Welcome to the class of Advanced Topics in Information Retrieval !

Min ZHANG （张敏）

z-m@tsinghua.edu.cn

Tea Time

IBM Watson DeepQA

z-m@tsinghua.edu.cn

Jeopardy: An American TV show

Requires the players to suss out the subtleties of language from jokes and puns to irony and anagrams

IBM Watson @ Jeopardy

- February 14, 15, and 16, 2011
  - *Jeopardy's* two biggest champions
  - Brad Rutter (right):
    - Won a whopping $3.25 million playing *Jeopardy*, the most cash ever awarded on the show.
    - He is a Johns Hopkins University dropout
  - Ken Jennings (left):
    - Holds the title for longest *Jeopardy* winning streak, with 74 consecutive wins in 2004.
    - He holds degrees in computer science and English, from Brigham Young University, and an international BA diploma from Seoul Foreign School.

IBM Watson won the Jeopardy

- Towards the Open Advancement of Question Answering Systems

Final:

$77,147

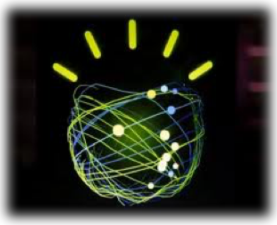
(5,000+35,734+41,413)

vs.

$21,600 &

$24,000.

IBM Watson



- In development for 4 years
- Runs on 90 Power 7 servers
  - Each: 4*8 power 7
- Does not connect to the Internet
- Search on a large scale knowledge base, not Internet
  - Search on billion pages within 3s
- Trained with previous questions and games
  - With Jeopardy players: 77 (2009) + 55 (2010, winners)
  - Lack of real-time learning ability
  - E.g. Category: US Cities
    - *Q: "Its largest airport was named for a World War II hero; its second largest, for a World War II battle."*
    - *A: "What is Chicago / Toronto?"*

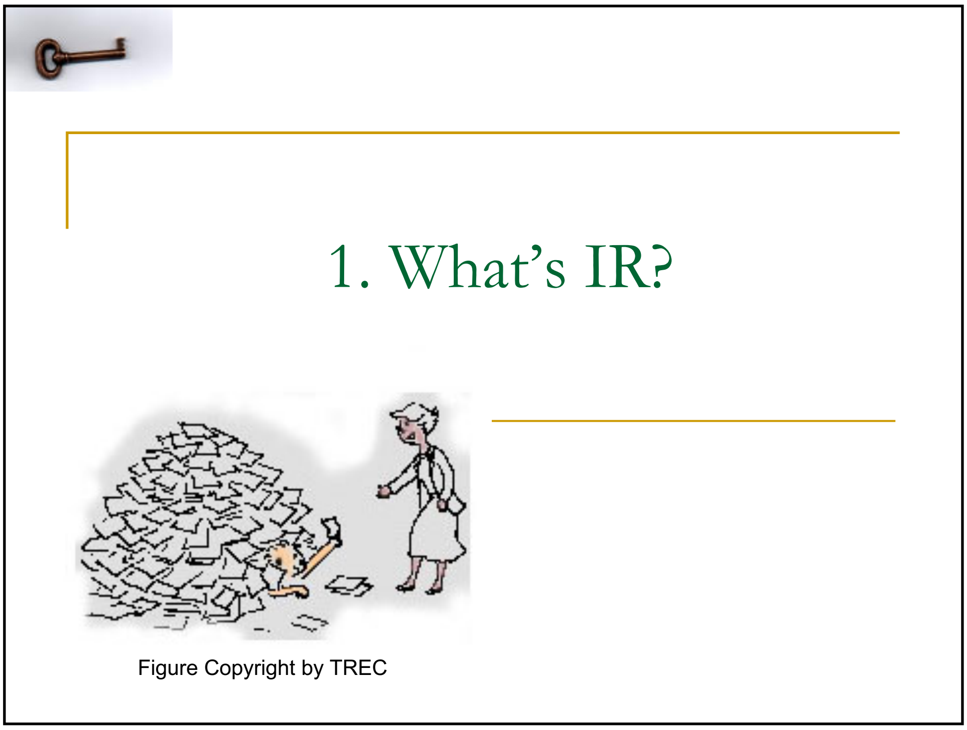
Technical requirements

- Answers to questions on any topic
  - Science, geography, popular culture …
- Accuracy: not only an answer, but a confident right answer
- Speed: within 3 second or less

