

## Verslag 3 – Collectieve Intelligentie

Groep 4: Dennis de Buck, Cesar Groot Kormelink, Enzo Delaney-Lamour en Maik Larooij

### 3.0 Taakverdeling (update)

We hebben de taakverdeling aangehouden die vorige week een update heeft gekregen.

### 3.1 Methode

#### **CBF**

De database van Yelp is erg groot. Je kan de gehele database niet gebruiken om een algoritme te maken. Om deze reden moest de data gesplitst worden. Wij hebben de data gebruikt uit een stad, Ajax. Deze stad heeft 3288 reviews. De stad Henderson, die is gebruikt om het algoritme vervolgens te testen, heeft 220.391 reviews. Dit is dezelfde stad als bij Collaborative filtering is gebruikt om te testen, zodat we de 2 algoritmes makkelijker met elkaar kunnen vergelijken. De similaritymaten die wij hebben gebruikt zijn Jaccard en Eskin. Deze moeten vergeleken worden. Om te kijken of de algoritmes wel zin hebben om te gebruiken worden ze eerst vergeleken met de baseline: random voorspellingen. Ook het gemiddelde nemen van elke rating en deze toekennen aan alle voorspellingen wordt gebruikt als baseline. Een goede maat om te kijken of een algoritme goed werkt is de MSE uitrekenen. Vervolgens wordt er een precision-recall curve gecreëerd in een plot, waarbij telkens de precision en recall worden weergegeven met een veranderende threshold.

#### **CF**

Om op een goede manier te evalueren hebben we gebruik gemaakt van een subset van de complete dataset omdat deze te groot was om helemaal te gebruiken. Deze subset bestond uit twee steden: Toronto en Henderson. De grotere stad Toronto is gebruikt om het model te trainen, waarna de tests zijn uitgevoerd op Henderson. Deze stad heeft in totaal 220.391 reviews om mee te werken, dit fijne aantal was de reden dat we deze stad kozen voor het testen. De grote stad Toronto, die is gebruikt om het model te maken, bevat 525.477 reviews. Om de verschillende similarities (Cosine, Euclidian, Minkowski) met elkaar te vergelijken moest er eerst gekeken worden hoe het algoritme scoort ten opzichte van een baseline. Wij hebben gekeken naar een random baseline voor het vergelijken van de MSE en later ook nog gebruik gemaakt van een voorspelling op basis van het gemiddelde per restaurant bij de precision recall curve. Het algoritme heeft eerst 200 random bedrijven uitgekozen, zoals beschreven in verslag 2. Goed om te noemen is dat de gehele evaluatie heeft plaatsgevonden over dezelfde bedrijven. Om te kijken of de resultaten van de stad Henderson niet compleet toevallig waren, is er ook getest op de stad Akron met 9193 reviews.

### 3.2 Resultaten

#### Collaborative filtering

Collaborative filtering (Henderson)	MSE
Euclidean distance	0,322
Cosine similarity	0,333
Minkowski distance	0,384
Random baseline	3,849
Mean baseline	1,223

*Tabel 1. MSE collaborative filtering*

Collaborative filtering (Akron)	MSE
Euclidean distance	0,391
Cosine similarity	0,266
Minkowski distance	0,381
Random baseline	4,143
Mean baseline	1,184

De verschillende methodes van similarity en aanbevelingen berekenen worden hier allemaal naast elkaar gezet op de mean squared error (mse). De mse is een kwaliteitsmaat waarbij de afstanden tussen de voorspelde rating en de echte rating gekwadrateerd, opgeteld en door het aantal afstanden gedeeld worden. Hoe dichterbij de nul in de buurt ligt hoe hoger de kwaliteit/accuraatheid van de aanbevelingen zijn. De mse's van de drie afstandsmaten worden vergeleken met een baseline van random voorspellingen omdat er zo bekeken kan worden of de methoden ook echt beter werken dan random aanbevelingen.

Deze precision-recall plots laten zien wat de precision en recall zijn van de verschillende methodes bij elke treshold. Op deze manier kan er worden gekeken welke treshold de hoogste combinatie van precision en recall op levert. Ook worden de verschillende vormen waarmee similarities en aanbevelingen zijn berekend hier met elkaar vergeleken op precision en recall.

