Information Storage and Retrieval

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

Content-Based Recommenders
4 April 2017

Today

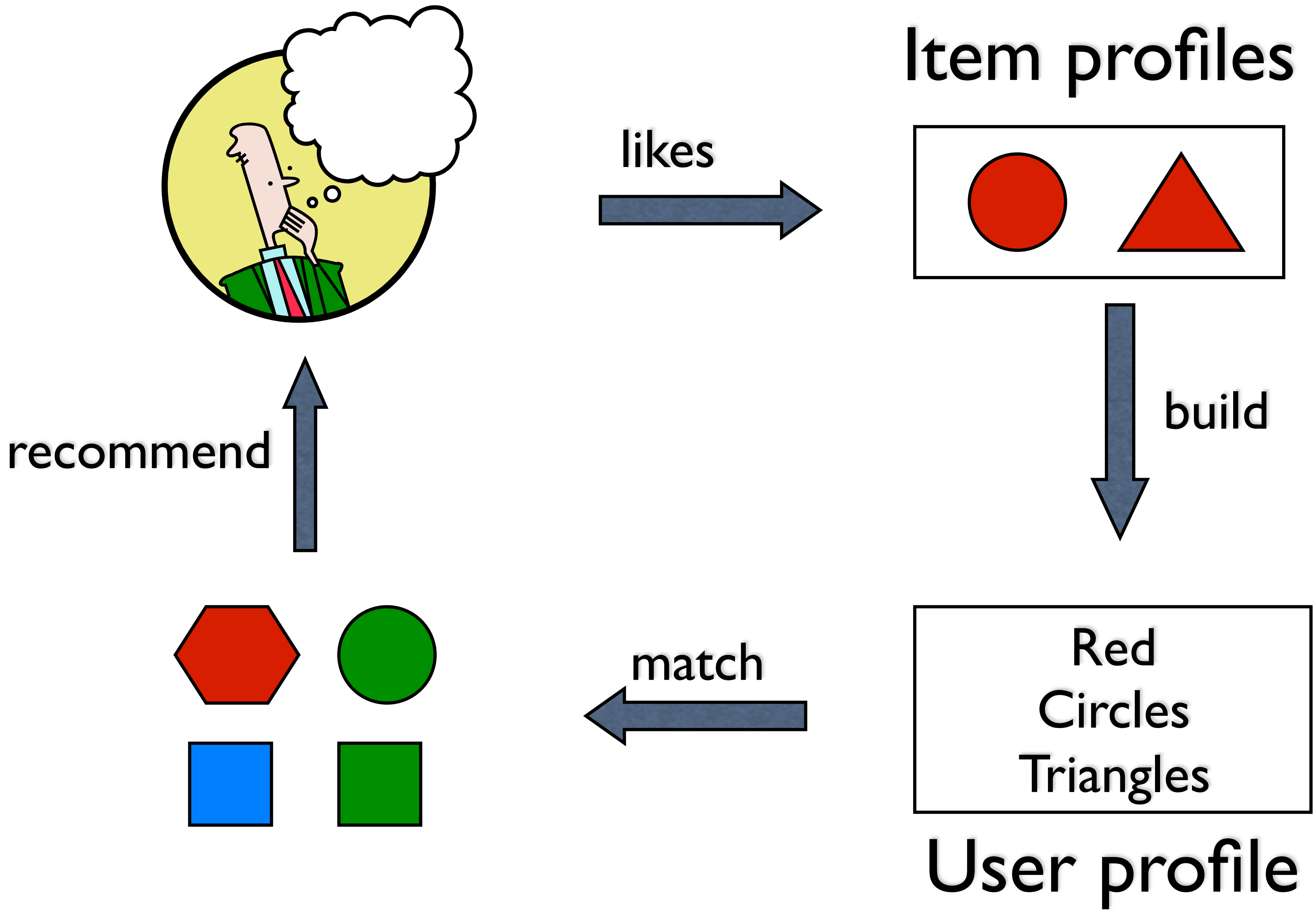
- Content-based recommenders
- Model-based recommenders
- (recap of where we are)
- Evaluation
- Attacks

Content-based recommendations

Content-based recommendations

- **Main idea**: recommend items to customer ***x*** similar to previous items rated highly by ***x***
- Movie recommendations
  - recommend movies with same actor(s), director, genre, …
- Websites, blogs, news
  - recommend other sites with "similar" content

Plan of action



Item Profiles

- For each item, create an item profile
- Profile is a set of features (vectors!)
    - movies: author, title, actor, director,…
    - text: set of "important" words in document
- How to pick important words?
    - Usual heuristic is TF.IDF

User profiles and prediction

- User profile possibilities:
  - Weighted average of rated item profiles
  - Variation: weight by difference from average rating for item
  - …
- Prediction heuristic
  - Given user profile **x** and item profile **i**, estimate
    - $u(\mathbf{x},\mathbf{i}) = \cos(\mathbf{x},\mathbf{i}) = \mathbf{x}.\mathbf{i}/(|\mathbf{x}||\mathbf{i}|)$

Advantages of Content-based Recs?

