# Evaluatie

Om te weten te komen of het ontstane systeem daadwerkelijk goede voorspellingen kan geven aan de gebruikers, is het nodig om het muziek recommendatie systeem te testen.

Daarnaast wilden wij graag weten in hoeverre de 2 gebruikte algoritmes invloed hebben op de graaf. Hiervoor hebben wij eerst de test uitgevoerd door enkel gebruik te maken van het 2e algoritme (sorteren op score). Vervolgens hebben we dezelfde test uitgevoerd met gebruikmaking van het 3e algoritme, welke het 1e en 2e algoritme combineert.

## Opzet van de test

Bij de test wordt er gekeken of de voorkeuren van een bepaald persoon naar voren komen na het invoeren van zijn voorkeuren in het systeem. Aangezien het idee is dat ons systeem steeds betere suggesties geeft naarmate er meer waarderingen zijn gegeven aan de tracks, zullen er meerdere malen waarderingen van de gebruiker ingevuld moeten worden. Uiteraard zal ook het basisprofiel van de testgebruiker ingevoerd worden.

Aangezien wij zelf aan de hand van de genres de artiesten selecteren, en aan de hand van deze artiesten hun beste track selecteren, ligt het voor de hand om te werken met de lijst van favoriete artiesten van de testgebruiker. Via Last.FM is dit gemakkelijk haalbaar door middel van ‘user.getTopArtists’. Daarom hebben we ervoor gekozen om testgebruikers van Last.FM te halen.

Dit resulteert in de volgende concrete opzet van de test:

1. Er wordt een willekeurige gebruiker van Last.FM gekozen, die minstens 3000 artiesten heeft beoordeeld.
2. De favoriete genres van deze gebruiker worden via user.getTopTags opgehaald. Deze worden ingevoerd bij de basis genrevoorkeuren van een nieuwe testgebruiker in ons systeem.
3. Aan de hand hiervan worden er een aantal suggesties gegeven door ons systeem. De artiesten behorend bij de suggesties worden opgezocht in de top 1000 van favoriete artiesten van de testgebruiker via user.getTopArtists. Vervolgens wordt geregistreerd welke suggesties wel en niet voorkomen in de top 1000 van de testgebruiker. Aan de hand hiervan wordt een percentage juiste suggesties bepaalt. Ook werd de positie van de artiest uit de top 1000 van de gebruiker genoteerd, en hiervan een gemiddelde bepaald.
4. De juiste suggesties krijgen een +1 waardering, de rest een -1 waardering.
5. Vervolgens worden stap 3 en 4 vier maal herhaald. Door het geven van de waarderingen zal de lijst suggesties steeds veranderen.
6. Om nog extra variatie aan te brengen, zullen er nu 15 suggesties toegevoegd worden middels de ‘Iets nieuws toevoegen’ optie. Ook hiervan wordt gekeken in hoeverre er suggesties wel of niet voorkomen in de top 1000 artiesten van de gebruiker, en worden er +1 en -1 waarderingen toegekend.
7. Vervolgens wordt nog een laatste maal stap 3 en 4 uitgevoerd. De resultaten hiervan zijn de eindresultaten van de test.

## Resultaten tests

De willekeurige gekozen gebruikers zijn ikea, irviniebrown en lucasm2.

“ikea” houdt vooral van Rock. Daarnaast vind “ikea” het Techno, Pop, Punk en Reggae genre ook wel aardig. Over de andere genres is geen mening bekend.

Gebruiker “irvinebrown” houdt van Rock, Soul en Jazz. Daarnaast is er voor bijna alle andere genres bekend hoe leuk hij deze vind.

De derde gebruiker, “lucasm2”, is een typische Amerikaan. Hij houdt van Rock, Punk en Techno. Daarnaast vind hij Reggae en Electronic ook erg leuk.

Deze 3 gebruikers kregen de volgende resultaten tijdens de twee tests:

**Ikea**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **2e algoritme** | | | **1e + 2e algoritme** | | |
| **Actie** | **Juist** | **Percentueel** | **Gemiddelde** | **Juist** | **Percentueel** | **Gemiddelde** |
| 1: Suggesties beoordelen | 4 | 4/19 = 21% | 759.75 | 4 | 4/19=21% | 545 |
| 2: Suggesties beoordelen | 6 | 6/24 = 25% | 642.17 | 9 | 9/22=41% | 457.33 |
| 3: Suggesties beoordelen | 8 | 8/19 = 42% | 602.25 | 12 | 12/24=50% | 405.92 |
| 4: Suggesties beoordelen | 9 | 9/21 = 43% | 624.44 | 12 | 12/22=55% | 405.92 |
| 5: 15x iets nieuws | 4 | 4/15 = 27% | 392.5 | 2 | 2/15=13.33% | 323 |
| 6: Suggesties beoordelen | 13 | 13/24=54% | 553.08 | 15 | 15/24=63% | 410.87 |

**Irvinebrown**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **2e algoritme** | | | **1e + 2e algoritme** | | |
| **Actie** | **Juist** | **Percentueel** | **Gemiddelde** | **Juist** | **Percentueel** | **Gemiddelde** |
| 1: Suggesties beoordelen | 5 | 5/19 = 26% | 305.2 | 5 | 5/19=26% | 177.4 |
| 2: Suggesties beoordelen | 6 | 6/19 = 32% | 275.83 | 6 | 6/19=32% | 169.33 |
| 3: Suggesties beoordelen | 9 | 9/24 = 38% | 300.44 | 9 | 9/19=47% | 277.67 |
| 4: Suggesties beoordelen | 10 | 10/24=42% | 305 | 9 | 9/24=38% | 229.44 |
| 5: 15x iets nieuws | 0 | 0/15 = 0% | - | 3 | 3/15=20% | 703.33 |
| 6: Suggesties beoordelen | 12 | 12/24= 50% | 308.9 | 13 | 13/26=50% | 296.54 |

