Introduction to Information Retrieval
http://informationretrieval.org

IIR 19: Web Search

Hinrich Schütze

Center for Information and Language Processing, University of Munich

2014-07-02
Overview

1. Recap
2. Big picture
3. Ads
4. Duplicate detection
5. Spam
6. Web IR
   - Queries
   - Links
   - Context
   - Users
   - Documents
   - Size
7. Size of the web
Outline

1 Recap
2 Big picture
3 Ads
4 Duplicate detection
5 Spam
6 Web IR
   • Queries
   • Links
   • Context
   • Users
   • Documents
   • Size
7 Size of the web
Indexing anchor text

- Anchor text is often a better description of a page’s content than the page itself.
- Anchor text can be weighted more highly than the text on the page.
- A Google bomb is a search with “bad” results due to maliciously manipulated anchor text.
  - [dangerous cult] on Google, Bing, Yahoo
Model: a web surfer doing a random walk on the web

Formalization: Markov chain

PageRank is the \textit{long-term visit rate} of the random surfer or the \textit{steady-state distribution}.

Need \textit{teleportation} to ensure well-defined PageRank

Power method to compute PageRank

  - PageRank is the principal left eigenvector of the transition probability matrix.
Computing PageRank: Power method

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$P_{t_1} (d_1)$</th>
<th>$P_{t_2} (d_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$P_{11} = 0.1$</td>
<td>$P_{12} = 0.9$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$P_{21} = 0.3$</td>
<td>$P_{22} = 0.7$</td>
</tr>
</tbody>
</table>

$t_0$    0   1   0.3   0.7  $= \vec{x}P$
$t_1$    0.3 0.7  0.24 0.76  $= \vec{x}P^2$
$t_2$    0.24 0.76 0.252 0.748 $= \vec{x}P^3$
$t_3$    0.252 0.748 0.2496 0.7504 $= \vec{x}P^4$
$t_\infty$  0.25 0.75 0.25 0.75  $= \vec{x}P^\infty$

PageRank vector $= \vec{\pi} = (\pi_1, \pi_2) = (0.25, 0.75)$

$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$
$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$
HITS: Hubs and authorities

- **Hubs**
  - www.bestfares.com
  - www.airlinesquality.com
  - blogs.usatoday.com/sky
  - aviationblog.dallasnews.com

- **Authorities**
  - www.aa.com
  - www.delta.com
  - www.united.com
HITS update rules

- $A$: link matrix
- $\vec{h}$: vector of hub scores
- $\vec{a}$: vector of authority scores
- **HITS algorithm:**
  - Compute $\vec{h} = A\vec{a}$
  - Compute $\vec{a} = A^T\vec{h}$
  - Iterate until convergence
  - Output (i) list of hubs ranked according to hub score and (ii) list of authorities ranked according to authority score
Outline

1 Recap
2 Big picture
3 Ads
4 Duplicate detection
5 Spam
6 Web IR
   - Queries
   - Links
   - Context
   - Users
   - Documents
   - Size
7 Size of the web
Web search overview
Search is a top activity on the web

How often do you use search engines on the Internet?

- Four or more times each day: 21.2%
- At least once every day: 35.1%
- Several times each week: 22.7%
- At least once each week: 10.3%
- Several times each month: 5.5%
- Less frequently: 3.9%
- Never: 1.2%
Without search engines, the web wouldn’t work
Without search engines, the web wouldn’t work

- Without search, **content is hard to find.**
Without search engines, the web wouldn’t work

- Without search, content is hard to find.
- Without search, there is no incentive to create content.
Without search engines, the web wouldn’t work

- Without search, content is hard to find.
- → Without search, there is no incentive to create content.
  - Why publish something if nobody will read it?
Without search, content is hard to find.

→ Without search, there is no incentive to create content.
  - Why publish something if nobody will read it?
  - Why publish something if I don’t get ad revenue from it?
Without search engines, the web wouldn’t work

- Without search, content is hard to find.
- → Without search, there is no incentive to create content.
  - Why publish something if nobody will read it?
  - Why publish something if I don’t get ad revenue from it?
- Somebody needs to pay for the web.
Without search, content is hard to find.

→ Without search, there is no incentive to create content.
  - Why publish something if nobody will read it?
  - Why publish something if I don’t get ad revenue from it?

Somebody needs to pay for the web.
  - Servers, web infrastructure, content creation
Without search engines, the web wouldn’t work

- Without search, content is hard to find.
- Without search, there is no incentive to create content.
  - Why publish something if nobody will read it?
  - Why publish something if I don’t get ad revenue from it?
- Somebody needs to pay for the web.
  - Servers, web infrastructure, content creation
  - A large part today is paid by search ads.
Without search engines, the web wouldn’t work

- Without search, **content is hard to find**.
- → Without search, there is **no incentive to create content**.
  - Why publish something if nobody will read it?
  - Why publish something if I don’t get ad revenue from it?
- Somebody needs to pay for the web.
  - Servers, web infrastructure, content creation
  - A large part today is paid by search ads.
  - **Search pays for the web.**
Interest aggregation
Interest aggregation

- Unique feature of the web: A small number of geographically dispersed people with similar interests can find each other.
Interest aggregation

- Unique feature of the web: A small number of geographically dispersed people with similar interests can find each other.
  - Elementary school kids with hemophilia
### Interest aggregation

- Unique feature of the web: A small number of geographically dispersed people with similar interests can find each other.
  - Elementary school kids with hemophilia
  - People interested in translating R5R5 Scheme into relatively portable C (open source project)
Interest aggregation

- Unique feature of the web: A small number of geographically dispersed people with similar interests can find each other.
  - Elementary school kids with hemophilia
  - People interested in translating R5R5 Scheme into relatively portable C (open source project)
  - Search engines are a key enabler for interest aggregation.
IR on the web vs. IR in general
On the web, search is not just a nice feature.
IR on the web vs. IR in general

- On the web, search is not just a nice feature.
  - Search is a key enabler of the web: ...
IR on the web vs. IR in general

- On the web, search is not just a nice feature.
  - Search is a key enabler of the web: ...
  - ...financing, content creation, interest aggregation etc.
IR on the web vs. IR in general

- On the web, search is not just a nice feature.
  - Search is a key enabler of the web: ...
  - ...financing, content creation, interest aggregation etc.

- The web is a chaotic und uncoordinated collection.
IR on the web vs. IR in general

- On the web, search is not just a nice feature.
  - Search is a key enabler of the web: ... 
  - ...financing, content creation, interest aggregation etc.

- The web is a chaotic and uncoordinated collection.

- No control / restrictions on who can author content
IR on the web vs. IR in general

- On the web, search is not just a nice feature.
  - Search is a key enabler of the web: ...
  - ...financing, content creation, interest aggregation etc.

- The web is a chaotic und uncoordinated collection.

- No control / restrictions on who can author content

- The web is very large.
IR on the web vs. IR in general

- On the web, search is not just a nice feature.
  - Search is a key enabler of the web: . . .
  - . . . financing, content creation, interest aggregation etc.
  - → look at search ads

- The web is a chaotic und uncoordinated collection.

- No control / restrictions on who can author content

- The web is very large.
IR on the web vs. IR in general

- On the web, search is not just a nice feature.
  - Search is a key enabler of the web: ...
  - ... financing, content creation, interest aggregation etc.
  → look at search ads

- The web is a chaotic und uncoordinated collection. → lots of duplicates – need to detect duplicates

- No control / restrictions on who can author content

- The web is very large.
IR on the web vs. IR in general

- On the web, search is not just a nice feature.
  - Search is a key enabler of the web: ...
  - ...financing, content creation, interest aggregation etc.
    → look at search ads
- The web is a chaotic and uncoordinated collection. → lots of duplicates – need to detect duplicates
- No control / restrictions on who can author content → lots of spam – need to detect spam
- The web is very large.
IR on the web vs. IR in general

- On the web, search is not just a nice feature.
  - Search is a key enabler of the web: ...  
  - ...financing, content creation, interest aggregation etc.
    → look at search ads

- The web is a chaotic und uncoordinated collection.  → lots of duplicates – need to detect duplicates

- No control / restrictions on who can author content  → lots of spam – need to detect spam

- The web is very large.  → need to know how big it is
Take-away today
Take-away today

- Big picture
Take-away today

- Big picture
- Ads – they pay for the web
Take-away today

- Big picture
- Ads – they pay for the web
- Duplicate detection – addresses one aspect of chaotic content creation
Take-away today

- Big picture
- Ads – they pay for the web
- Duplicate detection – addresses one aspect of chaotic content creation
- Spam detection – addresses one aspect of lack of central access control
Take-away today

- Big picture
- Ads – they pay for the web
- Duplicate detection – addresses one aspect of chaotic content creation
- Spam detection – addresses one aspect of lack of central access control
- Probably won’t get to today
  - Web information retrieval
  - Size of the web
Outline

1 Recap
2 Big picture
3 Ads
4 Duplicate detection
5 Spam
6 Web IR
   - Queries
   - Links
   - Context
   - Users
   - Documents
   - Size
7 Size of the web
First generation of search ads: Goto (1996)
First generation of search ads: Goto (1996)
First generation of search ads: Goto (1996)
First generation of search ads: Goto (1996)

- Buddy Blake bid the maximum ($0.38) for this search.
First generation of search ads: Goto (1996)

- Buddy Blake bid the maximum ($0.38) for this search.
- He paid $0.38 to Goto every time somebody clicked on the link.
First generation of search ads: Goto (1996)

- Buddy Blake bid the maximum ($0.38) for this search.
- He paid $0.38 to Goto every time somebody clicked on the link.
- Pages were simply ranked according to bid – revenue maximization for Goto.
First generation of search ads: Goto (1996)

- Buddy Blake bid the maximum ($0.38) for this search.
- He paid $0.38 to Goto every time somebody clicked on the link.
- Pages were simply ranked according to bid – revenue maximization for Goto.
- No separation of ads/docs. Only one result list!
First generation of search ads: Goto (1996)

- Buddy Blake bid the maximum ($0.38) for this search.
- He paid $0.38 to Goto every time somebody clicked on the link.
- Pages were simply ranked according to bid – revenue maximization for Goto.
- No separation of ads/docs. Only one result list!
- Upfront and honest. No relevance ranking, ...
First generation of search ads: Goto (1996)

- Buddy Blake bid the maximum ($0.38) for this search.
- He paid $0.38 to Goto every time somebody clicked on the link.
- Pages were simply ranked according to bid – revenue maximization for Goto.
- No separation of ads/docs. Only one result list!
- Upfront and honest. No relevance ranking, . . .
- . . .but Goto did not pretend there was any.
Second generation of search ads: Google (2000/2001)
Second generation of search ads: Google (2000/2001)

- Strict separation of search results and search ads
Two ranked lists: web pages (left) and ads (right)
Two ranked lists: web pages (left) and ads (right)

SogoTrade appears in ads.
Two ranked lists: web pages (left) and ads (right)

SogoTrade appears in search results.

SogoTrade appears in ads.
Two ranked lists: web pages (left) and ads (right)

SogoTrade appears in search results.

SogoTrade appears in ads.

Do search engines rank advertisers higher than non-advertisers?
Two ranked lists: web pages (left) and ads (right)

SogoTrade appears in search results.

SogoTrade appears in ads.

Do search engines rank advertisers higher than non-advertisers?

All major search engines claim no.
Do ads influence editorial content?
Do ads influence editorial content?

- Similar problem at newspapers / TV channels
Do ads influence editorial content?

- Similar problem at newspapers / TV channels
- A newspaper is reluctant to publish harsh criticism of its major advertisers.
Do ads influence editorial content?

- Similar problem at newspapers / TV channels
- A newspaper is reluctant to publish harsh criticism of its major advertisers.
- The line often gets blurred at newspapers / on TV.
Do ads influence editorial content?

- Similar problem at newspapers / TV channels
- A newspaper is reluctant to publish harsh criticism of its major advertisers.
- The line often gets blurred at newspapers / on TV.
- No known case of this happening with search engines yet?
How are the ads on the right ranked?

Discount Broker Reviews
Information on online discount brokers emphasizing rates, charges, and customer comments and complaints.
www.broker-reviews.us/ - 94k - Cached - Similar pages

Discount Broker Rankings (2008 Broker Survey) at SmartMoney.com
Discount Brokers. Rank/ Brokerage/ Minimum to Open Account, Comments, Standard Commis- sion*, Reduced Commission, Account Fee Per Year (How to Avoid), Avg. ...

Stock Brokers | Discount Brokers | Online Brokers
Most Recommended. Top 5 Brokers headlines. 10. Don’t Pay Your Broker for Free Funds May 15 at 3:39 PM. 5. Don’t Discount the Discounters Apr 18 at 2:41 PM ...
www.fool.com/investing/brokers/index.aspx - 44k - Cached - Similar pages

Discount Broker
Discount Broker - Definition of Discount Broker on Investopedia - A stockbroker who carries out buy and sell orders at a reduced commission compared to a ...
www.investopedia.com/terms/d/discountbroker.asp - 31k - Cashed - Similar pages

Discount Brokerage and Online Trading for Smart Stock Market ...
Online stock broker SogoTrade offers the best in discount brokerage investing. Get stock market quotes from this internet stock trading company.
www.sogotrade.com/ - 39k - Cached - Similar pages

15 questions to ask discount brokers - MSN Money
Jan 11, 2004 ... If you’re not big on hand-holding when it comes to investing, a discount broker can be an economical way to go. Just be sure to ask these ...
moneycentral.msn.com/content/Investing/StartInvesting/P66171.asp - 34k - Cashed - Similar pages

Rate #1 Online Broker
No Minimums. No Inactivity Fee Transfer to Firsttrade for Free!
www.ﬁrsttrade.com

Discount Broker
Commission free trades for 30 days.
No maintenance fees. Sign up now.
TDAMERITRADE.com

TradeKing - Online Broker
$4.95 per Trade, Market or Limit
SmartMoney Top Discount Broker 2001
www.TradeKing.com

Scottrade Brokerage
$7 Trades, No Share Limit. In-Depth Research. Start Trading Online Now!
www.Scottrade.com

Stock trades $1.50 - $3
100 free trades, up to $100 back for transfer costs, $500 minimum
www.sogotrade.com

$3.95 Online Stock Trades
Market/Limit Orders, No Share Limit and No Inactivity Fees
www.Marsco.com

INGDIRECT | ShareBuilder
Brokerage, Order Entry, et al.
### How are ads ranked?

<table>
<thead>
<tr>
<th>Recap</th>
<th>Big picture</th>
<th>Ads</th>
<th>Duplicate detection</th>
<th>Spam</th>
<th>Web IR</th>
<th>Size of the web</th>
</tr>
</thead>
</table>

Schütze: Web search
How are ads ranked?

- Advertisers bid for keywords – *sale by auction.*
How are ads ranked?

- Advertisers bid for keywords – sale by auction.
- Open system: Anybody can participate and bid on keywords.
How are ads ranked?

- Advertisers bid for keywords – **sale by auction**.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are **only charged when somebody clicks** on your ad.
How are ads ranked?

- Advertisers bid for keywords – sale by auction.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are only charged when somebody clicks on your ad.
- How does the auction determine an ad’s rank and the price paid for the ad?
How are ads ranked?

- Advertisers bid for keywords – *sale by auction*.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are *only charged when somebody clicks* on your ad.
- How does the auction determine an ad’s *rank* and the *price paid* for the ad?
- Basis is a *second price auction*, but with twists.
Ads
Duplicate detection
Spam
Web IR
Size of the web

How are ads ranked?

- Advertisers bid for keywords – sale by auction.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are only charged when somebody clicks on your ad.
- How does the auction determine an ad’s rank and the price paid for the ad?
- Basis is a second price auction, but with twists
- For the bottom line, this is perhaps the most important research area for search engines – computational advertising.
How are ads ranked?

- Advertisers bid for keywords – *sale by auction*.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are *only charged when somebody clicks* on your ad.
- How does the auction determine an ad’s *rank* and the *price paid* for the ad?
- Basis is a *second price auction*, but with twists
- For the bottom line, this is perhaps the most important research area for search engines – computational advertising.
  - Squeezing an additional fraction of *a cent* from each ad *means billions* of additional revenue for the search engine.
How are ads ranked?
How are ads ranked?

- First cut: according to bid price à la Goto
How are ads ranked?

- First cut: according to bid price à la Goto
  - Bad idea: open to abuse
How are ads ranked?

- First cut: according to bid price à la Goto
  - Bad idea: open to abuse
  - Example: query [treatment for cancer?] → how to write your last will
How are ads ranked?

- First cut: according to bid price à la Goto
  - Bad idea: open to abuse
  - Example: query [treatment for cancer?] → how to write your last will
  - We don’t want to show nonrelevant or offensive ads.
How are ads ranked?

- First cut: according to bid price à la Goto
  - Bad idea: open to abuse
  - Example: query [treatment for cancer?] → how to write your last will
  - We don’t want to show nonrelevant or offensive ads.
- Instead: rank based on bid price and relevance
How are ads ranked?

- First cut: according to bid price à la Goto
  - Bad idea: open to abuse
  - Example: query [treatment for cancer?] → how to write your last will
  - We don’t want to show nonrelevant or offensive ads.

- Instead: rank based on bid price and relevance

- Key measure of ad relevance: clickthrough rate
How are ads ranked?

- First cut: according to bid price à la Goto
  - Bad idea: open to abuse
  - Example: query [treatment for cancer?] → how to write your last will
  - We don’t want to show nonrelevant or offensive ads.

- Instead: rank based on bid price and relevance

- Key measure of ad relevance: clickthrough rate
  - clickthrough rate = CTR = clicks per impressions
How are ads ranked?

