Previous lecture

- **LSI (Latent Semantic Indexing)** - concepts
  - The term-document matrix is decomposed into three other matrices of a special form by use of Singular Value Decomposition (SVD)
  - The matrices show a breakdown of the original relationships into linearly independent components
  - Many of these components are very small and can be ignored - leading to an approximate model that contains fewer dimensions.
- **SVM (Support Vector Machines)** - classification

**LSI - reduced SVD**

- Reduce dimensionality => retain only $k$ largest singular values
- Saved space

$$A \approx A_k = U_k \Sigma_k V_k^T$$
LSI - Concept extraction

use rows of $\Sigma_k^{-1} U_k^T$ as concepts

<table>
<thead>
<tr>
<th>Concept 1</th>
<th>Concept 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>carlstrom</td>
<td>regia</td>
</tr>
<tr>
<td>rick</td>
<td>oct</td>
</tr>
<tr>
<td>amelnx</td>
<td>chrisp</td>
</tr>
<tr>
<td>advmar</td>
<td>problems</td>
</tr>
<tr>
<td>cuttings</td>
<td>pm</td>
</tr>
<tr>
<td>september</td>
<td>ip-forum</td>
</tr>
<tr>
<td>miller</td>
<td>stratification</td>
</tr>
<tr>
<td>re</td>
<td>uk</td>
</tr>
<tr>
<td>wants</td>
<td>bladderwort</td>
</tr>
<tr>
<td>aquatic</td>
<td>cuttings</td>
</tr>
<tr>
<td>rotundifolia</td>
<td></td>
</tr>
<tr>
<td>bladderwort</td>
<td></td>
</tr>
</tbody>
</table>

HARD to interpret

Text classification

- Goal: classify documents into predefined categories
- Examples
  - Subject classification: 'business', 'sports', 'engineering', ...
  - Review classification: 'positive' or 'negative'
  - Web page classification: 'Personal homepage' or others
- Approach: supervised machine learning ($\Rightarrow$ SVM)
  - Each predefined category needs a set of training documents
  - From training sets train a classifier
  - Use classifier to classify new documents

Automated Classification technologies

- Machine learning methods
  - Statistical models (Bayes, SVM, ...)
  - ANN
- Information Retrieval methods
  - Clustering (no predefined categories)
- Library Science methods
  - String matching + Thesaurus

SVM

- SVM maximize the margin around the separating hyper-plane
- Decision function specified by support vectors (from training examples)
- Quadratic programming problem

Hot text classification method
Lecture 5 agenda

Chapters 2, 11, 12 in “Modern Information Retrieval”

- Reiteration
- Web search
- Web search engines
- Web robots, crawler
- Focused Web crawling
- Web search vs Browsing
- Privacy, Filter bubble

Outline

- Reiteration
- Web search
- Web search engines
- Web robots, crawler
- Focused Web crawling
- Web search vs Browsing
- Privacy, Filter bubble

Why Web search ...

- Explosion of (digital) information within all types of information collections
- Harder and harder to follow information flow
- Faster way to find relevant information when its needed
- Challenges
  - Distributed, dynamic data
  - Large volume
  - Unstructured, heterogeneous data

Size of the Web

- no one knows
- estimates (text pages)
  - 2005 ‘more than 11.5 billion’
  - 2007 ‘more than 20 billion’
  - 2010 ‘20 - 55 billion’
- Google claims to know of $10^{12}$ unique URLs (text, images, ...)
Important questions

- How do I find relevant information?
- How do I navigate the digital information landscape?
- How structure and organize information to ease knowledge extraction?
- How to create collections, properly organized, with relevant material?
- How to keep collections updated?

Outline

- Reiteration
- Web search
  - Web search engines
  - Web robots, crawler
  - Focused Web crawling
  - Web search vs Browsing
  - Privacy, Filter bubble

Search Engine - Basic structure

- Database
- Web pages
- Web robot
- HTTP
- The Web
- Database Interface
- Query
- CGI–script
- Answer
- HTTP
- Web browser

Size of search engines

- not published
- guesses 1 - 20 - 50 billion pages
- overlap between search engines is small \(\approx 5 - 10\%\)

Software crawling the web (much like a human clicking on links)
- collect all found web-pages into a database (IR system)
- offer a web-interface to that database
Google

- started late 1990's
- Estimated 450,000 low-cost commodity servers (2006)
- 1 trillion links to web pages (July 2008)
- "over 8 billion web pages"
- estimate 40 billion pages?
- goal is to index all the world's data
- Google Flu Trends

Google Servers

The Joys of Real Hardware

Typical first year for a new cluster:

- 0.5 overheating (power down most machines in <5 mins, 1-2 days to recover)
- 1 PDU failure (~500-1000 machines suddenly disappear, ~6 hours to come back)
- 1 rack-move (plenty of warning, ~500-1000 machines powered down, ~6 hours)
- 1 network rewiring (rolling ~5% of machines down over 2-day span)
- 20 rack failures (40-80 machines instantly disappear, 1-6 hours to get back)
- 5 racks go wonky (40-80 machines see 50% packetloss)
- 8 network maintenances (4 might cause ~30-minute random connectivity losses)
- 12 router reloads (takes out DNS and external vips for a couple minutes)
- 3 router failures (have to immediately pull traffic for an hour)
- dozens of minor 30-second blips for dns
- 1000 individual machine failures
- thousands of hard drive failures
  slow disks, bad memory, misconfigured machines, flaky machines, etc.

Long distance links: wild dogs, sharks, dead horses, drunken hunters, etc.

Sideline - Large server-clusters

Twitter

- broadcast what's on your mind
- max 140 chars
- 27.3 M tweets per day (November, 2009)
- 250 M tweets per day (October, 2011)
- Twitter moods

Search engine examples

Google, Bing, Yahoo

Search Engine - Application
Overlap between search engines

Compare Google, Yahoo, and Ask Jeeves. Using 10316 queries and hits from first result page.

<table>
<thead>
<tr>
<th>Only in 1</th>
<th>Shared by 2</th>
<th>In all 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>85 %</td>
<td>12 %</td>
<td>3 %</td>
</tr>
</tbody>
</table>

MetaSearch engine Dogpile found 68 % of all results.

DOI: 10.1108/10662240610690034

MetaSearch Engine

- it's software that simultaneously search several individual search engines
- collecting, reviewing and ranking their answers
- and give them back in a merged/condensed form to the user
- they are not better than the quality of the search engine databases they obtain results from

MetaSearch engines

- Simultaneously search several individual search engines
- Query translation
- Result merging
  - Simple merge
  - Duplicate detection
  - Check availability of page
  - tf-idf/similarity ranking
  - Position based
MetaSearch Engine examples

Yippy, Dogpile, DuckDuckGo

Special (Vertical) search engines

- prices
  - ex: prisjakt, PriceRunner, ...
  - http://www.pricerunner.co.uk/
  - http://www.prisjakt.nu/

- jobs
  - ex: freejobsearch, jobspider, ...
  - http://freejobsearch.org/
  - http://www.jobspider.com/

- Housing
  - ex: rightmove, hemnet, bovision, ...
  - http://www.rightmove.co.uk/
  - http://www.hemnet.se/
  - http://bovision.se/

... and so on ...

Other Search Engines

Wolfram Alpha

Wolfram|Alpha introduces a fundamentally new way to get knowledge and answers — not by searching the web, but by doing dynamic computations based on a vast collection of built-in data, algorithms, and methods.
From http://www.wolframalpha.com/about.html
Outline

1. Reiteration
2. Web search
3. Web search engines
4. Web robots, crawler
5. Focused Web crawling
6. Web search vs Browsing
7. Privacy, Filter bubble

Web Robot - Basic architecture

Spider, Crawler, Robot, agent, ...

- Get URL
- Fetch Web page
- Analyze
- Save
- Database
- Frontier
- List of unvisited pages
- URLs
- Repository of visited pages
- Links
- Seed URLs

Web Robot - Types

- Personal Crawlers
- Vertical Search Engines
- General Search Engines
- Archive Crawlers
- Mirroring Systems
- Feed Crawlers

Quality

Freshness

Volume

Figure 12.2: Types of crawlers.


Web Robot - Ethics

- Important - BE NICE
- Do not overload network or server
- Robot exclusion protocol check for http://www.foobar.com/robots.txt
- HTML meta-tag ROBOTS

robots.txt:
User-agent: *
Disallow: /cgi-bin/
Disallow: /DATA/
Disallow: /Images/

<META NAME="ROBOTS" CONTENT="NOINDEX, NOFOLLOW">
Web Robot - Problems

- Network failures
- Erroneous URLs
- Unreachable servers
- Password protection
- Spider traps
- Recursive URLs
- Character set encodings
- Same page - different URLs - deduplication

Web Robot - More Problems

- Hidden Web
- Databases
- Dynamic scripts
- ... ?

Web Robot - Traversal algorithms

- Depth first (Stack, LIFO queue)
- Breadth first (FIFO queue)
- Relevance order (How?)

