ECE 498KL: eCrime and Internet Service Abuse

Web Search

Kirill Levchenko
November 8, 2018
Reading Questions

❖ How does a search engine know what is on your site?
❖ How does a search engine determine which sites to show in response to a query?
❖ How does a search engine determine a site’s position on a result page?
ECE 498 KL - eCrime and Internet Service Abuse
https://courses.engr.illinois.edu/ece498kl/fa2018/
Dec 6, TBA. Dec 11, TBA. Dec 14, Final Exam, Time: 8 A.M.. Location: ECEB 2015. These important Web sites are part of Assignment 1: Trimpuny Student Pages.

Trimpuny
studi... Trigwyny.org/
Future of Trimpuny - The future of Sales. Trimpuny is like nothing you've ever sold before. It practically flies off the shelf. Additionally, it is available in two ...

Kingston Daily Gleaner Newspaper Archives, Feb 16, 1888
https://newspaperarchive.com/jm/kingston/kingston/kingston-daily.../02.../page-1/  
Pillow Linens 7 UH Til* Northern Assurance trim puny af London and Aberdeen. LINEN huckabacks Linen Tolis ti CL Towels, worh la each, at As per dos, _ ' -a.

Arkansas City Traveler Archives, Nov 24, 1960, p. 16
... Style* Sixes IO to 18 Ladies' Quilted Robes # White, Rose and Blue SPECIAL SALE it” rn 0 Rayon with Lurex metallic trim Puny Cat Bow Tie, Floral trim motif.
Future home of Trimpuny - The future of Sales

Trimpony is like nothing you've ever sold before. It practically flies off the shelf. Additionally, it is available in two attractive colors.

- Green
- Light Green

All of your coursework to date has taught you how to design and build computer systems to do useful work. Along the way, whether you realized it or not, you made certain assumptions about how your computer system will be used. Unfortunately, many of the systems we build can also be used in unintended ways. Miscreants can exploit vulnerabilities in the design and implementation of computer systems for profit. In this class, you will learn how computer systems can be abused for illicit financial gain and how to anticipate and protect against such unintended uses.
Crawling

❖ How did Google know about Trimpuny?
❖ Search engines periodically crawl (visit) every page
❖ Crawler will download page content
  • May download images and other media
  • Normally will not execute JavaScript
    • Crawler is not a full browser
    • Executing JavaScript is resource-intensive
2. Google Indexing Factors

From my experience, factors that contribute to quicker crawling and indexing revolve around:

1. **Domain authority**: Score (on a 100-point scale) developed by Moz that predicts how well a website will rank on search engines.

2. **Page authority**: Score (on a 100-point scale) developed by Moz that predicts how well a specific page will rank on search engines.

3. **Content schedule**: The frequency at which you publish new content on your website.

4. **Popularity of website**: A combination of site traffic, CTR and time on site can contribute to quicker crawling and indexing.
Crawling

- *How did Google know about the Trimpuny page?*
- Start with a seed set of pages, follow links
- How do you come up with seed set?
Crawling

- Mentions in email, private messages, etc.
- Contact every IP address on port 80 and 443?
Crawling

- Site owner can restrict what crawler sees
- Robots exclusion standard defines how to do this
- `robots.txt` file tells search engine what it can do

```
User-agent: *
Disallow: /cgi-bin/
Disallow: /tmp/
Disallow: /junk/
```
Crawling Challenges

❖ Finding new content
❖ Staying up-to-date
  • How often do you re-crawl?
❖ Deep web
  • Pages behind forms (no direct links)
  • Pages requiring registration or payment
Indexing

- How did Google know to show Trimpuny in results?
- Search engines create an index of all Web pages
- Index maps words to pages containing them
  - Some words are not indexed (e.g. articles “a” and “the”)
  - May require language-specific processing
- Identify all pages containing search terms quickly
Indexing Challenges

- **Performance**: want to identify relevant pages quickly

- **Which terms to index?**
  - Language-specific processing
  - Stop-word elimination
  - **Word stemming** (remove inflection)
    - e.g. querying “cars” should also retrieve pages containing “car”
Query Processing

- Basic search engine query is a list of terms
- Identify pages *relevant* to query using index
- Some search engines allow modifiers and constraints
  - **AND** vs **OR** (require *all* or *some* of the terms to be present)
  - Exclude certain terms (e.g. should not contain term $X$)
  - Ordering constraints (e.g. term $Y$ before $Z$)
- Web search engines avoid complex query languages
Query Ranking

- **In what order are results shown?**
  - This is the most important question is search

- **Ultimate goal:** make users happy

- Users like *relevant* results

- Users like *authoritative* results

- Users hate *spam*
Query Relevance

- Pages may contain only **some** of the search terms
  - More distinct terms on page—more relevant?
  - More term occurrences on page—more relevant?

