Ranking of ads

- Goto model:
  - Rank according to how much advertiser pays
- Current model:
  - Balance auction price and relevance
  - Irrelevant ads (few click-throughs)
    - Decrease opportunities for relevant ads
    - Harm the user experience
  - Idea: Well-targeted advertising is good for everyone
Paying for advertisements

- CPM
  - “Cost Per Mil”
  - Pay for 1000 eyeballs
  - Important for branding campaigns

- CPC
  - “Cost per Click”
  - Pay for clicking on ads
  - Important for sales campaigns
Overview

- Introduction
- Classic Information Retrieval
- Web IR
- Sponsored Search
- Web Search Basics
  - Size of the Web
- Web Users
- Spam
Web Search Basics

The Web Corpus

- No design/coordination
- Distributed content creation, linking
- “Democratization of publishing”
- Content includes truth, lies, contradictions, etc.
- Unstructured Data (text, html)
- Semi-Structured (XML, annotated photos)
- Structured (Databases)
- Scale is much larger than previous text corpora
The Web Corpus

- Growth - slowing from “doubling every few months”, but still expanding
Dynamic Content

- Content can be dynamically generated
- There is no static HTML version
- Flight status information, evite responses
- Assembled on request ("?" in URL is a clue)
Dynamic Content

- Most (truly) dynamic content is ignored by web spiders
  - Too much to index
  - Static information is more important for search
  - Spider Traps look dynamic
- Actually a lot of “static” content is assembled on the fly
  - Also
    - ASP, PHP, JSP, ads, etc....
Web Search Basics

The Web as a graph

- Web pages are nodes
- Hyperlinks are directed edges
Web Search Basics

Characteristics of the web

• Significant Duplication
• 30%-40% is some studies [Brod97, Shiv99]
• www.copyscape.com
• High linkage
• more than 8 links per page on average
• Spam
• Billions of pages of it.
The User
flickr:crankyT

Search Results

Web Spider

The Web

Indexer

Indices

Ad Indices

Search Basics
How big is the web?

- What is measured?
  - Number of hosts
  - Number of "static" html pages
- Number of hosts - netcraft survey
  - Monthly report on hosts and servers
- Number of pages
  - Lots of estimates which warrant further discussion
How big is the web?

- Netcraft Web Server Survey
Rate of change

- [Cho00] 720k pages from 270 popular sites sample daily for 5 months in 1999
  - 40% changed weekly, 23% daily
- [Fett02] Massive study: 151M pages checked over a few months
  - Significant changes 7% weekly
  - Any change 25% weekly
Rate of change

- [Ntul04] 154 large sites recrawled from scratch weekly
  - 8% had new pages every week
  - 8% die
  - 5% new content
  - 25% new links per week
Size of the Web

Rate of change

- Fetterly et al. study in 2002
- 150 million pages over 11 weekly crawls
- Bucketed into 85 groups according to amount of change
Web Evolution

• The nature of the web is change
• Not much work on studying web evolution
  • Exception is Fetterly et. al, 2003
  • Some effort has been made to extrapolate from small samples using fractal models [Dill et. al. 2001]
Overview

- Introduction
- Classic Information Retrieval
- Web IR
- Sponsored Search
- Web Search Basics
- Size of the Web
- Web Users
- Spam
User Search Needs in Brod02/RL04

- Informational
  - Want to learn about something (~40%/65%)
- Navigational
  - Want to go to that page (~25%/15%)
- Transactional
  - Want to do something (~35%/20%)
    - Access a service, download, shop
- Others?
  - Exploration, social, etc...
Web Users

- Make ill defined queries
  - Short
    - Average in 2001: 2.54 terms (80% < 3 words)
    - Average in 1998: 2.35 terms (88% < 3 words) [Silv98]
  - Imprecise terms
  - Suboptimal syntax (no operators)
  - Low effort (spelling mistakes)
Web Users

- Wide Variance in
  - Needs
  - Expectations
  - Knowledge
  - Bandwidth
Web Users

• Behavior
  • 85% look over one result screen only
  • 78% of queries are not modified
  • Follow links ("the scent of information")
Power law

- Few popular broad queries
- Many rare specific queries
**Web Users**

**Top queries**

- Most are related to sex
- 2007 Who What How (Google)

<table>
<thead>
<tr>
<th>Who is...</th>
<th>What is...</th>
<th>How to...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. who is god</td>
<td>1. what is love</td>
<td>1. how to kiss</td>
</tr>
<tr>
<td>2. who is who</td>
<td>2. what is autism</td>
<td>2. how to draw</td>
</tr>
<tr>
<td>3. who is lookup</td>
<td>3. what is rss</td>
<td>3. how to knit</td>
</tr>
<tr>
<td>4. who is jesus</td>
<td>4. what is lupus</td>
<td>4. how to hack</td>
</tr>
<tr>
<td>5. who is it</td>
<td>5. what is sap</td>
<td>5. how to dance</td>
</tr>
<tr>
<td>6. who is buckethead</td>
<td>6. what is bluetooth</td>
<td>6. how to crochet</td>
</tr>
<tr>
<td>7. who is calling</td>
<td>7. what is emo</td>
<td>7. how to meditate</td>
</tr>
<tr>
<td>8. who is keppler</td>
<td>8. what is java</td>
<td>8. how to flirt</td>
</tr>
<tr>
<td>9. who is this</td>
<td>9. what is hpv</td>
<td>9. how to levitate</td>
</tr>
<tr>
<td>10. who is satan</td>
<td>10. what is gout</td>
<td>10. how to skateboard</td>
</tr>
</tbody>
</table>
How far do people look for results?

If you don't find what you are looking for, at what point do you move on either to another search engine or to another search on the same engine?

- After the first few entries: 22.6%
- After the first page: 18.6%
- After the first two pages: 25.8%
- After the first three pages: 14.7%
- More than three pages: 10.8%
- The whole list, unless it's dozens of pages: 7.4%

Number of Responses
How do users evaluate search engines?

- Quality of pages
- Classic IR relevance
- Also important:
  - Trust
  - Duplicate elimination
  - Readability
  - Fast Access
  - No pop-ups
How do users evaluate search engines?

• Precision is more important than recall
  • Precision:
  • How precise is a portal in locating relevant results?
  • Recall
  • How thorough is the coverage of available relevant results?
• Precision with 1 result, 10 results, 2-3 pages of results.
• When is recall important?
How do users evaluate search engines?

- Recall is sometimes important:
  - Googling for a new doctor
  - Googling a prospective employee
  - Googling your date
How do users evaluate search engines?

- Good U/I
- Simple
- No Clutter
- Pre and post processing tools
- Spell check ("Did you mean ....?")
- Suggested alternative searches
- Links to resources (maps, images, stock quotes)
- Able to deal with typical behavior
  - e.g., a URL typed into a search box
Loyalty to a given search engine

- iProspect Survey 4/2004

Which would you say best describes how you use search engines?

- I usually use the same search engine or directory (56.7%)
- I have several favorite search engines and use them interchangeably (30.5%)
- I use different search engines for different types of searches (12.8%)