- No need for data on other users
  - No cold-start or sparsity problems
- Able to recommend to users with unique tastes
- Able to recommend new and unpopular items
  - No first-rater problem
- Can provide explanations of recommended items by listing content-features that caused item to be recommended

Limitations of content-based approach

- Finding the appropriate features
    - e.g., images, movies, music
- Recommendations for new users
    - How to build a profile?
- Overspecialization
    - Never recommends items outside user's content profile
    - People might have multiple interests
    - Unable to exploit quality judgments of other users

Hybrid: Content + Collaborative

Hybrid Methods

- Implement two separate recommenders and combine predictions
- Add content-based methods to collaborative filtering
    - item profiles for new item problem
    - demographics to deal with new user problem

Recommendation Approach	Recommendation Technique	
	Heuristic-based	Model-based
Content-based	Commonly used techniques: • TF-IDF (information retrieval) • Clustering Representative research examples: • Lang 1995 • Balabanovic & Shoham 1997 • Pazzani & Billsus 1997	Commonly used techniques: • Bayesian classifiers • Clustering • Decision trees • Artificial neural networks Representative research examples: • Pazzani & Billsus 1997 • Mooney et al. 1998 • Mooney & Roy 1999 • Billsus & Pazzani 1999, 2000 • Zhang et al. 2002
Collaborative	Commonly used techniques: • Nearest neighbor (cosine, correlation) • Clustering • Graph theory Representative research examples: • Resnick et al. 1994 • Hill et al. 1995 • Shardanand & Maes 1995 • Breese et al. 1998 • Nakamura & Abe 1998 • Aggarwal et al. 1999 • Delgado & Ishii 1999 • Pennock & Horwitz 1999 • Sarwar et al. 2001	Commonly used techniques: • Bayesian networks • Clustering • Artificial neural networks • Linear regression • Probablistic models Representative research examples: • Billsus & Pazzani 1998 • Breese et al. 1998 • Ungar & Foster 1998 • Chien & George 1999 • Getoor & Sahami 1999 • Pennock & Horwitz 1999 • Goldberg et al. 2001 • Kumar et al. 2001 • Pavlov & Pennock 2002 • Shani et al. 2002 • Yu et al. 2002, 2004 • Hofmann 2003, 2004 • Marlin 2003 • Si & Jin 2003
Hybrid	Combining content-based and collaborative components using: • Linear combination of predicted ratings • Various voting schemes • Incorporating one component as a part of the heuristic for the other Representative research examples: • Balabanovic & Shoham 1997 • Claypool et al. 1999 • Good et al. 1999 • Pazzani 1999 • Billsus & Pazzani 2000 • Tran & Cohen 2000 • Melville et al. 2002	Combining content-based and collaborative components by: • Incorporating one component as a part of the model for the other • Building one unifying model Representative research examples: • Basu et al. 1998 • Condliff et al. 1999 • Soboroff & Nicholas 1999 • Ansari et al. 2000 • Popescul et al. 2001 • Schein et al. 2002

