ifis

Institut für Informationssysteme
Technische Universität Braunschweig

Information Retrieval and Web Search Engines

Lecture 13: Miscellaneous

Wolf-Tilo Balke
Muhammad Usman

Institut für Informationssysteme
Technische Universität Braunschweig

Miscellaneous

1. **Spamdexing**
2. Hardware for Large Scale Web Search
3. Metasearch
4. Privacy Issues

Spamdexing

- Spamdexing = The practice of **modifying the Web** to get certain **Web resources unjustifiably ranked high** on search engine result lists
- Often a synonym of SEO ("search engine optimization")

SPONSOR RESULTS

Seo
How to Generate Endless Leads & Instant Cash for Any Business.
EasyIncomeInstantly.com

Need Website Traffic I Will Help Deliver
Targeted Traffic Guaranteed Sign Ups Email Campaigns, Surveyed Leads.
www.**webtrafficus.com**

Search Engine Optimization - SEO Marketing
Search engine optimization, SEO, and **search engine** marketing guide. ... With **Search Engine Optimization** strategies being applied by more and more ...
www.**lilengine.com** - 62k - Cached

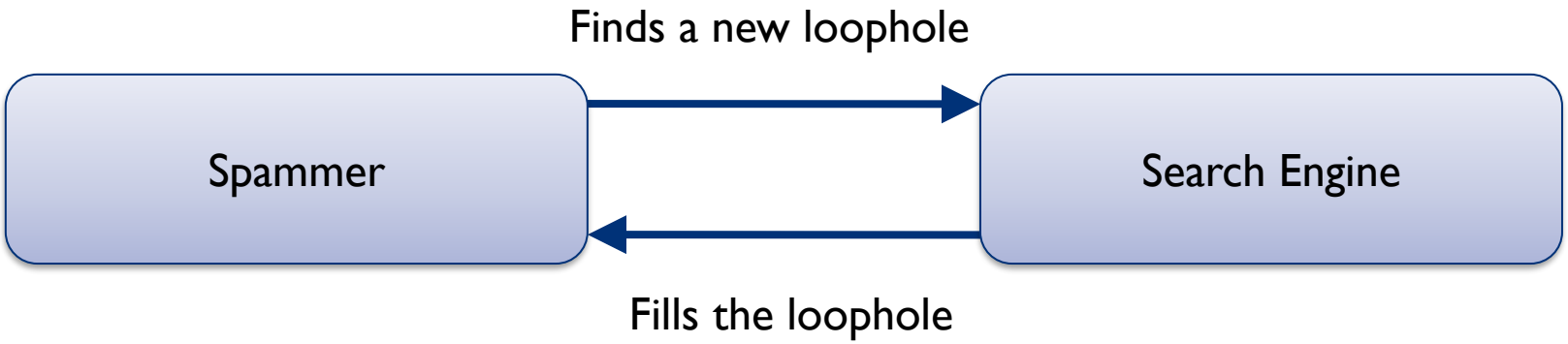
Category:**Search engine optimization** - Wikipedia, the free encyclopedia
Articles on people should be categorized in the "consultants" subcategory.
en.wikipedia.org/wiki/Category:**Search_engine_optimization** - Cached

Search Engine Optimization - Natural **Search** Placement & SEO Services ...
Offers a full suite of **search engine optimization** services that includes link popularity building, natural **search engine optimization**, and site analytics. SEO ...
www.**submitawebsite.com** - 72k - Cached

Search Engine Optimization Insights & SEO 2.0 Marketing by Blackwood.
Search Engine Marketing and **Optimization**, SEO tools and techniques by the industry's leading innovators @ BlackwoodProductions.com.
www.**blackwoodproductions.com** - Cached

Spamdexing

- Spamdexing usually means **finding weaknesses** in ranking algorithms and **exploiting them**
- Usually, it looks like this:



- There are two classes of spamdexing techniques:
  - **Content spam:** Alter a page's contents
  - **Link spam:** Alter the link structure between pages

Content Spam

Idea:

- Exploit TF–IDF

Method:

- Repeatedly place the keywords to be found in the text, title, or URI of your page
- Place the keywords in anchor texts of pages linking to your page
- Weave your content into high-quality content taken from (possibly a lot of) other pages

