Recommender-Systeme
Teil 2

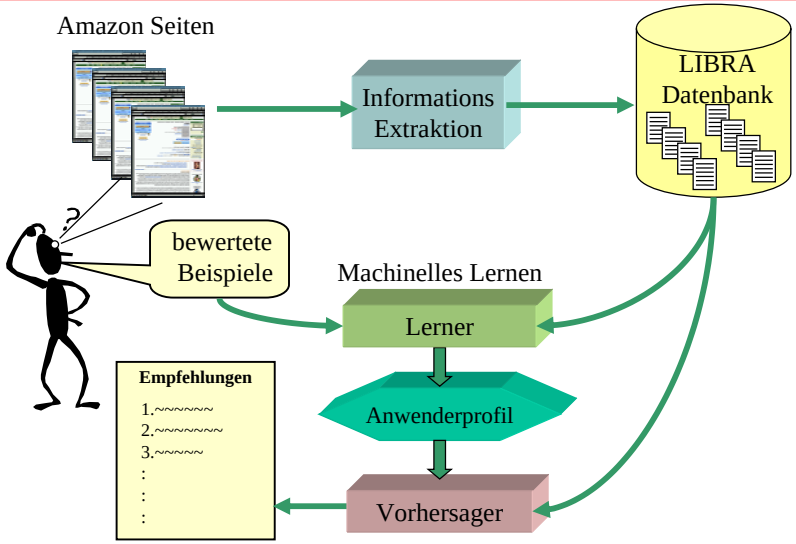
Kollaboratives Filtern &
inhaltsbasierte Empfehlungen

LIBRA
Learning Intelligent Book Recommending Agent

- Inhaltsbasierter Recommender für Bücher, der Informationen über Titel verwendet, die von Amazon extrahiert wurden.
- Nutzt Informations-Exktration-Methoden, um Text aus den folgenden Feldern zu bestimmen:
  - Autor
  - Titel
  - Redaktionnelle Reviews
  - Stellungnahmen von Kunden
  - Themenbezeichnungen (Schlagworte)
  - Zugehörige Autoren
  - Zugehörige Titel

http://hyperion.cs.utexas.edu:8090/

LIBRA System



Libra Übersicht

- Anwender bewertet ausgewählte Titel auf einer Skala von 1 bis 10.
- Libra verwendet einen naïven Bayesian Text-Kategorisierungs-Algorithmus, um ein Profil aus diesen bewerteten Beispielen zu erlernen.
  - Bewertung 6–10: positiv
  - Bewertung 1–5: negativ
- Das erlernte Profil wird dazu verwendet, alle anderen Bücher als Empfehlungen zu klassifizieren, wenn die posterior Wahrscheinlichkeit diese positiv klassifiziert.
- Der Anwender kann auch explizite positive/negative Schlüsselwörter liefern, die als Prior verwendet werden, um die Rolle dieser Merkmale bei der Kategorisierung zu beeinflussen.

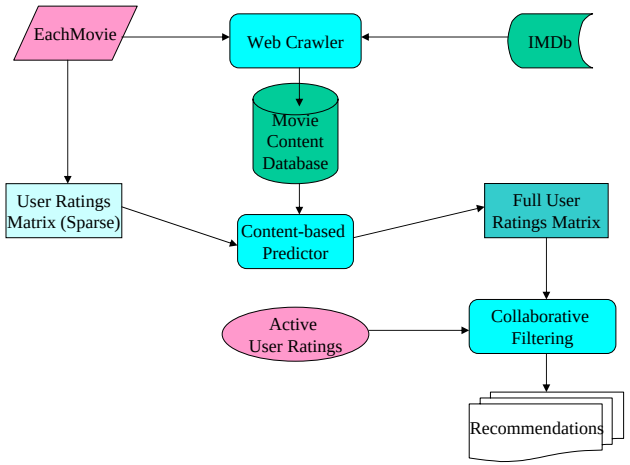
Verbinden von Inhalt und Kollaboration

- Inhaltsbasierte und kollaborative Methoden haben komplementäre Stärken und Schwächen.
- Kombiniere Methoden, um das Beste von beiden zu erhalten.
- Verschiedene hybride Ansätze:
  - Wende beide Methoden an und verknüpfe Empfehlungen.
  - Verwende kollaborative Daten als Inhalt.
  - Verwende inhaltsbasierte Vorhersage als weiteres Element beim kollaborativen Filtern.
  - Verwende eine inhaltsbasierte Vorhersage um kollaborative Daten zu vervollständigen.

