Autonomy definieert meaning based computing

IDOL wint zoekwedstrijd

Teus Molenaar

Na ruim een uur praten over Autonomy zegt Victor Cohen, general manager Noord-Europa, dat hij slechts het topje van de ijsberg heeft kunnen tonen. De software is zo veelomvattend dat alleen al de kernboodschap niet snel in één volzin is te vatten. Het bedrijf laat zien wat de volgende stap is in het omgaan met digitale informatie: meaning based computing. De mogelijkheid om betekenis toe te kennen aan digitale informatie in welke vorm en op welke plek dan ook.

Cohen stelt dat de mens zich de afgelopen veertig jaar aan de computer heeft moeten aanpassen, en dat Autonomy het mogelijk maakt dat de computer zich aan de mens aanpast. Dat is een uitkomst in een wereld die lijkt te bezwijken onder het gewicht van de enorme hoeveelheden digitale informatie. Niet alleen is de hoeveelheid magistraal toegenomen; ook de soortenrijkdom is weelderig. Begonnen we met gestructureerde data, netjes opgeborgen in kolommen en rijen in databases en makkelijk te raadplegen, tegenwoordig zijn er digitale tekstdocumenten, e-mails, telefoongesprekken, foto's, bewegend beeld, spreadsheets, nieuwsuitzendingen en dergelijke. Daarbij is de locatie van al die informatie de afgelopen jaren erg divers geworden: keurig op een server in het rekencentrum (al dan niet gevirtualiseerd), iets minder keurig op de lokale harde schijf, nog minder keurig op laptops, en in toenemende mate op slimme telefoons.
Om wijs te worden in die wirwar aan informatie heeft Autonomy het platform Intelligent Data Operating Layer (IDOL) gecreëerd. Dit vormt het kloppend hart van alle oplossingen die de onderneming naar de markt brengt. Met IDOL is het mogelijk die speld in de hooiberg te vinden. En dat volledig geautomatiseerd en razendsnel. Daarbij is Mike Lynch, mede-oprichter en CEO van Autonomy, schatplichtig aan de achttiende eeuwse presbyteriaanse dominee Thomas Bayes. Het theorema van Bayes is een belangrijke regel in de kansrekening. Zij wordt veel toegepast bij medische diagnostiek, in de (astro)fysica en civiele techniek. En dus ook bij het vinden van digitale informatie.

Historie van zoeken

Cohen neemt de geschiedenis door van zoeken in digitale bestanden. Het begon met het gebruik van steekwoorden. De computer heeft evenwel geen flauw benul van wat de betekenis is van een steekwoord. Zo zal 'hond' een enorme waslijst met resultaten opleveren, want ook 'zeehond', 'honds', 'schond' en noem maar op, bereikt het lijstje. Steekwoord+ moest hierop het antwoord zijn. Hierbij, zo legt Cohen uit, is een aantal regels opgesteld. Zo krijgt 'hond' meer punten als het in een titel van een document voorkomt, en minder als het bijvoorbeeld minder dan drie keer in de rest van een document staat. Nog steeds weet de computer niet wat de betekenis is van 'hond'. Bovendien moeten de regels handmatig worden bijgesteld als het nodig is; en dat vergt veel werk.
Op internet is vervolgens 'page rank' populair geworden. Als een begrip veelvuldig voorkomt op webpagina's en andere pagina's ernaar verwijzen, dan geeft dit een aardige indicatie. "Dat kan wel op internet werken, maar niet binnen de besloten omgeving van een organisatie, want daar zijn vaak geen onderlinge verwijzingen naar begrippen in documenten. Bovendien vind je dan alleen de populaire documenten en niet dat ene, zeer gespecialiseerde document waar verder blijkbaar niemand anders in is geïnteresseerd, maar jij juist wel wilt hebben."

> Sommige zoekmachines gaan bij het indexeren al na wat de relevantie is van een document

Federated Search was het antwoord om ook resultaten te krijgen uit bronnen die niet te indexeren zijn, of die je niet wenst te indexeren. Een zoekmachine vraagt dan aan alle originele repository's om de zoekopdracht aldaar uit te voeren, en brengt de uitkomsten integraal bijeen. Denk bijvoorbeeld aan content die is opgeslagen in repository's van content providers als Factiva en LexisNexis.
Cohen gaat nog een tijdje door met alle verschillende zoekmethoden die in de loop der tijd zijn ontwikkeld. Conceptual Search noemt hij een belangrijke doorbraak. "Daarbij begrijpt de computer het idee achter 'hond' en zal hij bijvoorbeeld ook gaan zoeken naar 'trouwe viervoeter' of alle hondenrassen. Ook begrippen die te maken hebben met 'hond', zoals 'uitlaten' krijgen een plek op de vindlijst."

Victor Cohen: "Het belangrijkste is dat de computer de context begrijpt van datgene waarnaar je op zoek bent".

