Welcome to the class of
Web Information Retrieval !

Min ZHANG (张敏)
z-m@tsinghua.edu.cn

Course basic information

Instructor

Instructor: Min ZHANG 张敏

- Associate Professor,
- IR group, DCST, Tsinghua Uni.
- Office Addr: Room 1-506B, Building FIT
- Tel: 62792595
- Email: z-m@tsinghua.edu.cn
- http://www.thuir.cn/group/~mzhang

TA: Chao WEI 魏超

- Email: weichao053825@163.com

And how about you?

Arrangement

- First 11 weeks
  - Lectures by the teacher on selected topics
  - With "Afternoon Tea Time" discussions on the news on Web/IR industry/research during the past week (by one student)
    - Send the email by Wed. noon 11:59am for applying a tea time show on Fri.
- The following 4 weeks -- A workshop
  - Lectures by the students
  - (tentative) 15 mins' talk + 15 mins' QA
- Last week
  - Course Overview, awards and celebration to "the Best Lecture".

Tentative Syllabus / Topics (1)

- Introduction to
  - The course
  - What's IR, Basic procedure
- Key techniques of an IR system
  - Data acquisition (esp. crawler)
  - Indexing
  - Weighting and Ranking models
  - Evaluation
- Web IR
  - Web –specific features
  - Link analysis
  - User behavior analysis
  - Challenges (e.g. scale, quality, anti-spam, multi-resource fusion, UI, etc)

Tentative Syllabus / Topics (2)

- Re-thinking of Evaluation
  - Methodology
  - Metrics
  - Web scale evaluation
- Visual IR
  - Low + high level features
  - content-based VIR
  - Semantic-based IR
  - HCI
- Social computing and search
- Other topics (optional)
  - Opinion Retrieval, Mobile search, etc

Course workshop

- Everybody need to give a lecture on the workshop.

  Your talk should include:

  - 1. Introduce the search engines you'd like to use in your motherland, and Compare it with Google, if it is not Google
    - If it's Google, then pls. compare it with another SE (e.g. Bing, Baidu, …)
  - 2. What are major problems on current SE (that makes you unsatisfied)?
  - 3. What is the ideal future SE in your mind?
    - You can even show a design with pictures, animation, …
  - Any other topics (optional) …

- Write & submit a 4~6 pages paper before the workshop (required)

Evaluation

(Subject to modifications)

- Workshop (~40%)
  - Evaluated by the other students (20%)
  - + by the teachers (20%)
  - We will have a best presentation award
- The paper (~20%)
- Homework (~ 40%)

Active thinking and discussions are highly encouraged !

References

- We're not having an official textbook
  - There isn't one with good coverage of all & only the topics we'll discuss
  - A changing field, advanced topics
- A list of references:
  - Books
    - R. Baeza-Yates and B. Ribeiro-Neto, **Modern Information Retrieval**
    - Christopher D. Manning, Prabhakar Raghavan ,Hinrich Schütze, **Introduction to information retrieval**
    - I. Witten, A. Moffat, and T. Bell, **Managing Gigabytes**
  - Proceedings of Conferences
    - SIGIR, WWW, WSDM, CIKM, TREC, ...
  - Very important: Web resources, Search engines

