'Big data vertalen naar statistieken is dé grote uitdaging'

CBS onderzoekt de mogelijkheden van big data

Het is een hot issue: big data. Ondernemingen zitten op extreme hoeveelheden bits en bytes. Digitale informatie die kan helpen bij het klantgerichter sturen van de onderneming. Ook het CBS heeft te maken met big data. Methodoloog Piet Daas onderzoekt, samen met een aantal collega's, hoe het statistiekbureau daarvan gebruik kan maken.

Auteur: *Jaap van Sandijk* Fotografie: *Sjoerd van der Hucht*

Het CBS is één van de eerste statistiekbureaus ter wereld dat de mogelijkheden van big data onderzoekt. Al sinds 2009 – ruim voordat de term big data in zwang raakte – zoekt het naar aanvullende, externe informatiebronnen. 'Alleen noemden we het toen nieuwe bronnen en niet big data', vertelt Piet Daas. 'Denk aan verkeerslussen bijvoorbeeld. Voor statistieken over verkeer en vervoer zijn daar interessante gegevens uit te halen.' Maar big data vertalen naar statistieken is erg lastig – en dat is nog zacht uitgedrukt. 'Dat is dé grote uitdaging', bevestigt de onderzoeksleider. 'Als CBS zijn we gewend om kleinere hoeveelheden data binnen te halen en te werken met steekproeven. Maar nu liggen enorme hoeveelheden data buiten voor het oprapen en gaan we van data-schaarste naar data-overdaad. We worden er echt door overdonderd. Eén van de belangrijkste vragen waar we nu in ons onderzoek voor staan, is: welke methode heb je nodig om uit een big databron een betrouwbare CBS-statistiek te maken?'

SOCIAL MEDIA MONITORING

Met de komst van het innovatieprogramma binnen het CBS in 2011 is de snelheid van het onderzoek van Daas en zijn collega's verhoogd. De resultaten van de eerste proeven zijn positief. 'Zo hebben we in samenwerking met Coosto, een Eindhovens bedrijf dat is gespecialiseerd in social media monitoring, het sentiment in bijna een miljard social mediaberichten bepaald en vergeleken met het consumentenvertrouwen van het CBS. Hierbij

Piet Daas, methodoloog bij het CBS

‘Het sentiment dat via social media wordt geuit, beweegt mee met het consumentenvertrouwen’

Symposium over big data

Het CBS verricht sinds 2009 onderzoek naar mogelijkheden om nieuwe databronnen zoals internet, smartphone-meetingen en grote en complexe bestanden (zogenaamde big data) te ontsluiten voor het maken van statistiek. Ook is er een initiatief om te kijken naar de eventuele voordelen van het opzetten van een eigen webpanel voor het CBS. Op 6 juni van dit jaar werd daarover bij het CBS een symposium gehouden, waar deze onderwerpen uitgebreid aan de orde kwamen. In het volgende nummer van het relatiemagazine kunt u hier meer over lezen.

bleek dat het sentiment dat via social media wordt geuit met het consumentenvertrouwen mee beweegt.' Dat lijkt vreemd, omdat de steekproeven waarop het consumentenvertrouwen is gebaseerd representatief zijn en social media niet. Maar, zo zegt Daas, hier komt de wet van de grote aantallen om de hoek kijken. Daarbij is het een voordeel dat in Nederland massaal gebruik wordt gemaakt van social media. 'De massa aan social mediaberichten maakt het mogelijk een stabiel cijfer op te leveren. De penetratie van social media in Nederland is bovendien hoog. Zeventig procent van de Nederlanders is er actief.' Over het gebruik van big data voor de statistiek bestaan wereldwijd nog nauwelijks theorieën. De bevindingen van Daas en zijn team zijn daarmee op z'n minst opmerkelijk. 'Toen ik die correlatie zag dacht ik: we hebben iets bijzonders te pakken, zeker omdat het reproduceerbaar is.'

