CS6120: Intelligent Media Systems

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

Today's Lecture

- Introduction to recommender systems
- Non-personalized recommender systems
  - Expert and user reviews
  - Aggregated opinion
    - sales, ratings,...
  - Market basket analysis
  - Business rules
- Ratings

INTRODUCTION TO RECOMMENDER SYSTEMS

Recommendations

- Recommendations help us to decide which goods, services or information to purchase or consume
- Sources of recommendations
  - salespeople, critics, guides, acquaintances,…
  - recommender systems
- Generically, the thing recommended is called an “item”

“Items”

- Physical products
  - books, phones
- Non-physical products
  - movies, music, ebooks, ringtones
- Services
  - a hotel to stay in, a restaurant, a school or university
- People
  - someone to date, a person to ‘friend’ or ‘follow’, an expert (e.g. a plumber, a dentist)
- Sources of information
  - news stories, web pages, a blog to read, recipes, lessons, tutorials
- Events, actions and activities
  - a museum to visit, a concert to go to, a job to apply for, an exercise regime to follow
- …and many more!

The way items are consumed affects the way we recommend them

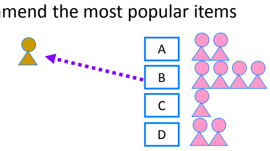
- The unit of recommendation
  - individual items, packages, sequences (e.g. playlists)
- The target consumer
  - individual users, small groups (e.g. families, housemates), larger groups (occupants of a shared space, communities)
- Level of interaction
  - passive, confirmation (e.g. skipping a song), selection from a list
- The nature of the item
  - high-value versus low-value
  - high consumption cost versus low consumption cost
  - rivalrous versus non-rival
  - perishable versus non-perishable
  - one-off consumption versus repeated consumption
- …and so on!

Recommender systems

- Non-personalized
- Demographic
- Personalized
  - content-based
  - collaborative
    - user-based
    - item-based
  - knowledge-based
  - social
- Hybrids and ensembles

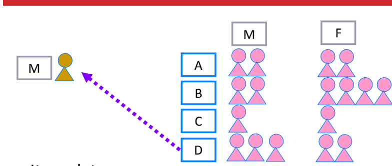
Non-personalized

- E.g. recommend the most popular items



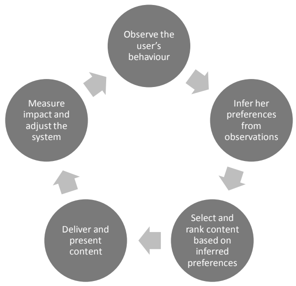
- Item data
  - store popularity data or other summary statistics
  - e.g. total sales, total visits, average rating
- User model
  - none
- Recommend
  - most popular item(s)

Demographic

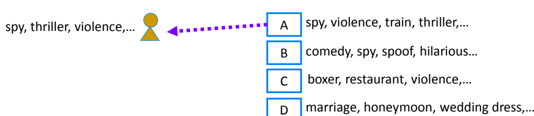


- Item data
  - store popularity data per demographic niche
- User model
  - demographic data
- Recommend
  - most popular item(s) for the niche to which the user belongs

Personalized

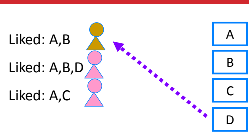


Content-based



- Item data
  - store descriptions, e.g. set of keywords
- User model
  - interests, also as set of keywords
- Recommend
  - item(s) whose descriptions are most similar to user's interests

Collaborative (user-based)



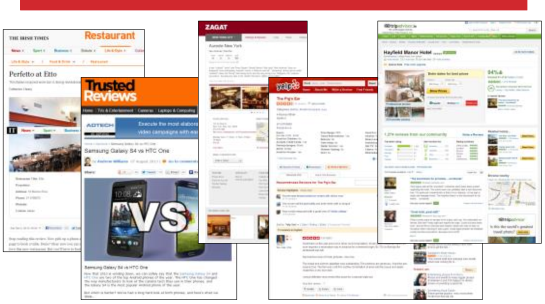
- Item data/User model
  - set of items the user has purchased/ clicked/ read/ rated highly
- Recommend
  - item(s) liked by a user who has a similar profile

