

Reformatorisch Dagblad

Structuur in bits en bytes

22-01-2008 09:55 | Jacob Siebelink



Oma van 84 krijgt een hartaanval. Via mobiel en e-mail worden niet alleen haar kinderen automatisch op de hoogte gesteld van het incident, maar krijgen ook ziekenhuis en verzekeringsmaatschappij een seintje. Een ambulance rijdt uit; agenda's van de kinderen worden automatisch afgestemd op de onverwachte situatie. Zo'n scenario zou in de nabije toekomst mogelijk moeten zijn, zo hypen veel media de nieuwe internettechnologie, web 3.0 of het semantische web. De werkelijkheid blijkt een stuk weerbarstiger.

Google is dom. Dat de vier letters b-a-n-k evengoed een zitmeubel als een financiële dienstverlener kan zijn, dat is voor de zoekmachine te hoog gegrepen. Niet handig, vindt Frank van Harmelen, hoogleraar kunstmatige intelligentie aan de Vrije Universiteit in

Amsterdam. „Zeker als je bedenkt dat de informatiebrij op het wereldwijde web alleen maar groter en ondoorzichtiger wordt.”

Dat het ook anders kan, bewees Van Harmelens collega hoogleraar Guus Schreiber eerder door bestaande hiërarchisch gestructureerde woordenboeken van musea en andere cultuurinstellingen aan elkaar te knopen en verbanden aan te brengen. „Als je in deze zoekmachine ”Van Gogh” invoert, zal hij een onderscheid laten zien tussen schilderijen die door Van Gogh zijn gemaakt, schilderijen waar Van Gogh op afgebeeld is, brieven van Van Gogh en boeken over de schilder”, zegt Van Harmelen. „Maar hij zal ook informatie geven over andere impressionistische schilders en tijdgenoten.”

Voordat deze zogenaamde semantische technologie klaargestoomd is voor het wereldwijde web en er sprake kan zijn van web 3.0, moeten onderzoekers nog veel werk verzetten, benadrukt Van Harmelen. „Er is in de wereldgeschiedenis nog nooit zo veel informatie bij elkaar gebracht. En de informatie is ook nog eens uiterst divers. De vraag is of er ooit een compleet hiërarchisch woordenboek komt voor het wereldwijde wilde web.”

Internettaal

Het semantische web drijft volgens Van Harmelen op twee technieken. „In de eerste plaats moet je labels aan documenten hangen, kernwoorden die iets zeggen over het document. Aan ”bank” kunnen bijvoorbeeld de woorden ”zitten” en ”meubel” toegekend worden. De computer zal deze meestal automatisch genereren.”

Maar gebruikers kunnen volgens Marianne Herbert van LogicaCMG ook zelf labels aan hun documenten toekennen, zoals bij web 2.0 al gebeurt. „Denk aan productwaarderingen. Uit onderzoek blijkt dat gebruikers graag bereid zijn hun mening ergens over te geven, als hen daar expliciet om gevraagd wordt.” Wel is het

volgens haar van belang om onderscheid te maken tussen waarderingen. „Een organisatie als de Consumentenbond spreekt met meer gezag dan een individuele gebruiker. Dat moet in een taxonomie wel duidelijk worden.”

Alle labels worden vervolgens ondergebracht in een hiërarchisch woordenboek. Van Harmelen en Schreiber ontwikkelden samen met collega's over de hele wereld een internettaal die de computer in staat stelt zo'n woordenboek te lezen. „De simpele variant van deze computertaal heet Resource Description Framework, kortweg RDF. Hierin kun je bijvoorbeeld aangeven dat een "bank" een "meubel" is of een "financiële instelling”.”

De farmaceutische industrie, die volgens Van Harmelen zeer geïnteresseerd is in semantische technieken, kan hiermee echter niet uit de voeten. „In RDF kun je de computer uitleggen dat de groep "patiënten" bestaat uit "mannen" en "vrouwen", maar niet dat een patiënt niet tegelijk man en vrouw kan zijn. Daarom ontwikkelden we Ontology Web Language, OWL. Als je in deze taal uitlegt dat een medicijn niet geschikt is voor zwangere patiënten, weet de computer direct dat mannen er geen problemen mee kunnen hebben.” Dat RDF naast OWL blijft bestaan, heeft volgens hem met de rekenkracht te maken die nodig is om met OWL te kunnen werken.

Privacy

Het toepassingsgebied van semantische technieken beperkt zich niet tot het verbeteren van de zoekmachines. „Nu is het zo dat je voor het bekijken van je bankrekening en je telefoonkosten naar twee verschillende websites moet. Erg onhandig”, vindt Marianne Herbert. „In de toekomst heb je een soort persoonlijke softwareagent die deze applicaties aan elkaar koppelt, samen met je elektronische agenda op de computer en andere data.” Volgens Van Harmelen wordt naar zo'n zogenaamde semantische desktop actief onderzoek gedaan. „Ik verwacht dat die met een paar jaar voor het grote publiek beschikbaar is.”

Semantische technieken maken het ook mogelijk om websites automatisch toe te snijden op het profiel van de gebruiker. Personalisatie, noemt Van Harmelen dat. „Als mijn zoontje de website van KPN bezoekt, krijgt hij hetzelfde op zijn scherm als wanneer ik naar deze website ga. Niet zo slim, want mijn zoontje heeft heel andere voorkeuren en een ander budget dan ik.” Herbert: „Als ik op vakantie wil, ontstaan er allemaal deelvragen: Is het hotel nog vrij, is het goed bereikbaar, is er wat te beleven voor de kinderen. Op basis van mijn voorkeuren zal een speciale website in de toekomst in één keer een concreet aanbod presenteren.”

