Item-based Collaborative Filtering

Paper presentation

Martin Krüger, Sebastian Kölle

28.04.2011

Seminar Collaborative Filtering

KDD Cup 2011: Aufgabenbeschreibung Track 1

Item-based Collaborative Filtering Recommendation Algorithms

Improved Neighborhood-based Collaborative Filtering

Verwendung des Algorithmus im KDD Cup 2011

KDD Cup 2011 – Track 1

Ziel Vorhersage von Nutzerbewertungen für Musikstücke, Alben, Künstler und Genres unter Verwendung bereits gegebener Bewertungen

Bewertung Root mean square error (RMSE)

$$\sqrt{\frac{\sum_{i=1}^{n} (r_i - \hat{r}_i)^2}{n}}$$

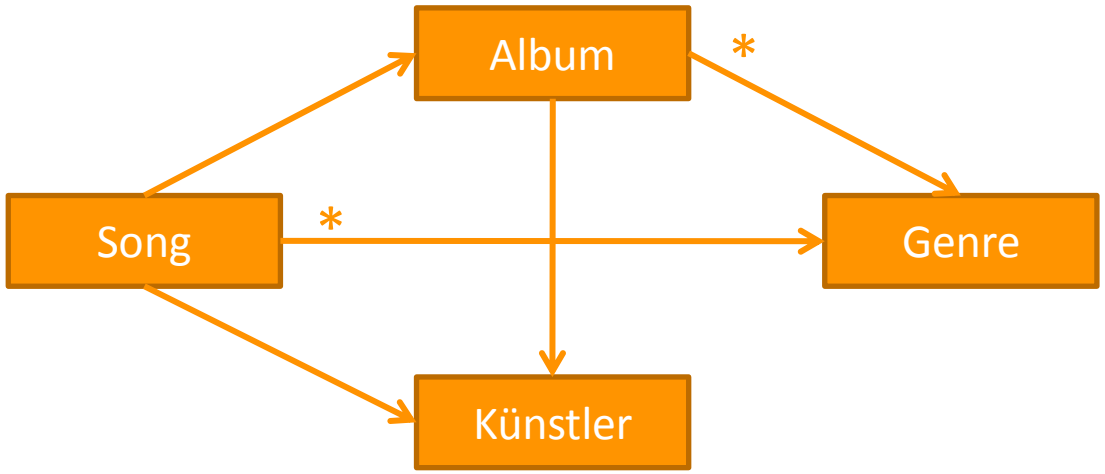
[4]

KDD Cup 2011 – Track 1

Daten

1.000.990	Nutzer
624.961	Items
262.810.175	Bewertungen (Training + Validierung + Test)

Items



[4]

KDD Cup 2011 – Track 1

Daten

1.000.990	Nutzer
624.961	Items
262.810.175	Bewertungen (Training + Validierung + Test)

Bewertungen

- Nutzer-ID
- Item-ID
- Bewertung [0; 100]
- Datum
- Uhrzeit

[4]