- Terms may appear in different **roles** and **places** in page
  - Terms in title or section headings—more relevant?
  - Terms in metadata—more relevant?
  - Terms in bold or larger font size—more relevant?
Query Relevance

- **External endorsement:** query terms in anchor text

  <A HREF="http://student1.sigpwny.org">Trimpuny</A>

- Associates target page with anchor term
Page Rank

❖ Relevance alone is not sufficient to get useful ranking

❖ PageRank ranks page based on number of links to page
  • Linking seen as *endorsement* of target page by linking page

❖ More links — higher target page rank

❖ Higher linking page rank — higher target page rank
Page Rank

- Classic PageRank is *independent* of query
- Rank of each page depends only on Web link graph
  - **Nodes**: Web pages
  - **Directed edge** from \( A \) to \( B \): page \( A \) links to page \( B \)
- Rank of page derived from rank of pages linking to it
  - Start with seed set of high-rank sites, find fixed-point
Result Ranking

- Result rank is combination of content rank (relevance) and page rank (authority)
- Many ways to combine content and page rank
  - Page rank may be query-dependent in modern search engines
  - Page rank may be content-dependent in modern search engines
- Today’s search engine use closely-guarded formula
  - The “secret sauce” of the search engine
- Search engine rank has tremendous commercial value
- Most users don’t go to page 2
- Top results clicked more often
U of I Car Pool
No reviews · Car Rental Agency
1701 S Oak St · (217) 333-3910
Open · Closes 10:30PM

Enterprise Rent-A-Car
4.3 ★★★★★ (94) · Car Rental Agency
1804 S Neil St · (217) 351-1400
New Cars, Used Cars, Car Reviews and News | Cars.com
https://www.cars.com/
Research and compare cars, find local dealers/sellers, calculate loan payments, find your car's value, sell or trade your car, get a service estimate, and much ...
Cars for Sale · Videos & Reviews · Car Dealers · Car Reviews

Top stories

This Is A Real Car That Kept Its Engine in a Box in the Trunk
Jalopnik
23 hours ago

Forget lap times; this car control class makes teen drivers safer
Ars Technica
1 day ago

Should you upgrade your car to HID headlights?
CNet
17 mins ago

→ More for car

Used Cars for Sale Online | Cars.com
https://www.cars.com/shopping/
Browse cars for sale on Cars.com. Shop the best deals near you on popular brands and body styles.
Used Cars for Sale Online | Cars.com
https://www.cars.com/shopping/
Browse cars for sale on Cars.com. Shop the best deals near you on popular brands and body styles. Get news and advice on car shopping, including current ...

Used Cars for Sale - CarMax
https://www.carmax.com/cars
Search for new and used cars at carmax.com. Use our car search or research makes and models with customer reviews, expert reviews, and more.

Car.com - We Do the Research, You Do the Driving
https://www.car.com/
Car.com is for people who need help finding the perfect vehicle. Choosing from thousands of cars is really confusing, we have the tools to help you make a ...

Car and Driver: New and Used Car Reviews, Car News and Prices
https://www.caranddriver.com/
Research new car reviews and car buying resources at Car and Driver. Our new car reviews and car buying resources help you make a smart purchase ...

Car - Wikipedia
https://en.wikipedia.org/wiki/Car
A car (or automobile) is a wheeled motor vehicle used for transportation. Most definitions of car say they run primarily on roads, seat one to eight people, have ...
The Value of Rank

Breakdown of Google Click-Through Rates in 2017 By Position

- Position 1 – 20.5%
- Position 2 – 13.32%
- Position 3 – 13.14%
- Position 4 – 8.98%
- Position 5 – 9.21%
- Position 6 – 6.73%
- Position 7 – 7.61%
- Position 8 – 6.92%
- Position 9 – 5.52%
- Position 10 – 7.95%
Search Engine Optimization

- **Search Engine Optimization (SEO):** Improving the site’s result rank in search engines
  - Improve content rank of page
  - Improve page rank of page

- **Whitehat SEO:** SEO permitted by search engines
- **Blackhat SEO:** SEO prohibited by search engines
Improving Content Rank

❖ Whitehat
  • Improve page structure
  • Add keywords metadata

❖ Blackhat
  • Keyword stuffing: place keywords for which you would like to rank in title, page body, metadata, etc.
Query Relevance

❖ Pages may contain only *some* of the search terms
  • More distinct terms on page — more relevant?
  • More term occurrences on page — more relevant?

