COSC 431
Information Retrieval

Andrew Trotman

Instructors and Support People

- Dr. Andrew Trotman
  - Office: 123A, Owheo
  - Email: andrew@cs.otago.ac.nz
  - No office hours, just “drop in”
- Department Kaiāwhina (Māori and Pacific Support):
  - Steven Mills (Office: 245, Owheo)
- Department Disability Support:
  - Contact the main office.

Learning Outcomes

- This paper will enable students to:
  - Implement a range of data structures and algorithms using the C programming language
  - Classify familiar algorithms in terms of efficiency and present big-O calculations in a clear and logical manner
  - Use proofs to support efficiency and effectiveness calculations
  - Critically evaluate the factors that should be taken into account when deciding on the data structures and/or algorithms to use for a given purpose
  - Demonstrate understanding of a variety of algorithm designs for optimisation

Academic Integrity and Academic Misconduct

- Academic integrity means being honest in your studying and assessments. It is the basis for ethical decision-making and behaviour in an academic context. Academic integrity is informed by the values of honesty, trust, responsibility, fairness, respect and courage. Students are expected to be aware of, and act in accordance with, the University’s Academic Integrity Policy.

- Academic Misconduct, such as plagiarism or cheating, is a breach of Academic Integrity and is taken very seriously by the University. Types of misconduct include plagiarism, copying, unauthorised collaboration, taking unauthorised material into a test or exam, impersonation, and assisting someone else’s misconduct. A more extensive list of the types of academic misconduct and associated processes and penalties is available in the University’s Student Academic Misconduct Procedures.

- It is your responsibility to be aware of and use acceptable academic practices when completing your assessments. To access the information in the Academic Integrity Policy and learn more, please visit the University’s Academic Integrity website at www.otago.ac.nz/study/academicintegrity or ask at the Student Learning Centre or Library. If you have any questions, ask your lecturer.

- Academic Integrity Policy
- Student Academic Misconduct Procedures

What Is Information Retrieval?

- Tradition spanning over 60 years
- An Application of Computer Science
  - NLP, DB, OS, AI, HCI, Graphics, Music, etc.
- History
  - Library Science / Online systems
  - CD-ROMs
  - Internet (Archie, WAIS, Web)
  - Modern OS (Index Server / Desktop Search)
- IR is not DB
  - What is relevance?
  - Application of AI

Course Details

- 20 point paper
  - So 200 hours in total
- 12 Lectures
  - 1 per week, Owheo G34
  - 11am-1pm Wednesdays
- HOW MANY HOURS IN YOUR TIME?
- Assessment
  - 1 programming assignment (20%)
    - Write your own search engine
    - Due April 3
  - 1 essay assignment (20%)
    - With presentation
    - Due 22 May
  - Exam (60%)

Course Details

- Resources
  - Journals and conferences:
    - JASIST / IP&M / TOIS / IR / SIGIR / CIKM / SPIRE
  - Digital libraries:
    - CiteSeer / Google Scholar / ScienceDirect / etc.
  - Evaluation Forums
    - TREC / CLEF / NTCIR / FIRE / etc.
  - Local resources
    - COSC 431 Website, http://www.cs.otago.ac.nz/cosc431/
    - Lecture notes
    - Science and Central Libraries

Course Outline

1. Course outline and Outstanding Issues in IR
2. Searching
3. Efficiency and Ranking
4. Evaluation
5. Phrase and Structured Search
6. Term Conflation
7. Relevance Feedback
8. Distributed Information Retrieval
9. Very Fast Search
10. Compression
11. XML-IR and Link Discovery
12. COSC431 Presentations / Revision

