Web Search

Introduction

The World Wide Web

- Developed by Tim Berners-Lee in 1990 at CERN to organize research documents available on the Internet.
- Combined idea of documents available by FTP with the idea of *hypertext* to link documents.
- Developed initial HTTP network protocol, URLs, HTML, and first "web server."

Web Pre-History

- Ted Nelson developed idea of hypertext in 1965.
- Doug Engelbart invented the mouse and built the first implementation of hypertext in the late 1960's at SRI.
- ARPANET was developed in the early 1970's.
- The basic technology was in place in the 1970's; but it took the PC revolution and widespread networking to inspire the web and make it practical.

Web Browser History

- Early browsers were developed in 1992 (Erwise, ViolaWWW).
- In 1993, Marc Andreessen and Eric Bina at UIUC NCSA developed the Mosaic browser and distributed it widely.
- Andreessen joined with James Clark (Stanford Prof. and Silicon Graphics founder) to form Mosaic Communications Inc. in 1994 (which became Netscape to avoid conflict with UIUC).
- Microsoft licensed the original Mosaic from UIUC and used it to build Internet Explorer in 1995.

Search Engine Early History

- By late 1980's many files were available by anonymous FTP.
- In 1990, Alan Emtage of McGill Univ. developed Archie (short for "archives")
  - Assembled lists of files available on many FTP servers.
  - Allowed regex search of these file names.
- In 1993, Veronica and Jughead were developed to search names of text files available through Gopher servers.

Web Search History

- In 1993, early web robots (spiders) were built to collect URL's:
  - Wanderer
  - ALIWEB (Archie-Like Index of the WEB)
  - WWW Worm (indexed URL's and titles for regex search)
- In 1994, Stanford grad students David Filo and Jerry Yang started manually collecting popular web sites into a topical hierarchy called Yahoo.

Web Search History (cont.)

- In early 1994, Brian Pinkerton developed WebCrawler as a class project at U Wash. (eventually became part of Excite and AOL).
- A few months later, Fuzzy Maudlin, a grad student at CMU developed Lycos. First to use a standard IR system as developed for the DARPA Tipster project. First to index a large set of pages.
- In late 1995, DEC developed Altavista. Used a large farm of Alpha machines to quickly process large numbers of queries. Supported boolean operators, phrases, and “reverse pointer” queries.

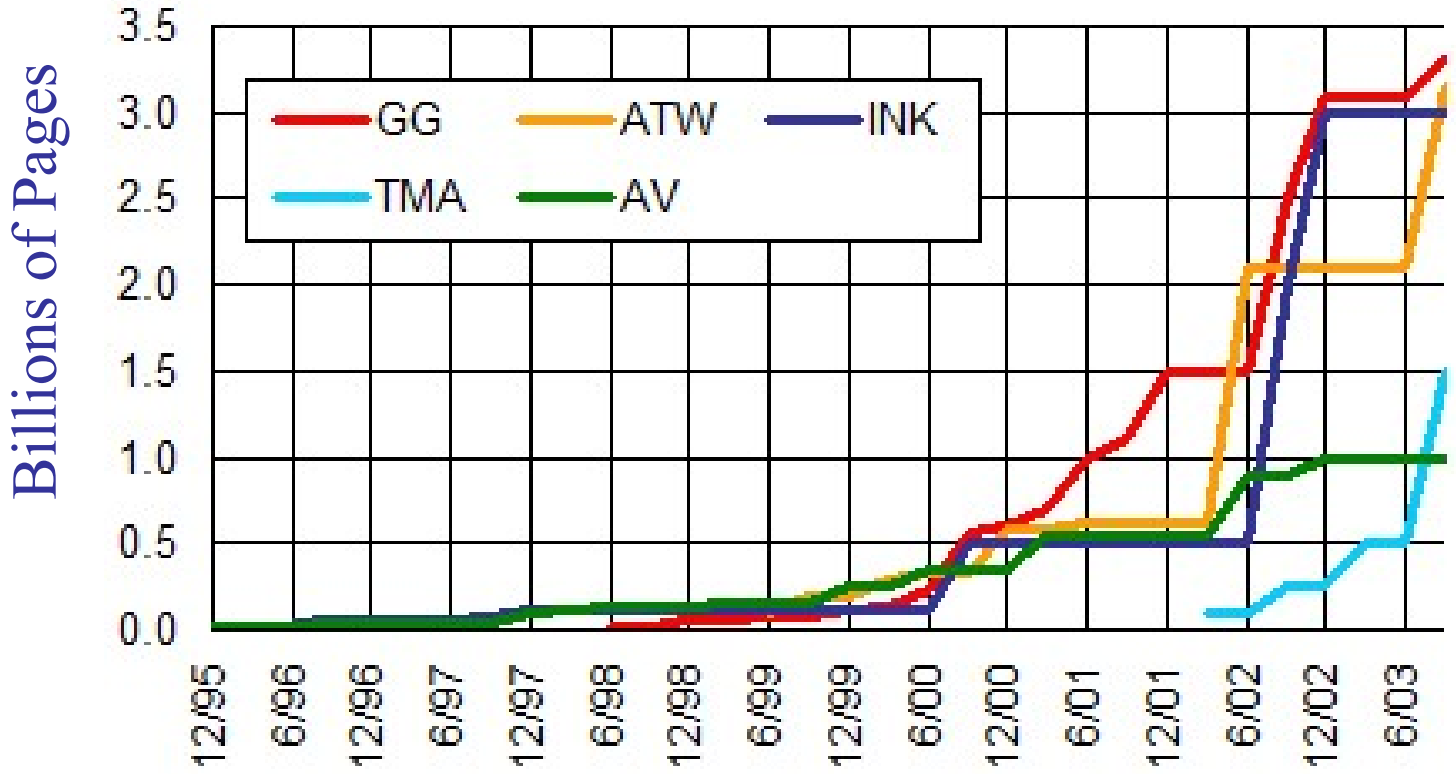
Web Search History (cont.)

