Challenges in Running a Commercial Web Search Engine

Amit Singhal
Overview

- Introduction/History
- Search Engine Spam
- Evaluation Challenge
- Google
Introduction

• **Crawling**
  – Follow links to find information

• **Indexing**
  – Record what words appear where

• **Ranking**
  – What information is a good match to a user query?
  – What information is inherently good?

• **Displaying**
  – Find a good format for the information

• **Serving**
  – Handle queries, find pages, display results
History

• The web happened (1992)
• Mosaic/Netscape happened (1993-95)
• Crawler happened (1994): M. Mauldin
• SEs happened 1994-1996
  – InfoSeek, Lycos, Altavista, Excite, Inktomi, …
• Yahoo decided to go with a directory
• Google happened 1996-98
  – Tried selling technology to other engines
  – SEs though search was a commodity, portals were in
• Microsoft said: whatever …
Present

- Most search engines have vanished
- Google is a big player
- Yahoo decided to de-emphasize directories
  - Buys three search engines
- Microsoft realized Internet is here to stay
  - Dominates the browser market
  - Realizes search is critical
History

• Early systems Information Retrieval based
  – Infoseek, Altavista, …

• Information Retrieval
  – Field started in the 1950s
  – Primarily focused on text search
  – Already had written-off directories (1960s)
  – Mostly uses statistical methods to analyze text
History

- IR necessary but not sufficient for web search
- Doesn’t capture authority
  - Same article hosted on BBC as good as a slightly modified copy on john-doe-news.com
- Doesn’t address web navigation
  - Query *ibm* seeks **www.ibm.com**
  - To IR **www.ibm.com** may look less topical than a quarterly report
History

• But there are links
  – Long history in citation analysis
  – Navigational tools on the web
  – Also a sign of popularity
  – Can be thought of as recommendations (source recommends destination)
  – Also describe the destination: anchor text
History

• Link analysis
  – Hubs and authority (Jon Kleinberg)
    • Topical links exploited
    • Query time approach
  – PageRank (Brin and Page)
    • Computed on the entire graph
    • Query independent
    • Faster if serving lots of queries
  – Others…
History

- Google showed link analysis can make a huge difference and is practical too
  - Everyone else followed
- Then there is the secret sauce
  - Link analysis
  - Information retrieval
  - Anchor text
  - Other stuff
History

• Interfaces

  – Many alternatives existed/exist
    • Simple ranked list
    • Keywords in context snippets (Google first SE to do this)
    • Topics/query suggestion tools (e.g. Vivisimo, Teoma)
    • Graphical, 2-D, 3-D

  – Simple and clean preferred by users
    • Like relevance ranking
    • Like keywords in context snippets
End Product

• As of today
  – Users give a 2-4 word query
  – SE gives a relevance ranked list of web pages
  – Most users click only on the first few results
  – Few users go below the fold
    • Whatever is visible without scrolling down
  – Far fewer ask for the next 10 results
Overview

- Introduction/History
- Search Engine Spam
- Evaluation Challenge
- Google
Oh No … This is REAL

- 80% of users use search engines to find sites

Figure 1: Online Tenure Impacts Search Use And Effectiveness

**Search use increases with online tenure**
“How have you typically found Web sites you have visited in the past month?”

Base: North American consumers (multiple responses accepted)
Enter the Greedy Spammer

- Users follow search results
- Money follows users, spam follows …
- There is value in getting ranked high
  - Affiliate programs
    - Siphon traffic from SEs to Amazon/eBay/…
      - Make a few bucks
    - Siphon traffic from SEs to a Viagra seller
      - Make $6 per sale
    - Siphon traffic from SEs to a porn site
      - Make $20-$40 per new member
Big Money

• Let’s do the math
• How much can the spam industry make by spamming search engines?
  – Assume 500M searches/day on the web
    • All search engines combined
  – Assume 5% commercially viable
    • Much more if you include porn queries
  – Assume $0.50 made per click (from 5c to $40)
  – $12.5M/day or about $4.5 Billion/year
How?

