



Faculteit Bedrijf en Organisatie

Personalized Search: Graph vs. OLAP Database

Shauni Van de Velde

Scriptie voorgedragen tot het bekomen van de graad van
professionele bachelor in de toegepaste informatica

Promotor:
Guy Dekoning
Co-promotor:
Nicolas Lierman

Instelling: MultiMinds

Academiejaar: 2019-2020

Tweede examenperiode

Faculteit Bedrijf en Organisatie

Personalized Search: Graph vs. OLAP Database

Shauni Van de Velde

Scriptie voorgedragen tot het bekomen van de graad van
professionele bachelor in de toegepaste informatica

Promotor:
Guy Dekoning
Co-promotor:
Nicolas Lierman

Instelling: MultiMinds

Academiejaar: 2019-2020

Tweede examenperiode

Woord vooraf

Samenvatting

In samenwerking met MultiMinds werd besloten onderzoek te voeren naar de haalbaarheid van

Inhoudsopgave

1	Inleiding	13
1.1	Probleemstelling	13
1.2	Onderzoeksvraag	14
1.3	Onderzoeksdoelstelling	14
1.4	Opzet van deze bachelorproef	14
2	Stand van zaken	15
2.1	Personalisatie	15
2.1.1	Personalisatie op basis van e-mail en sociale media	15
2.1.2	Personalisatie op basis van geografische locatie	15
2.1.3	Personalisatie op basis van IP-adres	16
2.1.4	Verwante inhoud personalisatie	17

2.2	Recommender Systems	17
2.2.1	Haalbaarheid	18
2.2.2	Algoritmen voor aanbevelingssystemen	18
2.3	Wat is Graph?	21
2.3.1	Graph	21
2.3.2	Neo4j	22
2.4	Wat is ElasticSearch?	22
2.5	Filter Bubble	22
3	Methodologie	25
4	Conclusie	27
A	Onderzoeksvoorstel	29
A.1	Introductie	29
A.2	State-of-the-art	29
A.3	Methodologie	30
A.4	Verwachte resultaten	31
A.5	Verwachte conclusies	31
	Bibliografie	33

Lijst van figuren

Lijst van tabellen

1. Inleiding

De inleiding moet de lezer net genoeg informatie verschaffen om het onderwerp te begrijpen en in te zien waarom de onderzoeksvraag de moeite waard is om te onderzoeken. In de inleiding ga je literatuurverwijzingen beperken, zodat de tekst vlot leesbaar blijft. Je kan de inleiding verder onderverdelen in secties als dit de tekst verduidelijkt. Zaken die aan bod kunnen komen in de inleiding (Pollefliet, 2011):

- context, achtergrond
- afbakenen van het onderwerp
- verantwoording van het onderwerp, methodologie
- probleemstelling
- onderzoeksdoelstelling
- onderzoeksvraag
- ...

1.1 Probleemstelling

De meeste e-commerce online platformen hebben reeds een zoekfunctie geïmplementeerd, maar deze is vaak beperkt tot enkel de productcatalogus. Rekeninghoudend met de algemene trend rond personalisering van de customer experience zou het ook aangewezen zijn om ook de zoekfunctionaliteit op e-commerce websites te personaliseren. Dit zou resulteren in een betere gebruikerservaring voor de klant en een hogere conversie voor het bedrijf.

1.2 Onderzoeksvraag

De onderzoeksvraag bestaat eruit om te ontdekken welke databanktechnologie de beste oplossing biedt om in real-time op grote schaal gepersonaliseerde zoekresultaten te kunnen leveren.

Belangrijke criteria hierbij zijn performantie, kwaliteit van de resultaten, en of het al dan niet mogelijk is om familierelaties te kunnen verwerken.

Concreet omvat dit onderzoek volgende onderzoeksvragen:

- Welke technologie biedt de mogelijkheid om een gepersonaliseerde zoekfunctie te implementeren
- Welke technologie biedt de beste resultaten op basis van performantie en kwaliteit
- Laten deze technologieën toe om rekening te houden met factoren die niet te maken hebben met historisch koopgedrag (bv. leeftijd, geslacht, gezinssamenstelling)

1.3 Onderzoeksdoelstelling

Het onderzoek heeft als doel om te ontdekken in hoeverre de persoonlijke data van een specifieke gebruiker en contextuele data kan ingeschakeld en gecombineerd worden met de data uit een productcatalogus om persoonlijke en relevante zoekresultaten te genereren. We spreken hier vooral over historische koopdata van de persoon zelf, maar ook demografische data zoals geslacht, leeftijd, gezinssamenstelling kunnen hierbij belangrijke factoren zijn.

1.4 Opzet van deze bachelorproef

De rest van deze bachelorproef is als volgt opgebouwd:

In Hoofdstuk 2 wordt een overzicht gegeven van de stand van zaken binnen het onderzoeksdomein, op basis van een literatuurstudie.

In Hoofdstuk 3 wordt de methodologie toegelicht en worden de gebruikte onderzoekstechnieken besproken om een antwoord te kunnen formuleren op de onderzoeksvragen.

In Hoofdstuk 4, tenslotte, wordt de conclusie gegeven en een antwoord geformuleerd op de onderzoeksvragen. Daarbij wordt ook een aanzet gegeven voor toekomstig onderzoek binnen dit domein.

2. Stand van zaken

2.1 Personalisatie

Personalisatie van webapplicaties en websites draait erom de bezoekers een op maat gemaakte ervaring aan te bieden. Dit kan op verschillende manieren toegepast worden, en kan meerdere doelen hebben. In dit hoofdstuk wordt er verder ingegaan op welke manieren personalisatie van websites wordt toegepast, en wat dit betekent voor zowel de bezoeker als het bedrijf zelf.