- In 1998, Larry Page and Sergey Brin, Ph.D. students at Stanford, started Google. Main advance is use of *link analysis* to rank results partially based on authority.
- Microsoft lauched MSN Search in 1998 based on Inktomi (started from UC Berkeley in 1996), changed to Live Search in 2007, and Bing in 2009.

Web Challenges for IR

- Distributed Data: Documents spread over millions of different web servers.
- Volatile Data: Many documents change or disappear rapidly (e.g. dead links).
- Large Volume: Billions of separate documents.
- Unstructured and Redundant Data: No uniform structure, HTML errors, up to 30% (near) duplicate documents.
- Quality of Data: No editorial control, false information, poor quality writing, typos, etc.
- Heterogeneous Data: Multiple media types (images, video, VRML), languages, character sets, etc.

Growth of Web Pages Indexed



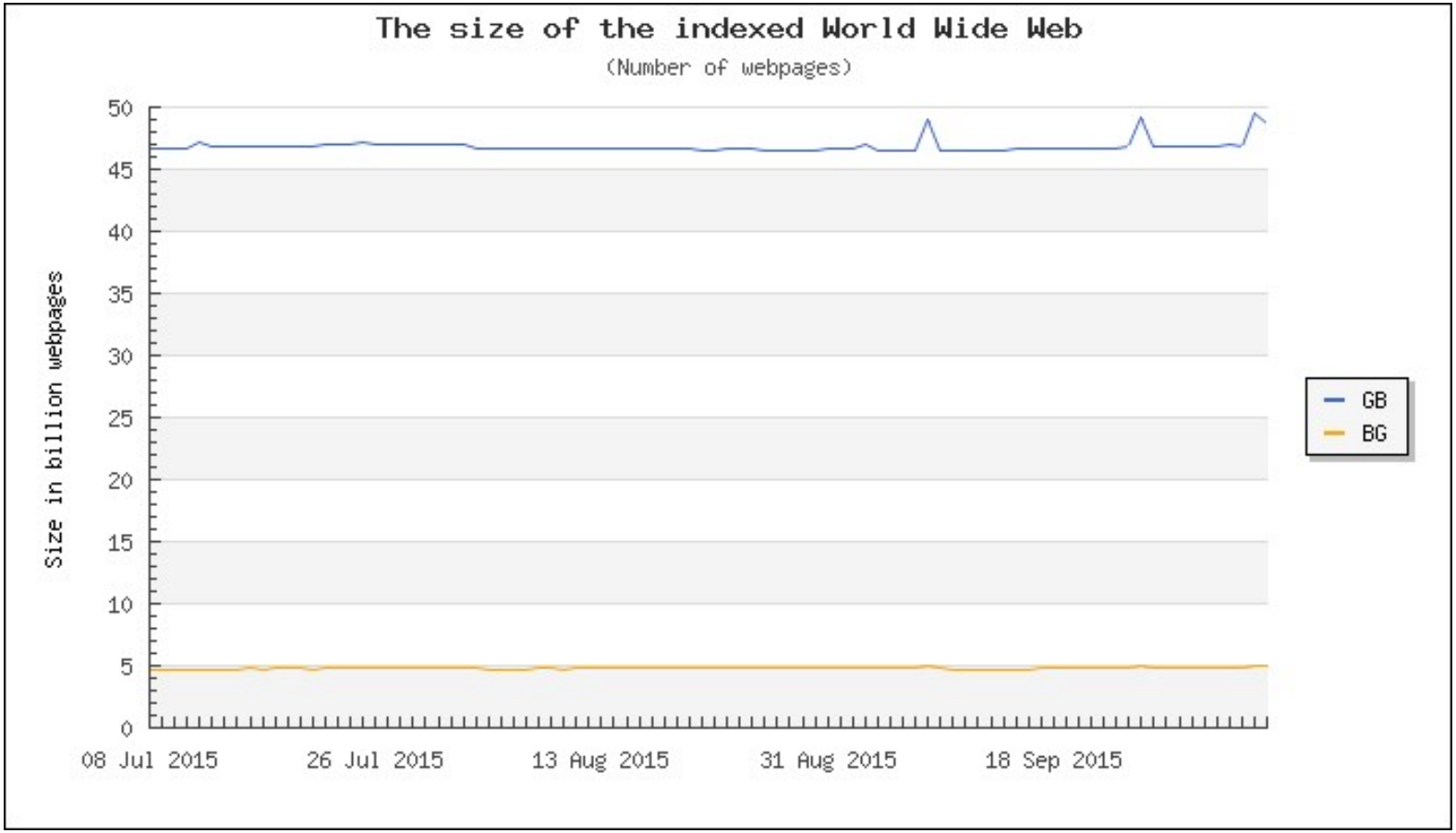
SearchEngineWatch

Google
Inktomi
AllTheWeb

Teoma
Altavista

Assuming 20KB per page,
1 billion pages is about 20 terabytes of data.

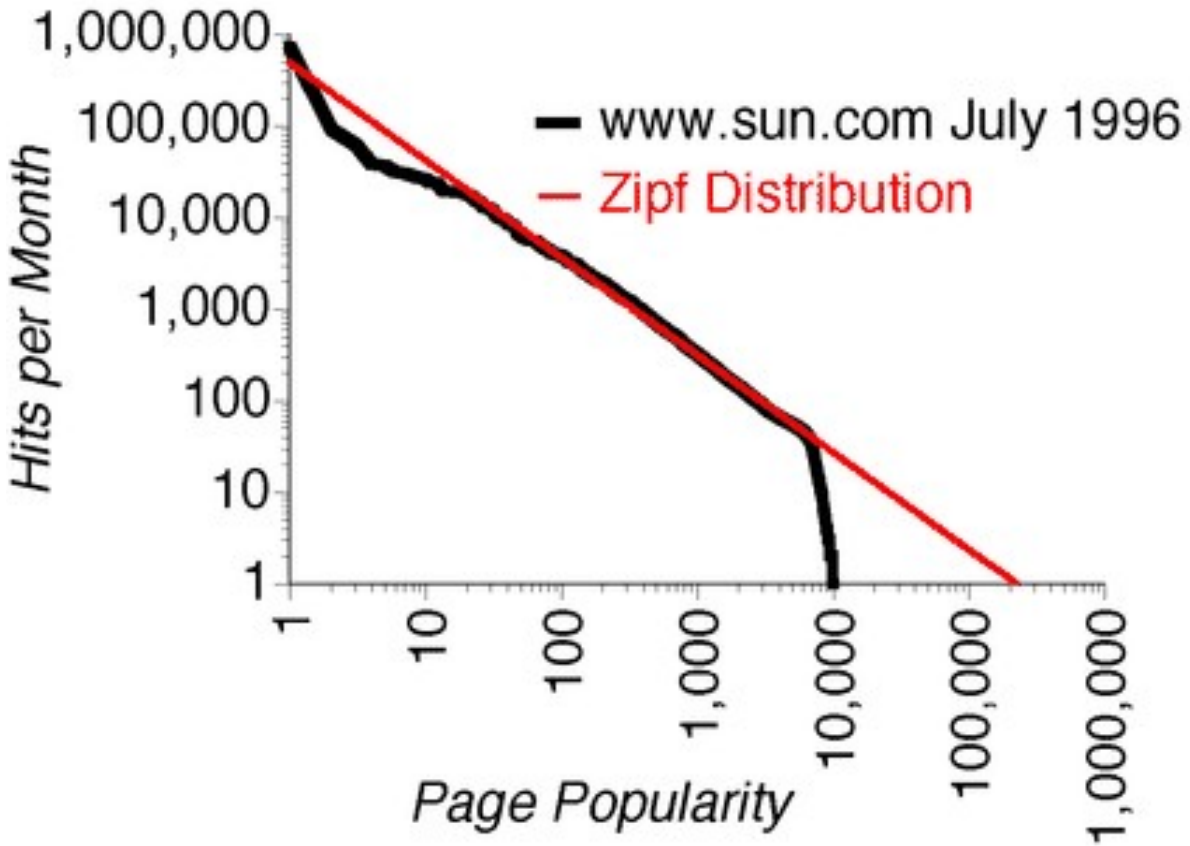
Current Size of the Web



Zipf's Law on the Web

- Number of in-links/out-links to/from a page has a Zipfian distribution.
- Length of web pages has a Zipfian distribution.
- Number of hits to a web page has a Zipfian distribution.