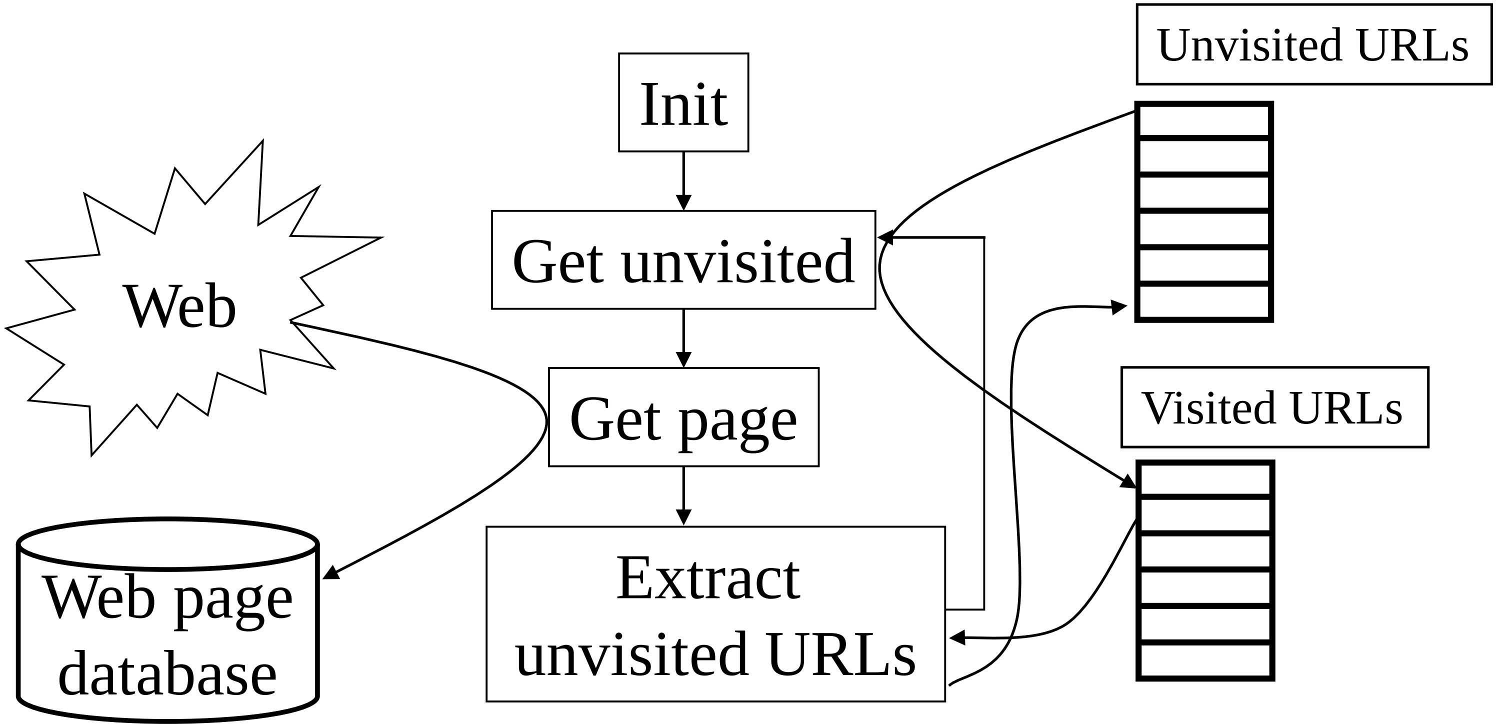
Search Engines: The Problems....

Web Crawling

- The Crawler, agent, robot, spider, wanderer, walker, etc.
  - Move around the web looking for pages not in the collection
  - ***CRAWLERS DO NOT MOVE,*** They are trawlers not crawlers
  - Run on local machines and download pages
  - This is just like you loading a page across the Internet

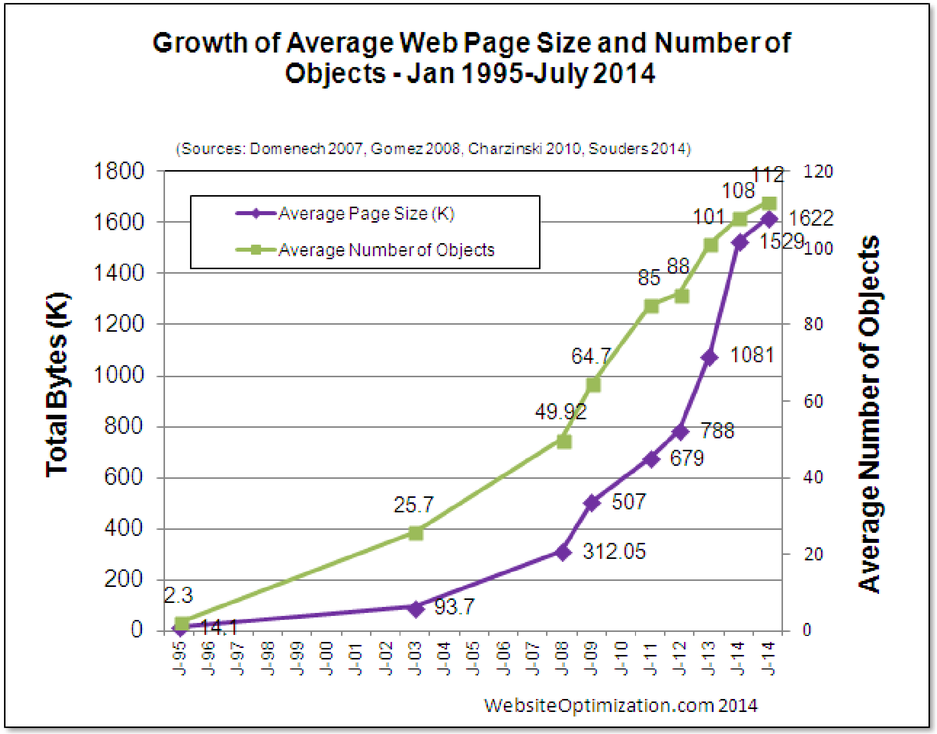
- Guidelines for robot behaviours
  - Robots.txt: placed at the server root
  - Indicates how a robot is allowed to trawl the site
  - But can be ignored,
    - But there might be consequences to ignoring it
      - Such as the site administrator blocking your IP address

