Met Twitter-data bovenop het nieuws

Wat gebeurt er in de wereld om ons heen? Nieuwsredacties zitten graag zo dicht mogelijk op het nieuws. Sandjai Bhulai, hoofddocent aan de VU in Amsterdam, werkte mee aan RTreporter, een systeem dat automatisch nieuwsonderwerpen uit Twitter-data filtert. Persbureaus zijn enthousiast. RTreporter is de reguliere nieuwskanalen vaak voor. Wat was er (wiskundig) voor nodig om dit te realiseren?

De vakgroep van Sandjai Bhulai, Stochastic Operations Research, werkt veel samen met het bedrijfsleven. Onderzoeksonderwerpen variëren van processen in de gezondheidszorg, luchtvaartplanning en planning in call centers tot social media. Bij al deze onderwerpen speelt toeval (stochastiek) een rol. Het Twitter-project waar Bhulai aan meewerkte is een mooi voorbeeld van wat er met data en wiskunde mogelijk is.

Schatkamer vol informatie

"Een paar jaar geleden werden we benaderd door een bedrijf dat op Twitter nieuws wilde halen voordat dat gebracht werd door de mainstream media", vertelt Bhulai. "Hun ervaring was dat nieuws vaak eerder op internet te vinden is dan in de redactiekamer."
Veel mensen twitteren als er iets gebeurt. Daardoor komen in de Twitter-datastroom dagelijks grote hoeveelheden 'nieuwsberichten' langs. Voor redacteurs is het een schatkamer vol informatie. Maar om uit alle berichten – plus alle andere nieuwskanalen – relevante nieuwsonderwerpen te zoeken, is zoeken naar een speld in een hooiberg. "De vraag was of we een systeem konden ontwikkelen dat automatisch nieuwsonderwerpen uit Twitter-data tevoorschijn haalt. Met steun uit het Stimuleringsfonds voor de Pers konden we deze vraag onderzoeken en hebben we de Real Time News Reporter ontwikkeld, kortweg RTreporter."

Twitter stelt ongeveer 1% van zijn data gratis beschikbaar voor derden. Dit zijn random berichten uit de totale stroom. Met deze data gingen Bhulai en zijn medewerkers aan de slag. "De vraag gaat over Nederlandse nieuwsonderwerpen. Daarom moeten we in de beschikbare data eerst de Nederlandse Twitter-berichten zoeken. Samen met een linguïstische groep hebben we 133 woorden geselecteerd – zoals als en ik – waarmee je Nederlandse berichten kunt herkennen. De gevonden berichten splitsen we op in woorden waarvan we tellen hoe vaak ze voorkomen. Berichten met woorden die vaker voorkomen dan normaal onderzoeken we nader. Bij deze berichten kijken we of we clusters kunnen ontdekken. De een heeft het over brand in Amsterdam, de ander over een fikkie op de Dam en weer een ander heeft het over 020. Deze berichten gaan allemaal over hetzelfde onderwerp. Om te clusteren gebruiken we vectoren in een meer-dimensionale ruimte. Elk bericht wordt weergegeven door een vector.
We onderzoeken of vectoren dicht bij elkaar liggen of niet. Aan de hand van drempelwaarden stellen we vast of termen wel of niet bij elkaar horen. Berichten met veel gemeenschappelijke termen liggen op kleine afstand van elkaar. Hetzelfde geldt als ze veel termen bevatten die in de grote bulk niet voorkomen."

Niet delen door nul

Het systeem houdt bij hoe snel een cluster groeit. Als dit snel toeneemt is de kans groot dat er iets aan de hand is. Snelheid en versnelling bepaal je met de afgeleiden van de datastroom. "Wiskundig gezien hebben we daarbij een praktisch probleem. Om de snelheid te bepalen, kijken we naar

$$\frac{\Delta x}{\Delta t}.$$

Maar Twitter rondt de timestamp van een bericht op secondes af. Je krijgt daardoor veel berichten op hetzelfde tijdstip zodat

$$\Delta t = 0.$$

Delen door nul kan niet. We lossen dit op door het tijdstip van elk bericht een klein beetje te verschuiven tot maximaal 1 seconde. Zo krijg je een gladde functie die je kunt differentiëren en zo kunnen we de snelheid en versnelling van clusters berekenen."

RTreporter in gebruik bij het ANP, het linkerscherm toont een top-10 van gedetecteerde onderwerpen, het rechterscherm toont de grootte van een geselecteerd cluster als functie van de tijd

Als een cluster snel groeit, betekent dat niet meteen dat hij nieuwswaarde heeft. “Overdag twitteren tieners veel mopjes naar elkaar”, legt Bhulai uit. “En op vaste tijdstippen van de dag zie je vaste clusters zoals ‘goedemorgen’, ‘slaap lekker’ of met commentaar op een populair tv-programma. Met trenddetectie filteren we deze berichten eruit. De top-10 van de (wiskundig gezien) opvallendste berichten wordt tenslotte zichtbaar gemaakt op een dashboard.” Vanaf dat moment neemt de redacteur het over van het systeem. Is een bericht de moeite waard om uit te diepen? Sturen we er een verslaggever op af? RTreporter blijkt een welkome aanvulling in de redactiekamer. Via de reguliere kanalen komen berichten vaak later binnen of helemaal niet. Vooral regionaal nieuws komt door het systeem beter aan bod.

“Ook met kleine hoeveelheden data kun je slimme en nuttige dingen doen.”

Steeds meer slimme toepassingen

Data wordt een steeds belangrijker onderdeel van onze samenleving. RTreporter is daar een mooi voorbeeld van. “Er wordt veel gesproken over ‘Big data’ maar je kunt het beter over ‘Smart data’ hebben”, vindt Bhulai. “Ook met kleine hoeveelheden data kun je slimme en nuttige dingen doen. Denk bijvoorbeeld aan computerondersteunde gezondheidszorg.” De vraag naar zorg neem toe, maar het aanbod is stabiel. Er ontstaat een mismatch tussen vraag en aanbod. Voor een deel kan dit opgevangen worden door slimmer met data om te gaan. Gegevens over bloeddruk, bewegen of de inname van medicatie kun je bijvoorbeeld via een smartphone bijhouden. Als er iets mis gaat, gaat er (bij de huisarts) een alarmfunctie af. Dit soort toepassingen levert nieuwe discussies. “Belangrijk wordt dan de vraag tot waar de arts de controle heeft en tot waar de patiënt”, aldus Bhulai.

Het vak data science is volop in ontwikkeling. Met behulp van wiskunde ontstaan er steeds meer slimme toepassingen van data. “Ik ben zelf nog regelmatig verrast door wat er in dit vakgebied mogelijk is”, constateert Bhulai.

Kijktip: Op https://www.youtube.com/watch?v=eTJLM8F1Xfk vindt u een TED-presentatie van Sandjai Bhulai over RTreporter. Aan het eind van zijn verhaal (rond minuut 9 van het filmpje) demonstreert Bhulai met hulp van het publiek real time de mogelijkheden van het systeem.