2.1.1 Personalisatie op basis van e-mail en sociale media

Zowat iedereen heeft wel te kampen met een overvloed aan e-mails in hun postvak van allerlei websites waar ze hun e-mailadres ooit hebben vrijgegeven. U zou denken dat deze door de meeste mensen simpelweg verwijderd worden, maar e-mailmarketing blijft een van de meest succesvolle marketingstrategieën (Dehkordi, Rezvani, Rahman, Nahid & Jouya, 2012). E-mailmarketing is relatief makkelijk te implementeren en vereist weinig technische investering, meestal wordt dit verwezenlijkt via systemen van derden zoals bijvoorbeeld MailChimp. Een nadeel van deze marketingvorm is dat het bedrijf continu bezig moet zijn met nieuwe inhoud te creëren voor deze e-mails, alsook op de website waar de marketingmails over gaan.

2.1.2 Personalisatie op basis van geografische locatie

Geografische personalisatie is het aanpassen van de website op basis van de locatie van de gebruiker. Gebruikers uit België die naar de website van een internationaal bedrijf surfen,

zullen dan worden omgeleid naar een Nederlandse of Franse versie van die website.

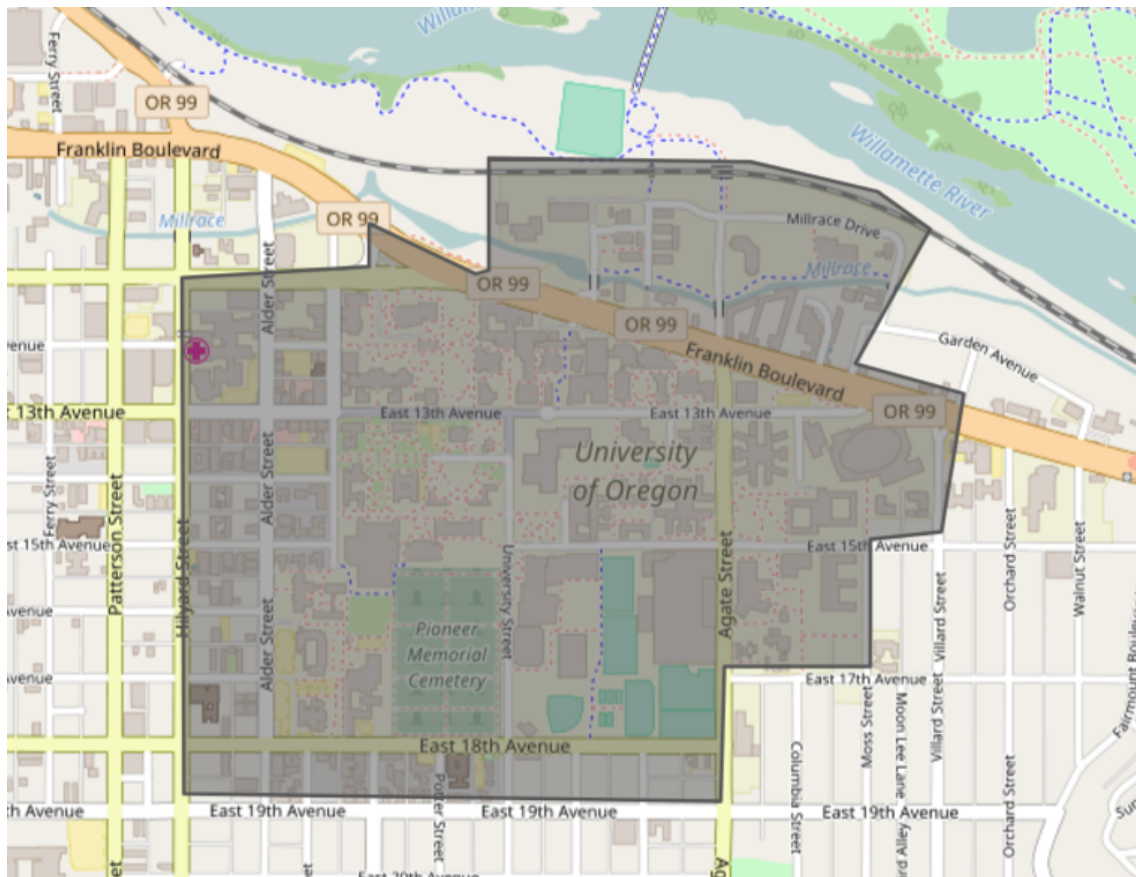
Geografische personalisatie kan ook gebruikt worden om de inhoud van een pagina aan te passen aan de hand van de locatie van de gebruiker, of om vertalingen aan te bieden. Een nadeel hiervan is dat mensen die op reis gaan het soms moeilijk zouden kunnen hebben om naar de juiste versie van de website te navigeren, aangezien het systeem de gebruiker zal willen omleiden naar de pagina of inhoud die voorzien is voor het land waar zij zich momenteel in bevinden. Eenzelfde probleem kan zich voordoen bij bedrijven die hun webverkeer omleiden via een ander land door middel van bijvoorbeeld een VPN.

Geografische personalisatie is ook relatief eenvoudig te implementeren en kan een grote troef zijn op de internationale markt.

2.1.3 Personalisatie op basis van IP-adres

Deze methode van personalisatie is wat minder opvallend, aangezien het bij de gemiddelde internetgebruiker weinig tot nooit zal voorkomen, aangezien zij het internet gebruiken via een serviceprovider zoals Telenet of Proximus.

Deze vorm van personalisatie wordt gebruikt om zakelijke gebruikers en bedrijven te kunnen identificeren op basis van hun IP-adres. Zo kan men zien of een bezoeker bij een bepaald bedrijf werkzaam is om deze direct aan te spreken op bijvoorbeeld de homepage.



Net zoals bij personalisatie op basis van locatie kan dit misleidende resultaten opleveren, bijvoorbeeld als de werknemer van thuis werkt of het IP-adres niet duidelijk aantoont vanuit welk bedrijf het webverkeer van de bezoeker afkomstig is. Ook voor performantie kan dit negatieve gevolgen hebben, aangezien deze vorm van personalisatie afhankelijk is van systemen van derden.

Verder moet er ook inhoud gecreëerd worden voor elk bedrijf dat men specifiek wil aanspreken. Dit is een tijdrovend proces, maar aangezien deze vorm van personalisatie weinig voorkomt, is het wel een troef waardoor het bedrijf zich kan onderscheiden van de meerderheid en zich kan laten opvallen.