Web Crawling Algorithm



Number of Web Pages

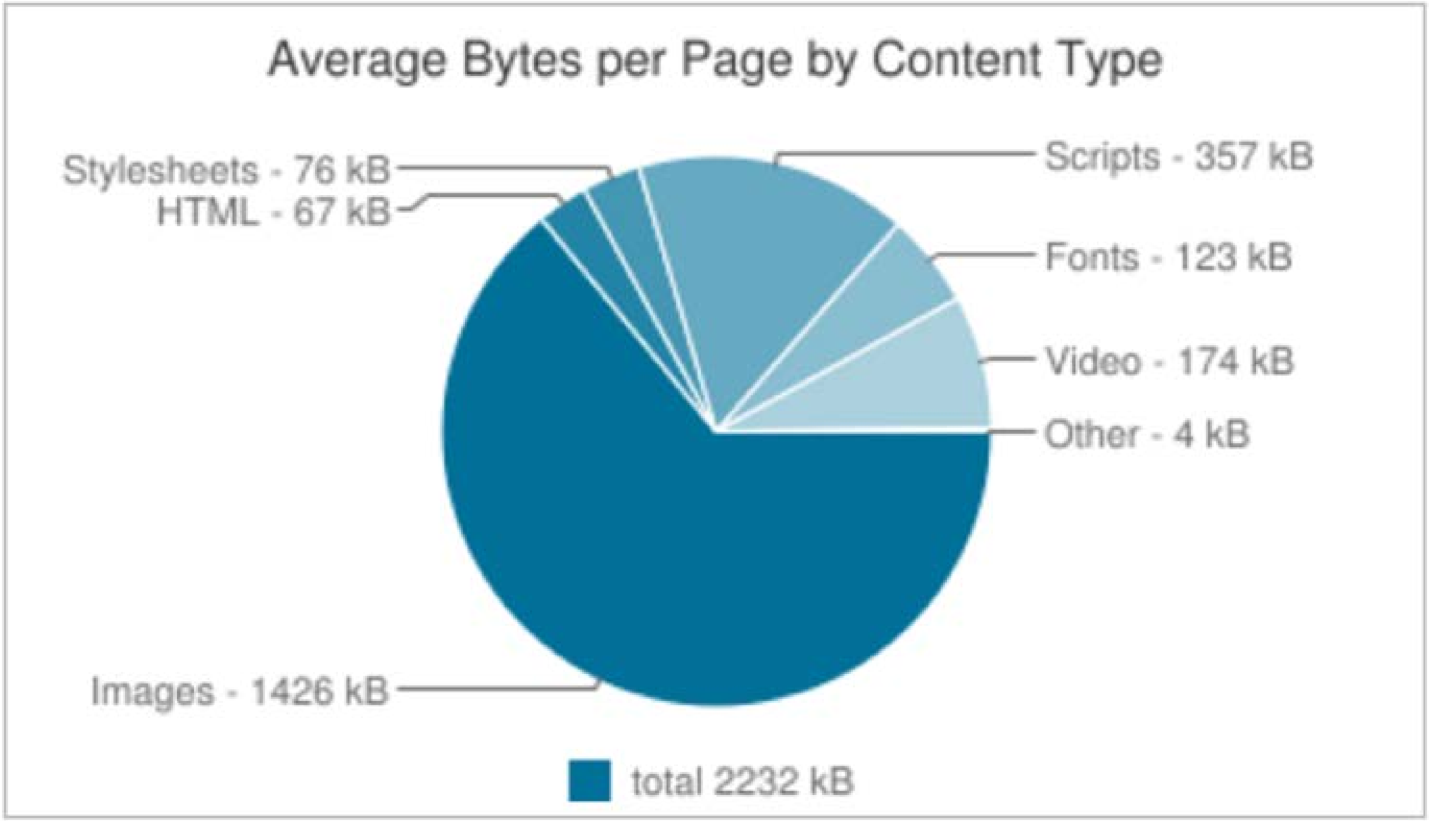
- Cuil:127 Billion pages (Mar-10)
- Google indexes between 40 and 50 Billion pages
- Google “knows about” 130 Trillion



http://www.websiteoptimization.com/speed/tweak/average-web-page/

The “Average” Page Size

Average Page Size – 2016



https://www.keycdn.com/support/the-growth-of-web-page-size/

How Much is That?

- How much?
  - 130 Trillion Pages
  - 2232 KB / page
  - Total Data:

- Problems:
  - How do they store that?
  - How do they crawl it?
  - How long does it take to download it?

