Information Retrieval and Web Search

Class Introduction

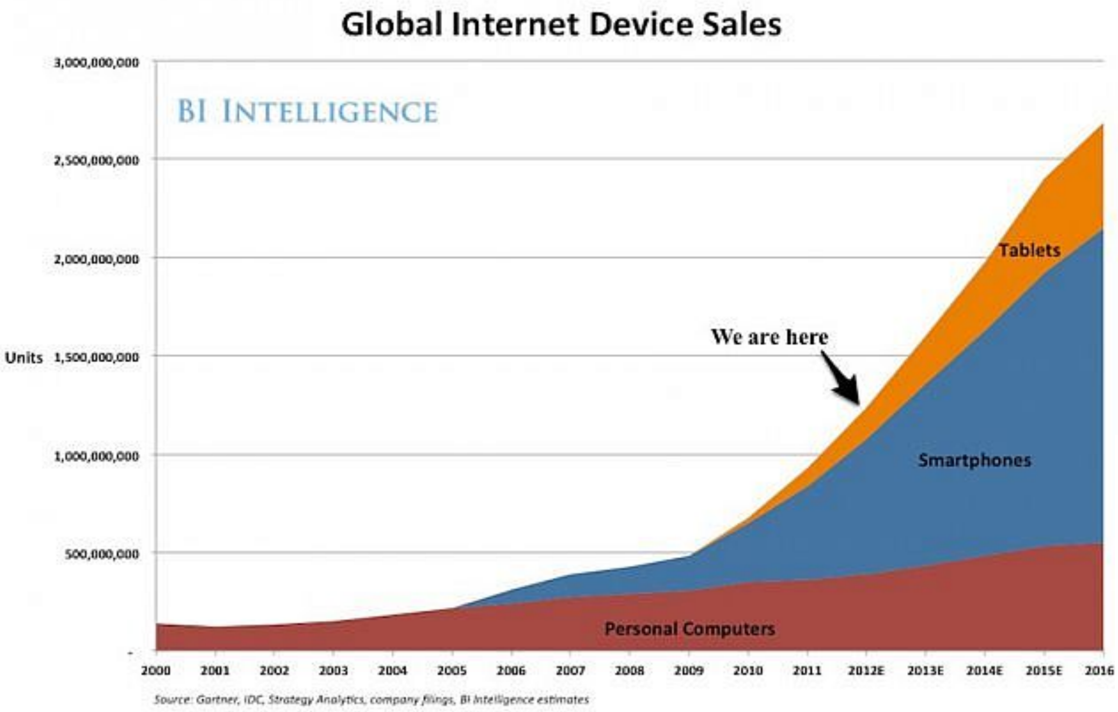
Tao Yang, 2017

http://www.cs.ucsb.edu/~tyang/class/293S17/

Introduction

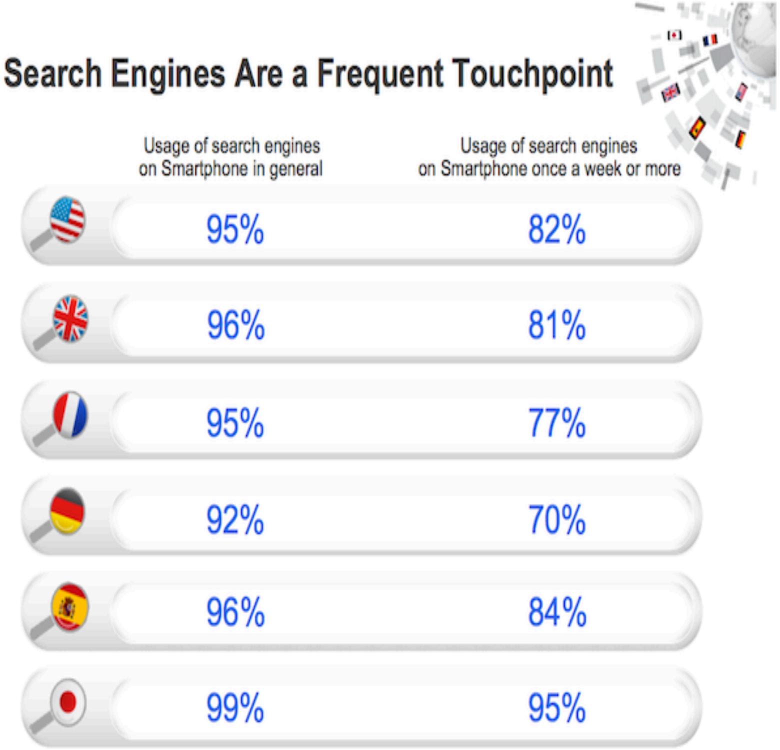
- **Internet users**
  - Interests/content
  - Importance of search engine traffic
  - Online advertisement
- **Class Topics**

