Prof. Dr. Franz Schweiggert
Ingo Melzer

Software Engineering Praxis (SS 2002)

Abgabetermin 17. Mai 2002 vor den Übungen

Was mag ich denn? (20 Punkte)

Es ist spannend, das Verhalten oder die Meinung von Personen durch Vergleich mit anderen zu schätzen. Dazu benötigt man eine Datenbank mit Wertung von Personen. Es sei im Folgenden $v_{i,j}$ die Wertung des Benutzers i zum Thema j. Ist I_i die Menge aller von i abgegebenen Wertungen, so ist seine mittlere Wertung $\bar{v}_i = \frac{1}{|I_i|} \sum_{j \in I_i} v_{i,j}$. Möchte man nun die Wertung der aktiven Person a zu j, also $p_{a,j}$ vorhersagen, so berechnet man $p_{a,j} = \bar{v}_a + \kappa \sum_{i=1}^{n} \omega(a,i)(v_{i,j} - \bar{v}_i)$. Die Gewichte $\omega(a,i)$ geben die Korrelation zwischen a und i wieder. Im Folgenden wählen wir $\kappa = \frac{1}{\sum_{i=1}^{n} |\omega(a,i)|}$. Eine der gebräuchlichsten Formeln zur Berechnung der Korrelation zwischen zwei Personen a und i ist der Pearson-Korrelationskoeffizient:

$$\omega(a,i) = \frac{\sum_k (v_{a,k} - \bar{v}_a)(v_{i,k} - \bar{v}_i)}{\sqrt{\sum_k (v_{a,k} - \bar{v}_a)^2 \sum_k (v_{i,k} - \bar{v}_i)^2}}$$

Die Summation über k geht über alle Themen, zu denen die Wertung von a und i vorliegt. Obiges wollen wir an einer kleinen Filmdatenbank testen. Auf unserem FTP-Server liegt eine Liste von Filmen (*films*). Zu etwa zehn dieser Filme sollte jeder Teilnehmer der Übungen bitte seine Meinung (Zahl 1 (schlecht) bis 10 (sehr gut)) abgeben. Rufen Sie dazu bitte das Skript `/www/turing/htdocs/sai/ss02/sep/addvote.pl` auf. Als Parameter muss der genaue Titel des Films und die Wertung angegeben werden. Alle Wertung werden in der Datei votes im gleichen Verzeichnis abgelegt.

Schreiben Sie nun ein Perl-Skript, das $p_{a,j}$, also die vermutliche Meinung der Person a zum Film j, mittels obiger Formel berechnet. Verwenden Sie die Option `-w`, deklarieren Sie alle Variablen und nutzen Sie keine globalen Variablen.

Obige Formeln stammen aus einem Paper von John Breese, David Heckermann und Carl Kadie mit dem Titel „Empirical Analysis of Predictive Algorithms for Collaborative Filtering“ von 1998.

Noch ein paar Tipps:

- Als Datenstruktur empfiehlt es sich, die Wertung der einzelnen Personen zu den Filmen als Hash von Hashes zu speichern.
- Es können Wertungen über 10 entstehen. Diese können einfach auf 10 gesetzt werden.
- Berechnen Sie die einzelnen Teile in eigenen Prozeduren.
- Der Befehl `addvote.pl 'Casablanca (1942)((aka Casablanca (1942)))' 10` sollte 10 Punkte bei Casablanca eintragen. Einfach mit Copy/Paste den Filmtitel aus *films* kopieren (im xterm mit mittlerer Maustaste). Achtung bei Hochkommata:
  `addvote.pl 'Schindler'\''s List (1993)((aka Schindlers Liste (1994)))' 8`
- Potenzen kann man in Perl mit `**` berechnen. Daher gilt: `2 ** 10 == 1024`
- Bitte gute und schlechte Filme bewerten. Lauter identische Eintragungen einer Person helfen der Vorhersage fast nicht.