Summary of the overview

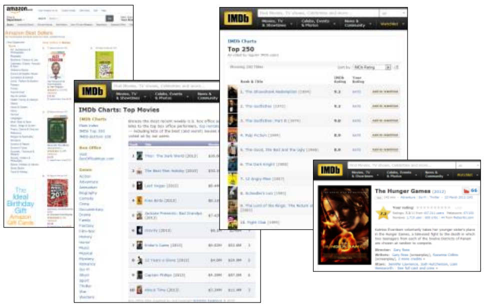
- Item descriptions
  - some systems require extensive item descriptions; others don't
- User models
  - all personalized systems require user models, but they differ in what they contain
- Matchmaking
  - 'all' personalized systems require a way to measure similarity

NON-PERSONALIZED RECOMMENDER SYSTEMS

Expert and User Reviews & Ratings



Popularity



The Problems with Averages, Part 1

- Both of these have the same average
  - 3, 3, 3, 3, 3, 3, 3, 3, 3, 3
  - 1, 1, 1, 1, 1, 5, 5, 5, 5, 5
- The first of these has a higher average
  - 5
  - 1, 5, 5, 5, 5, 5, 5, 5, 5, 5

http://xkcd.com/1098/

Ratings Distributions

- It can be important to show the distribution
  - the range of the opinions and their frequencies
  - e.g. number of ratings
  - e.g. summary percentages
  - e.g. histogram

The Problem with Averages, Part 2

Over time, the average ratings of excellent items fall

↓

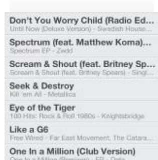
Why?

↑

Over time, the average ratings of mediocre items rise

Market Basket Analysis

- We have items
  - e.g. products
- And we have sets of items
  - sometimes called baskets
  - sometimes called transactions
- A *basket* is the set of items a user bought/consumed
  - on a particular occasion, or
  - over a period of time



Frequent Itemsets

- Informally, a set of items that appears in many baskets is said to be frequent
- Formally,
  - Let I be a set of items
  - The *support* for I, $support(I)$, is the number of baskets for which I is a subset
  - I is *frequent* if $support(I)$ is greater than or equal to s, the *support threshold*

Baskets and Itemsets: Examples

Baskets

- {bread, milk, bananas, butter}
- {marmite, tea, sugar, juice, bread, peppers, onions, bananas, milk, oranges, coffee, biscuits}
- {bread, wine, olives, cheese, juice, grapes, bananas, tomatoes}
- {tissues, cake, crumpets, margarine, bananas, eggs, nuts}
- {bread, milk, chicken, eggs, potatoes}
- {bananas, apples, bread, yoghurt, eggs, rice, oil, pasta, cheese}
- {bananas, milk, bread, cheese, juice, tuna, beans, leeks, mushrooms, carrots, tomatoes}
- {onions, tomatoes, lentils, crisps, bananas, milk, cheese}

Some itemsets and their support

itemset, I	$support(I)$	
{bread}	6	✓
{onions}	2	
{bread, milk}	4	✓
{bread, margarine}	0	
{bananas, onions}	1	
{bananas, bread, milk}	3	✓
{wine, olives, cheese}	1	

- If the support threshold, s, is 3, which of the itemsets above is frequent?

Finding Frequent Itemsets

Problem

- There are so many candidate itemsets
  - all singletons
  - all pairs
  - all triples
  - all sets of four,...
- With n items, there are 2^n candidate itemsets
  - e.g. if $n = 38$, then 274,877,906,944 candidates
  - e.g. if $n = 100{,}000$, then the number of candidates has 30,103 digits

Solutions

- Most solutions exploit the following to massively prune the candidates:
  - if I is not frequent, then no superset of I can be frequent
  - e.g. {onions} is not a frequent itemset, so {bananas, onions} cannot be a frequent itemset
- E.g. the A-Priori algorithm

Association Rules

Converting a Frequent Itemset to Association Rules

- In each rule, one of the items is placed on the right-hand side
- E.g. {bananas, bread, milk} becomes
  - if {bananas, bread} then milk
  - if {bananas, milk} then bread
  - if {bread, milk} then bananas