Brief introduction to IR foundations

Mainly Text IR

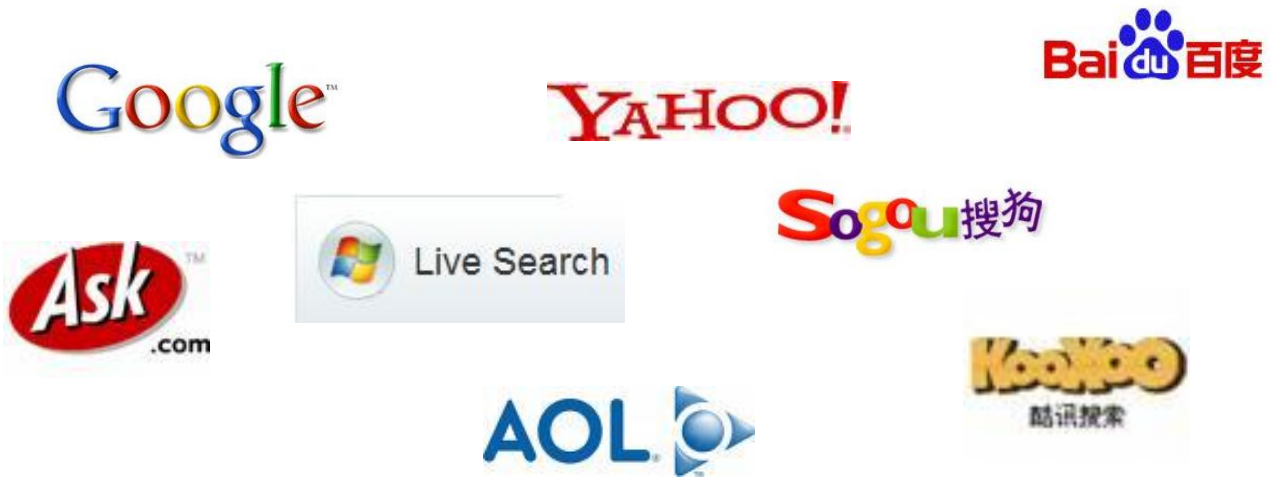
Visual IR will be discussed later by separate lectures

1. What's IR?

Figure Copyright by TREC

What is Information Retrieval (IR)?

- Narrow-sense:
  - IR = Search Engine Technologies (IR=Google, baidu, sogou, yahoo, live, …, library info systems)
  - IR = Text matching



What is Information Retrieval (IR)?

- Broad-sense: IR ~ **Information Management**
  - General problem: how to manage information?
  - How to **find** useful information? (retrieval)
    - e.g., google, baidu, sogou, kooxun, soso, yahoo, live search, ……
  - How to **organize** information? (classification)
    - e.g., automatically assign email to different folders
  - How to **discover** knowledge from the data? (mining)
    - e.g., discover correlation of events
- “搜索无处不在” -- by 李彦宏 in early years

IR foundations – what's IR?

- **Goal**:
  - Find documents *relevant* to an information need from a large document set

- And now:
  - Beyond relevance
  - Document: multi-modal
  - User's information need

Figure Copyright by TREC

IR is Hard!

- Under/over-specified query
  - Ambiguous: “buying CDs” (money or music?)
  - Incomplete: what kind of CDs?
  - What if “CD” is never mentioned in document?
- Vague semantics of documents
  - Ambiguity: word-sense, structural
    - e.g. “A fly flied into the kitchen.” “bank”
  - Incomplete: Inferences required
    - E.g. “windows”
- Also hard for human beings!
  - 80% agreement in human judgments

IR is "Easy"!

- IR CAN be easy in a particular case
  - Ambiguity in query/document is RELATIVE to the database
  - So, if the query is SPECIFIC enough, just one keyword may get all the relevant documents
- PERCEIVED IR performance is usually better than the actual performance
  - Users can NOT judge the completeness of an answer
  - E.g. Web Search vs. Machine Translation