- Defeat IR
  - Keyword stuffing
  - Crawlers declare that it is a SE spider
  - They dish us an “optimized” page
But that should be easy…

• Just detect keyword density
But that is easy too…

- Just detect that page is not about query
Legitimate NLP Parse

- Noun phrase to noun phrase

Taking Control

Once you recognize that herbs for migraine has an impact on your life, it is important to take control of managing your herbs for migraine. The first step is to schedule a specific appointment with your health care provider to discuss herbs for migraine. For a proper herbs for migraine diagnosis, illness severity assessment, and treatment plan, an appointment dedicated to herbs for migraine is necessary.

Taking Control

Once you recognize that Ocular Migraine Stroke has an impact on your life, it is important to take control of managing your Ocular Migraine Stroke. The first step is to schedule a specific appointment with your health care provider to discuss Ocular Migraine Stroke. For a proper Ocular Migraine Stroke diagnosis, illness severity assessment, and treatment plan, an appointment dedicated to Ocular Migraine Stroke is necessary.
But links should help…

- No one should link to these bad sites
  - Expired domains
    - The owner of a legitimate domain doesn’t renew it
    - Spammers grab it, it already has tons of incoming links
    - E.g., anchor text for
      - The War on Freedom
      - The War on Freedom: How and Why America was attacked
      - The War on Freedom
Get Links

Guestbooks
Get Links

Mailing lists

W3C home > Mailing lists > Public > html-tidy@w3.org > October to December 1999

Natural Viagra Alternative! (Adv.)

This message: [ Message body ] [ Respond ] [ More options ]
Related messages: [ Next message ] [ Previous message ]

From: <Viagra@weeklysex.com>
Date: Sat, 11 Dec 1999 17:27:02 +0000
To: <html-tidy@w3.org>
Message-Id: <16270281811057@napolionline.com>

The Ultimate Sensation!

Get your "Men's Ultra" Now, No Doctor's Prescription needed! 100 % Natural Special Sale Price $39.95 - 60 Caplets

Includes Shipping & Handling "30 Day Money Back Guarantee" Credit Cards, E Check

Received on Saturday, 11 December 1999 17:26:03 EST
Get Links

Link Exchange

related links

Accommodation Links

Rest of World

Accommodation Directories

Armchair Travel - Plan your next holiday, accommodation or travel arrangements
Cape Car Rentals - Your online resource for Car Rentals!
Pamela Lanier’s Bed and Breakfast Guide Online - Free searchable database of over 58,000 quality Bed & Breakfast Inns, Small Hotels, Golf Courses, Family Travel Accommodations and more!
www.alleuropesaccommodation.com
Redflag - Chinese hotel travel guide
www.1stvacations.com
www.allcheaphotels.net
Discount Hotels - Discount Hotel Reservations Worldwide - find great discounts for all hotels worldwide
www.amsterdamhotels-discount.com
www.1st-berlin-hotels.com
www.aboutanyhotel.com
www.travelwand.com
travelchiapas.com - Travel Information for Chiapas Mexico
State of Affairs

• There is big money in spamming SEs
• Easy to get links from good sites
• Easy to generate search algorithm friendly pages
• Any technique can be and will be attacked by spammers
• Have to make sense out of this chaos
We counter it well

- Most SEs are still very useful
  - Used over 500 million times every day
    - All search engines put together

- Our internal measurements show that we are winning

- Still need to be watchful
And then…
Overview

- Introduction/History
- Search Engine Spam
- Evaluation Challenge
- Google
Information Retrieval

• Test collection paradigm of evaluation
  – Static collection of documents (few million)
  – A set of queries (around 50-100)
  – Relevance judgments
  – Extensive judgments not possible (100x1,000,000)
  – Use pooling
    • Pool top 1,000 results from various techniques
    • Assume all possible relevant documents judged
    • Biased against revolutionary new methods
      – Judge new documents if needed
On the Web

- Collection is dynamic
  - 10-20% urls change every month
  - Spam methods are dynamic
  - **Need to keep the collection recent**
- Queries are also time sensitive
  - Topics are hot then not
  - Need to keep a representative sample
On the Web

- **Search space is HUGE**
  - Over 200 million queries a day
  - Over 100 million are unique
  - Need 2700 queries for a 5% (700 for 10%) improvement to be meaningful at 95% confidence

- **Search space is varied**
  - Serve 90 different languages
  - Can’t have a catastrophic failure in any
  - Monitoring every part of the system is non-trivial

- **IR style evaluation**
  - Incredibly expensive
  - Always out of date
On the Web

• But what about user behavior?
  – You can use clicks as supervision.

