Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Seminar Collaborative Filtering

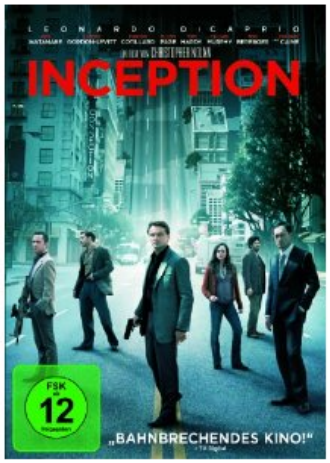
KDD Cup

Ziawasch Abedjan, Arvid Heise, Felix Naumann

Collaborative Filtering

Recommendation systems

Recommendation systems



Für Bildvergrößerung und andere Ansichten klicken

2 weitere Bilder und Videos ansehen

Eigene Bilder hinzufügen

Inception

Leonardo DiCaprio (Darsteller), Joseph Gordon-Levitt (Darsteller), Christopher Nolan (Regisseur) | Alterseinstufung: Freigegeben ab 12 Jahren | Format: DVD

★★★★☆ (505 Kundenrezensionen) Gefällt mir (39)

Preis: EUR 7,95 Kostenlose Lieferung ab EUR 20 (auch bei Verkäufern mit "Versand durch Amazon.de").
Alle Bücher und Blu-rays versandkostenfrei. Details
Alle Preisangaben inkl. MwSt.

Auf Lager.
Verkauf und Versand durch **Amazon.de**. Geschenkverpackung verfügbar.

Lieferung bis Dienstag, 12. April: Bestellen Sie in den nächsten 6 Stunden und 17 Minuten per **Overnight-Express**. Siehe Details.

68 neu ab EUR 4,71 50 gebraucht ab EUR 3,00 1 Sammlerstück(e) ab EUR 12,96

Leihen Sie DVDs & Blu-rays bei LOVEFiLM
LOVEFiLM ist Europas größte Online-Videothek mit einer Auswahl von über 40.000 Titeln. Testen Sie LOVEFiLM.de jetzt 14 Tage lang kostenlos und erhalten Sie einen 20 EUR Gutschein für Amazon.de. Alles zur Aktion bei LOVEFiLM.de

Weitere Ausgaben:	Preis:	Weitere Angebote:
Blu-ray	EUR 11,99	76 Angebote ab EUR 7,50

Kunden, die diesen Artikel gekauft haben, kauften auch

Salt (Deluxe Extended Edition) [Deluxe Edition] DVD ~ Angelina Jolie
★★★½☆ (146)
EUR 8,97