History of IR* on One Slide

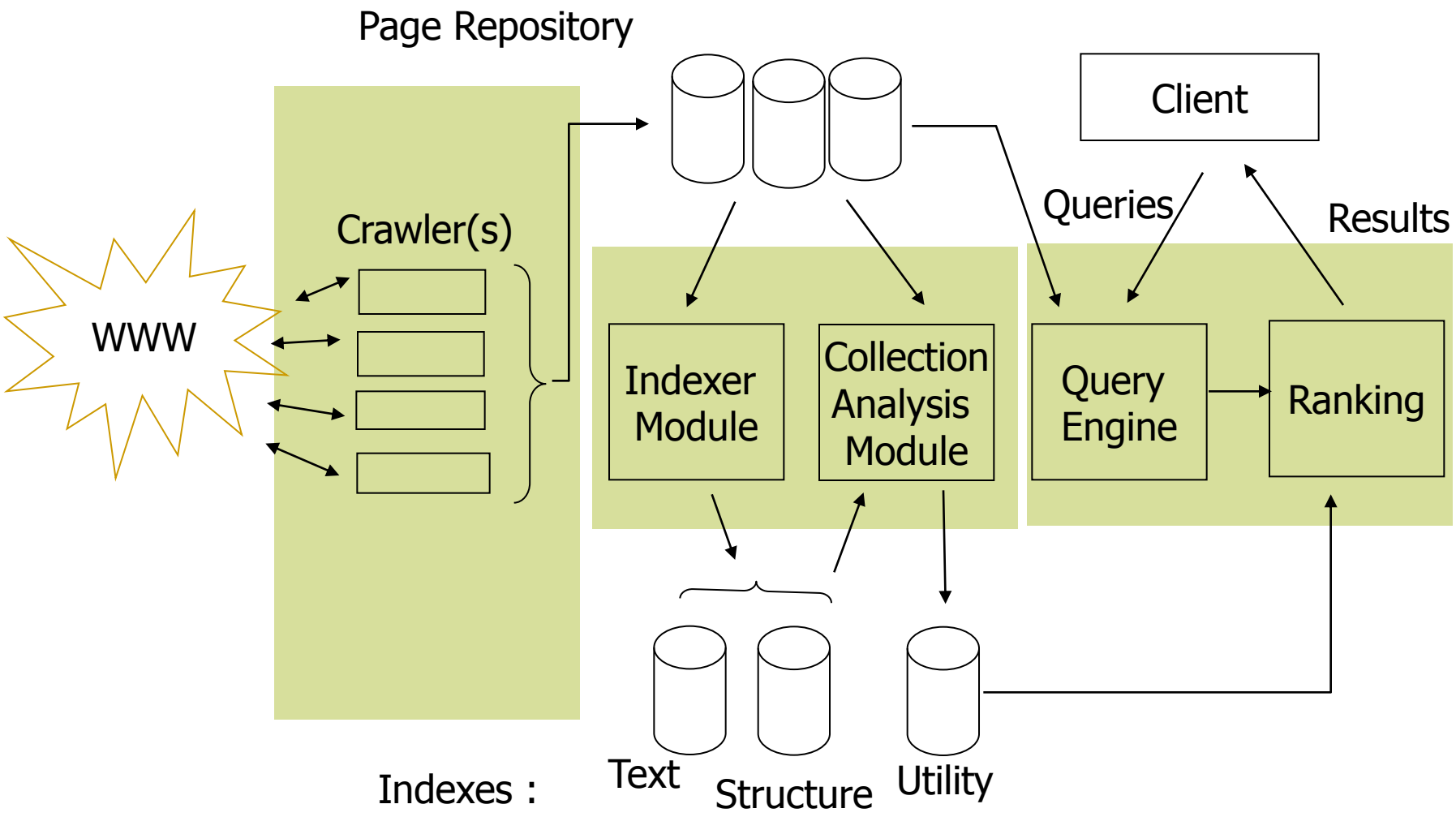
**(The history of Web Search will be discussed in later lectures)*

- Birth of IR
  - 1945: Vannevar Bush's article "As we may think"
  - 1957: H. P. Luhn's idea of word counting and matching
- Indexing & Evaluation Methodology (1960's)
  - Smart system (G. Salton's group)
  - Cranfield test collection (C. Cleverdon's group)
  - Indexing: automatic can be as good as manual (controlled vocabulary)
- IR Models (1970's & 1980's, late 1990's & early 2000's) …
- Large-scale Evaluation & Applications (1990's-Present)
  - TREC (D. Harman & E. Voorhees, NIST)， CLEF, NTCIR, …
  - Web search, PubMed, …
  - Boundary with related areas are disappearing

2. A general (Basic) IR procedure

Figure Copyright by TREC

Example: search engine architecture



Basic IR procedure

- Data acquisition
  - How to collect fulfill resources?
- Document and query indexing
  - How to represent their contents?
- Ranking
  - How to measure the (ordered) relevance between a document and the query?
- System evaluation
  - How good is a system? Are the retrieved documents relevant and useful?