- Advanced linguistic understanding
  - Parser complex sentences, recognize and understand jokes, metaphors, puns and riddles
- Real time analysis of questions
- Learn from mistakes
- Be prepared to handle the unexpected …

Techniques involved -- DeepQA

- A massively parallel probabilistic evidence-based architecture for answer questions
  - Non-database approach
  - Deep text analytics
    - NLP and statistical NLP
  - Formulating parallel hypotheses with confidence score
    - Voting, Question interpretation…
  - Search
  - Risk assessment
  - Hadoop and UIMA
- Difficulties/Problems in real application scenarios



1. What’s IR?

Figure Copyright by TREC

Review: What is Information Retrieval (IR)?

- Narrow-sense:
  - IR = Search Engine Technologies (i.e. IR =
    - Google, Yahoo, Bing, Ask, Baidu, Sogou, …
    - Library info search, enterprise search, in-site search, destop search…
    - PicSearch, Greplin, Blekko, SkyScanner, KooXoo, Qunar, …

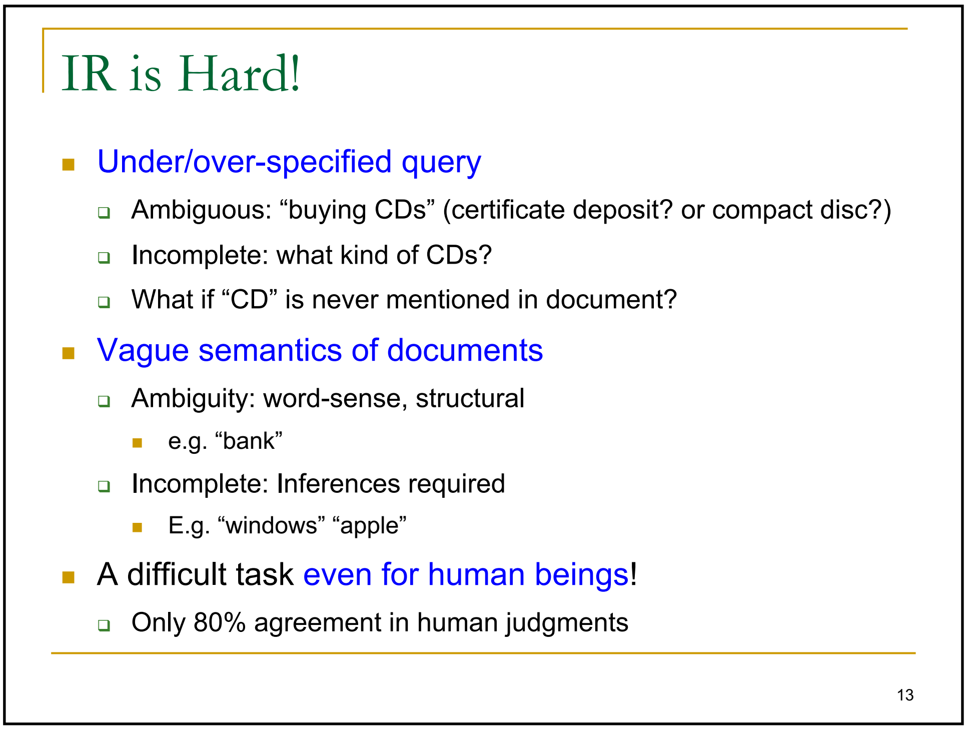
Review: What's IR? (cont.)

- Broad-sense: IR ~ **Information Management**
  - General problem: how to manage information?
  - How to **find** useful information? (retrieval & recommendation)
    - Beyond search engine:
    - e.g. in news feed, movie, travel, e-commerce, financial… scenarios
    - e.g. in social media platform, e.g. Twitter, Facebook, YouTube, WeChat, Weibo, Zhihu, ……
  - How to **organize** information? (classification & filtering)
    - e.g., automatically assign email to different folders
  - How to **discover** information (or even knowledge) from the data? (mining)
    - e.g., discover correlation of events

Review: What's IR? (cont.)

