Information Storage and Retrieval

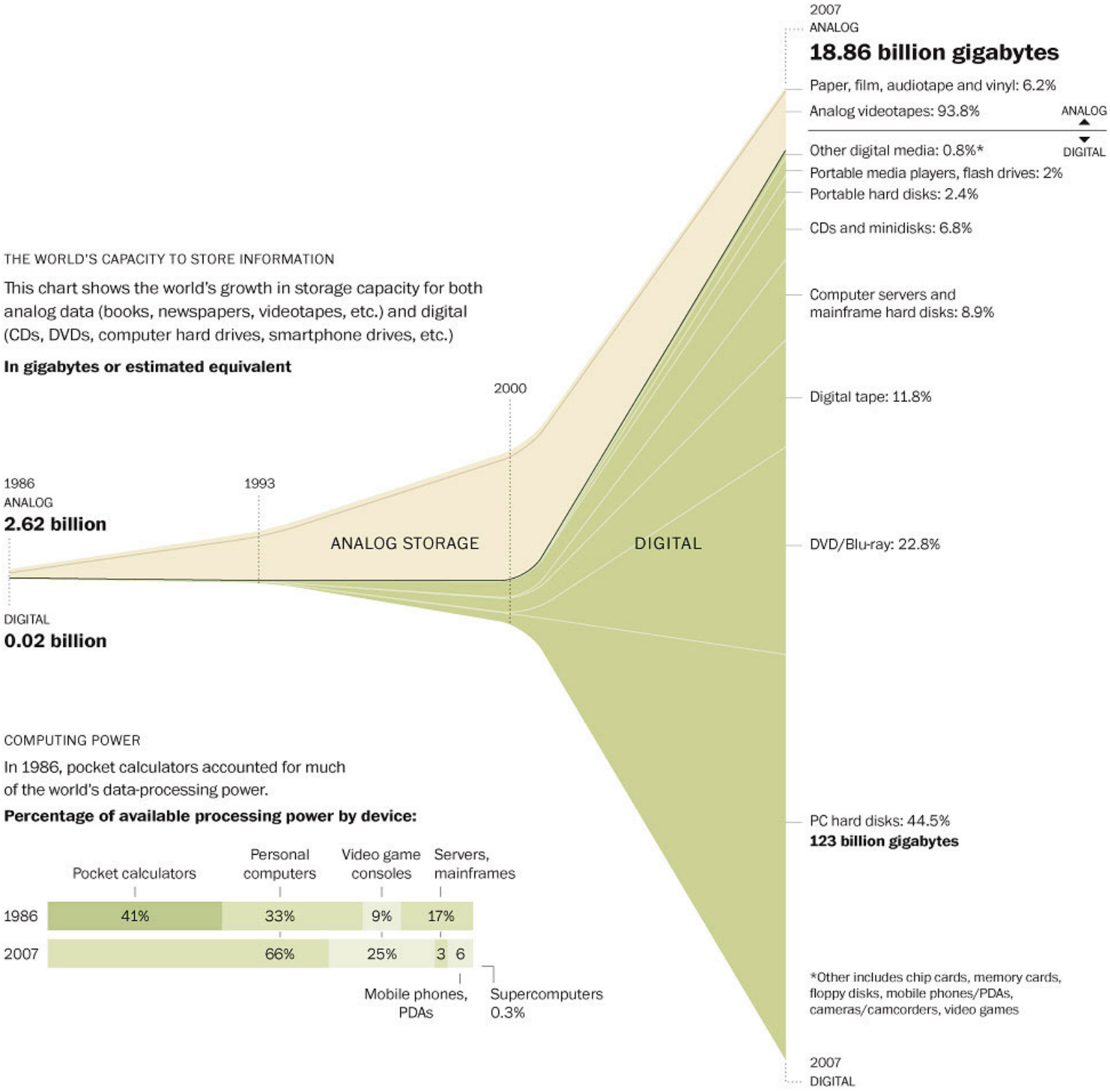
CSCE 670
Texas A&M University
Department of Computer Science & Engineering
Instructor: Prof. James Caverlee

Recommenders
21 March 2017

Today

- Recommendation systems: Intro
- Collaborative Filtering
  - Finding nearest neighbors
  - How to aggregate ratings?

