

Semantic web-pioniers in Nederland

AI, OO, RDF, integratie en meer

achtergrond

Het informatieaanbod is de laatste jaren enorm toegenomen. Om daarvan gebruik te kunnen maken zijn intelligente zoekmethoden nodig. De ontwikkeling daarvan geschiedt op basis van kennis over kunstmatige intelligentie (AI). Er bestaat al een W3C-taal (RDF) waarmee kennis opgeslagen en gemodelleerd kan worden en waarmee over die kennis geredeneerd kan worden. Aduna - een Nederlands bedrijf - heeft een query-taal voor RDF-databases ontwikkeld: SeRQL. Er is een redelijk grote kans dat deze taal tot W3C-standaard uit zal groeien.

Aduna is een voorbeeld van een bedrijf dat voortgekomen is uit wetenschappelijk onderzoek en dat nog steeds nauwe banden onderhoudt met de universiteit. De oprichter, Jos van der Meer, is na zijn studie gaan werken bij Sun, maar heeft later een bedrijf opgericht onder de naam VMX, en daarna Aduna, dat aanvankelijk Administrator heette.

Van der Meer: 'De naam Administrator illustreert mooi hoe wij op dat moment dachten. Wanneer je AI loslaat op beheer en krijg je Administrator: een samen-trekking van AI en administration. Later is ervoor gekozen om AI te gaan toepassen op informatie en software te gaan ontwikkelen die de semantische consistentie van informatie kan controleren.'

REDENEREN *Kunstmatige intelligentie wordt om ver-warring te voorkomen vaak aangeduid met de Engelse afkorting AI. Wat betekent het precies?*

Van Harmelen: 'Het is een vakgebied dat veel te lijden heeft onder zijn eigen naam. We zijn helemaal niet bezig met het bouwen van de ultieme intelligente computers, maar met het ontwikkelen van allerlei geavanceerde representatie-technieken die gebruikt worden binnen de informatica. Als je kijkt naar de geschiedenis van de AI, dan zie je dat AI allerlei technieken heeft opgeleverd die nu mainstream informatica zijn. De wortels van OO-programmeren liggen in kennisrepresentatie-technieken uit de jaren zestig, de wortels van alle functionele programmeertalen liggen in LISP, een programmeertaal die uit de AI komt. Ik denk dat het semantic web daar een goed voorbeeld van is. AI is goed in informatie modelleren en daarover redeneren, conclusies daaruit trekken. Het tegengestelde daarvan is een database, daarin sla je dingen op en het enige wat je eruit krijgt, is precies datgene wat je erin opgeslagen

hebt. Als je de stap maakt van databank naar kennisbank, dan is het idee dat je daar niet alleen informatie en kennis in opslaat, maar ook regels om over die kennis te redeneren. Vervolgens levert die kennisbank je allerlei conclusies op. Die kennisbank redeneert over die gegevens en komt met conclusies terug over die gegevens. Het woord "redeneren" gebruiken we, omdat dat losjes gebaseerd is op de menselijke metafoor. We beweren niet dat we met die machines menselijke intelligentie nabootsen. Het is wel een belangrijke stap dat die machine zelf, op basis van de regels en de feiten die je erin gestopt hebt, conclusies trekt en bijvoorbeeld inconsistenties kan achterhalen.'

Maar in veel applicaties doe je dat ook in zekere mate.

Van Harmelen: 'Zeker, het is een glijdende schaal. Er is geen harde grens waar je kunt zeggen: nu is het ineens AI geworden. AI zit wel heel erg aan het ene eind van die glijdende schaal en databases zitten aan het andere eind. Daar tussenin zijn allerlei mogelijkheden over hoe geavanceerd en "intelligent" je machine met die data omgaat.'

Van der Meer: 'Een van de kenmerken van een AI-toepassing is naar mijn mening dat je expliciet bent over welke regels je toepast. En veel van die programma's die jij noemt, zoals een Rich UI heeft zijn kennis impliciet in procedures ingebakken. Je kunt die kennis niet raadplegen en ook niet hergebruiken.'