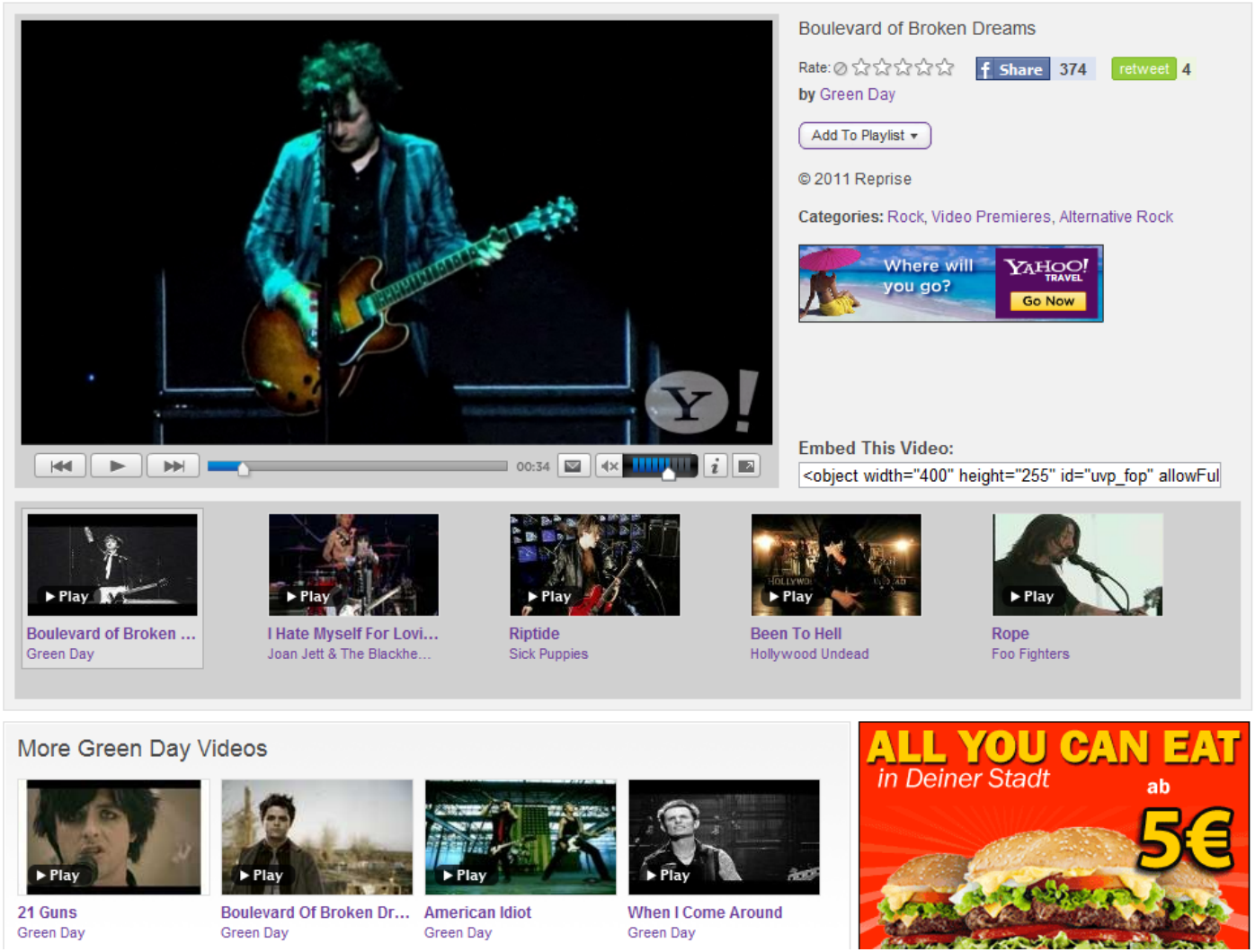
Knight and Day - Agentenpaar wider Willen (Exten... DVD ~ Tom Cruise
★★★★☆ (115)
EUR 7,97

Das A-Team - Der Film (Extended Cut) DVD ~ Liam Neeson
★★★½☆ (177)
EUR 14,99

Shutter Island DVD ~ Leonardo DiCaprio
★★★★☆ (195)
EUR 8,91

Avatar (Extended Collector's Edition, 3... DVD ~ Sam Worthington
★★★★☆ (241)
EUR 17,95

Recommendation systems



Recommendation systems

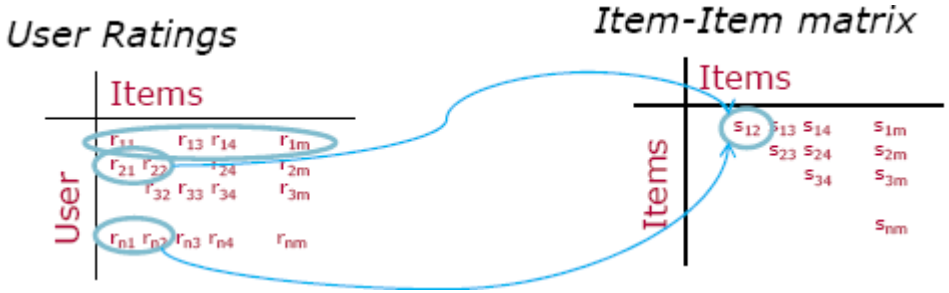
- Content-based recommendations
  - Relevance is based on content
  - Similar items
  - Uses also metadata
- Collaborative filtering
  - Recommend items suiting one’s taste
  - Based on community ratings
  - Serendipity

Collaborative Filtering Algorithms

- Memory-based CF
  - Item/user-based top N
  - New items and user can be added easily
  - Scales well on dense data sets
  - Limited scalability on sparse data sets
- Model-based CF
  - Singular Value Decomposition (SVD)
  - Better scalability on sparse data sets
  - Little evidence for the model
- Hybrid recommenders
  - Content-boosted CF
  - Mixture of memory-based and model-based techniques

Neighbor-based CF

- Process: filtering -> prediction -> recommendation
- User-based approach
  - Create cluster of similar users
  - Recommendations depend on ratings of similar users
  - Problem: missing ratings
- Item-based approach
  - Create item-item matrix based on rating similarities
  - Retrieve top K similar items and aggregating their ratings by the user



Singular Value Decomposition

- Reduce dimensions of user/item matrix by factorization
- Use overlapping features (e.g. gender, genre)