Some slides adapted from MMD: *Mining of Massive Datasets* and from Julian McAuley


THE WORLD'S CAPACITY TO STORE INFORMATION
This chart shows the world's growth in storage capacity for both analog data (books, newspapers, videotapes, etc.) and digital (CDs, DVDs, computer hard drives, smartphone drives, etc.)
In gigabytes or estimated equivalent
1986
ANALOG
2.62 billion
DIGITAL
0.02 billion
1993
2000
ANALOG STORAGE
DIGITAL
2007
ANALOG
18.86 billion gigabytes
Paper, film, audiotape and vinyl: 6.2%
Analog videotapes: 93.8%
ANALOG
DIGITAL
Other digital media: 0.8%*
Portable media players, flash drives: 2%
Portable hard disks: 2.4%
CDs and minidisks: 6.8%
Computer servers and mainframe hard disks: 8.9%
Digital tape: 11.8%
DVD/Blu-ray: 22.8%
PC hard disks: 44.5%
123 billion gigabytes
*Other includes chip cards, memory cards, floppy disks, mobile phones/PDAs, cameras/camcorders, video games
2007
DIGITAL
COMPUTING POWER
In 1986, pocket calculators accounted for much of the world's data-processing power.
Percentage of available processing power by device:
Pocket calculators
Personal computers
Video game consoles
Servers, mainframes
1986
41%
33%
9%
17%
2007
66%
25%
3
6
Mobile phones, PDAs
Supercomputers 0.3%


THE MOBILE WEB RECEIVES 217 NEW USERS.
YOUTUBE USERS UPLOAD 48 HOURS OF NEW VIDEO.
EMAIL USERS SEND 204,166,667 MESSAGES.
GOOGLE RECEIVES OVER 2,000,000 SEARCH QUERIES.
FACEBOOK USERS SHARE 684,478 PIECES OF CONTENT.
WORDPRESS USERS PUBLISH 347 NEW BLOG POSTS.
571 NEW WEBSITES ARE CREATED.
EVERY MINUTE of the DAY
CONSUMERS SPEND $272,070 ON WEB SHOPPING.
FOURSQUARE USERS PERFORM 2,083 CHECK-INS.
TWITTER USERS SEND OVER 100,000 TWEETS.
FLICKR
APPLE

How Many Products Does Amazon Sell?

by **Paul Grey** on 15 December 2013 in **E-Commerce**

Amazon.com is the self-styled "Greatest Store on Earth." It's been said that Amazon aims one day to sell everything to everyone.

Exactly how much choice do you have when shopping with Amazon?

Today Amazon sells over 200 million products in the USA, which are categorised into 35 departments. There are almost 5 million items in the Clothing department, almost 20 million in Sports & Outdoors, and over 4 million Office Products. There 7 million items in the Amazon Jewelry department, 24 million in Electronics, 1.4 million products in the Beauty department, 570 thousand Baby products, and 600 thousand Grocery items.

That's in the USA. This table lists my estimates of the number of products offered on the main Amazon websites around the world.

Amazon.com	USA	232 million
Amazon.co.uk	UK	132 million
Amazon.de	Germany	118 million

Apple: iTunes Now Has 20M Songs; Over 16B Downloads

Posted Oct 4, 2011 by ***Leena Rao*** *(@leenarao)*

At today's Apple iPhone event, newly appointed CEO Tim Cook revealed some staggering numbers on Apple's music downloads and songs in iTunes. iTunes now offers 20 million songs, which is up from 200,000 when iTunes was first launched eight years ago.

iTunes is the number one music store in the world with over 16 billion songs downloaded. It looks like Apple has seen about a billion downloads in the past four months, as the company revealed 15 billion song downloads in June. At the time, Apple revealed that it 225 million credit card accounts listed with iTunes.