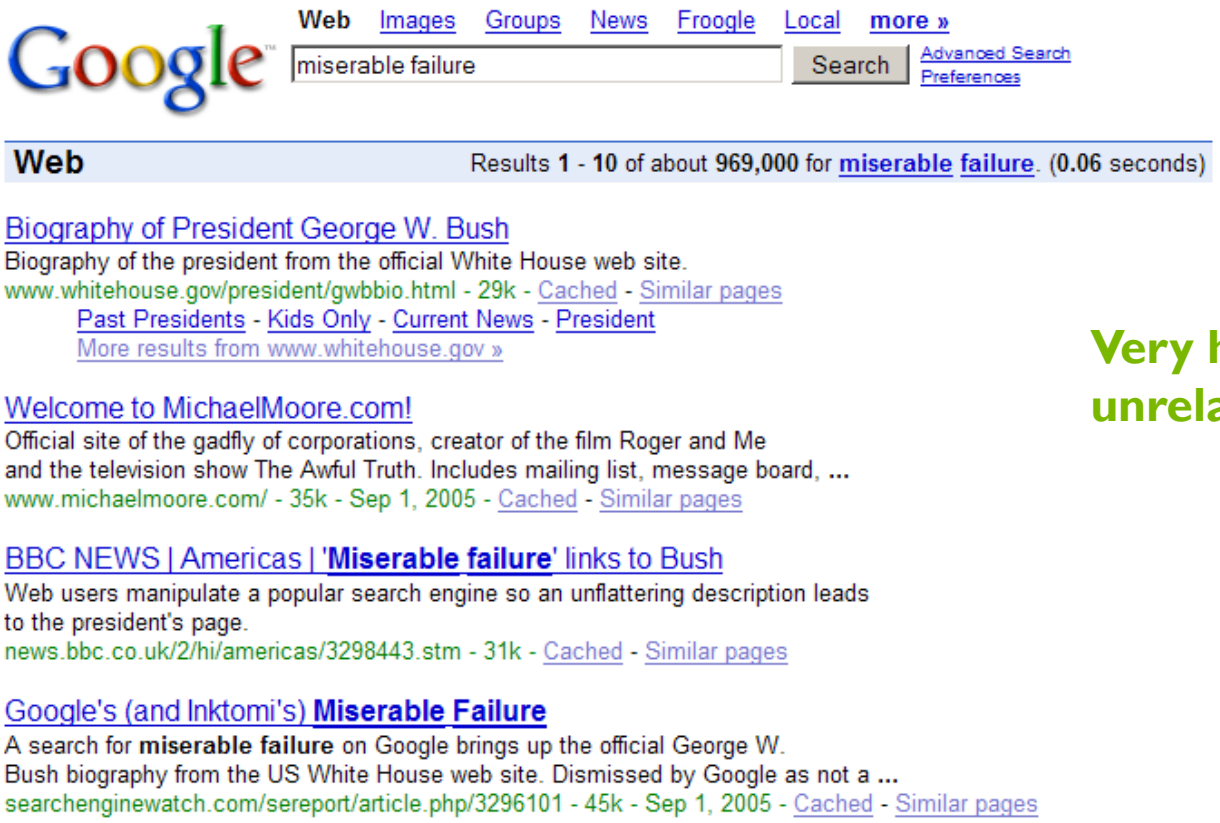
Countermeasures:

- Train **classification algorithms** to detect patterns that are "typical" for spam pages
- Most difficult part: Find suitable **features** to describe pages
  - Degree of similarity to other pages, degree of term repetitions, …

Content Spam

Example (Google bombing):