Zipfs Law and Web Page Popularity



“Small World” (Scale-Free) Graphs

- Social networks and six degrees of separation.
  - Stanley Milgram Experiment
- Power law distribution of in and out degrees.
- Distinct from purely random graphs.
- “Rich get richer” generation of graphs (preferential attachment).
- Kevin Bacon game.
  - Oracle of Bacon
- Erdos number.
- Networks in biochemistry, roads, telecommunications, Internet, etc are “small word”

Manual Hierarchical Web Taxonomies

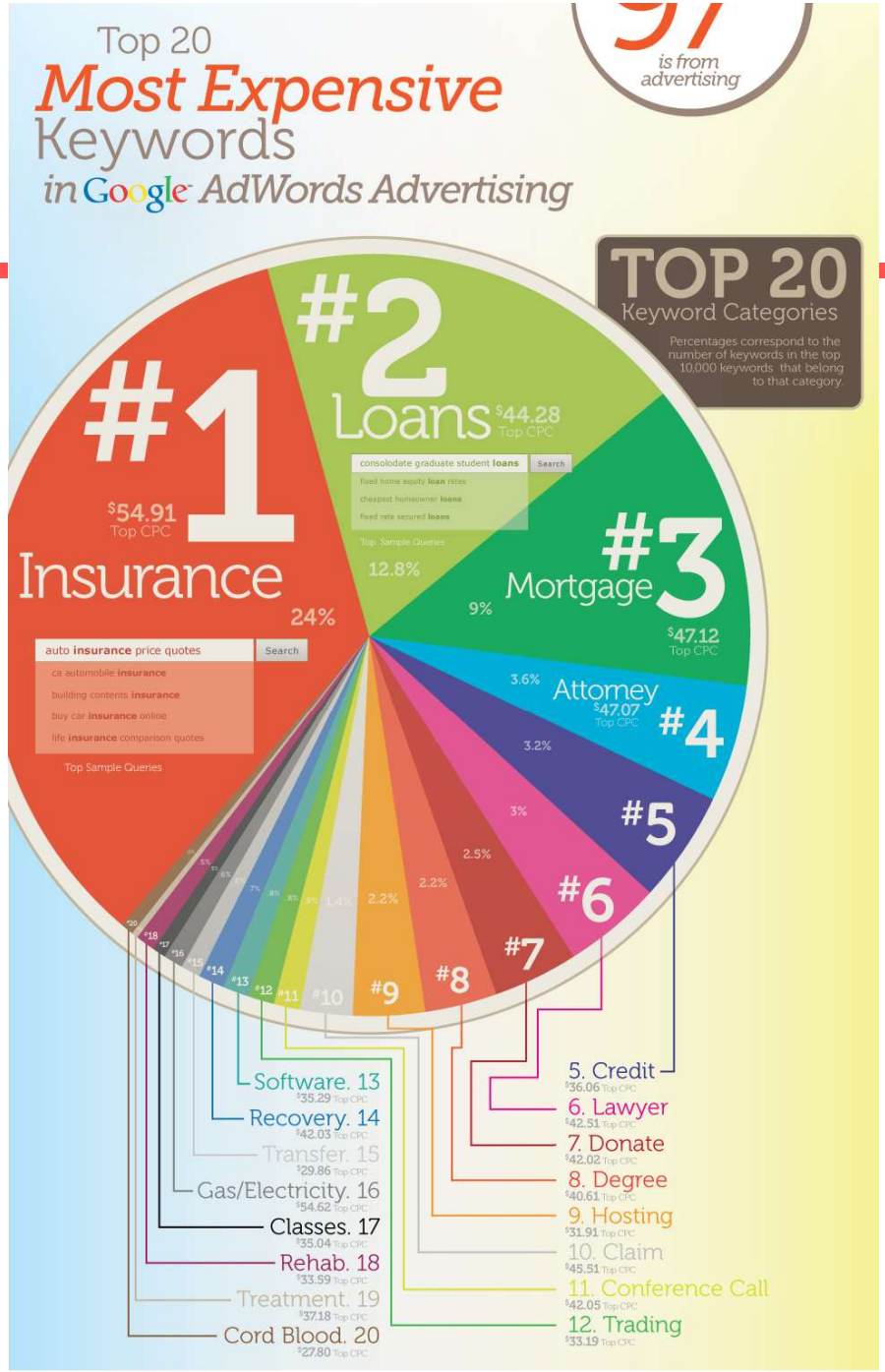
- Yahoo approach of using human editors to assemble a large hierarchically structured directory of web pages (closed in 2014).
- Open Directory Project is a similar approach based on the distributed labor of volunteer editors ("net-citizens provide the collective brain"). Used by most other search engines. Started by Netscape.
  - http://www.dmoz.org/