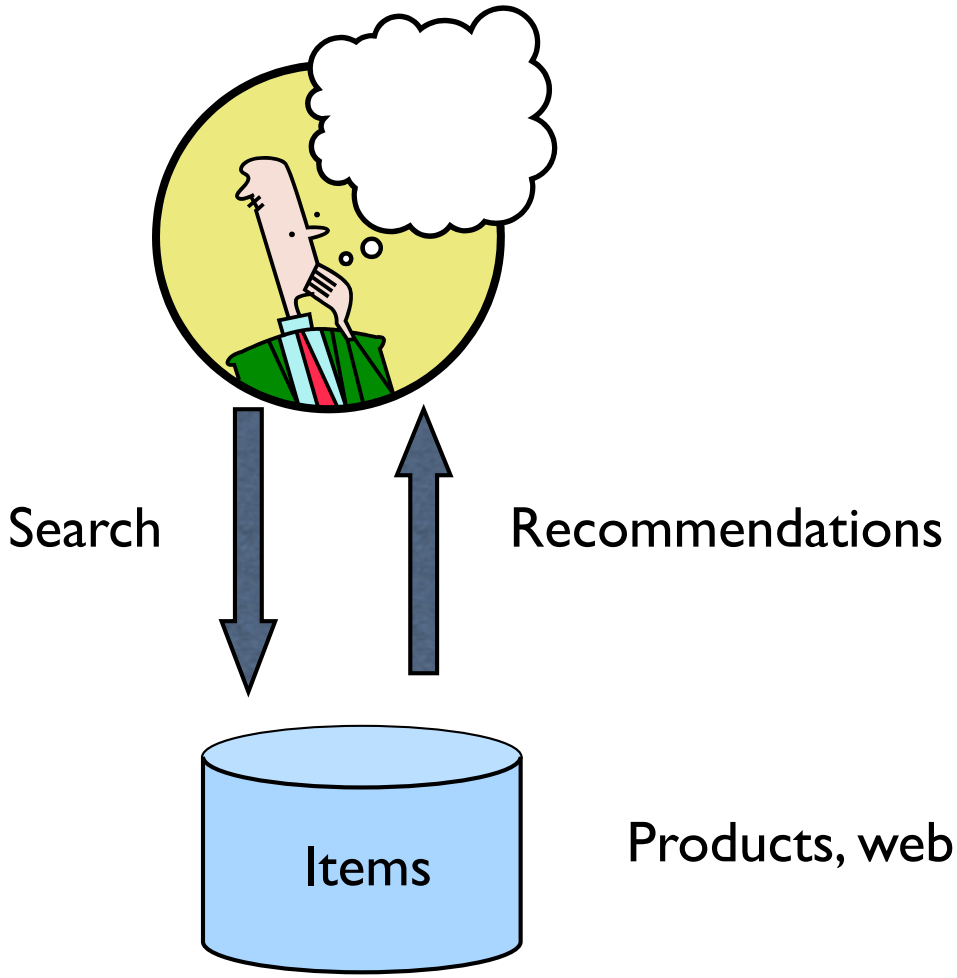
Interestingly, streaming service Spotify has around 15 million songs available, which isn't that much less.

Statistics

Viewership

- More than 1 billion unique users visit YouTube each month
- Over 6 billion hours of video are watched each month on YouTube—that's almost an hour for every person o
- 100 hours of video are uploaded to YouTube every minute
- 80% of YouTube traffic comes from outside the US
- YouTube is localized in 61 countries and across 61 languages
- According to Nielsen, YouTube reaches more US adults ages 18-34 than any cable network
- Millions of subscriptions happen each day. The number of people subscribing daily is up more than 3x since
  number of daily subscriptions is up more than 4x since last year

Recommendations



Products, web sites, blogs, news items, …

Why recommendation?

