Information filtering

Ian Ruthven
ir@cis.strath.ac.uk

52.475 MIA Lecture 6

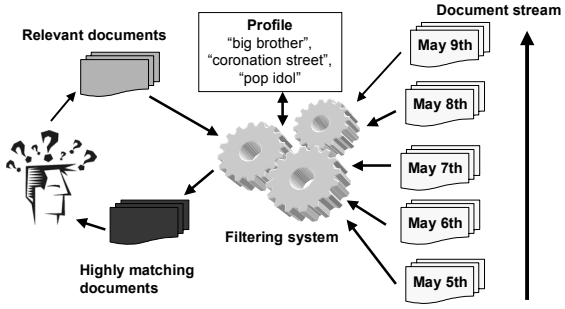
Introduction

- Last time
  - Relevance feedback
- This time
  - Different types of filtering application
    - Information filtering
    - Collaborative filtering
    - Use-based filtering

Information filtering

- Information retrieval/access
  - Static collection
    - Content doesn't change (much)
  - Queries
    - Short
    - Short-lived
- Information filtering
  - Dynamic collection
    - e.g. newswire, radio, television
  - Profiles/filters
    - Big
    - Long-lasting

Information filtering



Information filtering

- Traditional IR
  - Importance of *ranking*
    - "what is best order of documents?"
- Information filtering
  - Importance of *selection*
    - "which documents to show to user?"
    - Binary decision – show/don't show
      - Show too many – user gets swamped
      - Show too few – user misses relevant information
      - Importance of *threshold*

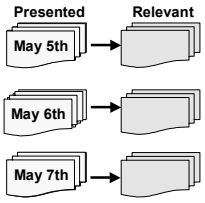
Threshold

Document	Retrieval score
D1	0.77
D2	0.65
D3	0.54
D4	0.42
…	…

Threshold of
0.8 will show no documents to user
0.6 will show D1, D2 to user
0.4 will show D1-D4 to user

Thresholding

- One method:
  - Start with guess
  - Then modify threshold based on previous thresholds
  - % of relevant presented
    - Low % increase
    - High % decrease
  - Different thresholds for different topics

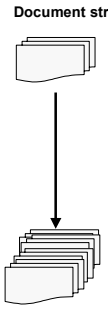


Filtering decisions

- Save previous documents or not?
- If yes
  - massive storage requirements
    - Store documents themselves
  - and processing requirements
    - Store and update indexes
  - but can re-run profiles
- If no
  - weighting problems
    - for ranking
      - » we only index small batches for ranking
    - for profile updating
      - » poor weight estimation (no *whole* collection)

Store

- If we store previous documents…
  - Index and rank new documents for presentation
  - Show some to user
  - Merge new documents and indices
  - Once user gives relevance info
    - Use RF to update profile
      - Add new concepts
      - Change weights
  - Possibly time-limited
    - difficult



Don't store

- If we don't store previous documents
  - Index and rank new documents for presentation
  - Show to user
  - ~~Merge new documents~~
  - Once user gives relevance info
    - Use RF to update profile
      - » Add new concepts
      - » Change weights
    - But only based on new documents
  - Very little information
    - Can store concept table
    - Can store part of document set

Term	*n*	*r*
a	2	1
am	1	0
be	1	0
cat	2	0
did	2	0
hunting	1	0
i	1	0
i	2	1
puddy	2	1
…	…	

Filtering decisions

- How much input does user have/need to have?
  - Lots? – e.g. reads and assesses every doc
    - Profile is very quick to change
      - But perhaps too quick
  - Little? – reads every 50/100/150, every week, …
    - Profile is slow to change
      - Can miss important changes
  - Usually dependent on application
    - How timely information is, how much information generated, etc

Filtering decisions

- Who uses filter?
  - 1 person
    - Very personalised
  - Group of people
    - Clusters of interests
    - Difficult to set thresholds
    - Docs usually ordered by date

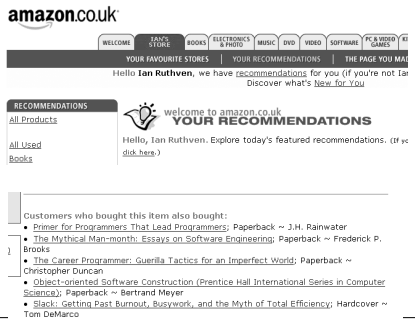
Profile
"big brother",
"coronation street",
"pop idol"