Keywords are placed in anchor texts of pages linking to your page



Very hard to detect if many unrelated people do it…

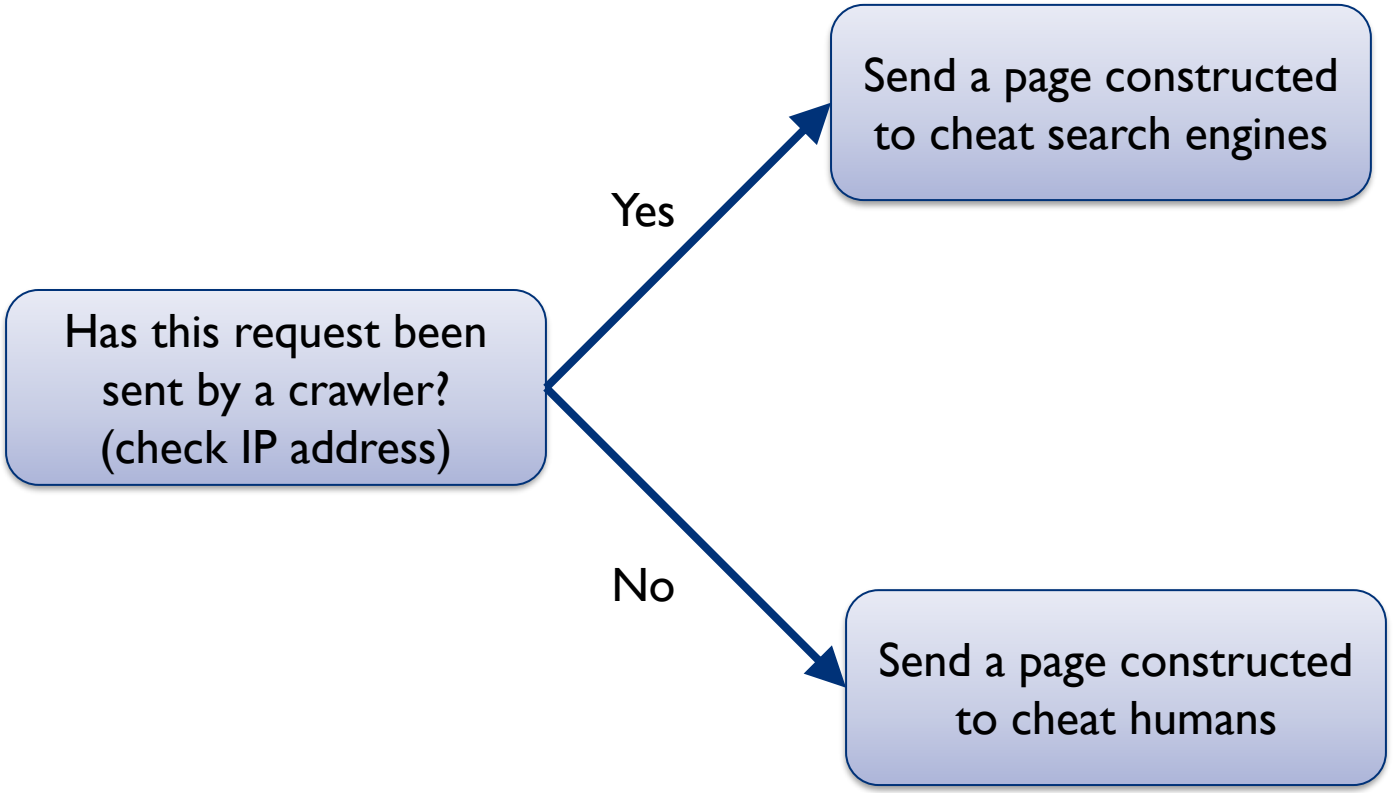
Content Spam

- There is a further way to detect content spam:
  - After a spammer has cheated the search engine, the same must be done for real users
  - Therefore, spammers try to hide the parts of their page used for spamdexing:
    - Place text behind images
    - Write text in the background color
    - Set the font size to 0
    - Dynamically delete text using scripts
    - Deliver different Web pages to Web crawlers (“cloaking”)
    - Immediately redirect to a different page (“doorway pages”)
    - …
  - Most of these techniques can be detected by search engines
    - But: This kind of analysis is quite expensive…

```
<body background=white>
<font color=white>text</font>
</body>
```

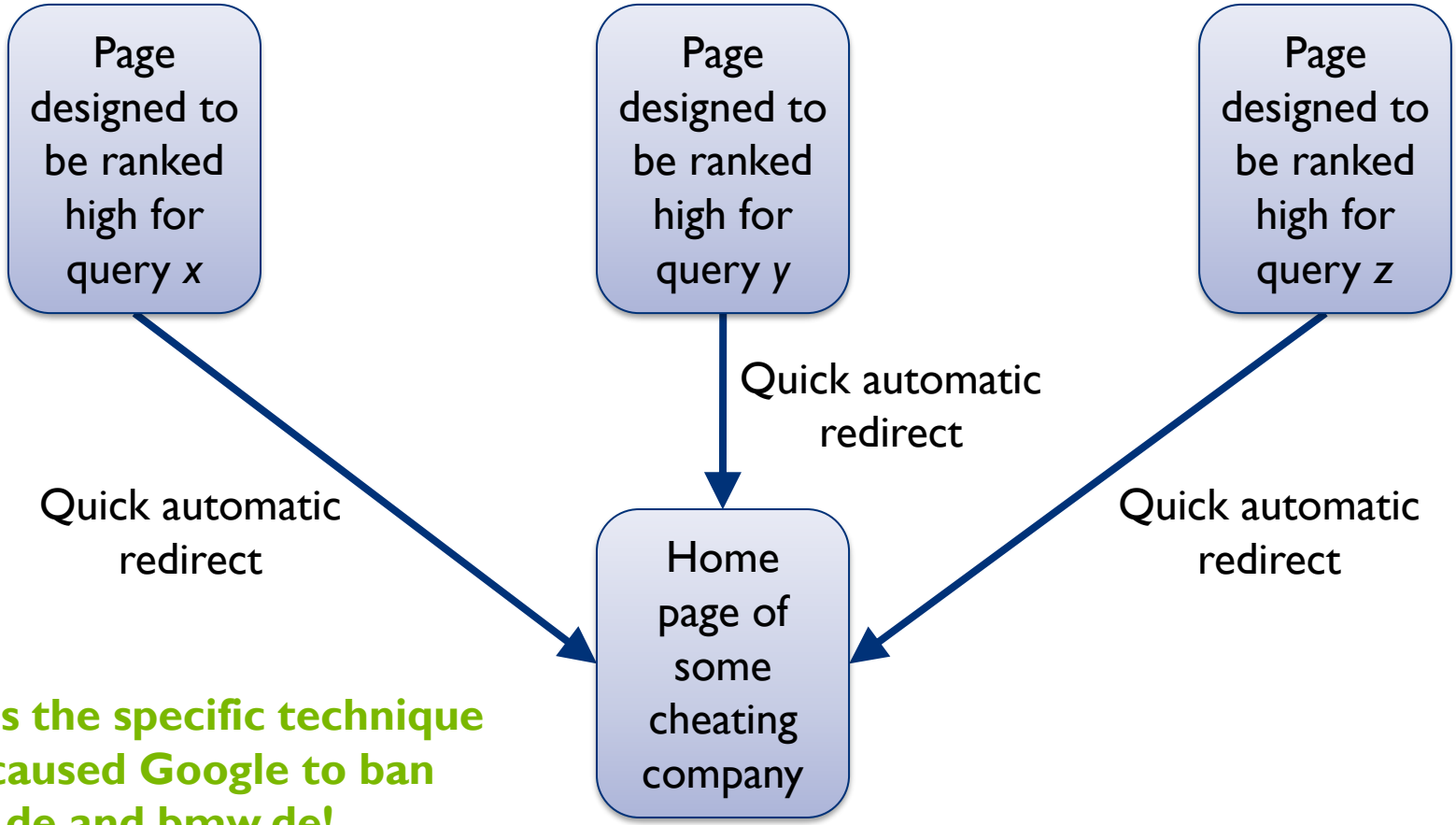
Content Spam

Cloaking:



Content Spam

Doorway pages:



This is the specific technique that caused Google to ban ricoh.de and bmw.de!

