14. November 2013

DATA MINING

- Versuch 3: Recommender Systeme mit Collaborative Filtering -

Teammitglieder

JanHorak

Ralf Palyov

Andreas Mayer

2 Durchführung Teil 1: Fiktive Filmbewertung

2.2 Ähnlichkeitsbestimmung

1 Welche Bedeutung hat der Übergabeparameter normed in der Funktion sim_euclid?

Die euklidische Distanz wird immer größer, aus je mehr Dimensionen die zu vergleichende Vektoren bestehen. Das ist der Fall, da eine Summe über die Differenzen der Werte der einzelnen Dimensionen gebildet wird.

Dieses Verhalten entspricht nicht unbedingt der Realität, da die Ähnlichkeit auf die Weise geringer wird, selbst wenn eine Dimension hinzu kommt deren Differenz genau dem bisherigen Differenzdurchschnitt der einzelnen Dimensionen entspricht.

Um dem entgegen zu wirken, wird hier durch die Anzahl aller Dimensionen (mit Wert ungleich 0) geteilt. Dadurch erhält man eine Distanz, die unabhängig von der Dimensionalität der Eingabevektoren ist.

2 Keine Frage für das Protokoll

3 Vergleichen Sie die beiden Ähnlichkeitsmaße. Welches Ähnlichkeitsmaß erscheint Ihnen für diesen Anwendungsfall sinnvoller und warum?

Die Pearson Korrelation ist hier besser geeignet um die Distanz zwischen den Vektoren zu berechnen.

Das liegt daran, dass die Pearson Korrelation, im Gegensatz zur euklidischen Distanz die Länge der Vektoren nicht berücksichtigt. Dies ist hier sinnvoll, da nicht jeder Kritiker die gleiche Messlatte für Filme hat. Das bedeutet, dass manche Kritiker Filme prinzipiell schlechter bewerten als andere Kritiker. Eigentlich interessiert aber wie gut ein Kritiker einen Film im Vergleich mit den anderen Filmen bewertet; Die Verwendung der absoluten Bewertungen verzerren das Ergebnis.