Content from the Web
Protocols + Crawlers

Class Overview

Other Cool Stuff
- Query processing
- Content Analysis
- Indexing
- Crawling
- Document Layer
- Network Layer

A Closeup View

10/7 - Crawlers
10/12 - IR
10/14 - No class
10/19 - IR, indexing
10/21 - Alta Vista Pagerank
10/26 - No class

Today

• Search Engine Overview
• HTTP
• Crawlers
• Server Architecture

Standard Web Search Engine Architecture

Crawling

Slide adapted from Marti Hearst / UC Berkeley
**Indexing**

- What data is necessary?
- Format?
- Compression?
- Efficient Creation

**Scalability**

- Relational DB

**Query Processing**

- Efficient Processing
- Ranking

**User Interface**

- Spell Checking
- Suggestions
- Faceted Interfaces
- Personalization

**Precision and Recall**

- **Precision**: fraction of retrieved docs that are relevant = \( P(\text{relevant}|\text{retrieved}) \)
- **Recall**: fraction of relevant docs that are retrieved = \( P(\text{retrieved}|\text{relevant}) \)

<table>
<thead>
<tr>
<th></th>
<th>Relevant</th>
<th>Not Relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieved</td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>Not Retrieved</td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>

- Precision \( P = \frac{tp}{tp + fp} \)
- Recall \( R = \frac{tp}{tp + fn} \)
**todo**

- Prec and recall when there are more than two classes

**But Really**

- Precision & Recall are too simple
- Evaluation is a very thorny problem

**Your Project Architecture?**

1. User query
2. Show results to user
3. Extract
4. Classify?
5. DocIds
6. Store documents, check for duplicates, extract links
7. Front end
8. Relational DB

**"Information Extraction"**

As a task: Filling slots in a database from sub-segments of text.

**Traditional, Supervised I.E.**

- Raw Data
- Labeled Training Data
- Learning Algorithm
- Extractor

**Kylin: Self-Supervised Information Extraction from Wikipedia**

From infoboxes to a training set:

- Clearfield County was created in 1804 from parts of Huntingdon and Lycoming Counties but was administered as part of Centre County until 1812.
- Its county seat is Clearfield.
- Clearfield County covers 2,972 km² (1,147 mi²) of it is land and 17 km² (7 mi²) of it (0.56%) is water.
- As of 2005, the population density was 28.2/km².
Architecture

Example Evaluation

Effect of custom lexicons

<table>
<thead>
<tr>
<th>Features</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>.49</td>
</tr>
<tr>
<td>+ Naïve Lexicons</td>
<td>.37</td>
</tr>
<tr>
<td>+ Lexions &amp; Crosstraining</td>
<td>.55</td>
</tr>
</tbody>
</table>

Project Ideas?

Outline

• Search Engine Overview
• HTTP
• Crawlers
• Server Architecture

Connecting on the WWW

What happens when you click?

• Suppose
  ~ You are at www.yahoo.com/index.html
  ~ You click on www.grippy.org/mattmarg/
• Browser uses DNS => IP addr for www.grippy.org
• Opens TCP connection to that address
• Sends HTTP request:

Request

Get /mattmarg/ HTTP/1.0
User-Agent: Mozilla/2.0 (Macintosh; I; PPC)
Accept: text/html; */*
Cookie: name = value
Referer: http://www.yahoo.com/index.html
Host: www.grippy.org
Expires: ...
If-modified-since: ...
HTTP Response

HTTP/1.0 200 Found
Date: Mon, 10 Feb 1997 23:48:22 GMT
Server: Apache/1.1.1 HotWired/1.0
Content-type: text/html
Last-Modified: Tues, 11 Feb 1999 22:45:55 GMT

- One click => several responses
- HTTP1.0: new TCP connection for each elt/page
- HTTP1.1: KeepAlive - several requests/connection

Response Status Lines

- 1xx Informational
- 2xx Success
  - 200 Ok
- 3xx Redirection
  - 302 Moved Temporarily
- 4xx Client Error
  - 404 Not Found
- 5xx Server Error

Logging Web Activity

- Most servers support “common logfile format” or “extended logfile format”
- Apache lets you customize format
- Every HTTP event is recorded
  - Page requested
  - Remote host
  - Browser type
  - Referring page
  - Time of day
- Applications of data-mining logfiles ??