2.1.4 Verwante inhoud personalisatie

Dit is de vorm van personalisatie die een grote meerwaarde zal leveren aan dit onderzoek. De meeste mensen hebben deze vorm al ondervonden op een webshop zoals Amazon of Bol.com. Deze vorm draait erom de gebruikers artikels aan te raden op basis van artikels of inhoud die ze al eerder bekeken hebben, alsook het gedrag van andere gebruikers.

De werking van het aanbevelingssysteem van Amazon is gebaseerd op enkele complexe algoritmen (Linden, Smith & York, 2003). Dit is natuurlijk verantwoord omdat zij op zeer grote schaal werken en veel geld hebben geïnvesteerd in de ontwikkeling van hun systeem.

In de realiteit hoeven de technologieën voor aanbevelingen van producten niet zo complex te zijn voor gewone webshops en bedrijven, vaak is het voldoende om relaties te creëren tussen artikels en op basis van deze relaties nieuwe artikels aan te raden aan de gebruikers.

Een voorbeeld van een relatie tussen twee artikels is de welbekende 'Anderen bekeken ook' blok die vaak zichtbaar is bij het bekijken van een detailpagina van een product. Een simpelere methode van dergelijke relaties is het aanbieden van verwante producten op basis van categorieën of tags. Tags zijn een manier om kenmerken van een product weer te geven die specifiek zijn dan een categorie. Een categorie kan dan 'schoenen' zijn, terwijl een tag 'lage sneakers' is.

2.2 Recommender Systems

Recommender Systems (Resnick & Varian, 1997) zijn aanbevelingen vanuit het systeem die rekening houden met de beschikbare informatie van gebruikers en hun voorkeuren om zo een filter te plaatsen op de informatie die weergegeven wordt. Verder zullen we deze benoemen aan de hand van hun Nederlandse naam 'aanbevelingssystemen'.

Aanbevelingssystemen worden vooral gebruikt in een e-commerce toepassingen waar een zeer groot en verscheiden aanbod aan producten is, en het al vaak lastig wordt om precieze aanbevelingen aan de klant te geven. Hierbij wordt allerlei informatie van een gebruiker verzameld, zoals historische aankopen, items op het verlanglijstje, items waar de gebruiker op geklikt heeft, etc.

In dit onderdeel zullen we kort wat dieper ingaan op de werking van dergelijke aanbevelingssystemen en de achterliggende algoritmen.

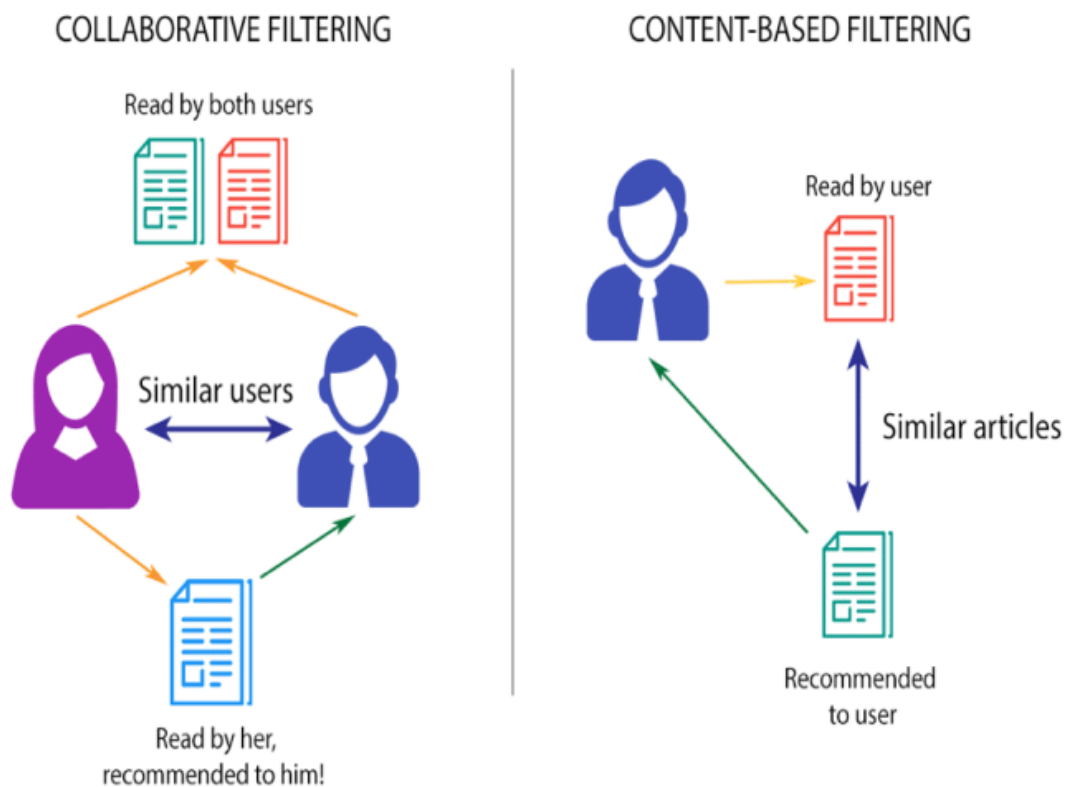
2.2.1 Haalbaarheid

Een degelijk systeem biedt een grote meerwaarde voor een bedrijf, zo heeft Netflix een competitie gehouden die 1 miljoen dollar bood aan degene die een aanbevelingssysteem kon maken dat 10% beter presteerde dan hun bestaande systeem. Deze wedstrijd liep van oktober 2006 tot minstens oktober 2011. De wedstrijd werd de Netflix, 2009 genoemd, Netflix stelde hiervoor een dataset beschikbaar. Enkele groepen hebben het doel behaald, maar het algoritme van de winnende groep is uiteindelijk nooit in productie gebracht, omdat de mogelijke opbrengst niet kon opwegen tegen de extra gevraagde rekenkracht.