KDD Cup 2011: Aufgabenbeschreibung Track 1

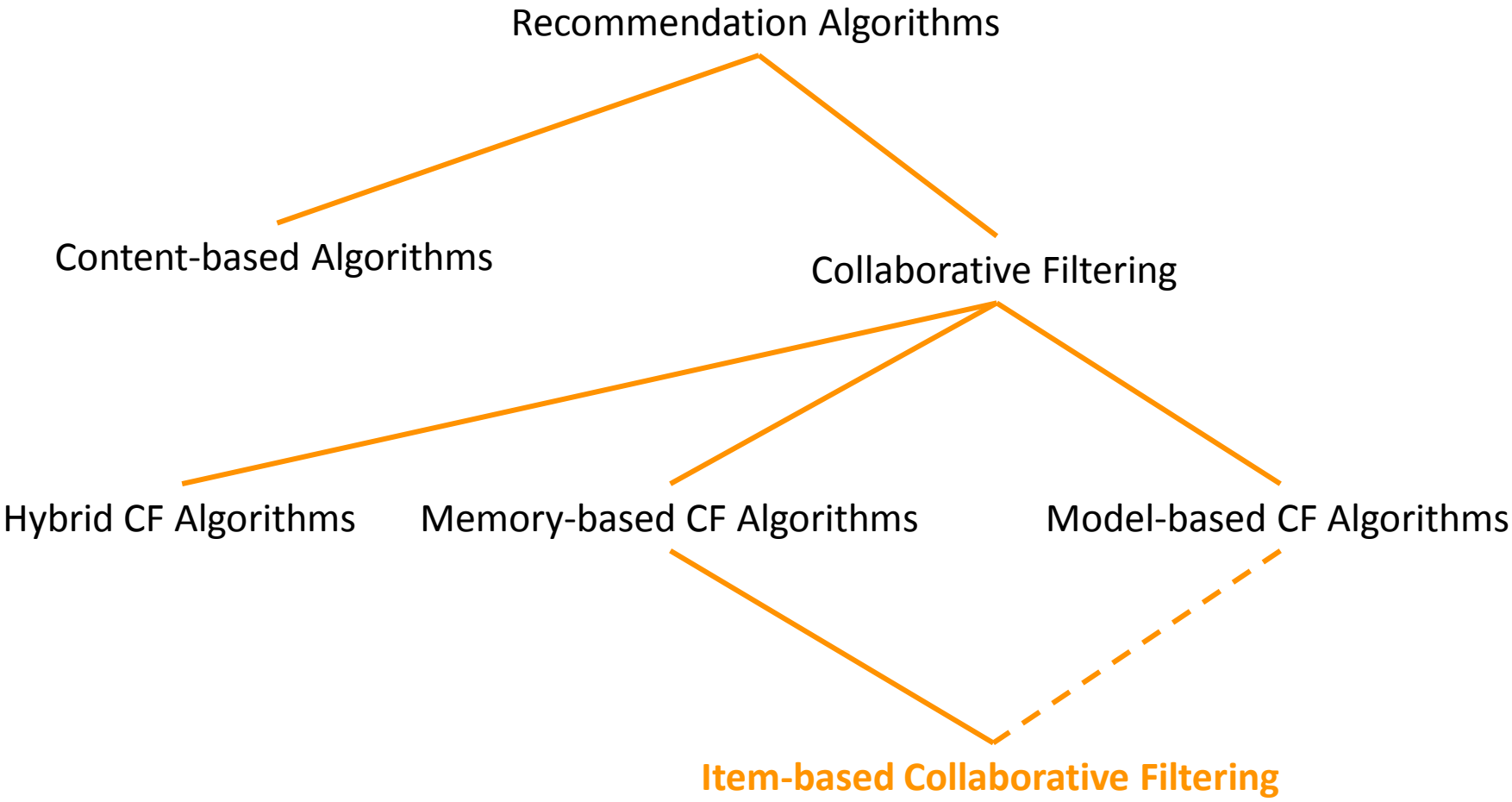
Item-based Collaborative Filtering Recommendation Algorithms

- Einordnung des Ansatzes
- Beschreibung des Algorithmus
- Vergleich: Item-based und User-based CF

Improved Neighborhood-based Collaborative Filtering

Verwendung des Algorithmus im KDD Cup 2011

Einordnung des Ansatzes



[1, S. 8], [3, S. 3]

Beitrag der Arbeit – Item-based CF

„Ein Benutzer bewertet ein neues Item so, wie er ähnliche Items schon früher bewertet hat.“

- Analyse des Item-based Algorithmus

- Vergleich verschiedener Implementierungen
  - Ähnlichkeitsmaße
  - Bewertungsermittlung
  - Größe der Nachbarschaft
  - Modellgröße

- Vergleich der Ergebnisse mit bekannten User-based Verfahren

[1, S.3]

Beschreibung des Algorithmus

Vorberechnung

Erstelle eine Item-Item Matrix, berechne dabei die Ähnlichkeit jedes Item-Paares unter Verwendung eines Ähnlichkeitsmaßes (*Cosinus-based*, *Correlation-based* oder *Adjusted Cosine similarity*).