Link Spam

Idea:

- Improve your page's rank by getting in-links from other pages

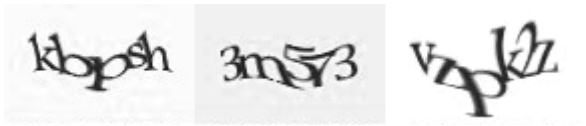
Method (comment spamming):

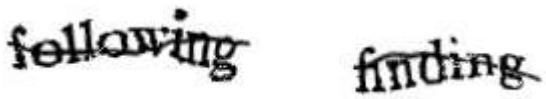
- Collect a list of high-quality sites that allow other users to post their own comments
  - Comments in blogs
  - Public forums
  - Wikis
- Write (a lot of) comments linking to your page
- This can easily be **automated** since most people use **standard software** for running their forums, blogs, …

- **Countermeasures:**
  - Require users to solve **CAPTCHAs**

Link Spam

CAPTCHAs:





- CAPTCHA: "Completely Automated Public Turing test to tell Computers and Humans Apart"
- Character recognition is easy for humans, but hard for machines

- **Countermeasures (taken by spammers):**
  - Build character recognition algorithms that are hand-tailored to the CAPTCHAs generated by standard CAPTCHA software
  - Let real humans solve CAPTCHAs (e.g. pay 1 cent per solution)

Link Spam

Method (link farms):

- Create a large group of pages that link to each other
- Or: Participate in **link exchange programs**
- Try to create link patterns that look "normal"
- Set out-links to topically related high-quality pages, which gives you high hub scores
  - This can be done e.g. by cloning directories like DMOZ
- This will consequently lead to high authority scores for your other pages

Link Spam

Method (honeypots):

- Create a set of pages (called honeypot) that provide some useful resource
  - Examples: Copies of Unix documentation pages or Wikipedia pages
- Insert hidden links to some target pages to be boosted
- This honeypot then attracts people to link to it, boosting indirectly the ranking of the target pages

Link Spam

Method (buy expired domains):

- Monitor DNS records for domains that will expire soon, i.e. whose registration has not been extended on time
- Buy such domains when they expire
- Replace their pages by pages with links to your own pages
- Using this technique you can get hold of all external links linking to the expired domain's pages

Link Spam

Countermeasures:

- In general, link spam is quite hard to detect
- **Heuristic:**
  Remove pages whose in-links look almost the same
  (can detect Google bombing and comment spamming)
- **Heuristic:**
  Remove (modified) copies of high-quality content
  (can detect honeypots)
- **Heuristic:**
  Create a white-list of pages known to be "good" and use the link-distance to these pages as an indicator of trustworthiness

“Best Practices”

- As long as you don’t want to sell Viagra or memberships in online casinos:

  Invest your time into creating good content!

- Usually, the costs of cheating search engines are higher than the benefits
  - Recall Google’s ban on bmw.de
- Therefore:
  - Create high-quality content
  - Follow the rule “link = recommendation” when creating links
  - Build crawler-friendly Web sites
  - Use “white hat” techniques like Google AdWords

SEO Contests

Detour

Hommingberger Gepardenforelle:

- Organizer: c't, German Magazine for Computer Technology.
- Goal: how search engines rank sites.
- Cut-off dates May 15-December,15 2005
- Strategies:
  - High-quality content
  - Stories on their websites
  - Illustrated sightings, Videos
  - Links and Backlinks
  - Terget terms in the domain preferred by the search engines

Schnitzelmitkartoffelsalat:

Webmaster Guidelines

Detour

- Quality Content: high-quality, informative, and original content that is relevant to your target audience.
- Use of Keywords: Proper use of keywords is encouraged, but keyword stuffing is discouraged.
- Structure and Navigation: logical site structure, making it easy for users to navigate your content.
- Avoiding Deceptive Practices: presenting different content to users and search engines
- Link Building: Guidelines on natural and ethical link building practices are provided, discouraging the use of manipulative techniques.

Miscellaneous

1. Spamdexing
2. **Hardware for Large Scale Web Search**
3. Metasearch
4. Privacy Issues

Google Servers

Detour

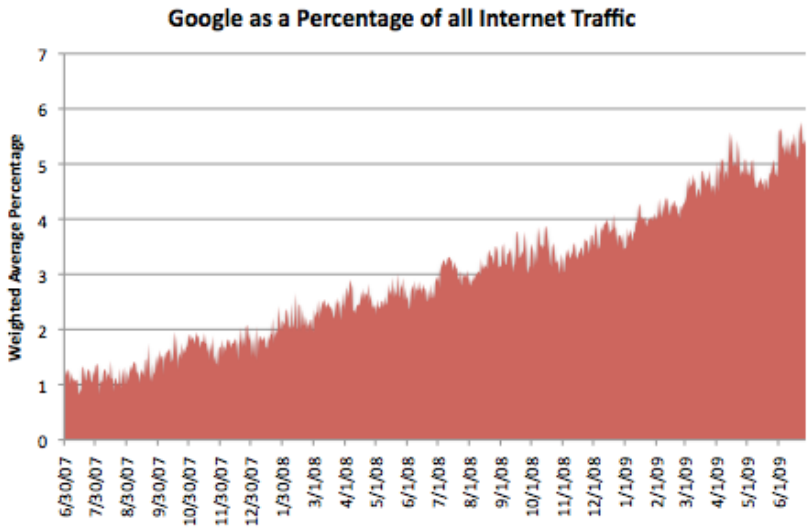
- …or how to build one of the **most powerful data centers** out of **crappy hardware**
  - For a long time, Google has jealously guarded the design of its data centers
    - In 2007 & 2009 some details have been revealed

