Recommender System auf Basis von Collaborative Filtering

Alex Egger

3. Juni 2015

Inhaltsverzeichnis

1 Einleitung

Empfehlungsdienste (engl. Recommender Systems, im Folgenden RS) sind Softwaresysteme, die zum Ziel haben, Vorhersagen über das Verhalten oder die Präferenz eines Benutzers zu treffen. Ihre Einsatzgebiete sind zahlreich und vielfältig. Ein berühmtes Beispiel stellt **Amazons Recommendation Engine** dar, welche bereits seit einigen Jahren erfolgreich im Einsatz ist. Weitere berühmte Beispiele stellen z.B. Netflix, oder MovieLens dar.

2 Theoretischer Hintergrund

RS sind eine Kategorie von Softwaresystemen, welche zum Oberbegriff **"Maschinelles Lernen"** gezählt werden.

> **Maschinelles Lernen** ist ein Oberbegriff für die "künstliche" Generierung von Wissen aus Erfahrung: Ein künstliches System lernt aus Beispielen und kann nach Beendigung der Lernphase verallgemeinern. Das heißt, es werden nicht einfach die Beispiele auswendig gelernt, sondern es "erkennt" Gesetzmäßigkeiten in den Lerndaten. So kann das System auch unbekannte Daten beurteilen (Lerntransfer).[1]

Es gibt unterschiedliche Ansätze zur Realisierung eines RS. Die drei wichtigsten Ansätze stellen folgende dar:

- Collabortive Filtering
- Content-Based Filtering
- Hybrid Recommender Systems

Im folgenden wird die Funktionsweise eines RS erklärt, welches auf Collaborative Filtering basiert.

2.1 Collaborative Filtering

Collaborative Filtering (im Folgenden CF) beschreibt den Prozess des Filterns von Informationen basierend auf der Zusammenarbeit mehrerer Datenquellen. CF-basierte Ansätze werden meist bei großen Datenmengen eingesetzt, so z.B. im Finanzbereich, oder im Bereich des E-Commerce.

2.1.1 Funktionsweise

CF basiert auf der fundamentalen Annahme, dass Personen sich in ihren Vorlieben ähneln. Dies bedeutet, wenn zwei Personen ähnliche Vorlieben haben, kann man annehmen, dass dies auch in Zukunft so bleiben wird. Dieses Prinzip nutzt CF aus, um ähnliche Personen zu finden, und ihnen die Präferenzen des jeweils anderen vorzuschlagen. Desweiteren ist meist eine aktive Teilnahme des Benutzers am Quantifizierungsprozess unerlässlich. Ein typischer Aufbau eines CF-basierten Systems könnte wie folgt aussehen:

[1] Von http://de.wikipedia.org/wiki/Maschinelles_Lernen, Stand: 28.05.15

1. Ein Benutzer drückt seine Präferenz bezüglich eines angebotenen Objektes aus. Dies kann z.B. eine Bewertung auf einer Skala (z.B: 1 - 5), oder das Drücken eines Buttons (z.B: 0 - 1) sein.

2. Das System vergleicht die Bewertung des Benutzers mit denen der anderen Benutzer und findet jene mit den ähnlichsten Vorlieben.

3. Das System findet ein Objekt, welches ähnliche Benutzer mit einer hohen Bewertung gekennzeichnet haben und welches noch nicht vom Benutzer bewertet wurde.

Desweiteren existieren zwei Funktionweisen von Collaborative Filtering, die unterschiedliche Sichtpunkte auf die gegebenen Daten haben.
User-Based Collaborative Filtering (im Folgenden UBCF) kann durch die folgenden zwei Schritte erklärt werden:

1. Man finde Benutzer, welche die selben Bewertungmuster haben, wie der aktuelle Benutzer.

2. Man nutze die Bewertungen der in Schritt 1 gefundenen Benutzer, um einen Vorschlag für den aktuellen Benutzer zu berechnen.

Item-Based Collaborative Filtering (im Folgenden IBCF) verhält sich ähnlich wie UBCF, aber in einer objekt-zentrischen Verhaltensweise:

1. Man erstelle eine Objekt-zu-Objekt-Matrix, welche Beziehungen zwischen Objektpaaren beschreibt.

2. Man erfährt die Vorlieben des aktuellen Benutzers, indem man seine Bewertungen mit der Matrix abgleicht.

2.1.2 Ähnlichkeit

In **UBCF** ist es meist nötig festzustellen, wie ähnlich sich zwei Benutzer sind. Davor werden verschiedene Methoden zur Festellung der Ähnlichkeit zweier Mengen eingesetzt.

