Overview

1. Web IR
   - Links
   - Queries
   - Context
   - Users
   - Documents
   - Size

2. Ads & Spam
   - Ads
   - Spam

Web search overview

Search is a top activity on the web

![Graph showing the frequency of search engine usage](image)
Without search engines, the web wouldn’t work

Without search, content is hard to find.

→ Without search, there is no incentive to create content.
  Why publish something if nobody will read it?
  Why publish something if I don’t get ad revenue from it?

Interest aggregation
  Unique feature of the web: A small number of geographically dispersed people with similar interests can find each other.
  Elementary school kids with hemophilia
  People interested in translating R5R5 Scheme into relatively portable C (open source project)
  Interest aggregation without search engines is not possible.

Somebody needs to pay for the web.
  Servers, web infrastructure, content creation
  A large part today is paid by search ads.

Web IR: Differences from traditional IR

Links: The web is a hyperlinked document collection.

Queries: Web queries are different, more varied and there are a lot of them. How many? 10^8 every day, approaching 10^9

Users: Users are different, more varied and there are a lot of them. How many? 10^9

Documents: Documents are different, more varied and there are a lot of them. How many? \approx 10^{11}. Indexed: 10^{10}

Context: Context is more important on the web than in many other IR applications.

Ads and spam
Search in a hyperlinked collection

- Web search in most cases is interleaved with navigation . . .
- . . . i.e., with following links.
- Different from most other IR collections

Bowtie structure of the web

- Strongly connected component (SCC) in the center
- Lots of pages that get linked to, but don’t link (OUT)
- Lots of pages that link to other pages, but don’t get linked to (IN)
- Tendrils, tubes, islands

Outline

1. Web IR
   - Links
   - Queries
   - Context
   - Users
   - Documents
   - Size

2. Ads & Spam
   - Ads
   - Spam
Query distribution (1)

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1 sex</td>
<td>16</td>
<td>crack</td>
<td>31</td>
<td>juegos</td>
<td>46</td>
</tr>
<tr>
<td>2 (artifact)</td>
<td>17</td>
<td>games</td>
<td>32</td>
<td>nude</td>
<td>47</td>
</tr>
<tr>
<td>3 (artifact)</td>
<td>18</td>
<td>pussy</td>
<td>33</td>
<td>music</td>
<td>48</td>
</tr>
<tr>
<td>4 porno</td>
<td>19</td>
<td>cracks</td>
<td>34</td>
<td>musica</td>
<td>49</td>
</tr>
<tr>
<td>5 mp3</td>
<td>20</td>
<td>lolita</td>
<td>35</td>
<td>anal</td>
<td>50</td>
</tr>
<tr>
<td>6 Halloween</td>
<td>21</td>
<td>britney spears</td>
<td>36</td>
<td>free6</td>
<td>51</td>
</tr>
<tr>
<td>7 sexo</td>
<td>22</td>
<td>ebay</td>
<td>37</td>
<td>avril lavigne</td>
<td>52</td>
</tr>
<tr>
<td>8 chat</td>
<td>23</td>
<td>sexe</td>
<td>38</td>
<td><a href="http://www.hotmail.com">www.hotmail.com</a></td>
<td>53</td>
</tr>
<tr>
<td>9 porn</td>
<td>24</td>
<td>Pamela Anderson</td>
<td>39</td>
<td>winzip</td>
<td>54</td>
</tr>
<tr>
<td>10 yahoo</td>
<td>25</td>
<td>warez</td>
<td>40</td>
<td>fuck</td>
<td>55</td>
</tr>
<tr>
<td>11 KaZaA</td>
<td>26</td>
<td>divx</td>
<td>41</td>
<td>wallpaper</td>
<td>56</td>
</tr>
<tr>
<td>12 xxx</td>
<td>27</td>
<td>gay</td>
<td>42</td>
<td>hotmail.com</td>
<td>57</td>
</tr>
<tr>
<td>13 Hentai</td>
<td>28</td>
<td>harry potter</td>
<td>43</td>
<td>postales</td>
<td>58</td>
</tr>
<tr>
<td>14 lyrics</td>
<td>29</td>
<td>playboy</td>
<td>44</td>
<td>shakira</td>
<td>59</td>
</tr>
<tr>
<td>15 hotmail</td>
<td>30</td>
<td>lolitas</td>
<td>45</td>
<td>traductor</td>
<td>60</td>
</tr>
</tbody>
</table>

More than 1/3 of these are queries for adult content. Does this mean that most people are looking for adult content?

