Rubriek: Interview

Tekst: Mirjam Hulsebos

Foto's: inzetjes van de drie profs en stockfoto

Hoe big wordt big brother?

De privacygevaren van big data

Big data is hot. Steeds meer bedrijven onderzoeken hoe zij voorheen losstaande databronnen kunnen combineren om zo tot nieuwe inzichten te komen. Die inzichten zijn vaak baanbrekend, maar er zit ook een groot gevaar aan: privacy. Aan welke nieuwe technologieën wordt op Nederlandse universiteiten gewerkt en welke nieuwe privacyrisico's introduceren die?

COMMIT is een publiek-private onderzoekscommunity die oplossingen ontwikkelt op basis van de nieuwste informatietechnologie. Veel van het onderzoek richt zich op big data analytics. BIM spreekt met drie onderzoekers over de kansen en de risico's die het combineren van meerdere databronnen met zich meebrengt.

Sentimentanalyse
Het COMMIT-project Information Retrieval for Information Services richt zich op de vraag hoe we op een geautomatiseerde manier relevantie kunnen halen uit big data. Maarten de Rijke, hoogleraar Information Retrieval aan de Universiteit van Amsterdam, houdt zich sinds 2011 onder meer bezig met dit project. Hij vertelt: "We hebben dit onderzoek opgedeeld in tien deelprojecten, die zich op drie gebieden richten. De eerste is tekstanalyse. Hieronder valt bijvoorbeeld sentimentanalyse, wat heel hot is onder marketeers. Ze gebruiken tooling om in de gaten te houden wat er op sicial media over hun bedrijf wordt gezegd, maar die is nog niet zo intelligent. Als iemand twittert: 'net een slechte wedstrijd gezien op een verder heel gaaf ABN AMRO toernooi', dan moet je eruit halen dat deze tweet gaat over tennis en niet over de bank. Daarna moet je onderscheiden dat de woorden 'gaaf toernooi' in deze tweet meer zeggen dan 'slechte wedstrijd'. Wij ontwikkelen software die tot dit soort intelligente analyses in staat is."

Hoe ver gaan we?
De tweede categorie projecten houdt zich bezig met de structuur van data. "Dan gaat het om onderlinge verwijzingen tussen documenten, zoals verwijzingen naar databronnen, maar ook verwijzingen naar andere websites, naar mensen en ga zo maar door. Als je door een topic map inzichtelijk kunt maken hoe documenten en databronnen zich tot elkaar verhouden, dan zegt dat veel over de relevantie van zo'n document", weet De Rijke.
De derde categorie projecten is gericht op multimediavraagstukken. Hij loopt naar zijn computer en toont een project wat zijn onderzoeksgroep vorig jaar opleverde: Streamwatchr.com, een analyse van alle muziektweets wereldwijd, waarbij de nummers waar het meest over getwitterd wordt op het scherm worden getoond middels de bijbehorende afbeelding (wat veertigplussers noemen: het platen- of CD-hoesje). Iedere seconde wijzigen ongeveer zes afbeeldingen op het scherm. Het ene moment is een song van John Legend nog populair, het volgende moment staat een

ons totaal onbekend Aziatisch nummer bovenaan. “Het leuke is dat we hier een functie aan hebben toegevoegd: anderen die hiernaar luisteren, luisterden ook naar dit nummer. Zo krijg je op basis van de muziek waar jij naar luistert of over Twittert een playlist waar ook nummers op staan die jij misschien helemaal niet kent, maar die wel aansluiten bij jouw muzieksmaak.”
Het is voor IT-managers en CIO’s geen verrassing dat een combinatie van deze drie categorieën nog rijkere informatie oplevert. Toch wil De Rijke wel waarschuwen. “Marketeers en IT-ers weten al lang hoeveel je over een individu kunt leren door zijn gedrag te volgen en data te combineren. Op basis van berichten op Facebook, likes en tweets kun je een heel goede voorspelling doen van iemands geslacht, seksuele voorkeur, inkomen en persoonskenmerken. We dachten misschien dat psychologen en psychiaters met privacygevoelige informatie werkten, maar ook een analyse van de berichten van iemand op social media levert een buitengewoon rijke indicatie op van iemands gemoedstoestand. De vraag is: hoever wil je als bedrijf met deze informatie over klanten gaan?”