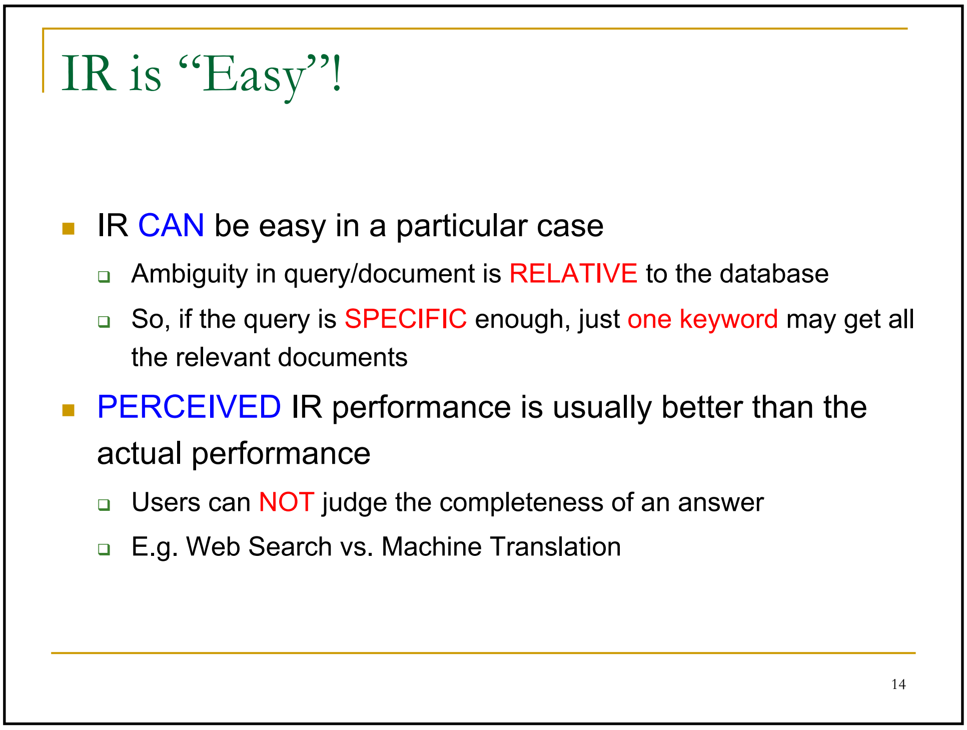
- Goal:
  - Find documents *relevant* to an information need from a large document set
- And now:
  - Beyond relevance
  - Multi-modal documents
  - Users' (implicit) information need
  - Heterogeneous environment

Figure Copyright by TREC



IR is Hard!

- Under/over-specified query
  - Ambiguous: “buying CDs” (certificate deposit? or compact disc?)
  - Incomplete: what kind of CDs?
  - What if “CD” is never mentioned in document?
- Vague semantics of documents
  - Ambiguity: word-sense, structural
    - e.g. “bank”
  - Incomplete: Inferences required
    - E.g. “windows” “apple”
- A difficult task even for human beings!
  - Only 80% agreement in human judgments



IR is “Easy”!

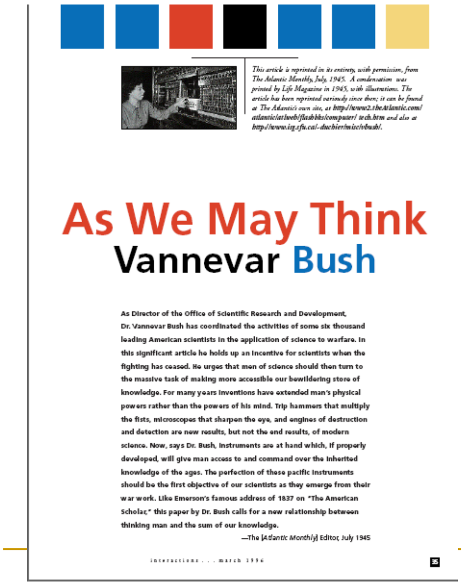
- IR CAN be easy in a particular case
  - Ambiguity in query/document is RELATIVE to the database
  - So, if the query is SPECIFIC enough, just one keyword may get all the relevant documents
- PERCEIVED IR performance is usually better than the actual performance
  - Users can NOT judge the completeness 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, 2009~2015) …
- Large-scale Evaluation & Applications (1990's~present)
  - TREC (D. Harman & E. Voorhees, NIST)， CLEF, NTCIR, …
- Web search (2000's ~ present)
  - Search engine companies, Boundary with related areas are disappearing
- Vertical Search, Knowledge, Social, User (2010's ~ present)