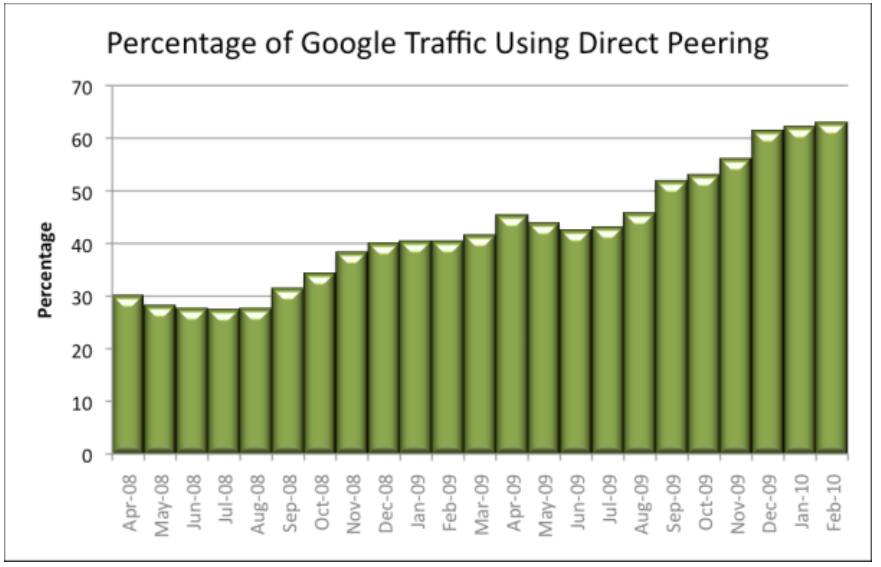
- The **Google Servers**
  - Google uses only custom built servers
  - Google is the world 4th largest server producer
    - They don't even sell servers…
    - In 2007, it was estimated that Google operates over **1,000,000 servers.**

Google Servers

Detour

- Data centers are connected to each other and major internet hubs via massive fiber lines
  - ~7% of all internet traffic is generated by Google
  - ~60% of that traffic connects directly to consumer networks without connecting to a global backbone
  - **If Google was an ISP, it would be the 3rd largest global carrier**





Google Servers

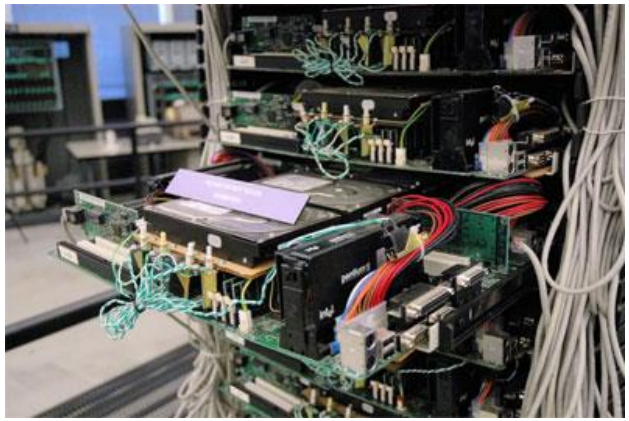
Detour

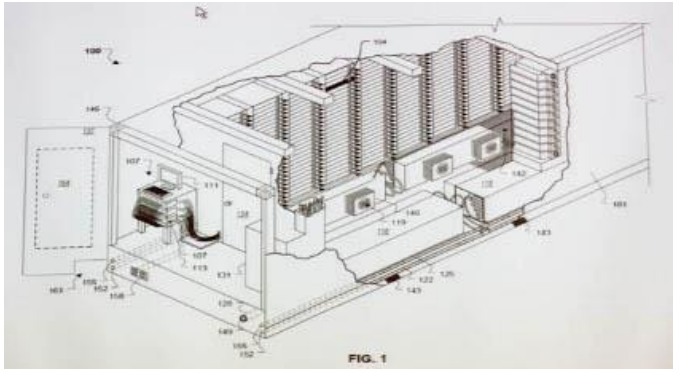
- Some Google datacenter facts and rumors
  - Construction of four new data centers (in 2007): **600 million dollars**
  - Annual operation costs in 2007: **2.4 billion dollars**
  - Energy consumption per data center: **50 to 100 megawatts**
    - The largest center in Oregon: **over 110 megawatts**
    - The whole region of **Braunschweig: 225 megawatts**

Google Servers

Detour

- Each server rack:
  40–80 **commodity class PC servers** with custom Linux
  - Slightly **outdated hardware**
  - **12V battery** to counter unstable power supplies
  - **No cases,** racks are setup in standard shipping containers and are just wired together
  - More info: http://www.youtube.com/watch?v=Ho1GEyftpmQ