Van Harmelen: 'Wanneer spreek je dus van een AI-toepassing: wanneer je je domein modelleert en ook de constraints op die regels modelleert en dat allemaal expliciet maakt en je gaat er dan met een normale programmeertaal over redeneren. Dan heb je een AI-toepassing. Je hebt geen AI-toepassing wanneer je al je kennis in C-methodes hebt uitgedrukt, zelfs al zou die toepassing zich hetzelfde gedragen. C-code is niet herbruikbaar, die heb je een keer gemaakt en als je die voor een andere toepassing wilt gebruiken dan is het altijd weer net verkeerdt. Het idee van tools als Sesame is dat je die kennis kunt representeren, opslaan in Sesame en daar vervolgens conclusies aan kunt verbinden: Sesame redeneert en geeft je dan nieuwe conclusies terug.'

SLIMMER DAN EEN DATABASE *Sesame blijkt oorspronkelijk door Aduna ontwikkeld te zijn binnen een Europees onderzoeksproject met een aantal verschillende partners, waaronder de VU, BT en Swiss Life. Het wordt verder ontwikkeld. Aduna heeft het intellectuele eigendomsrecht, maar het is inmiddels wel beschikbaar onder een open source licentie.*

Broekstra: 'Sesame is vergelijkbaar met een database, maar dan slimmer. Waar een database gaat over puur gegevens opslaan, gaat Sesame over kennis opslaan in zo'n taal als RDF. Je hebt een beschrijving in RDF, een Thesaurus bijvoorbeeld, die kun je in Sesame stoppen net zoals je klantgegevens in een database stopt. Sesame kan er dan over redeneren en zelf uitvinden dat een klant met een voorkeur voor Frankrijk interesse heeft in de Loire, omdat de Loire in Frankrijk ligt.'

Van der Meer: 'Een van de belangrijkste features van Sesame is dat daar dan ook weer een query-taal op gedefinieerd is. Net zoals je SQL hebt voor database heb je SeRQL, dat staat voor Sesame RDF query language. Via een api of soap - Sesame is beschikbaar als webservice - kun je communiceren met die informatie, de kennis die in zo'n Sesame-database zit eruit trekken en gebruiken in je eindgebruikerapplicatie.'

En SeRQL maakt daar deel van uit?

Van der Meer: 'SeRQL is onderdeel van Sesame, maar trekt op zichzelf aandacht omdat men probeert uit te

vinden of de taal geschikt is als basis voor standaardisatie van query-talen in het algemeen.'

Van Harmelen: 'Het W3C is nu zo ver dat het RDF gestandaardiseerd heeft, maar er is nog geen gestandaardiseerde querytaal voor. Het W3C kijkt naar SeRQL als een van de kandidaten voor het standaardisatieproces. Belangrijk is ook dat de ontwikkeling van Sesame als open source project ertoe geleid heeft dat het door allerlei partijen wereldwijd gebruikt wordt. Het is een van de twee meest gebruikte RDF-databases in de wereld.'

GOOGLE *Van der Meer vond het hergebruik van informatie het belangrijkste aspect van het semantic web.*

Van der Meer: 'Er zijn twee belangrijke kanten aan het semantic web: de ene kant is het redeneren over de data, de tweede kant die data op zo'n manier opslaan dat je ze kunt hergebruiken en dat onafhankelijke partijen toch elkaars data kunnen gebruiken. Ik denk dat het een vergissing is om net als in de AI-tijd te veel nadruk te leggen op het redeneren, het hergebruik van informatie is waarschijnlijk een miljoen keer belangrijker dan de redeneercapaciteiten.'

