Web Search Basics

Introduction to Information Retrieval
INF 141/ CS 121
Donald J. Patterson

Content adapted from Hinrich Schütze
http://www.informationretrieval.org
Web Search Basics

The Web as a graph

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

The User
flickr:crankyT

Search Results

Search Basics

Web Spider

The Web

Indexer

Indices

Ad Indices
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 ever week
  - 8% die
  - 5% new content
  - 25% new links per week
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
User Search Needs in Brod02/RL04

- Informational
- Want to learn about something (~40%/65%)
User Search Needs in Brod02/RL04

- Informational
  - Want to learn about something (~40%/65%)
- Navigational
  - Want to go to that page (~25%/15%)
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
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
Top queries

- Most are related to sex
- 2008 Who What How (Google)

Who is...
1. who is obama
2. who is mccain
3. who is palin
4. who is lil wayne
5. who is miley cyrus
6. who is dolla
7. who is jonas brothers
8. who is chris brown
9. who is biden
10. who is martin luther

What is...
1. what is love
2. what is life
3. what is java
4. what is sap
5. what is rss
6. what is scientology
7. what is autism
8. what is lupus
9. what is 3g
10. what is art

How to...
1. how to draw
2. how to kiss
3. how to write
4. how to cook
5. how to tie
6. how to hack
7. how to run
8. how to cite
9. how to paint
10. how to spell

Web Users

Top queries

- Live demo - WARNING this is not very safe....
- “Is it safe to”
- “Is it legal to”
- “why does”
- “why doesn’t”
- “why is”
- “why isn’t”
- “americans are”
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
True Example *

The User
flickr:crankyT

Stop the noisy fan in the courtyard

Task

Info about EPA regulations

Info Need

What are EPA rules on noise pollution?

Verbal Form

EPA Sound Pollution

Query

“To Google or to GoTo” Business Week Online 9/28/2001
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

**Web Users**

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%)

Number of Responses
Overview

• Introduction
• Classic Information Retrieval
• Web IR
• Sponsored Search
• Web Search Basics
  • Size of the Web
• Web Users
  • Helping the User
• Spam
Helping the user

True Example *

The User

Stop the noisy fan in the courtyard

Task

Info about EPA regulations

Info Need

What are EPA rules on noise pollution?

Verbal Form

EPA Sound Pollution

Query

“To Google or to GoTo” Business Week Online 9/28/2001
Answering “the need behind the query”

• The query is often an imprecise indicator of what the user really wants
• What can we do to get a better handle on the underlying information need?
• Query language
  • Adjust rank of English results for a Japanese query
• Use user context
  • In particular geographic context
Answering “the need behind the query”

- Guess what type of information the user wants
  - a web page?
  - a map?
  - a stock price?
  - what else?

- Correct queries
  - Suggest correct spellings
  - Suggest related searches (google-fu)
Helping the user

Examples - language

Google search for "kitakurihama" showing results including "kujukurihama" as a suggested alternative. Some links included are:
- MySpace.com - Nasty - Kitakurihama, Kanagawa - Hip Hop - www ...
- Kitakurihama
- YuDiary
Helping the user

Examples - query spelling

Google search for "informatics" with results indicating a typo. The search engine suggests the correct spelling "informatics" and provides results that include spellings like "informatics", "informatics", and "Auditing international informatics - BLogcu".
Helping the user

Examples - query expansion

Live Search: rock

Web results 1-10 of 441,000,000
See also: Images, Video, News, Maps, MSN, More

The Rock - Jellyfish.com/Movies
Shop and compare prices across stores. Rebates up to $1.10.

Rock.com® - The Official Site of Rock Music®
Free Email & Free Music Downloads by Rock.com. Get Free Internet Radio, Music Downloads & more. Get your free @rock.com email today
www.rock.com · Cached page

Rock music - Wikipedia, the free encyclopedia
Rock music is a form of popular music with a prominent vocal melody accompanied by guitar, drums, and bass. Many styles of rock music also use keyboard instruments such as organ...
en.wikipedia.org/wiki/Rock_music · Cached page

Rock and roll - Wikipedia, the free encyclopedia
Rock and roll is a genre of music that evolved in the United States in the late 1940s and early 1950s, and quickly spread to the rest of the world.
en.wikipedia.org/wiki/Rock_and_roll · Cached page
Show more results from en.wikipedia.org

Rock - Australia's Climbing Magazine
Query shortcuts

• Map: “irvine, ca 92614”
• Calculation: “5+4”
• Flight Info: “american airlines 715”
• Stock price: “msft”
• Unit conversion: “1 dollar in euro”
• Music: “White Stripes”
Helping the user

Examples - query shortcuts
Helping the user

Examples - query shortcuts
Examples - query aggregations