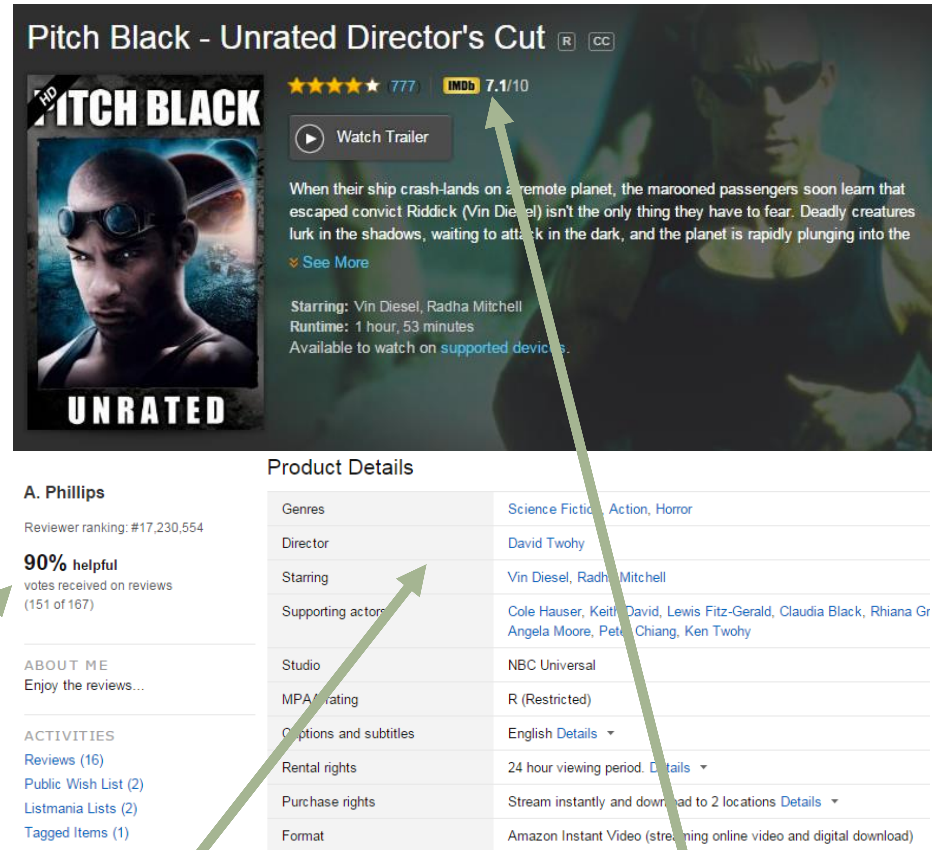
Model-Based Methods

Slides from Julian McAuley

Suppose we want to build a movie recommender

e.g. which of these films will I rate highest?

We already have a few tools in our "supervised learning" toolbox that may help us



$$f(\text{user features}, \text{movie features}) \overset{?}{\longrightarrow} \text{star rating}$$

$$f(\text{user features}, \text{movie features}) \xrightarrow{?} \text{star rating}$$

Movie features: genre, actors, rating, length, etc.

Product Details

Genres	Science Fiction, Action, Horror
Director	David Twohy
Starring	Vin Diesel, Radha Mitchell
Supporting actors	Cole Hauser, Keith David, Lewis Fitz-Gerald, Claudia Black, Rhiana Gr Angela Moore, Peter Chiang, Ken Twohy
Studio	NBC Universal
MPAA rating	R (Restricted)
Captions and subtitles	English Details
Rental rights	24 hour viewing period. Details
Purchase rights	Stream instantly and download to 2 locations Details
Format	Amazon Instant Video (streaming online video and digital download)

User features: age, gender, location, etc.

A. Phillips

Reviewer ranking: #17,230,554

90% helpful
votes received on reviews
(151 of 167)

ABOUT ME
Enjoy the reviews...

ACTIVITIES
Reviews (16)
Public Wish List (2)
Listmania Lists (2)
Tagged Items (1)

$$f(\text{user features}, \text{movie features}) \xrightarrow{?} \text{star rating}$$

With the models we've seen so far, we can build predictors that account for…

- Do women give higher ratings than men?
- Do Americans give higher ratings than Australians?
- Do people give higher ratings to action movies?
- Are ratings higher in the summer or winter?
- Do people give high ratings to movies with Vin Diesel?

So what **can't** we do yet?

$$f(\text{user features}, \text{movie features}) \stackrel{?}{\to} \text{star rating}$$

Consider the following linear predictor (e.g. from week 1):

$$f(\text{user features}, \text{movie features}) = \langle \phi(\text{user features}); \phi(\text{movie features}), \theta \rangle$$

But this is essentially just two separate predictors!

$$f(\text{user features}, \text{movie features}) =$$

$$= \underbrace{\langle \phi(\text{user features}), \theta_{\text{user}} \rangle}_{\text{user predictor}} + \underbrace{\langle \phi(\text{movie features}), \theta_{\text{movie}} \rangle}_{\text{movie predictor}}$$

That is, we're treating user and movie features as though they're **independent!**