REAL TIME BESCHIKBAAR

De kracht van big data is dat de informatie *real time* beschikbaar is, waardoor snel resultaten bekend zijn. Maar dan moet je uit die enorme berg data wel snel de juiste gegevens kunnen halen en deze kunnen 'lezen'. Daas: 'Daarvoor moeten nieuwe methoden en middelen worden ontwikkeld. Dat is de uitdaging waar we nu voor staan. In elk geval is visualisering heel belangrijk. Wij doen dat onder andere met *heat maps*, die op grafische wijze data weergeven.' Voor welke statistieken zou het gebruik van big data interessant zijn? 'Onder meer voor het consumentenvertrouwen, een belangrijke statistiek van het CBS', aldus Daas. 'Deze cijfers worden nu maandelijks gepubliceerd, maar je zou deze kunnen aanvullen met wekelijkse cijfers vanuit de big databron.' Aanvullen, zegt hij met nadruk, want big data zal altijd een extra bron zijn en de steekproeven niet zomaar kunnen vervangen. 'Je moet big data bevindingen altijd kunnen vergelijken met een steekproef. Als je niets hebt om mee te vergelijken, weet je niets.'

VERKEERSLUSSEN

Bij de inzet van data die door verkeerslussen worden gegenereerd, zijn de mogelijkheden legio. 'Elk kwartaal produceert het CBS verkeersindexcijfers', vertelt Daas. 'Deze geven een overzicht van de ontwikkeling van de verkeersdrukte op het totale rijks- en provinciale wegennet buiten de bebouwde kom. Dat gebeurt door de gegevens van een klein aantal lussen te gebruiken. Maar door de grote hoeveelheid data van alle verkeerslussen te gebruiken kunnen we deze cijfers niet alleen vaker publiceren, maar kunnen we ook met voertuigcategorieën werken. Zo kun je bijvoorbeeld een onderverdeling maken in grote en kleine auto's en categorieën per regio maken. In een regio waar de economie opbloeit zou je bijvoorbeeld meer vrachtwagens kunnen waarnemen.' Zoiets zou je zelfs op het water kunnen doen, weet Daas. 'Het aantal sensoren in dijken neemt toe. Die sensoren

‘Je moet big data bevindingen altijd kunnen vergelijken met een steekproef. Als je niets hebt om mee te vergelijken, weet je niets’

zijn in staat veranderingen in het vochtgehalte van dijken te meten, ook als die door bewegingen op het water worden veroorzaakt. Het zou heel mooi zijn als wij daar als CBS ook gebruik van kunnen maken. Big data heeft veel beloften.’ Bovendien, stelt de methodoloog, is het een bron die de lastendruk van bedrijven verder verlaagt.

PRIVACYWETGEVING

Het gebruik van big data door statistiekbureaus moet wel goed worden geregeld, waarschuwt Daas. ‘In de Verenigde Staten en in Engeland loopt men tegen juridische problemen aan, in verband met privacy-wetgeving. Ook voor ons geldt dat het juridisch in orde moet zijn. Doordat je heel veel data binnenhaalt, zullen ongetwijfeld de privacy-eisen worden verhoogd. Waarom zou het CBS mijn sociale mediaberichten moeten weten, zal de burger zich afvragen?’ Een andere hobbel wordt mogelijk gevormd door commerciële partijen. Overheidsorganisaties zullen hun gegevens gemakkelijk afstaan, zoals de verkeerslussen, maar hoe zit het met telecomproviders als ook de smartphone mogelijkerwijs wordt ingezet voor onderzoek? Die zouden hun data wel eens als een aantrekkelijke melkkoe kunnen zien.

Naast deze juridische en de eerder genoemde methodologische uitdagingen zijn er ook organisatorische. Want gebruik maken van big data heeft nogal wat gevolgen voor de interne structuur van het statistiekbureau, dat van oudsher gewend is om data zelf te verzamelen en binnen te halen. ‘Er zullen zwaardere computersystemen moeten komen en methoden ontwikkeld moeten worden om big data te koppelen aan het statistisch proces. En als je meer met big data gaat werken, zul je ook andere mensen nodig hebben. En dan bedoel ik niet alleen statistici, maar vooral ook *data scientists* – de statistici van de toekomst.’ Nu de eerste resultaten van het onderzoek bekend zijn, lijkt het onderzoek van Daas en zijn team in een stroomversnelling te zijn gekomen. En dat verhoogt de spanning, zegt hij opgetogen. ‘Er is behoefte aan snellere statistieken en big data spelen daarbij een belangrijke rol. Het is een databron die op ons afkomt. Daar moeten we zeker iets mee doen’, besluit de onderzoeker.

‘In de Verenigde Staten en in Engeland loopt men met het gebruik van big data tegen juridische problemen aan’