- First cut: according to bid price à la Goto
  - Bad idea: open to abuse
  - Example: query [treatment for cancer?] → how to write your last will
  - We don’t want to show nonrelevant or offensive ads.

- Instead: rank based on bid price and relevance

- Key measure of ad relevance: clickthrough rate
  - clickthrough rate = $\text{CTR} = \text{clicks per impressions}$

- Result: A nonrelevant ad will be ranked low.
How are ads ranked?

- First cut: according to bid price à la Goto
  - Bad idea: open to abuse
  - Example: query [treatment for cancer?] → how to write your last will
  - We don’t want to show nonrelevant or offensive ads.

- Instead: rank based on bid price and relevance

- Key measure of ad relevance: clickthrough rate
  - clickthrough rate = CTR = clicks per impressions

- Result: A nonrelevant ad will be ranked low.
  - Even if this decreases search engine revenue short-term
How are ads ranked?

- First cut: according to bid price à la Goto
  - Bad idea: open to abuse
  - Example: query [treatment for cancer?] → how to write your last will
  - We don’t want to show nonrelevant or offensive ads.

- Instead: rank based on bid price and relevance

- Key measure of ad relevance: clickthrough rate
  - clickthrough rate = CTR = clicks per impressions

- Result: A nonrelevant ad will be ranked low.
  - Even if this decreases search engine revenue short-term
  - Hope: Overall acceptance of the system and overall revenue is maximized if users get useful information.
How are ads ranked?

- First cut: according to bid price à la Goto
  - Bad idea: open to abuse
  - Example: query [treatment for cancer?] → how to write your last will
  - We don’t want to show nonrelevant or offensive ads.
- Instead: rank based on bid price and relevance
- Key measure of ad relevance: clickthrough rate
  - clickthrough rate = CTR = clicks per impressions
- Result: A nonrelevant ad will be ranked low.
  - Even if this decreases search engine revenue short-term
  - Hope: Overall acceptance of the system and overall revenue is maximized if users get useful information.
- Other ranking factors: location, time of day, quality and loading speed of landing page
How are ads ranked?

- First cut: according to bid price à la Goto
  - Bad idea: open to abuse
  - Example: query [treatment for cancer?] → how to write your last will
  - We don’t want to show nonrelevant or offensive ads.

- Instead: rank based on bid price and relevance

- Key measure of ad relevance: clickthrough rate
  - clickthrough rate = CTR = clicks per impressions

- Result: A nonrelevant ad will be ranked low.
  - Even if this decreases search engine revenue short-term
  - Hope: Overall acceptance of the system and overall revenue is maximized if users get useful information.

- Other ranking factors: location, time of day, quality and loading speed of landing page

- The main ranking factor: the query
Google AdWords demo
## Google’s second price auction

<table>
<thead>
<tr>
<th>advertiser</th>
<th>bid</th>
<th>CTR</th>
<th>ad rank</th>
<th>rank</th>
<th>paid</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$4.00</td>
<td>0.01</td>
<td>0.04</td>
<td>4</td>
<td>(minimum)</td>
</tr>
<tr>
<td>B</td>
<td>$3.00</td>
<td>0.03</td>
<td>0.09</td>
<td>2</td>
<td>$2.68</td>
</tr>
<tr>
<td>C</td>
<td>$2.00</td>
<td>0.06</td>
<td>0.12</td>
<td>1</td>
<td>$1.51</td>
</tr>
<tr>
<td>D</td>
<td>$1.00</td>
<td>0.08</td>
<td>0.08</td>
<td>3</td>
<td>$0.51</td>
</tr>
</tbody>
</table>
## Google’s second price auction

<table>
<thead>
<tr>
<th>advertiser</th>
<th>bid</th>
<th>CTR</th>
<th>ad rank</th>
<th>rank</th>
<th>paid</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$4.00</td>
<td>0.01</td>
<td>0.04</td>
<td>4</td>
<td>(minimum)</td>
</tr>
<tr>
<td>B</td>
<td>$3.00</td>
<td>0.03</td>
<td>0.09</td>
<td>2</td>
<td>$2.68</td>
</tr>
<tr>
<td>C</td>
<td>$2.00</td>
<td>0.06</td>
<td>0.12</td>
<td>1</td>
<td>$1.51</td>
</tr>
<tr>
<td>D</td>
<td>$1.00</td>
<td>0.08</td>
<td>0.08</td>
<td>3</td>
<td>$0.51</td>
</tr>
</tbody>
</table>

- **bid**: maximum bid for a click by advertiser
- **CTR**: click-through rate: when an ad is displayed, what percentage of time do users click on it? **CTR is a measure of relevance.**
- **ad rank**: bid \( \times \) CTR: this trades off (i) how much money the advertiser is willing to pay against (ii) how relevant the ad is
- **rank**: rank in auction
- **paid**: second price auction price paid by advertiser
### Google’s second price auction

<table>
<thead>
<tr>
<th>advertiser</th>
<th>bid</th>
<th>CTR</th>
<th>ad rank</th>
<th>rank</th>
<th>paid</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$4.00</td>
<td>0.01</td>
<td>0.04</td>
<td>4</td>
<td>(minimum)</td>
</tr>
<tr>
<td>B</td>
<td>$3.00</td>
<td>0.03</td>
<td>0.09</td>
<td>2</td>
<td>$2.68</td>
</tr>
<tr>
<td>C</td>
<td>$2.00</td>
<td>0.06</td>
<td>0.12</td>
<td>1</td>
<td>$1.51</td>
</tr>
<tr>
<td>D</td>
<td>$1.00</td>
<td>0.08</td>
<td>0.08</td>
<td>3</td>
<td>$0.51</td>
</tr>
</tbody>
</table>

Second price auction: The advertiser pays the minimum amount necessary to maintain their position in the auction (plus 1 cent).

\[
\text{price}_1 \times \text{CTR}_1 = \text{bid}_2 \times \text{CTR}_2 \quad \text{(this will result in rank}_1\text{=rank}_2\text{)}
\]

\[
\text{price}_1 = \frac{\text{bid}_2 \times \text{CTR}_2}{\text{CTR}_1}
\]

\[
\begin{align*}
\text{price}_1 &= \frac{\text{bid}_2 \times \text{CTR}_2}{\text{CTR}_1} = \frac{3.00 \times 0.03}{0.06} = 1.50 \\
\text{price}_2 &= \frac{\text{bid}_3 \times \text{CTR}_3}{\text{CTR}_2} = \frac{1.00 \times 0.08}{0.03} = 2.67 \\
\text{price}_3 &= \frac{\text{bid}_4 \times \text{CTR}_4}{\text{CTR}_3} = \frac{4.00 \times 0.01}{0.08} = 0.50
\end{align*}
\]
Keywords with high bids
Keywords with high bids

According to http://www.cwire.org/highest-paying-search-terms/

$69.1  mesothelioma treatment options
$65.9  personal injury lawyer michigan
$62.6  student loans consolidation
$61.4  car accident attorney los angeles
$59.4  online car insurance quotes
$59.4  arizona dui lawyer
$46.4  asbestos cancer
$40.1  home equity line of credit
$39.8  life insurance quotes
$39.2  refinancing
$38.7  equity line of credit
$38.0  lasik eye surgery new york city
$37.0  2nd mortgage
$35.9  free car insurance quote
Search ads: A win-win-win?
Search ads: A win-win-win?

- The **search engine** company gets revenue every time somebody clicks on an ad.
Search ads: A win-win-win?

- The search engine company gets revenue every time somebody clicks on an ad.
- The user only clicks on an ad if they are interested in the ad.
Search ads: A win-win-win?

- The **search engine** company gets revenue every time somebody clicks on an ad.
- The **user** only clicks on an ad if they are interested in the ad.
  - Search engines punish misleading and nonrelevant ads.
Search ads: A win-win-win?

- The **search engine** company gets revenue every time somebody clicks on an ad.
- The **user** only clicks on an ad if they are interested in the ad.
  - Search engines punish misleading and nonrelevant ads.
  - As a result, users are often satisfied with what they find after clicking on an ad.
Search ads: A win-win-win?

- The **search engine** company gets revenue every time somebody clicks on an ad.
- The **user** only clicks on an ad if they are interested in the ad.
  - Search engines punish misleading and nonrelevant ads.
  - As a result, users are often satisfied with what they find after clicking on an ad.
- The **advertiser** finds new customers in a cost-effective way.
Exercise
Exercise

Why is web search potentially more attractive for advertisers than TV spots, newspaper ads or radio spots?
Exercise

- Why is web search potentially more attractive for advertisers than TV spots, newspaper ads or radio spots?
- The advertiser pays for all this. How can the advertiser be cheated?
Exercise

Why is web search potentially more attractive for advertisers than TV spots, newspaper ads or radio spots?

The advertiser pays for all this. How can the advertiser be cheated?

Any way this could be bad for the user?
Exercise

- Why is web search potentially more attractive for advertisers than TV spots, newspaper ads or radio spots?
- The advertiser pays for all this. How can the advertiser be cheated?
- Any way this could be bad for the user?
- Any way this could be bad for the search engine?
Not a win-win-win: Keyword arbitrage
Not a win-win-win: Keyword arbitrage

- Buy a keyword on Google
Not a win-win-win: Keyword arbitrage

- Buy a keyword on Google
- Then redirect traffic to a third party that is paying much more than you are paying Google.
Not a win-win-win: Keyword arbitrage

- Buy a keyword on Google
- Then redirect traffic to a third party that is paying much more than you are paying Google.
  - E.g., redirect to a page full of ads
Not a win-win-win: Keyword arbitrage

- Buy a keyword on Google
- Then redirect traffic to a third party that is paying much more than you are paying Google.
  - E.g., redirect to a page full of ads
- This rarely makes sense for the user.
Not a win-win-win: Keyword arbitrage

- Buy a keyword on Google
- Then redirect traffic to a third party that is paying much more than you are paying Google.
  - E.g., redirect to a page full of ads
- This rarely makes sense for the user.
- Ad spammers keep inventing new tricks.
Not a win-win-win: Keyword arbitrage

- Buy a keyword on Google
- Then redirect traffic to a third party that is paying much more than you are paying Google.
  - E.g., redirect to a page full of ads
- This rarely makes sense for the user.
- Ad spammers keep inventing new tricks.
- The search engines need time to catch up with them.
Not a win-win-win: Violation of trademarks
Not a win-win-win: Violation of trademarks

Example: geico
Not a win-win-win: Violation of trademarks

- Example: geico
- During part of 2005: The search term “geico” on Google was bought by competitors.
Not a win-win-win: Violation of trademarks

- Example: geico
- During part of 2005: The search term “geico” on Google was bought by competitors.
- Geico lost this case in the United States.
Not a win-win-win: Violation of trademarks

- Example: geico
- During part of 2005: The search term “geico” on Google was bought by competitors.
- Geico lost this case in the United States.
- Louis Vuitton lost similar case in Europe.
Not a win-win-win: Violation of trademarks

- Example: geico

- During part of 2005: The search term “geico” on Google was bought by competitors.

- Geico lost this case in the United States.

- Louis Vuitton lost similar case in Europe.

- See http://google.com/tm_complaint.html
Example: geico
During part of 2005: The search term “geico” on Google was bought by competitors.
Geico lost this case in the United States.
Louis Vuitton lost similar case in Europe.
See http://google.com/tm_complaint.html
It’s potentially misleading to users to trigger an ad off of a trademark if the user can’t buy the product on the site.
Outline

1 Recap
2 Big picture
3 Ads
4 Duplicate detection
5 Spam
6 Web IR
   • Queries
   • Links
   • Context
   • Users
   • Documents
   • Size
7 Size of the web
The web is full of duplicated content.
Duplicate detection

- The web is full of duplicated content.
- More so than many other collections
Duplicate detection

- The web is full of duplicated content.
- More so than many other collections
- Exact duplicates
Duplicate detection

- The web is full of duplicated content.
- More so than many other collections
- Exact duplicates
  - Easy to eliminate
Duplicate detection

- The web is full of duplicated content.
- More so than many other collections
- Exact duplicates
  - Easy to eliminate
  - E.g., use hash/fingerprint
The web is full of duplicated content.

More so than many other collections

Exact duplicates
  - Easy to eliminate
  - E.g., use hash/fingerprint

Near-duplicates
Duplicate detection

- The web is full of duplicated content.
- More so than many other collections
- Exact duplicates
  - Easy to eliminate
  - E.g., use hash/fingerprint
- Near-duplicates
  - Abundant on the web
The web is full of duplicated content.

More so than many other collections

Exact duplicates
   - Easy to eliminate
   - E.g., use hash/fingerprint

Near-duplicates
   - Abundant on the web
   - Difficult to eliminate
Duplicate detection

- The web is full of duplicated content.
- More so than many other collections
- Exact duplicates
  - Easy to eliminate
  - E.g., use hash/fingerprint
- Near-duplicates
  - Abundant on the web
  - Difficult to eliminate
- For the user, it’s annoying to get a search result with near-identical documents.
Duplicate detection

- The web is full of duplicated content.
- More so than many other collections
- Exact duplicates
  - Easy to eliminate
  - E.g., use hash/fingerprint
- Near-duplicates
  - Abundant on the web
  - Difficult to eliminate
- For the user, it’s annoying to get a search result with near-identical documents.
  - Marginal relevance is zero: even a highly relevant document becomes nonrelevant if it appears below a (near-)duplicate.
Duplicate detection

• The web is full of duplicated content.
• More so than many other collections
• Exact duplicates
  • Easy to eliminate
  • E.g., use hash/fingerprint
• Near-duplicates
  • Abundant on the web
  • Difficult to eliminate

For the user, it’s annoying to get a search result with near-identical documents.

**Marginal relevance is zero**: even a highly relevant document becomes nonrelevant if it appears below a (near-)duplicate.

We need to eliminate near_duplicates.
Near-duplicates: Example
Near-duplicates: Example

For other persons named Michael Jackson, see Michael Jackson (disambiguation).

Michael Joseph Jackson (August 29, 1958 - June 25, 2009) was an American recording artist, entertainer and businessman. The seventh child of the Jackson family, he made his debut as an entertainer in 1968 as a member of The Jackson 5. He then began a solo career in 1971 with the album Got to Be There.
Exercise
How would you eliminate near-duplicates on the web?
Detecting near-duplicates
Detecting near-duplicates

- Compute similarity with an edit-distance measure
Detecting near-duplicates

- Compute similarity with an edit-distance measure
- We want “syntactic” (as opposed to semantic) similarity.
Detecting near-duplicates

- Compute similarity with an edit-distance measure
- We want “syntactic” (as opposed to semantic) similarity.
  - True semantic similarity (similarity in content) is too difficult to compute.
Detecting near-duplicates

- Compute similarity with an edit-distance measure
- We want "syntactic" (as opposed to semantic) similarity.
  - True semantic similarity (similarity in content) is too difficult to compute.
- We do not consider documents near-duplicates if they have the same content, but express it with different words.
Detecting near-duplicates

- Compute similarity with an edit-distance measure
- We want “syntactic” (as opposed to semantic) similarity.
  - True semantic similarity (similarity in content) is too difficult to compute.
- We do not consider documents near-duplicates if they have the same content, but express it with different words.
- Use similarity threshold $\theta$ to make the call “is/isn’t a near-duplicate”.
Detecting near-duplicates

- Compute similarity with an edit-distance measure
- We want “syntactic” (as opposed to semantic) similarity.
  - True semantic similarity (similarity in content) is too difficult to compute.
- We do not consider documents near-duplicates if they have the same content, but express it with different words.
- Use similarity threshold $\theta$ to make the call “is/isn’t a near-duplicate”.
- E.g., two documents are near-duplicates if similarity $> \theta = 80\%$. 
Represent each document as set of shingles
Represent each document as set of **shingles**

- A shingle is simply a **word n-gram**.
Represent each document as set of **shingles**

- A shingle is simply a **word n-gram**.
- Shingles are used as features to **measure syntactic similarity** of documents.
A shingle is simply a word n-gram.

Shingles are used as features to measure syntactic similarity of documents.

For example, for $n = 3$, “a rose is a rose is a rose” would be represented as this set of shingles:
Represent each document as set of **shingles**

- A shingle is simply a **word n-gram**.
- Shingles are used as features to **measure syntactic similarity** of documents.
- For example, for \( n = 3 \), “a rose is a rose is a rose” would be represented as this set of shingles:
  
  \[
  \{ \text{a-rose-is, rose-is-a, is-a-rose} \}
  \]
A shingle is simply a **word n-gram**.

Shingles are used as features to **measure syntactic similarity** of documents.