Sales of PCs/Mobile Devices



Global Internet Device Sales

Source: Gartner, IDC, Strategy Analytics, company filings, BI Intelligence estimates

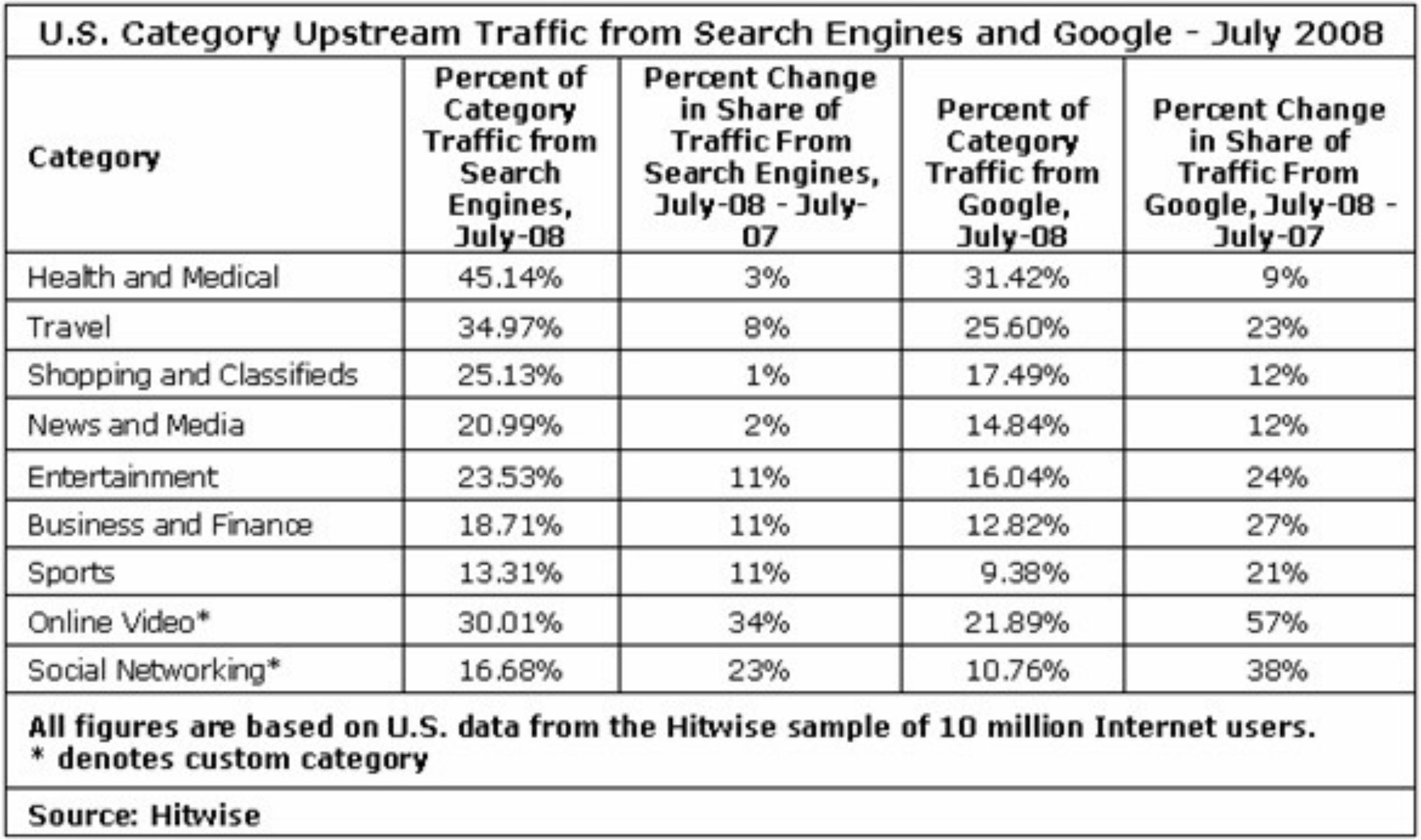


Search Engines Are a Frequent Touchpoint

	Usage of search engines on Smartphone in general	Usage of search engines on Smartphone once a week or more
🇺🇸	95%	82%
🇬🇧	96%	81%
🇫🇷	95%	77%
🇩🇪	92%	70%
🇪🇸	96%	84%
🇯🇵	99%	95%