Een aanbevelingssysteem moet dus niet enkel de juiste waarden kunnen aangeven, het moet ook haalbaar zijn qua rekenkracht en extra kosten. Het ontwerpen van dergelijk systeem is dus niet eenvoudig.

2.2.2 Algoritmen voor aanbevelingssystemen

We spreken over twee grote categorieën in algoritmen van aanbevelingssystemen: 'collaborative filtering'-algoritmen en 'content based'-algoritmen. Adomavicius en Tuzhilin, 2005 Door de evolutie en groeiende ontwikkeling van beter presterende systemen zijn er ondertussen ook enkele algoritmen die niet echt binnen een van deze twee koepels vallen. In de onderstaande figuur wordt een simpele representatie gegeven van de werking van deze soorten algoritmes.



Collaborative Filtering

Het kernidee bij Collaborative Filtering (**Schafera**) is om aanbevelingen te maken gebaseerd op de voorkeuren van een andere gebruiker met een gelijkaardig gedrag. Zoals de figuur hierboven aantoont dat als gebruiker 1 en 2 hetzelfde artikel gelezen hebben, gebruiker 2 een artikel als aanbeveling zal krijgen dat gelezen werd door gebruiker 1. Hetzelfde idee kan toegepast worden op allerlei interacties van de gebruikers met een website zoals likes, shares, verlanglijstjes, etc.

In de praktijk geeft deze techniek zeer goede resultaten, maar zoals verwacht brengt deze techniek ook enkele problemen met zich mee. Het meest merkwaardige probleem is een zogenaamde cold start, dit komt onder andere voor bij de eerste interactie van een nieuwe gebruiker met een applicatie die aanbevelingen biedt. Het algoritme heeft dan onvoldoende informatie om een correcte en nuttige aanbeveling op te leveren aan de gebruiker. Een andere oorzaak kan zijn dat er een nieuw product wordt toegevoegd aan het systeem, er kan dan nog niet geweten zijn welk type gebruiker hierin geïnteresseerd zou kunnen zijn.

Een andere factor voor het succes van een Collaborative Filtering algoritme is het aantal gebruikers van een systeem, met andere woorden, hoe meer gebruikers er zijn, hoe correcter de aanbevelingen aan een specifieke gebruiker zal zijn. (Sarwar, Karypis, Konstan & Riedl, 2001). Een gebruiker met ongewone interesses zal logischerwijs in dergelijk klein systeem weinig gelijkaardige gebruikers hebben, en zal dus ook geen optimale aanbevelingen krijgen.

Een groot voordeel van Collaborative Filtering is dat er absoluut geen kennis hoeft te zijn van de toepassing van het systeem, de aanbevelingen worden gegenereerd op basis van het gedrag van de gebruikers en zijn interesses. De producten of hun attributen moeten dus niet gekend zijn om aanbevelingen te kunnen geven, dat maakt het eenvoudiger om een aanbevelingssysteem met de techniek van Collaborative Filtering te implementeren.

Content Based

Content Based aanbevelingssystemen (**Lops2011**) maken, in tegenstelling tot Collaborative Filtering, wel gebruik van de specifieke producten binnen het systeem en hun attributen. Op basis van deze attributen en de interesse van de gebruiker daarin, wordt per gebruiker een profiel opgezet, elk attribuut krijgt dan een score toegekend, een hogere score betekent grotere interesse. Een attribuut van een product kan dan bijvoorbeeld 'schoenen' of 'PS4 games' zijn, of zelfs een filmgenre.

Een probleem van Content Based is, net zoals bij Collaborative Filtering, het cold start probleem. Als een nieuwe gebruiker het systeem gebruikt, is er voor deze gebruiker nog geen profiel opgesteld en kunnen er ook geen aanbevelingen gemaakt worden.

Een ander probleem van Content Based wordt overspecialisatie genoemd, dit treedt op wanneer het systeem eigenlijk té accurate aanbevelingen doet. Het gevolg hiervan is dat slechts enkele producten voldoen aan de verwachtingen van het systeem, waardoor er geen nieuwe aanbevelingen aan de gebruiker naar voor gebracht worden, en de gebruiker enkel producten zal zien die hij reeds bekeken heeft.

Het andere probleem dat optreedt bij Collaborative Filtering, namelijk dat bij het toevoegen van nieuwe producten niet geweten kan zijn welke gebruikers hierin geïnteresseerd zouden zijn, is niet van toepassing bij Content Based. Het systeem maakt gebruik van de attributen van producten, dus nieuwe producten kunnen meteen belanden in de aanbevelingen van gebruikers die reeds interesse getoond hebben in andere producten met die attributen. Ook het aantal gebruikers binnen een systeem vormt om dezelfde reden geen probleem bij Content Based aanbevelingssystemen.

Een grote boosdoener bij Content Based kan zijn dat producten slecht gelabeld zijn, en hun attributen onvoldoende passen bij wat het product effectief is. Hierdoor kan het systeem deze producten niet goed vergelijken met andere. Dit is vooral een probleem wanneer de attributen van de producten van verschillende bronnen afkomstig zijn, of manueel slecht opgesteld zijn.

Hybrides

Beide van de voorgenoemde technieken hebben elk hun eigen voor- en nadelen, alsook sterke en zwakke punten. Content Based heeft te kampen met overspecialisatie, maar is wel in staat om nieuwe producten meteen aan te bevelen aan de gebruikers. Collaborative Filtering heeft moeite met het aanbevelen van nieuwe producten en een cold start, maar heeft geen problemen in de aard van overspecialisatie. De logische redenering is dan

natuurlijk om deze twee soorten systemen te gaan combineren, kwestie van het beste van twee werelden te proberen bekomen. Dit worden hybride aanbevelingssystemen (Cano2017) genoemd.

