Een datavisualisatie:
Wie zijn de nakomelingen van een wetenschappelijke publicatie?

Project

Minor Programmeren

Onno Hartveldt

(10972935)

juni 2015

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Universiteit van Amsterdam

docent: dr. Thijs Coenen

begeleiding: Jelle van Assema BSc.

Inleiding

In dit verslag wordt ingegaan op het tot stand komen van een datavisualisatie en de verzameling van de data waarop de visualisatie gebaseerd is. De visualisatie is gemaakt met HTML, CSS, JavaScript en D3. De dataverzameling is gedaan met Python. De structuur van het verslag is in vier fases ingedeeld aan de hand van het Nested Model voor visualisatie [1]. In de fases wordt behandeld de centrale vraag, de herkomst van de data en de representatie ervan, het ontwerp en interactie en de implementatie. Tot slot wordt de visualisatie geëvalueerd.

Centrale vraag

De vraag die centraal staat in de visualisatie is welke wetenschappelijke artikelen vinden gedeeltelijk hun oorsprong in een referentie die je wilt gaan gebruiken in je eigen essay. Zo zou je de publicatie van interesse voor gebruik op waarde kunnen schatten en is er misschien een recenter artikel ook van toegevoegde waarde voor je essay.

Herkomst van de data

De data gebruikt voor de visualisatie is gebaseerd op informatie weergegeven op de website "Web of Science" [2], eigendom van de informatiedienst Thomson Reuters. De gegevens die van de website worden gehaald vinden zijn oorsprong in een opgegeven publicatie. Van alle publicaties die de opgegeven publicatie citeren wordt de titel, auteurs, DOI nummer en door welke publicaties dit artikel wordt gerefereerd opgeslagen. Voor alle publicaties waarvan de gegevens worden opgeslagen wordt hetzelfde proces herhaald.

Data representatie

De gegevens van de publicaties kunnen worden gerepresenteerd als graaf. Een graaf is een set van knopen die al dan niet met elkaar verbonden zijn. Een publicatie kan zich worden voorgesteld als een knoop en is verbonden met een andere knoop als de ene publicatie de ander citeert.

De gegevens van de publicaties zijn omgezet naar een lijst met knopen en een lijst met verbindingen. De knopen zijn een *dictionary* en bevatten de informatie uit de artikelen. De verbindingen zijn eveneens een *dictionary* en bevatten simpel weg de informatie van een bronknoop en een doelknoop.

Ontwerp en interactie

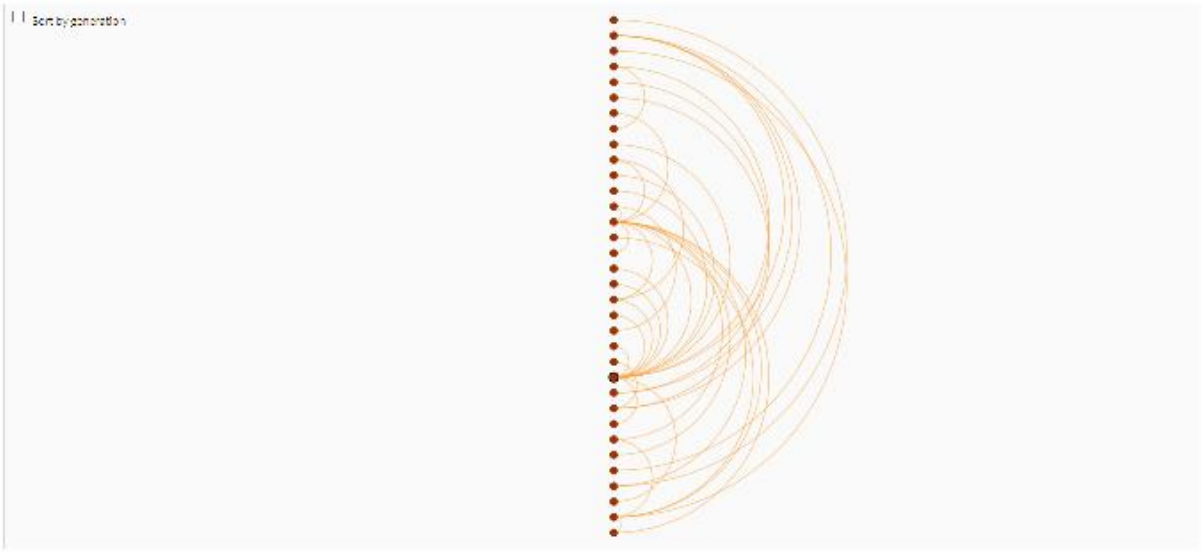
Een publicatie wordt gerepresenteerd door een knoop en is weergegeven als een cirkel. Een citatie is een verbinding tussen twee publicaties en wordt weergeven als een lijn. De informatie wordt weergeven in een Arc-diagram. De knopen worden onder elkaar gezet op alfabetische volgorde van de titel. Er wordt een gebogen lijn getekend tussen verbonden knopen. Een knoop met meer verbindingen dan andere knopen is mogelijk interessant door de hoeveelheid waarin het geciteerd wordt door andere publicaties.