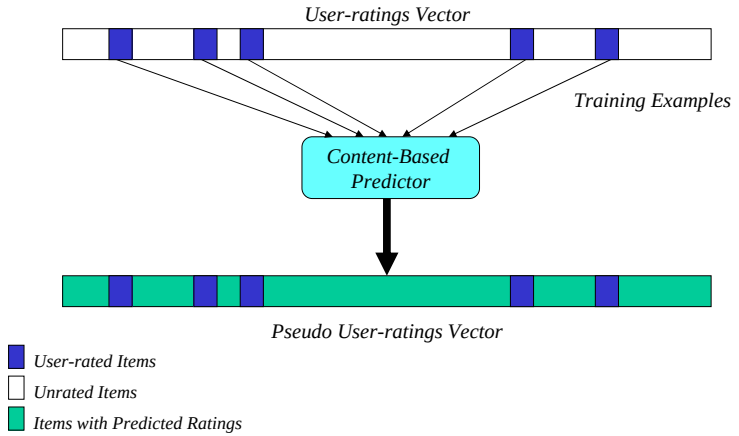
Movie Domain

- *EachMovie* Datensatz [Compaq Research Labs]
  - enthält Anwenderbewertungen für Filme auf einer Skala von 0–5.
  - 72,916 Anwender (durchschn. jeder 39 Bewertungen).
  - 1,628 Filme.
  - Spärliche besetzte Anwender-Bewertungsmatrix – (2.6% voll).
- Internet Filmdatenbank gecrawlt (*IMDb*)
  - Für jeden Film im *EachMovie* Datensatz gibt es einen Link zur IMDb, der zum Extrahieren weiterer Informationen genutzt wird
- Wesentliche Filminformationen sind:
  - Titel, Direktor, Rollenbesetzung, Genre, etc.
- Populäre Meinungen:
  - Anwenderstellungnahmen, Zeitungen und Newsgroup Reviews, etc.

Content-Boosted kollaboratives Filtern



Content-Boosted CF - I



Content-Boosted CF - II

- Berechne Pseudo Anwenderbewertungsmatrix
  - Full matrix – approximates actual full user ratings matrix
- Führe CF aus:
  - Unter Verwendung der Pearson Corr. zwischen pseudo Anwender-Bewertungsvektoren

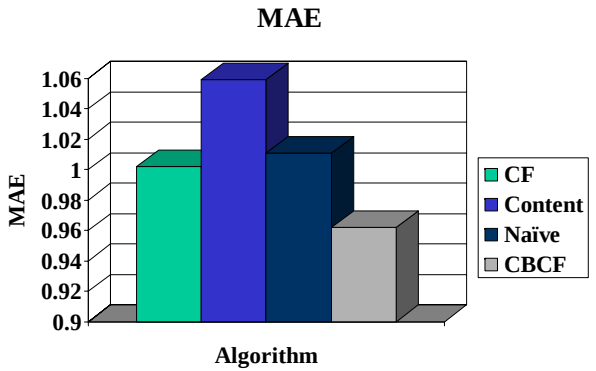
Experimentelle Methode

- Verwendete Subset jedes *EachMovie* (7,893 Anwender; 299,997 Bewertungen)
- Testmenge: 10% der Nutzer zufällig ausgewählt.
  - Testnutzer, die mindesten 40 Filme bewerteten.
  - Trainiere mit der restlichen Menge.
- Hold-out Menge: 25% Objekte jedes Testnutzer.
  - Sage die Bewertung für jedes Objekt in der Hold-out Menge voraus.
- Vergliche CBCF mit anderen Vorhersagesansätzen:
  - Reines CF
  - Rein inhaltsbasiertes CF
  - Naïve hybrid (mittle CF und inhaltsbasierte Vorhersage)

Maße

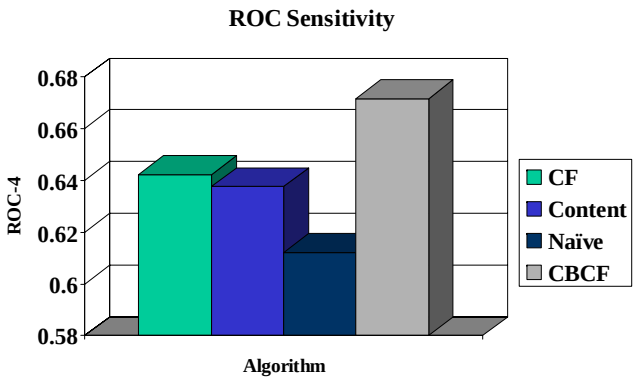
- Mittlerer absoluter Fehler (MAE)
  - Vergleicht numerische Vorhersagen mit Anwenderbewertungen
- ROC Empfindlichkeit [Herlocker 99]
  - Wie gut hilft die Vorhersage dem Nutzer *high-quality* Objekte auszuwählen
  - Bewertungen ≥ 4 werden als “gut” und < 4 als “schlecht” betrachtet
- Paired T-test als statistischen Signifikanztest

Ergebnisse - I



CBCF ist bedeutend besser (4% über CF) bei ($p < 0.001$)

Ergebnis - II



CBCF übertrifft Rest (5% Verbesserung über CF)

Aktives Lernen

- Wird verwendet, um die Anzahl der Trainingsbeispiele zu verringern.
- System fordert Bewertungen für spezifische Objekte, von denen es am meisten lernen würde.
- Verschiedene existierende Methoden:
  - Uncertainty sampling
  - Komitee-basierte Stichproben

Halb-überwachtes Lernen
(weakly supervised, Bootstrapping)

- Verwende die Fülle ungekennzeichneter Beispiele, um das Lernen durch eine kleine Menge von gekennzeichneten Daten zu unterstützen.
- Einige neue Methoden entwickelt:
  - Halb-überwacht EM (Erwartungsmaximierung)
  - Ko-Training
  - Transduktive SVM's

Schlussfolgerungen

- Empfehlen und Personalizierung sind wichtige Ansätze zur Bekämpfung der Informations-Überlast.
- Machinelles Lernen ist ein wichtiger Teil der Systeme für diese Aufgaben.
- Kollaboratives Filtern hat Probleme.
- Inhaltsbasierende Methoden sprechen diese Probleme an (haben aber eigene Probleme).
- Das Beste ist, beides zu integrieren.