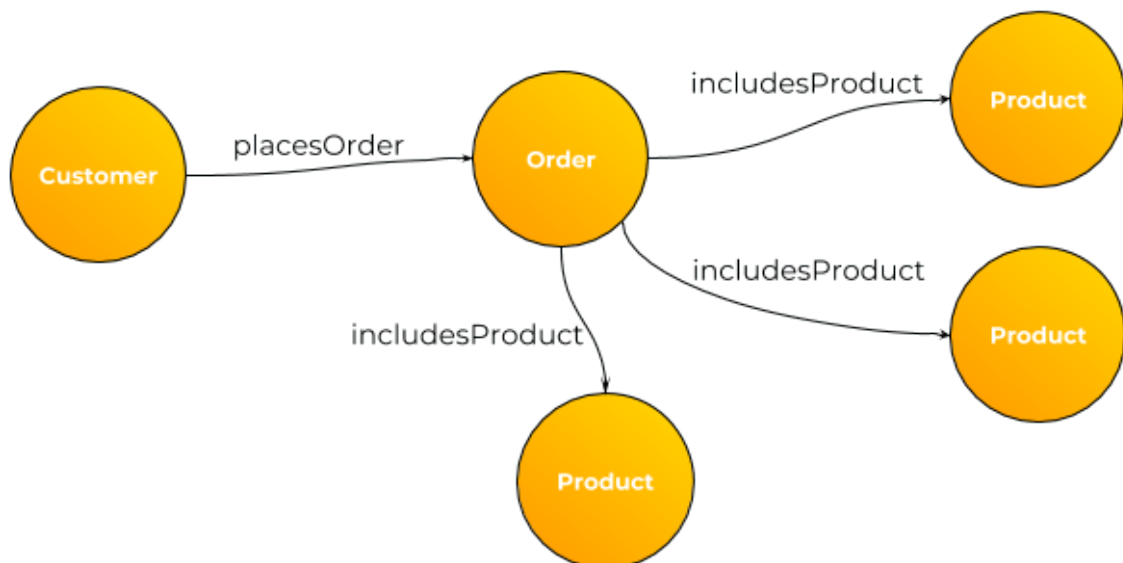
De manier van het opbouwen van een hybride systeem kan zijn dat beide methoden afzonderlijk worden uitgevoerd, en hun resultaten op het einde samengebundeld worden. Dit is de meest eenvoudige implementatie. Een andere manier van combineren kan zijn door de informatie binnen een Collaborative Filtering systeem aan te vullen met informatie uit de gebruikersprofielen. Hierdoor wordt de gelijkaardigheid van twee producten bepaald door zowel de inhoud en welke soorten typische gebruikers deze producten bekijken, kopen, leuk vinden, etc. De informatie van deze gebruikers kan dan bijvoorbeeld leeftijdsgroep, woonplaats, gezinssamenstelling, etc. voorstellen.

2.3 Wat is Graph?

In de context van deze bachelorproef zal er met 'Graph' steeds verwezen worden naar een Graph databank. In dit onderdeel zullen we wat dieper ingaan op wat dit soort databank precies inhoudt om een volledig begrip van de precieze werking te verzekeren.

2.3.1 Graph

Dit is de structuur die verder in dit onderzoek gebruikt zal worden. Deze structuur maakt gebruik van een wiskundige graaf om data op te slaan. Een graaf bestaat uit een aantal knopen (genaamd nodes) die al dan niet verbonden zijn. Een groot voordeel hiervan is dat deze databanken sneller zijn dan relationele databanken omdat ze snel naar een bepaalde node kunnen verwijzen, waarbij we bij een relationele databank een JOIN zouden moeten gebruiken.



2.3.2 Neo4j

Er bestaan verschillende platformen voor SQL, alsook verschillende platformen voor de hierboven beschreven NoSQL databanken. Neo4j is een van de meer bekende platformen voor graph databanken, dit platform zal gebruikt worden in dit onderzoek.

Walmart maakt gebruik van Neo4j om de aanbevelingen voor hun klanten op hun online webservices te optimaliseren (NeoTechnology, 2014). Zij gebruiken dit omdat graph databanken zeer snel over een gebruiker zijn koophistorie kunnen traverseren, en ook direct nieuwe mogelijke interesses kunnen halen uit het gedrag van de gebruiker. Daarmee wordt bedoeld dat er in real-time nieuwe connecties worden gelegd tussen de gebruiker en de producten, en hij de nieuwe aanbevelingen meteen zal zien, en niet enkele dagen of uren later. Er wordt dus historische data gematcht met real-time data, hier blinkt Neo4j in uit.

Neo4j maakt dus gebruik van wiskundige grafen om data weer te geven samen met hun onderlinge relaties. Een graaf kan gericht of ongericht zijn, ongericht wil zeggen dat er geen richting is waarin de relaties lopen, dus deze zijn onderling uitwisselbaar. Een gerichte graaf is dan een graaf waarbij de relaties in een specifieke richting lopen, zoals volgers op Twitter: Persoon A volgt persoon B, maar persoon B volgt niet persoon A.

2.4 Wat is Elasticsearch?

Elasticsearch op zichzelf is eigenlijk een zoekmachine die data kan analyseren, of dit nu nummers, tekst, gestructureerd of ongestructureerd is. Elasticsearch is een component van de Elastic stack, ook wel ELK-Stack genoemd. Dit staat voor Elasticsearch, Logstash en Kibana. Logstash wordt gebruikt om data te verwerken uit meerdere bronnen en te versturen naar Elasticsearch. Kibana is een tool om de gebruikers een visueel beeld te geven van de data in de vorm van grafieken of tabellen.

In Elasticsearch is het mogelijk gewichten toe te kennen aan de resultaten van een zoekopdracht, zo kunnen meer relevante resultaten hoger in de lijst staan, dit is een belangrijke factor in dit onderzoek, namelijk of deze technologie gebruikt kan worden in E-Commerce toepassingen. In het artikel van Vavliakis, Katsikopoulos en Symeonidis, 2019 wordt beweerd dat het systeem dat zij implementeerden op een performante manier de gewenste resultaten gaf, alsook dat dit een oplossing kan zijn voor real-time zoekopdrachten in commerciële toepassingen.