The goal of recommender systems is…

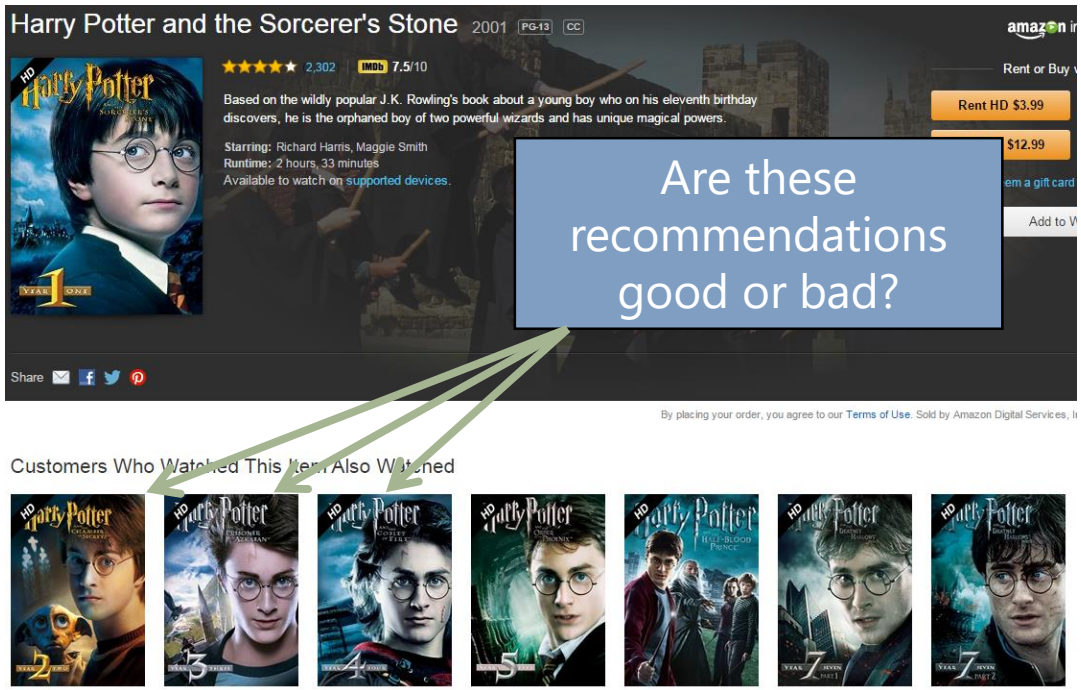
- To help people discover new content

Recommendations for You in Amazon Instant Video See more

Why recommendation?

The goal of recommender systems is…

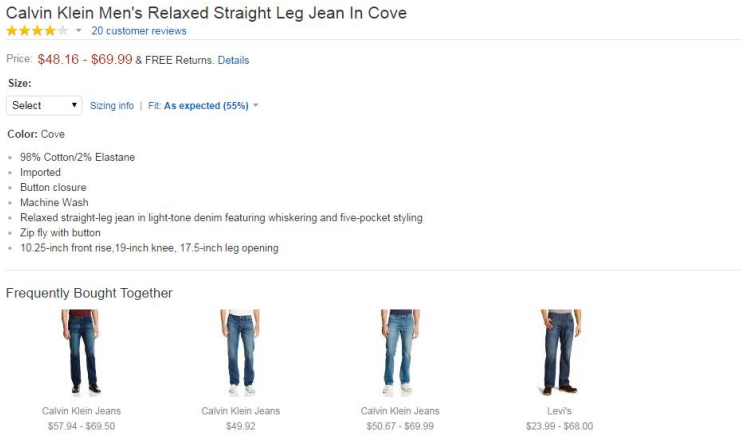
- To help us find the content we were already looking for



Why recommendation?