For example, for \( n = 3 \), “a rose is a rose is a rose” would be represented as this set of shingles:

\[
\{ \text{a-rose-is, rose-is-a, is-a-rose} \}
\]

We can map shingles to \( 1..2^m \) (e.g., \( m = 64 \)) by fingerprinting.
Represent each document as set of **shingles**

- A shingle is simply a **word n-gram**.
- Shingles are used as features to **measure syntactic similarity** of documents.
- For example, for $n = 3$, “a rose is a rose is a rose” would be represented as this set of shingles:
  - \{ a-rose-is, rose-is-a, is-a-rose \}
- We can map shingles to $1..2^m$ (e.g., $m = 64$) by fingerprinting.
- From now on: $s_k$ refers to the shingle’s fingerprint in $1..2^m$. 
Represent each document as set of **shingles**

- A shingle is simply a **word n-gram**.
- Shingles are used as features to **measure syntactic similarity** of documents.
- For example, for \( n = 3 \), “a rose is a rose is a rose” would be represented as this set of shingles:
  - \{ a-rose-is, rose-is-a, is-a-rose \}
- We can map shingles to \( 1..2^m \) (e.g., \( m = 64 \)) by fingerprinting.
- From now on: \( s_k \) refers to the shingle’s fingerprint in \( 1..2^m \).
- We define the similarity of two documents as the **Jaccard coefficient** of their shingle sets.
Recall: Jaccard coefficient
Recall: Jaccard coefficient

- A commonly used measure of overlap of two sets
Recall: Jaccard coefficient

- A commonly used measure of overlap of two sets
- Let $A$ and $B$ be two sets
Recall: Jaccard coefficient

- A commonly used measure of overlap of two sets
- Let $A$ and $B$ be two sets
- Jaccard coefficient:

$$\text{JACCARD}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

$(A \neq \emptyset \text{ or } B \neq \emptyset)$
Recall: Jaccard coefficient

- A commonly used measure of overlap of two sets
- Let $A$ and $B$ be two sets
- Jaccard coefficient:

$$\text{JACCARD}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

$(A \neq \emptyset \text{ or } B \neq \emptyset)$

- $\text{JACCARD}(A, A) = 1$
Recall: Jaccard coefficient

- A commonly used measure of overlap of two sets
- Let $A$ and $B$ be two sets
- Jaccard coefficient:

$$\text{JACCARD}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

$(A \neq \emptyset \text{ or } B \neq \emptyset)$

- $\text{JACCARD}(A, A) = 1$
- $\text{JACCARD}(A, B) = 0$ if $A \cap B = 0$
Recall: Jaccard coefficient

- A commonly used measure of overlap of two sets
- Let $A$ and $B$ be two sets
- Jaccard coefficient:

$$\text{JACCARD}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

$(A \neq \emptyset \text{ or } B \neq \emptyset)$

- $\text{JACCARD}(A, A) = 1$
- $\text{JACCARD}(A, B) = 0$ if $A \cap B = 0$
- $A$ and $B$ don’t have to be the same size.
Recall: Jaccard coefficient

- A commonly used measure of overlap of two sets
- Let $A$ and $B$ be two sets
- Jaccard coefficient:

$$\text{JACCARD}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

$(A \neq \emptyset \text{ or } B \neq \emptyset)$

- $\text{JACCARD}(A, A) = 1$
- $\text{JACCARD}(A, B) = 0$ if $A \cap B = 0$
- $A$ and $B$ don’t have to be the same size.
- Always assigns a number between 0 and 1.
Jaccard coefficient: Example
Jaccard coefficient: Example

Three documents:

- $d_1$: “Jack London traveled to Oakland”
- $d_2$: “Jack London traveled to the city of Oakland”
- $d_3$: “Jack traveled from Oakland to London”
Jaccard coefficient: Example

- Three documents:
  - $d_1$: “Jack London traveled to Oakland”
  - $d_2$: “Jack London traveled to the city of Oakland”
  - $d_3$: “Jack traveled from Oakland to London”

- Based on shingles of size 2 (2-grams or bigrams), what are the Jaccard coefficients $J(d_1, d_2)$ and $J(d_1, d_3)$?
Jaccard coefficient: Example

- Three documents:
  - \(d_1\): “Jack London traveled to Oakland”
  - \(d_2\): “Jack London traveled to the city of Oakland”
  - \(d_3\): “Jack traveled from Oakland to London”
- Based on shingles of size 2 (2-grams or bigrams), what are the Jaccard coefficients \(J(d_1, d_2)\) and \(J(d_1, d_3)\)?
  - \(J(d_1, d_2) = \frac{3}{8} = 0.375\)
Jaccard coefficient: Example

- Three documents:
  - $d_1$: “Jack London traveled to Oakland”
  - $d_2$: “Jack London traveled to the city of Oakland”
  - $d_3$: “Jack traveled from Oakland to London”
- Based on shingles of size 2 (2-grams or bigrams), what are the Jaccard coefficients $J(d_1, d_2)$ and $J(d_1, d_3)$?
  - $J(d_1, d_2) = 3/8 = 0.375$
  - $J(d_1, d_3) = 0$
Jaccard coefficient: Example

- Three documents:
  - $d_1$: “Jack London traveled to Oakland”
  - $d_2$: “Jack London traveled to the city of Oakland”
  - $d_3$: “Jack traveled from Oakland to London”

- Based on shingles of size 2 (2-grams or bigrams), what are the Jaccard coefficients $J(d_1, d_2)$ and $J(d_1, d_3)$?

- $J(d_1, d_2) = 3/8 = 0.375$

- $J(d_1, d_3) = 0$

- Note: very sensitive to dissimilarity
Represent each document as a sketch
Represent each document as a **sketch**

- The number of shingles per document is large.
Represent each document as a sketch

- The number of shingles per document is large.
- To increase efficiency, we will use a sketch, a cleverly chosen subset of the shingles of a document.
Represent each document as a **sketch**

- The number of shingles per document is large.
- To increase efficiency, we will use a **sketch**, a cleverly chosen **subset** of the shingles of a document.
- The size of a sketch is, say, $n = 200$ . . .
Represent each document as a sketch

- The number of shingles per document is large.
- To increase efficiency, we will use a sketch, a cleverly chosen subset of the shingles of a document.
- The size of a sketch is, say, \( n = 200 \ldots \)
- \( \ldots \) and is defined by a set of permutations \( \pi_1 \ldots \pi_{200} \).
Represent each document as a **sketch**

- The number of shingles per document is large.
- To increase efficiency, we will use a **sketch**, a cleverly chosen **subset** of the shingles of a document.
- The size of a sketch is, say, $n = 200$ . . .
- . . . and is defined by a set of permutations $\pi_1 \ldots \pi_{200}$.
- Each $\pi_i$ is a random permutation on $1..2^m$.
Represent each document as a sketch

- The number of shingles per document is large.
- To increase efficiency, we will use a sketch, a cleverly chosen subset of the shingles of a document.
- The size of a sketch is, say, \( n = 200 \ldots \)
- \ldots and is defined by a set of permutations \( \pi_1 \ldots \pi_{200} \).
- Each \( \pi_i \) is a random permutation on \( 1..2^m \)
- The sketch of \( d \) is defined as:
  \[
  \langle \min_{s \in d} \pi_1(s), \min_{s \in d} \pi_2(s), \ldots, \min_{s \in d} \pi_{200}(s) \rangle
  \]
  (a vector of 200 numbers).
Permutation and minimum: Example

document 1: \( \{ s_k \} \)

\[
\begin{align*}
1 & \rightarrow 2^m \\
1 & \rightarrow 2^m \\
1 & \rightarrow 2^m \\
1 & \rightarrow 2^m \\
1 & \rightarrow 2^m
\end{align*}
\]

document 2: \( \{ s_k \} \)

\[
\begin{align*}
1 & \rightarrow 2^m \\
1 & \rightarrow 2^m \\
1 & \rightarrow 2^m \\
1 & \rightarrow 2^m \\
1 & \rightarrow 2^m
\end{align*}
\]
Permutation and minimum: Example

document 1: \( \{s_k\} \)

document 2: \( \{s_k\} \)

\[
\begin{align*}
\text{s} & \quad \text{s} \\
1 & \rightarrow 2^m \\
& \quad \text{s} \\
1 & \rightarrow 2^m \\
& \quad \text{s} \\
1 & \rightarrow 2^m \\
& \quad \text{s} \\
1 & \rightarrow 2^m \\
& \quad \text{s}
\end{align*}
\]
Permutation and minimum: Example

document 1: \( \{s_k\} \)

document 2: \( \{s_k\} \)

\[ x_k = \pi(s_k) \]
Permutation and minimum: Example

Document 1: \( \{s_k\} \)

\[
x_k = \pi(s_k)
\]

Document 2: \( \{s_k\} \)

\[
x_k = \pi(s_k)
\]
Permutation and minimum: Example

document 1: \{s_k\}

\[ x_k = \pi(s_k) \]

\[ \min_{s_k} \pi(s_k) \]


document 2: \{s_k\}

\[ x_k = \pi(s_k) \]

\[ \min_{s_k} \pi(s_k) \]
Permutation and minimum: Example

**document 1: \( \{s_k\} \)**

\[
x_k = \pi(s_k)
\]

\[
\min_{s_k} \pi(s_k)
\]

We use \( \min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s) \) as a test for: are \( d_1 \) and \( d_2 \) near-duplicates?
Permutation and minimum: Example

document 1: \{s_k\}

document 2: \{s_k\}

\[ x_k = \pi(s_k) \]

\[ \min_{s_k} \pi(s_k) \]

We use \( \min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s) \) as a test for: are \( d_1 \) and \( d_2 \) near-duplicates? In this case: permutation \( \pi \) says: \( d_1 \approx d_2 \)
Computing Jaccard for sketches
Computing Jaccard for sketches

- Sketches: Each document is now a vector of $n = 200$ numbers.
Computing Jaccard for sketches

- Sketches: Each document is now a vector of $n = 200$ numbers.
- Much easier to deal with than the very high-dimensional space of shingles
Computing Jaccard for sketches

- Sketches: Each document is now a vector of \( n = 200 \) numbers.
- Much easier to deal with than the very high-dimensional space of shingles
- But how do we compute Jaccard?
Computing Jaccard for sketches (2)

How do we compute Jaccard?
Computing Jaccard for sketches (2)

- How do we compute Jaccard?
- Let $U$ be the union of the set of shingles of $d_1$ and $d_2$ and $I$ the intersection.
Computing Jaccard for sketches (2)

- How do we compute Jaccard?
- Let $U$ be the union of the set of shingles of $d_1$ and $d_2$ and $I$ the intersection.
- There are $|U|!$ permutations on $U$. 
Computing Jaccard for sketches (2)

- How do we compute Jaccard?
- Let $U$ be the union of the set of shingles of $d_1$ and $d_2$ and $I$ the intersection.
- There are $|U|!$ permutations on $U$.
- For $s' \in I$, for how many permutations $\pi$ do we have $\arg\min_{s \in d_1} \pi(s) = s' = \arg\min_{s \in d_2} \pi(s)$?
Computing Jaccard for sketches (2)

- How do we compute Jaccard?
- Let $U$ be the union of the set of shingles of $d_1$ and $d_2$ and $I$ the intersection.
- There are $|U|!$ permutations on $U$.
- For $s' \in I$, for how many permutations $\pi$ do we have $\arg \min_{s \in d_1} \pi(s) = s' = \arg \min_{s \in d_2} \pi(s)$?
- Answer: $(|U| - 1)!$
Computing Jaccard for sketches (2)

- How do we compute Jaccard?
- Let $U$ be the union of the set of shingles of $d_1$ and $d_2$ and $I$ the intersection.
- There are $|U|$! permutations on $U$.
- For $s' \in I$, for how many permutations $\pi$ do we have $\arg\min_{s \in d_1} \pi(s) = s' = \arg\min_{s \in d_2} \pi(s)$?
- Answer: $(|U| - 1)!$
- There is a set of $(|U| - 1)!$ different permutations for each $s$ in $I$. $\Rightarrow |I|(|U| - 1)!$ permutations make $\arg\min_{s \in d_1} \pi(s) = \arg\min_{s \in d_2} \pi(s)$ true
Computing Jaccard for sketches (2)

- How do we compute Jaccard?
- Let $U$ be the union of the set of shingles of $d_1$ and $d_2$ and $I$ the intersection.
- There are $|U|!$ permutations on $U$.
- For $s' \in I$, for how many permutations $\pi$ do we have $\arg \min_{s \in d_1} \pi(s) = s' = \arg \min_{s \in d_2} \pi(s)$?
- Answer: $(|U| - 1)!$
- There is a set of $(|U| - 1)!$ different permutations for each $s$ in $I$. $\Rightarrow |I|(|U| - 1)!$ permutations make $\arg \min_{s \in d_1} \pi(s) = \arg \min_{s \in d_2} \pi(s)$ true
- Thus, the proportion of permutations that make $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$ true is:

$$\frac{|I|(|U| - 1)!}{|U|!}$$
Computing Jaccard for sketches (2)

- How do we compute Jaccard?
- Let $U$ be the union of the set of shingles of $d_1$ and $d_2$ and $I$ the intersection.
- There are $|U|!$ permutations on $U$.
- For $s' \in I$, for how many permutations $\pi$ do we have $\arg\min_{s \in d_1} \pi(s) = s' = \arg\min_{s \in d_2} \pi(s)$?
- Answer: $(|U| - 1)!$
- There is a set of $(|U| - 1)!$ different permutations for each $s$ in $I$. $\Rightarrow |I|(|U| - 1)!$ permutations make $\arg\min_{s \in d_1} \pi(s) = \arg\min_{s \in d_2} \pi(s)$ true.
- Thus, the proportion of permutations that make $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$ true is:

$$\frac{|I|(|U| - 1)!}{|U|!}$$
Computing Jaccard for sketches (2)

- How do we compute Jaccard?
- Let $U$ be the union of the set of shingles of $d_1$ and $d_2$ and $I$ the intersection.
- There are $|U|!$ permutations on $U$.
- For $s' \in I$, for how many permutations $\pi$ do we have $\text{arg min}_{s \in d_1} \pi(s) = s' = \text{arg min}_{s \in d_2} \pi(s)$?

**Answer:** $(|U| - 1)!$

- There is a set of $(|U| - 1)!$ different permutations for each $s$ in $I$. $\Rightarrow |I|(|U| - 1)!$ permutations make $\text{arg min}_{s \in d_1} \pi(s) = \text{arg min}_{s \in d_2} \pi(s)$ true

- Thus, the proportion of permutations that make $\text{min}_{s \in d_1} \pi(s) = \text{min}_{s \in d_2} \pi(s)$ true is:

$$\frac{|I|(|U| - 1)!}{|U|!}$$
Computing Jaccard for sketches (2)

- How do we compute Jaccard?
- Let $U$ be the union of the set of shingles of $d_1$ and $d_2$ and $I$ the intersection.
- There are $|U|!$ permutations on $U$.
- For $s' \in I$, for how many permutations $\pi$ do we have $\arg\min_{s \in d_1} \pi(s) = s' = \arg\min_{s \in d_2} \pi(s)$?
- Answer: $(|U| - 1)!$
- There is a set of $(|U| - 1)!$ different permutations for each $s$ in $I$. $\Rightarrow |I|(|U| - 1)!$ permutations make $\arg\min_{s \in d_1} \pi(s) = \arg\min_{s \in d_2} \pi(s)$ true
- Thus, the proportion of permutations that make $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$ true is:

$$\frac{|I|(|U| - 1)!}{|U|!} = \frac{|I|}{|U|}$$
Computing Jaccard for sketches (2)

- How do we compute Jaccard?
- Let $U$ be the union of the set of shingles of $d_1$ and $d_2$ and $I$ the intersection.
- There are $|U|!$ permutations on $U$.
- For $s' \in I$, for how many permutations $\pi$ do we have $\arg\min_{s \in d_1} \pi(s) = s' = \arg\min_{s \in d_2} \pi(s)$?
- Answer: $(|U| - 1)!$.
- There is a set of $(|U| - 1)!$ different permutations for each $s$ in $I$. $\Rightarrow |I|(|U| - 1)!$ permutations make $\arg\min_{s \in d_1} \pi(s) = \arg\min_{s \in d_2} \pi(s)$ true.
- Thus, the proportion of permutations that make $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$ true is:

$$\frac{|I|(|U| - 1)!}{|U|!} = \frac{|I|}{|U|} = J(d_1, d_2)$$
Thus, the proportion of successful permutations is the Jaccard coefficient.
Estimating Jaccard

• Thus, the proportion of successful permutations is the Jaccard coefficient.
  • Permutation $\pi$ is successful iff $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$
Thus, the proportion of successful permutations is the Jaccard coefficient.

- Permutation $\pi$ is successful iff $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$

- Picking a permutation at random and outputting 1 (successful) or 0 (unsuccessful) is a Bernoulli trial.
Estimating Jaccard

- Thus, the proportion of successful permutations is the Jaccard coefficient.
  - Permutation $\pi$ is successful iff $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$
- Picking a permutation at random and outputting 1 (successful) or 0 (unsuccessful) is a Bernoulli trial.
- Estimator of probability of success: proportion of successes in $n$ Bernoulli trials. ($n = 200$)
Estimating Jaccard

- Thus, the proportion of successful permutations is the Jaccard coefficient.
  - Permutation $\pi$ is successful iff $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$
- Picking a permutation at random and outputting 1 (successful) or 0 (unsuccessful) is a Bernoulli trial.
- Estimator of probability of success: proportion of successes in $n$ Bernoulli trials. ($n = 200$)
- Our sketch is based on a random selection of permutations.
Estimating Jaccard

- Thus, the proportion of successful permutations is the Jaccard coefficient.
  - Permutation $\pi$ is successful iff $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$
- Picking a permutation at random and outputting 1 (successful) or 0 (unsuccessful) is a Bernoulli trial.
- Estimator of probability of success: proportion of successes in $n$ Bernoulli trials. ($n = 200$)
- Our sketch is based on a random selection of permutations.
- Thus, to compute Jaccard, count the number $k$ of successful permutations for $< d_1, d_2 >$ and divide by $n = 200$. 

Estimating Jaccard

- Thus, the proportion of successful permutations is the Jaccard coefficient.
  - Permutation $\pi$ is successful iff $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$
- Picking a permutation at random and outputting 1 (successful) or 0 (unsuccessful) is a Bernoulli trial.
- Estimator of probability of success: proportion of successes in $n$ Bernoulli trials. ($n = 200$)
- Our sketch is based on a random selection of permutations.
- Thus, to compute Jaccard, count the number $k$ of successful permutations for $<d_1,d_2>$ and divide by $n = 200$.
- $k/n = k/200$ estimates $J(d_1,d_2)$. 