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

interactions . . . march 1996

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.‘

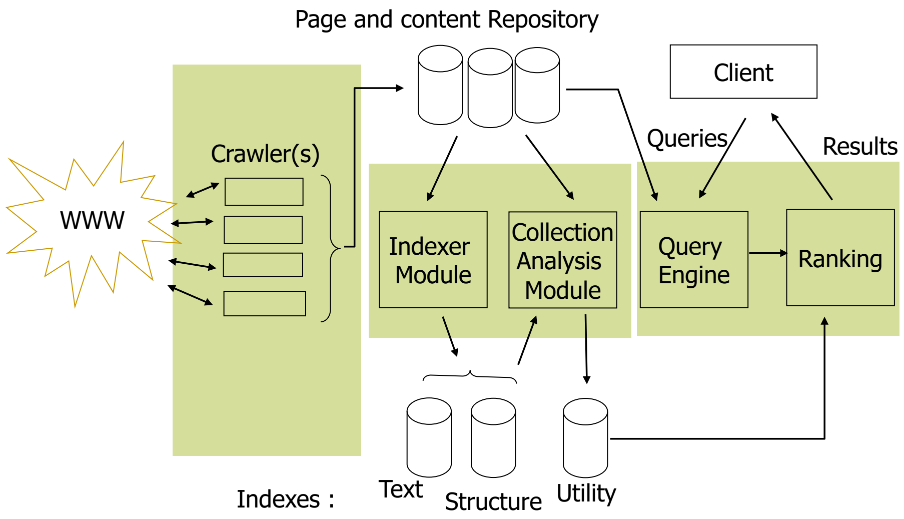
---- LUHN, H.P., 'The automatic creation of literature abstracts', *IBM Journal of Research and Development*, 2, 159-165 (1958).

◀

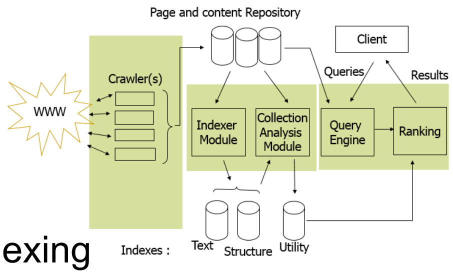
2. A general (Basic) IR procedure

Figure Copyright by TREC

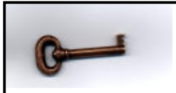
Example: search engine architecture



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?



3. Brief introduction to IR fundamentals – I

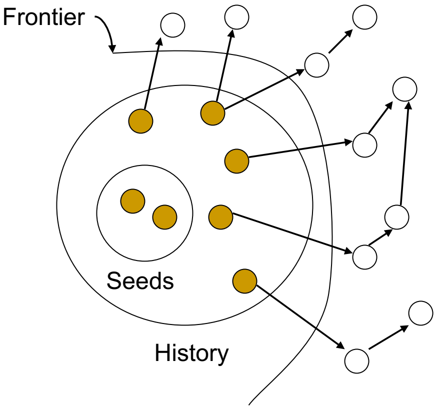
(Mainly Text IR;
Visual IR will be discussed by a specific lecture later)