Om dit voor elkaar te krijgen, moet de computer iets weten over de website én over de gebruiker. Het beschermen van privacy is dan een terechte zorg, erkent Van Harmelen. „Eerlijk gezegd weten we niet hoe we daarmee om moeten gaan. Enerzijds willen we privacy beschermen, anderzijds willen we wel een betere dienstverlening. Dat botst met elkaar. Als de tomtom je op de hoogte wil houden van files, moet het systeem wel weten waar je bent.” Overigens zijn gedachten over privacy sterk aan het verschuiven, constateert hij. „Jongeren zetten vrijwillig veel privacygevoelige info op internet. Denk aan Hyves. Dat was vroeger niet denkbaar.”

Hype

Het scenario van de oma van 84 die een hartaanval krijgt, doet Van Harmelen af als een hype. „Leuk om over te dromen, maar je moet niet net doen alsof dat om de hoek is. Mijn vakgebied, dat van de kunstmatige intelligentie, zit sinds jaar en dag vol met mensen die een veel te grote mond opzetten en van alles beloven, wat vervolgens niet uitkomt.”

De VU-onderzoeker formuleert daarom voorzichtig als hij een tipje van web 4.0 oplicht. „De volgende stap is dat digitale informatie de computer uitgaat en de fysieke omgeving koloniseert. Denk aan de koelkast die aan het internet hangt en automatisch boodschappen doet bij de buurtsuper.” Van web 4.0 zal dan overigens geen sprake

zijn, zegt hij. „Dat zou weer het web zijn, gescheiden van de fysieke wereld. Juist dat onderscheid tussen virtueel en fysiek zal in de toekomst steeds meer vervagen.”

„Computer kan zelf verbanden ontdekken”

Wetenschappers aan de Universiteit van Amsterdam zien niet zo veel heil in het van bovenaf opleggen van formele verbanden, zoals VU-onderzoeker Frank van Harmelen doet. „Mensen hebben een uitstekend middel om hun ideeën, ervaringen en emoties met anderen te delen, namelijk taal”, zegt hoogleraar informatica Maarten de Rijke van de UvA. Hij ziet meer in statistische methoden waarbij de computer dit taalgebruik op het web analyseert en zelf verbanden aanbrengt.

Beide onderzoekers hebben een eigen benadering om het probleem van het onoverzichtelijke web op te lossen. De wegen liggen echter minder ver uit elkaar dan op het eerste gezicht lijkt. „We zetten de tegenstellingen soms extra scherp aan”, zegt Van Harmelen. „Maar we weten dat de waarheid ergens in het midden zal liggen.”

Huidige zoekmachines zijn volgens De Rijke veel intelligenter dan Van Harmelen beweert. „Je zou het niet zeggen, maar Google kan wel degelijk een onderscheid zien tussen ”bank” en ”bank”. Dat is met statistische methoden gemakkelijk te achterhalen. Zoekmachines kijken niet alleen naar die vier letters, maar ook naar de context waarin ze voorkomen in documenten op internet. Je hebt helemaal geen zelfgebouwd hiërarchisch woordenboek nodig om de verschillende soorten banken vervolgens te clusteren. Een zoekmachine als www.vivisimo.nl biedt de resultaten al geclusterd aan en ook Google experimenteert hiermee.”

De Rijke pleit ervoor om de computer zelf hiërarchische verbanden te laten ontdekken. „De computer kan statistisch relaties typeren. Het boeiende is om algoritmen zo te ontwikkelen dat de computer zo veel mogelijk zelf doet en als hij onzeker is om menselijke hulp vraagt. Aan ons de uitdaging om het algoritme met de juiste voorbeelden te voeden, waarmee hij aan de slag kan.”

Wat semantische technieken van doen hebben met personalisatie van websites, zoals Van Harmelen noemt, ziet De Rijke niet. „Dat is een ontwerpersbeslissing van de makers van de website. Daar heb je niet zo veel kunstmatige intelligentie voor nodig. Je maakt afspraken over het uitwisselen van gegevens. Meer niet.” Interessanter wordt het volgens hem als de computer automatisch de leeftijd of de vakantievoorkeuren van nieuwe bezoekers kan vaststellen, zonder dat de bezoeker zelf uitgebreide formulieren hoeft in te vullen.

De toegevoegde waarde van semantische technieken ziet De Rijke met name in het ontwikkelen van een gestandaardiseerde computertaal die hiërarchische verbanden expliciet benoemt. „Het bouwen van zo’n woordenboek moet je de computer vervolgens zelf laten doen.”

Vijfentwintig jaar internet

Internet is deze maand jarig. Het Arpanet van het Amerikaanse ministerie van Defensie ging 25 jaar geleden over op het zogenaamde TCP/IP protocol, waarin beschreven staat hoe computers met elkaar via het wereldwijde web kunnen communiceren.

- Web 1.0: De eerste generatie van het www heeft een hoog brochuregehalte. Internetverkeer gaat in één richting, van websiteaanbieder naar gebruiker.

- Web 2.0: Geleidelijk aan kan de gewone gebruiker informatie toevoegen aan het web. Er ontstaan weblogs en mensen gaan onderling foto's delen via websites als www.flickr.com. Met web 2.0 wordt internet socialer. De techniek erachter is echter nog gelijk aan die van web 1.0.

- Web 3.0: Bij web 2.0 doet de gebruiker nog alle intelligente handelingen. Bij web 3.0 verandert dat. De computer interpreteert data en legt verbanden. Van web 2.0 naar web 3.0 is wél een technische stap.