http://www.businessinsider.com/the-future-of-mobile-deck-2012-3?op=1

Users' interests in information search



U.S. Category Upstream Traffic from Search Engines and Google - July 2008

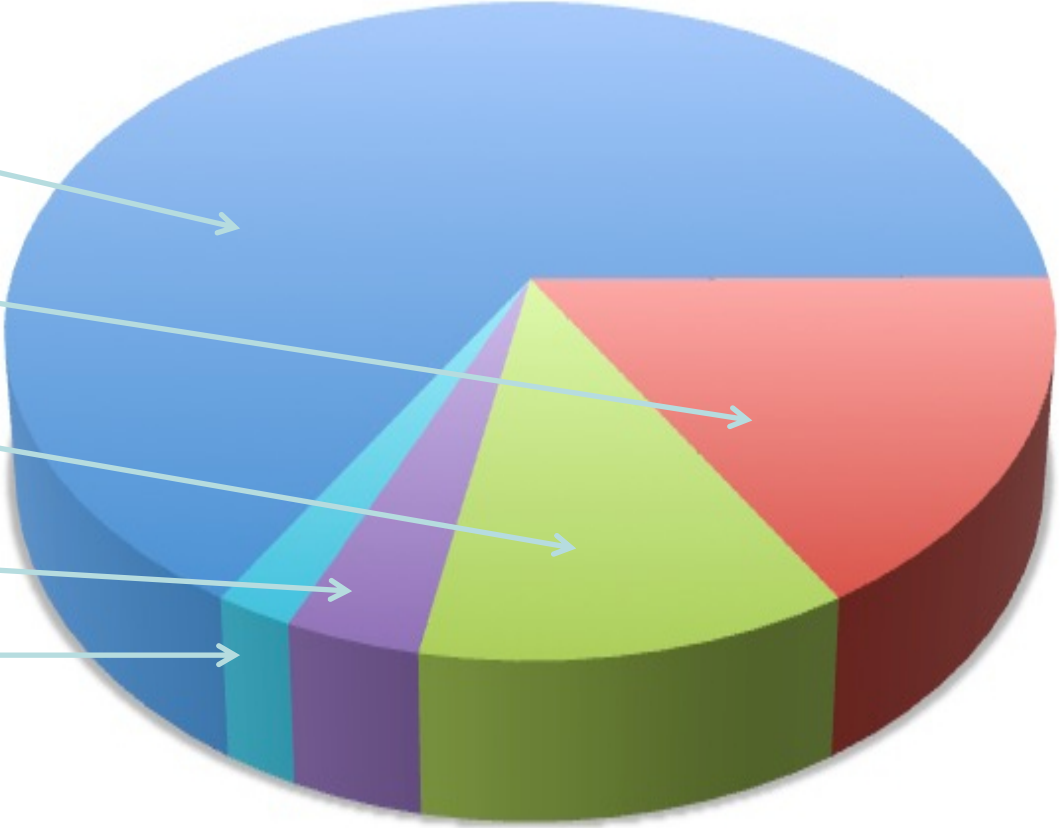
Category	Percent of Category Traffic from Search Engines, July-08	Percent Change in Share of Traffic From Search Engines, July-08 - July-07	Percent of Category Traffic from Google, July-08	Percent Change in Share of Traffic From Google, July-08 - July-07
Health and Medical	45.14%	3%	31.42%	9%
Travel	34.97%	8%	25.60%	23%
Shopping and Classifieds	25.13%	1%	17.49%	12%
News and Media	20.99%	2%	14.84%	12%
Entertainment	23.53%	11%	16.04%	24%
Business and Finance	18.71%	11%	12.82%	27%
Sports	13.31%	11%	9.38%	21%
Online Video*	30.01%	34%	21.89%	57%
Social Networking*	16.68%	23%	10.76%	38%

All figures are based on U.S. data from the Hitwise sample of 10 million Internet users.
*** denotes custom category**