❖ Terms may appear in different *roles* and *places* in page
  • Terms in title or section headings — more relevant?
  • Terms in metadata — more relevant?
  • Terms in bold or larger font size — more relevant?
Improving Content Rank

❖ **Whitehat**
  - Improve page structure
  - Add keywords metadata
  - **URL slugs:** e.g. https://www.vox.com/policy-and-politics/2018/11/8/18073006/jeff-sessions-firing-democracy-erosion

❖ **Blackhat**
  - **Keyword stuffing:** place keywords for which you would like to rank in title, page body, metadata, etc.
Improving Page Rank

❖ Whitehat
  • Provide useful content so that other sites link to you
    • e.g. Wikipedia

❖ Blackhat
  • Create links from other sites to your site (called backlinks)
Kingston Daily Gleaner Newspaper Archives, Feb 16, 1888
https://newspaperarchive.com/kingston-daily-gleaner-feb-16-1888-p-1/
Pillow Linens 7 Uf Till* Northern Assurance trim puny af London and Aberdeen. LINEN huckabacks
Linen Toliis ti CL Towels, worth la each, at As per dos, _ ' -a.

KWNITA 3G SmartPhone User Manual UM-V1 MobiWire SAS - FCC ID
https://fccid.io › MobiWire SAS › KWNITA ▼
Ilpliflon dilly" 011cm mun be mdmlnd between he ml": body Ind (M Mndnh hfludnfl m IMlIIMI Trim-puny
hon—clam uni-urn m dm!lumcmIlnq m- Arm-mu. Tummy ...

December-1943 to April-1944 › Page 48 - Fold3.com
... alterations can be acceonlxdied! ;; during this refit specially: conversion of 4A and 413 to fuel:\ ballast
and installation of new type centri¬ fugal trim puny).

Greensboro Daily News from Greensboro, North Carolina on March 19 ...
https://www.newspapers.com/newspage/71492861/ ▼
Sunday, March 19, 1911 AY MARCH 19. J9l1 1114:1 "MILL TO MAN- ii. rm tttt tt t a rm tt ck t t a tt tt t
tk ck it m t LI ILLS n i: t T M 0 The largest dealers in ...

Tyrone Daily Herald from Tyrone, Pennsylvania on September 19 ...
https://www.newspapers.com/newspage/14266210/ ▼
Trimpuny
student1.sigpwny.org/
Future home of Trimpuny - The future of Sales. Trimpony is like nothing you've ever sold before. It practically flies off the shelf. Additionally, it is available in two ...

Tyrone Daily Herald from Tyrone, Pennsylvania on September 19 ...
https://www.newspapers.com/newspage/14266210/
Tuesday, September 19, 1905 -Ki»*t tyivme Conhll No. Amnrln&n&l«lwatc* lIKAAt, CAtttAt I . HAM* l«, Knights of l'ythUia, meet* nt A. PI n't ic IJ'tm.mxri-f ywne ...

The Weekly Sentinel from Raleigh, North Carolina on December 15 ...
https://www.newspapers.com/newspage/67749938/ Tuesday, December 15, 1874 he J nprorlaUon is made, for additions annronrta- I line sxMvle bands, act of Congrrs I PH. WXtl. TT ft Tali's THE SENTINEL!

Trimpunt.nl - 't Trimpunt - hondentrimsalon uit Kesteren | QanAtor
https://vuinch.com/2938-trimpunt.nl.html

[PDF] Untitled - The Keep
media.thekelke.info/gb179/HASTINGS%20OBSERVER_19160205.pdf
-d# (trimpuny of. Gi ri) Gafdltn are busy eon- ertilrig t lir eur b n ora into a Hma|l emar-. I ehcy win'fl and ; ( T(<aing Jstojijn. Evyry- ihlng Ha# henn mud" as far ...
<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dec 6</td>
<td>TBA</td>
</tr>
<tr>
<td>Dec 11</td>
<td>TBA</td>
</tr>
<tr>
<td>Dec 14</td>
<td><strong>Final Exam</strong></td>
</tr>
</tbody>
</table>

These important Web sites are part of Assignment 1: [Trimpuny](#)
Discussion:
Is Blackhat SEO illegal?