Google Servers

Detour

- Google servers are highly **unstable** but also **very cheap**
  - **High "bang-for-buck" ratio**
- Typical **first year for a new cluster** (several racks):
  - ~0.5 **overheating**
    - Power down most machines in less than 5 minutes, ~1–2 days to recover
  - ~1 **PDU** (power distribution unit) **failure**
    - ~500–1000 machines suddenly disappear, ~6 hours to come back
  - ~1 **rack-move**
    - ~500–1000 machines powered down with plenty of warning, ~6 hours
  - ~1 **network rewiring**
    - Rolling ~5% of machines down over 2-day span

Google Servers

Detour

- ~20 **rack failures**
  - 40–80 machines instantly disappear, 1–6 hours to get back
- ~8 **network maintenance operations**
  - Might cause ~30-minute random connectivity losses
- ~12 **router reloads**
  - Takes out DNS and external VIPs (virtual IPs) for a couple minutes
- ~3 **router failures**
  - Traffic immediately pulled for an hour
- Dozens of minor 30-second **DNS blips**
- ~1000 **individual machine failures**
- Thousands of **hard drive failures**
- Countless **slow disks, bad memory, misconfigured machines**
- …

Google Servers

Detour

- Challenges to the data center software
  - Deal with all these hardware failures
  - Avoiding any data loss
  - Guarantee ~100% global uptime
  - Decrease maintenance costs to minimum
  - Allow flexible extension of data centers
  - **Solution:**

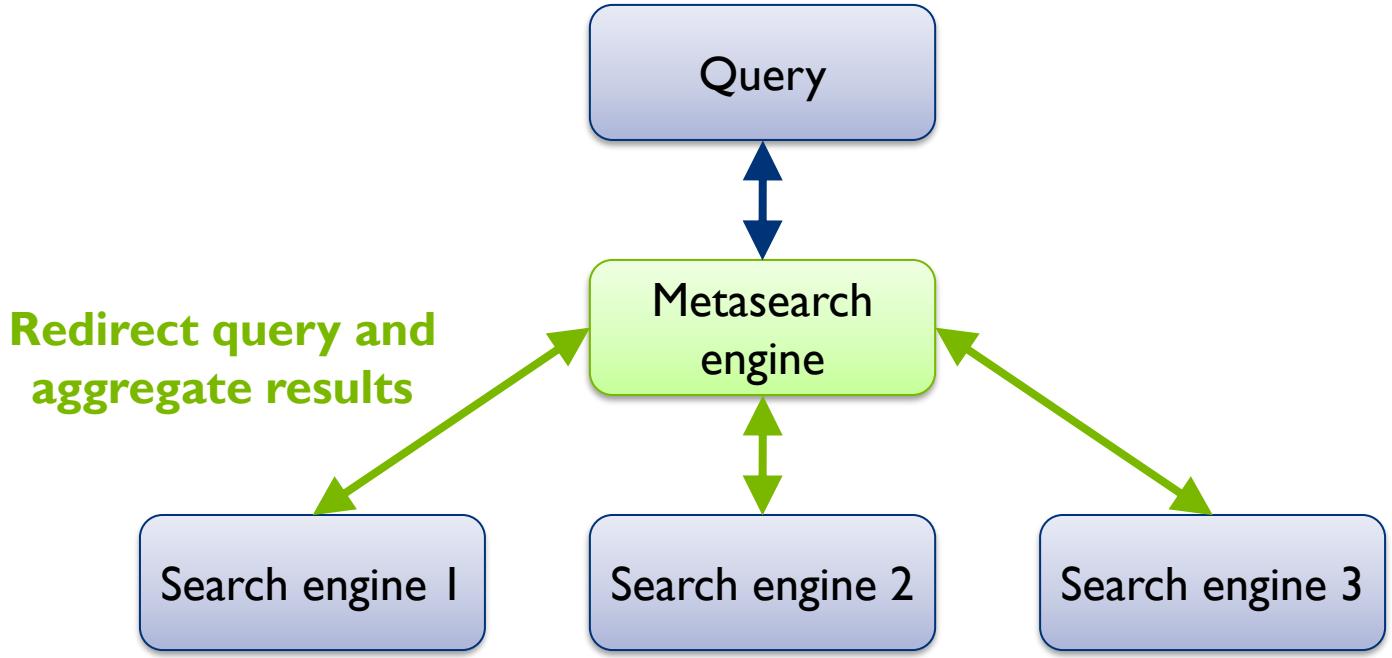
    Use **Google Cloud Platform (GCP)**
    - **GFS** (Google File System), **Google Big Table, BigQuery,…**
- More on Google Data Centers:
  - http://www.google.com/about/datacenters/

Miscellaneous

1. Spamdexing
2. Hardware for Large Scale Web Search
3. **Metasearch**
4. Privacy Issues

Metasearch

- **Idea:**
  - Given access to several search engines,
    each with its individual strengths and weaknesses,
    then combining their results could improve overall result quality
- **How it works:**



Metasearch

- A metasearch engine can **only access the result lists** returned by the individual engines
- It is not able to exploit any engine's internal information