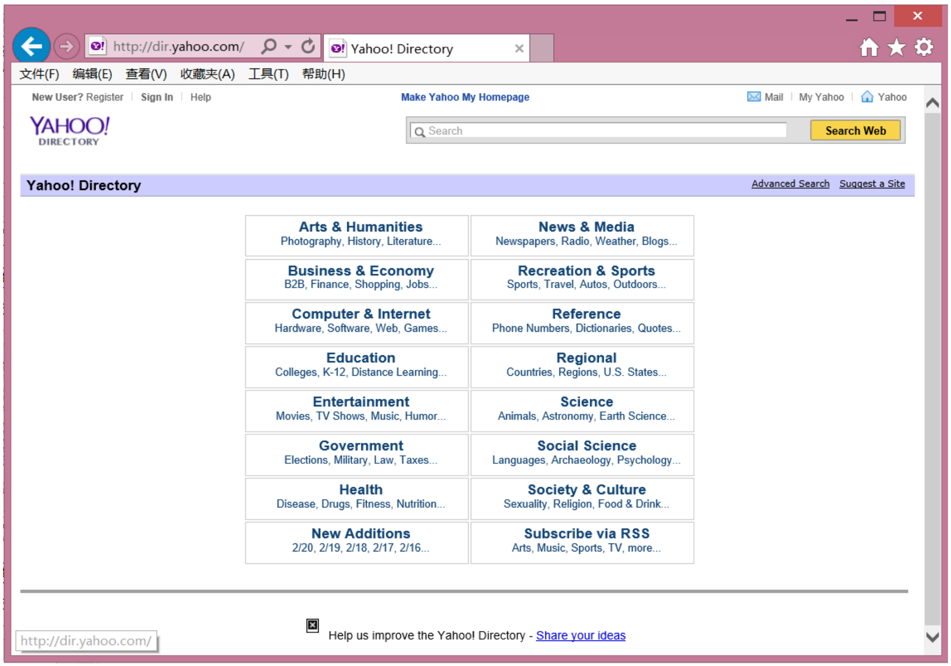
Outline

- 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

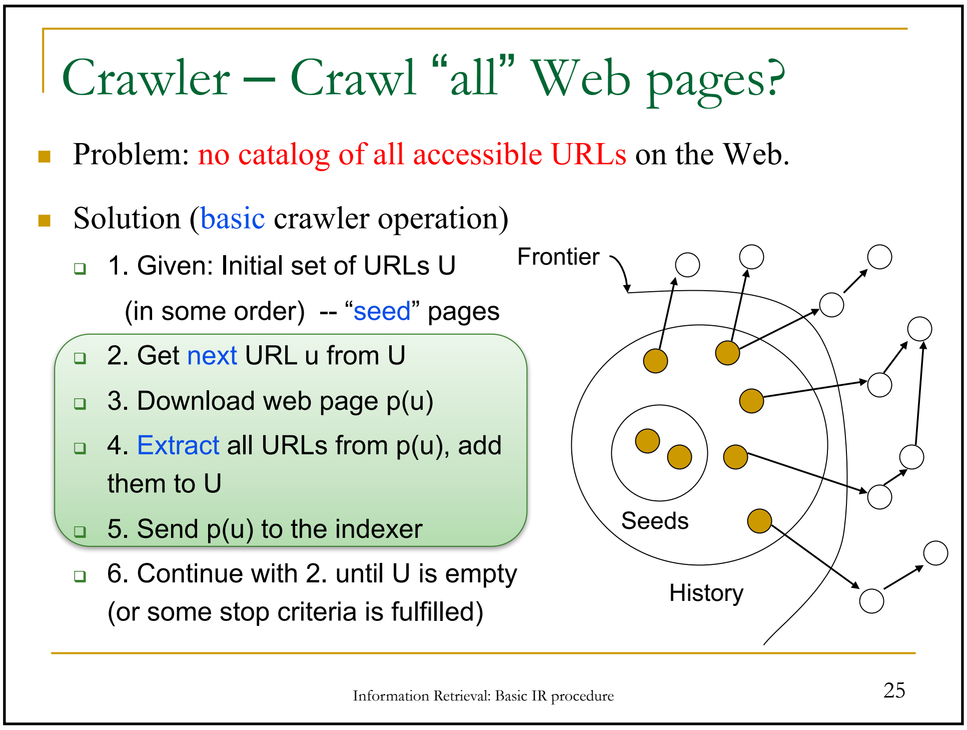




Yahoo! Directory

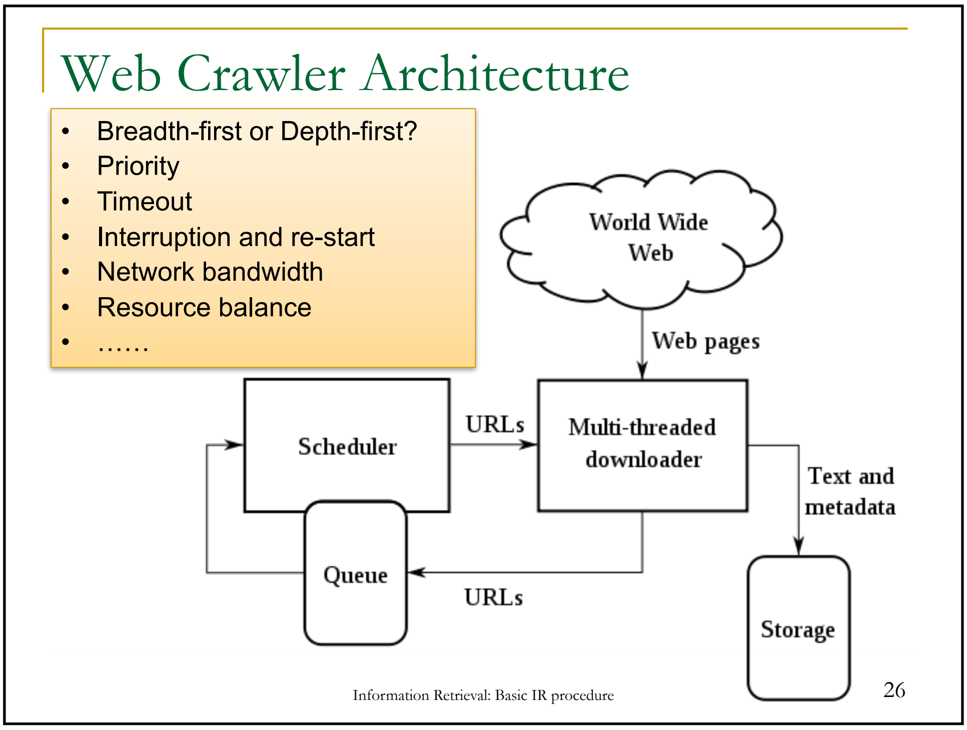
Arts & Humanities Photography, History, Literature...	**News & Media** Newspapers, Radio, Weather, Blogs...
Business & Economy B2B, Finance, Shopping, Jobs...	**Recreation & Sports** Sports, Travel, Autos, Outdoors...
Computer & Internet Hardware, Software, Web, Games...	