Information Retrieval and Web Search Engines

Lecture 10: Introduction to Web Retrieval
June 22, 2011

Wolf-Tilo Balke and Joachim Selke
Institut für Informationssysteme
Technische Universität Braunschweig
Exercise 7.1
Why is it a good idea to use a maximum-margin classifier?
Exercise 7.2
What is the kernel trick?
Exercise 7.3
What is Learning to Rank and in which way is it similar to traditional SVM classification?
Exercise 7.4

Describe the problem of overfitting. What can be done to avoid it?
Three models:

- **Google Mini**: Index up to 300,000 documents
- **GB-7007**: Index up to 10,000,000 documents
- **GB-9009**: Index up to 30,000,000 documents
Google Search Appliance

Features:
- Fully-automatic indexing
- Web-based search interface
- Supports 220 file types
- Searches databases and content management systems
- Accessible from outside via API calls
- Includes a Google t-shirt

Costs:
- Google Mini: Starts at $2,000
- Other versions: $30,000 and up
Thinking about how using the internet affects you overall... How much, if at all, has the internet improved...?

Source: pewinternet.org
Most People Use the Web

Change in internet use by age, 2000-2010 (US)

- Teens 12-17*: 93%
- All adults 18 and older: 79%
- Adults 18-29: 95%
- Adults 30-49: 87%
- Adults 50-64: 78%
- Adults 65 and older: 42%

* Teens data is from Sept-09.

Source: Pew Internet & American Life Project
### What Do People Do?

**Source:** pewinternet.org

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Email</td>
<td>90-100%</td>
<td>40-49%</td>
<td>80-89%</td>
<td>30-39%</td>
<td>70-79%</td>
<td>20-29%</td>
</tr>
<tr>
<td>Search</td>
<td>80-89%</td>
<td>30-39%</td>
<td>70-79%</td>
<td>20-29%</td>
<td>60-69%</td>
<td>10-19%</td>
</tr>
<tr>
<td>Health info</td>
<td>60-69%</td>
<td>10-19%</td>
<td>50-59%</td>
<td>0-9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social network sites</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Watch video</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Get news</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Travel reservations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buy a product</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listen to music</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Travel reservations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online classified</td>
<td>90-100%</td>
<td>40-49%</td>
<td>80-89%</td>
<td>30-39%</td>
<td>70-79%</td>
<td>20-29%</td>
</tr>
<tr>
<td>Bank online</td>
<td>80-89%</td>
<td>30-39%</td>
<td>70-79%</td>
<td>20-29%</td>
<td>60-69%</td>
<td>10-19%</td>
</tr>
<tr>
<td>Govt website</td>
<td>60-69%</td>
<td>10-19%</td>
<td>50-59%</td>
<td>0-9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Play games</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read blogs</td>
<td>90-100%</td>
<td>40-49%</td>
<td>80-89%</td>
<td>30-39%</td>
<td>70-79%</td>
<td>20-29%</td>
</tr>
<tr>
<td>Financial info</td>
<td>80-89%</td>
<td>30-39%</td>
<td>70-79%</td>
<td>20-29%</td>
<td>60-69%</td>
<td>10-19%</td>
</tr>
<tr>
<td>Rate things</td>
<td>60-69%</td>
<td>10-19%</td>
<td>50-59%</td>
<td>0-9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Religious info</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online auction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Podcasts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Donate to charity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blog</td>
<td>90-100%</td>
<td>40-49%</td>
<td>80-89%</td>
<td>30-39%</td>
<td>70-79%</td>
<td>20-29%</td>
</tr>
<tr>
<td>Virtual worlds</td>
<td>80-89%</td>
<td>30-39%</td>
<td>70-79%</td>
<td>20-29%</td>
<td>60-69%</td>
<td>10-19%</td>
</tr>
<tr>
<td>Virtual worlds</td>
<td>60-69%</td>
<td>10-19%</td>
<td>50-59%</td>
<td>0-9%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Key:** % of internet users in each generation who engage in this online activity.
Web Search is Essential

• Without Web search, **content** cannot be found
  – Why create online content if nobody will read it?
  – Only for very popular topics, Web search can be replaced by Web directories like DMOZ

• Without Web search, there would be less **collaboration**
  – How to find people with similar interests and problems?
  – What open source projects would be possible without Web search? What about the Social Web?

