

Van data naar betekenis: hergebruik van data beter faciliteren?

 ictmagazine.nl/ict-breed/van-data-naar-betekenis-hergebruik-van-data-beter-faciliteren/

Big data is hot. De grote vraag waar organisaties mee worstelen is hoe ze alle databronnen op een eenvoudige en kostenefficiënte wijze aan elkaar kunnen koppelen, zonder dat er betekenissen verloren gaan of data verkeerd wordt geïnterpreteerd. Het COMMIT-project 'From data to semantics' probeert een antwoord te vinden op die vraag.

Het project kijkt vooral naar wetenschappelijke data, maar de resultaten zijn ook in andere sectoren toepasbaar. De wetenschap is gebouwd op een systeem van publicaties die verwijzen naar andere publicaties. Steeds vaker klinkt echter de roep dat publicaties ook zouden moeten verwijzen naar de onderliggende databronnen. En dat via zo'n verwijzing andere wetenschappers eenvoudig toegang zou moeten hebben tot die databronnen. Sinds de affaire Diederik Stapel is die roep ook in andere kringen groter geworden. Frank van Harmelen, principal investigator van het project 'From data to semantics', gaat het echter niet zozeer om het kunnen controleren van de onderliggende data, maar vooral om de rijkdom die het biedt als onderzoekers hun data heel makkelijk ter beschikking zouden kunnen stellen aan anderen.



Van Harmelen is professor of Knowledge Representation and Reasoning aan de VU in Amsterdam en houdt zich dagelijks bezig met de vraag hoe je databestanden beter geschikt kunt maken voor hergebruik met behulp van metadata. "Er wordt wereldwijd heel veel data verzameld. Dat doen wij als wetenschappelijke onderzoekers, maar ook bedrijven en instellingen verzamelen data. Denk eens aan alle gegevens die een ziekenhuis vastlegt, of een gemeente. Als je die data op een slimme manier met anderen kunt delen en verschillende databronnen met elkaar in verband kunt brengen, dan leidt dat tot nieuwe inzichten. Daarvan is iedereen wel overtuigd. Nu gaat het er alleen nog om zo'n slimme manier te ontwikkelen."

Automatisch metadata toekennen

En dat is waar postdoctoraal onderzoeker aan de VU Rinke Hoekstra zich mee bezighoudt. Hij is projectleider van het project 'From data to semantics'. Hoekstra: "Uit een onderzoek naar wetenschappelijke workflows blijkt dat 60 procent van de tijd van wetenschappers gaat zitten in datapreparatie en het opnieuw creëren van data die door een ander ook al eens is verzameld. Ook kenniswerkers in het bedrijfsleven verspillen veel tijd met het zoeken naar informatie, tot wel 40 procent van hun tijd. Ons onderzoek richt zich op de vraag hoe je die tijd drastisch kunt verkorten door databronnen makkelijker extern te publiceren, ze beter vindbaar te maken en ze makkelijker bruikbaar te maken."

Dat laatste is het meest ingewikkeld, stelt Hoekstra. "Nu werkt iedere wetenschapper nog op zijn eigen vierkante millimeter en gebruikt zijn eigen terminologie, die voor andere wetenschappers niet altijd duidelijk is. Als je een spreadsheet ziet met een kolom G en in de vakjes enen en nullen, bedenk dan maar eens dat G staat voor gender. En koppel dat dan maar eens aan een andere bron waarbij de tabelnaam 'geslacht' is en in de vakjes M/V staat, of M/F. Wij werken aan een systeem dat data automatisch zo beschrijft met metadata dat ook computers het kunnen lezen. Want het is hopeloos om wereldwijd in de academische wereld afspraken te maken hoe we iets vastleggen, zo'n top-down benadering gaat niet werken. Daarom kiezen wij voor een bottom-up aanpak: iedere wetenschapper mag met zijn eigen terminologie werken en het systeem zorgt ervoor dat deze worden geïntegreerd."

In een van de tools waar Hoekstra aan werkt geven gebruikers door simpelweg slepen aan wat de

namen zijn van de datavelden. De machine vindt daar automatisch de juiste metadata-omschrijving bij, die de onderzoeker alleen nog maar hoeft te controleren. De 'human in the loop'-aanpak is daarbij belangrijk, benadrukt Hoekstra. "Het woord 'geslacht' in een dataset over varkensgriep staat waarschijnlijk voor een datum. Onze tool herkent dit soort dingen wel, maar het is wel goed dat nog door mensen te laten controleren."

