Introduction to Information Retrieval
http://informationretrieval.org

IIR 19: Web Search Basics

Hinrich Schütze

Institute for Natural Language Processing, Universität Stuttgart

2008.07.07
Overview

1. Web IR
   - Links
   - Queries
   - Context
   - Users
   - Documents
   - Size

2. Ads & Spam
   - Ads
   - Spam
Web search overview
Web IR Ads & Spam

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%

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

- 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
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?

- 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)
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?

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 without search engines is not possible.
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?

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 without search engines is not possible.

- Somebody needs to pay for the web.
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?

- 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 without search engines is not possible.

- 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?
- 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 without search engines is not possible.
- Somebody needs to pay for the web.
  - Servers, web infrastructure, content creation
  - A large part today is paid by search ads.
Outline

1. Web IR
   - Links
   - Queries
   - Context
   - Users
   - Documents
   - Size

2. Ads & Spam
   - Ads
   - Spam
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.
- 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.
- 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?
- 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? $10^8$ every day, approaching $10^9$
- Users: Users are different, more varied and there are a lot of them. How many? $10^9$
- Documents: Documents are different, more varied and there are a lot of them. How many? $\approx 10^{11}$. Indexed: $10^{10}$
- Context: Context is more important on the web than in many other IR applications.
- Ads and spam
Outline

1 Web IR
   - Links
   - Queries
   - Context
   - Users
   - Documents
   - Size

2 Ads & Spam
   - Ads
   - Spam
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
Outline

1 Web IR
   - Links
   - Queries
   - Context
   - Users
   - Documents
   - Size

2 Ads & Spam
   - Ads
   - Spam
# Query distribution (1)


<table>
<thead>
<tr>
<th>Rank</th>
<th>Query</th>
<th>Rank</th>
<th>Query</th>
<th>Rank</th>
<th>Query</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><a href="http://www.hotmail.com">www.hotmail.com</a></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>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>Caramail</td>
<td>47</td>
<td>msn</td>
<td>48</td>
<td>jennifer lopez</td>
</tr>
<tr>
<td>49</td>
<td>tits</td>
<td>50</td>
<td>free porn</td>
<td>51</td>
<td>cheats</td>
</tr>
<tr>
<td>52</td>
<td>yahoo.com</td>
<td>53</td>
<td>eminem</td>
<td>54</td>
<td>Christina Aguilera</td>
</tr>
<tr>
<td>55</td>
<td>incest</td>
<td>56</td>
<td>letras de canciones</td>
<td></td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>hardcore</td>
<td>58</td>
<td>weather</td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>wallpapers</td>
<td>60</td>
<td>lingerie</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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  www.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.
Query distribution (1)


<table>
<thead>
<tr>
<th></th>
<th>Query</th>
<th></th>
<th>Query</th>
<th></th>
<th>Query</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>
<td>46</td>
</tr>
<tr>
<td>2</td>
<td>(artifact)</td>
<td>17</td>
<td>games</td>
<td>32</td>
<td>nude</td>
<td>47</td>
</tr>
<tr>
<td>3</td>
<td>(artifact)</td>
<td>18</td>
<td>pussy</td>
<td>33</td>
<td>music</td>
<td>48</td>
</tr>
<tr>
<td>4</td>
<td>porno</td>
<td>19</td>
<td>cracks</td>
<td>34</td>
<td>musica</td>
<td>49</td>
</tr>
<tr>
<td>5</td>
<td>mp3</td>
<td>20</td>
<td>lolita</td>
<td>35</td>
<td>anal</td>
<td>50</td>
</tr>
<tr>
<td>6</td>
<td>Halloween</td>
<td>21</td>
<td>britney spears</td>
<td>36</td>
<td>free6</td>
<td>51</td>
</tr>
<tr>
<td>7</td>
<td>sexo</td>
<td>22</td>
<td>ebay</td>
<td>37</td>
<td>avril lavigne</td>
<td>52</td>
</tr>
<tr>
<td>8</td>
<td>chat</td>
<td>23</td>
<td>sexe</td>
<td>38</td>
<td><a href="http://www.hotmail.com">www.hotmail.com</a></td>
<td>53</td>
</tr>
<tr>
<td>9</td>
<td>porn</td>
<td>24</td>
<td>Pamela Anderson</td>
<td>39</td>
<td>winzip</td>
<td>54</td>
</tr>
<tr>
<td>10</td>
<td>yahoo</td>
<td>25</td>
<td>warez</td>
<td>40</td>
<td>fuck</td>
<td>55</td>
</tr>
<tr>
<td>11</td>
<td>KaZaA</td>
<td>26</td>
<td>divx</td>
<td>41</td>
<td>wallpaper</td>
<td>56</td>
</tr>
<tr>
<td>12</td>
<td>xxx</td>
<td>27</td>
<td>gay</td>
<td>42</td>
<td>hotmail.com</td>
<td>57</td>
</tr>
<tr>
<td>13</td>
<td>Hentai</td>
<td>28</td>
<td>harry potter</td>
<td>43</td>
<td>postales</td>
<td>58</td>
</tr>
<tr>
<td>14</td>
<td>lyrics</td>
<td>29</td>
<td>playboy</td>
<td>44</td>
<td>shakira</td>
<td>59</td>
</tr>
<tr>
<td>15</td>
<td>hotmail</td>
<td>30</td>
<td>lolitas</td>
<td>45</td>
<td>traductor</td>
<td>60</td>
</tr>
</tbody>
</table>