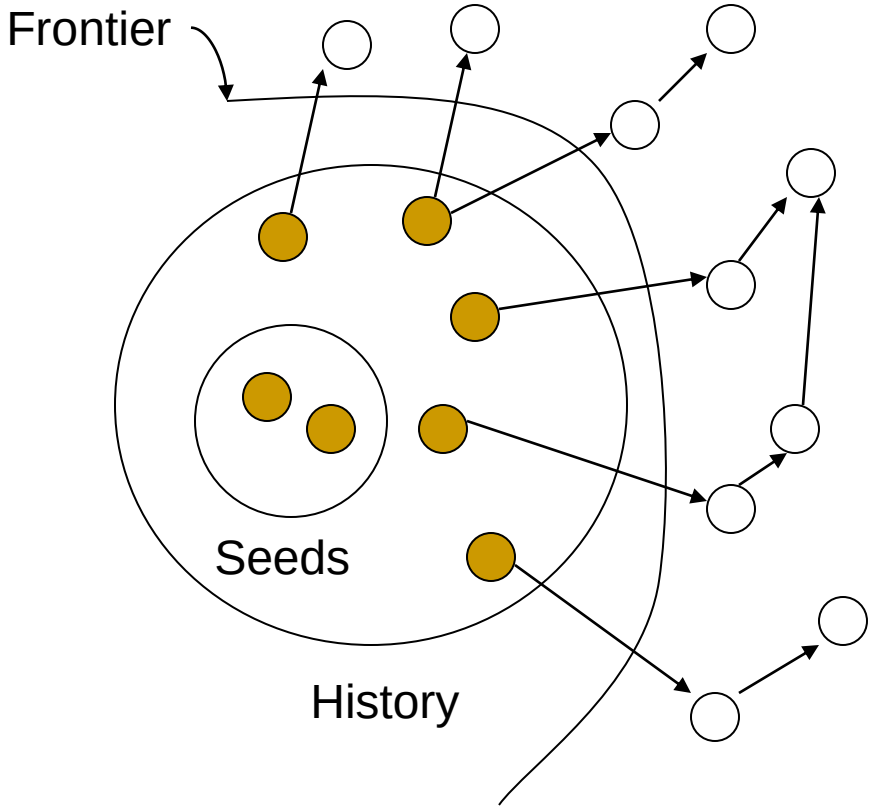
Outline

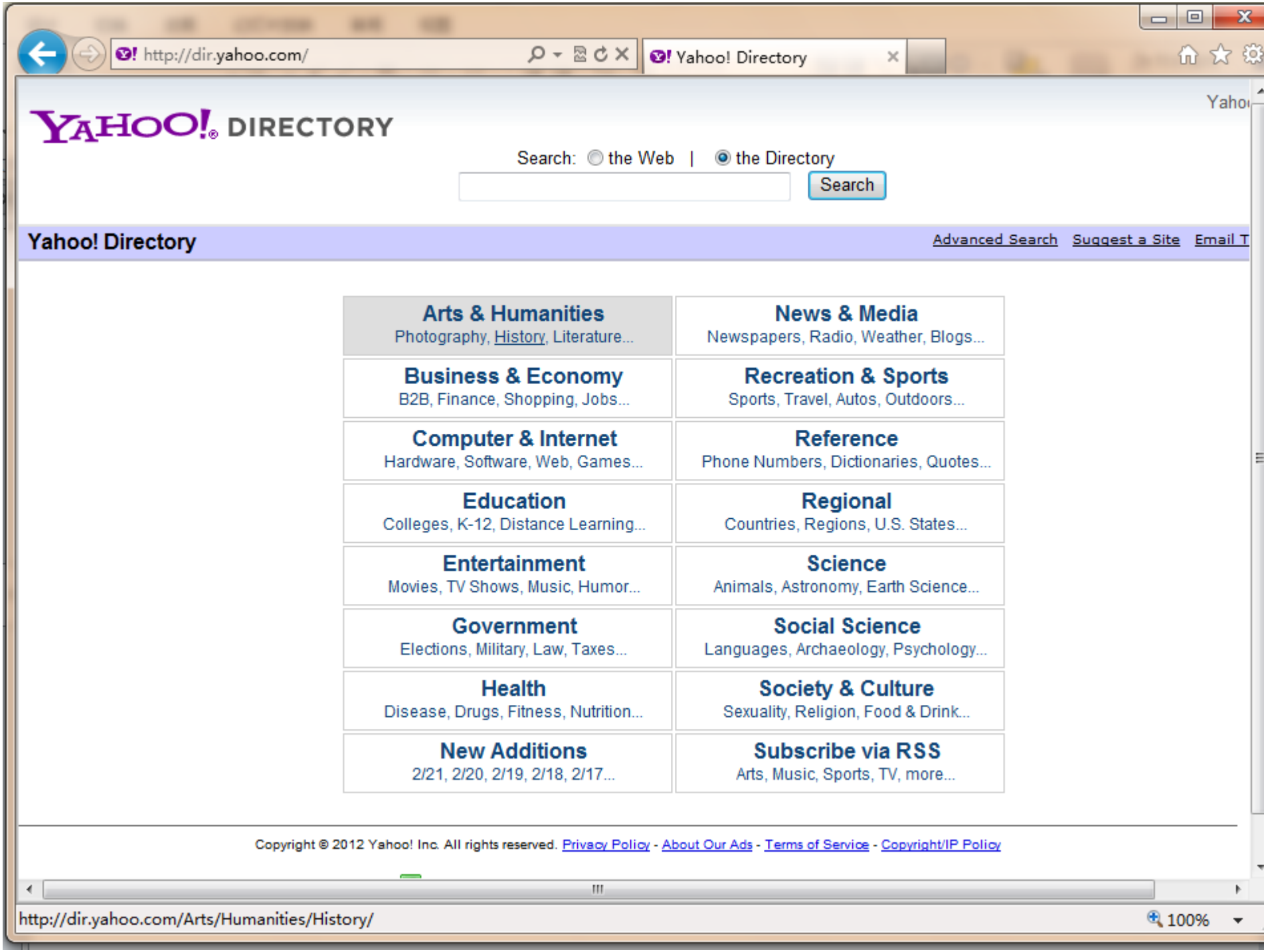
- What is IR?
- Basic IR procedure
  - Data acquisition – on the Web: Crawler
  - Indexing
  - Ranking
  - System evaluation