De visuele variabelen van Bertin die gebruikt zijn in de Arc-diagram zijn als volgt [3]. Er worden twee verschillende vormen associatief gebruikt. De cirkel representeert de publicatie en de gebogen lijn de verbinding. Het verschil in grootte van de cirkel is associatief, de enkele grotere cirkel is oorspronkelijk de publicatie van interesse. De kleur intensiteit wordt als ordening gebruikt. Des te intenser de kleur des te meer van belang het object van interesse is. Zowel de knoop als de verbinding kan van kleur intensiteit wisselen. De positie van de knopen wordt na het sorteren als ordening gebruikt. De bovenste knoop komt voort uit de verbonden knoop er onder.

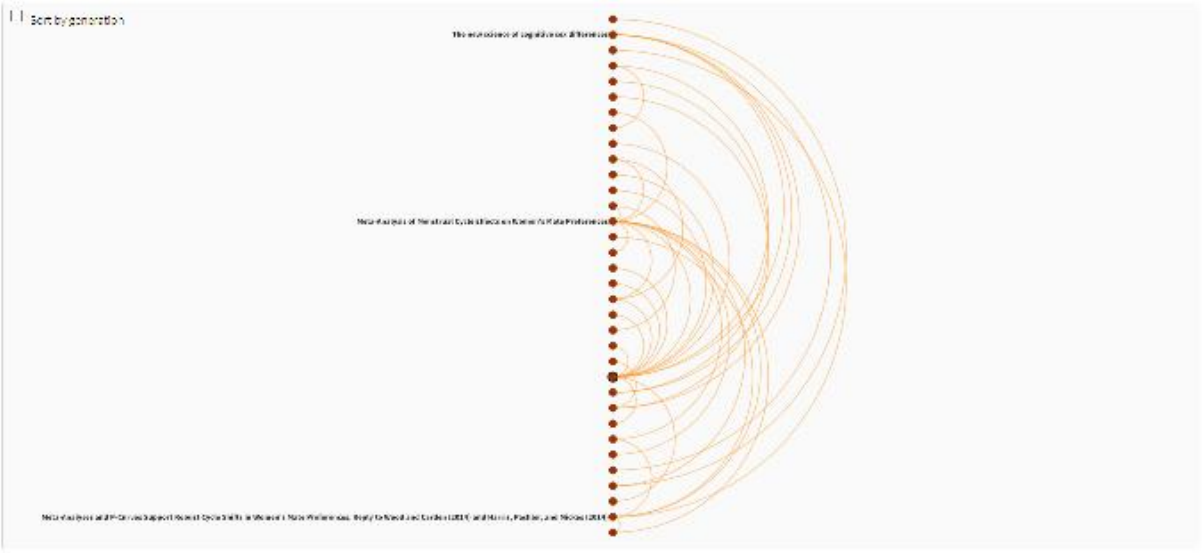
De visualisatie heeft naast het uiterlijk van de Arc-diagram vier mogelijke interacties om de onderlinge relaties van de publicaties te ontdekken. Wanneer de muis een knoop raakt komt de titel

van het artikel tevoorschijn. Bij het verwijderen van de muis van de knoop verdwijnt de titel weer gelijktijdig.