$$s_{ij} = \frac{\sum_{u \in U} (r_{ui} - \overline{r}_u)(r_{uj} - \overline{r}_u)}{\sqrt{\sum_{u \in U} (r_{ui} - \overline{r}_u)^2} \cdot \sqrt{\sum_{u \in U} (r_{uj} - \overline{r}_u)^2}}$$

Vorhersage

Gegeben: User **u**, Item **i**. Gesucht: Rating $\mathbf{r_{ui}}$

1. Finde die **K** zu **i** ähnlichsten Nachbarn **N(i;u)**, die von **u** bewertet wurden.
2. Berechne den gewichteten Mittelwert auf Basis der Ähnlichkeiten oder berechne das Rating mit einem Regressionsmodell.

$$r_{ui} = \frac{\sum_{j \in N(i;u)} s_{ij} r_{uj}}{\sum_{j \in N(i;u)} |s_{ij}|}$$

[1, S. 6 ff]

Beispiel – Vorberechnung

	User_1	User_2	User_3	User_4
Item_1	2	3	4	3
Item_2	1	3	5	2
Item_3	2	3	3	3
Item_4	???	3	4	2

	Item_1	Item_2	Item_3	Item_4
Item_1	0	3		
Item_2	3	0		
Item_3			0	
Item_4				0

Beispiel – Vorberechnung

	User_1	User_2	User_3	User_4
Item_1	2	3	4	3
Item_2	1	3	5	2
Item_3	2	3	3	3
Item_4	???	3	4	2

	Item_1	Item_2	Item_3	Item_4
Item_1	0	3	4	
Item_2	3	0		
Item_3	4		0	
Item_4				0

Beispiel – Vorberechnung

	User_1	User_2	User_3	User_4
Item_1	2	3	4	3
Item_2	1	3	5	2
Item_3	2	3	3	3
Item_4	???	3	4	2

	Item_1	Item_2	Item_3	Item_4
Item_1	0	3	4	2
Item_2	3	0		
Item_3	4		0	
Item_4	2			0

Beispiel – Vorberechnung

	User_1	User_2	User_3	User_4
Item_1	2	3	4	3
Item_2	1	3	5	2
Item_3	2	3	3	3
Item_4	???	3	4	2

	Item_1	Item_2	Item_3	Item_4
Item_1	0	3	4	1
Item_2	3	0	4	1
Item_3	4	4	0	2
Item_4	1	1	2	0

Beispiel – Vorhersage

	User_1	User_2	User_3	User_4
Item_1	2	3	4	3
Item_2	1	3	5	2
Item_3	2	3	3	3
Item_4	???	3	4	2

	Item_1	Item_2	Item_3	Item_4
Item_1	0	3	4	1
Item_2	3	0	4	1
Item_3	4	4	0	2
Item_4	1	1	2	0

Beispiel – Vorhersage

	User_1	User_2	User_3	User_4
Item_1	2	3	4	3
Item_2	1	3	5	2
Item_3	2	3	3	3
Item_4	???	3	4	2

	Item_1	Item_2	Item_3	Item_4
Item_1	0	3	4	1
Item_2	3	0	4	1
Item_3	4	4	0	2
Item_4	1	1	2	0

Beispiel – Vorhersage

	User_1	User_2	User_3	User_4
Item_1	2	3	4	3
Item_2	1	3	5	2
Item_3	2	3	3	3
Item_4	???	3	4	2

	Item_1	Item_2	Item_3	Item_4
Item_1	0	3	4	1
Item_2	3	0	4	1
Item_3	4	4	0	2
Item_4	1	1	2	0

Beispiel – Vorhersage

	User_1	User_2	User_3	User_4
Item_1	2	3	4	3
Item_2	1	3	5	2
Item_3	2	3	3	3
Item_4	1.5	3	4	2

	Item_1	Item_2	Item_3	Item_4
Item_1	0	3	4	1
Item_2	3	0	4	1
Item_3	4	4	0	2
Item_4	1	1	2	0

Vergleich Item-based und User-based CF

Sparsity

User-based: Die Nutzer, welche als „Jury“ für eine Bewertung herangezogen werden, decken nur einen Teil aller Bewertungen ab.