**Reference** Phone Numbers, 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 2/20, 2/19, 2/18, 2/17, 2/16...	**Subscribe via RSS** Arts, Music, Sports, TV, more...

Help us improve the Yahoo! Directory - Share your ideas



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

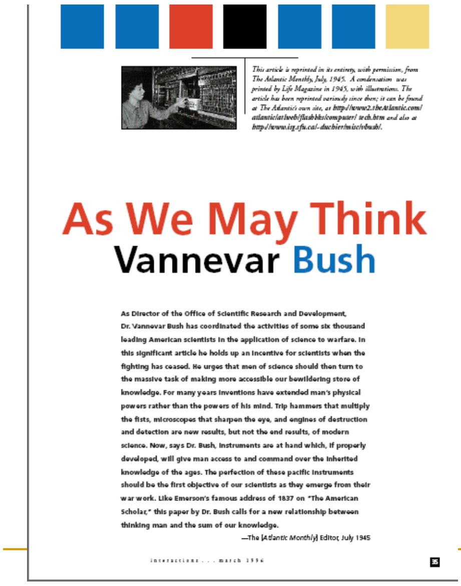
- Breadth-first or Depth-first?
- Priority
- Timeout
- Interruption and re-start
- Network bandwidth
- Resource balance
- ……

"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

interactions . . . march 1996

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).