Types of queries / user needs in web search

- **Informational user needs**: I need information on something. “low hemoglobin"
- We called this “information need” earlier in the class.
- On the web, information needs proper are only a subclass of user needs.
- Other user needs: Navigational and transactional
- **Navigational user needs**: I want to go to this web site. “hotmail”, “myspace”, “United Airlines"
- **Transactional user needs**: I want to make a transaction.
  - Buy something: “MacBook Air"
  - Download something: “Acrobat Reader"
  - Chat with someone: “live soccer chat"
- Difficult problem: How can the search engine tell what the user need or intent for a particular query is?

Outline

1 Web IR
   - Links
   - Queries
   - Context
   - Users
   - Documents
   - Size

2 Ads & Spam
   - Ads
   - Spam
User intent: Answering the need behind the query

- What can we do to guess user intent?
  - Guess user intent independent of context:
    - Spell correction
    - Precomputed “typing” of queries (next slide)
  - Better: Guess user intent based on context:
    - Geographic context (slide after next)
    - Context of user in this session (e.g., previous query)
    - Context provided by personal profile (Yahoo/MSN do this, Google claims it doesn’t)

The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)
  - User (located in Palo Alto)
- Locating the user
  - IP address
  - Information provided by user (e.g., in user profile)
  - Mobile phone
- Geo-tagging: Parse text and identify the coordinates of the geographic entities
  - Example: East Palo Alto CA → Latitude: 37.47 N, Longitude: 122.14 W
  - Important NLP problem

Guessing of user intent by “typing” queries

- Calculation: 5+4
- Unit conversion: 1 kg in pounds
- Currency conversion: 1 euro in kronor
- Tracking number: 8167 2278 6764
- Flight info: LH 454
- Area code: 650
- Map: columbus oh
- Stock price: msft
- Albums/movies etc: coldplay

How do we use context to modify query results?

- Result restriction: Don’t consider inappropriate results
  - For user on google.fr . . .
  - . . . only show .fr results
- Ranking modulation: use a rough generic ranking, rerank based on personal context
- Contextualization / personalization is an area of search with a lot of potential for improvement.
Outline

1. Web IR
   - Links
   - Queries
   - Context
   - Users
   - Documents
   - Size

2. Ads & Spam
   - Ads
   - Spam

Users of web search

- Use short queries (average < 3)
- Rarely use operators
- Don’t want to spend a lot of time on composing a query
- Only look at the first couple of results
- Want a simple UI, not a search engine start page overloaded with graphics
- Extreme variability in terms of user needs, user expectations, experience, knowledge, ...
  - Industrial/developing world, English/Estonian, old/young, rich/poor, differences in culture and class
- One interface for hugely divergent needs

How do users evaluate search engines?

- Classic IR relevance (as measured by $F$) can also be used for web IR.
- Equally important: Trust, duplicate elimination, readability, loads fast, no pop-ups
- On the web, precision is more important than recall.
  - Precision at 1, precision at 10, precision on the first 2-3 pages
  - But there is a subset of queries where recall matters.

Web information needs that require high recall?