Item-based: Es können **alle** Bewertungen zur Ähnlichkeitsbestimmung herangezogen werden.

Scalability

- Potentiell weniger Items als User im System
- Beziehungen zwischen Items statischer als zwischen Usern
  - z.B. neue Bewertungen eines Nutzers haben große Auswirkungen auf seine „Jury“-Auswahl, aber kaum auf die Ähnlichkeit von Items.
  - Vorberechnung der Ähnlichkeiten bei Item-based ist möglich.

[1, S. 5], [3, S. 2ff]

KDD Cup 2011: Aufgabenbeschreibung Track 1

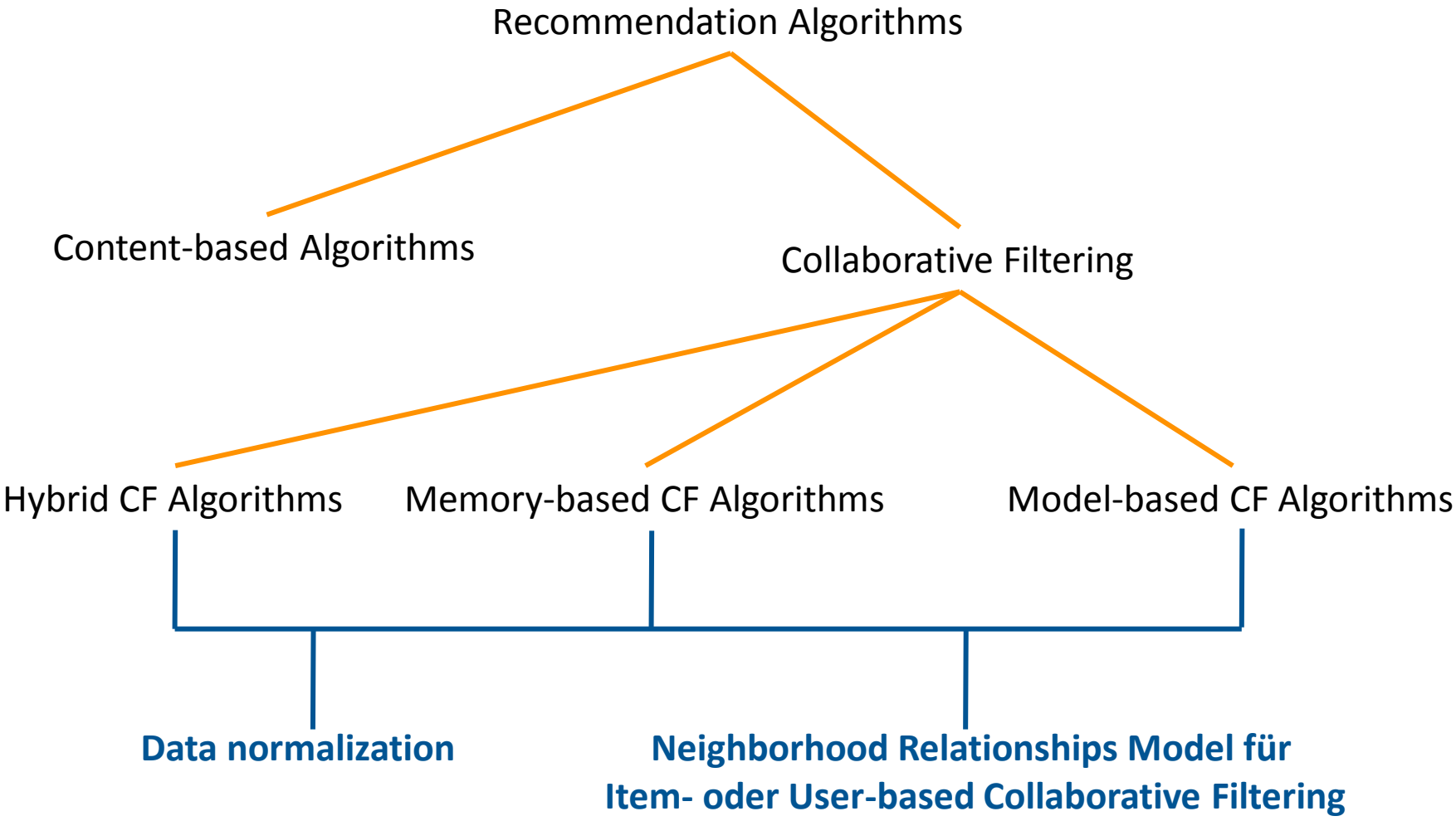
Item-based Collaborative Filtering Recommendation Algorithms