Confidence of an Association Rule

- if I then j:
  - if all the items I appear in a basket, then item j is likely to appear in that basket as well
- How likely?
  - the confidence of a rule

$$confidence(\text{if } I \text{ then } j) = \frac{support(I \cup \{j\})}{support(I)}$$

Confidence of an Association Rule

Baskets

- {bread, milk, bananas, butter}
- {marmite, tea, sugar, juice, bread, peppers, onions, bananas, milk, oranges, coffee, biscuits}
- {bread, wine, olives, cheese, juice, grapes, bananas, tomatoes}
- {tissues, cake, crumpets, margarine, bananas, eggs, nuts}
- {bread, milk, chicken, eggs, potatoes}
- {bananas, apples, bread, yoghurt, eggs, rice, oil, pasta, cheese}
- {bananas, milk, bread, cheese, juice, tuna, beans, leeks, mushrooms, carrots, tomatoes}
- {onions, tomatoes, lentils, crisps, bananas, milk, cheese}

Confidence examples

$$confidence(\text{if } \{bread\} \text{ then } eggs)$$

$$= \frac{support(\{bread, eggs\})}{support(\{bread\})}$$

$$confidence(\text{if } \{bread, milk\} \text{ then } bananas)$$

$$= \frac{support(\{bread, milk, bananas\})}{support(\{bread, milk\})}$$

Confidence and Interest

- Confidence is not enough
  - e.g. *confidence(*if *{tomatoes}* then *bananas*) is high ($^3/_3$)
  - but that's because bananas are so common (7 out of 8 baskets)
- We want the idea that I affects *j*
  - the *interest* of a rule

$$interest(\text{if } I \text{ then } j) = confidence(\text{if } I \text{ then } j) - \frac{support(\{j\})}{number\ of\ baskets}$$

  - if interest is zero (or close to zero), then I doesn't affect *j*
  - if greater than zero, then I in some sense causes *j*
  - if less than zero, then I in some sense discourages *j*

Interest of an Association Rule

Baskets

- {bread, milk, bananas, butter}
- {marmite, tea, sugar, juice, bread, peppers, onions, bananas, milk, oranges, coffee, biscuits}
- {bread, wine, olives, cheese, juice, grapes, bananas, tomatoes}
- {tissues, cake, crumpets, margarine, bananas, eggs, nuts}
- {bread, milk, chicken, eggs, potatoes}
- {bananas, apples, bread, yoghurt, eggs, rice, oil, pasta, cheese}
- {bananas, milk, bread, cheese, juice, tuna, beans, leeks, mushrooms, carrots, tomatoes}
- {onions, tomatoes, lentils, crisps, bananas, milk, cheese}

Interest examples

$$interest(\text{if } \{tomatoes\} \text{ then } bananas)$$

$$= confidence - \frac{support(\{bananas\})}{number\ of\ baskets}$$

$$confidence(\text{if } \{cheese\} \text{ then } tomatoes)$$

$$= confidence - \frac{support(\{tomatoes\})}{number\ of\ baskets}$$

Using Interesting Association Rules

Physical retail

- Promote I (e.g. advertise, offer vouchers)
  - but put the price of j up
- if {diapers} then beer
  - Fact or fable?

On-line retail

- When people show interest in I (browse, place in basket)
  - make it easy to buy j
  - e.g. "These are frequently bought together"

Business Rules

- Manually-created rules
  - mostly, non-personalized (or having little personalization)
- E.g. rules that push certain products
  - end of the line stock
  - loss-leaders
  - headline products
  - sequels

RATINGS

Ratings

- Scalar
  - often numeric, e.g. 1-5
  - but always ordered,
    e.g. *strongly disagree, disagree, neutral, agree, strongly agree*
  - not too few values: why?
  - not too many values: why?

7 :	BOOM! One of my FAVORITE few! Can't live without it.
6 :	Solid. They are up there.
5 :	Good Stuff.
4 :	Doesn't turn me on, doesn't bother me.
3 :	Eh. Not really my thing.
2 :	Barely tolerable.
1 :	Pass the earplugs.