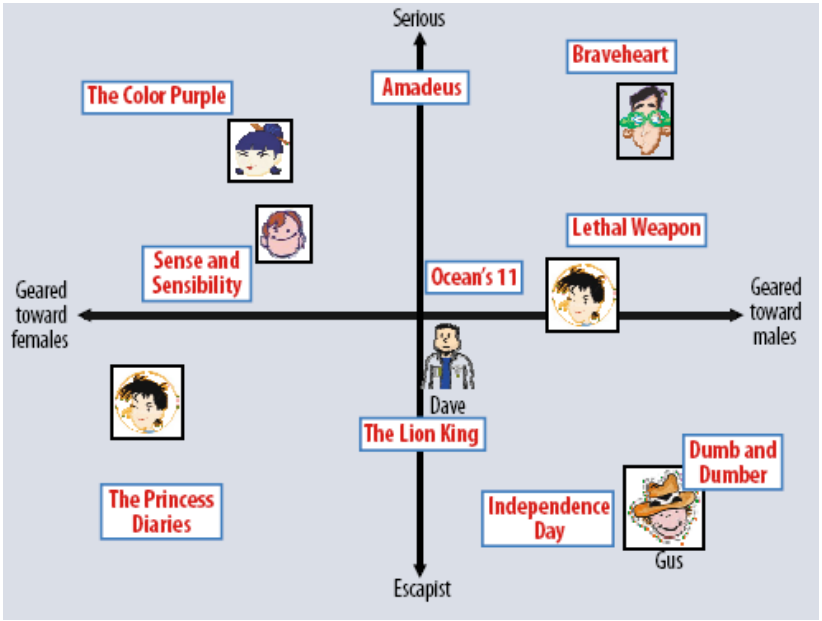
user	items		
	r_{11}		r_{1n}
	r_{i1}	r_{ij}	
	r_{m1}		r_{mn}

=

user	features	
	f_{11}	f_{12}
	f_{m1}	f_{m2}

X

feature	items			
	g_{11}	g_{12}	...	g_{1n}
	g_{21}	g_{22}		g_{2n}



Source: Koren, Yehuda, Bell, Robert & Volinsky, Chris: "Matrix Factorization Techniques for Recommender Systems"

KDD Cup

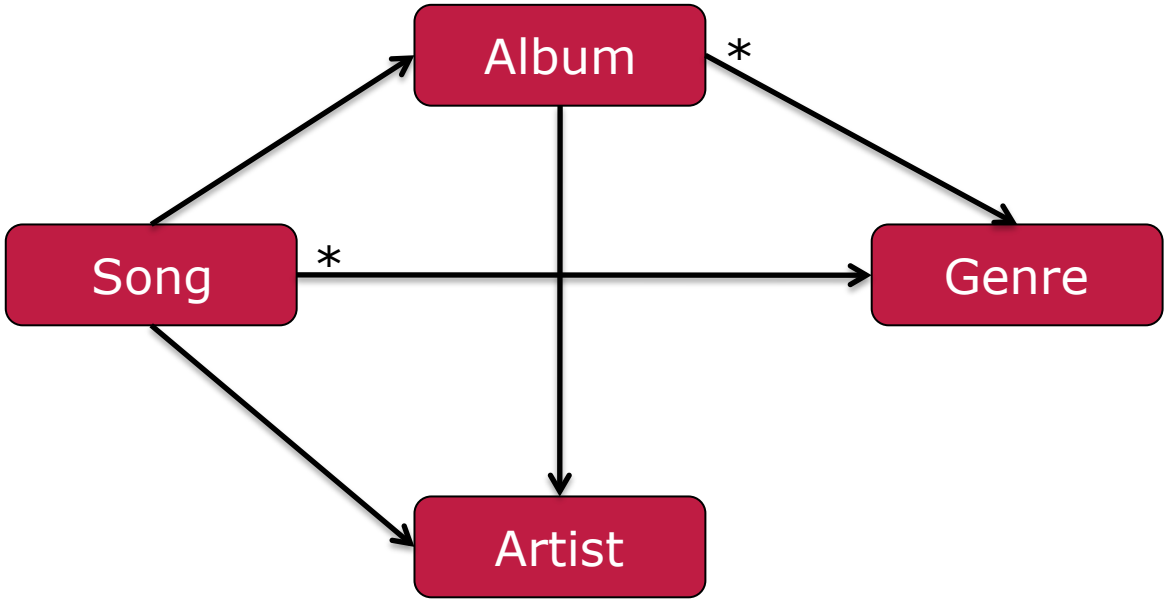
Overview

- ACM SIGKDD
  - Knowledge Discovery and Data Mining
- Variety of topics
  - 2010 - Student performance evaluation
  - 2008 - Breast cancer
  - 2007 - **Consumer recommendations**
  - 2004 - Particle physics; plus protein homology prediction
  - 2001 - Molecular bioactivity; plus protein locale prediction
  - 1997 - Direct marketing for lift curve optimization
- 2 tracks with three places