• Without Web search, **bills** cannot be paid
  – Infrastructure, servers, and content cost a lot of money
  – This is largely paid by search ads
A typical Web search engine:

- **Users**
- **The Web**
- **User interface**
- **Business model**
- **Retrieval algorithms**
- **Web crawler**
- **Indexer**
Lecture 10: 
Introduction to Web Retrieval

1. **Web Retrieval vs. Classical IR**
2. What Does the Web Look Like?
3. How Do Users Use the Web?
Web Retrieval vs. Classical IR

• **Heterogeneity**
  – Many different users, topics, languages, document types, …
  – Websites are not classical documents (dynamic content, …)
  – Open platform: variety of authors, opinions, writing styles, …

• **Hyperlinks**
  – Documents are connected and refer to each other

• **Problem size**
  – Many documents, many queries, high percentage of volatile data

• **Spam**
  – Evil forces are around

• **Business model**
  – Web search is expensive
# Heterogeneity of Users

## World internet usage (all numbers in millions):

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>767</td>
<td>1038</td>
<td>5</td>
<td>119</td>
</tr>
<tr>
<td>Asia</td>
<td>3557</td>
<td>3880</td>
<td>114</td>
<td>922</td>
</tr>
<tr>
<td>Europe</td>
<td>729</td>
<td>816</td>
<td>105</td>
<td>476</td>
</tr>
<tr>
<td>Middle East</td>
<td>176</td>
<td>216</td>
<td>3</td>
<td>69</td>
</tr>
<tr>
<td>North America</td>
<td>307</td>
<td>347</td>
<td>108</td>
<td>272</td>
</tr>
<tr>
<td>Latin America</td>
<td>511</td>
<td>597</td>
<td>18</td>
<td>216</td>
</tr>
<tr>
<td>Oceania</td>
<td>30</td>
<td>35</td>
<td>8</td>
<td>21</td>
</tr>
<tr>
<td><strong>World</strong></td>
<td><strong>6007</strong></td>
<td><strong>6929</strong></td>
<td><strong>361</strong></td>
<td><strong>2095</strong></td>
</tr>
</tbody>
</table>

Source: internetworldstats.com
**Heterogeneity of Users (2)**

- Web users are not all alike
- Demographics of US Internet users (2008):

<table>
<thead>
<tr>
<th>Age</th>
<th>Use the Internet</th>
</tr>
</thead>
<tbody>
<tr>
<td>18–29</td>
<td>87%</td>
</tr>
<tr>
<td>30–49</td>
<td>82%</td>
</tr>
<tr>
<td>50–64</td>
<td>72%</td>
</tr>
<tr>
<td>65+</td>
<td>41%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Household income (per year)</th>
<th>Use the Internet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than $30,000</td>
<td>57%</td>
</tr>
<tr>
<td>$30,000–$49,999</td>
<td>77%</td>
</tr>
<tr>
<td>$50,000–$74,999</td>
<td>90%</td>
</tr>
<tr>
<td>$75,000+</td>
<td>94%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total adults</th>
<th>Use the Internet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women</td>
<td>75%</td>
</tr>
<tr>
<td>Men</td>
<td>73%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Education</th>
<th>Use the Internet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than high school</td>
<td>35%</td>
</tr>
<tr>
<td>High school</td>
<td>67%</td>
</tr>
<tr>
<td>Some college</td>
<td>85%</td>
</tr>
<tr>
<td>College+</td>
<td>95%</td>
</tr>
</tbody>
</table>

Source: pewinternet.org
Some statistics about the Web’s languages:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>72%</td>
<td>2,567,509</td>
</tr>
<tr>
<td>German</td>
<td>7%</td>
<td>808,044</td>
</tr>
<tr>
<td>Japanese</td>
<td>6%</td>
<td>523,639</td>
</tr>
<tr>
<td>Spanish</td>
<td>3%</td>
<td>402,430</td>
</tr>
<tr>
<td>French</td>
<td>3%</td>
<td>709,312</td>
</tr>
<tr>
<td>Italian</td>
<td>2%</td>
<td>499,234</td>
</tr>
<tr>
<td>Dutch</td>
<td>2%</td>
<td>481,064</td>
</tr>
<tr>
<td>Chinese</td>
<td>2%</td>
<td>205,047</td>
</tr>
<tr>
<td>Korean</td>
<td>1%</td>
<td>75,184</td>
</tr>
<tr>
<td>Portuguese</td>
<td>1%</td>
<td>429,730</td>
</tr>
</tbody>
</table>

Sources: oclc.org and wikipedia.org
Some file types a search engine should be able to process:

application/ms-excel (different versions), application/ms-powerpoint (different versions), application/msword (different versions), application/pdf (different versions), application/postscript, application/x-dvi, application/x-tar, application/x-zip-compressed, text/html (different versions and encodings), text/plain (different encodings), text/rtf, application/xml, text/xml, application/xhtml+xml, application/docbook+xml, application/x-shockwave-flash, …

— Images, videos, audio, executable code?
Web search engines are used for different purposes and within different contexts.