2.5 Filter Bubble

Een filter bubble, ook wel informatieluchtbel genaamd, is een gevolg van het personaliseren van zoekopdrachten. Personalisatie is het proberen bepalen welke zaken een gebruiker zou willen zien, aan de hand van een algoritme. Een filter bubble betekent dat door deze personalisatie, een bijvoorbeeld geen producten zal zien die hem niet interesseren, in dit geval is dat een voordeel. Als we kijken in de context van controversiële onderwerpen,

zorgt deze bubbel ervoor dat de gebruiker eigenlijk geen informatie zal zien die niet bij zijn standpunt past, de gebruikers worden dus afgesloten in luchtbel bestaande uit enkel hun visie of standpunt. (Pariser, 2011)

De relevantie van dit fenomeen voor dit onderzoek valt op het feit dat bij de Content Based algoritmen die eerder besproken werden, we met een gelijkaardig probleem te maken hadden. Bij deze aanbevelingssystemen kan er overspecialisatie optreden, waardoor de gebruiker enkel producten aanbevolen krijgt die reeds gezien zijn, of té gelijkaardig zijn, en er dus geen nieuwe verrassende producten aangeboden worden.

3. Methodologie

Etiam pede massa, dapibus vitae, rhoncus in, placerat posuere, odio. Vestibulum luctus commodo lacus. Morbi lacus dui, tempor sed, euismod eget, condimentum at, tortor. Phasellus aliquet odio ac lacus tempor faucibus. Praesent sed sem. Praesent iaculis. Cras rhoncus tellus sed justo ullamcorper sagittis. Donec quis orci. Sed ut tortor quis tellus euismod tincidunt. Suspendisse congue nisl eu elit. Aliquam tortor diam, tempus id, tristique eget, sodales vel, nulla. Praesent tellus mi, condimentum sed, viverra at, consectetur quis, lectus. In auctor vehicula orci. Sed pede sapien, euismod in, suscipit in, pharetra placerat, metus. Vivamus commodo dui non odio. Donec et felis.

Etiam suscipit aliquam arcu. Aliquam sit amet est ac purus bibendum congue. Sed in eros. Morbi non orci. Pellentesque mattis lacinia elit. Fusce molestie velit in ligula. Nullam et orci vitae nibh vulputate auctor. Aliquam eget purus. Nulla auctor wisi sed ipsum. Morbi porttitor tellus ac enim. Fusce ornare. Proin ipsum enim, tincidunt in, ornare venenatis, molestie a, augue. Donec vel pede in lacus sagittis porta. Sed hendrerit ipsum quis nisl. Suspendisse quis massa ac nibh pretium cursus. Sed sodales. Nam eu neque quis pede dignissim ornare. Maecenas eu purus ac urna tincidunt congue.

Donec et nisl id sapien blandit mattis. Aenean dictum odio sit amet risus. Morbi purus. Nulla a est sit amet purus venenatis iaculis. Vivamus viverra purus vel magna. Donec in justo sed odio malesuada dapibus. Nunc ultrices aliquam nunc. Vivamus facilisis pellentesque velit. Nulla nunc velit, vulputate dapibus, vulputate id, mattis ac, justo. Nam mattis elit dapibus purus. Quisque enim risus, congue non, elementum ut, mattis quis, sem. Quisque elit.

Maecenas non massa. Vestibulum pharetra nulla at lorem. Duis quis quam id lacus dapibus interdum. Nulla lorem. Donec ut ante quis dolor bibendum condimentum. Etiam egestas

tortor vitae lacus. Praesent cursus. Mauris bibendum pede at elit. Morbi et felis a lectus interdum facilisis. Sed suscipit gravida turpis. Nulla at lectus. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Praesent nonummy luctus nibh. Proin turpis nunc, congue eu, egestas ut, fringilla at, tellus. In hac habitasse platea dictumst.

Vivamus eu tellus sed tellus consequat suscipit. Nam orci orci, malesuada id, gravida nec, ultricies vitae, erat. Donec risus turpis, luctus sit amet, interdum quis, porta sed, ipsum. Suspendisse condimentum, tortor at egestas posuere, neque metus tempor orci, et tincidunt urna nunc a purus. Sed facilisis blandit tellus. Nunc risus sem, suscipit nec, eleifend quis, cursus quis, libero. Curabitur et dolor. Sed vitae sem. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Maecenas ante. Duis ullamcorper enim. Donec tristique enim eu leo. Nullam molestie elit eu dolor. Nullam bibendum, turpis vitae tristique gravida, quam sapien tempor lectus, quis pretium tellus purus ac quam. Nulla facilisi.

4. Conclusie

Curabitur nunc magna, posuere eget, venenatis eu, vehicula ac, velit. Aenean ornare, massa a accumsan pulvinar, quam lorem laoreet purus, eu sodales magna risus molestie lorem. Nunc erat velit, hendrerit quis, malesuada ut, aliquam vitae, wisi. Sed posuere. Suspendisse ipsum arcu, scelerisque nec, aliquam eu, molestie tincidunt, justo. Phasellus iaculis. Sed posuere lorem non ipsum. Pellentesque dapibus. Suspendisse quam libero, laoreet a, tincidunt eget, consequat at, est. Nullam ut lectus non enim consequat facilisis. Mauris leo. Quisque pede ligula, auctor vel, pellentesque vel, posuere id, turpis. Cras ipsum sem, cursus et, facilisis ut, tempus euismod, quam. Suspendisse tristique dolor eu orci. Mauris mattis. Aenean semper. Vivamus tortor magna, facilisis id, varius mattis, hendrerit in, justo. Integer purus.