KDD Cup 2011

- On Yahoo Music Dataset
  - Artists, Albums, Songs, Genres

- Track 1: predict user rating
- Track 2: decide whether a user rates a song or not



Track 1

- Typical collaboration filtering task
- Predict the rating of a specific user for unrated songs

- Includes hierarchy of items
  - User might have rated other songs of same album/artist
  - No user might have rated the song but the same album/artist
- Includes time stamp of rating
  - Rating behavior might have changed over time
  - Older songs rated differently than newer songs?

#Users	#Items	#Ratings	#Train Ratings	#Validation Ratings	#Test Ratings
1,000,990	624,961	262,810,175	252,800,275	4,003,960	6,005,940

Track 2

- Predict if a user would rate a given song highly or not at all
- Need a model for rate behavior
  - Machine learning?

- Ratings on songs only
- No timestamps
- Given six songs, which three will most likely not be rated

#Users	#Items	#Ratings	#Train Ratings	#Test Ratings	#Users
249,012	296,111	62,551,438	61,944,406	607,032	249,012

Seminar

Approaches Track 1

- Koren, Y., Bell, R. & Volinsky, C.: "Matrix Factorization Techniques for Recommender Systems"
  - Similar to Singular Value Decomposition
  - Create small abstracted matrix
  - Soft clustering

- Sarwar, B., Karypis, G., Konstan, J. & Reidl, J.: "Item-based collaborative filtering recommendation algorithms"
  - Classic approach
  - Improvements in Robert M. Bell & Yehuda Koren: "Improved Neighborhood-based Collaborative Filtering"
  - Need to be adjusted to include time stamp and hierarchy

Approaches Track 2

- Kurucz, M., Benczúr, A.A., Kiss, T., Nagy, I., Szabó, A. & Torma, B.: "Who Rated What: a Combination of SVD, Correlation and Frequent Sequence Mining"
  - Combination of four predictions (only two relevant)
    - SVD and base prediction
  - Needs estimation on how many ratings

- Sueiras, J.: "A classical predictive modeling approach for Task “Who rated what?” of the KDD CUP 2007"
  - Machine learning (logistic regression)
  - User-specific, movie-related, and user-movie pair features
  - SVD for movie-related

Application

- Send mail to ziawasch.abedjan@hpi.uni-potsdam.de
- Subject: [CF Seminar]
- Deadline: April 18th
- Notification: April 19th

- Limited to 8 participants = 4 teams
  - Random selection if more applicants

- Send top 3 wishes on tasks and approaches
  - We are open to other approaches

Time Schedule

- April 14th: first seminar, topic presentations
- April **16th**: **application deadline**, team/paper preferences
- April **17th**: team/paper notification

- April 21th: mandatory consultation
- April 28th: **paper presentation**

- May 12th: initial implementation/idea presentation
- June 9th: intermediate presentation/project consolidation
- June 23th: **final presentations**
- June 30th: KDD cup **submission deadline** (results only)

- July 9th: hand in **short paper** (6 pages)

Grading

- 3 LP (half semester, project seminar)
- Paper and final presentations
- Participation during intermediate presentations / discussions
- Implementation strategies/proposed extensions
- Short paper
- Bonus: good results in KDD cup

Options

- Mahout or stand-alone
- Single or combined repository
  - We recommend cooperation
  - Many similar issues

References

- Official KDD Cup: http://kddcup.yahoo.com/
- Su, Xiaoyuan & Khoshgoftaar, Taghi M.: "A survey of collaborative filtering techniques"

- Koren, Yehuda; Bell, Robert & Volinsky, Chris: "Matrix Factorization Techniques for Recommender Systems"
- Sarwar, Badrul; Karypis, George, Konstan, Joseph & Reidl, John: "Item-based collaborative filtering recommendation algorithms"
- Improvements in Robert M. Bell & Yehuda Koren: "Improved Neighborhood-based Collaborative Filtering"
- Miklós Kurucz, András A. Benczúr, Tamás Kiss, István Nagy, Adrienn Szabó & Balázs Torma: "Who Rated What: a Combination of SVD, Correlation and Frequent Sequence Mining"
- Jorge Sueiras: "A classical predictive modeling approach for Task "Who rated what?" of the KDD CUP 2007"