Crawler – Crawl "all" Web pages?

- Problem: no catalog of all accessible URLs on the Web.
- Solution (basic crawler operation)
  - 1. Given: Initial set of URLs U

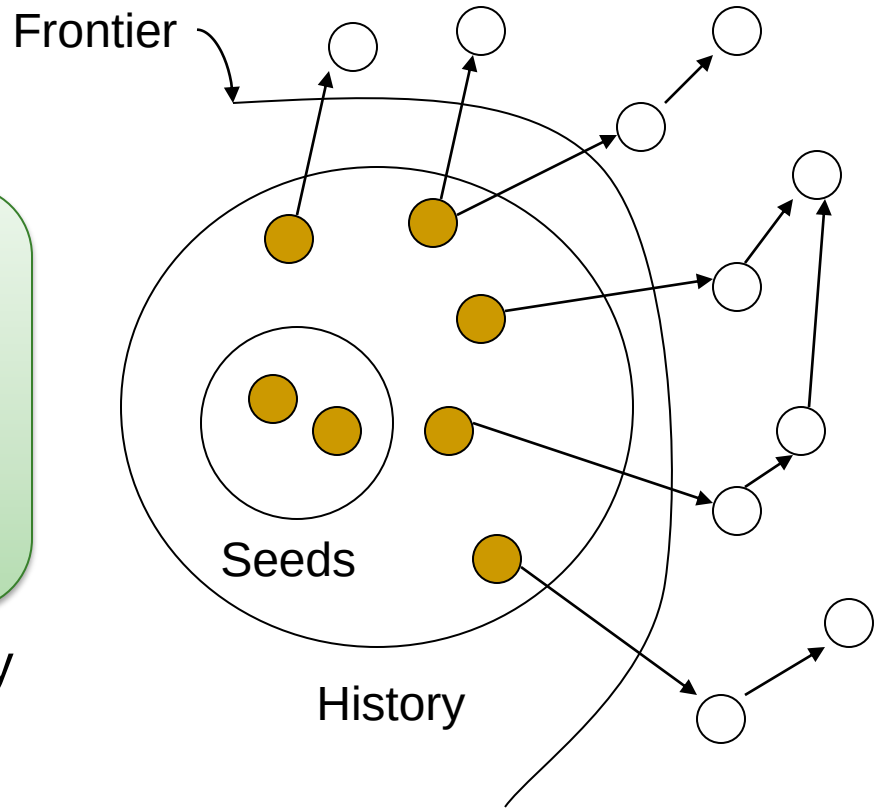
    (in some order) -- "seed" pages
  - 2. Get next URL u from U
  - 3. Download web page p(u)