Business Models for Web Search

- Advertisers pay for banner ads on the site that do not depend on a user's query.
  - CPM: Cost Per Mille (thousand impressions). Pay for each ad display.
  - CPC: Cost Per Click. Pay only when user clicks on ad.
  - CTR: Click Through Rate. Fraction of ad impressions that result in clicks throughs. CPC = CPM / (CTR * 1000)
  - CPA: Cost Per Action (Acquisition). Pay only when user actually makes a purchase on target site.
- Advertisers bid for "keywords". Ads for highest bidders displayed when user query contains a purchased keyword.
  - PPC: Pay Per Click. CPC for bid word ads (e.g. Google AdWords).

History of Business Models

- Initially, banner ads paid thru CPM were the norm.
- GoTo Inc. formed in 1997 and originates and patents bidding and PPC business model.
- Google introduces AdWords in fall 2000.
- GoTo renamed Overture in Oct. 2001.
- Overture sues Google for use of PPC in Apr. 2002.
- Overture acquired by Yahoo in Oct. 2003.
- Google settles with Overture/Yahoo for 2.7 million shares of Class A common stock in Aug. 2004.


Top 20
Most Expensive
Keywords
in Google AdWords Advertising
97 is from advertising
TOP 20
Keyword Categories
Percentages correspond to the number of keywords in the top 10,000 keywords that belong to that category.
#1
Insurance
$54.91 Top CPC
24%
auto insurance price quotes
Search
ca automobile insurance
building contents insurance
buy car insurance online
life insurance comparison quotes
Top Sample Queries
#2
Loans
$44.28 Top CPC
consolodate graduate student loans
Search
fixed home equity loan rates
cheapest homeowner loans
fixed rate secured loans
Top Sample Queries
12.8%
#3
Mortgage
$47.12 Top CPC
9%
#4
Attorney
$47.07 Top CPC
3.6%
#5
3.2%
#6
3%
#7
2.5%
#8
2.2%
#9
2.2%
#10
1.4%
#11
#12
#13
#14
#15
#16
#17
#18
#20
5. Credit
$36.06 Top CPC
6. Lawyer
$42.51 Top CPC
7. Donate
$42.02 Top CPC
8. Degree
$40.61 Top CPC
9. Hosting
$31.91 Top CPC
10. Claim
$45.51 Top CPC
11. Conference Call
$42.05 Top CPC
12. Trading
$33.19 Top CPC
Software. 13
$35.29 Top CPC
Recovery. 14
$42.03 Top CPC
Transfer. 15
$29.86 Top CPC
Gas/Electricity. 16
$54.62 Top CPC
Classes. 17
$35.04 Top CPC
Rehab. 18
$33.59 Top CPC
Treatment. 19
$37.18 Top CPC
Cord Blood. 20
$27.80 Top CPC
presented by WordStream
developed by nowsourcing.com

Affiliates Programs

- If you have a website, you can generate income by becoming an *affiliate* by agreeing to post ads relevant to the topic of your site.
- If users click on your impression of an ad, you get some percentage of the CPC or PPC income that is generated.
- Google introduces AdSense affiliates program in 2003.

Automatic Document Classification

- Manual classification into a given hierarchy is labor intensive, subjective, and error-prone.
- Text categorization methods provide a way to automatically classify documents.
- Best methods based on training a *machine learning* (*pattern recognition*) system on a labeled set of examples (*supervised learning*).
- Text categorization is a topic we will discuss later in the course.

Automatic Document Hierarchies

- Manual hierarchy development is labor intensive, subjective, and error-prone.
- It would nice to automatically construct a meaningful hierarchical taxonomy from a corpus of documents.
- This is possible with hierarchical text clustering (unsupervised learning).
  - Hierarchical Agglomerative Clustering (HAC)
- Text clustering is a another topic we will discuss later in the course.

Web Search Using IR