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

Search Engines Are a Frequent Touchpoint

<table>
<thead>
<tr>
<th>Region</th>
<th>Usage of search engines on Smartphone in general</th>
<th>Usage of search engines on Smartphone once a week or more</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>95%</td>
<td>82%</td>
</tr>
<tr>
<td></td>
<td>96%</td>
<td>81%</td>
</tr>
<tr>
<td></td>
<td>95%</td>
<td>77%</td>
</tr>
<tr>
<td></td>
<td>92%</td>
<td>70%</td>
</tr>
<tr>
<td></td>
<td>96%</td>
<td>84%</td>
</tr>
<tr>
<td></td>
<td>99%</td>
<td>95%</td>
</tr>
</tbody>
</table>

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

http://www.businessinsider.com/the-future-of-mobile-deck-2012-3?op=1
### Users’ interests in information search

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

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

<table>
<thead>
<tr>
<th>Content type</th>
<th>Amount of content produced per day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Published content</td>
<td>3-4 GB</td>
</tr>
<tr>
<td>Professional web content</td>
<td>~ 2 GB</td>
</tr>
<tr>
<td>User generated content</td>
<td>8-10 GB</td>
</tr>
<tr>
<td>Private text content</td>
<td>~ 3 TB (300x more)</td>
</tr>
<tr>
<td>Upper bound on typed content</td>
<td>~700 TB (~200x more)</td>
</tr>
</tbody>
</table>

[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**: 4,488 Visits
- **10.00% Direct Traffic**: 1,766 Visits
- **12.94% Campaigns**: 2,285 Visits

### Search Traffic

<table>
<thead>
<tr>
<th>Source</th>
<th>Visits</th>
<th>% Visits</th>
</tr>
</thead>
<tbody>
<tr>
<td>google</td>
<td>8,795</td>
<td>96.57%</td>
</tr>
<tr>
<td>bing</td>
<td>105</td>
<td>1.16%</td>
</tr>
<tr>
<td>yahoo</td>
<td>96</td>
<td>1.05%</td>
</tr>
<tr>
<td>search</td>
<td>38</td>
<td>0.42%</td>
</tr>
<tr>
<td>ask</td>
<td>28</td>
<td>0.31%</td>
</tr>
<tr>
<td>aol</td>
<td>14</td>
<td>0.15%</td>
</tr>
<tr>
<td>avg</td>
<td>9</td>
<td>0.10%</td>
</tr>
<tr>
<td>images google</td>
<td>9</td>
<td>0.10%</td>
</tr>
<tr>
<td>search-results</td>
<td>5</td>
<td>0.05%</td>
</tr>
<tr>
<td>babylon</td>
<td>3</td>
<td>0.03%</td>
</tr>
</tbody>
</table>

### Referral Traffic

### Direct Traffic

### Landing Page
2012 Survey: Web Search Importance for Business

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

- More: 72%
- Same: 26%
- Less: 2%

N = 358

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

- Much more important: 53%
- Same: 31%
- More important: 16%
- Less important: 0%

N = 359
Online advertising market, Worldwide
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
  ▪ 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

• **Text Mining**
  - Duplicate analysis. Text Categorization and Clustering
  - Question answering/deep learning, Recommendation

• **Advertisement**

• **Systems Support**
  - Online servers and offline computation. MapReduce.
  - 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