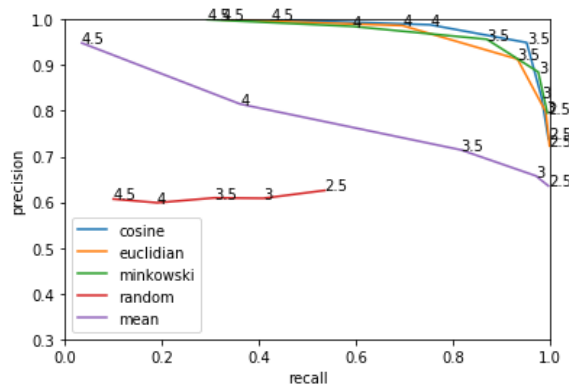


Figure 2. Precision/recall curve (Henderson)

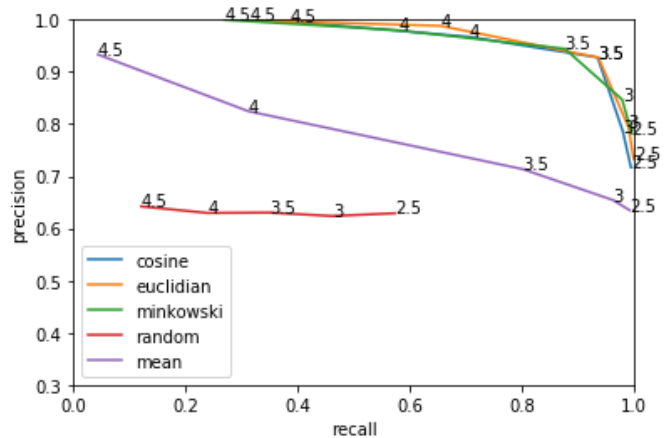


Figure 1. Precision/recall curve (Akron)

## Content based filtering

Door de Jaccard similarity index te gebruiken zijn alle bedrijven vergeleken op basis van categorieën. Deze methode voor het doen van aanbevelingen leverde de volgende MSE op:

Content based filtering (Ajax)	MSE
Jaccard	0,429
Eskin	1,698
Random baseline	3,560
Mean baseline	0.971

Tabel 2. MSE content based filtering

Content based filtering (Henderson)	MSE
Jaccard	0,429
Random baseline	3,560
Mean baseline	0.971

Ook is gekeken of we een andere similarity maat konden gebruiken die betere resultaten kon opleveren. Zo kwamen we bij de Eskin similarity maat. Met een test naar de effectiviteit van het algoritme met deze similarity maat konden we conclusies trekken over het gebruik van complexere similarity maten zoals Eskin en Goodall. Eskin bleek voor grotere steden niet goed te werken omdat het algoritme dan te lang moet laden. De jaccard index gaf dan voldoende informatie.

In figuur 3 is de precision/recall curve te zien voor het content based algoritme. Op basis van dit plot kunnen er conclusies worden getrokken over een ideale treshold waar een voorspelde rating aan moet voldoen.

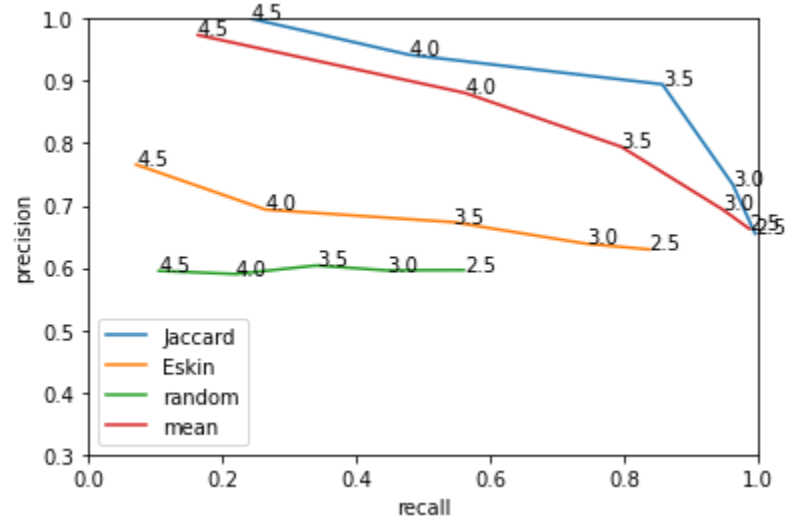


Figure 3. Precision/recall curve (Ajax)