Euklidischer Abstand Der **Euklidsche Abstand** zweier Punkte in einem n-dimensionalen Raum wird wie folgt berechnet:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

Jaccard-Index Eine weiter Möglichkeit besteht im **Jaccard-Index (auch Jaccard-Koeffizient**. Er ist wie folgt definiert:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

Kosinus-Ähnlichkeit **Kosinus-Ähnlichkeit** ist ein Maß für die Ähnlichkeit zweier Vektoren. Sie beschreibt den Kosinus des Winkels zwischen beiden Vektoren und nimmt damit natürlich bei gleichen Vektoren den Wert 1 an. Für Vektoren die weiter von einander entfernt sind, nimmt der Wert immer weiter ab. Sie ist wie folgt definiert:

$$\cos(\phi) = \frac{a \cdot b}{\|a\| \cdot \|b\|} \tag{3}$$

2.1.3 Probleme

CF wird meist in Umgebungen mit sehr großen Datenmengen eingesetzt. Dies führt dazu, dass neue Benutzer erst genügend Bewertungen abgeben müssen, bevor für sie sinnvolle Vorhersagen getroffen werden können. Das gleiche gilt für neue Objekte, welche noch nicht genügend Bewertungen haben.[2] Ein weiteres Problem stellt die Skalierung von CF-basiertes System dar. Da solche Systeme oft im Web-Bereich eingesetzt werden, müssen sie sehr hohen Anforderungen an die Reaktionszeit entsprechen.

3 Implementierung

3.1 Datenquellen

3.2 Auswahl der Filter-Methode

Das gewählte Framework bietet verschiedene Methoden zur Filterung. Folgende sind wählbar:

- **UBCF** - User-Based Collaborative Filtering
- **IBCF** - Item-Based Collaborative Filtering
- **POPULAR** - Filtert basierend auf der Popularität von Objekten
- **RANDOM** - Filtert Objekte zufällig

Im Folgenden wurden die verschiedenen Ansätze bei gleichen Bedingungen und Daten eingesetzt. In Abbildung 1 ist zu sehen, wie sich die verschiedenen Methoden verhalten. **TPR (True Positive Rate)** beschreibt hier bei die Rate an Treffern, die für den Benutzer wirklich relevant sind. **FPR (False Positive Rate)** beschreibt die Rate an Treffern, die nicht relevant sind. Die Datenpunkte beschreiben dabei die Anzahl der zusuchenden Objekte. Aus Abbildung 1 kann somit abgelesen werden, dass **UBCF** sowohl die beste **TPR**, als auch die niedrigste **FPR** hat. **UBCF** wurde somit als Algorithmus zur Filterung für das RS ausgewählt.

[2] Siehe: http://en.wikipedia.org/wiki/Cold_start

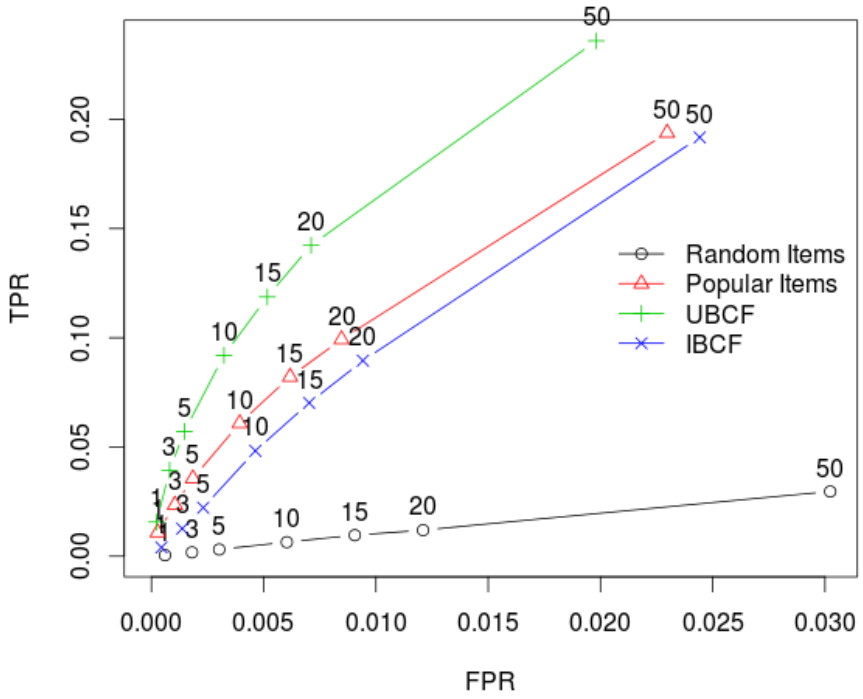

Abbildung 1: Vergleich der Trefferquoten verschiedener Filter-Methoden.

4 Quellen