But these predictors should (obviously?) **not** be independent

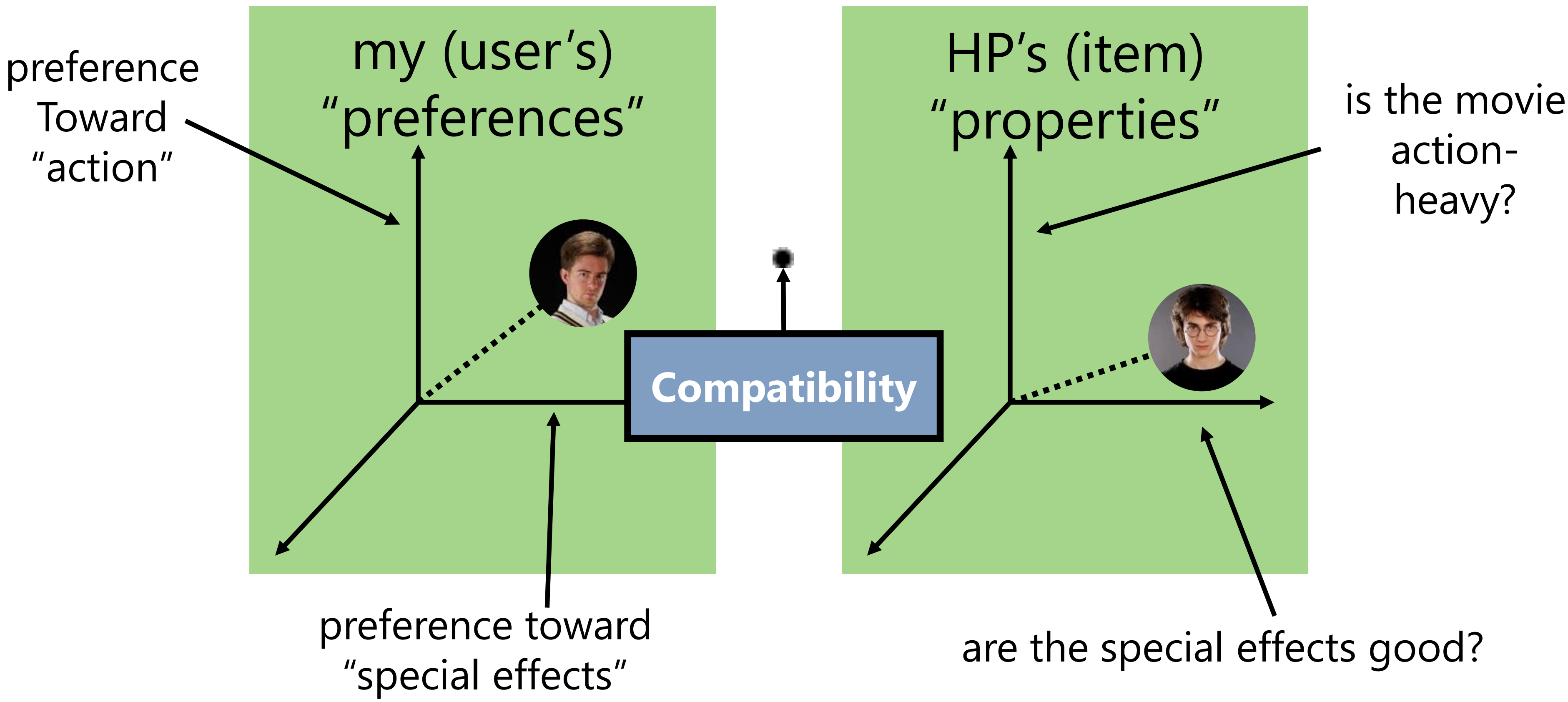
$$f(\text{user features}, \text{movie features}) = f(\text{user}) + f(\text{movie})$$

do I tend to give high ratings?

does the population tend to give high ratings to this genre of movie?

But what about a feature like "do **I** give high ratings to **this genre** of movie"?

Recommender Systems go beyond the methods we've seen so far by trying to model the **relationships** between people and the items they're evaluating



Recap

- Ratings-based
  - Baseline (overall average + user-bias + item-bias)
  - Collaborative filtering (user-user, item-item)
  - Latent factor approaches (SVD)
- Content-based
- Hybrid collaborative + content
- Model-based

Evaluation

Evaluating Predictions

- Compare predictions with known ratings
  - Root-mean-square error (RMSE)
- Another approach:
  - Coverage
    - Number of items/users for which system can make predictions
  - Precision
    - Accuracy of predictions
  - Receiver operating characteristic (ROC)
    - Tradeoff curve between false positives and false negatives

Problems with Measures

- Narrow focus on accuracy sometimes misses the point
  - Prediction Diversity
  - Prediction Context
  - Order of predictions

Extending capabilities

- Multidimensionality of recommendations
- Multi-criteria ratings
- Non-intrusiveness
- Flexibility
- Effectiveness of recommendations