Improved Neighborhood-based Collaborative Filtering

- Einordnung des Ansatzes
- Beitrag der Arbeit
- Entfernen globaler Effekte
- Neighborhood relationships model

Verwendung des Algorithmus im KDD Cup 2011

Einordnung des Ansatzes



[1, S. 3f], [2, S. 8], [3]

Beitrag der Arbeit (I)

Zwei einheitliche Lösungsverfahren für Probleme, die beim Collaborative Filtering auftreten

1. **Daten in der User-Item-Matrix sind nicht normalisiert**

Beispiel User A bewertet Filme nur im Bereich [0; 40],
User B ausschließlich im Bereich [60; 100]
→ Bewertungen der User A und B sind nicht vergleichbar

Beitrag Ein Verfahren, um verschiedene „globale Effekte“ aus den Daten zu entfernen, bevor der eigentliche CF-Algorithmus angewandt wird.

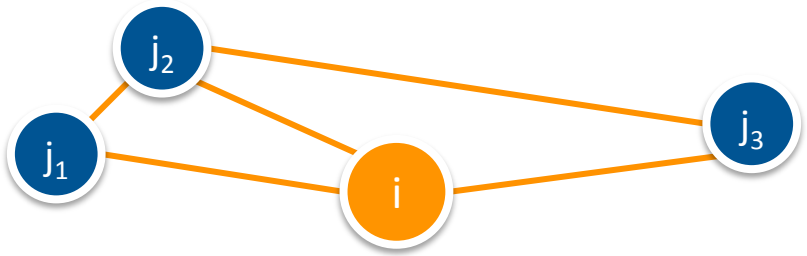
[2, S. 9f]

Beitrag der Arbeit (II)

Zwei einheitliche Lösungsverfahren für Probleme, die beim Collaborative Filtering auftreten

2. **Wahl der Gewichte zur Prognose eines Ratings auf Basis seiner Nachbarn**

$$r_{ui} \leftarrow \sum_{j \in N(i;u)} w_{ij} r_{uj} \qquad w_{ij} = \frac{s_{ij}}{\sum_{k \in N(i;u)} |s_{ik}|}$$



Beitrag Eine Modellierung der gesuchten Gewichte durch Angabe eines geeigneten Optimierungsproblems und ein Verfahren zum bestimmen der Gewichte.

[2, S. 10ff]

Entfernen globaler Effekte (I)

Beispiele

User effect	ein User bewertet im Durschnitt schlechter als andere
Movie effect	ein Film wird im Durchschnitt besser bewertet als andere

Problem Die einzelnen Wertungen sind nicht vergleichbar!

Effect	RMSE	Improvement
Overall mean	1.1296	NA
Movie effect	1.0527	.0769
User effect	0.9841	.0686
User×Time(user)$^{1/2}$	0.9809	.0032
User×Time(movie)$^{1/2}$	0.9786	.0023
Movie×Time(movie)$^{1/2}$	0.9767	.0019
Movie×Time(user)$^{1/2}$	0.9759	.0008
User×Movie average	0.9719	.0040
User×Movie support	0.9690	.0029
Movie×User average	0.9670	.0020
Movie×User support	0.9657	.0013

[2, S. 9f]