- ?
- Has this idea been patented?
- Searching for info on a prospective financial advisor
- Searching for info on a prospective employee
- Searching for info on a date
Web documents: different from other IR collections

- Distributed content creation: no design, no co-ordination
  - “Democratization of publishing”
  - Result: extreme heterogeneity of documents on the web
- Unstructured (text, html), semistructured (html, xml), structured/relational (databases)
- Dynamically generated content

Dynamic content

- Dynamic pages are generated from scratch when the user requests them – usually from underlying data in a database.
- Example: current status of flight LH 454

Dynamic content (2)

- Most (truly) dynamic content is ignored by web spiders.
  - It’s too much to index it all.
- Actually, a lot of “static” content is also assembled on the fly (asp, php etc.: headers, date, ads etc)
Web pages change frequently (Fetterly 1997)

Multilinguality

- Documents in a large number of languages
- Queries in a large number of languages
- First cut: Don’t return English results for a Japanese query
- However: Frequent mismatches query/document languages
- Many people can understand, but not query in a language
- Translation is important.
- Google example: “Beaujolais Nouveau -wine”

Duplicate documents

- Significant duplication – 30%–40% duplicates in some studies
- Duplicates in the search results were common in the early days of the web.
- Today’s search engines eliminate duplicates very effectively.
- Key for high user satisfaction

Trust

- For many collections, it is easy to assess the trustworthiness of a document.
  - A collection of Reuters newswire articles
  - A collection of TASS (Telegraph Agency of the Soviet Union) newswire articles from the 1980s
  - Your Outlook email from the last three years
- Web documents are different: In many cases, we don’t know how to evaluate the information.
- Hoaxes abound.
Growth of the web

The web keeps growing.
But growth is no longer exponential?

Size of the web: Issues

- What is size? Number of web servers? Number of pages?
  Terabytes of data available?
- Some servers are seldom connected.
  - Example: Your laptop running a web server
  - Is it part of the web?
- The “dynamic” web is infinite.
  - Any sum of two numbers is its own dynamic page on Google.
    (Example: “2+4”)

“Search engine index contains \( N \) pages”: Issues

- Can I claim a page is in the index if I only index the first 4000 bytes?
- Can I claim a page is in the index if I only index anchor text pointing to the page?
  - There used to be (and still are?) billions of pages that are only indexed by anchor text.
How would you estimate the number of pages indexed by a web search engine?

Simple method for determining a lower bound

- OR-query of frequent words in a number of languages
- http://ifnlp.org/ir/sizeoftheweb.html
- According to this query: Size of web $\geq 21,450,000,000$ on 2007.07.07 and $\geq 25,350,000,000$ on 2008.07.03
- But page counts of google search results are only rough estimates.

Outline

1. Web IR
   - Links
   - Queries
   - Context
   - Users
   - Documents
   - Size

2. Ads & Spam
   - Ads
   - Spam
First generation of search ads: Goto (1996)

- No separation of ads/docs. Just one result!
- Buddy Blake bid the maximum ($0.38) for this search.
- He paid $0.38 to Goto every time somebody clicked on the link.
- Upfront and honest. No relevance ranking, but Goto did not pretend there was any.

Second generation of search ads: Google (2000/2001)

- Strict separation of search results and search ads

Ranking of advertisers in search results
- SogoTrade appears in search results.
- SogoTrade appears in ads.
- Do search engines rank advertisers higher than non-advertisers?
- All major search engines claim no.

Do ads influence editorial content?
- Similar problem at newspapers / TV channels
- A newspaper is reluctant to publish harsh criticism of its major advertisers.
- The line often gets blurred at newspapers / on TV.
- No known case of this happening with search engines yet.
How are ads placed?