Van Harmelen: 'Die informatie is herbruikbaar geworden, doordat het in dat soort standaardrepresentaties is opgeslagen. We hebben al een heel goed voorbeeld van herbruikbare informatie op het web en dat is HTML. Dat is nu echter alleen maar handwerk, mijn pc en zelfs Google heeft geen idee waar die pagina's over gaan. Google doet gewoon letterherkenning, of dat nu een pagina van een reisbureau is of van een districtsbestuur, geen idee, terwijl dat soort representatietalen als RDF die informatie in een herbruikbare manier opslaan.'

Waar staat RDF voor?

Van Harmelen: 'RDF betekent Resource Description Framework: geen sexy acroniem, een taal om informatie in op te slaan. Het is een heel eenvoudige model-

leertaal. Je kunt daarin praten over dingen en over relaties tussen en eigenschappen van dingen, zodat er uiteindelijk een wereldwijd netwerk van informatiepagina's ontstaat, op dezelfde manier waarop er nu een netwerk is van HTML-pagina's die aan elkaar gelinkt zijn. Het enige maar wel heel belangrijke verschil is, dat het wereldwijde netwerk van HTML-pagina's leuk is, zolang je Engels kent en plaatjes begrijpt. Onze computer kan er niets mee beginnen. Het idee van het wereldwijde web van RDF-databronnen die naar elkaar verwijzen, is dat computers die gegevens kunnen interpreteren en daar op een nuttige, slimme en intelligente manier combinaties van kunnen maken.'

ALGORITMEN *De Chomskyaanse transformationele generatieve grammatica kent ook semantiek (betekenisleer). Ik ben niet meer op de hoogte van de laatste ontwikkelingen op dat gebied, maar ik zou me kunnen voorstellen dat je daarmee ook zou kunnen redeneren.*

Van Harmelen: 'Je wilt juist dat je over die kennis kunt redeneren. Barcelona ligt in Catalonië, Catalonië ligt in Spanje, dus ligt Barcelona in Spanje. Dat is een conclusie waar wij niet echt van omvallen, maar die mijn computer niet kan trekken. Sesame kan dat wel, omdat dat een regel is die je kunt vaststellen, exclusief representatie. Dat is een stukje semantiek. Je maakt dus inderdaad precies zoals in die oude droom van de TGG niet alleen de syntax machinaal toegankelijk, maar ook de semantiek. Tot nu toe heb ik echter alleen voorbeelden gezien van conclusies die worden toegekend aan informatie op basis van regels, niet op basis van het begrijpen van taal.'

Van der Meer: 'Je hebt dus behoefte aan modellen van een domein. Je hebt heel veel van dat soort modellen nodig en je kunt ze allemaal met de hand gaan maken, maar het is natuurlijk wel heel erg aantrekkelijk om de algoritmen die in de AI ontwikkeld zijn te gebruiken om op zijn minst een voorzet kunnen doen van een modellering. Die twee werelden raken daar wel weer aan elkaar, waarbij die natuurlijke taalherkenning helpt om het model te formuleren.'

INNOVATIE *Voor Frank van Harmelen is de samenwerking met Aduna van groot belang:*

Van Harmelen: 'Het aardige is dat dit soort technieken en toepassingen - die binnen Aduna centraal staan en die hun toepassingen vinden voor de klanten van Aduna - heel direct hun wortels hebben in dat lange termijn onderzoek dat AI heet. De grote IT-partijen zijn vaak het meest conservatief: hoe groter een bedrijf, hoe technologisch conservatiever, dat is de regel. De innovatie zit bij de kleine bedrijven.'

Voor ons bij de universiteit zijn het soort banden zoals die met Aduna dan ook van levensbelang, want wij moeten aantonen wat ons onderzoek oplevert voor de maatschappij. Net zo goed als het omgekeerde het geval is.'