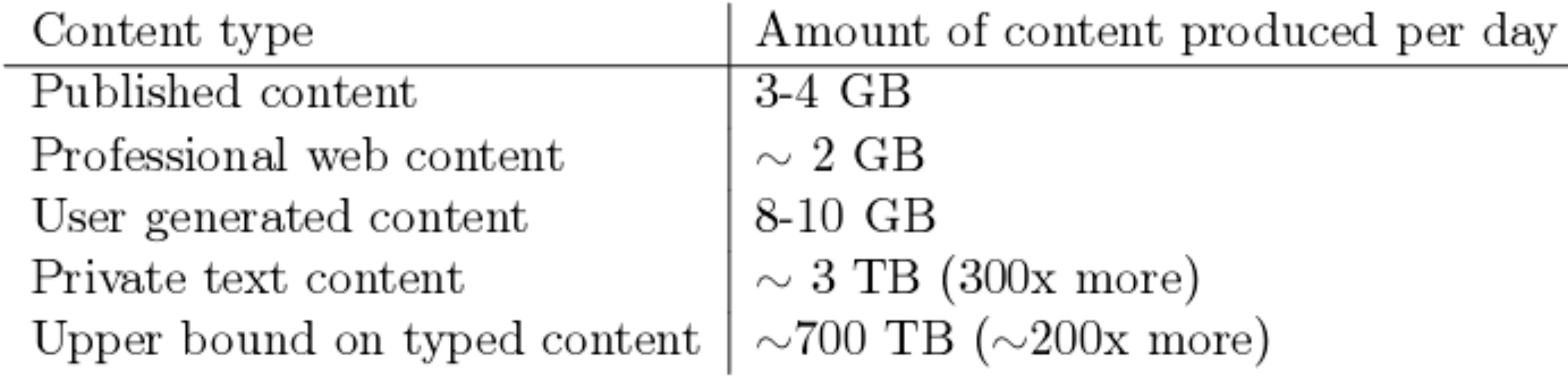
Source: Hitwise

Web Search Engine Market in USA (Jan 2016)