Sociale bottlenecks

Eén van de partners in het COMMIT-project is Elsevier. De vakgroep Knowledge Representation and Reasoning werkt al jarenlang samen met de onderzoeksafdeling van deze uitgever, maar in het kader van het 'Data to Semantics'-project zitten de onderzoekers sinds kort ook aan tafel met de product managers. Van Harmelen: "Nu verkoopt Elsevier PDF's van publicaties. Die publicaties hebben veel meer waarde als je daar de bijbehorende data bij kunt geven. Wij gaan in een periode van slechts anderhalve maand een demo voor ze bouwen die werkt op hun platform. En dat kan, want de benodigde technologie is inmiddels beschikbaar. De bottlenecks zijn veel meer van sociale aard. Er zijn onderzoekers die zeggen: ik heb jarenlang aan mijn onderzoek gewerkt, waarom zou ik die data nu ineens met Jan en alleman gaan delen? Ook zijn ze 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, heb je er geen controle meer op. Wij vinden: dat moet je accepteren, 'data wants to be free'. En steeds meer onderzoekersponsors, zoals NWO en de EU, vinden dat ook: ze eisen tegenwoordig open access voor onderzoeksdata. 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 doordat er een privacygevaar aan zit, zegt Hoekstra. "Het koppelen van verschillende databronnen kan gevoelige informatie opleveren."

Het koppelen van verschillende databronnen kan daarnaast ook licentietechnisch problemen geven. "Onze ervaring is dat het technisch steeds minder een probleem is om databronnen te delen en aan elkaar te koppelen, maar dat de struikelblokken van organisatorische en sociale aard zijn," concludeert Van Harmelen. "We hebben als maatschappij op deze vraagstukken nog geen goed antwoord 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."

Toepassingen in andere sectoren

De meerwaarde van het openbaar publiceren van data en het automatisch toekennen van gestandaardiseerde metadata is helder voor de wetenschappelijke wereld, maar is ook in andere sectoren zijn er toepassingen. Zo werken leden van Van Harmelens onderzoeksgroep samen met het AMC. Elk academisch ziekenhuis moet jaarlijks zo'n 2000 getallen aanleveren die de kwaliteit van de zorg meten (zogenoemde kwaliteitsindicatoren). Die indicatoren worden nu met de hand berekend, elk jaar weer. De uitdaging hier is het koppelen van interne databronnen in het ziekenhuis op zo'n manier dat de kwaliteitsindicatoren automatisch kunnen worden berekend.

Een andere partner is de gemeente Amsterdam in samenwerking met de Waag Society. De VU helpt data van de gemeente toegankelijk te maken voor de brandweer. Het gaat bijvoorbeeld om informatie over wegoptbrekingen, zodat de brandweer bij een uitruk real-time de juiste routeinformatie krijgt. Ook gaat het om informatie over verleende bouwvergunningen, zodat de brandweer weet dat er een dakkapel op zolder is gemaakt ten behoeve van een slaapkamer. Of dat er een invalidetoilet in het pand aanwezig is, zodat ze bedacht zijn op rolstoelgebruikers. En het gaat om informatie over de wijze van isolatie, want bepaalde isolatie is brandgevaarlijk.

Semantische interoperabiliteit

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. Van Harmelen: “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.”

En dat is goed nieuws voor iedereen die aan de slag wil met big data, want het probleem daarbij zit meestal niet in de V van Volume, maar in de V van Variety. “De semantische methode zorgt ervoor dat je een grote diversiteit aan bronnen aan elkaar kunt koppelen teneinde op zoek te gaan naar patronen en afwijkingen daarin, zonder dat je data eerst moet prepareren en in een standaardformat moet gieten,” aldus Van Harmelen. Hiermee winnen organisaties tijd en sparen ze investeringen uit. Deze semantische technologie is inmiddels in gebruik bij alle grote search-engines (Google, Yahoo, Bing), bij nieuws- en mediabedrijven (New York Times, BBC), bij overheden (bijvoorbeeld de UK en de US), bij internetwinkels (Sears, KMart) en in de industrie (Renault, Volkswagen).

Wat is COMMIT?

COMMIT is een publiek-private onderzoekscommunity die oplossingen ontwikkelt op basis van de nieuwste informatietechnologie om beter tegemoet te komen aan de uitdagingen binnen onze maatschappij. Eén van die uitdagingen is om data die overal wordt verzameld toegankelijk te maken voor hergebruik, zodat ook anderen die datasets kunnen gebruiken voor onderzoek en beslissingsondersteuning in allerlei organisaties. Andere onderzoeksgebieden zijn onder meer sensornetwerken ter verhoging van de publieke veiligheid en virtuele werelden om eenzaamheid bij onder meer ouderen tegen te gaan. Kijk op www.commit-nl.nl voor meer informatie.

Lees meer

artikelen over het topic:

[BI / BPM](#)

artikelen over de sector:

[Business](#)

artikelen via de tag: