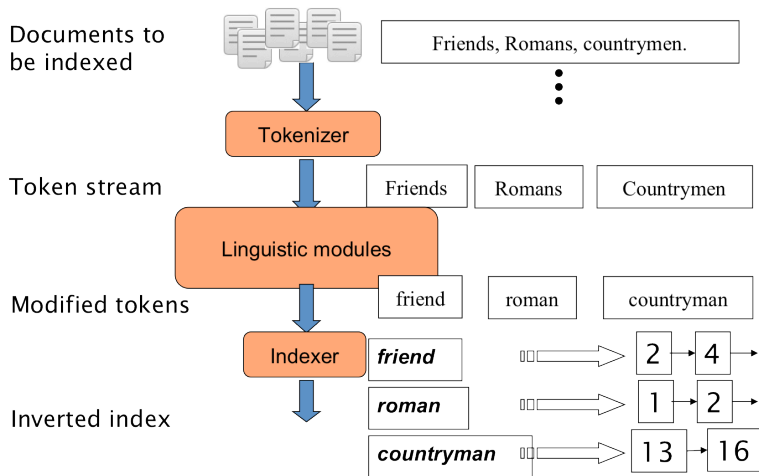
## what is an index

### INDEX

- ABC, 164, 321*n*  
academic journals, 262, 280–82  
Adobe eBook Reader, 148–53  
advertising, 36, 45–46, 127, 145–46, 167–68, 321*n*  
Africa, medications for HIV patients in, 257–61  
Agee, Michael, 223–24, 225  
agricultural patents, 313*n*  
Aibo robotic dog, 153–55, 156, 157, 160  
AIDS medications, 257–60  
air traffic, land ownership vs., 1–3  
Akerlof, George, 232  
Alben, Alex, 100–104, 105, 198–99, 295, 317*n*  
alcohol prohibition, 200  
*Alice's Adventures in Wonderland* (Carroll), 152–53  
Anello, Douglas, 60  
animated cartoons, 21–24  
antiretroviral drugs, 257–61  
Apple Corporation, 203, 264, 302  
architecture, constraint effected through, 122, 123, 124, 318*n*  
archive.org, 112  
    *see also* Internet Archive  
archives, digital, 108–15, 173, 222, 226–27  
Aristotle, 150  
Armstrong, Edwin Howard, 3–6, 184, 196  
Arrow, Kenneth, 232  
art, underground, 186  
artists:  
    publicity rights on images of, 317*n*  
    recording industry payments to, 52, 58–59, 74, 195, 196–97, 199, 301, 329*n*–30*n*



## index construction process



## Initial stages of text processing

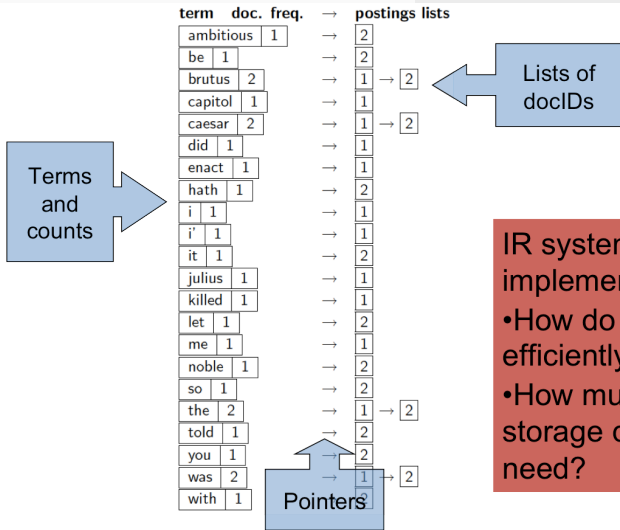
- Tokenization
  - Cut character sequence into word tokens
- Normalization
  - Map text and query term to same form
    - You want **U.S.A.** and **USA** to match
- Stemming
  - We may wish different forms of a root to match
    - **authorize, authorization**
- Stop words
  - We may want to omit very common words (modern methods may not)
    - **the, a, to, of**

# postings

Term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
i	1
i	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

term	doc. freq.	→	postings lists
ambitious	1	→	2
be	1	→	2
brutus	2	→	1 → 2
capitol	1	→	1
caesar	2	→	1 → 2
did	1	→	1
enact	1	→	1
hath	1	→	2
i	1	→	1
i'	1	→	1
it	1	→	2
julius	1	→	1
killed	1	→	1
let	1	→	2
me	1	→	1
noble	1	→	2
so	1	→	2
the	2	→	1 → 2
told	1	→	2
you	1	→	2
was	2	→	1 → 2
with	1	→	2

- Multiple term entries in a single document are merged.
- Split into Dictionary and Postings
- Doc. frequency information is added.



IR system implementation

- How do we index efficiently?
- How much storage do we need?

## query processing

- Consider processing the query:
  - Brutus AND Caesar
- Locate Brutus in the Dictionary;
- Retrieve its postings.
- Locate Caesar in the Dictionary;
- Retrieve its postings.
- Merge the two postings (intersect the document sets):

brutus → 1 2 4 11 31 45 173 174

caesar → 1 2 4 5 6 16 57 132

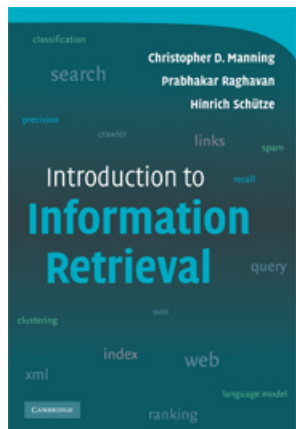
# Outline

- 1 what is IR
- 2 course schedule
- 3 grading scheme

## tentative schedule

- boolean model
- text transformation
- build a search engine using Lucene
- vector space model
- representation learning
- evaluation methods in information retrieval
- link analysis and PageRank
- document classification
- document clustering
- web crawling. Data cleaning (e.g. near-duplicate detection)

## Text Book



[IIR] Introduction to Information Retrieval, by C. Manning, P. Raghavan, and H. Schütze. Cambridge University Press, 2008.



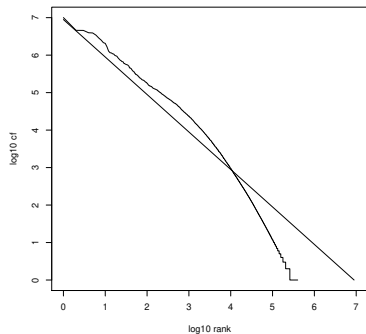
## Other reference books

- SE** Search Engines: Information Retrieval in Practice, by Bruce Croft, Donald Metzler and Trevor Strohman.
- MIR** Modern Information Retrieval, by R. Baeza-Yates and B. Ribeiro-Neto. 2-nd edition 2010.
- MMD** Anand Rajaraman and Jeff Ullman, Mining of massive datasets , 2013.

## IIR 02: The term vocabulary and postings lists

- Phrase queries: “STANFORD UNIVERSITY”
- Proximity queries: GATES NEAR MICROSOFT
- We need an index that captures **position information** for phrase queries and proximity queries.

## statistic properties of text



- Zipf's law, heaps' law, power law.
- the mechanism: Yule process, Preferential attachment

## IIR 06: Scoring, term weighting and the vector space model

- Ranking search results
  - Boolean queries only give inclusion or exclusion of documents.
  - For ranked retrieval, we measure the proximity between the query and each document.
  - One formalism for doing this: [the vector space model](#)
- Key challenge in ranked retrieval: evidence accumulation for a term in a document
  - 1 vs. 0 occurrence of a query term in the document
  - 3 vs. 2 occurrences of a query term in the document
  - Usually: more is better
  - But by how much?
  - Need a scoring function that translates frequency into score or weight

## Language models

- assign a probability to a sequence of  $m$  words by means of a probability distribution.
- How to compute this joint probability:

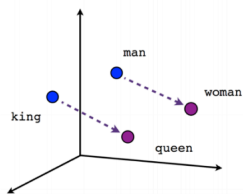
$$P(\textit{its, water, is, so, transparent, that}) \quad (1)$$

$$P(w_1 w_2 \dots w_n) = \prod P(w_i)? \quad (2)$$

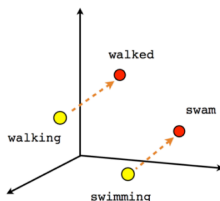
## Text classification & Naive Bayes

- Text classification = assigning documents automatically to predefined classes
- Examples:
  - CS vs. Non-CS papers
  - Papers in Software Engineering vs. Database
  - positive/negative reviews
  - Spams
- Naive Bayes (Multinomial and Bernoulli model), Support vector machine, feature selection, representation learning, neural networks.

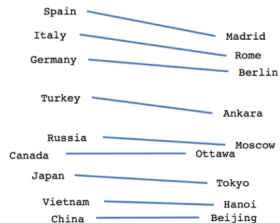
# Neural network based representation learning



Male-Female



Verb tense



Country-Capital

Answer analogical questions, e.g

$$Man : Woman = King : ?$$

The answer will be Queen.

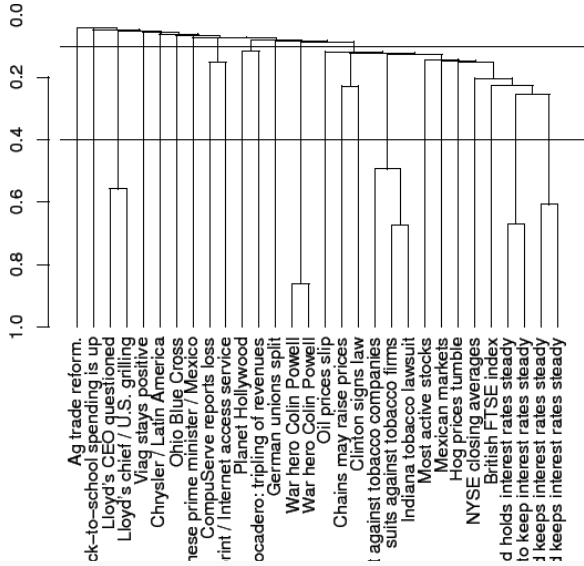
- word2vec,
- Pretrained LLMs, BERT, LLaMA.
- Fine-tuning LLMs

# clustering

- Flat clustering
- Hierarchical agglomerative clustering (HAC)
- Single-link and complete-link clustering
- Centroid and group-average agglomerative clustering (GAAC)
- Bisecting K-means
- How to label clusters automatically



# HAC

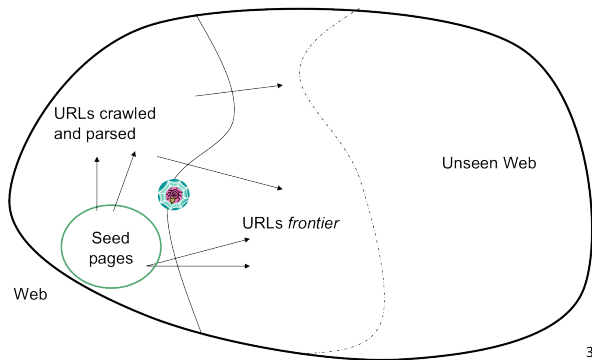


# Latent Semantic Indexing

- how to find semantically related documents?
- matrix decomposition
- SVD

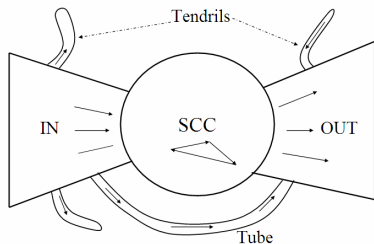
$$\begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{bmatrix} \begin{bmatrix} \bullet & & & & \\ & \bullet & & & \\ & & \bullet & & \\ & & & \bullet & \\ & & & & \bullet \end{bmatrix} \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix}$$
$$\begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{bmatrix} = \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} \begin{bmatrix} \bullet & & \\ & \bullet & \\ & & \bullet \end{bmatrix} \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{bmatrix}$$

# Crawling



## Link analysis / PageRank

- which web page is more important?
- who are in a community?
- PageRank algorithm
- graph analysis and mining. Graph Convolution Neural Network .



# Outline

- 1 what is IR
- 2 course schedule
- 3 grading scheme

## marking scheme

- exam 50%
- project 50%

## project

- build searching engine
- Similar to google but domain specific
  - on CS papers
  - provide better search experience
- enhance the search engine by adding one or more features, such as:
  - semantic search
  - classification
  - clustering (returning results (papers) are clustered into several areas)
  - ranking (ranked by PageRank algorithm)
  - personalization
  - recommendation (recommend most similar papers)
  - ...

## The project

- 10%: Phase one. A generic search engine for academic papers.
  - Workable search engine and basic extensions.
  - One page report and class presentation.
  - Earlier presenters choose the features they want.
  - Later presenters need to implement and present different features.
- 15% Phase two. Add one feature on the search engine. e.g.
  - Rank documents using the PageRank algorithm using citation data
  - Return results by categories (By running clustering algorithms)
  - Search for the most similar papers (e.g., running doc2vec)
  - ...
- 25% Phase three: Integrate two or more features into a real search engine.