Indexing

- Parse and build an “Inverted File” index
  - Similar to indexes seen in databases (in COSC344)
- Problems:
  - Index does not fit in memory
  - The index is bigger than OS limits (volume size, file size, etc.)
  - What’s a word (accents, specials, spellings, Americanisms)
  - How do you (quickly) search an index that is so large?

Searching

- Searching
  - Parsing Query
    - No real query language, (but some have Boolean (+/-))
    - 2.5 words per query (on average)

- Problems:
  - Phrase search
  - Fielded (Advanced) search
  - Misspelled words

Results / More Like This

- Results
  - Translate HTML into standard presentation format
    - Snippets, etc.
  - Relevance Ranking
    - PageRank / Newsgroups / Images

- More Like This
  - Clustering,
  - Document Similarity

Scale

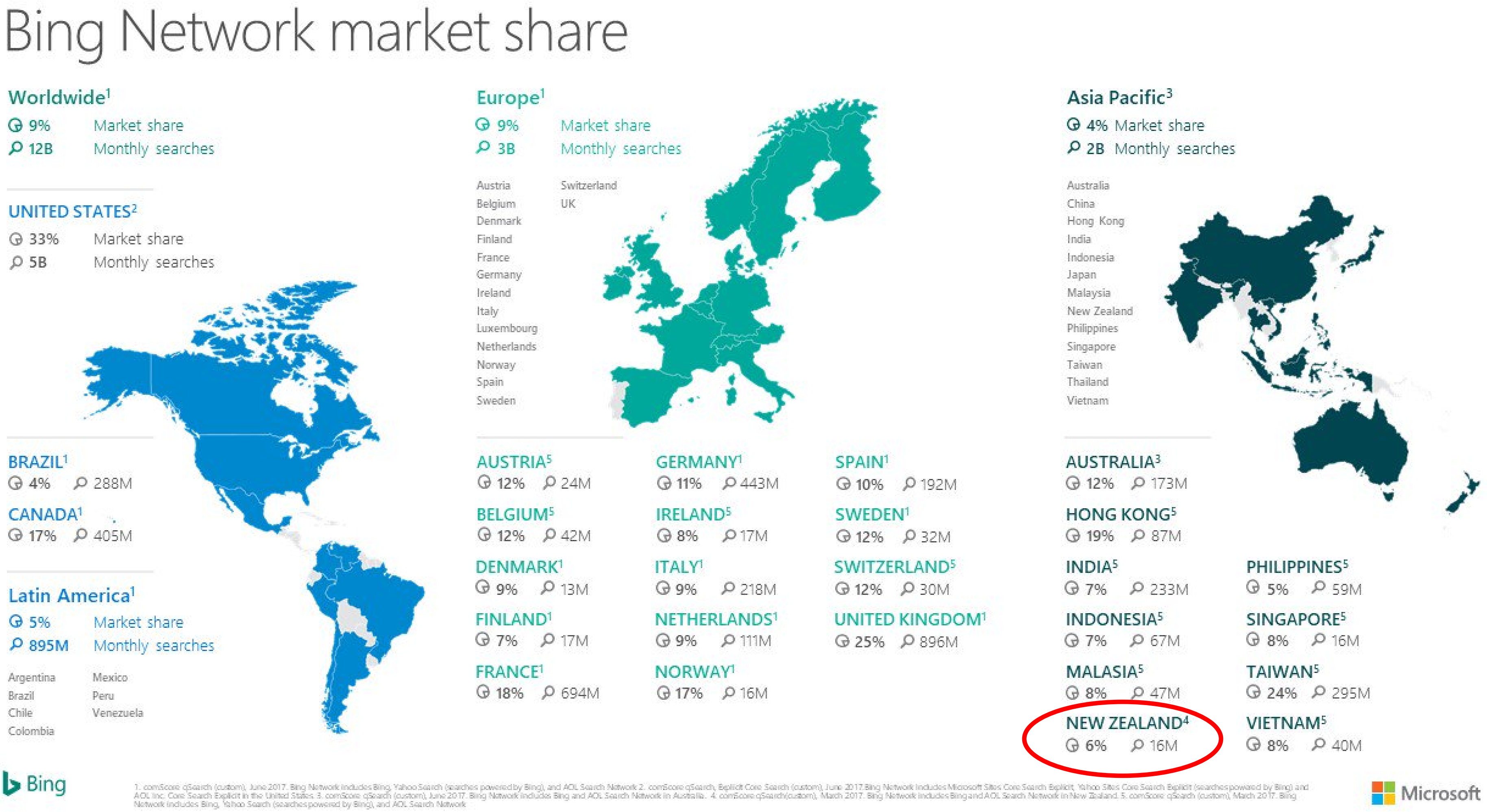
- Problems:
  - – Google Core Search (excludes YouTube, etc.) gets ~10 Billion queries per month from Desktop Search. That's about 4,000 queries per second.
  - – Remember – approximately 60% of traffic is from mobile.
  - – 40% of people search only on a smartphone