- Therefore, we have to solve the following problem:
  - **Given:** A set of *k* individual ordered result lists of size *n*
  - **Task:** Aggregate these *k* rankings into a single ranking
  - Of course, some **constraints** should hold here that define which properties a "good" aggregate should have
- This is a well-known problem from **social choice theory** having a lot of different solutions

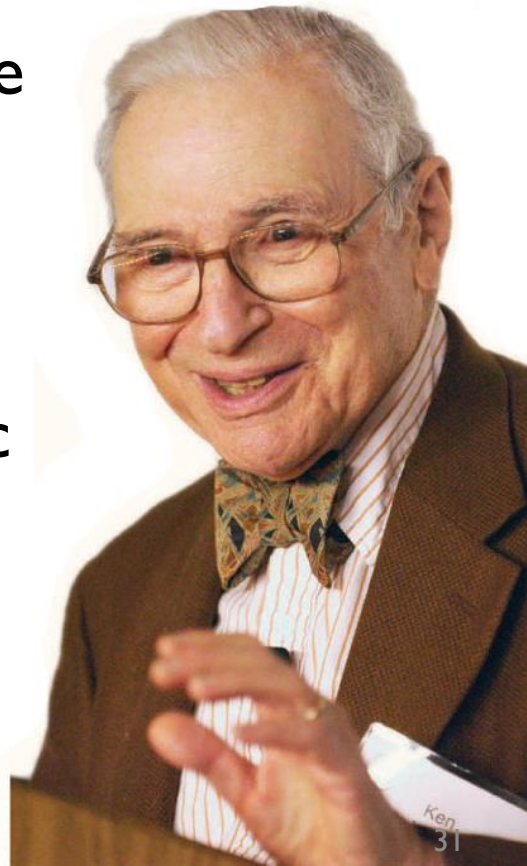
What's a Good Aggregate?

- **Pareto efficiency:**
  If every individual engine ranks a certain page higher than another, then so must the aggregate ranking
- **Non-dictatorship:**
  The aggregate ranking is not just always the same as a certain fixed individual engine's ranking
- **Independence of irrelevant alternatives:**
  If page *A* is ranked higher than page *B* in the aggregate ranking, then adding a new page *C* to each of the input rankings will not change *A*'s relation to *B*
- **Maybe some more…**

What's a Good Aggregate?

- These three constraints sound completely reasonable
- Clearly, any "reasonable" aggregation algorithm should adhere to these constraints

- In fact, "should" means "cannot" in this case
- **Kenneth Arrow's impossibility theorem (1951):**
  "If there are more than two different pages to be ranked, then there is no deterministic aggregation algorithm that satisfies Pareto efficiency, non-dictatorship, and independence of irrelevant alternatives"



What's a Good Aggregate?

- Thus, whatever method we choose to solve our aggregation problem, it will have severe weaknesses
  - Fortunately, in our case, the aggregate ranking will not be used for anything of importance, so violations are not that critical…
- There are many different aggregation methods available, two of which we will discuss briefly:
  - Majority rule
  - The Borda count
- **Let's assume that any page being ranked by at least one individual engine, is ranked by all of them**
  - In fact, this usually is not true
  - But it is possible to extend the methods to handle this problem

Majority Rule

- For any pair of pages (*a*, *b*),
  count how many search engines rank *a* higher than *b*
- If the majority of engines ranks *a* higher than *b*,
  then place *a* before *b* in the aggregate ranking
  - Ties also can be handled but let's assume that
    the number of search engines is odd
- Construct the aggregate ranking from this comparisons
- **Example:**

Engine 1	Engine 2	Engine 3
A	A	B
B	C	A
C	B	C

2 of 3: A > B
3 of 3: A > C
2 of 3: B > C

Aggregate
A
B
C

Majority Rule

- One important drawback of majority vote are **cycles**
- **Example:**

Engine 1	Engine 2	Engine 3
A	C	B
B	A	C
C	B	A

2 engines rate A > B
2 engines rate B > C
2 engines rate C > A

- There are **many methods** available to break cycles…

The Borda Count

- The **Borda count** avoids cycles
- Every engine assigns a numerical score to each page:
  - The best page gets a score of *n* (if there are n pages in total)
  - The second-best page gets a score of *n* − 1, …
- The final ranking is created by adding all scores

Engine 1	Score
A	3
B	2
C	1

Engine 2	Score
A	3
C	2
B	1

Engine 3	Score
B	3
A	2
C	1

For each page, add up its individual scores

Aggregate	Score
A	8
B	6
C	4

The Borda Count

- **Advantages** of the Borda count:
  - It is easy to compute
  - It can handle page that have not been ranked by all engines
    - E.g. assign the page a score of 0 if is has not been included in the ranking
  - It allows for ties in the aggregate ranking
  - It is easy to weight the individual engine's importance
    - Multiply the scores assigned by "good" engines by a factor larger than 1
    - Multiply the scores assigned by "bad" engines by a factor smaller than 1
- **Disadvantage:**
  - It assumes a uniform degradation of relevance in each ranking