Hergebruik data introduceert privacyrisico
Dat is ook een vraag die wetenschappers bezighoudt, bijvoorbeeld als zij onderzoek doen naar ontstaansoorzaken van ziekten. Dit gebeurt altijd op basis van geanonimiseerde bronnen, maar door bronnen te koppelen kan data soms ineens wél worden toegeschreven aan één persoon. Daarmee introduceert big data analytics een privacyrisico dat er voorheen niet was, ziet ook Frank van Harmelen, professor of Knowledge Representation and Reasoning aan de VU in Amsterdam.
Hij houdt zich bezig met de vraag hoe je databestanden beter geschikt kunt maken voor hergebruik met behulp van metadata. Dit is ook het doel van het COMMIT-project dat hij leidt en dat getiteld is: ‘From data to semantics’. “Tot nu toe was de makkelijkste manier om databronnen te combineren het opslaan van de informatie in een centraal datawarehouse. Het is ook mogelijk koppelingen te maken tussen meerdere applicaties, maar dat wordt al snel een berg spaghetti, waardoor niemand meer weet welke applicaties nu precies op welke manier met welke andere applicaties zijn verbonden en welke informatie automatisch van het ene in het andere systeem wordt overgenomen. Wat wij ontwikkelen, met vele collega’s wereldwijd, is een methode die werkt als het web: je verwijst onderling naar elkaar zonder dat je vooraf tijd hoeft te investeren in precieze afspraken over de manier van registreren (gebruik je M/V, M/F of 0/1 om geslacht aan te duiden). Je kunt daardoor veel sneller en tegen veel lagere kosten gebruikmaken van meerdere databronnen, interne en externe.”

Privacy en semantiek
Je zou zeggen dat onderzoekers niet kunnen wachten tot deze technologie breed beschikbaar komt, want dan kunnen ze veel eenvoudiger databronnen gaan combineren. Toch voelt Van Harmelen nog veel weerstand. “Sommige onderzoekers zijn bang dat anderen de data gaan gebruiken voor andere doelen dan waarvoor de onderzoeker ze heeft verzameld. Want als je je data onder open access ter beschikking stelt, dan heb je er geen controle meer op. Wij vinden: dat moet je accepteren, ‘data wants to be free’. Bovendien heb je er als onderzoeker zelf ook baat bij dat databronnen toegankelijk worden, want jij krijgt dan ook toegang tot de data van andere onderzoekers. Toch is deze zienswijze nog niet algemeen geaccepteerd.”
Dat komt mede door het privacygevaar dat eraan zit. “Het koppelen van verschillende databronnen kan gevoelige informatie opleveren. Denk bijvoorbeeld aan data die in de ene bron anomiem is, maar die opeens kan worden toegeschreven aan een

specifieke persoon als de bron wordt gekoppeld aan een andere bron. We hebben als maatschappij nog geen antwoord op deze vraagstukken, en dat komt er misschien ook wel nooit. Je zult het voorlopig per situatie moeten bekijken en ook moeten vertrouwen op de oprechte bedoelingen van onderzoekers."

GPS-tracking

Ook Maurice van Keulen is huiverig voor de gevaren. Hij leidt de onderzoeksgroep Data Management Technologie aan de Universiteit Twente en werkt onder meer aan het COMMIT-project TimeTrails, dat zich richt op het toevoegen van tijd- en locatiedata aan andere databronnen. De Universiteit Twente werkt in dit project samen met het Centrum voor Wiskunde & Informatica (CWI) en de Universiteit Utrecht. Van Keulen: "Steeds meer mensen hebben een telefoon met GPS en daarmee laten ze een spoor na. Marketeers zijn bijzonder geïnteresseerd in dat digitale spoor, want daarmee kun je het gedrag van mensen analyseren en ze nog gerichtere aanbiedingen doen. Het GPS-spoor is alleen niet altijd even nauwkeurig. Het wordt bepaald op basis van coördinaten. Eén van de onderzoeken die wij hebben gedaan heeft betrekking op de technologie waarmee je de route die mensen afleggen het meest nauwkeurig in kaart kunt brengen, zodat je bijvoorbeeld niet alleen de looproute van iemand door een winkelstraat ziet, maar ook heel precies kunt identificeren in welke winkels iemand binnen is geweest en hoe lang. Deze data koppelen we aan andere data, bijvoorbeeld Twitterfeeds. Zo zien we wie waar welke twitterberichten verstuurde."