Rankings - Comscore, Inc. (Units are Millions)

#	Core Search Entity	Dec-19	Jan-20	Percent Change
1	Google Sites Core Search	10,826	11,914	10%
2	Microsoft Sites Core Search	4,426	4,825	9%
3	Verizon Media	1,992	2,159	8%
4	Ask Network Core Search	145	155	7%

See: https://www.comscore.com/Insights/Rankings

Bing’s Scale (17 Aug 2017)



Bing Network market share

Worldwide[1]
9% Market share
12B Monthly searches

UNITED STATES[2]
33% Market share
5B Monthly searches

	Market share	Monthly searches
BRAZIL[1]	4%	288M
CANADA[1]	17%	405M

Latin America[1]
5% Market share
895M Monthly searches

Argentina
Brazil
Chile
Colombia
Mexico
Peru
Venezuela

Europe[1]
9% Market share
3B Monthly searches

Austria
Belgium
Denmark
Finland
France
Germany
Ireland
Italy
Luxembourg
Netherlands
Norway
Spain
Sweden
Switzerland
UK

	Market share	Monthly searches
AUSTRIA[5]	12%	24M
BELGIUM[5]	12%	42M
DENMARK[1]	9%	13M
FINLAND[1]	7%	17M
FRANCE[1]	18%	694M
GERMANY[1]	11%	443M
IRELAND[5]	8%	17M
ITALY[1]	9%	218M
NETHERLANDS[1]	9%	111M
NORWAY[1]	17%	16M
SPAIN[1]	10%	192M
SWEDEN[1]	12%	32M
SWITZERLAND[5]	12%	30M
UNITED KINGDOM[1]	25%	896M

Asia Pacific[3]
4% Market share
2B Monthly searches

Australia
China
Hong Kong
India
Indonesia
Japan
Malaysia
New Zealand
Philippines
Singapore
Taiwan
Thailand
Vietnam

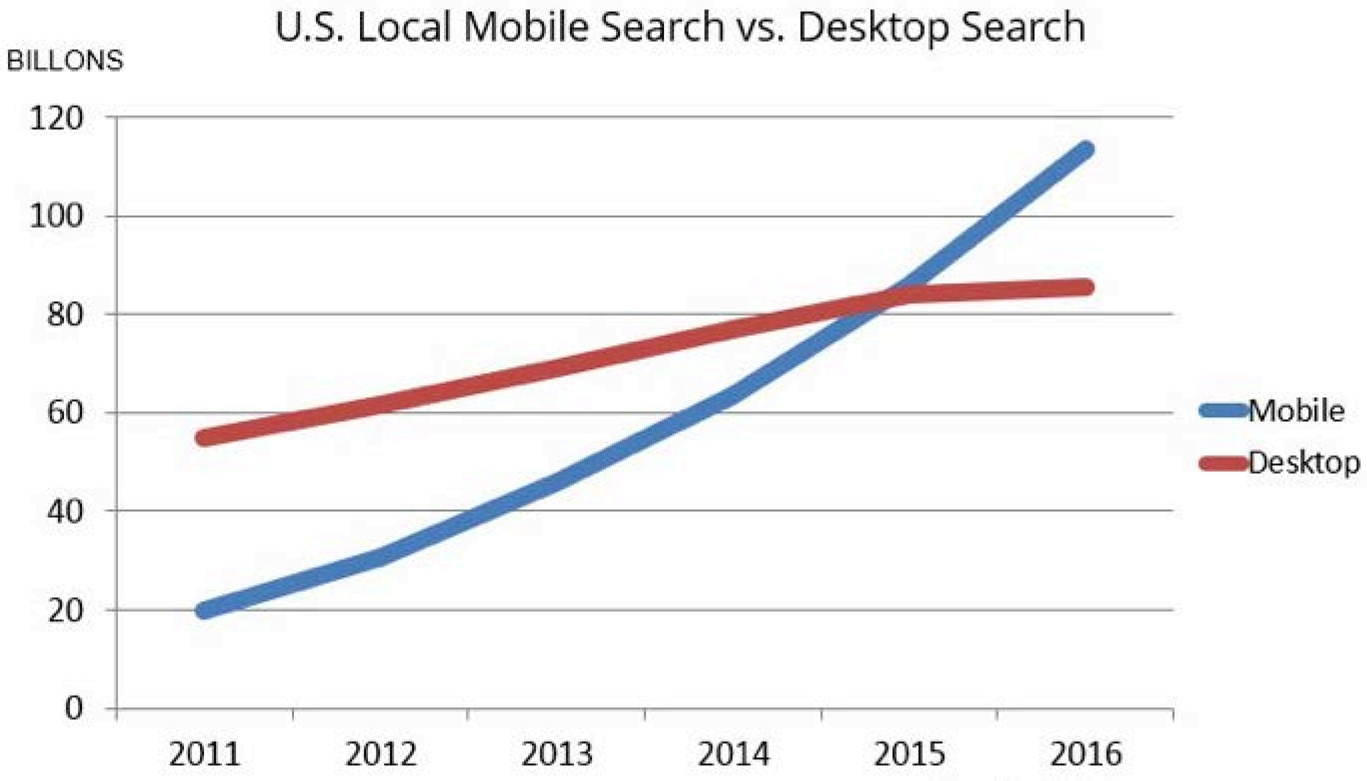
	Market share	Monthly searches
AUSTRALIA[3]	12%	173M
HONG KONG[5]	19%	87M
INDIA[5]	7%	233M
INDONESIA[5]	7%	67M
MALASIA[5]	8%	47M
NEW ZEALAND[4]	6%	16M
PHILIPPINES[5]	5%	59M
SINGAPORE[5]	8%	16M
TAIWAN[5]	24%	295M
VIETNAM[5]	8%	40M