Interactie 1

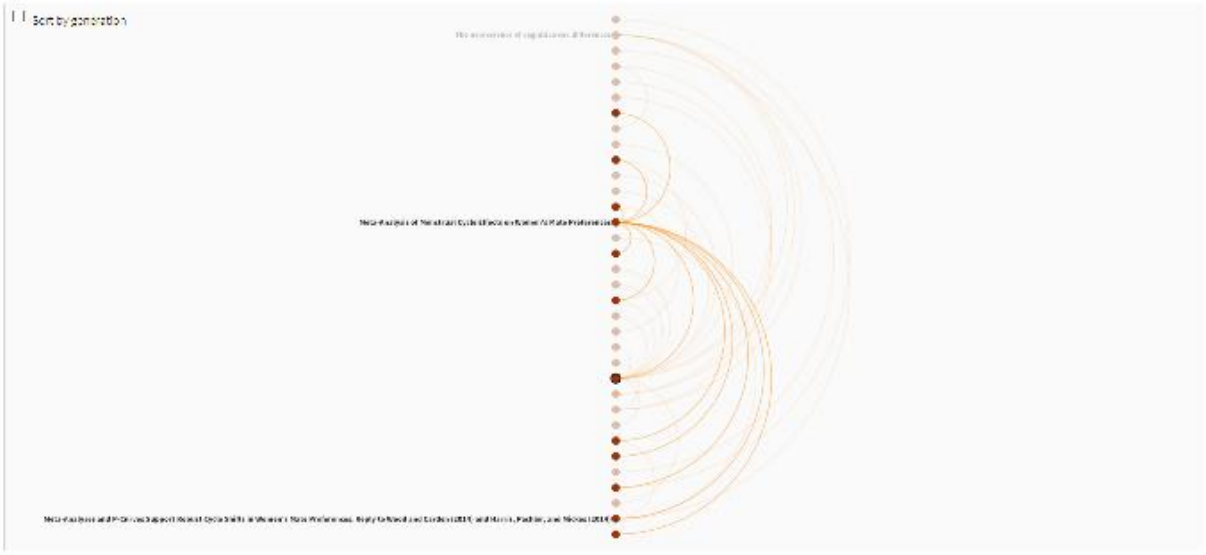


Wanneer een publicatie mogelijk interessant genoeg is om te markeren, kan door met de muis een enkele klik te geven op de knoop de titel langdurig zichtbaar gemaakt worden. Nogmaals een enkele klik zorgt ervoor dat de titel weer van het scherm verdwijnt.



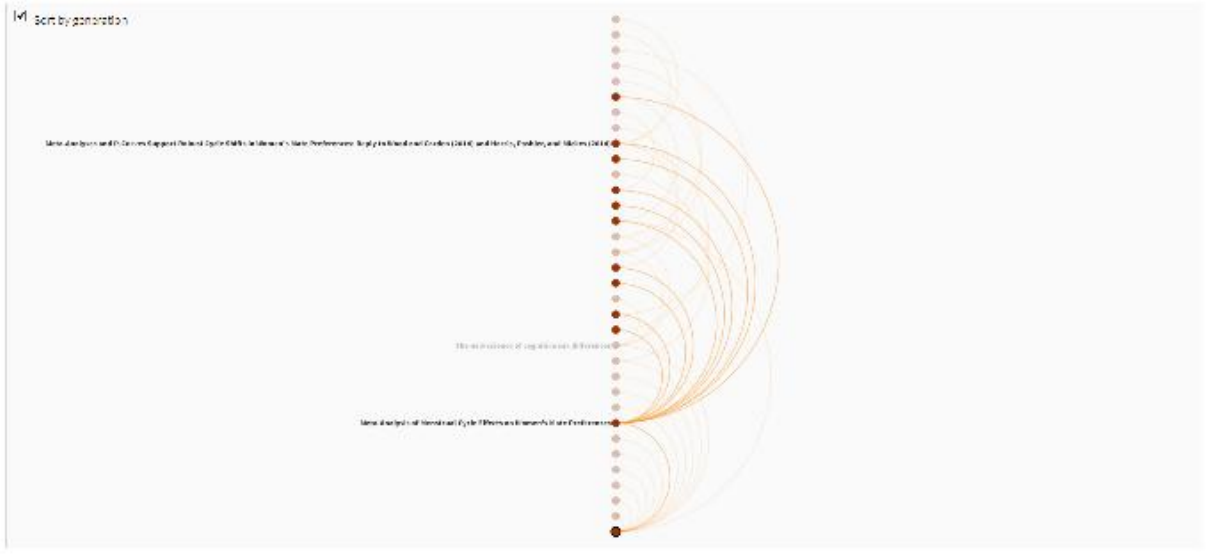
Interactie 2

Om de relaties van een publicatie te benadrukken kan dubbel geklikt worden op de desbetreffende knoop. De andere knopen worden dan in kleur intensiteit geminderd. Nogmaals een dubbel klik op een knoop en de intensiteit wordt weer hersteld.