HTTP Methods

- GET
  - Bring back a page
- HEAD
  - Like GET but just return headers
- POST
  - Used to send data to server to be processed (e.g. CGI)
    - Different from GET:
      - A block of data is sent with the request, in the body, usually with extra headers like Content-Type: and Content-Length:
        - Request URL is not a resource to retrieve; it’s a program to handle the data being sent
        - HTTP response is normally program output, not a static file.
  - PUT, DELETE, ...

HTTPS

- Secure connections
- Encryption: SSL/TLS
- Fairly straightforward:
  - Agree on crypto protocol
  - Exchange keys
  - Create a shared key
  - Use shared key to encrypt data
- Certificates

Cookies

- Small piece of info
  - Sent by server as part of response header
  - Stored on disk by browser; returned in request header
  - May have expiration date (deleted from disk)
- Associated with a specific domain & directory
  - Only given to site where originally made
  - Many sites have multiple cookies
  - Some have multiple cookies per page!
- Most Data stored as name=value pairs
- See
  - C:\Program Files\Netscape\Users\default\cookies.txt
  - C:\WINDOWS\Cookies
- Uses??
ToDo

- Add a slide on HTML
- How embedded elements create new HTTP requests
- Maybe put at very beginning?

Web Bugs (eavesdropping)

- A graphic on a Web page (or in email)
  - Allows monitor person reading the content
- Often invisible
  - 1-by-1 pixel in size.
  - Same color as background
- Represented as HTML IMG tags.
- Ubiquitous
  - Aka clear GIF, 1-by-1 GIF, invisible GIF, and beacon GIF

Information sent to server when Web bug is viewed

- The IP address of the fetching
- The URL of the page holding the Web
- The URL of the Web bug image
- The time the Web bug was viewed
- The type of browser
- A previously set cookie value
  - Note: bug can be stored on 3rd party server

Uses of Web bugs

- Ad networks can use Web bugs to add information to a personal profile of what sites a person is visiting.
  - The personal profile is identified by the browser cookie of an ad network.
  - At some later time, this personal profile which is stored in a data base server belonging to the ad network, determines what banner ad one is shown.
  - provide independent accounting of # people e visiting the Web site.
  - gather statistics about Web browser

What kinds of uses does a Web bug have in an Email message?

- A Web bug can be used to find out if a particular Email message has been read by someone and if so, when the message was read.
- A Web bug can provide the IP address of the recipient if the recipient is attempting to remain anonymous.
- Within an organization, A Web bug can give an idea how often a message is being forwarded and read.
**Why are Web bugs used in "junk" Email messages?**

- To measure how many people have viewed the same Email message in a marketing campaign.
- To detect if someone has viewed a junk Email message or not. People who do not view a message are removed from the list for future mailings.
- To synchronize a Web browser cookie to a particular Email address. This trick allows a Web site to know the identity of people who come to the site at a later date.

**What companies have used Web bugs in Email marketing campaigns?**

- Barnes and Noble
- eToys
- Cooking.com
- Microsoft
- InfoBeat

**Response Status Lines**

- 1xx Informational
- 2xx Success
  - 200 Ok
- 3xx Redirection
  - 302 Moved Temporarily
- 4xx Client Error
  - 404 Not Found
- 5xx Server Error

**redirect example**

```
GET / HTTP/1.1
Host: astrology.yahoo.com
```

```
HTTP/1.1 301 Moved Permanently
Location: http://shine.yahoo.com/astrology/
```

**common uses**

1. redirect from blah.com to www.blah.com
2. missing trailing slash
3. tracking internal traffic
4. tracking outbound traffic
5. prettier URLs, preserve old URLs
6. connecting web sites
7. ads
8. authentication

**use 8: authentication**

- cookies are used for authentication
  - cookies can only be set on the page’s domain
- how authenticate someone on domain A if they’re currently on domain B?
  - redirects
- authentication is often on https servers
- how authenticate someone on https if they’re currently on http?
  - redirects

Slide content from www.privacyfoundation.org
Slide by Steve Souders (Google, Stanford)
use 7: ads

- how do you count an ad impression?
  - when a page containing an ad is served?
    count it on the publisher’s backend
  - when a page containing an ad arrives at the client?
    send a beacon from the client
  - when the content of the ad (image, Flash) is requested from the advertiser?
    count it on the advertiser’s backend
  - after the content arrives?
    send a beacon from the client

- redirects can help count when content is served and reconcile the two parties