Beveiliging

Maar nog zijn we er niet: beveiliging speelt evenzeer een rol bij het zoeken naar informatie. "Niet alle documenten mogen door iedereen worden gezien. Gemiddeld mag binnen een organisatie een persoon één op de tienduizend documenten inzien.
Van sommige documenten mogen mensen niet eens het bestaan vermoeden," legt Cohen uit. "Daarom is Secure Search uitgevonden, waarbij rekening wordt gehouden met wie wat mag vinden. Het is niet eenvoudig daarvoor een waterdicht systeem te maken zonder de snelheid van het zoeken te vertragen, noch het netwerkverkeer."
Om sneller met resultaten te komen, gaan sommige zoekmachines bij het indexeren al na wat de relevantie is van een document. Als op dat moment evenwel iets over het hoofd wordt gezien, zal bij een latere zoekopdracht belangrijke informatie niet worden gevonden, omdat niet het gehele document wordt nageplozen.
Tot nu toe hebben we het steeds gehad over tekstdocumenten, maar in de loop der jaren zijn foto's, tekeningen en bewegend beeld erbij gekomen. De verwachting is dat deze informatiestromen sterk gaan groeien nu iedereen bij wijze van spreken met zijn telefoon een foto of filmpje kan maken. Voor bijvoorbeeld schade-experts of makelaars is beeld een handig hulpmiddel. Hier zijn weer andere zoektechnieken voor nodig.
"Het belangrijkste is," stelt Cohen, "dat de computer de context begrijpt van datgene waarnaar je op zoek bent. En bovendien ook snapt dat ik een heel ander woord kan gebruiken voor hetzelfde begrip dan jij."

Bedrijfsregels

En dan komt het hoge woord eruit: meaning based computing. Daar komt nog veel meer bij kijken dan Cohen tot dan toe heeft verteld. "Je kunt bijvoorbeeld ons platform voeden met de bedrijfsregels die binnen jouw organisatie gelden. Elke e-mail die je verstuurt, wordt dan real-time gescand en op de bedrijfsregels beoordeeld. Als je gevoelige informatie wilt verzenden, dan komt automatisch de vraag of je dat wel wilt doen, gezien de inhoud van het bericht. Het zal worden geblokkeerd; geheel automatisch."
Het systeem is bovendien niet passief en wacht niet tot een zoekopdracht leven in de brouwerij brengt. "Nee, ons systeem leert aan welke begrippen jij belang hecht en zal je een notitie melden als een nieuw document dat voor jou van belang kan zijn, ergens in de organisatie is ontstaan."
Autonomy indexeert alle informatie op het moment van creatie; razendsnel. Bovendien is het systeem (waarvan IDOL de basis vormt) geschikt voor organisaties die wereldwijd opereren en talrijke vestigingen hebben.
"Het gaat erom dat alles snel, grondig en automatisch gebeurt. En bovendien taalonafhankelijk is," vertelt Cohen. "Wij kijken naar de betekenis van een begrip, de eigenschappen die erbij horen. Dat is niet aan taal gebonden."

Vrijheidsbeneming

IDOL vormt de kern van het productportfolio van Autonomy. Daar omheen is een groep producten gebouwd die (soms specifiek voor een bepaalde beroepsgroep, zoals juristen met hun eigen taxonomie) helpen informatie beschikbaar te maken en te analyseren.

> Ons systeem analyseert gegevens en draagt informatie aan op grond van gebruikersprofielen

"Tegenwoordig is tachtig procent van alle gegevens binnen een organisatie ongestructureerd. En twintig procent zit keurig in databases. Het is juist die tachtig procent waar ook waardevolle gegevens zitten en die tijdig beschikbaar moeten zijn voor bijvoorbeeld toezichthoudende instanties. Je kunt je – op straffe van vrijheidsbeneming – niet verschuilen achter het niet kunnen vinden van informatie binnen de bedrijfssystemen. Dat alleen al is een reden om informatiebeheer goed op orde te hebben. Een extra reden is dat ons systeem gegevens analyseert en informatie

aandraagt op grond van gebruikersprofielen," zegt Cohen.
De eDiscovery-oplossing van Autonomy (met IDOL als kern) beschikt over alle zoekmogelijkheden die hierboven zijn beschreven. Het systeem kan duizend file-formaten aan, variërend van e-mail, audio, (bewegende) beelden en fileservers tot aan databases en web 2.0-bestanden. Meaning based computing betekent dat automatisch verbanden worden gevonden tussen verspreid opgeslagen informatie. Cohen geeft een voorbeeld: "Als je bezig bent een e-mail op te stellen, dan kan de inhoud daarvan automatisch worden gekoppeld aan een telefoongesprek over dat onderwerp of een bepaalde notitie van een collega over hetzelfde onderwerp."
Die zoektocht hoeft zich overigens niet te beperken tot de informatie die ergens binnen de eigen bedrijfssystemen (wereldwijd) ligt opgeslagen, maar is uit te breiden met informatiediensten buiten de organisatie waarop een abonnement is afgesloten. Het belangrijkste is dat dit alles razendsnel en grondig gebeurt. Analisten als Gartner, Forrester en Ovum zijn van mening dat Autonomy op het vlak van meaning based computing richtinggevend is.

