CS6120: Intelligent Media Systems

Dr. Derek Bridge
School of Computer Science & Information Technology
UCC

User Models

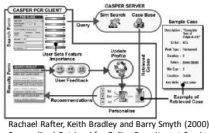
- For personalization, we need *user models* which contain some or all of:
  - Identification data
    - user id, email address, name, IP addresses used,…
  - Demographic data
    - age, sex, country, education, occupation,…
  - Profile data
    - descriptions of goals/interests, lists of liked/disliked items,…

How is User Model Data Obtained?

- Identification/demographic data
  - Fill in a form when you register
  - But note that demographic data can probably be inferred from profile data!
- Profile data
  - Explicit
    - E.g. list your interests when you register
    - E.g. rate or review items
  - Implicit
    - Inferred from your other actions (clicks, searches, etc.)

Where is User Model Data Collected & Stored?

- Client-side
  - User downloads & installs a browser plug-in or an app
  - It records user's actions, infers preferences, and uses inferred preferences to modify its output
- E.g. on the client, CASPER re-ranked search results using profile data that was stored only on the client



Rachael Rafter, Keith Bradley and Barry Smyth (2000)
Personalised Retrieval for Online Recruitment Services.
Proceedings of the 22nd Annual Colloquium on Information Retrieval

- More often, the client-side software does no more than share the data with the server at intervals
  - E.g. scrobbling by last.fm and Soundwave

Where is User Model Data Collected & Stored?

- Server-side
  - the server records the user's requests (clicks, searches,…)
- Server-side is still probably more common but:
  - a narrower range of actions can be observed (e.g. can't observe bookmarking, scrolling, requests satisfied by browser cache)
  - data is less accurate (e.g. have to factor our transmission times)
  - there is the problem of identifying the user

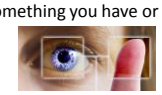
User Identification & Authentication

Identification

- establishing the identity of the user ("which user is this?")

Authentication

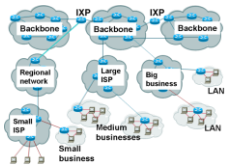
- verifying that the user is who s/he claims to be
- by means of credentials (e.g. something you have or know)



- Surely it's just a matter of logging-in!
  - identification: user id
  - authentication: password
- On the web, this does not solve the identification problem…

The Internet

- A global network of networks that inter-operate using the TCP/IP stack of communications protocols



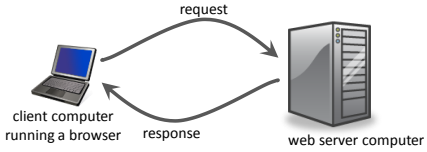
- (Simplification) Each device connected to the internet has a unique address: IP address

The Web

- Internet applications (uses), e.g.:
  - the web
  - email
  - internet telephony
- The web = hypertext + the internet
  - invented by Tim Berners-Lee in 1989/1990
  - he specified core technologies (HTML, URL, HTTP)
  - he built first web browser and web server

The Web

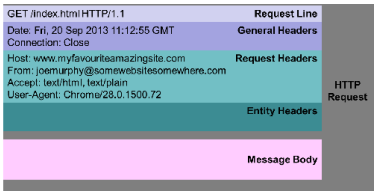
- The web has a client-server architecture
  - clients make requests
  - servers send responses



- Requests and responses use the HTTP

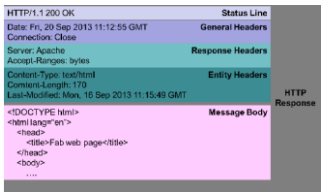
HTTP Request

- Suppose you enter `http://www.myfavouriteamazingsite.com/index.html` into the Location box of your browser or click on a link or submit a form
- Your browser creates and sends an HTTP request, e.g.:



HTTP Response

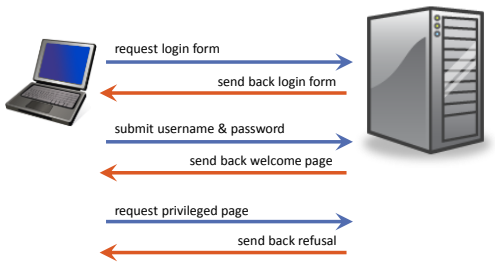
- The server receives the request and sends back an HTTP response, e.g.