Vivamus adipiscing. Curabitur imperdiet tempus turpis. Vivamus sapien dolor, congue venenatis, euismod eget, porta rhoncus, magna. Proin condimentum pretium enim. Fusce fringilla, libero et venenatis facilisis, eros enim cursus arcu, vitae facilisis odio augue vitae orci. Aliquam varius nibh ut odio. Sed condimentum condimentum nunc. Pellentesque eget massa. Pellentesque quis mauris. Donec ut ligula ac pede pulvinar lobortis. Pellentesque euismod. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent elit. Ut laoreet ornare est. Phasellus gravida vulputate nulla. Donec sit amet arcu ut sem tempor malesuada. Praesent hendrerit augue in urna. Proin enim ante, ornare vel, consequat ut, blandit in, justo. Donec felis elit, dignissim sed, sagittis ut, ullamcorper a, nulla. Aenean pharetra vulputate odio.

Quisque enim. Proin velit neque, tristique eu, eleifend eget, vestibulum nec, lacus. Vivamus odio. Duis odio urna, vehicula in, elementum aliquam, aliquet laoreet, tellus. Sed velit. Sed vel mi ac elit aliquet interdum. Etiam sapien neque, convallis et, aliquet vel, auctor non, arcu. Aliquam suscipit aliquam lectus. Proin tincidunt magna sed wisi. Integer blandit

lacus ut lorem. Sed luctus justo sed enim.

Morbi malesuada hendrerit dui. Nunc mauris leo, dapibus sit amet, vestibulum et, commodo id, est. Pellentesque purus. Pellentesque tristique, nunc ac pulvinar adipiscing, justo eros consequat lectus, sit amet posuere lectus neque vel augue. Cras consectetur libero ac eros. Ut eget massa. Fusce sit amet enim eleifend sem dictum auctor. In eget risus luctus wisi convallis pulvinar. Vivamus sapien risus, tempor in, viverra in, aliquet pellentesque, eros. Aliquam euismod libero a sem.

Nunc velit augue, scelerisque dignissim, lobortis et, aliquam in, risus. In eu eros. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Curabitur vulputate elit viverra augue. Mauris fringilla, tortor sit amet malesuada mollis, sapien mi dapibus odio, ac imperdiet ligula enim eget nisl. Quisque vitae pede a pede aliquet suscipit. Phasellus tellus pede, viverra vestibulum, gravida id, laoreet in, justo. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Integer commodo luctus lectus. Mauris justo. Duis varius eros. Sed quam. Cras lacus eros, rutrum eget, varius quis, convallis iaculis, velit. Mauris imperdiet, metus at tristique venenatis, purus neque pellentesque mauris, a ultrices elit lacus nec tortor. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent malesuada. Nam lacus lectus, auctor sit amet, malesuada vel, elementum eget, metus. Duis neque pede, facilisis eget, egestas elementum, nonummy id, neque.

A. Onderzoeksvoorstel

Het onderwerp van deze bachelorproef is gebaseerd op een onderzoeksvoorstel dat vooraf werd beoordeeld door de promotor. Dat voorstel is opgenomen in deze bijlage.

A.1 Introductie

De meeste e-commerce online platformen hebben reeds een zoekfunctie ingebouwd, maar deze is vaak beperkt tot enkel de productcatalogus. Rekening houdend met de algemene trend rond personalisering van de customer experience zou het ook aangewezen zijn om ook de search op de e-commerce site te personaliseren. Dit zou resulteren in een betere experience voor de klant en een hogere conversie voor het bedrijf. In deze bachelorproef wordt onderzocht in hoeverre de persoonlijke data van een specifieke gebruiker en contextuele data kan ingeschakeld en gecombineerd worden met de data uit de productcatalogus, om zo persoonlijke en relevante zoekresultaten te genereren. Met deze persoonlijke data wordt bedoeld de historische aankoopdata van de persoon zelf, maar ook bepaalde demografische data zoals geslacht, leeftijd, gezinssamenstelling, etc. De onderzoeksvraag kan dus als volgt geformuleerd worden: Welk databankmodel presteert het best om gepersonaliseerde zoekresultaten te bekomen?

A.2 State-of-the-art

Personalized Search (Pitkow e.a., 2002) verwijst naar het zoeken op het web waarbij de resultaten afhankelijk zijn van de interesses en voorkeuren van de gebruiker die verder gaan dan de query zelf.

Personalisatie in e-commerce toepassingen biedt een groot voordeel aan bedrijven. De loyaliteit van klanten wordt veel sterker als deze gebruik maken van gepersonaliseerde features. (Telang & Mukhopadhyay, 2005) Een zoekfunctie is een voorbeeld van zo een feature. Recommender Systems (Resnick & Varian, 1997) zijn aanbevelingen vanuit het systeem die rekening houden met de beschikbare informatie van gebruikers en hun voorkeuren om zo een filter te plaatsen op de informatie die weergegeven wordt.

De studie van Diehl (2003) onderzocht het effect van gepersonaliseerde zoekresultaten op de kwaliteit van keuzes die klanten maken, en vond een positieve correlatie. De studie ontdekte dat het verlagen van search cost (Smith, Venkatraman & Dholakia, 1999) leidde tot minder kwaliteitsvolle keuzes. De reden daarvoor is dat klanten slechtere beslissingen maken als de search cost lager ligt omdat zij minder ideale opties aangeboden krijgen. Personalized Search en Recommender System zorgen voor een enorme verbetering in de kwaliteit van de keuzes die de klant maakt, en verminderen het aantal producten die deze klant bekijkt alvorens hij/zij gevonden heeft wat hij/zij nodig heeft.

Een gevolg van gepersonaliseerde zoekopdrachten is dat we een Filter Bubble (Pariser, 2011) creëren. Het verlaagt de kans dat nieuwe informatie gevonden wordt doordat de resultaten van een zoekopdracht partijdig zijn en eerder wijzen naar dingen die de gebruiker reeds gezien heeft. Dit concept wordt een Filter Bubble genoemd omdat gebruikers eigenlijk geïsoleerd worden in hun eigen wereldje, waar ze enkel de informatie te zien krijgen die ze willen zien. Als we deze gebruikers met hun bubbels in groepen opdelen, verkrijgen we wel het probleem dat deze een vertekend beeld op de realiteit krijgen, zij krijgen bijvoorbeeld in het nieuws slechts het deel te zien dat voor hun interessant is. Als we deze lijn doortrekken naar de klanten van e-commerce websites, zullen zij ook slechts de merken te zien krijgen waar hun voorkeur naar uit gaat. Hierdoor verminder je de kans dat ze een nieuw merk ontdekken of een ander product uitproberen. Als we terug refereren naar de studie van Diehl (2003), dan kunnen we afleiden dat dit een positief effect zal hebben op de klanttevredenheid.

