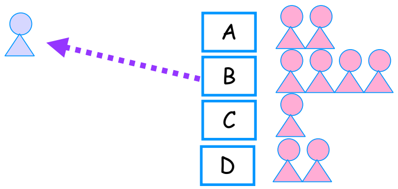
Lecture 2:
Personalization Infrastructure

Derek Bridge and Adrian O'Riordan

A classification

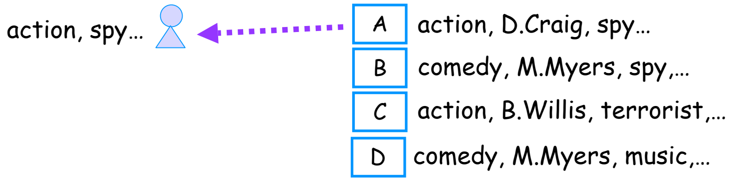
- Types of, e.g., movie recommender:
  - Non-personalized
  - Personalized
    - Content-based
    - Collaborative
    - Demographic
    - Hybrid
- Personalized systems must build *user models*
  - information about the user from which the system can infer his/her goals, interests, etc.

Non-personalized



- Items
  - store *popularity* data
  - e.g. total sales; total visits; average rating
- Users
  - nothing
- Recommend
  - most popular item

Content-based - first variant



action, D.Craig, spy…

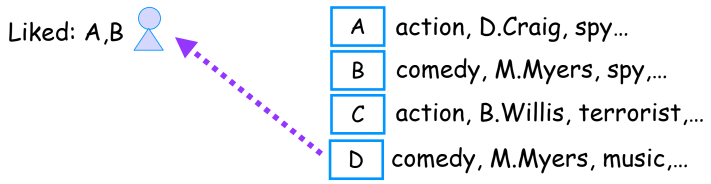
comedy, M.Myers, spy,…

action, B.Willis, terrorist,…

comedy, M.Myers, music,…

- Items
  - store item *descriptions*
  - e.g. bag of keywords, or something more structured
- Users
  - store a *profile* of short- or long-term interests
  - similarly, either keywords, or something more structured
- Recommend
  - item whose description is most *similar* to user's profile

Content-based - second variant



Liked: A,B

A action, D.Craig, spy…

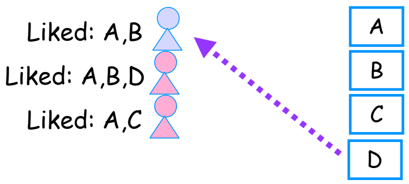
B comedy, M.Myers, spy,…

C action, B.Willis, terrorist,…

D comedy, M.Myers, music,…

- Items
  - store item *descriptions*
  - e.g. bag of keywords, or something more structured
- Users
  - store a *profile* of items the user has purchased/ clicked/ read/ rated highly
- Recommend
  - item whose description is most *similar* to one of the items in the user's profile

Collaborative (user-based)



Liked: A,B

Liked: A,B,D

Liked: A,C

A

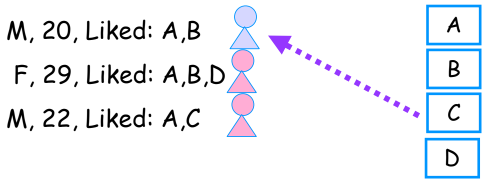
B

C

D

- Items
  - nothing
- Users
  - store a *profile* of items the user has purchased/ clicked/ read/ rated highly
- Recommend
  - item liked by a user who has a *similar* profile

Demographic



M, 20, Liked: A,B

F, 29, Liked: A,B,D

M, 22, Liked: A,C

A

B

C

D

- Items
  - nothing
- Users
  - store a *description* of the user and a *profile* of items the user has purchased/ clicked/ read/ rated highly
- Recommend
  - item liked by a user who has a *similar* description

Summary of system types

- Item descriptions
  - some systems require extensive item descriptions; others don't
- User models
  - all personalized systems require user models
- Matchmaking
  - all personalized systems require some form of similarity

User models

- Identificatory
  - user id, email address, name, IP address
- Demographic description
  - age, sex, country, education level, occupation
- Profile
  - long-term or short-term
  - description or descriptions of goals/interests or lists of items liked/disliked

User models

- How is the data collected
  - Explicit
    - e.g. user fills in a form
    - good for demographic info (but users lie!)
    - can be used for short- and long-term profiles
  - Implicit
    - observe the user's activities (browsing, searching, purchasing,…)
    - can be good for long-term profiles