The goal of recommender systems is…

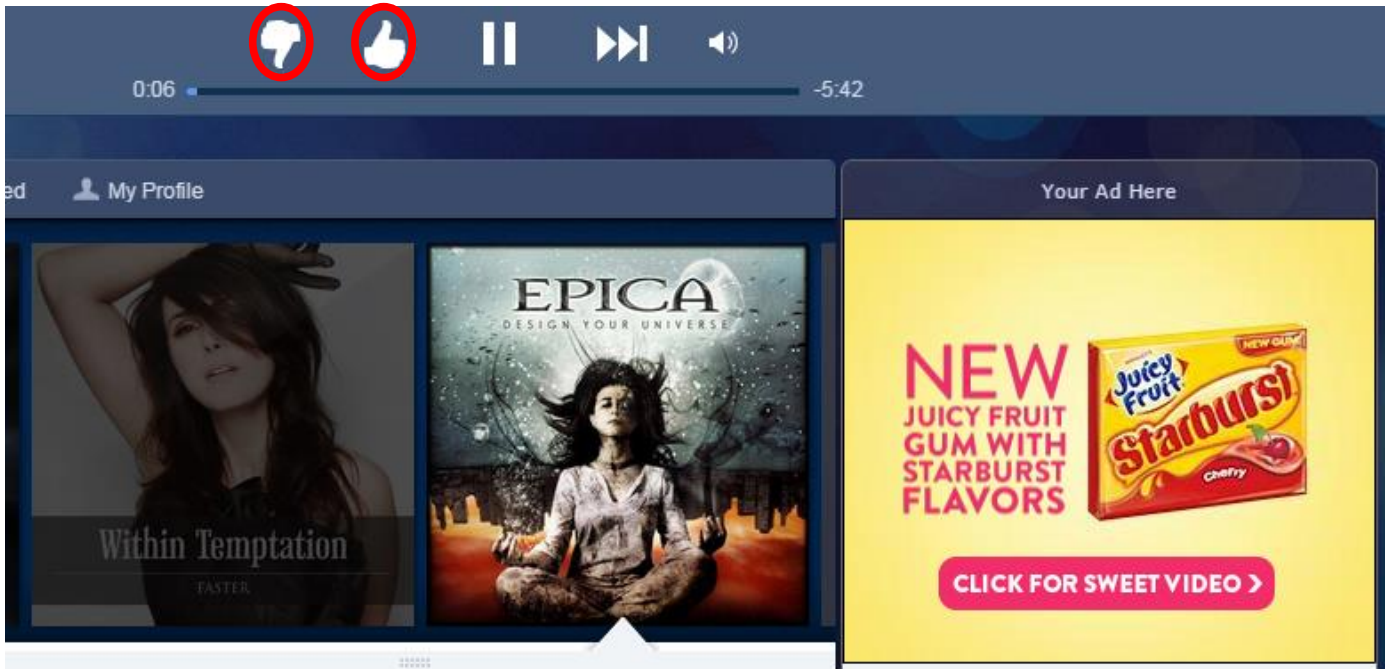
- To discover which things go together



Why recommendation?

The goal of recommender systems is…

- To personalize user experiences in response to user feedback



Why recommendation?

The goal of recommender systems is…

- To identify things that we **like**

Examples of recommenders?

Recommendation Types

- Editorial and hand curated
  - List of favorites
  - Lists of "essential" items
- Simple aggregates
  - Top 10, Most Popular, Recent Uploads
- **Tailored to individual users**
  - **Amazon, Netflix, ...**

$$$


Netflix Prize: Home
http://www.netflixprize.com/
netflix prize
Getting Started
Latest Headlines
NETFLIX
Netflix Prize
Home
Rules
Leaderboard
Register
Update
Submit
Download
NETFLIX
Browse
Recommendations
Friends
Queue
Buy DVDs
Movies For You
Welcome!
The Netflix Prize seeks to substantially improve the accuracy of predictions about how much someone is going to love a movie based on their movie preferences. Improve it enough and you win one (or more) Prizes. Winning the Netflix Prize improves our ability to connect people to the movies they love.
Read the Rules to see what is required to win the Prizes. If you are interested in joining the quest, you should register a team.
You should also read the frequently-asked questions about the Prize. And check out how various teams are doing on the Leaderboard.
Good luck and thanks for helping!
FAQ | Forum | Netflix Home
© 1997-2006 Netflix, Inc. All rights reserved.
Done

Formal Model

- X = set of Customers
- S = set of Items
- Utility function $u: X \times S \rightarrow R$
  - R = set of ratings
  - R is a totally ordered set
  - e.g., 0-5 stars, real number in [0,1]

Utility Matrix

	Spotlight	The Revenant	Mad Max: Fury Road	Room
Alice	1		0.2	
Bob		0.5		0.3
Carol	0.2		1	
David				0.4

Key Problems

- Gathering "known" ratings for matrix
- Extrapolate unknown ratings from known ratings
  - Mainly interested in high unknown ratings
  - Don't care about finding what you **don't like**, but rather what you do like
- Evaluating extrapolation methods
  - How do we know if we've done a good job?