"football focus"
"match of the day"
"grandstand

Information filtering/collaborative filtering

- Information filtering
  - Based on similarity of *topic*
    - "*If you like these document on Eastenders, you will like any document on Eastenders*"
      - If lots of relevant documents contain the concept Eastenders, add the concept Eastenders to the profile
- Collaborative filtering
  - Based on similarity of people's *taste*
    - "*If Fabio likes Scooby Doo, I will like Scooby Doo*"

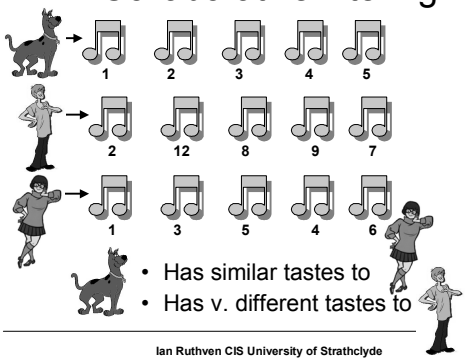
Collaborative filtering



Collaborative filtering

- Collaborative filtering based on user profiles tracking
  - What books/cds/dvds/etc people have bought
  - What web pages they have visited
  - What food they buy (supermarkets)
- Aim: to recommend items to people who share similar tastes
  - Also called *recommender* systems

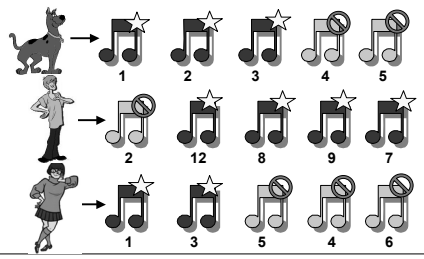
Collaborative filtering



- Has similar tastes to
- Has v. different tastes to

Ratings

- Can also include *ratings*
  - Explicit vs implicit



Collaborative filtering

- Similar to other filtering/retrieval applications
  - Profile = previous purchases
  - Matching = profile to profile
- But usually not content-based matching
  - Identifiers
    - Profile = relevant document identifiers *not* description of relevant documents
  - No indexing

Two problems

- size of data
  - Based on most similar person
    - Need large group
    - Need relatively large number of ratings/purchases
- 'cold-start' problem
  - New items to database
    - Cannot recommend without previous use/purchase
    - Cannot purchase without recommendation
  - New people
  - Usually need some hybrid system
    - Querying/browsing for new items, recommendation for others

Use-based filtering

- Information filtering
  - *"If you like these document on Eastenders, you will like any document on Eastenders*
  - Based on *topic*
- Collaborative filtering
  - "*If Fabio likes Scooby Doo, I will like Scooby Doo*"
  - Based on *taste*
- Use-based filtering
  - "*If I spend ages reading about Elmer Fudd, then I like information about Elmer Fudd*"
  - Based on *behaviour*

Use-based filtering

- Systems filter behaviour
  - Usually web-based
  - Examine your actions
    - Bookmarking, saving, printing, etc (explicit)
    - Time taken to read, scrolling, etc (implicit)
  - Develops a profile from 'interesting' pages
  - Use profile to retrieve new pages
    - Long-term – based on history
    - Interactive – whilst using web (session-based)

Use-based filtering

- Similar to RF
  - Indexed representations of documents
  - Relevance assessments = behaviour
  - Modified query = modified 'interest' profile
- Problems
  - Profiles can be very messy
  - Interaction style is very variable
    - Even within individuals
  - Can be intrusive

	IR	IF	CF	UBF
Collection	Fixed (usually)	Dynamic (quickly, replacement)	Dynamic (slowly, usually extended)	Dynamic (slowly, usually extended)
Search statement	Query (short-lived, representation)	Profile (long-lived, representation)	Profile (long-lived, object ids)	Profile (long-lived, representation)
Information need change	Quick	Slow	Slow	Quick?
Matching	Content	Content	Id's	Content
Output	Ranked list of items	Items above threshold	Unseen similar items	Unseen similar items

Summary

- All types of filtering use similar concepts to IR
  - Matching, relevance, relevance feedback
  - All have a number of application areas
    - Information filtering
      - Personalised information services
      - Intelligence services
    - Collaborative filtering
      - E-commerce
      - Complex searches (jobs, homes, online dating)
    - Use-based filtering
      - Alternative to search engines
      - Advertising