User models

- Where is the data collected & stored
  - Client-side
    - e.g. user downloads and installs a plugin for his/her browser, and it records user's actions
  - Server-side
    - the server records the user's requests
- Issues
  - server-side, a narrower range of actions can be observed (e.g. can't observe bookmarking in browser, e.g. won't receive requests that were handled by a cache)
  - server-side, data is less accurate (e.g. have to factor out transmission times)
  - server-side, have the problem of identifying the user (who does this request come from)
  - but client-side requires user to install software
  - hence, server-side is far more common

Some News Stories on Identity Theft and On-line Privacy

How worried should people be? - BBC News 2007

"The loss, by HM Revenue and Customs, of computer discs with details of about 25m child benefit recipients, is the biggest loss yet of personal data in the UK. Now the Chancellor, Alistair Darling, is trying to reassure millions of families whose details have been lost."

AT&T and Other ISPs May Be Getting Ready to Filter - New York Times, 2008

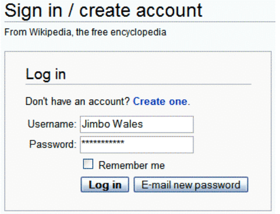
"Network-level filtering means your Internet service provider.. could soon start sniffing your digital packets, looking for material that infringes on someone's copyright."

User Identification and Authentication (I&A)

- User Identification – the established relationship between a User and a particular Registration, e.g. on-line it can be username, logon-id
  - Digital identifier usually string or token that is unique within a given scope
- Authentication - the act of establishing or confirming someone as authentic, that is, that they are who they claim to be, e.g. password, certification from a trusted third party
  - Usually something you know, something you have or something you are
- One familiar use of authentication is in access control, e.g. restricting access to on-line bank account details

Technology Options: Log ins

- Process by which individual access to a computer system
- Authenticate the identity of any computer user
- Log in (sign in) / Log out to website
- Need to create an account (get a user id and password)
- To *log out* (also to *log off*, *sign out*, or *sign off*) is to close off one's access



Sign in / create account

From Wikipedia, the free encyclopedia

Log in

Don't have an account? Create one.

Username: Jimbo Wales

Password: ***********

□ Remember me

User Identification

- IP addresses – use unique address devices use to identify and communicate with each other to try to identify people
- Cookies Tracking
  - (HTTP) Cookies – tracks user on-line within and between sessions
  - can be used for personalisation
  - Session ids – id for duration of browsing session
- Other methods – URL query string; hidden form fields

Criteria

- Can it 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?
- How much effort does it require of the user?

IP addresses

- Can be assigned to any device, e.g. PC, printer, router
- managed and created by the Internet Assigned Numbers Authority (IANA)
- IPv4 has 32 bits; IPv6 has 128 bits
- IP address can be expressed on a dotted-decimal notation where a period separates each 8 bits. Example: IP address 143.239.1.2 has the corresponding domain name student.cs.ucc.ie (task: use nslookup to query a domain name)
- Static or Dynamic IP: DHCP (Dynamic Host Configuration Protocol) centralises IP address management by automatically assigning IP addresses to clients when they log in

IP Addresses for I&A

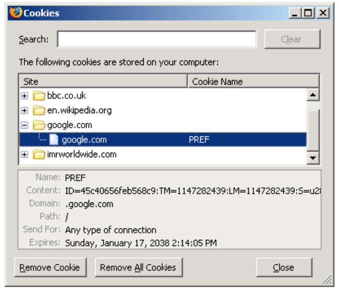
- Identifies user within-session and between-sessions provided static IP address assignment is used
- Only one user can use each machine
- No (legal) certification
- The user cannot use a variety of machines
- May only see proxy
- But no effort involved

HTTP Cookies

- Small text file sent by a server to a web browser and then sent back by the browser each time it accesses that server
- Developed at Netscape Communications in 1994
- Internet spec RFC 2109 in 1997
- HTTP cookies are used by Web servers to differentiate users and to maintain data related to the user during navigation, possibly across multiple visits
- Used to maintain site preferences, record contents of shopping cart, for personalisation
- Cookies can only be sent to the server setting them or one in the same domain

Cookies – The Details

- Limit on size and number of cookies -internet browser is expected to be able to store at least 300 cookies of 4 kilobytes each, and at least 20 cookies per server or domain.
  - Actually Firefox 2.0 and IE 7 will allow 50 per domain
- Contain name/value pair, and optionally expiration date, a path, a domain name, and whether the cookie is intended only for encrypted connections.