Van der Meer: 'Wat we als eerste commercialiseren, zijn de ideeën en de algoritmen die je nu kunt realiseren. Wij werden in het semantic web geconfronteerd met informatie die veel minder eenduidig te classificeren is dan in de Google-ranking. Wanneer je mensen dus wilt laten interacteren met die informatie moet je die op een betekenisvolle manier ordenen. Dus moet je werken aan representatievormen, interactievormen, waarin je veel complexere boodschappen kunt overdragen zonder dat dat voor de gebruiker complex wordt. Daar komt die visualisatie-techniek uit voort en ook het navigeren op heel veel dimensies. We hebben dus belang-

rijke ideeën over het navigeren over complexe informatie toegepast in problemen van nu.'

Van Harmelen: 'De Guided Exploration van Aduna levert niet alleen een kenniswolkje van alles waar dat woord in voorkomt, maar je krijgt ook sugges-

ties hoe je dat nog verder zou kunnen verfijnen. Dat zijn allemaal pogingen om nu al te scoren met dat gestructureerd zoeken en navigeren en expliciet representeren van structuren, ook zonder expliciet gerepresenteerde kennis. Daar waar die kennis al wel zo beschikbaar is, maken we er natuurlijk ook ogenblikkelijk gebruik van. Een voorbeeld daarvan is een thesaurus voor geneesmiddelen en ziekten voor Elsevier, met 50.000 termen en 190.000 synoniemen. Die hebben we in RDF gebracht en gekoppeld aan andere software, met visualisatietechnieken vastgeknoopt aan vijftien miljoen abstracts uit de medische literatuur database: een prachtige literatuurnavigatie-engine. Eigenlijk is dat de wereld waarin Aduna zou willen leven.'

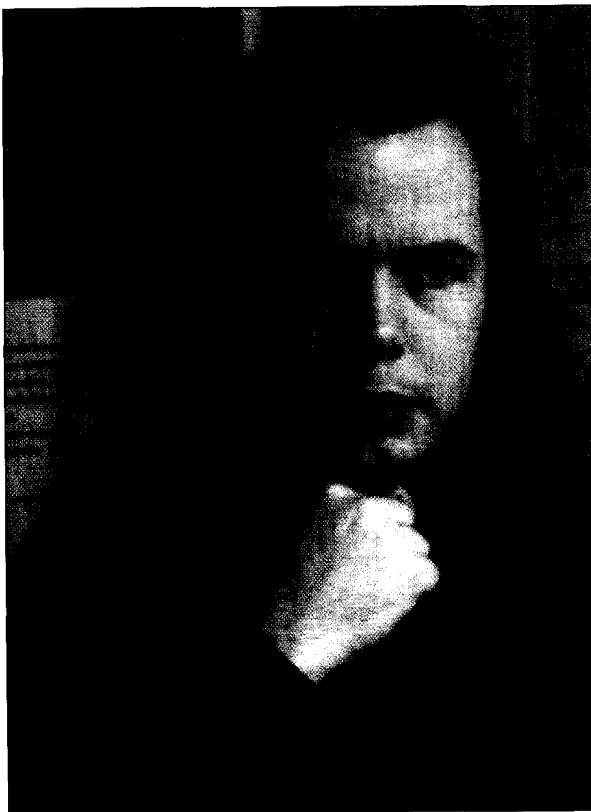
REPRESENTATIE *Wanneer we verder naar de toekomst kijken, zou dan SeRQL na verloop van tijd niet geschikt worden om webservices te bevragen naar hun eigenschappen, om de juiste webservice voor je applicatie te vinden?*

Van der Meer: 'Als je veel webservices hebt, krijg je inderdaad behoefte aan een formele representatie van de eigenschappen van die webservice, zodat je al redenerende de juiste webservice kunt vinden.'

Van Harmelen: 'Die semantische representatie sla je op in Sesame en vervolgens ga je redeneren. Welke van die duizend webservices heb ik nodig voor dit doel? RDF en webservices, dat zit natuurlijk heel dicht tegen elkaar aan. Een heel belangrijke rol voor RDF is ook data- en applicatie-integratie. Een goed voorbeeld is Boeing. Dat is een enorm bedrijf en die hebben ook geen idee hoeveel applicaties ze in huis hebben. Boeing heeft heel erg ingezet op semantic webtechnologie, op RDF als representatietaal om integratie van hun databronnen en applicaties voor elkaar te krijgen. Ze maken dan bijvoorbeeld RDF-schillen om databases heen en op het RDF-niveau heb je expliciete beschrijvingen van die databases gekregen. Die knoop je dan vervolgens aan elkaar. Dan krijg je integratiemogelijkheden die je helemaal niet voorzien had.'