There are four main types of queries:

- **Informational queries:**
  Find general information about some topic, e.g., “Web search”

- **Navigational queries:**
  Find a specific website, e.g., “Facebook”

- **Transactional queries:**
  Find websites providing some service, e.g., “Adobe Reader download”

- **Connectivity queries:**
  Find connected pages, e.g., “link:www.tu-bs.de” (finds all pages that link to http://www.tu-bs.de)
Heterogeneity of Queries (2)

Ask.com’s **top searches** for the week ending Jan 16\textsuperscript{th}, 2008:

1. MySpace
2. Facebook
3. YouTube
4. Angelina Jolie
5. Online Dictionary
6. Craigslist
7. eBay
8. Wikipedia
9. eMail
10. How to get pregnant

Information Retrieval and Web Search Engines — Wolf-Tilo Balke and Joachim Selke — Technische Universität Braunschweig
Again, some statistics...

<table>
<thead>
<tr>
<th>Activity</th>
<th>Percent of Internet users who report doing this “yesterday”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use the Internet</td>
<td>72%</td>
</tr>
<tr>
<td>Use a search engine to find information</td>
<td>49%</td>
</tr>
<tr>
<td>Get news</td>
<td>36%</td>
</tr>
<tr>
<td>Check the weather</td>
<td>30%</td>
</tr>
<tr>
<td>Look for info on a hobby or interest</td>
<td>29%</td>
</tr>
<tr>
<td>Surf the Web for fun</td>
<td>28%</td>
</tr>
<tr>
<td>Do any type of research for your job</td>
<td>23%</td>
</tr>
<tr>
<td>Use an online social networking site</td>
<td>19%</td>
</tr>
<tr>
<td>Research for school or training</td>
<td>16%</td>
</tr>
</tbody>
</table>

Source: pewinternet.org
• http://www.google.com/trends
• http://www.google.com/trends/hottrends
• Web documents can **link** to each other
• **Links are not created randomly**

Two different topics?

This page seems to be interesting
Number of Queries

• How many queries a search engine has to process?
• Here are some numbers from 2005:

<table>
<thead>
<tr>
<th></th>
<th>Average number of queries per second</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>700</td>
</tr>
<tr>
<td>Yahoo</td>
<td>600</td>
</tr>
<tr>
<td>MSN</td>
<td>300</td>
</tr>
<tr>
<td>AOL</td>
<td>180</td>
</tr>
</tbody>
</table>

• 700 queries per second are...
  – …around 60 million queries per day
  – …around 22 billion queries per year

• Today, it’s definitely much more than 700 queries…
Index Size

- How large is a typical Web search engine’s index?
- Here are some recent estimates from worldwidewebsize.com (as of June 2010):

<table>
<thead>
<tr>
<th></th>
<th>Number of indexed Web pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yahoo</td>
<td>50,000,000,000</td>
</tr>
<tr>
<td>Google</td>
<td>16,700,000,000</td>
</tr>
<tr>
<td>Bing</td>
<td>8,100,000,000</td>
</tr>
<tr>
<td>Ask</td>
<td>1,700,000,000</td>
</tr>
</tbody>
</table>

- By the way: Where did they get these numbers from?
The authors of worldwidewebsize.com describe their estimation method as follows:

- Obtain **word frequencies** from a large offline text collection
  - More than 1 million web pages from DMOZ
  - Can be considered a representative sample of the World Wide Web

- **Send 50 randomly chosen words** to the search engine
  - “Randomly” = selected evenly across logarithmic frequency intervals

- For each word, **record the number of Web pages found**

- **Estimate the index size** using these numbers by exploiting the **relative word frequencies** of the background corpus
When operating a search engine, you need a crawler.

The crawler must continuously feed the indexer with new or updated information:
- New Web pages
- Deleted Web pages
- Updated Web pages

How much data must be transferred for doing this?

Some recent numbers from netcompetition.org:
- Within the US part of the Internet, Google transfers around 60 petabytes per month: 60,000,000,000 megabytes!