### 3.3 Bespreking

#### CBF

De gebruikte similarity maten geven beide een lagere MSE dan de random voorspellingen. Dit betekent dat het algoritme wat ontworpen is beter zal werken dan op willekeurige wijze bedrijven voorspellen. Tussen de twee similarity maten zat ook een groot verschil. De Jaccard similarity kwam uit op een MSE van 0.429, terwijl de Eskin similarity uitkwam op 1,698. De Eskin similarity werkt zelfs minder accuraat dan de baseline van de gemiddelde rating. Hiermee wordt aangetoond dat de Jaccard similarity de juiste similarity maat is voor ons algoritme. De beste verhouding tussen precision en recall kwam uit bij een threshold van 3.5, met een precision van 0.9 en een recall van 0.85. De voorspellingen van het Collaborative filtering algoritme waren wat accurater dan CBF. De voorspelling was juist dat CBF beter zou werken dan CF, omdat er relatief weinig reviews per bedrijf waren, en elk bedrijf eigenlijk meerdere categories had.

#### CF

De precision recall curve laat ons zien welke treshold het beste werkt voor dit algoritme. In dit geval hebben wij, gebaseerd op de resultaten van beide test-steden, dit geschat op een treshold van 3.5. Cosine similarity en euclidian distance scores in Henderson ongeveer even goed. Minkowski scoort tegen verwachting in wat lager. Er is te zien dat de geprogrammeerde algoritmes in ieder geval aanzienlijk beter scoren dan de baselines, bij MSE én precision/recall. Op basis hiervan schatten wij dat ons algoritme redelijk accurate voorspellingen doen. Het is

moeilijk om te zeggen welk van de drie similarity maten nou het beste presteert. Het lijkt erop dat Minkowski het ietsjes slechter doet dan Cosine en Euclidian. Het heeft een hogere MSE en de precision en recall liggen bij de gekozen treshold van 3.5 ook lager. Cosine en Euclidian scoren erg gelijkwaardig en zullen wat ons betreft allebei goed kunnen worden ingezet. Kijkende naar de resultaten uit Akron, is het enige opvallende dat de Cosine similarity ineens een stuk lager uitvalt en dus duidelijk hier het beste werkt. De rest van de resultaten zijn gelijkwaardig. Wellicht heeft de verdeling van reviews ervoor gezorgd dat Cosine similarity hier goede resultaten behaald. Er zou op nog meer steden kunnen worden getest of dat ook bij deze steden het geval is, wellicht is Cosine dus de beste oplossing.

### **Gemeenschappelijke conclusie**

We kunnen geen goede uitspraak doen over of één van de algoritmes beter is. De algoritmes maken beide gebruik van compleet andere filters om een aanbeveling te doen. Daarom is de dataset om te vergelijken niet hetzelfde. We kunnen zeggen dat beide algoritmes prima werken en ze in een combinatie gebruikt kunnen worden. Er moet bij collaborative filtering bijvoorbeeld rekening gehouden worden met het 'cold start probleem', waarbij nieuwe gebruikers nog geen reviews hebben geschreven. Op deze uitzondering zou in principe goed het content based filtering algoritme worden toegepast. Er moet dan wel een bepaalde interactie hebben plaatsgevonden als een zoekopdracht of een click. Daarnaast neemt het collaborative filtering algoritme gebruikers met minder dan vijf reviews niet mee, ook voor deze gebruikers zou het content based algoritme kunnen werken. Een vervolgvraag die voortkomt uit onze onjuiste hypothese is de volgende: werken deze algoritmes ook wanneer er gekeken wordt naar alle soorten bedrijven, in plaats van alleen bedrijven?