More than 1/3 of these are queries for adult content. Does this mean that most people are looking for adult content?
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
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

- **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. Web IR
   - Links
   - Queries
   - Context
   - Users
   - Documents
   - Size

2. Ads & Spam
   - Ads
   - Spam
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:
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)
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

- 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

- 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?

- 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 Web IR
   - Links
   - Queries
   - Context
   - Users
     - Documents
     - Size

2 Ads & Spam
   - Ads
   - Spam
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?

- 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 IR  Ads & Spam

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
Outline

1. Web IR
   - Links
   - Queries
   - Context
   - Users
   - Documents
   - Size

2. Ads & Spam
   - Ads
   - Spam
Web documents: different from other IR collections

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

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

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

- Distributed content creation: no design, no co-ordination
  - “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 co-ordination
  - “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

Browser → Application server → Back-end databases
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)

- 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.
Dynamic content (2)

- 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)
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

- 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
For many collections, it is easy to assess the trustworthiness of a document.
Trust

- 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
Trust

- 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
Trust

- 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. Web IR
   - Links
   - Queries
   - Context
   - Users
   - Documents
   - Size

2. Ads & Spam
   - Ads
   - Spam
Growth of the web
Growth of the web

- The web keeps growing.
Growth of the web

- The web keeps growing.
- But growth is no longer exponential?
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”)
“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 would you estimate the number of pages indexed by a web search engine?
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 Web IR
   - Links
   - Queries
   - Context
   - Users
   - Documents
   - Size

2 Ads & Spam
   - Ads
   - Spam
Outline

1 Web IR
   - Links
   - Queries
   - Context
   - Users
   - Documents
   - Size

2 Ads & Spam
   - Ads
   - Spam
First generation of search ads: Goto (1996)
First generation of search ads: Goto (1996)

- No separation of ads/docs. Just one result!
First generation of search ads: Goto (1996)

- No separation of ads/docs. Just one result!
- Buddy Blake bid the maximum ($0.38) for this search.
First generation of search ads: Goto (1996)

- No separation of ads/docs. Just one result!
- 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)

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

- Strict separation of search results and search ads
## Ranking of advertisers in search results

<table>
<thead>
<tr>
<th>Discount Broker Reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information on online discount brokers emphasizing rates, charges, and customer comments and complaints. <a href="http://www.broker-reviews.us">www.broker-reviews.us</a> - 94k - Cached - Similar pages</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Discount Broker Rankings (2008 Broker Survey) at SmartMoney.com</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Stock Brokers</th>
<th>Discount Brokers</th>
<th>Online Brokers</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Discount Broker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discount Broker - Definition of Discount Broker on Investopedia - A stockbroker who carries out buy and sell orders at a reduced commission compared to a ... [<a href="http://www.investopedia.com/terms/d/discountbroker.asp">www.investopedia.com/terms/d/discountbroker.asp</a> - 31k - Cached - Similar pages](<a href="http://www.investopedia.com/terms/d/discountbroker.asp">http://www.investopedia.com/terms/d/discountbroker.asp</a> - 31k - Cached - Similar pages)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Discount Brokerage and Online Trading for Smart Stock Market ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online stock broker SogoTrade offers the best in discount brokerage investing. Get stock market quotes from this internet stock trading company. <a href="http://www.sogotrade.com">www.sogotrade.com</a> - 39k - Cached - Similar pages</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>15 questions to ask discount brokers - MSN Money</th>
</tr>
</thead>
<tbody>
<tr>
<td>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 - Cached - Similar pages](<a href="http://moneycentral.msn.com/content/investing/Startinvesting/P66171.asp">http://moneycentral.msn.com/content/investing/Startinvesting/P66171.asp</a> - 34k - Cached - Similar pages)</td>
</tr>
</tbody>
</table>
Ranking of advertisers in search results