Now you know why Web search is expensive...
• The Web grows fast (exponentially?)…
• The total number of hostnames:

Source: netcraft.com

• A Web search engine must **scale well** to keep up
We have seen: Web search is complicated and expensive
  – Exception: Local search functionality for a single Web site
You cannot run a Web search engine for free
  – Hardware, traffic, development, …
What could be a reasonable business model here?
  – Advertising model
  – Subscription model
  – Community model
  – Infomediary model
• The advertising model
  – You get paid for showing other people’s ads on your search result pages
  – Used by Google and most other search engines
  – To make this work, your search engine must attract a lot of people and placement of ads must be personalized
  – If your search engine fails at the former, there are other ways: In Microsoft’s “Live Search cashback” program, people earn some money if they buy products found via Live Search’s ads

Get cashback from Live Search!
Use Live Search to find cashback savings from the online stores you know and trust.
See how this works
• The subscription model
  – Customers pay for using your search engine
  – To make this work, your search engine must be really good
  – More popular: Rent your technology to other companies; many search engines use this model
  – Example: t-online.de’s search functionality is provided by Google
• The community model
  – Let users participate in product development
  – This lowers costs and often increases product quality
  – Pay your bills by ads and donations
  – Example: Wikia Search, in which users can directly annotate or even modify search results (discontinued in May 2009)
The infomediary model

- Users can use your search engine for free but agree to participate in “market studies”
- The users’ search behavior is analyzed to yield individual “user profiles” and to distill overall search trends
- This information is sold to other companies, which can use it to optimize their own advertising strategies
- This model usually comes along with severe legal issues regarding the users’ privacy
- Examples: No search engine would tell about…
Google’s ad program is called AdWords
It’s very successful

- 99% of Google’s revenue is derived from its advertising programs
- In 2007, Google had 1 million advertisers
  2003: 89,000   2005: 360,000
  2004: 201,000   2006: 600,000
- In 2007, on average, each advertiser spent $16,000 a year on Google ads
- In 2010, Google earned $28 billion with ads

https://adwords.google.com/select/KeywordToolExternal
Google’s Business Model

Keywords with high bids (according to cwire.org):

<table>
<thead>
<tr>
<th>Bid</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>$69.10</td>
<td>mesothelioma treatment options</td>
</tr>
<tr>
<td>$66.46</td>
<td>mesothelioma risk</td>
</tr>
<tr>
<td>$65.85</td>
<td>personal injury lawyer michigan</td>
</tr>
<tr>
<td>$65.74</td>
<td>michigan personal injury attorney</td>
</tr>
<tr>
<td>$62.59</td>
<td>student loans consolidation</td>
</tr>
<tr>
<td>$61.44</td>
<td>car accident attorney los angeles</td>
</tr>
<tr>
<td>$61.26</td>
<td>mesothelioma survival rate</td>
</tr>
<tr>
<td>$60.96</td>
<td>treatment of mesothelioma</td>
</tr>
<tr>
<td>$59.44</td>
<td>online car insurance quotes</td>
</tr>
<tr>
<td>$59.39</td>
<td>arizona dui lawyer</td>
</tr>
<tr>
<td>$59.04</td>
<td>mesothelioma article</td>
</tr>
</tbody>
</table>
• There are cheaper ways than AdWords to get your page on Google’s result pages…
• Just let your page look as if it would be highly relevant…
• The general term for such techniques is “spamdexing”
Lecture 10: Introduction to Web Retrieval

1. Web Retrieval vs. Classical IR
2. What Does the Web Look Like?
3. How Do Users Use the Web?
• In 2002, (Fetterly et al., 2004) crawled a set of around 151 million HTML pages once every week, over a span of 11 weeks.

• Amongst others, they tried to answer the following questions:
  – How large is a Web page (measured in bytes)?
  – How large is a Web page (measured in words)?
  – How much does a Web page change (within a week)?
How large is a Web page (measured in bytes)?

“19” means a page size of $2^{19}$ bytes.
How large is a Web page (measured in words)?

“3” means a page size of $2^3$ words
How much does a Web page change (within a week)?
• In 1993, measuring the Web’s size has been easy
  – Every web page corresponded to a file on some server
  – There was almost no duplicate content
  – There was no spam
  – Most Web servers have been known explicitly

• Estimation of 1993:
  – 100 servers
  – 200,000 documents
  – 4,000,000 pages

• Today, estimating the Web’s size is more difficult
First problem: **What pages counts as “the Web”?**

**How to handle duplicates?**
What pages counts as “the Web”? 

How to handle spam?
What content counts as “the Web”?