The following cookies are stored on your computer:

Site	Cookie Name
bbc.co.uk	
en.wikipedia.org	
google.com	
google.com	PREF
imrworldwide.com	

Name: PREF
Content: ID=45c40656feb568c9:TM=1147282439:LM=1147282439:S=u2
Domain: .google.com
Path: /
Send For: Any type of connection
Expires: Sunday, January 17, 2038 2:14:05 PM

Session Identification

- an HTTP exchange between a browser and a remote host may include an HTTP cookie which identifies state, such as a unique session ID
- A session ID is a unique number that a Web site's server assigns a specific user for the duration of that user's visit.
- Can be stored as cookie or Session Identification URI

  SID:*type*:*realm*:*identifier[-thread][:count]*

  SID:ANON:www.w3.org:j6oAOxCWZh/CD723LGeXlf-01:34
- Server can use simple method like incrementing a number or an algorithm that factors in other info such as date/time.
- Not secure – can be stolen (session hijacking)

Cookies for I&A

- cookies can expire or deleted; user can refuse cookies
- Identifies user within-session and between-sessions
- More than one user can use the same machine provided they each have a different local user id
- The user cannot use a variety of machines (or browsers)
- Cookies can be hijacked or "poisoned": using https overcomes this
- Low effort: simple, well established

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

On-line Privacy

- The ability to control what information one reveals about oneself over the Internet
- Prevent personal details being shared with third parties
- Levels:
  - Total Anonymity hard to achieve
  - "Pseudonymity" - state of disguised identity resulting from the use of a pseudonym
  - Use acceptable levels for convenience
- Marketers' dream or infringement of civil liberties?
- Concerns: Unsolicited marketing, usage logging

Privacy concerns: example study

Concerns are evident from studies (Kobsa's summary in 2007 of various studies form late 1990s to early 2000s)

- More than a quarter said they would never provide personal info
- Quite a few indicate having supplied false information
- 80-90% are concerned if a business shares their info for a purpose different from the original purpose
- Many are concerned about being tracked and some disable cookies or delete them regularly
- More willing to supply basic demographic info (age, etc) than hobbies than online behaviour than identification than esp. financial identification; but for any piece of info values outside the norm are less likely to be disclosed
- Older, more educated and wealthier have more concern. No clear gender effects
- But many do not follow through on their stated concerns, e.g. people stating a concern but buying online

Allaying concerns

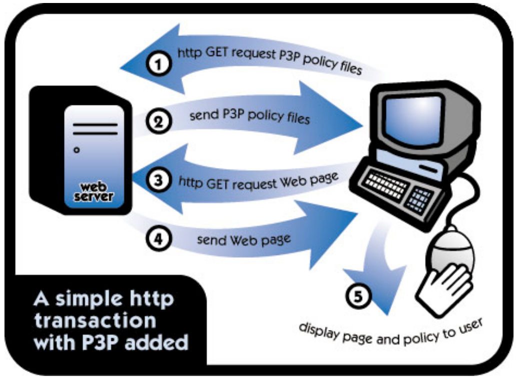
- Users more likely to supply data if they can see the benefit in doing so usually in terms of personalization
    - but maybe in terms of prestige, social interaction, reward
- they know how it will be used and shared (*scrutability*);
    - they trust the service provider, based on the quality of the website, the reputation of the provider, the presence of a privacy policy

Privacy-enhancing technology

- Pseudonyms- Use of unique and persistent identifiers that hides true identity; but…
    - identity may be guessable from pseudonym and, e.g., IP address, proxy server
    - identity may be guessable from the descriptive parts of a profile (e.g. keywords, ratings)
    - for the above reasons, or for other legal or policy reasons, users may have to register their pseudonym and true identity with a registrar
- Client-side personalization
    - Data is collected, stored and processed client-side

Personalizing privacy

- Privacy concerns differ between users so need personalization of privacy
- One step might be the Platform for Privacy Preferences (P3P)


1 http GET request P3P policy files
2 send P3P policy files
3 http GET request Web page
4 send Web page
5 display page and policy to user
web server
A simple http transaction with P3P added

Case study: Google

- Google's Privacy Policy
  - Highlights: http://www.google.com/privacy.html
  - In full: http://www.google.com/privacypolicy.html
  - FAQ: http://www.google.com/privacy_faq.html
- April 2007: Google agrees to buy online banner ad company DoubleClick for $1.3billion
  - Google and DoubleClick both run an ad service business. Google operates AdSense to dish up ads to Web sites in its publisher network, whereas DoubleClick offers an ad serving and ad management tool called Dart, geared toward publishers, advertisers, and corporate customers. The companies differ in that Google's pay-per-click text ads are generated from keyword searches, while DoubleClick puts banner ads on sites.