SogoTrade appears in ads.
Ranking of advertisers in search results

SogoTrade appears in search results.

SogoTrade appears in ads.
Ranking of advertisers in search results

SogoTrade appears in search results.

SogoTrade appears in ads.

Do search engines rank advertisers higher than non-advertisers?
Ranking of advertisers in search results

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?

- 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 ads placed?

- Advertisers bid for keywords.
How are ads placed?

- Advertisers bid for keywords.
- Open system: Anybody can participate and bid on keywords.
How are ads placed?

- Advertisers bid for keywords.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are only charged when somebody clicks on your ad.
How are ads placed?

- Advertisers bid for keywords.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are only charged when somebody clicks on your ad.
- How does the advertiser determine its bid price and how does the search engine take this into account in ranking and billing?
How are ads placed?

- Advertisers bid for keywords.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are only charged when somebody clicks on your ad.
- How does the advertiser determine its bid price and how does the search engine take this into account in ranking and billing?
  - Basis is a second price auction, but with twists
How are ads placed?

- Advertisers bid for keywords.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are only charged when somebody clicks on your ad.
- How does the advertiser determine its bid price and how does the search engine take this into account in ranking and billing?
  - Basis is a second price auction, but with twists
  - Lots of interesting work on this
How are ads placed?

- Advertisers bid for keywords.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are only charged when somebody clicks on your ad.
- How does the advertiser determine its bid price and how does the search engine take this into account in ranking and billing?
  - Basis is a second price auction, but with twists
  - Lots of interesting work on this
  - Squeeze an additional fraction of a cent from each ad means billions of additional revenue for the search engine.
How are ads placed?

- Advertisers bid for keywords.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are only charged when somebody clicks on your ad.
- How does the advertiser determine its bid price and how does the search engine take this into account in ranking and billing?
  - Basis is a second price auction, but with twists
  - Lots of interesting work on this
  - Squeeze an additional fraction of a cent from each ad means billions of additional revenue for the search engine.
  - Many interesting jobs in this area
Keywords with high bids

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

- $69.1  mesothelioma treatment options
- $65.85  personal injury lawyer michigan
- $62.59  student loans consolidation
- $61.44  car accident attorney los angeles
- $59.44  online car insurance quotes
- $59.39  arizona dui lawyer
- $57.87  michigan car accident attorney
- $56.59  free auto insurance quote
- $53.17  personal injury lawyers los angeles
- $52.31  free online auto insurance quote
- $50.4  accident attorney michigan
- $50.35  michigan auto accident attorney
- $49.25  accident helpline
- $47.74  automobile accident lawyers
- $47.49  dui defense attorneys
- $46.44  asbestos cancer
- $46.34  arizona dui
- $45.8  business liability insurance quote
- $43.86  loan consolidation
- $42.98  student loan consolidation
- $40.7  dui defense lawyers
- $40.1  home equity line of credit
- $39.81  life insurance quotes
- $39.78  criminal lawyers new york
- $39.32  loan federal consolidation
- $39.23  refinancing
- $38.72  equity line of credit
- $37.96  lasik eye surgery new york city
- $37  2nd mortgage
- $35.9  free car insurance quote
- $35.86  pennsylvania medical malpractice attorney
- $35.86  medical malpractice ohio
- $35.71  automobile insurance quote
- $35.4  loan consolidating
- $35.34  commercial insurance quote
- $35.33  tax attorney
- $35.15  home equity loans
- $34.81  instant auto insurance quotes
- $34.8  home equity loan rates
- $34.79  home owners insurance quotes
- $34.71  home equity line
- $34.53  compensation solicitors
- $34.38  automobile insurance quotes
- $34.37  term insurance quotes
- $34.26  instant car insurance quotes
- $34.02  auto insurance online quote
- $33.49  new york criminal attorney
- $33.45  secured loan
- $33.44  equity lines
- $33.41  criminal lawyer new york
- $33.36  refinance mortgage
- $33.12  equity loan rates
- $32.46  equity line
- $32.45  home equity credit
- $32.02  loan consolidate
- $31.98  secured loan consolidation
- $31.93  laser hair removal new york city
- $31.51  home equity rates
- $31.37  free credit report com
Google AdWords demo
How are ads ranked?