Schütze: Web search
Implementation
We use hash functions as an efficient type of permutation:

\[ h_i : \{1..2^m\} \rightarrow \{1..2^m\} \]
Implementation

- We use hash functions as an efficient type of permutation:
  \[ h_i : \{1..2^m\} \rightarrow \{1..2^m\} \]
- Scan all shingles \( s_k \) in union of two sets in arbitrary order
Implementation

- We use hash functions as an efficient type of permutation:
  \[ h_i : \{1..2^m\} \rightarrow \{1..2^m\} \]
- Scan all shingles \( s_k \) in union of two sets in arbitrary order
- For each hash function \( h_i \) and documents \( d_1, d_2, \ldots \): keep slot for minimum value found so far
Implementation

- We use **hash functions** as an efficient type of permutation:
  \[ h_i : \{1..2^m\} \rightarrow \{1..2^m\} \]
- Scan all shingles \( s_k \) in union of two sets in arbitrary order
- For each hash function \( h_i \) and documents \( d_1, d_2, \ldots \): keep slot for minimum value found so far
- If \( h_i(s_k) \) is lower than minimum found so far: update slot
Example
### Example

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$

$g(x) = (2x + 1) \mod 5$
Example

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$

$g(x) = (2x + 1) \mod 5$

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>h</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example

<table>
<thead>
<tr>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$

$g(x) = (2x + 1) \mod 5$
Example

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$

$g(x) = (2x + 1) \mod 5$

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
</tbody>
</table>

$h(1) = 1$
$g(1) = 3$
$h(2) = 2$
$g(2) = 0$
$h(3) = 3$
$g(3) = 2$
$h(4) = 4$
$g(4) = 4$
$h(5) = 0$
$g(5) = 1$
Example

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$

$g(x) = (2x + 1) \mod 5$

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
</tbody>
</table>

$h(1) = 1$

$g(1) = 3$

$h(2) = 2$

$g(2) = 0$

$h(3) = 3$

$g(3) = 2$

$h(4) = 4$

$g(4) = 4$

$h(5) = 0$

$g(5) = 1$
### Example

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$

$g(x) = (2x + 1) \mod 5$

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$h(1) = 1$</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>$g(1) = 3$</td>
<td>3</td>
<td>–</td>
</tr>
<tr>
<td>$h(2) = 2$</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$g(2) = 0$</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$h(3) = 3$</td>
<td>3</td>
<td>–</td>
</tr>
<tr>
<td>$g(3) = 2$</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>$h(4) = 4$</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$g(4) = 4$</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$h(5) = 0$</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$g(5) = 1$</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
Example

<table>
<thead>
<tr>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1 0</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0 1</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1 1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1 0</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0 1</td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$

$g(x) = (2x + 1) \mod 5$

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h(1)$ = 1</td>
<td>1 1</td>
<td>–</td>
</tr>
<tr>
<td>$g(1)$ = 3</td>
<td>3 3</td>
<td>–</td>
</tr>
<tr>
<td>$h(2)$ = 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g(2)$ = 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h(3)$ = 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g(3)$ = 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h(4)$ = 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g(4)$ = 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h(5)$ = 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g(5)$ = 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example

\[
\begin{array}{c|cc|c|c}
 & d_1 & d_2 & d_1 \text{ slot} & d_2 \text{ slot} \\
\hline
h & \infty & \infty \\
g & \infty & \infty \\
\hline
h(1) = 1 & 1 & 1 & - & \infty \\
g(1) = 3 & 3 & 3 & - & \infty \\
\hline
h(2) = 2 & 1 & 1 & - & \infty \\
g(2) = 0 & 3 & 3 & - & \infty \\
\hline
h(3) = 3 & 1 & 1 & - & \infty \\
g(3) = 2 & 3 & 3 & - & \infty \\
\hline
h(4) = 4 & 1 & 1 & - & \infty \\
g(4) = 4 & 3 & 3 & - & \infty \\
\hline
h(5) = 0 & 1 & 1 & - & \infty \\
g(5) = 1 & 3 & 3 & - & \infty \\
\end{array}
\]

\[
h(x) = x \mod 5
\]

\[
g(x) = (2x + 1) \mod 5
\]
### Example

<table>
<thead>
<tr>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$

$g(x) = (2x + 1) \mod 5$

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$h(1) = 1$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$g(1) = 3$</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$h(2) = 2$</td>
<td>$-\infty$</td>
<td>$-\infty$</td>
</tr>
<tr>
<td>$g(2) = 0$</td>
<td>$-\infty$</td>
<td>$-\infty$</td>
</tr>
<tr>
<td>$h(3) = 3$</td>
<td>$-\infty$</td>
<td>$-\infty$</td>
</tr>
<tr>
<td>$g(3) = 2$</td>
<td>$-\infty$</td>
<td>$-\infty$</td>
</tr>
<tr>
<td>$h(4) = 4$</td>
<td>$-\infty$</td>
<td>$-\infty$</td>
</tr>
<tr>
<td>$g(4) = 4$</td>
<td>$-\infty$</td>
<td>$-\infty$</td>
</tr>
<tr>
<td>$h(5) = 0$</td>
<td>$-\infty$</td>
<td>$-\infty$</td>
</tr>
<tr>
<td>$g(5) = 1$</td>
<td>$-\infty$</td>
<td>$-\infty$</td>
</tr>
</tbody>
</table>
### Example

<table>
<thead>
<tr>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$

$g(x) = (2x + 1) \mod 5$

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$h(1) = 1$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$g(1) = 3$</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$h(2) = 2$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g(2) = 0$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$h(3) = 3$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g(3) = 2$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$h(4) = 4$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g(4) = 4$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$h(5) = 0$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g(5) = 1$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
</tbody>
</table>
Example

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$

$g(x) = (2x + 1) \mod 5$

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$h(1) = 1$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$g(1) = 3$</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$h(2) = 2$</td>
<td>$-1$</td>
<td>2</td>
</tr>
<tr>
<td>$g(2) = 0$</td>
<td>$-3$</td>
<td>0</td>
</tr>
<tr>
<td>$h(3) = 3$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g(3) = 2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h(4) = 4$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g(4) = 4$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h(5) = 0$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g(5) = 1$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example

\[
\begin{array}{cc}
d_1 & d_2 \\
\hline
s_1 & 0 \hspace{1cm} 1 \\
s_2 & 1 \hspace{1cm} 0 \\
s_3 & 1 \hspace{1cm} 1 \\
s_4 & 1 \hspace{1cm} 0 \\
s_5 & 1 \hspace{1cm} 0 \\
\end{array}
\]

\begin{align*}
h(x) &= x \mod 5 \\
g(x) &= (2x + 1) \mod 5
\end{align*}

\[
\begin{array}{|c|c|c|}
\hline
 & d_1 \text{ slot} & d_2 \text{ slot} \\
\hline
h & \infty & \infty \\
g & \infty & \infty \\
\hline
h(1) &= 1 & 1 & 1 & \infty \\
g(1) &= 3 & 3 & 3 & \infty \\
\hline
h(2) &= 2 & – & 1 & 2 & 2 \\
g(2) &= 0 & – & 3 & 0 & 0 \\
\hline
h(3) &= 3 & & & \\
g(3) &= 2 & & & \\
\hline
h(4) &= 4 & & & \\
g(4) &= 4 & & & \\
\hline
h(5) &= 0 & & & \\
g(5) &= 1 & & & \\
\hline
\end{array}
\]
Example

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$

$g(x) = (2x + 1) \mod 5$

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
</tbody>
</table>

$h(1) = 1$

$g(1) = 3$

$h(2) = 2$

$g(2) = 0$

$h(3) = 3$

$g(3) = 2$

$h(4) = 4$

$g(4) = 4$

$h(5) = 0$

$g(5) = 1$
Example

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$

$g(x) = (2x + 1) \mod 5$

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$h(1) = 1$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$g(1) = 3$</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$h(2) = 2$</td>
<td>$- 1$</td>
<td>2</td>
</tr>
<tr>
<td>$g(2) = 0$</td>
<td>$- 3$</td>
<td>0</td>
</tr>
<tr>
<td>$h(3) = 3$</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$g(3) = 2$</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$h(4) = 4$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g(4) = 4$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h(5) = 0$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g(5) = 1$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Example

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
</tbody>
</table>

- $h(1) = 1$  \quad $g(1) = 3$
- $h(2) = 2$  \quad $g(2) = 0$
- $h(3) = 3$  \quad $g(3) = 2$
- $h(4) = 4$  \quad $g(4) = 4$
- $h(5) = 0$  \quad $g(5) = 1$
### Example

<table>
<thead>
<tr>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$

$g(x) = (2x + 1) \mod 5$

<table>
<thead>
<tr>
<th>$d_1$</th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
</tbody>
</table>

| $h(1) = 1$ | $1$ | $\infty$ |
| $g(1) = 3$ | $3$ | $\infty$ |
| $h(2) = 2$ | $-1$ | $2$ |
| $g(2) = 0$ | $-3$ | $0$ |
| $h(3) = 3$ | $3$ | $3$ |
| $g(3) = 2$ | $2$ | $2$ |
| $h(4) = 4$ | $2$ | $0$ |
| $g(4) = 4$ | $2$ | $0$ |
| $h(5) = 0$ | $2$ | $0$ |
| $g(5) = 1$ | $2$ | $0$ |
**Example**

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
<td>0</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
<td>1</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
<td>1</td>
<td>$-\infty$</td>
<td>2</td>
</tr>
<tr>
<td>$h(x) = x \mod 5$</td>
<td></td>
<td></td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>$g(x) = (2x + 1) \mod 5$</td>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>$h(1) = 1$</td>
<td>1</td>
<td>1</td>
<td>$-\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g(1) = 3$</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$h(2) = 2$</td>
<td>$-1$</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>$g(2) = 0$</td>
<td>$-3$</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>$h(3) = 3$</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>$g(3) = 2$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>$h(4) = 4$</td>
<td>4</td>
<td></td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>$g(4) = 4$</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h(5) = 0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g(5) = 1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Example

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$

$g(x) = (2x + 1) \mod 5$

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$h(1) = 1$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$g(1) = 3$</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$h(2) = 2$</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>$g(2) = 0$</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>$h(3) = 3$</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>$g(3) = 2$</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$h(4) = 4$</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>$g(4) = 4$</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>$h(5) = 0$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$g(5) = 1$</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
### Example

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$

$g(x) = (2x + 1) \mod 5$

<table>
<thead>
<tr>
<th>$h(x)$</th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h(1) = 1$</td>
<td>1 1</td>
<td>$\infty$ $\infty$</td>
</tr>
<tr>
<td>$g(1) = 3$</td>
<td>3 3</td>
<td>$\infty$ $\infty$</td>
</tr>
<tr>
<td>$h(2) = 2$</td>
<td>$-1$</td>
<td>2 2</td>
</tr>
<tr>
<td>$g(2) = 0$</td>
<td>$-3$</td>
<td>0 0</td>
</tr>
<tr>
<td>$h(3) = 3$</td>
<td>3 1</td>
<td>3 2</td>
</tr>
<tr>
<td>$g(3) = 2$</td>
<td>2 2</td>
<td>2 0</td>
</tr>
<tr>
<td>$h(4) = 4$</td>
<td>4 1</td>
<td>$-$</td>
</tr>
<tr>
<td>$g(4) = 4$</td>
<td>4 2</td>
<td>$-$</td>
</tr>
<tr>
<td>$h(5) = 0$</td>
<td>$-$</td>
<td>$-$</td>
</tr>
<tr>
<td>$g(5) = 1$</td>
<td>$-$</td>
<td>$-$</td>
</tr>
</tbody>
</table>
Example

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$

$g(x) = (2x + 1) \mod 5$

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$h(1) = 1$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$g(1) = 3$</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$h(2) = 2$</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>$g(2) = 0$</td>
<td>-3</td>
<td>-3</td>
</tr>
<tr>
<td>$h(3) = 3$</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$g(3) = 2$</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$h(4) = 4$</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>$g(4) = 4$</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>$h(5) = 0$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g(5) = 1$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Example

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$

$g(x) = (2x + 1) \mod 5$

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$h(1) = 1$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$g(1) = 3$</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$h(2) = 2$</td>
<td>-1</td>
<td>2</td>
</tr>
<tr>
<td>$g(2) = 0$</td>
<td>-3</td>
<td>0</td>
</tr>
<tr>
<td>$h(3) = 3$</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>$g(3) = 2$</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$h(4) = 4$</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>$g(4) = 4$</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>$h(5) = 0$</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>$g(5) = 1$</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Example

\[
\begin{array}{c|c|c}
& d_1 & d_2 \\
\hline
s_1 & 1 & 0 \\
s_2 & 0 & 1 \\
s_3 & 1 & 1 \\
s_4 & 1 & 0 \\
s_5 & 0 & 1 \\
\end{array}
\]

\[
h(x) = x \mod 5 \\
g(x) = (2x + 1) \mod 5
\]

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$h(1) = 1$</td>
<td>1 1</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g(1) = 3$</td>
<td>3 3</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$h(2) = 2$</td>
<td>$\infty$</td>
<td>2 2</td>
</tr>
<tr>
<td>$g(2) = 0$</td>
<td>3 $\infty$</td>
<td>0 0</td>
</tr>
<tr>
<td>$h(3) = 3$</td>
<td>3 1</td>
<td>3 2</td>
</tr>
<tr>
<td>$g(3) = 2$</td>
<td>2 2</td>
<td>2 0</td>
</tr>
<tr>
<td>$h(4) = 4$</td>
<td>4 1</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g(4) = 4$</td>
<td>4 2</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$h(5) = 0$</td>
<td>$\infty$</td>
<td>0</td>
</tr>
<tr>
<td>$g(5) = 1$</td>
<td>$\infty$</td>
<td>1</td>
</tr>
</tbody>
</table>
### Example

\[
\begin{array}{c|c|c|c}
& d_1 & d_2 & \text{d}_1 \text{ slot} & \text{d}_2 \text{ slot} \\
\hline
h & \infty & \infty & \\
g & \infty & \infty & \\
\hline
h(1) = 1 & 1 & 1 & - \infty \\
g(1) = 3 & 3 & 3 & - \infty \\
\hline
h(2) = 2 & -1 & 2 & 2 \\
g(2) = 0 & -3 & 0 & 0 \\
\hline
h(3) = 3 & 3 & 1 & 3 \\
g(3) = 2 & 2 & 2 & 0 \\
\hline
h(4) = 4 & 4 & 1 & -2 \\
g(4) = 4 & 4 & 2 & -0 \\
\hline
h(5) = 0 & -1 & - & 0 \\
g(5) = 1 & -2 & 1 & \\
\end{array}
\]

\[h(x) = x \mod 5\]
\[g(x) = (2x + 1) \mod 5\]
Example

<table>
<thead>
<tr>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$

$g(x) = (2x + 1) \mod 5$

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$h(1) = 1$</td>
<td>1 1</td>
<td>– $\infty$</td>
</tr>
<tr>
<td>$g(1) = 3$</td>
<td>3 3</td>
<td>– $\infty$</td>
</tr>
<tr>
<td>$h(2) = 2$</td>
<td>– 1</td>
<td>2 2</td>
</tr>
<tr>
<td>$g(2) = 0$</td>
<td>– 3</td>
<td>0 0</td>
</tr>
<tr>
<td>$h(3) = 3$</td>
<td>3 1</td>
<td>3 2</td>
</tr>
<tr>
<td>$g(3) = 2$</td>
<td>2 2</td>
<td>2 0</td>
</tr>
<tr>
<td>$h(4) = 4$</td>
<td>4 1</td>
<td>– 2</td>
</tr>
<tr>
<td>$g(4) = 4$</td>
<td>4 2</td>
<td>– 0</td>
</tr>
<tr>
<td>$h(5) = 0$</td>
<td>– 1</td>
<td>0 0</td>
</tr>
<tr>
<td>$g(5) = 1$</td>
<td>– 2</td>
<td>1 0</td>
</tr>
</tbody>
</table>
Example

\[ d_1 \quad d_2 \]
\[
\begin{array}{ccc}
 s_1 & 1 & 0 \\
 s_2 & 0 & 1 \\
 s_3 & 1 & 1 \\
 s_4 & 1 & 0 \\
 s_5 & 0 & 1 \\
\end{array}
\]
\[ h(x) = x \mod 5 \]
\[ g(x) = (2x + 1) \mod 5 \]

<table>
<thead>
<tr>
<th>( h(x) )</th>
<th>( d_1 ) slot</th>
<th>( d_2 ) slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>( h(1) = 1 )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( g(1) = 3 )</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>( h(2) = 2 )</td>
<td>(-)</td>
<td>1</td>
</tr>
<tr>
<td>( g(2) = 0 )</td>
<td>(-)</td>
<td>3</td>
</tr>
<tr>
<td>( h(3) = 3 )</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>( g(3) = 2 )</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>( h(4) = 4 )</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>( g(4) = 4 )</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>( h(5) = 0 )</td>
<td>(-)</td>
<td>1</td>
</tr>
<tr>
<td>( g(5) = 1 )</td>
<td>(-)</td>
<td>2</td>
</tr>
</tbody>
</table>

final sketches
Example

<table>
<thead>
<tr>
<th>d₁</th>
<th>d₂</th>
<th>d₁ slot</th>
<th>d₂ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>h</td>
<td>∞</td>
<td>∞</td>
<td></td>
</tr>
<tr>
<td>g</td>
<td>∞</td>
<td>∞</td>
<td></td>
</tr>
<tr>
<td>h(1) = 1</td>
<td>1</td>
<td>1</td>
<td>– ∞</td>
</tr>
<tr>
<td>g(1) = 3</td>
<td>3</td>
<td>3</td>
<td>– ∞</td>
</tr>
<tr>
<td>h(2) = 2</td>
<td>– 1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>g(2) = 0</td>
<td>– 3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>h(3) = 3</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>g(3) = 2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>h(4) = 4</td>
<td>4</td>
<td>1</td>
<td>– 2</td>
</tr>
<tr>
<td>g(4) = 4</td>
<td>4</td>
<td>2</td>
<td>– 0</td>
</tr>
<tr>
<td>h(5) = 0</td>
<td>– 1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>g(5) = 1</td>
<td>– 2</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

final sketches
Example

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_2$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$h(x) = x \mod 5$
$g(x) = (2x + 1) \mod 5$