Bing

[1] comScore qSearch (custom), June 2017. Bing Network includes Bing, Yahoo Search (searches powered by Bing), and AOL Search Network 2. comScore qSearch, Explicit Core Search (custom), June 2017.Bing Network includes Microsoft Sites Core Search Explicit, Yahoo Sites Core Search Explicit (searches powered by Bing) and AOL Inc. Core Search Explicit in the United States. 3. comScore qSearch (custom), June 2017. Bing Network includes Bing and AOL Search Network in Australia. 4. comScore qSearch(custom), March 2017. Bing Network includes Bing and AOL Search Network in New Zealand. 5. comScore qSearch (custom), March 2017. Bing Network includes Bing, Yahoo Search (searches powered by Bing), and AOL Search Network

Microsoft

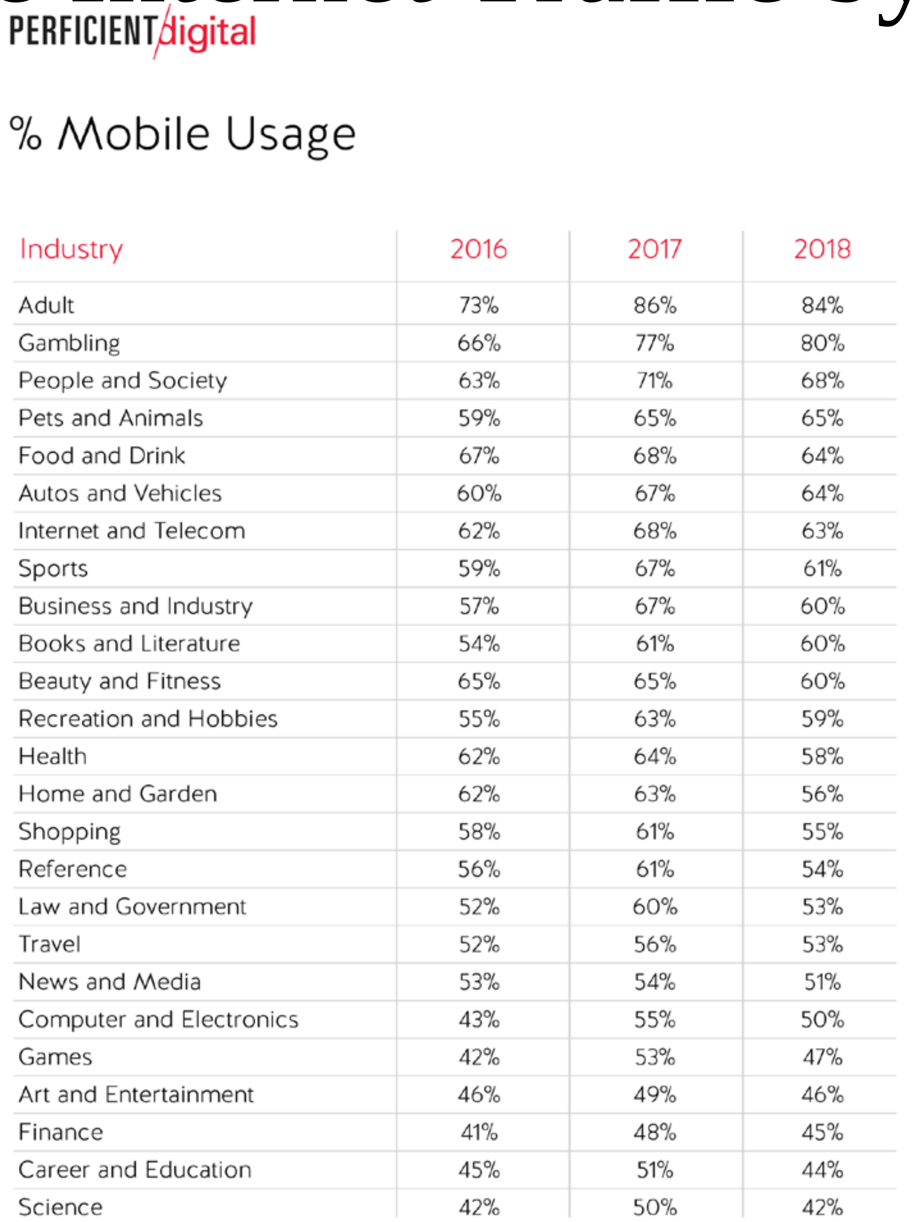
Mobile vs Desktop Search



https://www.smartinsights.com/search-engine-marketing/search-engine-statistics/attachment/us-mobile-search-vs-desktop-search-2017/

https://www.perficientdigital.com/insights/our-research/mobile-vs-desktop-usage-study