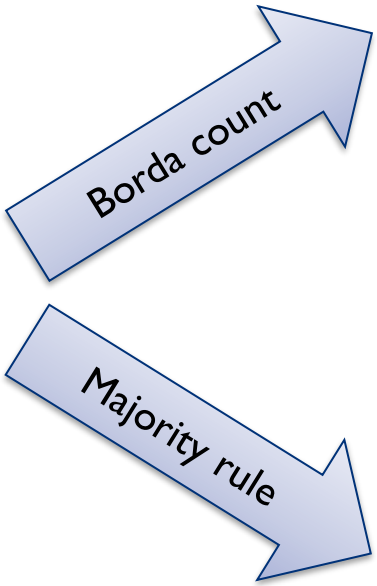
The Borda Count

Borda count vs. majority rule:

Engine 1
A
B
D
C

Engine 2
C
A
B
D

Engine 3
B
C
A
D



Aggregate	Score
A	9
B	9
C	8
D	4

2 of 3: A > B
2 of 3: B > C
2 of 3: C > A
3 of 3: A > D
3 of 3: B > D
2 of 3: C > D

Measures of Agreement

- Sometimes it is useful to **measure the agreement** between **two** search engines
  - Search engines that often yield very similar rankings should be considered as dependent
  - Therefore, they should get a lower influence at aggregation
- One of the most popular measures is **Kendall's τ:**
  - For each pair of pages (a, b) ranked by both engines, determine if both engines agree in their relative ranking or if one engine ranks a higher than b and the other ranks b higher than a
  - Basically, Kendall's τ is the ratio of agreeing pairs compared to all pairs ranked by both engines

Kendall's τ

- Define:
  - m: The number of pages ranked by both engines
  - p_+: The number of agreeing pairs of pages ranked by both engines
  - p_-: The number of disagreeing pairs of pages ranked by both engines
- Then, Kendall's τ is:

$$\tau = \frac{P_+ - P_-}{\binom{m}{2}} = \frac{2 \cdot (P_+ - P_-)}{m \cdot (m-1)}$$

- **Example:**

Engine 1	Engine 2
A	A
B	C
C	B

$m = 3$
$p_+ = 2$ (A, B) and (A, C)
$p_- = 1$ (B, C)

Therefore, $\tau = 1 / 3$

Metasearch

- Today, metasearch is well-suited for answering very special queries with **maximum recall**
- Unfortunately, it fails to increase result quality for most other queries…
- Why?
- Metasearch works best if…
  - The engines used are completely independent
  - The engines used are all of similar (high) quality
- The reality:
  - Most search engines use similar methods, thus being dependent
  - There are just a few good engines and many bad ones

Miscellaneous

1. Spamdexing
2. Hardware for Large Scale Web Search
3. Metasearch
4. **Privacy Issues**

AOL's Query Log

- On August 6, 2006, AOL tried to help IR researchers
- They released very private data about its users (without their permission)
  - 20 million Web queries from 650,000 AOL users
  - All searches from those users for a three month period
  - Whether they clicked on a result
  - Where this result appeared on the result page
- Of course, the data has been made anonymous by replacing each AOL username by a random ID number

- Unfortunately, this did not work too well…
- Let's look at some examples

AOL's Query Log

Detour

- User 311045:
  - how to change **brake pads** on **scion xb**
  - 2005 us open cup **florida** state champions
  - how to get revenge on a ex
  - how to get revenge on a ex girlfriend
  - how to get revenge on a friend who f---ed you over
  - **replacement bumper** for scion xb
  - **florida** department of law enforcement
  - crime stoppers **florida**

AOL's Query Log

Detour

- User 11574916:
  - **cocaine** in urine
  - Asian mail order brides
  - states reciprocity with **florida**
  - Florida dui laws
  - **extradtion from new york to florida**
  - mail order brides from largos
  - will one be extradited for a **dui**
  - **cooking jobs** in french quarter new orleans
  - will i be extradited from ny to fl on a dui charge

AOL's Query Log

Detour

- User 3540871:
  - i have an **interview at comcast** and i need help
  - cheap rims for a **ford focus**
  - how can i get a job in **joliet il** with a **theft** on my background
  - i need to trace a cellular location
  - i need to know if my **spouse is cheating** and i need to do a cellular trace for free
  - jobs with no background checks
  - how can i get a job with a conviction
  - motels in joliet il
  - motels in gurnee il area for under 40 dollars
  - my baby's father physically abuses me

AOL's Query Log

Detour

- User 17556639:
  - how to kill your wife
  - wife killer
  - how to kill a wife
  - dead people
  - pictures of dead people
  - killed people
  - murder photo
  - steak and cheese
  - decapatated photos
  - car crashes3
  - car crash photo

AOL's Query Log

Detour

- User 4417749:

  http://www.nytimes.com/2006/08/09/technology/09aol.html

"I feel violated, my privacy has been invaded and I'm not in control of my own data"

AOL's Query Log

Detour

- This has just been a very small sample…
- AOL removed the data on August 7, 2006, from their Web site, one day after its release
- They apologized: **"This was a screw up"**

- However, **the data is still out there…**

  https://jeffhuang.com/search_query_logs/

  …and probably always will be

The Netflix Data Set

Detour

- Netflix, America's largest online DVD rental service, had similar problems
- They released data about what DVDs has been rented by each user, along with the users' movie ratings
- As with the AOL data set, user IDs have been replaced by random numbers
- Unfortunately, researchers have been able to reconstruct the identity of some customers by comparing their movie ratings with reviews written at imdb.com, a public movie database

End of Semester…