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$g$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$h(1) = 1$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$g(1) = 3$</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$h(2) = 2$</td>
<td>–</td>
<td>1</td>
</tr>
<tr>
<td>$g(2) = 0$</td>
<td>–</td>
<td>3</td>
</tr>
<tr>
<td>$h(3) = 3$</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>$g(3) = 2$</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$h(4) = 4$</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>$g(4) = 4$</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>$h(5) = 0$</td>
<td>–</td>
<td>1</td>
</tr>
<tr>
<td>$g(5) = 1$</td>
<td>–</td>
<td>2</td>
</tr>
</tbody>
</table>

final sketches
Example

\[
\begin{align*}
  d_1 & \quad d_2 \\
  s_1 & \quad 1 \quad 0 \\
  s_2 & \quad 0 \quad 1 \\
  s_3 & \quad 1 \quad 1 \\
  s_4 & \quad 1 \quad 0 \\
  s_5 & \quad 0 \quad 1 \\
  h(x) &= x \mod 5 \\
  g(x) &= (2x + 1) \mod 5
\end{align*}
\]

\[
\begin{align*}
  \min(h(d_1)) &= 1 \neq 0 = \\
  \min(h(d_2)) &= \\
  \min(g(d_1)) &= 2 \neq 0 = \\
  \min(g(d_2)) &= \\
  \hat{J}(d_1, d_2) &= \frac{0+0}{2} = 0
\end{align*}
\]

<table>
<thead>
<tr>
<th></th>
<th>(d_1) slot</th>
<th>(d_2) slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>(h)</td>
<td>(\infty)</td>
<td>(\infty)</td>
</tr>
<tr>
<td>(g)</td>
<td>(\infty)</td>
<td>(\infty)</td>
</tr>
<tr>
<td>(h(1) = 1)</td>
<td>1 1</td>
<td>- (\infty)</td>
</tr>
<tr>
<td>(g(1) = 3)</td>
<td>3 3</td>
<td>- (\infty)</td>
</tr>
<tr>
<td>(h(2) = 2)</td>
<td>- 1</td>
<td>2 2</td>
</tr>
<tr>
<td>(g(2) = 0)</td>
<td>- 3</td>
<td>0 0</td>
</tr>
<tr>
<td>(h(3) = 3)</td>
<td>3 1</td>
<td>3 2</td>
</tr>
<tr>
<td>(g(3) = 2)</td>
<td>2 2</td>
<td>2 0</td>
</tr>
<tr>
<td>(h(4) = 4)</td>
<td>4 1</td>
<td>- 2</td>
</tr>
<tr>
<td>(g(4) = 4)</td>
<td>4 2</td>
<td>- 0</td>
</tr>
<tr>
<td>(h(5) = 0)</td>
<td>- 1</td>
<td>0 0</td>
</tr>
<tr>
<td>(g(5) = 1)</td>
<td>- 2</td>
<td>1 0</td>
</tr>
</tbody>
</table>

final sketches
Exercise

<table>
<thead>
<tr>
<th>Recap</th>
<th>Big picture</th>
<th>Ads</th>
<th>Duplicate detection</th>
<th>Spam</th>
<th>Web IR</th>
<th>Size of the web</th>
</tr>
</thead>
</table>

Schütze: Web search
Exercise

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$s_2$</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$s_3$</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$h(x) = 5x + 5 \mod 4$

$g(x) = (3x + 1) \mod 4$

Estimate $\hat{J}(d_1, d_2)$, $\hat{J}(d_1, d_3)$, $\hat{J}(d_2, d_3)$
Solution (1)
Solution (1)
### Solution (1)

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$s_2$</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$s_3$</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$h(x) = 5x + 5 \mod 4$

$g(x) = (3x + 1) \mod 4$

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
<th>$d_3$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h(1) = 2$</td>
<td>$- \infty$</td>
<td>$2$</td>
<td>$2$</td>
</tr>
<tr>
<td>$g(1) = 0$</td>
<td>$- \infty$</td>
<td>$0$</td>
<td>$0$</td>
</tr>
<tr>
<td>$h(2) = 3$</td>
<td>$3$</td>
<td>$3$</td>
<td>$-2$</td>
</tr>
<tr>
<td>$g(2) = 3$</td>
<td>$3$</td>
<td>$3$</td>
<td>$-0$</td>
</tr>
<tr>
<td>$h(3) = 0$</td>
<td>$-3$</td>
<td>$0$</td>
<td>$0$</td>
</tr>
<tr>
<td>$g(3) = 2$</td>
<td>$-3$</td>
<td>$2$</td>
<td>$0$</td>
</tr>
<tr>
<td>$h(4) = 1$</td>
<td>$1$</td>
<td>$1$</td>
<td>$-0$</td>
</tr>
<tr>
<td>$g(4) = 1$</td>
<td>$1$</td>
<td>$1$</td>
<td>$-0$</td>
</tr>
</tbody>
</table>
Solution (1)

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$s_2$</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$s_3$</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
h(x) = 5x + 5 \mod 4
\]
\[
g(x) = (3x + 1) \mod 4
\]

<table>
<thead>
<tr>
<th></th>
<th>$d_1$ slot</th>
<th>$d_2$ slot</th>
<th>$d_3$ slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$s_2$</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$h(1) = 2$</td>
<td>$-\infty$</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$g(1) = 0$</td>
<td>$-\infty$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$h(2) = 3$</td>
<td>3</td>
<td>3</td>
<td>$-2$</td>
</tr>
<tr>
<td>$g(2) = 3$</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>$h(3) = 0$</td>
<td>$-3$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$g(3) = 2$</td>
<td>$-3$</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>$h(4) = 1$</td>
<td>1</td>
<td>1</td>
<td>$-0$</td>
</tr>
<tr>
<td>$g(4) = 1$</td>
<td>1</td>
<td>1</td>
<td>$-0$</td>
</tr>
</tbody>
</table>

final sketches
Solution (2)
Solution (2)

\[
\hat{J}(d_1, d_2) = \frac{0 + 0}{2} = 0
\]

\[
\hat{J}(d_1, d_3) = \frac{0 + 0}{2} = 0
\]

\[
\hat{J}(d_2, d_3) = \frac{0 + 1}{2} = 1/2
\]
Shingling: Summary
Shingling: Summary

- Input: $N$ documents
Shingling: Summary

- Input: $N$ documents
- Choose n-gram size for shingling, e.g., $n = 5$
Shingling: Summary

- Input: $N$ documents
- Choose n-gram size for shingling, e.g., $n = 5$
- Pick 200 random permutations, represented as hash functions
Shingling: Summary

- Input: $N$ documents
- Choose n-gram size for shingling, e.g., $n = 5$
- Pick 200 random permutations, represented as hash functions
- Compute $N$ sketches: $200 \times N$ matrix shown on previous slide, one row per permutation, one column per document
Shingling: Summary

- Input: \( N \) documents
- Choose n-gram size for shingling, e.g., \( n = 5 \)
- Pick 200 random permutations, represented as hash functions
- Compute \( N \) sketches: \( 200 \times N \) matrix shown on previous slide, one row per permutation, one column per document
- Compute \( \frac{N \cdot (N-1)}{2} \) pairwise similarities
Shingling: Summary

- Input: \( N \) documents
- Choose n-gram size for shingling, e.g., \( n = 5 \)
- Pick 200 random permutations, represented as hash functions
- Compute \( N \) sketches: \( 200 \times N \) matrix shown on previous slide, one row per permutation, one column per document
- Compute \( \frac{N \cdot (N-1)}{2} \) pairwise similarities
- Transitive closure of documents with similarity > \( \theta \)
Shingling: Summary

- Input: $N$ documents
- Choose n-gram size for shingling, e.g., $n = 5$
- Pick 200 random permutations, represented as hash functions
- Compute $N$ sketches: $200 \times N$ matrix shown on previous slide, one row per permutation, one column per document
- Compute $\frac{N \cdot (N-1)}{2}$ pairwise similarities
- Transitive closure of documents with similarity $> \theta$
- Index only one document from each equivalence class
Efficient near-duplicate detection
Efficient near-duplicate detection

Now we have an extremely efficient method for estimating a Jaccard coefficient for a single pair of two documents.
Now we have an extremely efficient method for estimating a Jaccard coefficient for a single pair of two documents.

But we still have to estimate $O(N^2)$ coefficients where $N$ is the number of web pages.
Efficient near-duplicate detection

- Now we have an extremely efficient method for estimating a Jaccard coefficient for a single pair of two documents.
- But we still have to estimate $O(N^2)$ coefficients where $N$ is the number of web pages.
- Still intractable
Efficient near-duplicate detection

- Now we have an extremely efficient method for estimating a Jaccard coefficient for a single pair of two documents.
- But we still have to estimate $O(N^2)$ coefficients where $N$ is the number of web pages.
- Still intractable
- One solution: locality sensitive hashing (LSH)
Now we have an extremely efficient method for estimating a Jaccard coefficient for a single pair of two documents.

But we still have to estimate $O(N^2)$ coefficients where $N$ is the number of web pages.

Still intractable

One solution: locality sensitive hashing (LSH)

Another solution: sorting (Henzinger 2006)
Take-away today

- Big picture
- Ads – they pay for the web
- Duplicate detection – addresses one aspect of chaotic content creation
- Spam detection – addresses one aspect of lack of central access control
- Probably won’t get to today
  - Web information retrieval
  - Size of the web
Outline

1 Recap
2 Big picture
3 Ads
4 Duplicate detection
5 Spam
6 Web IR
   • Queries
   • Links
   • Context
   • Users
   • Documents
   • Size
7 Size of the web
The goal of spamming on the web
The goal of spamming on the web

- You have a page that will generate lots of revenue for you if people visit it.
The goal of spamming on the web

- You have a page that will generate lots of revenue for you if people visit it.
- Therefore, you would like to direct visitors to this page.
The goal of spamming on the web

- You have a page that will generate lots of revenue for you if people visit it.
- Therefore, you would like to direct visitors to this page.
- One way of doing this: get your page ranked highly in search results.
You have a page that will generate lots of revenue for you if people visit it.

Therefore, you would like to direct visitors to this page.

One way of doing this: get your page ranked highly in search results.

Exercise: How can I get my page ranked highly?
Spam technique: Keyword stuffing / Hidden text
Spam technique: Keyword stuffing / Hidden text

- Misleading meta-tags, excessive repetition
Spam technique: Keyword stuffing / Hidden text

- Misleading meta-tags, excessive repetition
- Hidden text with colors, style sheet tricks etc.
Spam technique: Keyword stuffing / Hidden text

- Misleading meta-tags, excessive repetition
- Hidden text with colors, style sheet tricks etc.
- Used to be very effective, most search engines now catch these
Keyword stuffing
Keyword stuffing

Something is happening in general sales-and-decreasing put consigo underwater health insurance. It paid companies may every tax deferred. boy each arm...
Spam technique: Doorway and lander pages
Spam technique: Doorway and lander pages

- Doorway page: optimized for a single keyword, redirects to the real target page
Spam technique: Doorway and lander pages

- **Doorway page**: optimized for a single keyword, redirects to the real target page
- **Lander page**: optimized for a single keyword or a misspelled domain name, designed to attract surfers who will then click on ads
Lander page
Number one hit on Google for the search “composita”
Lander page

- Number one hit on Google for the search “composita”
- The only purpose of this page: get people to click on the ads and make money for the page owner
Spam technique: Duplication
Spam technique: Duplication

- Get good content from somewhere (steal it or produce it yourself)
Spam technique: Duplication

- Get good content from somewhere (steal it or produce it yourself)
- Publish a large number of slight variations of it
Spam technique: Duplication

- Get good content from somewhere (steal it or produce it yourself)
- Publish a large number of slight variations of it
- For example, publish the answer to a tax question with the spelling variations of “tax deferred” on the previous slide
Spam technique: Cloaking
Spam technique: Cloaking

- Serve fake content to search engine spider
Spam technique: Cloaking

- Serve fake content to search engine spider
- So do we just penalize this always?
Spam technique: Cloaking

- Serve fake content to search engine spider
- So do we just penalize this always?
- No: legitimate uses (e.g., different content to US vs. European users)
Spam technique: Link spam
Spam technique: Link spam

- Create lots of links pointing to the page you want to promote
Spam technique: Link spam

- Create lots of links pointing to the page you want to promote
- Put these links on pages with high (or at least non-zero) PageRank
Spam technique: Link spam

- Create lots of links pointing to the page you want to promote
- Put these links on pages with high (or at least non-zero) PageRank
  - Newly registered domains (domain flooding)
Spam technique: Link spam

- Create lots of links pointing to the page you want to promote
- Put these links on pages with high (or at least non-zero) PageRank
  - Newly registered domains (domain flooding)
  - A set of pages that all point to each other to boost each other’s PageRank (mutual admiration society)
Spam technique: Link spam

- Create lots of links pointing to the page you want to promote
- Put these links on pages with high (or at least non-zero) PageRank
  - Newly registered domains (domain flooding)
  - A set of pages that all point to each other to boost each other’s PageRank (mutual admiration society)
  - Pay somebody to put your link on their highly ranked page (“schuetze horoskop” example)
Spam technique: Link spam

- Create lots of links pointing to the page you want to promote
- Put these links on pages with high (or at least non-zero) PageRank
  - Newly registered domains (domain flooding)
  - A set of pages that all point to each other to boost each other’s PageRank (mutual admiration society)
  - Pay somebody to put your link on their highly ranked page ("schuetze horoskop" example)
  - Leave comments that include the link on blogs
SEO: Search engine optimization
SEO: Search engine optimization

- Promoting a page in the search rankings is not necessarily spam.
SEO: Search engine optimization

- Promoting a page in the search rankings is not necessarily spam.
- It can also be a legitimate business – which is called SEO.
Promoting a page in the search rankings is not necessarily spam.

It can also be a legitimate business – which is called SEO.

You can hire an SEO firm to get your page highly ranked.
Promoting a page in the search rankings is not necessarily spam.

It can also be a legitimate business – which is called SEO.

You can hire an SEO firm to get your page highly ranked.

There are many legitimate reasons for doing this.
SEO: Search engine optimization

- Promoting a page in the search rankings is not necessarily spam.
- It can also be a legitimate business – which is called SEO.
- You can hire an SEO firm to get your page highly ranked.
- There are many legitimate reasons for doing this.
  - For example, Google bombs like *Who is a failure?*
SEO: Search engine optimization

- Promoting a page in the search rankings is not necessarily spam.
- It can also be a legitimate business – which is called SEO.
- You can hire an SEO firm to get your page highly ranked.
- There are many legitimate reasons for doing this.
  - For example, Google bombs like *Who is a failure?*
- And there are many legitimate ways of achieving this:
Promoting a page in the search rankings is not necessarily spam.

It can also be a legitimate business – which is called SEO.

You can hire an SEO firm to get your page highly ranked.

There are many legitimate reasons for doing this.

- For example, Google bombs like *Who is a failure?*

And there are many legitimate ways of achieving this:

- Restructure your content in a way that makes it easy to index
SEO: Search engine optimization

- Promoting a page in the search rankings is not necessarily spam.
- It can also be a legitimate business – which is called SEO.
- You can hire an SEO firm to get your page highly ranked.
- There are many legitimate reasons for doing this.
  - For example, Google bombs like *Who is a failure?*
- And there are many legitimate ways of achieving this:
  - Restructure your content in a way that makes it easy to index
  - Talk with influential bloggers and have them link to your site
Promoting a page in the search rankings is not necessarily spam.

It can also be a legitimate business – which is called SEO.

You can hire an SEO firm to get your page highly ranked.

There are many legitimate reasons for doing this.

