Social Data Analysis

ten Thij, M.C.

2018

document version
Publisher's PDF, also known as Version of record

Samenvatting

In de afgelopen jaren is de manier waarop mensen met elkaar communiceren op significante wijze veranderd, vooral dankzij de opkomst van sociale media. Deze verandering heeft het internet veranderd in een persoonlijker en meer interactief medium, waar sociaal netwerken een van de top-activiteiten is. De overvloed aan informatie die wordt gegenereerd op deze online platforms, heeft geleid tot veel nieuwe ontwikkelingen en onderzoeksgebieden in de wetenschap, zoals "complex science" en "computational social science". Ons doel is om deze overvloed aan beschikbare informatie, die we sociale data noemen, te gebruiken om meerwaarde te creëren. In dit proefschrift verkennen we daarom het veld van sociale data analytics. In het bijzonder bekijken we drie subgebieden van dit veld. Het eerste subgebied dat we bekijken is een van de prominente velden in sociale data analyse, dat zich richt op de vraag: "Wat drijft trend gedrag in online platformen?" Het tweede subgebied behelst het voorspellen van toekomstige gebeurtenissen op basis van huidig gedrag en activiteit op het platform. Het derde subgebied betreft het verwerken van informatie uit online platforms in real-time.

Ons werk op het gebied van trendgedrag richt zich op de onderliggende netwerken die de sociale verbindingen tussen de gebruikers van Twitter beschrijven. In het bijzonder, beschouwen we een zogeheten graaf van gebruikers, waar een verbinding aangeeft dat een van de gebruikers een bericht van de andere gebruiker middels een retweet verspreid heeft. We onderzoeken dus de dynamiek in retweet grafen, met het doel om de effecten van het pieken in activiteit na te bootsen, om zo de aard van trends te onderzoeken. Aangezien we trendgedrag willen modelleren zonder additionele informatie te gebruiken over de gebruikers die de informatie delen, gebruiken we zogeheten willekeurige grafen om een model te ontwerpen dat de ontwikkeling van trend gedrag vangt en het verspreiden van informatie door het sociale netwerk nabootst. We richten ons dus alleen op de verspreidingspatronen van de trends en niet op de inhoud van het populaire bericht. Ons model, dat we het *retweet graph model* (RGM) noemen, is een random graaf model, dat de mogelijke veranderingen in de retweet graaf beschrijft bij het verschijnen van nieuwe berichten

over het onderwerp dat we analyseren. In onze analyse nemen we aan dat het verspreidingspatroon van een enkel bericht volgens een standaard patroon gaat, waarvoor we het zogeheten superstar model gebruiken. Op basis van de definitie van het RGM, leiden we de verwachting en de variantie van zowel de gemiddelde graad als het aantal retweets per bericht analytisch af.

Het is bovendien belangrijk om te weten hoe waarschijnlijk het is dat een bericht frequent gedeeld gaat worden. Daarom analyseren we de graad verdeling van het RGM. Omdat de opzet van het model verscheidene tijdsafhankelijke onderdelen bevat, analyseren we een versimpelde versie van het RGM. Voor deze versimpelde versie leiden we af dat gebruikers met een hoge graad niet ongewoon zijn. Door deze uitkomsten vervolgens te vergelijken met simulatieresultaten van het RGM, vinden we dat het versimpelde RGM vergelijkbare graadverdelingen geeft als het RGM.

Als een onderwerp trending is, wordt het door een grote groep besproken op dat moment. De wetenschap hoe groot die groep is, en of dit een enkele groep is of meerdere groepen zijn, kan waardevolle inzichten bieden in de verspreiding van trends in een sociaal netwerk. Daarom bestuderen we ook de verdeling van de component grootte van het RGM. We tonen aan dat deze verdeling afhankelijk van de parameterwaarden óf gelijk is aan óf benaderd kan worden door de verdeling van de urn grootte in het Pólya proces. Verder bestuderen we ook de grootte van het grootste component (LCC) van de retweet graaf. Voor de LCC tonen we aan dat deze LCC onder bepaalde condities een zogeheten giant component is. Bovendien levert onze analyse vier verschillende regionen van gedrag voor de grootte van de LCC op.