Outline

- Reiteration
- Web search
- Web search engines
- Web robots, crawler
- Focused Web crawling
- Web search vs Browsing
- Privacy, Filter bubble
Focused Crawling

- Get URL
- Fetch Web page
- Analyze
- Focus filter
- List of unvisited pages
- Save

Focus:
- Domain
- Project
- Country
- Region
- Topic
- Subject

Conditions

- Page is about Carnivorous plants
  - automated subject classification
- There are many pages on the Internet
  - where to start?
  - look only at interesting links
  - take the most important pages first

Topic-specific Web-crawling

- Problem
  - Construct a topic specific search-engine (ex. Carnivorous plants)
- Solution
  - Make a Web-crawler walk through Internet and collect all pages with topic ‘Carnivorous plants’
  - easier said than done!

Automated Classification technologies

- Machine learning methods
  - Statistical models (Bayes, SVM, ...)
  - ANN
- Information Retrieval methods
  - Clustering (no predefined categories)
- Library Science methods
  - String matching + Thesaurus
**Conditions**

- **Page is about Carnivorous plants**
  \[\Rightarrow\] automated subject classification
- **There are many pages on the Internet**
  \[\Rightarrow\] where to start?
  \[\Rightarrow\] look only at interesting links
  \[\Rightarrow\] take the most important pages first

**Basic Algorithm**

Add good start pages (seeds) to frontier

**LOOP:**

- Choose a page among links
- Page OK?
  - Save page
  - Add all links to frontier
- Go to LOOP

**Save (database(s))**:
- All relevant pages (search engine database)
- All analyzed pages (seen pages)
- All new links (frontier)
Focused Crawling

- **Get URL**
- **Fetch Web page**
- **Analyze**
- **Focus filter**
- **List of unvisited pages**
- **Seed URLs**

Problems I
- Which new page?

Problems II
- Isolated pages

Problems III
- Non relevant pages “blocking”
Conditions

- Page is about Carnivorous plants
  - automated subject classification
- There are many pages on the Internet
  - where to start?
  - look only at interesting links
  - take the most important pages first

Compromises

- Precision/recall
- Completeness/speed

Outline

1. Reiteration
2. Web search
3. Web search engines
4. Web robots, crawler
5. Focused Web crawling
6. Web search vs Browsing
7. Privacy, Filter bubble

Browsing

- No idea how formulate a query
- Willing to invest some time
- Structure: flat vs hierarchy
  - Manual vs automatic classification
  - Lack of standard classification/terminology
- Precision - NOT recall
Browsing vs search

- Search
  - LOTS of data
  - Unstructured
  - Unrelated items clutter results
- Browsing
  - Small amounts of data
  - Hierarchically structured
  - Quality assessed

Browsing examples

Dmoz (ODP), Yahoo! Directory

Outline

- Reiteration
- Web search
- Web search engines
- Web robots, crawler
- Focused Web crawling
- Web search vs Browsing
- Privacy, Filter bubble

Filter bubble

- What do search engines or social sites know about me?
- At least location, search history, click history, likes, and more . . .
- Personalize what’s shown (search results, . . . ) using this info
- Show us what we want/like to see - algorithmically
- . . . and not what’s relevant (who decides that?)

Problem?
**Filter bubble example I**

**Google Search for Egypt**  
Scott: Egyptian Protests  
Daniel: Travel Information

From http://www.thefilterbubble.com/what-is-the-internet-hiding-lets-find-out

---

**Filter bubble example II**

**Bing Search for "Climate Change" - International Comparison**  
US: Informational Sites  
EU: Climate Action Sites

From http://www.thefilterbubble.com/what-is-the-internet-hiding-lets-find-out

---

**ToS-DR**

Terms-of-Service – Didn’t Read; http://tos-dr.info/

- you give Google (and those we work with) a worldwide license to use, host, store, reproduce, modify, create derivative works (such as those resulting from translations, adaptations or other changes we make so that your content works better with our Services), communicate, publish, publicly perform, publicly display and distribute such content.

- Facebook: you grant us a non-exclusive, transferable, sub-licensable, royalty-free, worldwide license to use any IP content that you post on or in connection with Facebook (IP License).

---

**Privacy**

- Search history, clicks, photos, documents, comments, …
- leads to a profile
- that can be used by ads or sold, or even stolen
- which might lead to it ending up in unwanted places
- and used against you

Beware!
Be aware!
Infinity i-Kitchen – intelligent fridge runs Linux
http://www.geek.com/articles/chips/this-intelligent-fridge-runs-linux-on-an-arm-chip-20101126/

Read:
http://www.scientificamerican.com/article.cfm?id=long-live-the-web