A.3 Methodologie

Om de onderzoeksvraag te beantwoorden wordt er een simpele webapplicatie opgezet waar een zoekterm kan ingevoerd worden. Deze webapplicatie zal louter gebruikt worden om met behulp van een tekstveld een query op een databank uit te voeren. Ook worden er twee databankmodellen ontworpen, een model dat gebruik maakt van Neo4j (Graph databank platform), en een model dat gebruik maakt van Elasticsearch (OLAP databank platform) en Kibana om de resultaten te visualiseren. Graph is een databankstructuur van het type NoSQL, dit wil algemeen zeggen dat ze geen gebruik maken van SQL. Bij Graph databanken worden gegevens voorgesteld door een geheel van entiteiten en verbindingen, alsook vrije relaties tussen deze entiteiten. Kortom is dit, zoals de naam al laat vermoeden, een graaf. OLAP is de afkorting voor Online Analytical Processing, dit is een technologie die geoptimaliseerd is voor het uitvoeren van query's en rapporten in plaats van transacties. Deze zullen elk met hun eigen API communiceren, en in beide modellen wordt dezelfde gebruiker- en productdata ingevoerd. Bij het opstellen van de modellen wordt mogelijk

al duidelijk of één van de modellen niet in staat zal zijn om dezelfde functionaliteiten te hebben als het andere, en dan zal moeten afgewogen worden of de voordelen van het ene model opwegen tegenover het andere model. Als beide modellen dezelfde functionaliteit kunnen bereiken, wordt er bekeken welke het meest performante is. Mogelijks zou het ook haalbaar zijn om beide manieren te combineren op voorwaarde dat de responstijd binnen de acceptabele norm valt.

A.4 Verwachte resultaten

Er wordt verwacht dat er ofwel een duidelijk verschil merkbaar is in performantie tussen de twee verschillende modellen. Mogelijk is dat één van de twee modellen totaal niet haalbaar is om efficiënt een link mee te leggen tussen bijvoorbeeld familieleden, in dit geval bekijken we of het mogelijk is deze twee modellen samen uit te voeren, als dit een resultaat biedt dat binnen de norm valt qua performantie, dan is dit ook een mogelijke oplossing. Het kan zich ook voordoen dat beide modellen vrij performant en efficiënt de query's kunnen verwerken, in dit geval worden de voor- en nadelen alsook de moeilijkheidsgraad van implementatie afgewogen

A.5 Verwachte conclusies

Er wordt verwacht dat het Graph model makkelijker te implementeren zal zijn en beter zal presteren als we rekening willen houden met het aankoopgedrag van vrienden en familie. Indien dit ook mogelijk is bij een OLAP-model, verwachten we dat Graph nog steeds beter zal presteren. Indien we hier geen rekening mee houden zal het OLAP-model beter presteren, aangezien dit model aangepast is aan grote hoeveelheden data waar complexe zoekopdrachten op kunnen uitgevoerd worden.

Bibliografie

- Adomavicius, G. & Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 734–749.
- Dehkordi, G. J., Rezvani, S., Rahman, M. S., Nahid, F. F. N. & Jouya, S. F. (2012). A conceptual study on E-marketing and its operation on firm's promotion and understanding customer's response. *International Journal of Business and Management*, 7(19), 114.
- Diehl, K. (2003). Personalization and decision support tools: Effects on search and consumer decision making. *ADVANCES IN CONSUMER RESEARCH*, VOL 30, 30, 166–167.
- Linden, G., Smith, B. & York, J. (2003). Amazon.Com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing*, 7(1), 76–80. doi:10.1109/MIC.2003.1167344
- NeoTechnology. (2014). Case Study: Walmart uses Neo4j to give customers best web experience through relevant and personal recommendations. Verkregen van http://dev.assets.neo4j.com.s3.amazonaws.com/wp-content/uploads/Neo4j_CS_Walmart.pdf?_ga=1.189104616.2108618949.1430736703
- Netflix. (2009). The Netflix Prize. Verkregen 29 maart 2020, van <https://www.netflixprize.com>
- Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Group, The.
- Pitkow, J., Schütze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., ... Breuel, T. (2002). Personalized search. *Communications of the ACM*, 45(9). doi:10.1145/567498.567526
- Pollefiët, L. (2011). *Schrijven van verslag tot eindwerk: do's en don'ts*. Gent: Academia Press.

- Resnick, P. & Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, 40(3), 56–58. doi:10.1145/245108.245121
- Sarwar, B., Karypis, G., Konstan, J. & Riedl, J. (2001). Item-Based Collaborative Filtering Recommendation Algorithms. In *Proceedings of the 10th International Conference on World Wide Web* (pp. 285–295). WWW '01. doi:10.1145/371920.372071
- Smith, G. E., Venkatraman, M. P. & Dholakia, R. R. (1999). Diagnosing the search cost effect: Waiting time and the moderating impact of prior category knowledge. *Journal of Economic Psychology*, 20(3), 285–314. doi:10.1016/s0167-4870(99)00010-0
- Telang, R. & Mukhopadhyay, T. (2005). Drivers of Web portal use. *Electronic Commerce Research and Applications*, 4(1), 49–65. doi:10.1016/j.elerap.2004.10.004
- Vavliakis, K. N., Katsikopoulos, G. & Symeonidis, A. L. (2019). E-commerce Personalization with Elasticsearch. *Procedia Computer Science*, 151, 1128–1133. The 10th International Conference on Ambient Systems, Networks and Technologies (ANT 2019) / The 2nd International Conference on Emerging Data and Industry 4.0 (EDI40 2019) / Affiliated Workshops. doi:<https://doi.org/10.1016/j.procs.2019.04.160>