- For example, Google bombs like *Who is a failure?*

And there are many legitimate ways of achieving this:

- Restructure your content in a way that makes it easy to index
- Talk with influential bloggers and have them link to your site
- Add more interesting and original content
The war against spam
The war against spam

- Quality indicators
The war against spam

- Quality indicators
  - Links, statistically analyzed (PageRank etc)
The war against spam

- Quality indicators
  - Links, statistically analyzed (PageRank etc)
  - Usage (users visiting a page)
The war against spam

Quality indicators
- Links, statistically analyzed (PageRank etc)
- Usage (users visiting a page)
- No adult content (e.g., no pictures with flesh-tone)
The war against spam

- Quality indicators
  - Links, statistically analyzed (PageRank etc)
  - Usage (users visiting a page)
  - No adult content (e.g., no pictures with flesh-tone)
  - Distribution and structure of text (e.g., no keyword stuffing)
The war against spam

- Quality indicators
  - Links, statistically analyzed (PageRank etc)
  - Usage (users visiting a page)
  - No adult content (e.g., no pictures with flesh-tone)
  - Distribution and structure of text (e.g., no keyword stuffing)

- Combine all of these indicators and use machine learning
The war against spam

- Quality indicators
  - Links, statistically analyzed (PageRank etc)
  - Usage (users visiting a page)
  - No adult content (e.g., no pictures with flesh-tone)
  - Distribution and structure of text (e.g., no keyword stuffing)
- Combine all of these indicators and use machine learning
- Editorial intervention
The war against spam

Quality indicators

- Links, statistically analyzed (PageRank etc)
- Usage (users visiting a page)
- No adult content (e.g., no pictures with flesh-tone)
- Distribution and structure of text (e.g., no keyword stuffing)

Combine all of these indicators and use machine learning

Editorial intervention

- Blacklists
The war against spam

- Quality indicators
  - Links, statistically analyzed (PageRank etc)
  - Usage (users visiting a page)
  - No adult content (e.g., no pictures with flesh-tone)
  - Distribution and structure of text (e.g., no keyword stuffing)

- Combine all of these indicators and use machine learning

- Editorial intervention
  - Blacklists
  - Top queries audited
The war against spam

**Quality indicators**
- Links, statistically analyzed (PageRank etc).
- Usage (users visiting a page).
- No adult content (e.g., no pictures with flesh-tone).
- Distribution and structure of text (e.g., no keyword stuffing).

**Combine all of these indicators and use machine learning.**

**Editorial intervention**
- Blacklists.
- Top queries audited.
- Complaints addressed.
The war against spam

- Quality indicators
  - Links, statistically analyzed (PageRank etc)
  - Usage (users visiting a page)
  - No adult content (e.g., no pictures with flesh-tone)
  - Distribution and structure of text (e.g., no keyword stuffing)

- Combine all of these indicators and use machine learning

- Editorial intervention
  - Blacklists
  - Top queries audited
  - Complaints addressed
  - Suspect patterns detected
Webmaster guidelines
Webmaster guidelines

- Major search engines have guidelines for webmasters.
Webmaster guidelines

- Major search engines have guidelines for webmasters.
- These guidelines tell you what is legitimate SEO and what is spamming.
Webmaster guidelines

- Major search engines have guidelines for webmasters.
- These guidelines tell you what is legitimate SEO and what is spamming.
- Ignore these guidelines at your own risk.
Webmaster guidelines

- Major search engines have guidelines for webmasters.
- These guidelines tell you what is legitimate SEO and what is spamming.
- Ignore these guidelines at your own risk
- Once a search engine identifies you as a spammer, all pages on your site may get low ranks (or disappear from the index entirely).
Webmaster guidelines

- Major search engines have guidelines for webmasters.
- These guidelines tell you what is legitimate SEO and what is spamming.
- Ignore these guidelines at your own risk.
- Once a search engine identifies you as a spammer, all pages on your site may get low ranks (or disappear from the index entirely).
- There is often a fine line between spam and legitimate SEO.
Webmaster guidelines

- Major search engines have guidelines for webmasters.
- These guidelines tell you what is legitimate SEO and what is spamming.
- Ignore these guidelines at your own risk
- Once a search engine identifies you as a spammer, all pages on your site may get low ranks (or disappear from the index entirely).
- There is often a fine line between spam and legitimate SEO.
- Scientific study of fighting spam on the web: *adversarial information retrieval*
Outline

1 Recap
2 Big picture
3 Ads
4 Duplicate detection
5 Spam
6 Web IR
   - Queries
   - Links
   - Context
   - Users
   - Documents
   - Size
7 Size of the web
Web IR: Differences from traditional IR
Web IR: Differences from traditional IR

- Links: The web is a hyperlinked document collection.
Web IR: Differences from traditional IR

- Links: The web is a hyperlinked document collection.
- Queries: Web queries are different, more varied and there are a lot of them.
Web IR: Differences from traditional IR

- Links: The web is a hyperlinked document collection.
- Queries: Web queries are different, more varied and there are a lot of them.
- Users: Users are different, more varied and there are a lot of them.
Web IR: Differences from traditional IR

- Links: The web is a hyperlinked document collection.
- Queries: Web queries are different, more varied and there are a lot of them.
- Users: Users are different, more varied and there are a lot of them.
- Documents: Documents are different, more varied and there are a lot of them.
Web IR: Differences from traditional IR

- **Links**: The web is a hyperlinked document collection.
- **Queries**: Web queries are different, more varied and there are a lot of them.
- **Users**: Users are different, more varied and there are a lot of them.
- **Documents**: Documents are different, more varied and there are a lot of them.
- **Context**: Context is more important on the web than in many other IR applications.
Web IR: Differences from traditional IR

- **Links:** The web is a hyperlinked document collection.
- **Queries:** Web queries are different, more varied and there are a lot of them. How many?
- **Users:** Users are different, more varied and there are a lot of them. How many?
- **Documents:** Documents are different, more varied and there are a lot of them. How many?
- **Context:** Context is more important on the web than in many other IR applications.
- **Ads and spam**
Web IR: Differences from traditional IR

- Links: The web is a hyperlinked document collection.
- Queries: Web queries are different, more varied and there are a lot of them. How many? $\approx 10^9$
- Users: Users are different, more varied and there are a lot of them. How many? $\approx 10^9$
- Documents: Documents are different, more varied and there are a lot of them. How many? $\approx 10^{11}$
- Context: Context is more important on the web than in many other IR applications.
- Ads and spam
Outline

1. Recap
2. Big picture
3. Ads
4. Duplicate detection
5. Spam
6. Web IR
   - Queries
   - Links
   - Context
   - Users
   - Documents
   - Size
7. Size of the web
Query distribution (1)
Query distribution (1)


<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sex</td>
<td>16</td>
<td>crack</td>
<td>31</td>
<td>juegos</td>
</tr>
<tr>
<td>2</td>
<td>(artifact)</td>
<td>17</td>
<td>games</td>
<td>32</td>
<td>nude</td>
</tr>
<tr>
<td>3</td>
<td>(artifact)</td>
<td>18</td>
<td>pussy</td>
<td>33</td>
<td>music</td>
</tr>
<tr>
<td>4</td>
<td>porno</td>
<td>19</td>
<td>cracks</td>
<td>34</td>
<td>musica</td>
</tr>
<tr>
<td>5</td>
<td>mp3</td>
<td>20</td>
<td>lolita</td>
<td>35</td>
<td>anal</td>
</tr>
<tr>
<td>6</td>
<td>Halloween</td>
<td>21</td>
<td>britney spears</td>
<td>36</td>
<td>free6</td>
</tr>
<tr>
<td>7</td>
<td>sexo</td>
<td>22</td>
<td>ebay</td>
<td>37</td>
<td>avril lavigne</td>
</tr>
<tr>
<td>8</td>
<td>chat</td>
<td>23</td>
<td>sexe</td>
<td>38</td>
<td>hotmail.com</td>
</tr>
<tr>
<td>9</td>
<td>porn</td>
<td>24</td>
<td>Pamela Anderson</td>
<td>39</td>
<td>winzip</td>
</tr>
<tr>
<td>10</td>
<td>yahoo</td>
<td>25</td>
<td>warez</td>
<td>40</td>
<td>fuck</td>
</tr>
<tr>
<td>11</td>
<td>KaZaA</td>
<td>26</td>
<td>divx</td>
<td>41</td>
<td>wallpaper</td>
</tr>
<tr>
<td>12</td>
<td>xxx</td>
<td>27</td>
<td>gay</td>
<td>42</td>
<td>hotmail.com</td>
</tr>
<tr>
<td>13</td>
<td>Hentai</td>
<td>28</td>
<td>harry potter</td>
<td>43</td>
<td>postales</td>
</tr>
<tr>
<td>14</td>
<td>lyrics</td>
<td>29</td>
<td>playboy</td>
<td>44</td>
<td>shakira</td>
</tr>
<tr>
<td>15</td>
<td>hotmail</td>
<td>30</td>
<td>lolitas</td>
<td>45</td>
<td>traductor</td>
</tr>
</tbody>
</table>

More than 1/3 of these are queries for adult content.
Query distribution (1)


1 sex 16 crack 31 juegos 46 Caramail
2 (artifact) 17 games 32 nude 47 msn
3 (artifact) 18 pussy 33 music 48 jennifer lopez
4 porno 19 cracks 34 musica 49 tits
5 mp3 20 lolita 35 anal 50 free porn
6 Halloween 21 britney spears 36 free6 51 cheats
7 sexo 22 ebay 37 avril lavigne 52 yahoo.com
8 chat 23 sexe 38 hotmail.com 53 eminem
9 porn 24 Pamela Anderson 39 winzip 54 Christina Aguilera
10 yahoo 25 warez 40 fuck 55 incest
11 KaZaA 26 divx 41 wallpaper 56 letras de canciones
12 xxx 27 gay 42 hotmail.com 57 hardcore
13 Hentai 28 harry potter 43 postales 58 weather
14 lyrics 29 playboy 44 shakira 59 wallpapers
15 hotmail 30 lolitas 45 traductor 60 lingerie

More than 1/3 of these are queries for adult content. Exercise: Does this mean that most people are looking for adult content?
Query distribution (2)
Queries have a power law distribution.
Queries have a power law distribution.
Recall Zipf’s law: a few very frequent words, a large number of very rare words.
Query distribution (2)

- Queries have a power law distribution.
- Recall Zipf’s law: a few very frequent words, a large number of very rare words
- Same here: a few very frequent queries, a large number of very rare queries
Query distribution (2)

- Queries have a power law distribution.
- Recall Zipf’s law: a few very frequent words, a large number of very rare words
- Same here: a few very frequent queries, a large number of very rare queries
- Examples of rare queries: search for names, towns, books etc
Query distribution (2)

- Queries have a power law distribution.
- Recall Zipf’s law: a few very frequent words, a large number of very rare words
- Same here: a few very frequent queries, a large number of very rare queries
- Examples of rare queries: search for names, towns, books etc
- The proportion of adult queries is much lower than 1/3
Types of queries / user needs in web search
Types of queries / user needs in web search

- **Informational user needs:** I need information on something. “low hemoglobin”
Types of queries / user needs in web search

- **Informational user needs:** I need information on something. “low hemoglobin”
- We called this “information need” earlier in the class.
Types of queries / user needs in web search

- **Informational user needs**: I need information on something. “low hemoglobin”

- We called this “information need” earlier in the class.

- On the web, information needs proper are only a subclass of user needs.
Types of queries / user needs in web search

- **Informational user needs**: I need information on something. “low hemoglobin”
- We called this “information need” earlier in the class.
- **On the web, information needs proper are only a subclass of user needs.**
- Other user needs: Navigational and transactional
Types of queries / user needs in web search

- **Informational user needs:** I need information on something. “low hemoglobin”
  
- We called this “information need” earlier in the class.
  
- **On the web, information needs proper are only a subclass of user needs.**
  
- **Other user needs:** Navigational and transactional
  
- **Navigational user needs:** I want to go to this web site. “hotmail”, “myspace”, “United Airlines”
Types of queries / user needs in web search

- **Informational user needs:** I need information on something. “low hemoglobin”
- **We called this “information need” earlier in the class.**
- **On the web, information needs proper are only a subclass of user needs.**
- **Other user needs:** Navigational and transactional
  - **Navigational user needs:** I want to go to this web site. “hotmail”, “myspace”, “United Airlines”
  - **Transactional user needs:** I want to make a transaction.
Types of queries / user needs in web search

- **Informational user needs**: I need information on something. “low hemoglobin”
- We called this “information need” earlier in the class.
- On the web, information needs proper are only a subclass of user needs.
- **Other user needs**: Navigational and transactional
  - **Navigational user needs**: I want to go to this web site. “hotmail”, “myspace”, “United Airlines”
  - **Transactional user needs**: I want to make a transaction.
    - Buy something: “MacBook Air”
Types of queries / user needs in web search

- **Informational user needs:** I need information on something.
  “low hemoglobin”
- We called this “information need” earlier in the class.
- **On the web, information needs proper are only a subclass of user needs.**
- **Other user needs:** Navigational and transactional
- **Navigational user needs:** I want to go to this web site.
  “hotmail”, “myspace”, “United Airlines”
- **Transactional user needs:** I want to make a transaction.
  - Buy something: “MacBook Air”
  - Download something: “Acrobat Reader”
Types of queries / user needs in web search

- **Informational user needs**: I need information on something. “low hemoglobin”
- We called this “information need” earlier in the class.
- **On the web, information needs proper are only a subclass of user needs.**
- **Other user needs**: Navigational and transactional
  - **Navigational user needs**: I want to go to this web site. “hotmail”, “myspace”, “United Airlines”
  - **Transactional user needs**: I want to make a transaction.
    - Buy something: “MacBook Air”
    - Download something: “Acrobat Reader”
    - Chat with someone: “live soccer chat”
Types of queries / user needs in web search

- **Informational user needs**: I need information on something. “low hemoglobin”
- We called this “information need” earlier in the class.
- **On the web, information needs proper are only a subclass of user needs.**
- **Other user needs**: Navigational and transactional
  - **Navigational user needs**: I want to go to this web site. “hotmail”, “myspace”, “United Airlines”
  - **Transactional user needs**: I want to make a transaction.
    - Buy something: “MacBook Air”
    - Download something: “Acrobat Reader”
    - Chat with someone: “live soccer chat”
- **Difficult problem**: How can the search engine tell what the user need or intent for a particular query is?
Outline

1 Recap
2 Big picture
3 Ads
4 Duplicate detection
5 Spam
6 Web IR
   - Queries
   - Links
     - Context
     - Users
     - Documents
   - Size
7 Size of the web
Search in a hyperlinked collection
Search in a hyperlinked collection

- Web search in most cases is interleaved with navigation...
Search in a hyperlinked collection

- Web search in most cases is interleaved with navigation...
- ...i.e., with following links.
Search in a hyperlinked collection

- Web search in most cases is interleaved with navigation . . .
- . . . i.e., with following links.
- Different from most other IR collections
Kinds of behaviors we see in the data

- Short / Nav
- Topic exploration
- Topic switch
- Methodical results exploration
- Query reform
- Multitasking
- Task 2
- Stacking behavior
Bowtie structure of the web
Bowtie structure of the web

- Strongly connected component (SCC) in the center
Bowtie structure of the web

- Strongly connected component (SCC) in the center
- Lots of pages that get linked to, but don’t link (OUT)
Bowtie structure of the web

- Strongly connected component (SCC) in the center
- Lots of pages that get linked to, but don’t link (OUT)
- Lots of pages that link to other pages, but don’t get linked to (IN)
Bowtie structure of the web

- Strongly connected component (SCC) in the center
- Lots of pages that get linked to, but don’t link (OUT)
- Lots of pages that link to other pages, but don’t get linked to (IN)
- Tendrils, tubes, islands
User intent: Answering the need behind the query
User intent: Answering the need behind the query

- What can we do to guess user intent?
User intent: Answering the need behind the query

- What can we do to guess user intent?
- Guess user intent independent of context:
User intent: Answering the need behind the query

- What can we do to guess user intent?
- Guess user intent independent of context:
  - Spell correction
User intent: Answering the need behind the query

What can we do to guess user intent?

Guess user intent independent of context:
  - Spell correction
  - Precomputed “typing” of queries (next slide)
User intent: Answering the need behind the query

- What can we do to guess user intent?
- Guess user intent independent of context:
  - Spell correction
  - Precomputed “typing” of queries (next slide)
- Better: Guess user intent based on context:
User intent: Answering the need behind the query

- What can we do to guess user intent?
- Guess user intent independent of context:
  - Spell correction
  - Precomputed “typing” of queries (next slide)
- Better: Guess user intent based on context:
  - Geographic context (slide after next)
User intent: Answering the need behind the query

- What can we do to guess user intent?
- Guess user intent independent of context:
  - Spell correction
  - Precomputed “typing” of queries (next slide)
- Better: Guess user intent based on context:
  - Geographic context (slide after next)
  - Context of user in this session (e.g., previous query)
User intent: Answering the need behind the query

- What can we do to guess user intent?
- Guess user intent independent of context:
  - Spell correction
  - Precomputed “typing” of queries (next slide)
- Better: Guess user intent based on context:
  - Geographic context (slide after next)
  - Context of user in this session (e.g., previous query)
  - Context provided by personal profile (Yahoo/MSN do this, Google claims it doesn’t)
Guessing of user intent by “typing” queries
Guessing of user intent by “typing” queries

- Calculation: $5 + 4$
Guessing of user intent by “typing” queries

- Calculation: $5+4$
- Unit conversion: 1 kg in pounds
Guessing of user intent by “typing” queries

- Calculation: 5+4
- Unit conversion: 1 kg in pounds
- Currency conversion: 1 euro in kronor
Guessing of user intent by “typing” queries

- Calculation: $5 + 4$
- Unit conversion: 1 kg in pounds
- Currency conversion: 1 euro in kronor
- Tracking number: 8167 2278 6764
Guessing of user intent by “typing” queries

- Calculation: 5+4
- Unit conversion: 1 kg in pounds
- Currency conversion: 1 euro in kronor
- Tracking number: 8167 2278 6764
- Flight info: LH 454
Guessing of user intent by “typing” queries

- Calculation: 5+4
- Unit conversion: 1 kg in pounds
- Currency conversion: 1 euro in kronor
- Tracking number: 8167 2278 6764
- Flight info: LH 454
- Area code: 650
Guessing of user intent by “typing” queries

- Calculation: 5+4
- Unit conversion: 1 kg in pounds
- Currency conversion: 1 euro in kronor
- Tracking number: 8167 2278 6764
- Flight info: LH 454
- Area code: 650
- Map: columbus oh
Guessing of user intent by “typing” queries

- Calculation: 5+4
- Unit conversion: 1 kg in pounds
- Currency conversion: 1 euro in kronor
- Tracking number: 8167 2278 6764
- Flight info: LH 454
- Area code: 650
- Map: columbus oh
- Stock price: msft
Guessing of user intent by “typing” queries

- Calculation: 5+4
- Unit conversion: 1 kg in pounds
- Currency conversion: 1 euro in kronor
- Tracking number: 8167 2278 6764
- Flight info: LH 454
- Area code: 650
- Map: columbus oh
- Stock price: msft
- Albums/movies etc: coldplay
The spatial context: Geo-search
The spatial context: Geo-search

- Three relevant locations
The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)
The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)
  - User (located in Palo Alto)
The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)
  - User (located in Palo Alto)

- Locating the user
The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)
  - User (located in Palo Alto)

- Locating the user
  - IP address
The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)
  - User (located in Palo Alto)

- Locating the user
  - IP address
  - Information provided by user (e.g., in user profile)
The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)
  - User (located in Palo Alto)

- Locating the user
  - IP address
  - Information provided by user (e.g., in user profile)
  - Mobile phone
The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)
  - User (located in Palo Alto)

- Locating the user
  - IP address
  - Information provided by user (e.g., in user profile)
  - Mobile phone

- Geo-tagging: Parse text and identify the coordinates of the geographic entities
The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)
  - User (located in Palo Alto)

- Locating the user
  - IP address
  - Information provided by user (e.g., in user profile)
  - Mobile phone

- Geo-tagging: Parse text and identify the coordinates of the geographic entities
  - Example: East Palo Alto CA → Latitude: 37.47 N, Longitude: 122.14 W
The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)
  - User (located in Palo Alto)

- Locating the user
  - IP address
  - Information provided by user (e.g., in user profile)
  - Mobile phone

- **Geo-tagging:** Parse text and identify the coordinates of the geographic entities
  - Example: East Palo Alto CA → Latitude: 37.47 N, Longitude: 122.14 W
  - Important NLP problem
How do we use context to modify query results?
How do we use context to modify query results?

