# Graaf voor suggesties

Tekst over de graaf in het algemeen..

## Initialisatie van de graaf

De graaf wordt op twee manieren geïnitialiseerd: via de rating data uit de Yahoo dataset en via de voorkeuren die de gebruiker zelf doorgeeft.

### Via Yahoo

In de Yahoo dataset zit een klein gedeelte van de waarderingen die Yahoo gebruikers hebben gegeven aan muzieknummers. Elke gebruiker uit de dataset heeft minstens 10 songs een waardering toegekend. De waardering bestaat uit een getal tussen de 1 en 5. Een 1 geeft hierbij een lage waardering aan, een 5 juist een hele hoge.

De songs zelf hebben willekeurige nummers gekregen, waardoor deze niet zijn te herleiden naar een daadwerkelijke artiest en nummer of MBID. Wel is er een koppeltabel toegevoegd waarin de songs zijn gekoppeld aan een bepaald genre. Deze informatie kan worden gebruikt om de genres uit de graaf te initialiseren.

Hiervoor is eerst een gedeelte van de data (30 miljoen waarderingen) uit de tekstbestanden van Yahoo via een Java programma ingelezen. Deze data is vervolgens in een database gezet. In de database worden de volgende gegevens opgeslagen:

* Genre id
* Genre naam
* Het aantal gebruikers dat per genre heeft gestemd
* De waarderingen van alle gebruikers voor een bepaald genre

De gemiddelde waardering wordt hiermee per genre bepaalt door het algoritme dat IMBD gebruikt:

(v ÷ (v+m)) × R + (m ÷ (v+m)) × C

Met:

R = gemiddelde van het genre, ofwel alle waarderingen gedeeld door het aantal gebruikers  
 v = het aantal gebruikers dat heeft gestemd  
 m = minimum aantal stemmen die nodig zijn (wij hebben gekozen voor 1500)  
 C = de gemiddelde waardering van alle genres bij elkaar

Door het bepalen van het gemiddelde op deze manier, worden genres waar maar een beperkt aantal mensen op stemmen (vaak de wat alternatievere) weggelaten. Ook krijgen de genres die juist door heel veel mensen worden beoordeeld, relatief gezien een hogere beoordeling.

De gemiddelde waardering die hiermee wordt berekend, is uiteraard een score tussen de 1 en 5. Deze score wordt met de factor 30 gewogen en vervolgens afgerond. Van al deze gemiddelde waarderingen wordt het gemiddelde genomen. Dit wordt vervolgens van de score afgetrokken.

De wegingen die daaruit komen (ongeveer variërend tussen de +15 en -30) worden opgeteld bij de nodes in de graaf, welke een standaard waarde van 100 hebben.

### Via de gebruikerswaarderingen