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

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

- First cut: according to bid price
  - Bad idea: open to abuse
  - Example: query *accident* → ad *buy a new car*
How are ads ranked?

- First cut: according to bid price
  - Bad idea: open to abuse
  - Example: query *accident* → ad *buy a new car*
- Instead: rank based on bid price *and* relevance
How are ads ranked?

- First cut: according to bid price
  - Bad idea: open to abuse
  - Example: query *accident* → ad *buy a new car*
- 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
  - Bad idea: open to abuse
  - Example: query *accident* → ad *buy a new car*
- Instead: rank based on bid price *and* relevance
- Key measure of ad relevance: clickthrough rate
- Result: A non-relevant ad will be ranked low.
How are ads ranked?

- First cut: according to bid price
  - Bad idea: open to abuse
  - Example: query *accident* → ad *buy a new car*
- Instead: rank based on bid price *and* relevance
- Key measure of ad relevance: clickthrough rate
- Result: A non-relevant ad will be ranked low.
  - Even if this decreases search engine revenue short-term
How are ads ranked?

- First cut: according to bid price
  - Bad idea: open to abuse
  - Example: query *accident* → ad *buy a new car*
- Instead: rank based on bid price and relevance
- Key measure of ad relevance: clickthrough rate
- Result: A non-relevant 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
  - Bad idea: open to abuse
  - Example: query *accident* → ad *buy a new car*
- Instead: rank based on bid price and relevance
- Key measure of ad relevance: clickthrough rate
- Result: A non-relevant 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
  - Bad idea: open to abuse
  - Example: query *accident* → ad *buy a new car*
- Instead: rank based on bid price and relevance
- Key measure of ad relevance: clickthrough rate
- Result: A non-relevant 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 factor of course is the query – and possibly earlier queries in the same session.
Ranking of advertisers in search results

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 - Cached - 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 - Cach ed - Similar pages

Sponsored Links

Rated #1 Online Broker
No Minimums. No Inactivity Fee Transfer to Fistrade for Free!
www.fistrade.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
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.
  - Being willing to pay for ads on a search engine is a quality signal (one of many) that users take into account.
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.
  - Being willing to pay for ads on a search engine is a quality signal (one of many) that users take into account.
- The **advertiser** finds new customers in a cost-effective way.
The appeal of search ads to advertisers

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

- Why is web search potentially more attractive for advertisers than TV spots, newspaper ads or radio spots?
- Someone who just searched for “Saturn Aura Sport Sedan” is infinitely more likely to buy one than a random person watching TV.
The appeal of search ads to advertisers

- Why is web search potentially more attractive for advertisers than TV spots, newspaper ads or radio spots?
- Someone who just searched for “Saturn Aura Sport Sedan” is infinitely more likely to buy one than a random person watching TV.
- Most importantly, the advertiser only pays if the customer took an action indicating interest (i.e., clicking on the ad).
But frequently it’s not a win-win-win

- Example: keyword arbitrage
But frequently it’s not a win-win-win

- Example: keyword arbitrage
  - Buy a keyword at Google
But frequently it’s not a win-win-win

- Example: keyword arbitrage
  - Buy a keyword at Google
  - Then redirect traffic to a third party that is paying much more than you had to pay to Google
But frequently it’s not a win-win-win

Example: keyword arbitrage
  - Buy a keyword at Google
  - Then redirect traffic to a third party that is paying much more than you had to pay to Google
  - This rarely makes sense for the user.
But frequently it’s not a win-win-win

- Example: keyword arbitrage
  - Buy a keyword at Google
  - Then redirect traffic to a third party that is paying much more than you had to pay to Google
  - This rarely makes sense for the user.

- Ad spammers keep inventing new tricks.
But frequently it’s not a win-win-win

- Example: keyword arbitrage
  - Buy a keyword at Google
  - Then redirect traffic to a third party that is paying much more than you had to pay to Google
  - This rarely makes sense for the user.
- Ad spammers keep inventing new tricks.
- The search engines need time to catch up with them.
Who owns a search term?