**Lucasm2**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **2e algoritme** | | | **1e + 2e algoritme** | | |
| **Actie** | **Juist** | **Percentueel** | **Gemiddelde** | **Juist** | **Percentueel** | **Gemiddelde** |
| 1: Suggesties beoordelen | 13 | 13/19 =68% | 265 | 7 | 7/19=37% | 224 |
| 2: Suggesties beoordelen | 17 | 17/21 =81% | 264.42 | 10 | 10/22=45% | 270.5 |
| 3: Suggesties beoordelen | 21 | 21/24= 88% | 279.05 | 15 | 15/22=68% | 262.6 |
| 4: Suggesties beoordelen | 21 | 21/25= 84% | 279.05 | 16 | 16/23=70% | 248.88 |
| 5: 15x iets nieuws | 4 | 4/15 = 27% | 335.25 | 6 | 6/15=40% | 409.33 |
| 6: Suggesties beoordelen | 26 | 26/29= 90% | 315 | 23 | 23/28=82% | 287.65 |

## Evaluatie test met 2e algoritme

Bij het 2e algoritme wordt er enkel gekeken naar de waarden van de nodes (welke genres respresenteren), waarna de hoogste genres worden teruggegeven als resultaat. Hiermee worden vervolgens de suggesties gegenereerd.

Door enkel gebruik te maken van deze manier, zullen de connecties tussen de nodes niet gebruikt worden, waardoor de graaf niet optimaal wordt benut. Het gevolg hiervan is, is dat de voorkeuren van de gebruiker niet optimaal worden meegenomen in het bepalen van de suggesties. Daardoor worden er over het algemeen niet heel veel nieuwe suggesties gegeven van gerelateerde genres. Bij de test met “lucasm2” kwam er in shift 4 bijvoorbeeld slechts 1 (foutieve) suggestie bij, de rest bleef ongewijzigd.

Het lijkt erop dat het systeem op deze manier vooral perfect is voor de typische Amerikaan (“lucasm2”). Aangezien de voorkeuren van de gemiddelde Amerikaan vooraf worden ingeladen aan de hand van de Yahoo data, is de keuze van lucasm2 al snel goed zichtbaar in de suggesties.

Verder is te zien dat er aan het einde respectievelijk 50, 54 en 90% van de suggesties goed zijn. Dit zijn naar ons idee op zich al prima percentages voor het suggestiesysteem.

## Evaluatie test met het volledige algoritme

In de 2e test worden de beide algoritmes geïmplementeerd, ofwel er wordt zowel gekeken naar het Minimum Spanning Tree algoritme, welke kijkt naar de lengtes van de verbindingen tussen de genres, als naar de hoogste waarden van de genres zelf.

Hierdoor is er meer variatie zichtbaar bij de suggesties. Ook worden er meer nieuwe suggesties aangedragen. Een nadeel hiervan is wel dat het vaak langer duurt voordat het ‘genre-profiel’ van een gebruiker goed zichtbaar wordt in de suggesties. Uiteindelijk is het geleverde profiel echter wel vollediger dan door enkel gebruik te maken van het 2e algoritme.

Bij gebruiker “ikea” leidt dit tot een betere inschatting van het aantal goede genres, waarschijnlijk omdat er voor deze gebruiker vooraf niet een duidelijk basisprofiel kon worden ingevoerd. Hierdoor bleven de resultaten eerder rond hetzelfde genre hangen, terwijl met het volledige systeem het resultaat veel diverser was.

Bij de gebruikers “Irvinebrown” en “Lucasm2” leidt het volledige systeem echter niet tot een verbetering, zelfs een verslechtering bij “Lucasm2”. Bij beiden stijgt het aantal goede suggesties echter nog steeds. Dit zou erop duiden dat, naarmate er vaker gebruik wordt gemaakt van het systeem en er vaker waarderingen worden gegeven aan de muziek, er uiteindelijk een breder en beter genre-profiel zal ontstaan met meer juiste suggesties.

Verder valt er te zien dat in de 2e versie van de test, er overal lagere gemiddeldes bij de eindscores worden gehaald. Dit betekent dat de suggesties over het algemeen beter passen bij de gebruiker dan de eerder aangedragen suggesties met de 1e test. Ook dit is een verbetering van het systeem.

Uiteindelijk heeft de test 50%, 63% en 82% juiste suggesties aangedragen. Dit zijn naar ons idee prima percentages voor het systeem dat wij hebben gemaakt.

## Evaluatie van de evaluatie

Tijdens deze evaluatie zijn er slechts de gegevens van 3 willekeurige gebruikers gebruikt. Alhoewel dit een redelijk inzicht geeft in hoe ons systeem presteert, kunnen we niet met absolute zekerheid zeggen dat het altijd zal werken, aangezien de testgroep nogal klein is. Het testen viel echter niet heel gemakkelijk te automatiseren, waardoor de tests handmatig zijn uitgevoerd. Dit is een redelijk tijdsintensieve manier om te testen, waardoor ervoor is gekozen om slechts te testen met 3 gebruikers en de verdere tijd te steken in verbetering van ons systeem.

Verder is het 1e algoritme (de MST variant) niet apart getest. Voor een volledig beeld van de werking van de 2 verschillende algoritmes zou dit nog handig zijn om toe te voegen aan de evaluatie.