Gathering Ratings

- Explicit
  - Ask people to rate items
  - Doesn't work well in practice – people can't be bothered
- Implicit
  - Learn ratings from user actions
    - e.g., purchase implies high rating
  - What about low ratings?

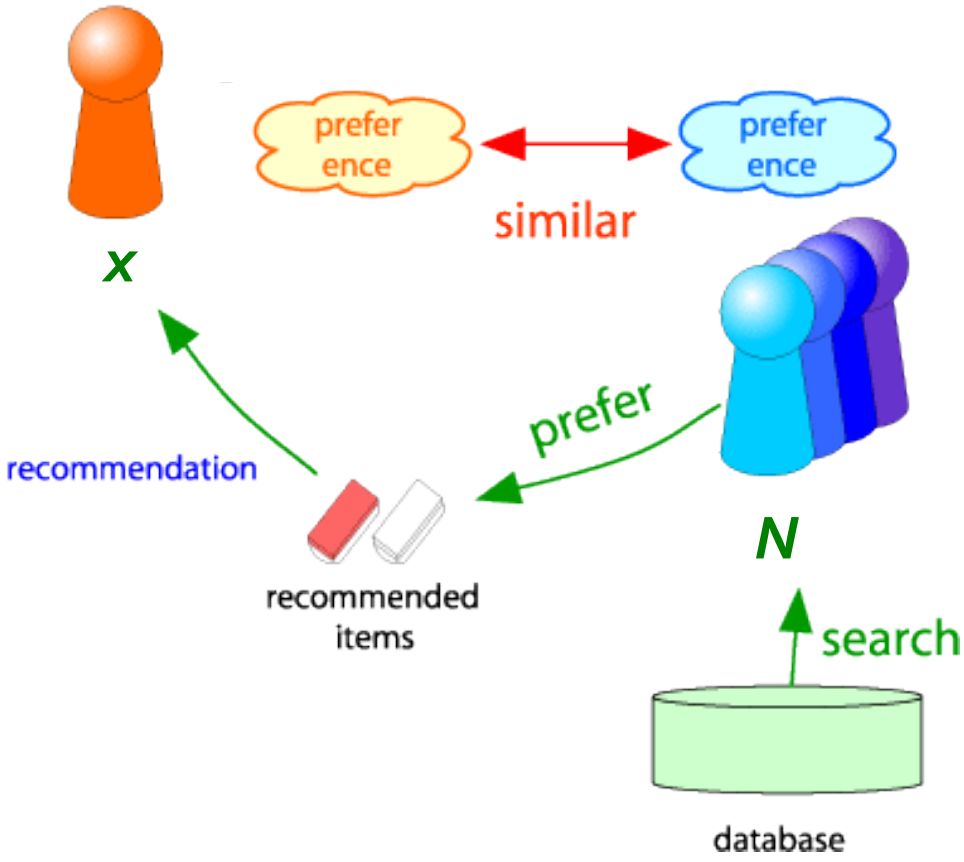
Extrapolating Utilities

- Key problem: matrix U is sparse
  - most people have not rated most items
  - Cold start:
    - New items have no ratings
    - New users have no history
- Three main approaches
  - Collaborative
  - Content-based
  - Latent factor models

Collaborative Recommendations

Collaborative Filtering

- Consider user ***x***
- Find set ***N*** of other users whose ratings are "similar" to ***x***'s ratings
- Estimate ***x***'s ratings based on ratings of users in ***N***



Finding Similar Users

- Let r_x be the vector of user x's ratings
- Jaccard!
  - But it ignores the value of the rating
- Cosine similarity measure
  - sim(x,y) = cos(r_x, r_y)
  - But it treats missing ratings as "negative"
- Pearson correlation coefficient
  - S_{xy} = items rated by both users x and y

$$sim(x,y) = \frac{\sum_{s \in S_{xy}} (r_{xs} - \bar{r_x})(r_{ys} - \bar{r_y})}{\sqrt{\sum_{s \in S_{xy}} (r_{xs} - \bar{r_x})^2 (r_{ys} - \bar{r_y})^2}}$$