Tekst en fotografie: Dré de Man

“We hebben belangrijke ideeën over het navigeren over complexe informatie toegepast in problemen van nu”



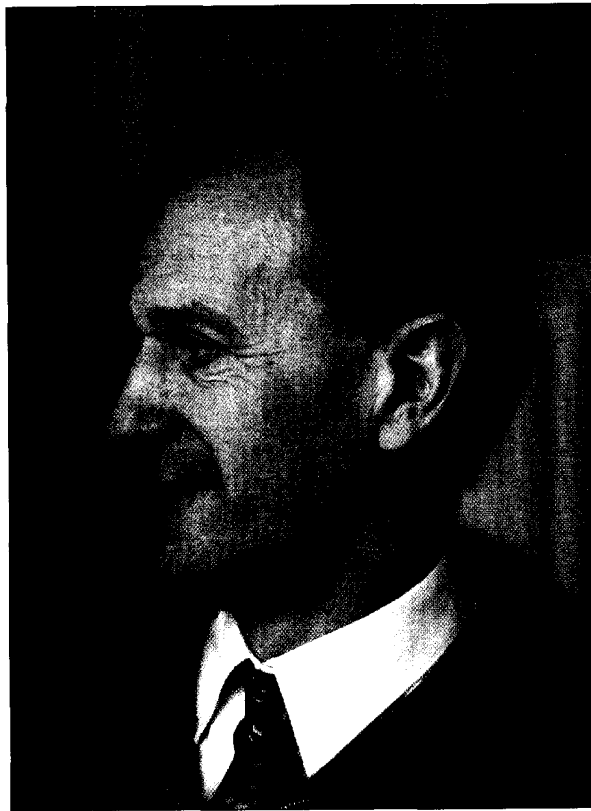
Jeen Broekstra, promovendus aan de vakgroep kunstmatige intelligentie van de VU en ontwikkelaar bij Aduna Software

Het navolgende gesprek vond enkele maanden geleden plaats. Een van de drie geïnterviewden - *Jos van der Meer* - overleed in de periode tussen het interview en de publicatie ervan. In overleg is besloten publicatie toch door te laten gaan. Jos van der Meer was oprichter van Aduna. Hij studeerde wiskunde en na afronding daarvan informatica aan de TU Twente. Zijn levenswerk wordt binnen Aduna voortgezet.

Frank van Harmelen is hoogleraar bij de vakgroep kunstmatige intelligentie aan de faculteit exacte wetenschappen van de Vrije universiteit. Een groot deel van zijn onderzoek gaat over allerlei nieuwe ontwikkelingen rond het semantic web. Hij werkt samen met Aduna.

Jeen Broekstra is promovendus aan de vakgroep kunstmatige intelligentie aan de faculteit exacte wetenschappen van de Vrije Universiteit, en tevens werknemer van Aduna Software, als ontwikkelaar. Zijn promotieonderzoek gaat over het RDF-specifiek opslaan, query'en en definiëren met semantic webtalen. Binnen Aduna is hij ontwikkelaar van een architectuur in Java, Sesame geheten die zich bezig houdt met intelligente storage en query'en van informatie.

“Het hergebruik van informatie is waarschijnlijk een miljoen keer belangrijker dan de redeneercapaciteiten”



Jos van der Meer was oprichter van Aduna. Hij overleed enige tijd nadat dit interview had plaatsgevonden



Frank van Harmelen, hoogleraar bij de vakgroep kunstmatige intelligentie aan de VU