Black box

Eén van de klanten is de Inspectie van het ministerie van Sociale Zaken. Zij leveren diverse fraudepreventiediensten, onder meer aan gemeenten. Gemeenteambtenaren kunnen bij de Inspectie het risicoprofiel opvragen van iemand die een uitkering aanvraagt. De onderzoeksgroep van Van Keulen verrijkt dat profiel op basis van internetbronnen, zoals bijvoorbeeld Twitter en Facebook.

Dat is nog een hele klus, want Twitteraccounts lijken lang niet altijd op iemands echte naam. Bovendien zijn honderden mensen die Henk Jansen of Jan de Vries heten. Hoe weet je zeker dat je de juiste hebt? Daarmee kom je meteen ook op het vlak van ethiek. "Je verzamelt ongevraagd een heleboel informatie over mensen, ook over mensen die nog nooit een bijstandsuitkering hebben aangevraagd en bij wie het niet zou opkomen om daar ooit mee te frauderen maar die toevallig dezelfde naam hebben als iemand die wel een hoog risicoprofiel heeft", zegt Van Keulen. "Daarom werken wij nauw samen met een ethisch adviseur. Zij wordt bij alle COMMIT-projecten betrokken, wijst aan waar de gevoeligheden liggen en denkt mee over oplossingen om die te omzeilen. Wij vertalen die oplossingen vervolgens weer in technologie."

Van Keulen is helder. "We werken op het randje." En dat zit hem duidelijk niet lekker. "Mijn ultieme droom is dat we een black box kunnen ontwikkelen waarin we alle persoonsgegevens verzamelen en die volautomatisch analyses doet, waarbij hij alleen resultaten terugkoppelt als een persoon verdacht is. In geval van een concrete verdenking heb je juridisch namelijk de mogelijkheid om zo'n persoon verder te onderzoeken en te volgen."

Technisch levert deze droom behoorlijk wat uitdagingen op, want hoe kun je de kwaliteit van de analyses garanderen als niemand het kan controleren? Maar ook juridisch moeten er nog de nodige hordes worden genomen, want de huidige wet zegt dat je alleen persoonsgegevens mag opslaan als die persoon daar toestemming voor heeft gegeven.

Impact op de maatschappij

Voorlopig is zo'n black box dus nog een utopie. Vandaar dat Van Keulen, wiens hart bij het onderwerp fraudepreventie ligt, het vooralsnog met iets minder gecompliceerde maar daarom niet minder leuke projecten moet doen. Zijn ogen glinsteren als hij vertelt: "Onze UT-onderzoeksgroep benaderd door de Milieudienst Rijnmond om ze te helpen illegale lozingen op te sporen met behulp van social media. Ze gebruiken nu al sensoren om bijvoorbeeld stank te identificeren. Dat gaan wij koppelen aan tweets van mensen die klagen over stank, of die iets zeggen over oppervlaktewater dat vervuild lijkt. We staan nog aan het begin van dit project, maar het lijkt veelbelovend."

Hij is ook helder waarom hij zoveel heeft met dit soort thema's. "Locatiegebonden data zijn tot nu toe vooral het speelveld geweest van marketeers die meer inzicht willen krijgen in klantgedrag. Dat is heel mooi, maar als het gaat om het opsporen van fraude of illegale lozingen dan praat je over een impact op de maatschappij."

De bevlogenheid spat ervan af als hij zegt: "Mijn drive is anderen in staat stellen coole dingen te doen met big data. Als je ziet dat AIO's nu vaak drie van de vier jaar bezig zijn met wat ik noem 'datageneuzel', van vind ik dat echt dood- en doodzonde. Als wij die tijd met een jaar kunnen reduceren door ze in staat te stellen databronnen makkelijker te combineren en beter om te gaan met onzekere data, dan verdubbelt hun onderzoektijd."

Het is ook duidelijk dat er voorlopig nog geen antwoord is op de privacyvraagstukken die met deze nieuwe technologische mogelijkheden verbonden zijn. Daarom roept Frank van Harmelen op: "Mensen die data gebruiken – of het nu onderzoekers zijn of marketeers - moeten zich van de risico's bewust zijn. Een breed maatschappelijk debat helpt dat bewustzijn aan te wakkeren." Dit nummer van BIM is een aanzet daartoe.