Sentiment onderzoeken

Wie de producten van Autonomy wil gebruiken, kan natuurlijk zelf de Autonomy software aanschaffen, in het rekencentrum plaatsen en bepaalde componenten (zoals eDiscovery, records management of archiving) daarop aansluiten. Maar het is ook mogelijk het informatiebeheer als een dienst af te nemen bij Autonomy. De software en de content staan in dat geval in het rekencentrum van Autonomy.
Volgens Cohen kiezen nogal wat klanten voor deze optie. "Autonomy heeft het grootste data-archief ter wereld. We hebben tien Petabyte op onze servers staan. Zo beheren wij bijvoorbeeld alle e-mails van de City Group."
Het aanbod van Autonomy is zo breed, dat de voorbeelden over de voordelen van meaning based computing blijven komen. Zo vertelt Cohen dat bedrijven het bijvoorbeeld inzetten om na te gaan hoe over hen wordt geschreven, gesproken, dan wel gefilmd in digitale, sociale netwerken. "Wij zoeken ook op context en het sentiment dat bij bepaalde begrippen hoort. Op die manier kun je analyseren hoe jouw organisatie wordt beoordeeld door het publiek; je kunt zelfs nagaan welk type mensen zich positief, dan wel negatief over jouw bedrijf uitlaat. Mocht het nodig zijn, dan kun je actie ondernemen."

Klinkende namen

Implementatie van Autonomy is niet gemakkelijk. Daarvoor is wel een consultant of een partner van Autonomy nodig, zegt Cohen. "Of je kunt gebruik maken van onze eigen raadgevers. Je moet namelijk heel goed nadenken over de inrichting van het systeem, over de beveiliging en over wat je ermee wilt bereiken. Onze partners zorgen ervoor dat alle aspecten aan bod komen. In Nederland zijn bijvoorbeeld Getronics, KnowledgePlaza, Logica en Emid onze partner. Sommige zijn gespecialiseerd in bepaalde branches; daar kun je rekening mee houden. Vooral het inregelen van het systeem kost tijd."
In het klantenbestand komen klinkende namen voor. Denk aan Homeland Security van de Verenigde Staten (het overheidsorgaan dat terroristische aanvallen moet voorkomen), General Motors dat Autonomy Virage inzet om informatie te beheren en analyseren uit de duizenden uren videobeelden die het bedrijf maakt van gesprekken met klanten over auto's (die meningen worden voorgelegd aan het ontwerpteam van de autofabrikant) en de Rechtbank van Amsterdam.

> Je kunt je niet verschuilen achter het niet kunnen vinden van informatie binnen bedrijfssystemen

Concurrenten ziet Cohen op de verschillende productonderdelen waarop Autonomy actief is, zoals Hummingbird voor document management in de juridische sector, of Symantec voor de archivering van e-mail. "Maar er is geen enkel bedrijf dat zoals wij het complete pakket aan meaning based computing aanbiedt. Wij zijn echt de pioniers op dit terrein."

Teus Molenaar is freelance journalist.

Geboren op de universiteit

Het geboortekaartje meldt 1996. De wieg stond op de vermaarde universiteit van het Britse stadje Cambridge. Geestelijk vaders zijn Mike Lynch en Richard Gaunt. De eerste is nog steeds de CEO en de tweede is tegenwoordig lid van de raad van bestuur van de onderneming. Bij het consultatiebureau bleek steeds dat de groeicurve de gemiddelde lijn ver oversteeg.
Inmiddels is het bedrijf de tweede, grootste softwarefabrikant van Europa (met SAP als nummer 1) met wereldwijd talrijke kantoren.
In 1990 had Lynch het bedrijfje Neurodynamics opgericht dat zich vooral op theoretisch vlak bezig hield met bijvoorbeeld patroonherkenning. Op praktisch vlak is destijds apparatuur voor vingerafdrukherkenning ontwikkeld. Lynch heeft de beginselen van Neurodynamics toegepast op digitale informatie. Met de acquisitie in 2005 van Verity versterkte de onderneming haar positie op het toneel van Pan-Enterprise Search. Andere belangrijke overnames waren Virage (2003, videobewerking), Zantaz (2007, e-mail archivering) en Interwoven (2009, ECM).
Nog wat cijfertjes: 22 procent van de winst gaat naar Onderzoek & Ontwikkeling; ongeveer 65 procent van de omzet komt uit de VS, 2000 medewerkers, 20.000 klanten, waarvan ongeveer 600 in de Benelux, 400 implementatie-partners, 400 OEM-overeenkomsten, onder meer met IBM en HP.