Entfernen globaler Effekte (II)

$$r_{ui}(0) = r_{ui}$$

Rating von User **u** für Item **i**

$$r_{ui}(n) = r_{ui}(n-1) - \text{effect}_n$$

$$\text{effect}_n = \Theta_u x_{ui} + \text{error}$$

„Variable of interest“

Beispiele **x_{ui} = 1**

x_{ui} = sqrt(Anzahl Tage zwischen der ersten Bewertung von **u** und der Bewertung für **i** (zentriert für **u**)**)**

$$\Theta_u = \frac{n_u \cdot \sum_i r_{ui}(n-1)x_{ui}}{(n_u + \alpha) \cdot \sum_i x_{ui}^2}$$

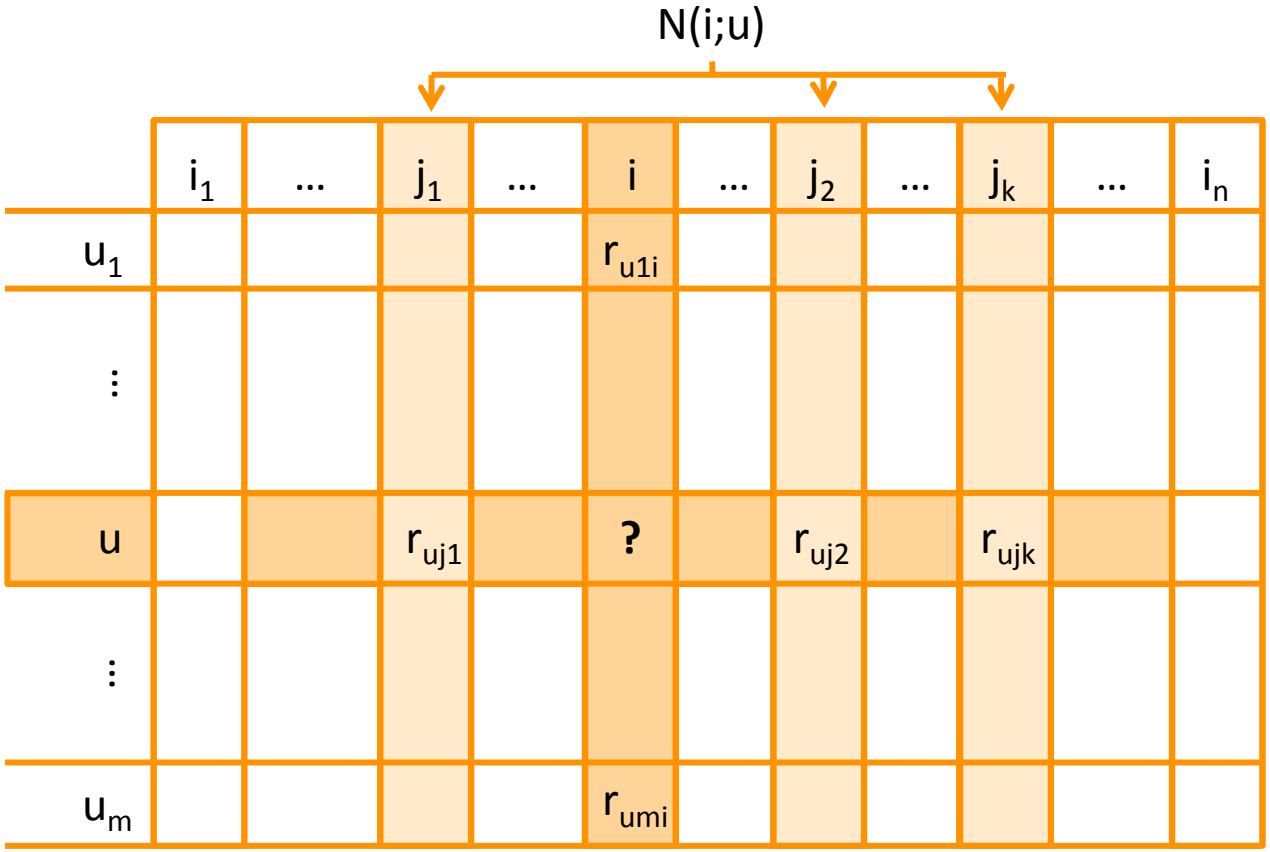
Anzahl der Ratings von User **u**

Verwendung für

- die Auswahl der Nachbarn
- die Berechnung des Ratings (→ globale Effekte müssen wieder addiert werden!)