http://dir.yahoo.com/
Yahoo! Directory
YAHOO! DIRECTORY
Search: the Web | the Directory
Search
Yahoo! Directory
Advanced Search
Suggest a Site
Email T
Arts & Humanities
Photography, History, Literature...
News & Media
Newspapers, Radio, Weather, Blogs...
Business & Economy
B2B, Finance, Shopping, Jobs...
Recreation & Sports
Sports, Travel, Autos, Outdoors...
Computer & Internet
Hardware, Software, Web, Games...
Reference
Phone Numbers, Dictionaries, Quotes...
Education
Colleges, K-12, Distance Learning...
Regional
Countries, Regions, U.S. States...
Entertainment
Movies, TV Shows, Music, Humor...
Science
Animals, Astronomy, Earth Science...
Government
Elections, Military, Law, Taxes...
Social Science
Languages, Archaeology, Psychology...
Health
Disease, Drugs, Fitness, Nutrition...
Society & Culture
Sexuality, Religion, Food & Drink...
New Additions
2/21, 2/20, 2/19, 2/18, 2/17...
Subscribe via RSS
Arts, Music, Sports, TV, more...
Copyright © 2012 Yahoo! Inc. All rights reserved. Privacy Policy - About Our Ads - Terms of Service - Copyright/IP Policy
http://dir.yahoo.com/Arts/Humanities/History/
100%

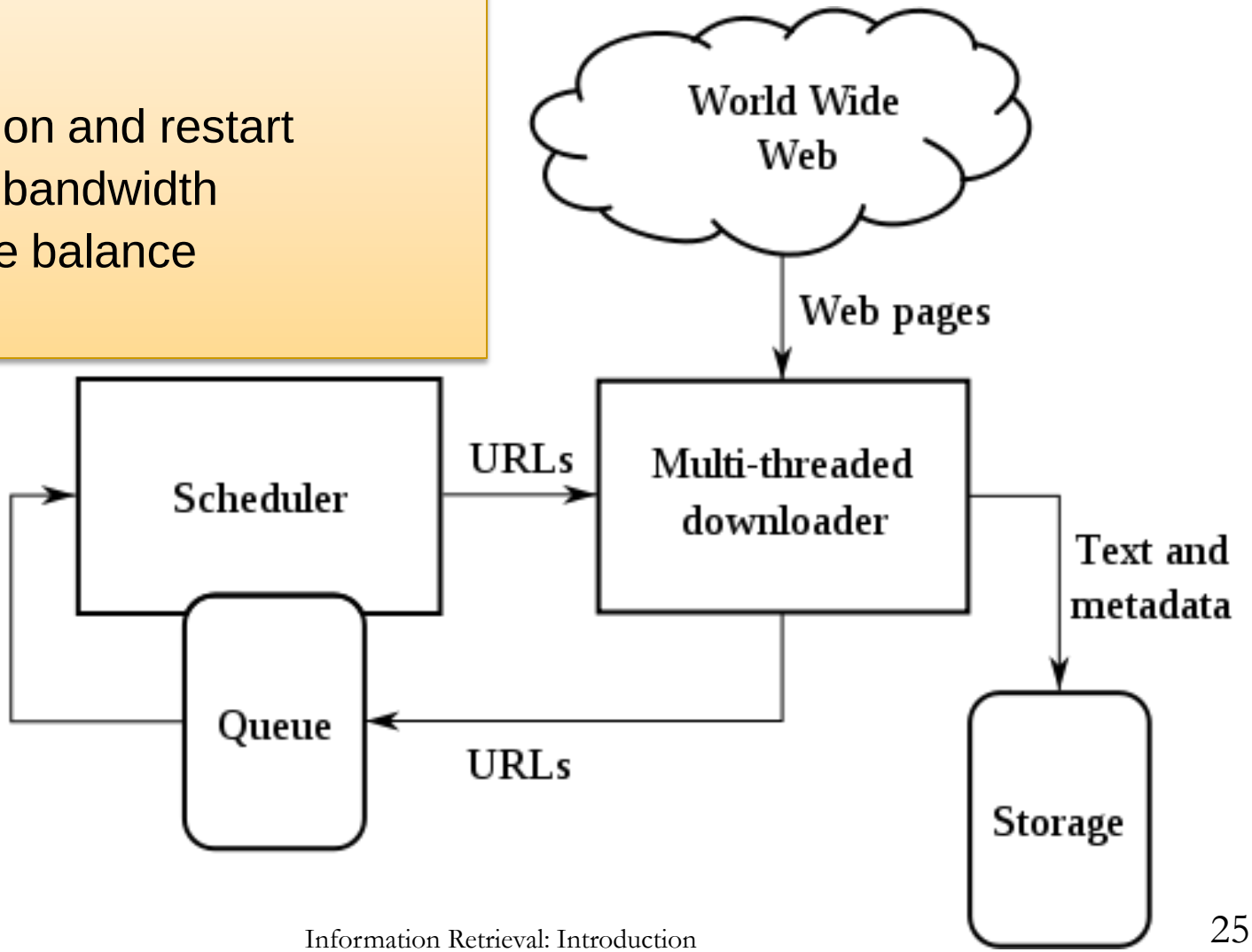
Crawler – Crawl "all" Web pages?