- Google: 63.8%
- Bing: 21.3%
- Yahoo: 12.4%
- Ask: 1.7%
- AOL: 0.9%



Content trend and ownership

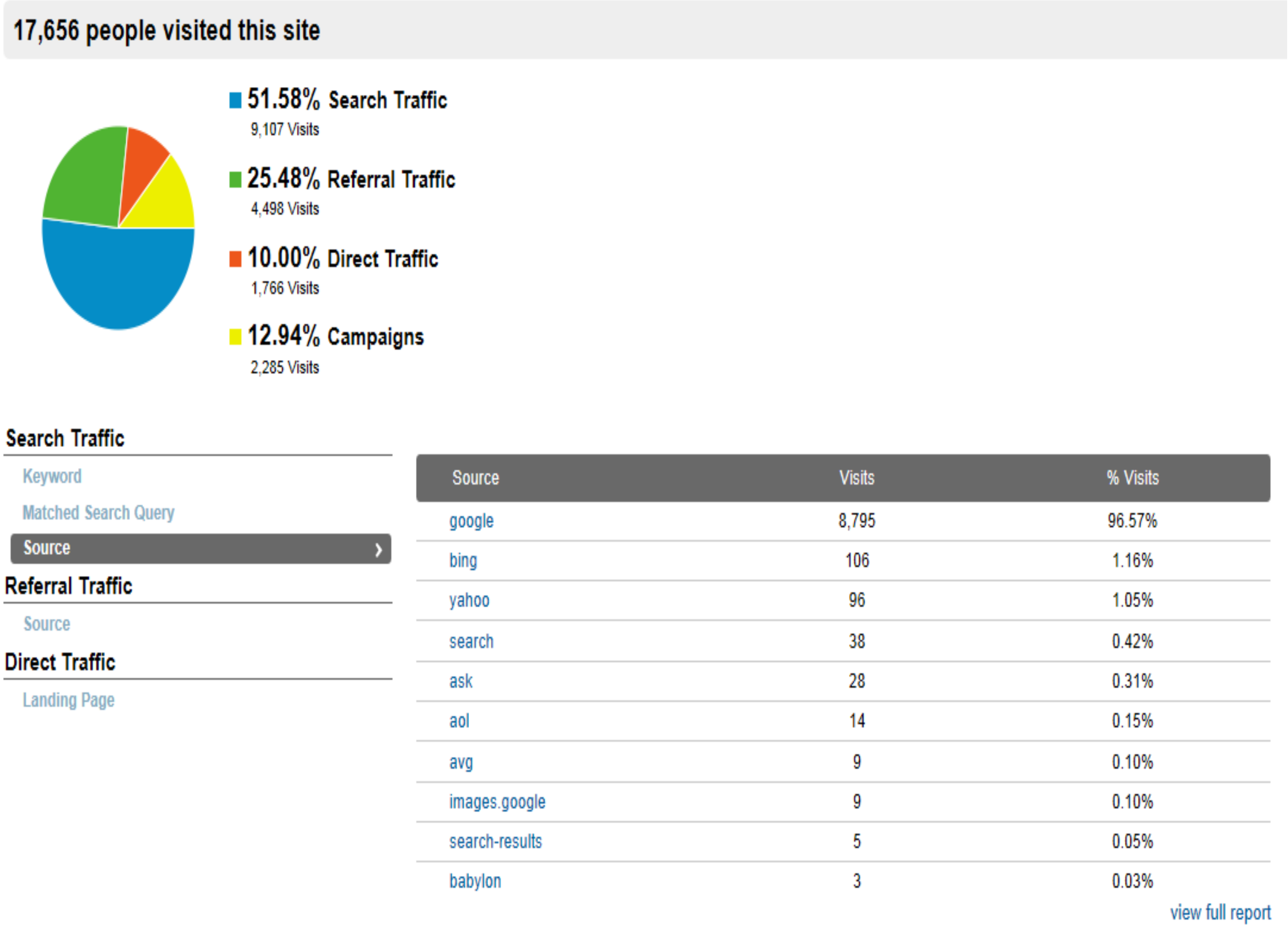


Content type	Amount of content produced per day
Published content	3-4 GB
Professional web content	~ 2 GB
User generated content	8-10 GB
Private text content	~ 3 TB (300x more)
Upper bound on typed content	~700 TB (~200x more)

[Ramakrishnan and Tomkins 2007]

- **Content consumption is fragmenting – nobody owns more than 10% of WWW pageviews**
- **No single place will own all the content**

Search Traffic is Important for Business



17,656 people visited this site

🟦 **51.58%** Search Traffic
9,107 Visits

🟩 **25.48%** Referral Traffic
4,498 Visits

🟧 **10.00%** Direct Traffic
1,766 Visits

🟨 **12.94%** Campaigns
2,285 Visits

Search Traffic

Keyword

Matched Search Query

Source ›

Referral Traffic

Source

Direct Traffic

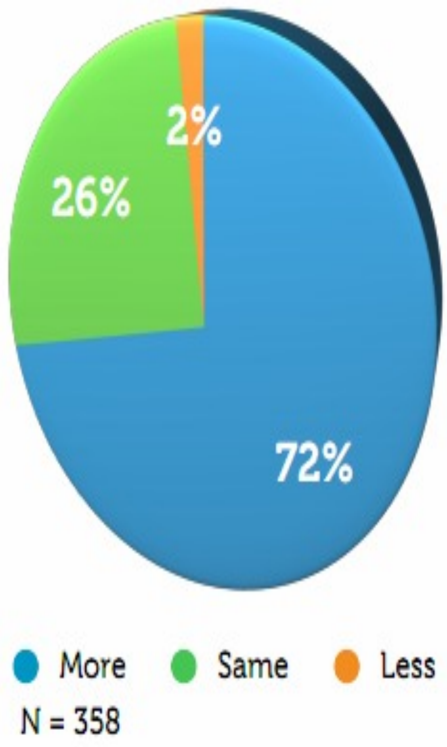
Landing Page

Source	Visits	% Visits
google	8,795	96.57%
bing	106	1.16%
yahoo	96	1.05%
search	38	0.42%
ask	28	0.31%
aol	14	0.15%
avg	9	0.10%
images.google	9	0.10%
search-results	5	0.05%
babylon	3	0.03%