Mobile Internet Traffic by Category



PERFICIENT digital

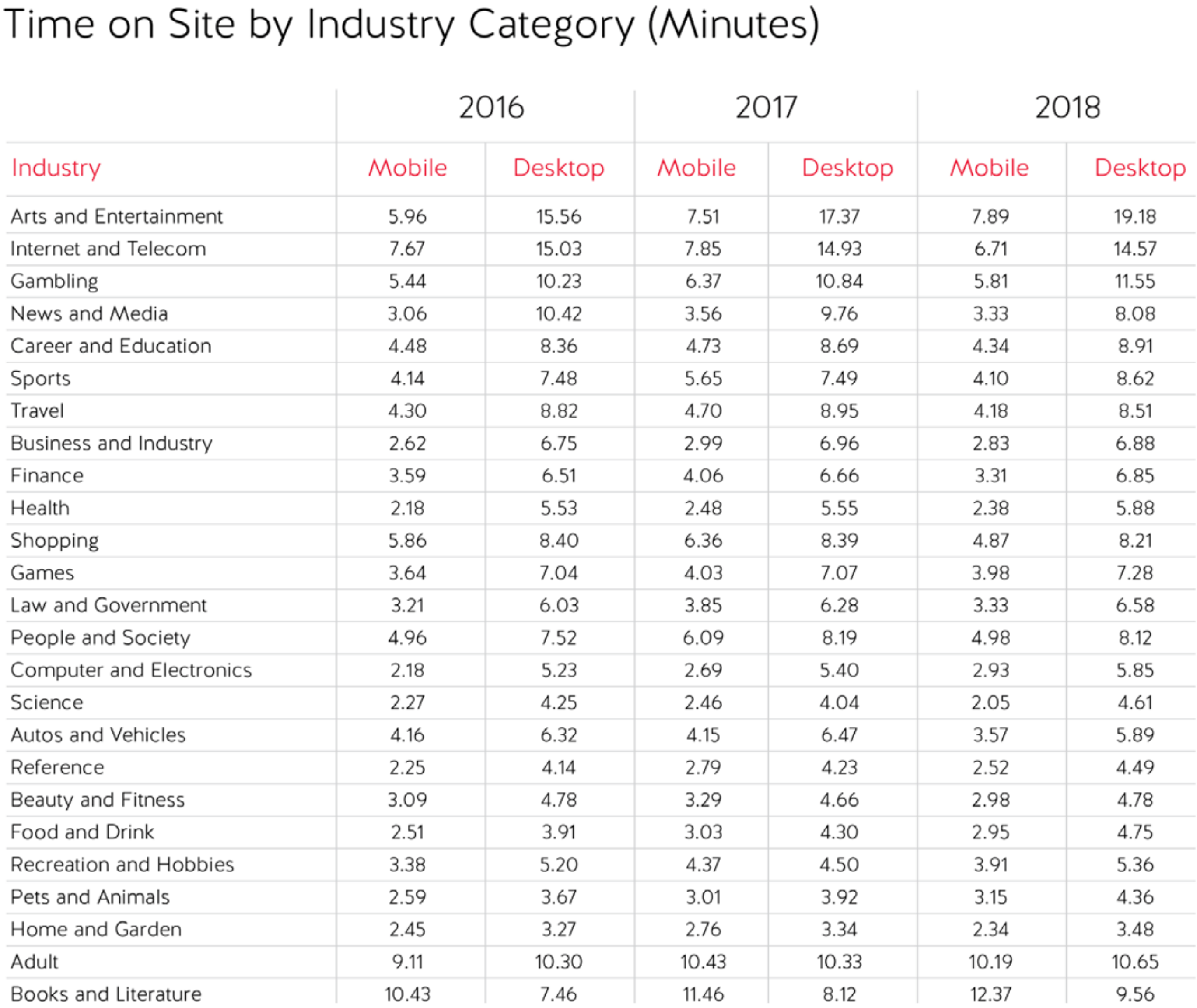
% Mobile Usage

Industry	2016	2017	2018
Adult	73%	86%	84%
Gambling	66%	77%	80%
People and Society	63%	71%	68%
Pets and Animals	59%	65%	65%
Food and Drink	67%	68%	64%
Autos and Vehicles	60%	67%	64%
Internet and Telecom	62%	68%	63%
Sports	59%	67%	61%
Business and Industry	57%	67%	60%
Books and Literature	54%	61%	60%
Beauty and Fitness	65%	65%	60%
Recreation and Hobbies	55%	63%	59%
Health	62%	64%	58%
Home and Garden	62%	63%	56%
Shopping	58%	61%	55%
Reference	56%	61%	54%
Law and Government	52%	60%	53%
Travel	52%	56%	53%
News and Media	53%	54%	51%
Computer and Electronics	43%	55%	50%
Games	42%	53%	47%
Art and Entertainment	46%	49%	46%
Finance	41%	48%	45%
Career and Education	45%	51%	44%
Science	42%	50%	42%

https://www.perficientdigital.com/insights/our-research/mobile-vs-desktop-usage-study

Time on Site

PERFICIENT digital



Time on Site by Industry Category (Minutes)

	2016		2017		2018	
Industry	Mobile	Desktop	Mobile	Desktop	Mobile	Desktop
Arts and Entertainment	5.96	15.56	7.51	17.37	7.89	19.18
Internet and Telecom	7.67	15.03	7.85	14.93	6.71	14.57
Gambling	5.44	10.23	6.37	10.84	5.81	11.55
News and Media	3.06	10.42	3.56	9.76	3.33	8.08
Career and Education	4.48	8.36	4.73	8.69	4.34	8.91
Sports	4.14	7.48	5.65	7.49	4.10	8.62
Travel	4.30	8.82	4.70	8.95	4.18	8.51
Business and Industry	2.62	6.75	2.99	6.96	2.83	6.88
Finance	3.59	6.51	4.06	6.66	3.31	6.85
Health	2.18	5.53	2.48	5.55	2.38	5.88
Shopping	5.86	8.40	6.36	8.39	4.87	8.21
Games	3.64	7.04	4.03	7.07	3.98	7.28
Law and Government	3.21	6.03	3.85	6.28	3.33	6.58
People and Society	4.96	7.52	6.09	8.19	4.98	8.12
Computer and Electronics	2.18	5.23	2.69	5.40	2.93	5.85
Science	2.27	4.25	2.46	4.04	2.05	4.61
Autos and Vehicles	4.16	6.32	4.15	6.47	3.57	5.89
Reference	2.25	4.14	2.79	4.23	2.52	4.49
Beauty and Fitness	3.09	4.78	3.29	4.66	2.98	4.78
Food and Drink	2.51	3.91	3.03	4.30	2.95	4.75