HTTP is Stateless

- HTTP is a *stateless protocol*:
  - each request is independent: by default, the server has no memory of previous requests
- Adequate for HTTP's original purpose:
  - a client contacts a server and requests a document
  - the server sends the requested document to the client
- Inadequate for situations where it can be useful to recognize repeat contacts, e.g.
  - a client which has contacted the server in the past
  - a sequence of requests from the same client within a short period of time (a *session*)

Logging-in (done wrong)



User Identification on the Web

- IP addresses
  - keep a record of client's IP addresses
- Cookies
  - Send the client a user identifier which it sends back to you

- Other methods
  - URL rewriting
    - include a user identifier in the query part of a URL
  - Hidden fields
    - include a non-visible field in forms so that a user identifier gets submitted with the form data

User Identification Criteria

- Can the technique identify the user correctly within-sessions or between-sessions or both?
- Can it identify the user correctly if more than one user uses the same machine?
- Can it identify the user correctly if s/he uses a different machine? Or a different browser on the same machine?
- What (else) can go wrong?

IP Addresses for User Identification

- None of the methods is perfect!
- What are the problems of trying to identify users using IP addresses?

Cookies for User Identification

- A cookie is a small amount of data (a name/value pair)
  - E.g. `id=cust123`
  - Each cookie can be no more than 4kb in size
- If a browser has sent a request to a server, the server can include a cookie in its response (in a header line)
- If the browser has cookies enabled, it stores the cookie
- Next time the browser sends a request to the same server (or one in the same domain), it includes the cookie in its request (a header line)
- This enables the server to know that it has previously received requests from this client

Cookies Example, Part I

- Your browser sends a request to `www.amazon.co.uk`

```
GET /index.html HTTP/1.1
...
```

- Software on the server stores information, e.g. in its database, about your visit
- The server's response includes a cookie

```
HTTP/1.1 200 OK
Set-Cookie: id=cust123; path=/; domain=.amazon.co.uk
...
```

- If cookies are enabled in your browser, your browser stores the cookie

Cookies Example, Part II

- On a subsequent occasion, you visit `www.amazon.co.uk` again
- Your browser includes the cookie in the request

```
GET /index.html HTTP/1.1
Cookie: id=cust123
...
```

- Hence, software on the server knows that you made requests on previous occasions and can use the cookie data, e.g. to look you up in its database

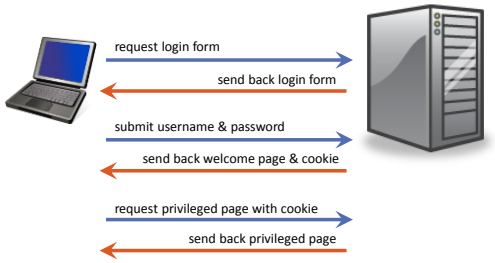
Cookies for User Identification

- What are the problems of trying to identify users using cookies?

Logging-in for User Identification

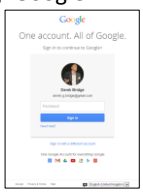
- Logging-in gives us
  - identification
  - authentication
- But note that logging-in is not an alternative to cookies
  - it typically relies on cookies
  - hence it has the same problems

Logging-in (done right)



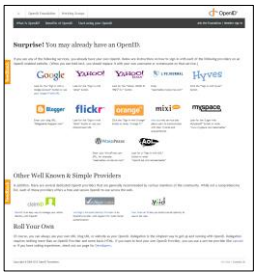
Log-in, please

- Sites want you to register and log-in
  - makes identification easier
  - thus can build more reliable user models
- E.g. Google



OpenID

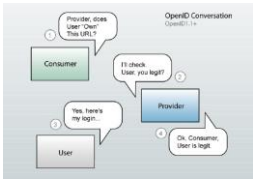
- A solution to the problems of
  - too many usernames
  - too many passwords
  - username already taken
- Register a single identity with an Identity Provider



OpenID

OpenID

- To log-in to an OpenID-enabled site:
  - enter your OpenID into the site's login form
  - you are redirected to your Identity Provider
  - enter your OpenID username and password
  - confirm that the original site can use your identity
  - you are redirected back to the original site

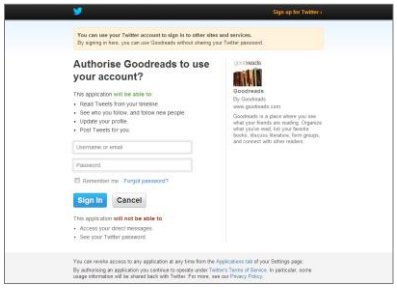


http://www.slideshare.net/steveivy/openid-oauth-an-introduction

OAuth

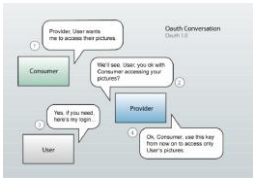
- OAuth is an open standard for authorization
  - not authentication
- A solution to the problem of granting a third-party access to your resources without sharing your password
- Like a valet key but for web services