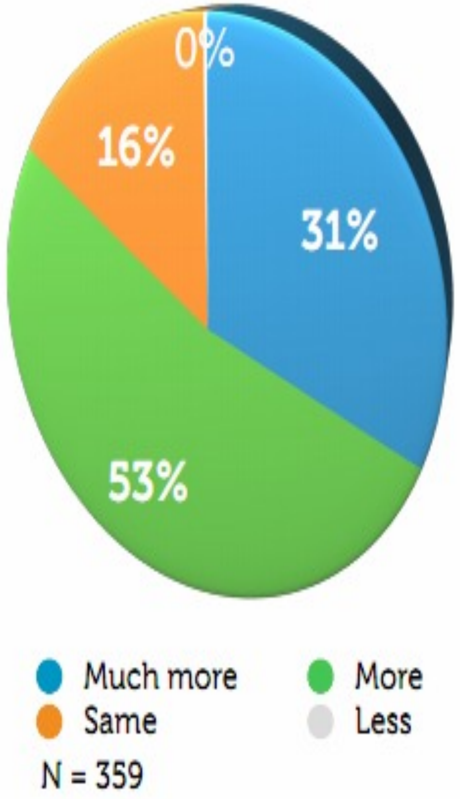
view full report

2012 Survey: Web Search Importance for Business

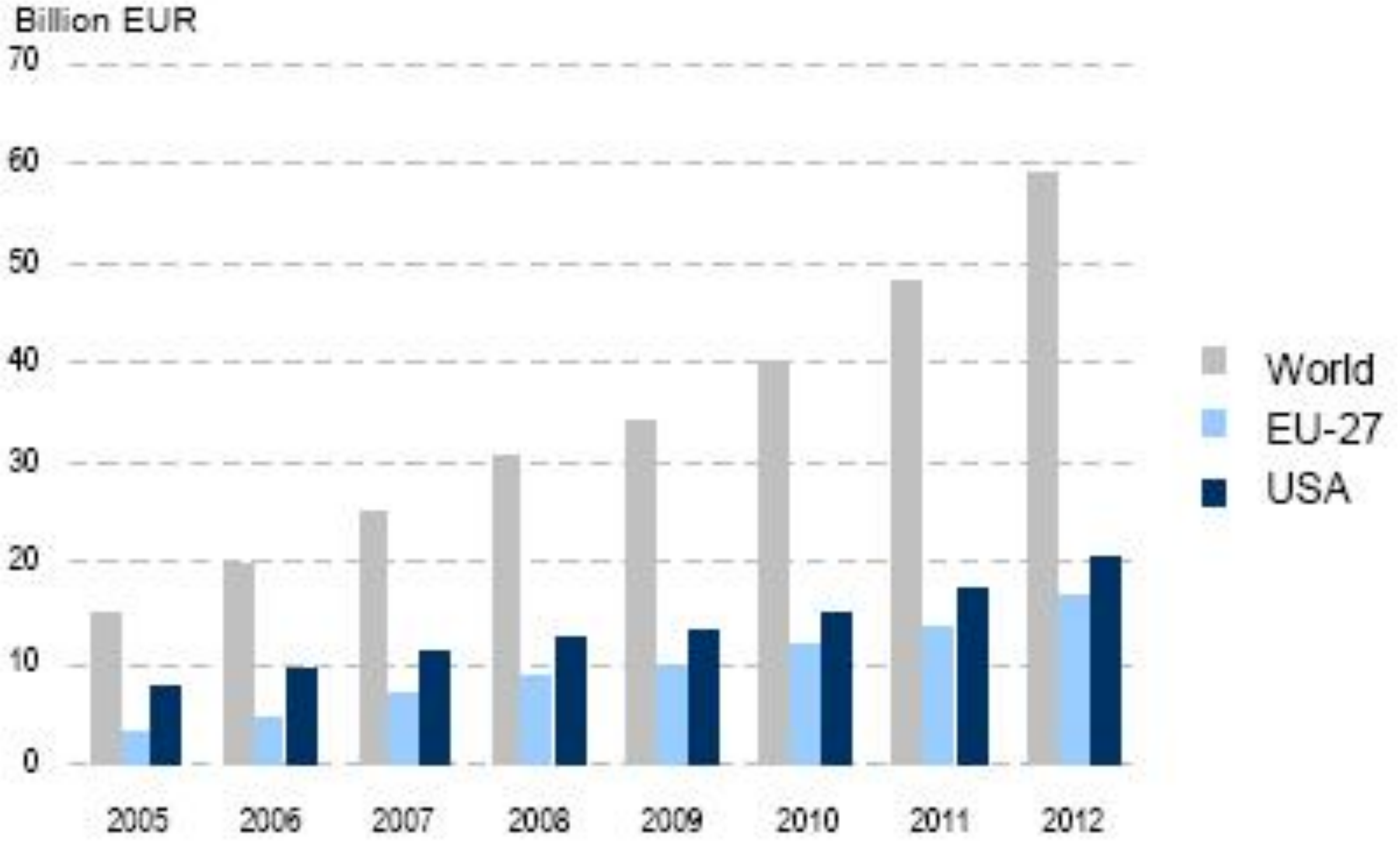
Will you spend more or less on search engine marketing technology in 2012 vs 2011?