- Binary
  - two values, e.g. *+/-, agree/disagree, good/bad, like/dislike*
- Unary
  - one value, e.g. to indicate that a link was clicked, a song was played, an item was purchased,…

Numeric Ratings Matrix

	Alien	Brazil	Crash	Dumbo	E.T.	Fargo
Ben		2	5	3	1	2
Col	5	5		3	4	
Deb					3	
Edd	5	4	2	4	3	3
Flo	2	5	4	4		

Binary Ratings Matrix

	Item1	Item2	Item3	Item4	Item5	Item6
User1		☹	☺	☺	☹	☹
User2	☺	☺		☺	☺	
User3					☺	
User4	☺	☺	☹	☺	☺	☺
User5	☺	☺	☺			

Unary Ratings Matrix

	Item1	Item2	Item3	Item4	Item5	Item6
User1	✓			✓	✓	✓
User2					✓	
User3			✓		✓	
User4	✓	✓		✓		✓
User5			✓	✓	✓	

Ratings Sparsity

- In all cases, a user may have no rating for an item (shown as blank or as ⊥)
- Ratings density: proportion of entries in the matrix ≠⊥
- In most commercial scenarios
  - very large number of items, e.g. thousands, even millions
  - even the most active user likely to have rated < 1%
  - hence, very sparse
    - e.g. MovieLens test data: 6.3%
    - e.g. PTV data: 0.3%
- Makes it very hard to find similar rows or columns

Explicit vs Implicit Ratings

Explicit ratings

- User is asked to provide the rating directly
- Usually numeric; sometimes binary

Implicit ratings

- Based on user actions
  - clicks, read time, searches, purchases,...
- Often only unary
  - not easy to infer negative opinions
  - not easy to infer different degrees on a rating scale

Which is More Accurate?

Explicit ratings
- Previously thought to be more accurate
- But may be inadvertently inaccurate:
  - are you consistent?
  - recently consumed, consumed in the past, not yet consumed?
- May be deliberately inaccurate
  - privacy concerns
  - attempts to bias the system or counteract perceived bias
  - jokes
  - 'posturing'

Implicit ratings
- Unary ratings based on clicks, plays, skips, etc. are fairly accurate
  - but some are done mistakenly
- Larger volumes of implicit ratings
  - reduce effects of noisy data
- Inferred binary or numeric ratings are more problematic
  - e.g. reading time as a measure of interest

Willingness to Rate

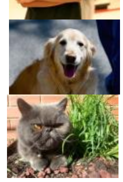
Explicit ratings
- Imposes a cost on users
- Some systems offer incentives for ratings, e.g. discounts
- On the other hand, some users enjoy providing and sharing feedback
  - prestige
  - social interaction
  - the system acts as an extension of their memory

Implicit ratings
- No cost to the user
- But privacy concerns are greater (esp. since the inferences may not be correct and are rarely transparent)
  - e.g. *If TiVo thinks you are gay, here's how to set it straight* (WSJ, 26/10/2002)
  - users may avoid being tracked

Variations in Numeric Ratings

- Some people use the whole scale
- Some don't

- Some people are more positive
- Others are more negative



Z-Scores

- We can normalize ratings to counteract such difference
- E.g. convert to z-scores
  - for a particular user u, first compute average of his/her ratings, $\bar{r_u}$, and their standard deviation, σ_u
  - then to normalize a rating r,

$$z = \frac{r - \bar{r_u}}{\sigma_u}$$

Z-Scores: Example

	Alien	Brazil	Crash	Dumbo	E.T.	Fargo
Ben		2	5	3	1	2

- Average: $\frac{2+5+3+1+2}{5} = 2.6$
- Standard deviation:

$$\sqrt{\frac{(2-2.6)^2+(5-2.6)^2+(3-2.6)^2+(1-2.6)^2+(2-2.6)^2}{5}} = \sqrt{\frac{9.2}{5}} = 1.36$$

- E.g. Ben's normalized rating for Brazil: $\frac{2-2.6}{1.36} = -0.44$

	Alien	Brazil	Crash	Dumbo	E.T.	Fargo
Ben		-0.44	1.76	0.29	-1.18	-0.44