Jaccard

$$\text{Jaccard}(A, B) =$$

$$\text{Jaccard}(U_i, U_j) =$$

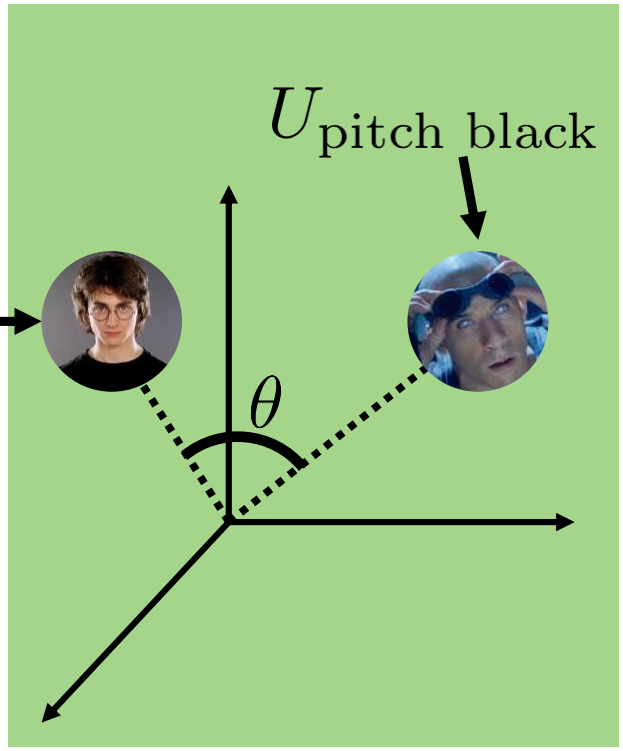
→ Maximum of 1 if the two users purchased **exactly the same** set of items
(or if two items were purchased by the same set of users)

→ Minimum of 0 if the two users purchased **completely disjoint** sets of items
(or if the two items were purchased by completely disjoint sets of users)

Cosine

$U_{\text{harry potter}}$

(vector representation of users who purchased harry potter)



$$\cos(\theta) = 1$$

(theta = 0) → *A* and *B* point in exactly the same direction

$$\cos(\theta) = -1$$

(theta = 180) → *A* and *B* point in opposite directions (won't actually happen for 0/1 vectors)

$$\cos(\theta) = 0$$

(theta = 90) → *A* and *B* are orthogonal

Cosine

Why cosine?

- Unlike Jaccard, works for arbitrary vectors
- E.g. what if we have **opinions** in addition to purchases?

$$R = \begin{pmatrix} 1 & 0 & \cdots & 1 \\ 0 & 0 & & 1 \\ \vdots & & \ddots & \vdots \\ 1 & 0 & \cdots & 1 \end{pmatrix} \rightarrow \begin{pmatrix} -1 & 0 & \cdots & 1 \\ 0 & 0 & & -1 \\ \vdots & & \ddots & \vdots \\ 1 & 0 & \cdots & -1 \end{pmatrix}$$

bought and **liked**

didn't buy

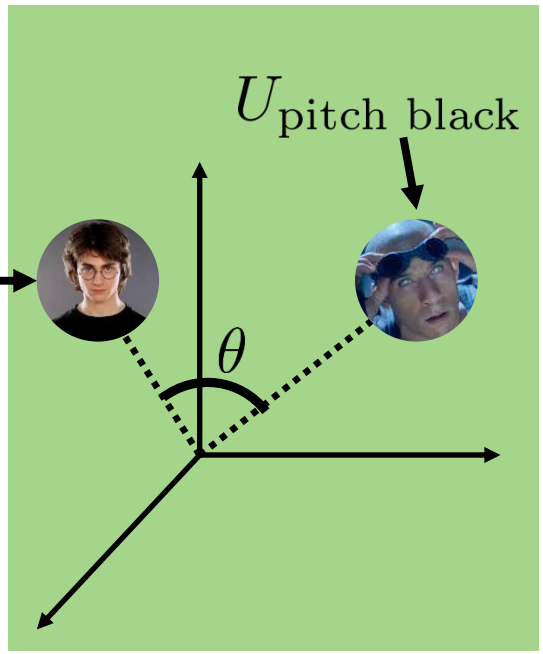
bought and **hated**

Cosine (with ratings)