OAuth



OAuth

- E.g. allowing Goodreads to post tweets
  - Goodreads requests a token from Twitter
  - Twitter asks user for authorization (and log-in)
  - Twitter sends a token to Goodreads
  - Subsequently Goodreads includes the token in requests it makes to Twitter
- Less simplified treatment: http://blog.varonis.com/introduction-to-oauth/



http://www.slideshare.net/steveivy/openid-oauth-an-introduction

Third-Party Cookies

- Suppose adpub.com is a web advertising company
- Suppose news.com's web page contains an ad

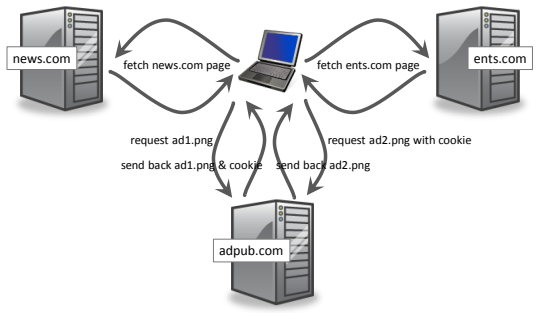
```
<!DOCTYPE html>
<html>
  …
  <img src="http://adpub.com/ad1.png" />
  …
</html>
```

- Also ents.com's web page contains an ad

```
<!DOCTYPE html>
<html>
  …
  <img src="http://adpub.com/ad2.png" />
  …
</html>
```

- Consider a user who first visits news.com's page and then ents.com's page

Third-Party Cookies



Third-Party Cookies

- They are regular cookies
  - but they come from a third-party
- Advertising companies use them to track users across the web
  - e.g. doubleclick, BlueKai, exelate
  - e.g. can serve you tailored ads
- Most browsers allow you to block them

SOME THOUGHTS ON PRIVACY

Privacy

By Peter Steiner , *The New Yorker* , July 5, 1993

"On the Internet, nobody knows you're a dog."

Privacy

"How the hell does Facebook know that I'm a dog."

On-line Privacy

- The ability to control which of the data one reveals about oneself on the web is…
  - …collected & stored
  - …used, e.g., for personalization
  - …shared with designated others (e.g. friends)
  - …shared publically
  - …shared with third-parties (other companies, government agencies)

There's no such thing as a free lunch

- "If you're not paying, you're the product"
- A deal…with the devil?

  - you give your data; you surrender your privacy
  - in exchange, you get a free service, possibly an adaptive one

Preventing Tracking

- You could avoid sites that use cookies
  - EU sites require sites to obtain your consent to cookies
- You could disable cookies
  - but this may make some sites unusable
- You could 'go incognito' (private browsing)
  - disables browser cache & history; deletes cookies
- You could disable third-party cookies
- You could use your browser's DoNotTrack facility
- You can opt-out of individual advertising companies (e.g. BlueKai)

Privacy Policies

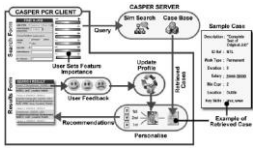
- Enhance trust
  - concise, clear policies
  - transparency over changes
- Make profile data
  - scrutable
  - editable

Pseudonyms

- Hide true identity
- But
  - identity might be guessable, e.g.:
    - from IP address
    - by matching profiles across services
  - makes bullying, stalking, etc. easier
- For legal or policy reasons, some services insist that users register and even use a true identity
  - e.g. early days of Google+

Client-Side User Models

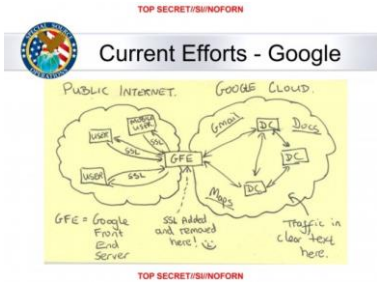
- If more systems were purely client-side, individuals would have more control



Rachael Rafter, Keith Bradley and Barry Smyth (2000)
Personalised Retrieval for Online Recruitment Services.
Proceedings of the 22nd Annual Colloquium on Information Retrieval

Surveillance

Surveillance



Sousveillance

- More people are wearing cameras & other sensors
  - e.g. for lifelogging
- Hence, our behaviour is being captured by those around us
  - without consent