Interactie 3

De laatste interactie die mogelijk is betreft de sortering van de knopen. In een verse pagina zijn de knopen ongerelateerd aan de relatie tot andere publicaties door gesorteerd te zijn op de lengte van de titel. Onbewust kan dit worden gebruikt om positie veranderingen te detecteren. Na het markeren van de *checkbox* veranderen de positie van de knopen. De knopen sorteren zich per generatie. De oorsprong publicatie staat onderaan met de publicaties, die het artikel citeren, direct daarboven, zij vormen generatie 1. Daarboven de publicaties die volgen uit generatie 1 artikelen, etc. De meest recent gepubliceerde artikelen staan dus bovenaan.



Interactie 4

Implementatie

Dataverzameling: de informatie is verkregen door een python script die de informatie over artikelen van de website haalt en opslaat als een JSON bestand. Het script maakt gebruik van twee verschillende bibliotheken om de interactie met de website mogelijk te maken, Pattern.web [3] en Selenium [4]. Een combinatie van deze bibliotheken zijn gebruikt om met Pattern.web HTML elementen op te slaan en met Selenium over JavaScript gegenereerde delen van de website te kunnen navigeren. Het script bestaat uit een deel wat ruwe data opslaat vanaf de website en een deel wat de ruwe data omzet naar knopen en verbindingen. Het eerste deel heeft functies om per publicatie de informatie op te slaan, een functie om een lijst met URL's te verzamelen van artikelen die de huidige publicatie refereren en een recursieve functie om de informatie van alle publicaties te verzamelen die voortvloeiden uit de publicatie die als uitgangspunt is genomen. Dit resulteert in een lijst met publicaties die worden geciteerd. Het tweede deel zet die lijst om in knopen ook als de publicatie nog niet wordt geciteerd, voor die knoop geldt dat het slechts één verbinding heeft.

Visualisatie: het algoritme bestaat uit statische functies, zoals het tekenen van de knopen en de verbindingen en dynamische functies. De dynamische functies zijn het sorteren van de artikelen en de locatie op het scherm die daar mee gemoeid is veranderen, het oplichten van verbonden artikelen en het langdurig weergeven van een titel van een potentieel interessante publicatie.

Validatie

De evaluatie van de visualisatie wordt behandeld in vier fases [1], centrale vraag, de herkomst van de data en de representatie ervan, het ontwerp en interactie en de implementatie. Een enkele factor heeft invloed op alle vier de fases van het ontwikkelen en verbeteren van de visualisatie. Het aantal publicaties dat zijn oorsprong vindt in een enkele publicatie kan heel snel extreem groot worden. Het gevolg voor de centrale vraag is dan of het nog wel nut heeft om zoveel publicaties te bekijken? De herkomst van de data is al snel storend, de interactie met de website kost erg veel tijd en is instabiel. Een verbetering zou een API zijn, maar dat is kostbaar. Het ontwerp om alle knopen onder elkaar weer te geven schiet te kort boven een groot aantal. Het combineren van een verticale weergave met een horizontale weergave zou uitkomst kunnen bieden. Zo zou een planare weergave van de graaf kunnen ontstaan. De implementatie van de filter algoritme op bijvoorbeeld het alleen weergeven van knopen met meer dan een verbinding. Of filter op een maximaal aantal generaties en het domein beperken van de publicatie data.

Blijft het aantal nakomelingen dat volgt uit een wetenschappelijke publicatie klein dan kan je met deze visualisatie een inzicht krijgen wie het zijn en hun relaties tot andere publicaties.

Referenties

[1] T. Munzner. A Nested Model for Visualisation Design and Validation. In IEEE Transactions on visualisation and computer graphics 15(6), pagina's 931-928, 2009.
[2] http://apps.webofknowledge.com/
[3] M.S.T. Carpendale. Considering Visual Variables as a Basis for Information Visualisation
[4] http://www.clips.ua.ac.be/pattern
[5] https://selenium-python.readthedocs.org/