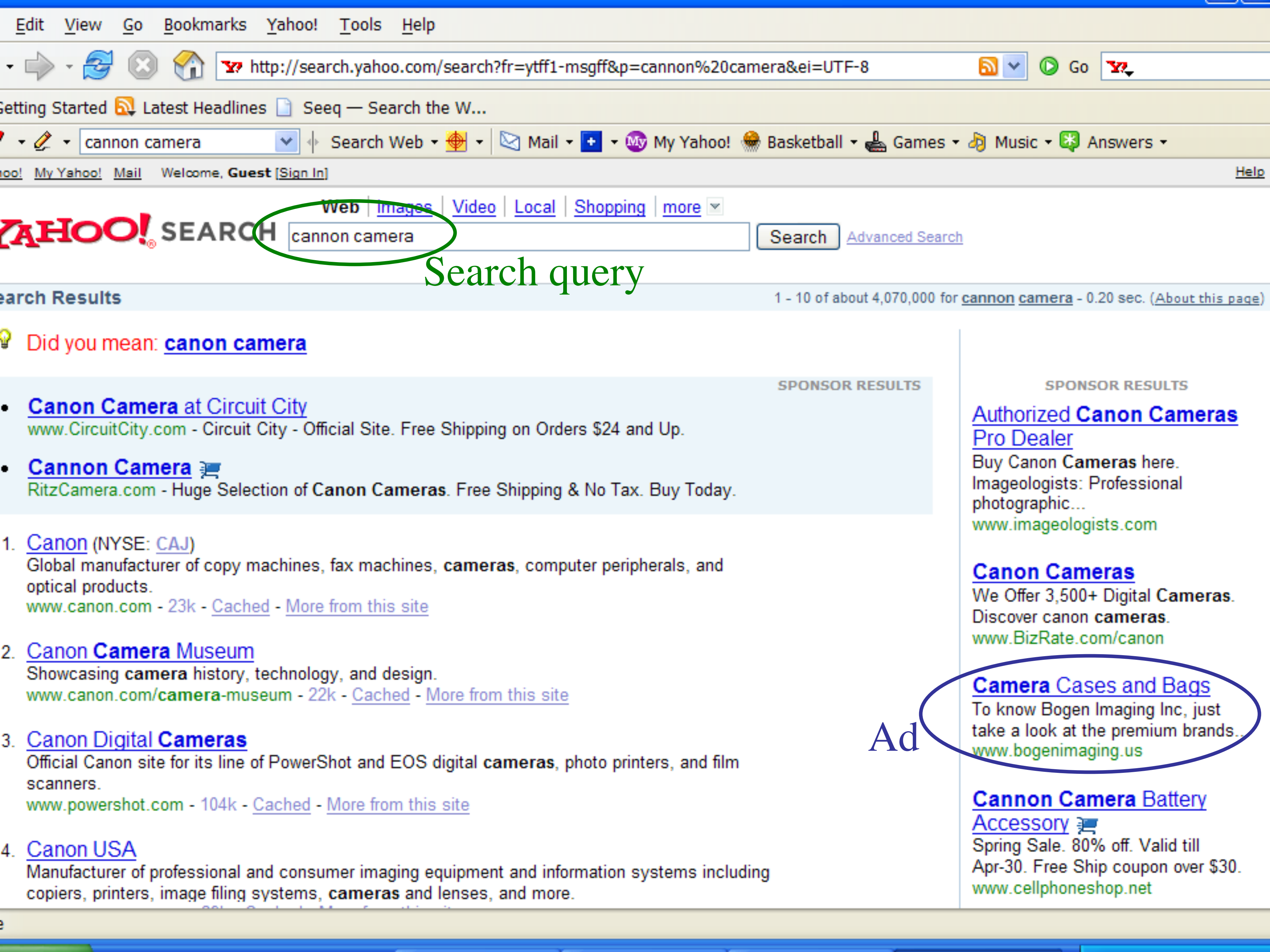


How important will social signals (Likes, Tweets, Google +1s) be to your SEO in 2012 vs 2011?