- Advertisers bid for keywords.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are only charged when somebody clicks on your ad.
- How does the advertiser determine its bid price and how does the search engine take this into account in ranking and billing?
  - Basis is a second price auction, but with twists
  - Lots of interesting work on this
  - Squeeze an additional fraction of a cent from each ad means billions of additional revenue for the search engine.
  - Many interesting jobs in this area

Keywords with high bids

According to http://www.cwire.org/highest-paying-search-terms/

<table>
<thead>
<tr>
<th>Bid Price</th>
<th>Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>$69.1</td>
<td>mesothelioma treatment options</td>
</tr>
<tr>
<td>$68.35</td>
<td>personal injury lawyer michigan</td>
</tr>
<tr>
<td>$62.99</td>
<td>student loans consolidation</td>
</tr>
<tr>
<td>$61.44</td>
<td>car accident attorney los angeles</td>
</tr>
<tr>
<td>$58.44</td>
<td>home liability insurance quote</td>
</tr>
<tr>
<td>$59.39</td>
<td>arizona dui lawyer</td>
</tr>
<tr>
<td>$57.87</td>
<td>michigan car accident attorney</td>
</tr>
<tr>
<td>$53.99</td>
<td>free auto insurance quote</td>
</tr>
<tr>
<td>$53.17</td>
<td>personal injury lawyers los angeles</td>
</tr>
<tr>
<td>$52.35</td>
<td>free online auto insurance quote</td>
</tr>
<tr>
<td>$50.4</td>
<td>accident attorney michigan</td>
</tr>
<tr>
<td>$50.35</td>
<td>michigan auto accident attorney</td>
</tr>
<tr>
<td>$49.25</td>
<td>accident helpline</td>
</tr>
<tr>
<td>$47.74</td>
<td>automobile accident lawyer</td>
</tr>
<tr>
<td>$47.49</td>
<td>dui defense attorneys</td>
</tr>
<tr>
<td>$46.44</td>
<td>asbestos cancer</td>
</tr>
<tr>
<td>$46.34</td>
<td>arizona dui</td>
</tr>
<tr>
<td>$45.5</td>
<td>business liability insurance quote</td>
</tr>
<tr>
<td>$43.86</td>
<td>loan consolidation</td>
</tr>
<tr>
<td>$42.98</td>
<td>student loan consolidation</td>
</tr>
<tr>
<td>$40.7</td>
<td>dui defense lawyers</td>
</tr>
<tr>
<td>$40.1</td>
<td>home equity line of credit</td>
</tr>
<tr>
<td>$39.65</td>
<td>life insurance quotes</td>
</tr>
<tr>
<td>$39.78</td>
<td>criminal lawyers new york</td>
</tr>
<tr>
<td>$39.52</td>
<td>loan federal consolidation</td>
</tr>
<tr>
<td>$39.23</td>
<td>refinancing</td>
</tr>
<tr>
<td>$38.72</td>
<td>equity line of credit</td>
</tr>
<tr>
<td>$37.95</td>
<td>lasik eye surgery new york city</td>
</tr>
<tr>
<td>$37.3</td>
<td>2nd mortgage</td>
</tr>
<tr>
<td>$35.9</td>
<td>free car insurance quote</td>
</tr>
</tbody>
</table>

Google AdWords demo

How are ads ranked?

- First cut: according to bid price
  - Bad idea: open to abuse
  - Example: query accident → ad buy a new car
- Instead: rank based on bid price and relevance
- Key measure of ad relevance: clickthrough rate
- Result: A non-relevant ad will be ranked low.
  - Even if this decreases search engine revenue short-term
  - Hope: Overall acceptance of the system and overall revenue is maximized if users get useful information.
- Other ranking factors: location, time of day, quality and loading speed of landing page
- The main factor of course is the query – and possibly earlier queries in the same session.
Ranking of advertisers in search results

Search ads: A win-win-win?

- The search engine company gets revenue every time somebody clicks on an ad.
- The user only clicks on an ad if they are interested in the ad.
  - Search engines punish misleading and nonrelevant ads.
  - As a result, users are often satisfied with what they find after clicking on an ad.
  - Being willing to pay for ads on a search engine is a quality signal (one of many) that users take into account.
- The advertiser finds new customers in a cost-effective way.