Recreation and Hobbies	3.38	5.20	4.37	4.50	3.91	5.36
Pets and Animals	2.59	3.67	3.01	3.92	3.15	4.36
Home and Garden	2.45	3.27	2.76	3.34	2.34	3.48
Adult	9.11	10.30	10.43	10.33	10.19	10.65
Books and Literature	10.43	7.46	11.46	8.12	12.37	9.56

Other Problems

- Other Data Sources:
  - Newsgroups (USENET, etc.)
  - Images
    - Anchor text
    - Image content retrieval
  - Headline News
    - No incoming links so how to rank?
- User issues:
  - Interfaces, behavior, profiles

- How can we determine which search engine is “best”?
  - Can this be quantified?

CiteSeer, eBay, and Amazon

- CiteSeer
  - Why does citation ranking not work?
  - How should structured documents be handled?
  - How does it find academic documents?
- Amazon.com
  - How does the recommender system work?
  - How many books (CDs, etc.) in amazon.com?
  - How many users?
- eBay
  - Turns over 20% of its “documents” every day
  - How to manage change in the data structures
- Money
  - How does money change the way searching works?
  - What are the commercial secrets?

Facebook / Twitter / Social Media

- How can the index be kept up to date when new "documents" are added every second

- What is a document?

- How do you identify spam in tweets?

Other Data Types

- TEXT
  - SGML / XML / Structured documents / Blogs
- Spoken Word
- Video (e.g. CNN News Feed)
- Music
  - Query by hum
  - AppleMusic (etc.) Recommended music
- Images
  - Query by example
- Chemical Structures
- Mixed types
  - Digital libraries
  - Human genome / bioinformatics

Readings

- S. Brin, L. Page (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30(1-7):107-117

- L.A. Barroso, J. Dean, U. Holzle (2003). Web search for a planet: The Google cluster architecture, IEEE Micro23(2):22-28.

- J. Wang, E. Lo, M. L. Yiu, J. Tong, G. Wang, X. Liu (2013) The Impact of Solid State Drive on Search Engine Cache Management, Proceedings of SIGIR 2013, pp 693-702