Online advertising market, Worldwide




Edit View Go Bookmarks Yahoo! Tools Help
http://search.yahoo.com/search?fr=ytff1-msgff&p=cannon%20camera&ei=UTF-8
Go
etting Started
Latest Headlines
Seeq — Search the W...
cannon camera
Search Web
Mail
My Yahoo!
Basketball
Games
Music
Answers
oo! My Yahoo! Mail Welcome, Guest [Sign In]
Help
Web Images Video Local Shopping more
YAHOO! SEARCH
Search
Advanced Search

cannon camera

Search query

earch Results

1 - 10 of about 4,070,000 for cannon camera - 0.20 sec. (About this page)

Did you mean: **canon camera**

SPONSOR RESULTS

- **Canon Camera** at Circuit City
  www.CircuitCity.com - Circuit City - Official Site. Free Shipping on Orders $24 and Up.
- **Cannon Camera**
  RitzCamera.com - Huge Selection of **Canon Cameras**. Free Shipping & No Tax. Buy Today.

1. Canon (NYSE: CAJ)
   Global manufacturer of copy machines, fax machines, **cameras**, computer peripherals, and optical products.
   www.canon.com - 23k - Cached - More from this site
2. Canon **Camera** Museum
   Showcasing **camera** history, technology, and design.
   www.canon.com/**camera**-museum - 22k - Cached - More from this site
3. Canon Digital **Cameras**
   Official Canon site for its line of PowerShot and EOS digital **cameras**, photo printers, and film scanners.
   www.powershot.com - 104k - Cached - More from this site
4. Canon USA
   Manufacturer of professional and consumer imaging equipment and information systems including copiers, printers, image filing systems, **cameras** and lenses, and more.

SPONSOR RESULTS

Authorized **Canon Cameras** Pro Dealer
Buy Canon **Cameras** here. Imageologists: Professional photographic...
www.imageologists.com

Canon Cameras
We Offer 3,500+ Digital **Cameras**. Discover canon **cameras**.
www.BizRate.com/canon

Ad

Camera Cases and Bags
To know Bogen Imaging Inc, just take a look at the premium brands...
www.bogenimaging.us

Cannon Camera Battery Accessory
Spring Sale. 80% off. Valid till Apr-30. Free Ship coupon over $30.
www.cellphoneshop.net

Course Objectives

- Practice and experience for building search services and developing related mining applications
  - Broad topics in web mining and search engines, advertisement
  - Algorithms & System support
- **Workload:**
  - Group project   (2 persons).
    - paper reviewing and presentation
    - Implementation/evaluation. Report.
  - 2 group HW exercises (Tentatively, Lucene/Solr search, Hadoop log analysis)
  - Exam vs 2 exams.

Course Topics

- **Information Retrieval & Web Search**
  - Indexing, Compression, and Online Search
  - Ranking methods with text/ link/click analysis. Machine learning.
- **Text Mining**
  - Duplicate analysis. Text Categorization and Clustering
  - Qestion answering/deep learning, Recommendation
- **Advertisement**
- **Systems Support**
  - Online servers and offline computation. MapReduce.
  - Caching. Crawling and document parsing.
  - Open source systems

Expected Work

- Tentatively Project 50%. Take-home exam 40%. 10% HW exercise.
- Timeline
  - Feb 2: 1-page project proposal (plain email text).
  - Week of Feb:
    - Meet with me and select paper(s) for reviewing.
    - Demo for HW 1
  - Mid of Feb:
    - Exam 1. Project progress & related papers presentation
  - End of Feb. HW2
    - Then schedule second meeting with me on HW2 and proj
  - Mid of March:
    - Project demo/interview
    - Final project slides/report.
  - Exam 2. Problems based on class presentation/references/HW.

Class Computing Resource & Info

- **www.cs.ucsb.edu/~tyang/class/293S17**
- **Comet supercomputer accounts:**
- **CSIL sandbox disk space**
  - /cs/sandbox/class/293SIR
  - /cs/sandbox/student/<username>
- **Class discussion group** at Google.com (we will send an invitation based on the class list).
  - https://groups.google.com/d/forum/cs290s17-ir