- Example: geico
Who owns a search term?

- Example: geico
- During part of 2005: The search term “geico” on Google was bought by competitors.
Who owns a search term?

- Example: geico
- During part of 2005: The search term “geico” on Google was bought by competitors.
- Geico lost this case in the United States.
Who owns a search term?

- Example: geico
- During part of 2005: The search term “geico” on Google was bought by competitors.
- Geico lost this case in the United States.
- Currently in the courts: Louis Vuitton case in Europe
Who owns a search term?

- Example: geico
- During part of 2005: The search term “geico” on Google was bought by competitors.
- Geico lost this case in the United States.
- Currently in the courts: Louis Vuitton case in Europe
- See http://google.com/tm_complaint.html
Outline

1 Web IR
   - Links
   - Queries
   - Context
   - Users
   - Documents
   - Size

2 Ads & Spam
   - Ads
   - Spam
The goal of spamming on the web

- You have a page that will generate lots of revenue for you if people visit it.
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.
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.
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.
- How can I get my page ranked highly?
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

Something else, this page is about keyword stuffing. It's a practice where web pages are overloaded with keywords to try to improve search engine rankings. It's considered a black hat SEO technique. The text on this page is a mix of unrelated words and phrases, possibly generated automatically by a machine. It's not clear what the intended message is, and it may not be relevant to the topic of keyword stuffing.

---

**Schütze: Web search basics**

---

57 / 66
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
Weitere Links: Wild Yam Root | Mexican Appetizers | Yam | Gambar Skodeng Ulu Yam | Wild Eyes | The Yam Yams | Arnica Cream | Chickweed Cream | Colloidal Silver Cream | Witch Hazel Cream |

**COMPOSITA.COM**

Sponsored Links

- **Wild Russian Girls**
  Plenty of Russian Girls interested in building a Happy Marriage.
  [uk.anastasia-international.com](http://uk.anastasia-international.com)

- **Wild Yam 10%**
  By HPLC, Supply 500Kg/mon from 100% natural herb
  [www.tonsonido.com](http://www.tonsonido.com)

- **Suche dir eine Frau aus**
  Sofort Kontakte zu Frauen Ohne Anmeldung, kostenlos starten!
  [www.SMS-Contacts.de/Sexy](http://www.SMS-Contacts.de/Sexy)

- **Yamaha Boats For Sale**
  Find, Buy and Sell the Right Boat! Free Text/Email Alert Service
  [rightboat.com/adverts/Yamaha.html](http://rightboat.com/adverts/Yamaha.html)

- **Wild Yam Root**
  Harvested at height of potancy. 20 Year, Family Run Herb Company.

**WEITERE LINKS**

- Wild Yam Root
- Mexican Appetizers
- Yam
- Gambar Skodeng Ulu Yam
- Wild Eyes
- The Yam Yams
- Arnica Cream
- Chickweed Cream
- Colloidal Silver Cream
- Witch Hazel Cream
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

- 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

Is this a search engine crawler?

- Yes: Serve misleading content
- No: Serve spam
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

- 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

- 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.
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.
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.
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
**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
  - Add more interesting and original content
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

- 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
Resources

- Chapter 19 of IIR
Resources

- Chapter 19 of IIR
- Resources at http://ifnlp.org(ir
Resources

- Chapter 19 of IIR
- Resources at http://ifnlp.org/ir
- Size of the web queries
Resources

- Chapter 19 of IIR
- Resources at http://ifnlp.org/ir
- Size of the web queries
- Trademark issues (Geico and Vuitton cases)
Resources

- Chapter 19 of IIR
- Resources at http://ifnlp.org/ir
- Size of the web queries
- Trademark issues (Geico and Vuitton cases)
- How ads are priced
Resources

- Chapter 19 of IIR
- Resources at http://ifnlp.org/ir
- Size of the web queries
- Trademark issues (Geico and Vuitton cases)
- How ads are priced
- How search engines fight webspam
Resources

- Chapter 19 of IIR
- Resources at http://ifnlp.org/ir
- Size of the web queries
- Trademark issues (Geico and Vuitton cases)
- How ads are priced
- How search engines fight webspam
- Adversarial IR site at Lehigh