- Result restriction: Don’t consider inappropriate results
How do we use context to modify query results?

- Result restriction: Don’t consider inappropriate results
  - For user on google.fr ...
How do we use context to modify query results?

- Result restriction: Don’t consider inappropriate results
  - For user on google.fr . . .
  - . . . only show .fr results
How do we use context to modify query results?

- Result restriction: Don’t consider inappropriate results
  - For user on google.fr . . .
  - . . . only show .fr results

- Ranking modulation: use a rough generic ranking, rerank based on personal context
How do we use context to modify query results?

- Result restriction: Don’t consider inappropriate results
  - For user on google.fr ...
  - ... only show .fr results

- Ranking modulation: use a rough generic ranking, rerank based on personal context

- Contextualization / personalization is an area of search with a lot of potential for improvement.
Outline

1 Recap
2 Big picture
3 Ads
4 Duplicate detection
5 Spam
6 Web IR
   - Queries
   - Links
   - Context
   - Users
   - Documents
   - Size
7 Size of the web
Users of web search
Users of web search

- Use short queries (average < 3)
Users of web search

- Use short queries (average < 3)
- Rarely use operators
Users of web search

- Use short queries (average $< 3$)
- Rarely use operators
- Don’t want to spend a lot of time on composing a query
Users of web search

- Use short queries (average < 3)
- Rarely use operators
- Don’t want to spend a lot of time on composing a query
- Only look at the first couple of results
Users of web search

- Use short queries (average < 3)
- Rarely use operators
- Don’t want to spend a lot of time on composing a query
- Only look at the first couple of results
- Want a simple UI, not a search engine start page overloaded with graphics
Users of web search

- Use short queries (average < 3)
- Rarely use operators
- Don’t want to spend a lot of time on composing a query
- Only look at the first couple of results
- Want a simple UI, not a search engine start page overloaded with graphics
- Extreme variability in terms of user needs, user expectations, experience, knowledge, ...
Users of web search

- Use short queries (average < 3)
- Rarely use operators
- Don’t want to spend a lot of time on composing a query
- Only look at the first couple of results
- Want a simple UI, not a search engine start page overloaded with graphics
- Extreme variability in terms of user needs, user expectations, experience, knowledge, ...
  - Industrial/developing world, English/Estonian, old/young, rich/poor, differences in culture and class
Users of web search

- Use short queries (average < 3)
- Rarely use operators
- Don’t want to spend a lot of time on composing a query
- Only look at the first couple of results
- Want a simple UI, not a search engine start page overloaded with graphics
- Extreme variability in terms of user needs, user expectations, experience, knowledge, . . .
  - Industrial/developing world, English/Estonian, old/young, rich/poor, differences in culture and class
- One interface for hugely divergent needs
How do users evaluate search engines?
How do users evaluate search engines?

- Classic IR relevance (as measured by $F$) can also be used for web IR.
How do users evaluate search engines?

- Classic IR relevance (as measured by $F$) can also be used for web IR.
- Equally important: Trust, duplicate elimination, readability, loads fast, no pop-ups
How do users evaluate search engines?

- Classic IR relevance (as measured by $F$) can also be used for web IR.
- Equally important: Trust, duplicate elimination, readability, loads fast, no pop-ups
- On the web, precision is more important than recall.
How do users evaluate search engines?

- Classic IR relevance (as measured by $F$) can also be used for web IR.
- Equally important: Trust, duplicate elimination, readability, loads fast, no pop-ups
- On the web, precision is more important than recall.
  - Precision at 1, precision at 10, precision on the first 2-3 pages
How do users evaluate search engines?

- Classic IR relevance (as measured by $F$) can also be used for web IR.

- Equally important: Trust, duplicate elimination, readability, loads fast, no pop-ups

- On the web, precision is more important than recall.
  - Precision at 1, precision at 10, precision on the first 2-3 pages
  - But there is a subset of queries where recall matters.
Web information needs that require high recall
Web information needs that require high recall

- Has this idea been patented?
Web information needs that require high recall

- Has this idea been patented?
- Searching for info on a prospective financial advisor
Web information needs that require high recall

- Has this idea been patented?
- Searching for info on a prospective financial advisor
- Searching for info on a prospective employee
Web information needs that require high recall

- Has this idea been patented?
- Searching for info on a prospective financial advisor
- Searching for info on a prospective employee
- Searching for info on a date
Web documents: different from other IR collections
Web documents: different from other IR collections

- Distributed content creation: no design, no coordination
Web documents: different from other IR collections

- Distributed content creation: no design, no coordination
  - “Democratization of publishing”
Web documents: different from other IR collections

- Distributed content creation: no design, no coordination
  - “Democratization of publishing”
  - Result: extreme heterogeneity of documents on the web
Web documents: different from other IR collections

- Distributed content creation: no design, no coordination
  - “Democratization of publishing”
  - Result: extreme heterogeneity of documents on the web
- Unstructured (text, html), semistructured (html, xml), structured/relational (databases)
Web documents: different from other IR collections

- Distributed content creation: no design, no coordination
  - “Democratization of publishing”
  - Result: extreme heterogeneity of documents on the web
- Unstructured (text, html), semistructured (html, xml), structured/relational (databases)
- Dynamically generated content
Dynamic content
Dynamic content

- Dynamic pages are generated from scratch when the user requests them – usually from underlying data in a database.
Dynamic pages are generated from scratch when the user requests them – usually from underlying data in a database.

Example: current status of flight LH 454
Dynamic content (2)
Dynamic content (2)

- Most (truly) dynamic content is ignored by web spiders.
Dynamic content (2)

- Most (truly) dynamic content is ignored by web spiders.
  - It’s too much to index it all.
Most (truly) dynamic content is ignored by web spiders.

- It’s too much to index it all.

- Actually, a lot of “static” content is also assembled on the fly (asp, php etc.: headers, date, ads etc)
Web pages change frequently (Fetterly 1997)
Web pages change frequently (Fetterly 1997)
Multilinguality
Multilinguality

- Documents in a large number of languages
Multilinguality

- Documents in a large number of languages
- Queries in a large number of languages
Multilinguality

- Documents in a large number of languages
- Queries in a large number of languages
- First cut: Don’t return English results for a Japanese query
Multilinguality

- Documents in a large number of languages
- Queries in a large number of languages
- First cut: Don’t return English results for a Japanese query
- However: Frequent mismatches query/document languages
Multilinguality

- Documents in a large number of languages
- Queries in a large number of languages
- First cut: Don’t return English results for a Japanese query
- However: Frequent mismatches query/document languages
- Many people can understand, but not query in a language
Multilinguality

- Documents in a large number of languages
- Queries in a large number of languages
- First cut: Don’t return English results for a Japanese query
- However: Frequent mismatches query/document languages
- Many people can understand, but not query in a language
- Translation is important.
Multilinguality

- Documents in a large number of languages
- Queries in a large number of languages
- First cut: Don’t return English results for a Japanese query
- However: Frequent mismatches query/document languages
- Many people can understand, but not query in a language
- Translation is important.
- Google example: “Beaujolais Nouveau -wine”
Duplicate documents
Duplicate documents

- Significant duplication – 30%–40% duplicates in some studies
Duplicate documents

- Significant duplication – 30%–40% duplicates in some studies
- Duplicates in the search results were common in the early days of the web.
Duplicate documents

- Significant duplication – 30%–40% duplicates in some studies
- Duplicates in the search results were common in the early days of the web.
- Today’s search engines eliminate duplicates very effectively.
Duplicate documents

- Significant duplication – 30%–40% duplicates in some studies
- Duplicates in the search results were common in the early days of the web.
- Today’s search engines eliminate duplicates very effectively.
- Key for high user satisfaction
Trust
For many collections, it is easy to assess the trustworthiness of a document.
For many collections, it is easy to assess the trustworthiness of a document.

- A collection of Reuters newswire articles
For many collections, it is easy to assess the trustworthiness of a document.

- A collection of Reuters newswire articles
- A collection of TASS (Telegraph Agency of the Soviet Union) newswire articles from the 1980s
For many collections, it is easy to assess the trustworthiness of a document.

- A collection of Reuters newswire articles
- A collection of TASS (Telegraph Agency of the Soviet Union) newswire articles from the 1980s
- Your Outlook email from the last three years
For many collections, it is easy to assess the trustworthiness of a document.

- A collection of Reuters newswire articles
- A collection of TASS (Telegraph Agency of the Soviet Union) newswire articles from the 1980s
- Your Outlook email from the last three years

Web documents are different: In many cases, we don’t know how to evaluate the information.
For many collections, it is easy to assess the trustworthiness of a document.

- A collection of Reuters newswire articles
- A collection of TASS (Telegraph Agency of the Soviet Union) newswire articles from the 1980s
- Your Outlook email from the last three years

Web documents are different: In many cases, we don’t know how to evaluate the information.

Hoaxes abound.
Outline

1. Recap
2. Big picture
3. Ads
4. Duplicate detection
5. Spam
6. Web IR
   - Queries
   - Links
   - Context
   - Users
   - Documents
   - Size
7. Size of the web
Growth of the web
Growth of the web

- The web keeps growing.
The web keeps growing.
But growth is no longer exponential?
Size of the web: Issues
Size of the web: Issues

- What is size? Number of web servers? Number of pages? Terabytes of data available?
Size of the web: Issues

- What is size? Number of web servers? Number of pages? Terabytes of data available?
- Some servers are seldom connected.
Size of the web: Issues

- What is size? Number of web servers? Number of pages? Terabytes of data available?
- Some servers are seldom connected.
  - Example: Your laptop running a web server
Size of the web: Issues

- What is size? Number of web servers? Number of pages? Terabytes of data available?
- Some servers are seldom connected.
  - Example: Your laptop running a web server
  - Is it part of the web?
Size of the web: Issues

- What is size? Number of web servers? Number of pages? Terabytes of data available?
- Some servers are seldom connected.
  - Example: Your laptop running a web server
  - Is it part of the web?
- The “dynamic” web is infinite.
Size of the web: Issues

- What is size? Number of web servers? Number of pages? Terabytes of data available?
- Some servers are seldom connected.
  - Example: Your laptop running a web server
  - Is it part of the web?
- The “dynamic” web is infinite.
  - Any sum of two numbers is its own dynamic page on Google. (Example: “2+4”)

Schütze: Web search
“Search engine index contains $N$ pages”: Issues
“Search engine index contains $N$ pages”: Issues

- Can I claim a page is in the index if I only index the first 4000 bytes?
“Search engine index contains $N$ pages”: Issues

- Can I claim a page is in the index if I only index the first 4000 bytes?
- Can I claim a page is in the index if I only index anchor text pointing to the page?
Search engine index contains $N$ pages: Issues

- Can I claim a page is in the index if I only index the first 4000 bytes?
- Can I claim a page is in the index if I only index anchor text pointing to the page?
  - There used to be (and still are?) billions of pages that are only indexed by anchor text.
Simple method for determining a lower bound
Simple method for determining a lower bound

- OR-query of frequent words in a number of languages
Simple method for determining a lower bound

- OR-query of frequent words in a number of languages
- http://ifnlp.org/ir/sizeoftheweb.html
Simple method for determining a lower bound

- OR-query of frequent words in a number of languages
- http://ifnlp.org/ir/sizeoftheweb.html
- According to this query: Size of web $\geq 21,450,000,000$ on 2007.07.07 and $\geq 25,350,000,000$ on 2008.07.03
Simple method for determining a lower bound

- OR-query of frequent words in a number of languages
- http://ifnlp.org/ir/sizeoftheweb.html
- According to this query: Size of web $\geq 21,450,000,000$ on 2007.07.07 and $\geq 25,350,000,000$ on 2008.07.03
- But page counts of google search results are only rough estimates.
Outline

1 Recap
2 Big picture
3 Ads
4 Duplicate detection
5 Spam
6 Web IR
   • Queries
   • Links
   • Context
   • Users
   • Documents
   • Size
7 Size of the web
Size of the web: Who cares?
Size of the web: Who cares?

- Media
Size of the web: Who cares?

- Media
- Users
Size of the web: Who cares?

- Media
- Users
  - They may switch to the search engine that has the best coverage of the web.
Size of the web: Who cares?

- Media
- Users
  - They may switch to the search engine that has the best coverage of the web.
  - Users (sometimes) care about recall. If we underestimate the size of the web, search engine results may have low recall.
Size of the web: Who cares?

- Media
- Users
  - They may switch to the search engine that has the best coverage of the web.
  - Users (sometimes) care about recall. If we underestimate the size of the web, search engine results may have low recall.
- Search engine designers (how many pages do I need to be able to handle?)
Size of the web: Who cares?

- Media
- Users
  - They may switch to the search engine that has the best coverage of the web.
  - Users (sometimes) care about recall. If we underestimate the size of the web, search engine results may have low recall.
- Search engine designers (how many pages do I need to be able to handle?)
- Crawler designers (which policy will crawl close to $N$ pages?)
What is the size of the web? Any guesses?
Simple method for determining a lower bound
Simple method for determining a lower bound

- OR-query of frequent words in a number of languages
Simple method for determining a lower bound

- OR-query of frequent words in a number of languages
Simple method for determining a lower bound

- OR-query of frequent words in a number of languages
- \url{http://ifnlp.org/lehre/teaching/2007-SS/ir/sizeoftheweb.html}
- According to this query: Size of web $\geq 21,450,000,000$ on 2007.07.07
Simple method for determining a lower bound

- OR-query of frequent words in a number of languages
- According to this query: Size of web $\geq 21,450,000,000$ on 2007.07.07
- Big if: Page counts of google search results are correct. (Generally, they are just rough estimates.)
Simple method for determining a lower bound

- OR-query of frequent words in a number of languages
- According to this query: Size of web $\geq 21,450,000,000$ on 2007.07.07
- Big if: Page counts of google search results are correct. (Generally, they are just rough estimates.)
- But this is just a lower bound, based on one search engine.
Simple method for determining a lower bound

- OR-query of frequent words in a number of languages
- According to this query: Size of web $\geq 21,450,000,000$ on 2007.07.07
- Big if: Page counts of google search results are correct. (Generally, they are just rough estimates.)
- But this is just a lower bound, based on one search engine.
- How can we do better?
Size of the web: Issues
Size of the web: Issues

- The “dynamic” web is infinite.
The “dynamic” web is infinite.

- Any sum of two numbers is its own dynamic page on Google.
  (Example: “2+4”)
Size of the web: Issues

- The “dynamic” web is infinite.
  - Any sum of two numbers is its own dynamic page on Google. (Example: “2+4”)
  - Many other dynamic sites generating infinite number of pages
The “dynamic” web is infinite.
- Any sum of two numbers is its own dynamic page on Google. (Example: “2+4”)
- Many other dynamic sites generating infinite number of pages

The static web contains duplicates – each “equivalence class” should only be counted once.
Size of the web: Issues

- The “dynamic” web is infinite.
  - Any sum of two numbers is its own dynamic page on Google.
    (Example: “2+4”)
  - Many other dynamic sites generating infinite number of pages

- The static web contains duplicates – each “equivalence class” should only be counted once.

- Some servers are seldom connected.
The “dynamic” web is infinite.
  - Any sum of two numbers is its own dynamic page on Google.
    (Example: “2+4”)
  - Many other dynamic sites generating infinite number of pages

The static web contains duplicates – each “equivalence class” should only be counted once.

Some servers are seldom connected.
  - Example: Your laptop
Size of the web: Issues

- The “dynamic” web is infinite.
  - Any sum of two numbers is its own dynamic page on Google. (Example: “2+4”)
  - Many other dynamic sites generating infinite number of pages

- The static web contains duplicates – each “equivalence class” should only be counted once.