[2, S. 9f], [7]

Neighborhood relationships model



$$r_{ui} \leftarrow \sum_{j \in N(i;u)} w_{ij} r_{uj}$$

$$\min_{w} \sum_{v \neq u} \left(r_{vi} - \sum_{j \in N(i;u)} w_{ij} r_{vj} \right)^2$$

[2, S. 10ff]

KDD Cup 2011: Aufgabenbeschreibung Track 1

Item-based Collaborative Filtering Recommendation Algorithms

Improved Neighborhood-based Collaborative Filtering

Verwendung des Algorithmus im KDD Cup 2011

Herausforderungen für den KDD Cup

- Großer Speicherbedarf: mindestens 364 GB für vollständige Speicherung der Item-Item- und Â-Matrix (ein Byte pro Paar – sehr optimistisch …)
  - Sampling für die Entwicklung!
  - Space-time-tradeoff?
- Berücksichtigung der Hierarchie
  - Reduzierung des Speicheraufwandes: Vergleich ausschließlich Items gleichen Genres?
  - Verbesserung der Vorhersagen: Macht z.B. die gute Bewertung eines Albums durch einen Nutzer es wahrscheinlicher, dass er auch die einzelnen Titel gut bewertet?
- Herausfinden der optimalen Parameterwerte für die gegebenen Daten
  - Testumgebung, die das systematische Ausprobieren mit verschiedenen Samples erlaubt

Quellen

[1] Sarwar B., et al: *Item-based Collaborative Filtering Recommendation Algorithms*, Proceedings of the 10th international conference on World Wide Web ACM, 2001

[2] Bell, R. M., Koren, Y.: *Improved Neighborhood-based Collaborative Filtering* Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining ACM, 2007

[3] Su, X., Khoshgoftaar, T. M.*: A survey of collaborative filtering techniques* Advances in Artificial Intelligence, Vol. 2009 Hindawi Publishing Corporation, 2009

[4] KDD Cup Website, http://kddcup.yahoo.com/datasets.php#

[5] Website von Yehuda Koren, http://research.yahoo.com/Yehuda_Koren

[6] Website von Robert M. Bell, http://www.research.att.com/people/Bell_Robert_M/index.html

[7] Freeman, L.: Algorithm Analysis Blog, http://algorithmsanalyzed.blogspot.com/

Anhang

Vergleich Item-based und User-based CF (II)

Cold-Start Problem

User-based: Mindestanzahl an Bewertungen nötig, um ähnliche User bestimmen zu können.

Item-based: Mindestanzahl an Bewertungen in der Nachbarschaft eines Items notwendig, um Vorhersage treffen zu können.

[1, S. 5], [3, S. 2ff]

Die Autoren

Robert M. Bell
AT&T Labs Research

PHD in Statistik

Yehuda Koren
Yahoo! Research Israel
(vorher AT&T Labs Research)

KDD 2009 best research paper award: collaborative filtering with temporal dynamics

- Mitglieder der Teams „BellKor“ / „BellKor Pragmatic Chaos“:
  - Netflix Progress Price / Netflix Grand Prize
- Co-Organisatoren des aktuellen KDD Cups

[5],[6]

Bewertung der Arbeiten

- Item-based Collaborative Filtering Recommendation Algorithms
  - Fehlende Erklärung, wie das Model für die Experimente erzeugt wird

- Improved Neighborhood-based Collaborative Filtering
  - Ungenaue Beschreibung der Algorithmus zur Berechnung der globalen Effekte (*residuals* und *ratings*)
  - Nicht nachvollziehbare Schritte in der Herleitung von Formeln
  - Fehlende Beschreibung, wie die beiden Beiträge der Arbeit zusammenspielen
  - Verweise für verwendete Formeln/Algorithmen: komplette Bücher