- Problem: no catalog of all accessible URLs on the Web.
- Solution (basic crawler operation)
  - 1. Given: Initial set of URLs U (in some order) -- "seed" pages
  - 2. Get next URL u from U
  - 3. Download web page p(u)
  - 4. Extract all URLs from p(u), add them to U
  - 5. Send p(u) to the indexer
  - 6. Continue with 2. until U is empty (or some stop criteria is fulfilled)



Web Crawler Architecture

- Breadth-first or Depth-first?
- Priority
- Timeout
- Interruption and restart
- Network bandwidth
- Resource balance
- ……



"It is fairly easy to build a slow crawler that downloads a few pages per second for a short period of time, building a high-performance system that can download hundreds of millions of pages over several weeks presents a number of challenges in system design, I/O and network efficiency, and robustness and manageability."

Eichmann, D. (1994). The RBSE spider: balancing effective search against Web load. In Proceedings of the First World Wide Web Conference, Geneva, Switzerland.

APPENDIX

As we may think

This article is reprinted in its entirety, with permission, from The Atlantic Monthly, July, 1945. A condensation was printed by Life Magazine in 1945, with illustrations. The article has been reprinted variously since then; it can be found at The Atlantic's own site, at http://www2.theAtlantic.com/atlantic/atlweb/flashbks/computer/tech.htm and also at http://www.isg.sfu.ca/~duchier/misc/vbush/.

As We May Think

Vannevar Bush

As Director of the Office of Scientific Research and Development, Dr. Vannevar Bush has coordinated the activities of some six thousand leading American scientists in the application of science to warfare. In this significant article he holds up an incentive for scientists when the fighting has ceased. He urges that men of science should then turn to the massive task of making more accessible our bewildering store of knowledge. For many years inventions have extended man's physical powers rather than the powers of his mind. Trip hammers that multiply the fists, microscopes that sharpen the eye, and engines of destruction and detection are new results, but not the end results, of modern science. Now, says Dr. Bush, instruments are at hand which, if properly developed, will give man access to and command over the inherited knowledge of the ages. The perfection of these pacific instruments should be the first objective of our scientists as they emerge from their war work. Like Emerson's famous address of 1837 on "The American Scholar," this paper by Dr. Bush calls for a new relationship between thinking man and the sum of our knowledge.

—The [*Atlantic Monthly*] Editor, July 1945

Set a goal of fast access to the contents of the world's libraries:

- A 1M book library

LUHN H.P.

- LUHN, H.P., 'A statistical approach to mechanised encoding and searching of library information', *IBM Journal of Research and Development,* 1, 309-317 (1957).

- 'It is here proposed that the frequency of word occurrence in an article furnishes a useful measurement of word significance. It is further proposed that the relative position within a sentence of words having given values of significance furnish a useful measurement for determining the significance of sentences. The significance factor of a sentence will therefore be based on a combination of these two measurements.'

---- LUHN, H.P., 'The automatic creation of literature abstracts', *IBM Journal of Research and Development*, 2, 159-165 (1958).