- Some servers are seldom connected.
  - Example: Your laptop
  - Is it part of the web?
“Search engine index contains $N$ pages”: Issues
Can I claim a page is in the index if I only index the first 4000 bytes?
“Search engine index contains $N$ pages”: Issues

- Can I claim a page is in the index if I only index the first 4000 bytes?
- Can I claim a page is in the index if I only index anchor text pointing to the page?
“Search engine index contains \( N \) pages”: Issues

- Can I claim a page is in the index if I only index the first 4000 bytes?
- Can I claim a page is in the index if I only index anchor text pointing to the page?
  - There used to be (and still are?) billions of pages that are only indexed by anchor text.
How can we estimate the size of the web?
Sampling methods
Sampling methods

- Random queries
Sampling methods

- Random queries
- Random searches
Sampling methods

- Random queries
- Random searches
- Random IP addresses
Sampling methods

- Random queries
- Random searches
- Random IP addresses
- Random walks
Variant: Estimate relative sizes of indexes
Variant: Estimate relative sizes of indexes

- There are significant differences between indexes of different search engines.
Variant: Estimate relative sizes of indexes

- There are significant differences between indexes of different search engines.
- Different engines have different preferences.
Variant: Estimate relative sizes of indexes

- There are significant differences between indexes of different search engines.
- Different engines have different preferences.
  - max url depth, max count/host, anti-spam rules, priority rules etc.
Variant: Estimate relative sizes of indexes

- There are significant differences between indexes of different search engines.
- Different engines have different preferences.
  - max url depth, max count/host, anti-spam rules, priority rules etc.
- Different engines index different things under the same URL.
There are significant differences between indexes of different search engines.

Different engines have different preferences.
- max url depth, max count/host, anti-spam rules, priority rules etc.

Different engines index different things under the same URL.
- anchor text, frames, meta-keywords, size of prefix etc.
Relative Size from Overlap [Bharat & Broder, 98]

Sample URLs randomly from A
Check if contained in B
and vice versa

\[ A \cap B = \frac{1}{2} \times \text{Size A} \]
\[ A \cap B = \frac{1}{6} \times \text{Size B} \]

\[ \frac{1}{2} \times \text{Size A} = \frac{1}{6} \times \text{Size B} \]
\[ \therefore \text{Size A} / \text{Size B} = \frac{1}{2} / \frac{1}{6} = 1/3 \]

Each test involves: (i) Sampling (ii) Checking
Sampling URLs
Sampling URLs

- Ideal strategy: Generate a random URL
Sampling URLs

- Ideal strategy: Generate a random URL
- Problem: Random URLs are hard to find (and sampling distribution should reflect "user interest")
Sampling URLs

- Ideal strategy: Generate a random URL
- Problem: Random URLs are hard to find (and sampling distribution should reflect “user interest”)
- Approach 1: Random walks / IP addresses
Sampling URLs

- Ideal strategy: Generate a random URL
- Problem: Random URLs are hard to find (and sampling distribution should reflect “user interest”)
- Approach 1: Random walks / IP addresses
  - In theory: might give us a true estimate of the size of the web (as opposed to just relative sizes of indexex)
Sampling URLs

- Ideal strategy: Generate a random URL
- Problem: Random URLs are hard to find (and sampling distribution should reflect “user interest”)
- Approach 1: Random walks / IP addresses
  - In theory: might give us a true estimate of the size of the web (as opposed to just relative sizes of indexex)
- Approach 2: Generate a random URL contained in a given engine
Sampling URLs

- Ideal strategy: Generate a random URL
- Problem: Random URLs are hard to find (and sampling distribution should reflect “user interest”)
- Approach 1: Random walks / IP addresses
  - In theory: might give us a true estimate of the size of the web (as opposed to just relative sizes of indexex)
- Approach 2: Generate a random URL contained in a given engine
  - Suffices for accurate estimation of relative size
Random URLs from random queries
Random URLs from random queries

- Idea: Use vocabulary of the web for query generation
Random URLs from random queries

- Idea: Use vocabulary of the web for query generation
- Vocabulary can be generated from web crawl
Random URLs from random queries

- Idea: Use vocabulary of the web for query generation
- Vocabulary can be generated from web crawl
- Use conjunctive queries $w_1 \ AND \ w_2$
Random URLs from random queries

- Idea: Use vocabulary of the web for query generation
- Vocabulary can be generated from web crawl
- Use conjunctive queries $w_1 \text{ AND } w_2$
  - Example: vocalists AND rsi
Random URLs from random queries

- Idea: Use vocabulary of the web for query generation
- Vocabulary can be generated from web crawl
- Use conjunctive queries $w_1 \text{ AND } w_2$
  - Example: vocalists AND rsi
- Get result set of one hundred URLs from the source engine
Random URLs from random queries

- Idea: Use vocabulary of the web for query generation
- Vocabulary can be generated from web crawl
- Use conjunctive queries $w_1$ AND $w_2$
  - Example: vocalists AND rsi
- Get result set of one hundred URLs from the source engine
- Choose a random URL from the result set
Random URLs from random queries

- Idea: Use vocabulary of the web for query generation
- Vocabulary can be generated from web crawl
- Use conjunctive queries $w_1$ AND $w_2$
  - Example: vocalists AND rsi
- Get result set of one hundred URLs from the source engine
- Choose a random URL from the result set
- This sampling method induces a weight $W(p)$ for each page $p$. 
Random URLs from random queries

- Idea: Use vocabulary of the web for query generation
- Vocabulary can be generated from web crawl
- Use conjunctive queries $w_1 \text{ AND } w_2$
  - Example: vocalists AND rsi
- Get result set of one hundred URLs from the source engine
- Choose a random URL from the result set
- This sampling method induces a weight $W(p)$ for each page $p$.
- Method was used by Bharat and Broder (1998).
Checking if a page is in the index
Checking if a page is in the index

- Either: Search for URL if the engine supports this
Checking if a page is in the index

- Either: Search for URL if the engine supports this
- Or: Create a query that will find doc $d$ with high probability
Checking if a page is in the index

- Either: Search for URL if the engine supports this
- Or: Create a query that will find doc $d$ with high probability
  - Download doc, extract words
Checking if a page is in the index

- Either: Search for URL if the engine supports this
- Or: Create a query that will find doc $d$ with high probability
  - Download doc, extract words
  - Use 8 low frequency word as AND query
Checking if a page is in the index

- Either: Search for URL if the engine supports this
- Or: Create a query that will find doc $d$ with high probability
  - Download doc, extract words
  - Use 8 low frequency word as AND query
  - Call this a strong query for $d$
Checking if a page is in the index

- Either: Search for URL if the engine supports this
- Or: Create a query that will find doc $d$ with high probability
  - Download doc, extract words
  - Use 8 low frequency word as AND query
  - Call this a strong query for $d$
  - Run query
Checking if a page is in the index

Either: Search for URL if the engine supports this
Or: Create a query that will find doc $d$ with high probability

- Download doc, extract words
- Use 8 low frequency word as AND query
- Call this a strong query for $d$
- Run query
- Check if $d$ is in result set
Checking if a page is in the index

- Either: Search for URL if the engine supports this
- Or: Create a query that will find doc $d$ with high probability
  - Download doc, extract words
  - Use 8 low frequency word as AND query
  - Call this a **strong query** for $d$
  - Run query
  - Check if $d$ is in result set

- Problems
Checking if a page is in the index

- Either: Search for URL if the engine supports this
- Or: Create a query that will find doc \( d \) with high probability
  - Download doc, extract words
  - Use 8 low frequency word as AND query
  - Call this a strong query for \( d \)
  - Run query
  - Check if \( d \) is in result set

Problems
- Near duplicates
Checking if a page is in the index

- Either: Search for URL if the engine supports this
- Or: Create a query that will find doc \( d \) with high probability
  - Download doc, extract words
  - Use 8 low frequency word as AND query
  - Call this a strong query for \( d \)
  - Run query
  - Check if \( d \) is in result set

Problems
- Near duplicates
- Redirects
Checking if a page is in the index

Either: Search for URL if the engine supports this
Or: Create a query that will find doc $d$ with high probability

- Download doc, extract words
- Use 8 low frequency word as AND query
- Call this a strong query for $d$
- Run query
- Check if $d$ is in result set

Problems

- Near duplicates
- Redirects
- Engine time-outs
Computing Relative Sizes and Total Coverage [BB98]

\( a = \text{AltaVista}, \ e = \text{Excite}, \ h = \text{HotBot}, \ i = \text{Infoseek} \)

\( f_{xy} = \) fraction of \( x \text{ in } y \)

- Six pair-wise overlaps
  \[ f_{ah} * a - f_{ha} * h = \varepsilon_1 \]
  \[ f_{ai} * a - f_{ia} * i = \varepsilon_2 \]
  \[ f_{ae} * a - f_{ea} * e = \varepsilon_3 \]
  \[ f_{hi} * h - f_{ih} * i = \varepsilon_4 \]
  \[ f_{he} * h - f_{eh} * e = \varepsilon_5 \]
  \[ f_{ei} * e - f_{ie} * i = \varepsilon_6 \]

- Arbitrarily, let \( a = 1 \).

- We have 6 equations and 3 unknowns.
- Solve for \( e, \ h \) and \( i \) to minimize \( \Sigma \varepsilon_i^2 \)
- Compute engine overlaps.
- Re-normalize so that the total joint coverage is 100%
Advantages & disadvantages

- Statistically sound under the induced weight.
- Biases induced by random query
  - Query Bias: Favors content-rich pages in the language(s) of the lexicon
  - Ranking Bias: Solution: Use conjunctive queries & fetch all
  - Checking Bias: Duplicates, impoverished pages omitted
  - Document or query restriction bias: engine might not deal properly with 8 words conjunctive query
  - Malicious Bias: Sabotage by engine
  - Operational Problems: Time-outs, failures, engine inconsistencies, index modification.
Random searches
Random searches

- Choose random searches extracted from a search engine log (Lawrence & Giles 97)
Random searches

- Choose random searches extracted from a search engine log (Lawrence & Giles 97)
- Use only queries with small result sets
Random searches

- Choose random searches extracted from a search engine log (Lawrence & Giles 97)
- Use only queries with small result sets
- For each random query: compute ratio size($r_1$)/size($r_2$) of the two result sets
Random searches

- Choose random searches extracted from a search engine log (Lawrence & Giles 97)
- Use only queries with small result sets
- For each random query: compute ratio $\frac{\text{size}(r_1)}{\text{size}(r_2)}$ of the two result sets
- Average over random searches
Advantages & disadvantages
Advantages & disadvantages

- Advantage
Advantages & disadvantages

- Advantage
  - Might be a better reflection of the human perception of coverage
Advantages & disadvantages

- Advantage
  - Might be a better reflection of the human perception of coverage

- Issues
Advantages & disadvantages

- **Advantage**
  - Might be a better reflection of the human perception of coverage

- **Issues**
  - Samples are correlated with source of log (unfair advantage for originating search engine)
Advantages & disadvantages

- **Advantage**
  - Might be a better reflection of the human perception of coverage

- **Issues**
  - Samples are correlated with source of log (unfair advantage for originating search engine)
  - Duplicates
Advantages & disadvantages

- **Advantage**
  - Might be a better reflection of the human perception of coverage

- **Issues**
  - Samples are correlated with source of log (unfair advantage for originating search engine)
  - Duplicates
  - Technical statistical problems (must have non-zero results, ratio average not statistically sound)
Random searches [Lawr98, Lawr99]

- 575 & 1050 queries from the NEC RI employee logs
- 6 Engines in 1998, 11 in 1999
- Implementation:
  - Restricted to queries with < 600 results in total
  - Counted URLs from each engine after verifying query match
  - Computed size ratio & overlap for individual queries
  - Estimated index size ratio & overlap by averaging over all queries
Queries from Lawrence and Giles study

- adaptive access control
- neighborhood preservation topographic
- hamiltonian structures
- right linear grammar
- pulse width modulation neural
- unbalanced prior probabilities
- ranked assignment method
- internet explorer favourites importing
- karvel thornber
- zili liu
- softmax activation function
- bose multidimensional system theory
- gamma mlp
dvi2pdf
- john oliensis
- rieke spikes exploring neural
- video watermarking
counterpropagation network
- fat shattering dimension
- abelson amorphous computing
Random IP addresses [Lawrence & Giles ‘99]

- Generate random IP addresses
- Find a web server at the given address
  - If there’s one
- Collect all pages from server.

http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003447
Random IP addresses [ONei97, Lawr99]
Random IP addresses [ONei97, Lawr99]

[Lawr99] exhaustively crawled 2500 servers and extrapolated
Random IP addresses [ONei97,Lawr99]

- [Lawr99] exhaustively crawled 2500 servers and extrapolated
- Estimated size of the web to be 800 million
Advantages and disadvantages
Advantages and disadvantages

- Advantages
Advantages and disadvantages

Advantages

- Can, in theory, estimate the size of the accessible web (as opposed to the (relative) size of an index)
Advantages and disadvantages

**Advantages**

- Can, in theory, estimate the size of the accessible web (as opposed to the (relative) size of an index)
- Clean statistics
Advantages and disadvantages

- Advantages
  - Can, in theory, estimate the size of the accessible web (as opposed to the (relative) size of an index)
  - Clean statistics
  - Independent of crawling strategies
Advantages and disadvantages

- **Advantages**
  - Can, in theory, estimate the size of the accessible web (as opposed to the (relative) size of an index)
  - Clean statistics
  - Independent of crawling strategies

- **Disadvantages**
Advantages and disadvantages

- **Advantages**
  - Can, in theory, estimate the size of the accessible web (as opposed to the (relative) size of an index)
  - Clean statistics
  - Independent of crawling strategies

- **Disadvantages**
  - Many hosts share one IP (→ oversampling)
Advantages and disadvantages

**Advantages**
- Can, in theory, estimate the size of the accessible web (as opposed to the (relative) size of an index)
- Clean statistics
- Independent of crawling strategies

**Disadvantages**
- Many hosts share one IP (→ oversampling)
- Hosts with large web sites don’t get more weight than hosts with small web sites (→ possible undersampling)
Advantages and disadvantages

- **Advantages**
  - Can, in theory, estimate the size of the accessible web (as opposed to the (relative) size of an index)
  - Clean statistics
  - Independent of crawling strategies

- **Disadvantages**
  - Many hosts share one IP (→ oversampling)
  - Hosts with large web sites don’t get more weight than hosts with small web sites (→ possible undersampling)
  - Sensitive to spam (multiple IPs for same spam server)
Advantages and disadvantages

**Advantages**
- Can, in theory, estimate the size of the accessible web (as opposed to the (relative) size of an index)
- Clean statistics
- Independent of crawling strategies

**Disadvantages**
- Many hosts share one IP (→ oversampling)
- Hosts with large web sites don’t get more weight than hosts with small web sites (→ possible undersampling)
- Sensitive to spam (multiple IPs for same spam server)
- Again, duplicates
Random walks
[Henzinger et al. WWW9]

- View the Web as a directed graph
- Build a random walk on this graph
  - Includes various “jump” rules back to visited sites
    - Does not get stuck in spider traps!
    - Can follow all links!
  - Converges to a stationary distribution
    - Must assume graph is finite and independent of the walk.
    - Conditions are not satisfied (cookie crumbs, flooding)
    - Time to convergence not really known
- Sample from stationary distribution of walk
- Use the “strong query” method to check coverage by SE
Dependence on seed list

- How well connected is the graph? [Broder et al., WWW9]
Advantages & disadvantages

■ Advantages
  ■ “Statistically clean” method at least in theory!
  ■ Could work even for infinite web (assuming convergence) under certain metrics.

■ Disadvantages
  ■ List of seeds is a problem.
  ■ Practical approximation might not be valid.
  ■ Non-uniform distribution
    ■ Subject to link spamming
Conclusion
Many different approaches to web size estimation.
Conclusion

- Many different approaches to web size estimation.
- None is perfect.
Many different approaches to web size estimation.
None is perfect.
The problem has gotten much harder.
Many different approaches to web size estimation.
None is perfect.
The problem has gotten much harder.
There hasn’t been a good study for a couple of years.
Conclusion

- Many different approaches to web size estimation.
- None is perfect.
- The problem has gotten much harder.
- There hasn’t been a good study for a couple of years.
- Great topic for a thesis!
Chapter 19 of IIR
Resources

• Chapter 19 of IIR
• Resources at http://cislmu.org
Resources

- Chapter 19 of IIR
- Resources at http://cislmu.org
  - Hal Varian explains Google second price auction: http://www.youtube.com/watch?v=K7I0a2PVhPQ
Resources

- Chapter 19 of IIR
- Resources at http://cislmu.org
  - Hal Varian explains Google second price auction: http://www.youtube.com/watch?v=K7l0a2PVhPQ
  - Size of the web queries
Chapter 19 of IIR

Resources at http://cislmu.org

- Hal Varian explains Google second price auction:
  http://www.youtube.com/watch?v=K7l0a2PVhPQ
- Size of the web queries
- Trademark issues (Geico and Vuitton cases)
Chapter 19 of IIR

Resources at http://cislmu.org

- Hal Varian explains Google second price auction: http://www.youtube.com/watch?v=K7I0a2PVhPQ
- Size of the web queries
- Trademark issues (Geico and Vuitton cases)
- How ads are priced
Resources

- Chapter 19 of IIR
- Resources at http://cis1mu.org
  - Hal Varian explains Google second price auction: http://www.youtube.com/watch?v=K7I0a2PVhPQ
  - Size of the web queries
  - Trademark issues (Geico and Vuitton cases)
  - How ads are priced