• Clicks
  – Incredibly noisy
  – A click on a result does not mean a vote for it
    • The destination may just be a traffic peddler
    • User taken to some other site
    • If anything, this (clicked) result was BAD
Blue and Gold Fleet

Homeport
Blue and Gold Fleet, exclusive provider of transportation to Alcatraz Island, is a passenger vessel company and tour excursion operator providing vessel ...
www.blueandgoldfleet.com/ - 11k - Cached

Blue & Gold Fleet
Main Menu: Blue & Gold Fleet ... Blue & Gold Fleet also operates the Alameda/Oakland Ferry and the Vallejo Baylink Ferry. Schedules ...
198.94.156.9/BlueGoldFleet/ - 4k - Cached

Blue & Gold Fleet Bicyclist's Guide
Blue & Gold Fleet Bicyclist's Guide. Main Menu: Blue & Gold Fleet
: Bicyclist's Guide You can find out more about bicycling in the ...
198.94.156.9/BlueGoldFleet/bicycles.html - 3k - Cached
[ More results from 198.94.156.9 ]

Transit 511.org
... Alameda/Oakland Ferry Alatamont Commuter Express (ACE) American Canyon Transit Angel Island/Tiburon Ferry BART Benicia Transit Blue & Gold Fleet Brentwood Dimes ...
www.transitinfo.org/providers/coverage.asp?mc=InitialMap&cid=BG - 36k - Cached

San Francisco Blue and Gold Fleet
thespacebetween.tvopad.com/photos/san francisco/ blue and gold fleet.html - 10k - Cached
We do Very Well

• Continually evaluate our system
  – In multiple languages
  – Tests valid over large traffic
  – Caught many possible disasters

• Constantly launch changes/products
  – Stemming, Google News, Froogle, Usenet, …
Overview

- Introduction/History
- Search Engine Spam
- Evaluation Challenge
  - Finding Needles in a 20 TB Haystack, 200M times per day
Past

1995 research project at Stanford University
Lego Disk Case

One of our earliest storage systems
Peak of google.stanford.edu
Growth

- Nov. 98: 10,000 queries on 25 computers
- Apr. 99: 500,000 queries on 300 computers
- Sept. 99: 3M queries on 2,100 computers
Servers 1999
Datacenters now

And 3 days later…
Where the users are...
What can we learn...

- Structure of Web
- Interests of Users
- Trends and Fads
- Languages
- Concepts
- Relationships
Spelling Correction: Britney Spears
Google

• Ethics
  – No pay for inclusion (in index)
  – No pay for placement (in ranking)
  – Clearly demarked results and ads
  – 20% engineer time doing random stuff
    • Out came news, froogle, orkut
  – Users come first
Recent launches...

[Google search results for 'cert advisory']

CERT Coordination Center
... January 29, 2004 Changes to CERT Advisories CERT advisories have become a core component of US-CERT’s Technical Cyber Security Alerts. ...
Description: The CERT Coordination Center studies Internet security vulnerabilities, provides incident response...
Category: Computers > Security
www.cert.org/ - 29k - Cached - Similar pages

CERT® Advisories
... self-propagating malicious code that exploits certain configurations of Microsoft Windows susceptible to the vulnerability described in CERT advisory CA-2001 ...
www.cert.org/advisories/ - 101k - Cached - Similar pages
[ More results from www.cert.org ]
Recent launches...

1 acre = 43 560 sq feet

Welcome to the Harvard Illinois Economic Development Corporation...
... commercial, retail, 1 Acre, 7,000 SQ FT, B-3, N/A, N/A, Details: commercial.

CERT - Computer Emergency Response Team

HB 455 - Peace officers; define; incl cert co probation officers
Some perks…
Our Chef Charlie...
Thank You...

Amit Singhal