E.g. our previous example, now with "thumbs-up/thumbs-down" ratings

$U_{\text{harry potter}}$

(vector representation of users' **ratings of** Harry Potter)



$$\cos(\theta) = 1$$

(theta = 0) → Rated by the same users, and they all agree

$$\cos(\theta) = -1$$

(theta = 180) → Rated by the same users, but they **completely disagree** about it

$$\cos(\theta) = 0$$

(theta = 90) → Rated by different sets of users

Pearson correlation

What if we have numerical ratings (rather than just thumbs-up/down)?

$$R = \begin{pmatrix} -1 & 0 & \cdots & 1 \\ 0 & 0 & & -1 \\ \vdots & & \ddots & \vdots \\ 1 & 0 & \cdots & -1 \end{pmatrix} \rightarrow \begin{pmatrix} 4 & 0 & \cdots & 2 \\ 0 & 0 & & 3 \\ \vdots & & \ddots & \vdots \\ 5 & 0 & \cdots & 1 \end{pmatrix}$$

bought and **liked**

didn't buy

bought and **hated**

Pearson correlation

What if we have numerical ratings (rather than just thumbs-up/down)?

- We wouldn't want 1-star ratings to be parallel to 5-star ratings
- So we can subtract the average – values are then **negative** for below-average ratings and **positive** for above-average ratings

items rated by both users

average rating by user *v*

$$\mathrm{Sim}(u,v) = \frac{\sum_{i \in I_u \cap I_v} (R_{u,i} - \bar{R_u})(R_{v,i} - \bar{R_v})}{\sqrt{\sum_{i \in I_u \cap I_v} (R_{u,i} - \bar{R_u})^2 \sum_{i \in I_u \cap I_v} (R_{v,i} - \bar{R_v})^2}}$$

Example

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

- Intuitively we want: sim(A,B) > sim(A,C)
- Jaccard: 1/5 < 2/4
- Cosine: 0.386 > 0.322
  - Considers missing ratings as "negative"
  - Solution: subtract the (row) mean

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	$2/3$			$5/3$	$-7/3$		
B	$1/3$	$1/3$	$-2/3$				
C				$-5/3$	$1/3$	$4/3$	
D		0					0

- sim(A,B) vs sim (A,C)
  - 0.092 > -0.559

- So far, we can find a rating, but how do we actually generate recommendations?

How to aggregate ratings?

$$r_{c,s} = \operatorname{aggr}_{c' \in \hat{C}} r_{c',s},$$

"similar" users to c

How to aggregate ratings?

$$\text{(a) } r_{c,s} = \frac{1}{N} \sum_{c' \in \hat{C}} r_{c',s},$$

$$\text{(b) } r_{c,s} = k \sum_{c' \in \hat{C}} sim(c, c') \times r_{c',s},$$

$$\text{(c) } r_{c,s} = \bar{r}_c + k \sum_{c' \in \hat{C}} sim(c, c') \times (r_{c',s} - \bar{r}_{c'}),$$

Many other tricks possible

Item-Item Collaborative Filtering

- So far: User-user collaborative filtering
- Another view
  - For item s, find other similar items
  - Estimate rating for item based on ratings for similar items
  - Can use same similarity metrics and prediction functions as in user-user model
- In practice, it has been observed that item-item often works better than user-user

Pros/cons of collaborative filtering

- Works for any kind of item
  - No feature selection needed
- Cold start:
  - Need enough users in the system to find a match
- Sparsity:
  - The user/ratings matrix is sparse
  - Hard to find users that have rated the same items
- First rater:
  - Cannot recommend an item that has not been previously rated
  - New items, esoteric items
- Popularity bias:
  - Cannot recommend items to someone with unique taste
  - Tends to recommend popular items