Een analyse van verschillende retweet grafen laat zien dat de grootte van de LLC en de lijndichtheid, die gedefinieerd is als het aantal lijnen gedeeld door het aantal gebruikers, de meeste voorspellende waarde hebben met betrekking tot de piek activiteit in Twitter. In de ontwikkeling van de lijndichtheid vinden we een interessant fenomeen dat we de *densification* van de LCC noemen. Deze densification geeft het moment weer dat meerdere componenten van de retweet graph aan elkaar verbonden worden in de LCC. Dit moment correspondeert met het moment dat meerdere groepen die het onderwerp bespreken, bij elkaar komen. Deze perceptie kan vervolgens tot een versnelling van de discussie leiden, waardoor trendgedrag ontstaat. Tot slot analyseren we de voorspellende kracht van ons model op basis van data die verworven is uit Twitter. Op basis van deze data kunnen vervolgens de parameterwaarden van het RGM geschat worden, waaruit blijkt dat het model de lijndichtheid van de LCC het beste vangt. Verder zijn de voorspellingen het beste voor datasets waar er slechts één enkele piek in de activiteit is.

Het analyseren van de activiteit dynamiek van online platformen, het tweede aspect dat we analyseren, heeft geleid tot interessante resultaten in de zoektocht naar het begrip van menselijke interacties. In ons onderzoek bekijken we de dynamiek in de activiteit op twee platformen: Twitter en Wikipedia. Als een bijproduct van onze analyse, tonen we aan dat beide platforms een duidelijk dagritme in hun globale activiteit vertonen.

We bestuderen de invloed van het promoten van een pagina op het aantal

keren dat deze pagina bekeken wordt op Wikipedia. Nadat we meerdere talen met elkaar vergeleken hebben (namelijk Engels, Spaans, Duits en Nederlands), tonen we aan dat de activiteit exponentieel afneemt en gemodelleerd kan worden met slechts twee parameters. Vervolgens tonen we aan dat ons model gelinkt kan worden aan het zogeheten Poisson proces. Ook presenteren we een algoritme om de precieze verspreiding van een bericht in Twitter te achterhalen. Aangezien we ontdekken dat een groot deel van de retweets direct vanaf het originele bericht komen, onderstreept deze analyse de eerdere aanname in het RGM waarbij we het superstar model gebruikten voor de verspreiding van een enkel bericht.

In het derde deel van dit proefschrift kijken we naar de mogelijkheid om meerwaarde te halen uit een real-time informatiestroom van sociale media berichten voor de tuin- en uitbouw sector, waarbij we ons in eerste instantie richten op berichten uit Twitter. De tuin- en uitbouw sector is een traditionele sector, waarin kwekers gericht zijn op productie en waarin handelaren hun eigen transacties gebruiken als grootste informatiebron. Het bijhouden van wanneer en op welke manier de producten uit de sector besproken worden op sociale media, is een belangrijke toevoeging om actief te luisteren naar de klant. Dit kan naast de huidige technieken die in deze industrie gebruikt worden.

Onze eerste onderzoeken leverden twee resultaten op; we vonden een correlatie tussen verkoopcijfers en het aantal Twitter berichten van twee producten en we toonden aan dat de informatie uit sociale media gebruikt kan worden voor effectmetingen van campagnes. Deze resultaten geven aan dat er meerwaarde zit om sociale media berichten te gebruiken om de dagelijkse gang van zaken te kunnen ondersteunen. Om de berichten die deze informatie bevatten in real-time binnen te kunnen halen, hebben we een systeem genaamd de *Hortiradar*, ontwikkeld. Dit systeem gebruikt technieken uit het vakgebied genaamd "Natural Language Processing" om de data in real-time te verwerken en vervolgens op product-niveau te visualiseren. Bovendien ontwikkelen we het *storify* algoritme dat onderwerpen detecteert in de berichtenstroom en de ontwikkeling van deze onderwerpen verder volgt. Een vergelijking van de resultaten van ons algoritme en een clustering op de complete dataset laat zien dat ons algoritme net zo goed werkt. Vervolgens bestuderen we de invloed van de parameters van het algoritme op de resultaten. Op basis hiervan vinden we de beste parameter instelling voor ons algoritme, die we vervolgens implementeren in de Hortiradar.

Op basis van de terugkoppeling van onze industriepartners, geeft de Hortiradar een goed overzicht van wat er op dit moment besproken wordt met betrekking tot de producten uit de sector. Vooral in het geval wanneer een van de producten in de media genoemd wordt, kan de Hortiradar een goed inzicht geven in de huidige publieke opinie die op Twitter geuit wordt.