The appeal of search ads to advertisers

- Why is web search potentially more attractive for advertisers than TV spots, newspaper ads or radio spots?
- Someone who just searched for “Saturn Aura Sport Sedan” is infinitely more likely to buy one than a random person watching TV.
- Most importantly, the advertiser only pays if the customer took an action indicating interest (i.e., clicking on the ad).

But frequently it’s not a win-win-win

- Example: keyword arbitrage
  - Buy a keyword at Google
  - Then redirect traffic to a third party that is paying much more than you had to pay to Google
  - This rarely makes sense for the user.
- Ad spammers keep inventing new tricks.
- The search engines need time to catch up with them.
Who owns a search term?

- Example: geico
- During part of 2005: The search term "geico" on Google was bought by competitors.
- Geico lost this case in the United States.
- Currently in the courts: Louis Vuitton case in Europe
- See http://google.com/tm_complaint.html

Outline

1. Web IR
   - Links
   - Queries
   - Context
   - Users
   - Documents
   - Size

2. Ads & Spam
   - Ads
   - Spam

The goal of spamming on the web

- You have a page that will generate lots of revenue for you if people visit it.
- Therefore, you would like to direct visitors to this page.
- One way of doing this: get your page ranked highly in search results.
- How can I get my page ranked highly?

Spam technique: Keyword stuffing / Hidden text

- Misleading meta-tags, excessive repetition
- Hidden text with colors, style sheet tricks etc.
- Used to be very effective, most search engines now catch these
Keyword stuffing

Spam technique: Doorway and lander pages

- Doorway page: optimized for a single keyword, redirects to the real target page
- Lander page: optimized for a single keyword or a misspelled domain name, designed to attract surfers who will then click on ads

Lander page

Spam technique: Duplication

- Get good content from somewhere (steal it or produce it yourself)
- Publish a large number of slight variations of it
- For example, publish the answer to a tax question with the spelling variations of “tax deferred” on the previous slide

Number one hit on Google for the search “composita”
The only purpose of this page: get people to click on the ads and make money for the page owner
Spam technique: Cloaking

Serve fake content to search engine spider
So do we just penalize this always?
No: legitimate uses (e.g., different content to US vs. European users)

SEO: Search engine optimization

Promoting a page in the search rankings is not necessarily spam.
It can also be a legitimate business – which is called SEO.
You can hire an SEO firm to get your page highly ranked.
There are many legitimate reasons for doing this.
  - For example, Google bombs like *Who is a failure?*
And there are many legitimate ways of achieving this:
  - Restructure your content in a way that makes it easy to index
  - Talk with influential bloggers and have them link to your site
  - Add more interesting and original content

The war against spam

Quality indicators
  - Links, statistically analyzed (PageRank etc)
  - Usage (users visiting a page)
  - No adult content (e.g., no pictures with flesh-tone)
  - Distribution and structure of text (e.g., no keyword stuffing)
Combine all of these indicators and use machine learning
Editorial intervention
  - Blacklists
  - Top queries audited
  - Complaints addressed
  - Suspect patterns detected

Spam technique: Link spam

Create lots of links pointing to the page you want to promote
Put these links on pages with high (or at least non-zero) PageRank
  - Newly registered domains (domain flooding)
  - A set of pages that all point to each other to boost each other’s PageRank (mutual admiration society)
  - Pay somebody to put your link on their highly ranked page (“schuetze horoskop” example)
  - Leave comments that include the link on blogs
Webmaster guidelines

- Major search engines have guidelines for webmasters.
- These guidelines tell you what is legitimate SEO and what is spamming.
- Ignore these guidelines at your own risk
- Once a search engine identifies you as a spammer, all pages on your site may get low ranks (or disappear from the index entirely).
- There is often a fine line between spam and legitimate SEO.
- Scientific study of fighting spam on the web: adversarial information retrieval

Resources

- Chapter 19 of IIR
- Resources at http://ifnlp.org/ir
- Size of the web queries
- Trademark issues (Geico and Vuitton cases)
- How ads are priced
- How search engines fight webspam
- Adversarial IR site at Lehigh