How many different pages should we count in this case?
What content counts as “the Web”?

How to handle sites that require users to login?
Now, what pages should be counted?

- **Duplicates:**
  Ignore them!

- **Spam:**
  Ignore it!

- **Dynamic Web pages (e.g. database interfaces):**
  Count them but try to focus on the actual information; maybe it is better to count in megabytes instead of pages…

- **(More or less public) private pages:**
  Count them if they can be accessed by a large number of people

Well, now we have defined what should be counted

**But… How to do it?**
• **How to find all Web pages?**
  – Just follow the links…
• What about pages nobody links to?
• How to detect duplicates?
• How to detect spam?
• How to crawl Web sites with dynamic pages?
• How to access (more or less public) private pages?

A lot of interesting questions to be solved by Web crawlers and indexers!

Let’s answer them next week…
Let’s assume for now, that we have some Web crawler that can automatically solve all these problems as good as currently possible.

Then, calculating the Web’s size is easy: Simply crawl the complete Web and count its number of pages or its size in megabytes!

Bad news: This doesn’t work due to the Web’s enormous size
- It would either take forever or require an enormous effort
- The Web has changed completely until the crawl is finished

Any better ideas?
A better approach is called “mark and recapture”: Take two (large) random samples of the Web and compute the Web’s total size by looking at the overlap.

**Idea:**

- Let $f$ be the number of pages found in the first crawl
- Let $s$ be the number of pages found in the second crawl
- Let $b$ be the number of pages found in both crawls
- Then, the probability of finding a page from the first crawl by randomly choosing a single Web page is about $\frac{b}{s}$
- On the other hand, this probability is about $\frac{f}{t}$, where $t$ denotes the (unknown) total number of Web pages
- Taken together, we get $t = \frac{f \cdot s}{b}$
In practice, one takes random samples from the index of different search engines.

Of course, we cannot assume anymore that these draws have been independent.

There are more advanced methods to account for this...

In 2005, the Web has been estimated to contain at least 11.5 billion pages.

Nobody knows exactly…
Of course, these estimates only cover the so-called "surface web," i.e., the part of the Web that can be accessed automatically by current Web crawlers. Even today’s best Web crawlers cannot find pages without in-links or all pages that have been generated dynamically.

The term "Deep Web" refers to all web pages that currently are not indexed by any Web search engine.

There are different estimates on the Deep Web’s size. The Deep Web is 15–500x as large as the surface Web.
Some types of “deep resources”:

- Dynamic content that cannot be accessed automatically, e.g. pages that are generated dynamically after filling out **Web forms**
- Unlinked or private content
- “Scripted” content, which requires code execution (e.g. Java, JavaScript, or Flash)
- “Strange” file formats not handled by current search engines
The Web Graph

- We can view the **static Web** consisting of static HTML pages together with the hyperlinks between them as a **directed graph**
  - Each Web page is a node
  - Each hyperlink is a directed edge
- The hyperlinks into a page are called **in-links**
- The hyperlinks out of a page are called **out-links**
• There is evidence that these links are not randomly distributed
• The distribution of in-links seems to follow a power law
  – The total number of pages having exactly $k$ in-links is proportional to $1 / k^{2.1}$
• Furthermore, several studies have suggested that the Web graph has a bowtie shape:
The Web Graph (3)

Note: The numbers given are as of 2000
Lecture 10: Introduction to Web Retrieval

1. Web Retrieval vs. Classical IR
2. What Does the Web Look Like?
3. How Do Users Use the Web?
Page Popularity

Page popularity is approximately Zipf distributed:

The Zipf curve is a straight line in log–log scale

The end of the “long tail” is absent

Source: useit.com
Incoming traffic from other sites follows Zipf's law:

Source: useit.com
• Several studies analyzed users’ query behavior:
  – The **average length** of a query is **2.4 terms**
  – About **half of all queries** consist of a **single term**
  – About **half of the users** looked only at the **first 20 results**
  – Less than 5% of users use advanced search features (e.g., Boolean operators)
  – About **20%** of all queries contain a **geographic term**
  – About **a third of the queries** from the same user were **repeated queries**; about 90% of the time the user would click on the same result
  – **Term frequency distributions** conform to the **power law**
• Web crawling
• Duplicate detection