Autonomy-baas Mike Lynch: Verity was een goede acquisitie

Email- en telefoon verkeer, audio en video. Dat worden volgens Autonomy-CEO Mike Lynch de expansiegebieden van het computer-zoeken. Lynch was half november korte tijd in Nederland. InformatieProfessional sprak met hem.

Eric Sieverts en Martijn de Groot

Foto: johannes Abeling

Mike Lynch

IP: Autonomy gaf deze maand klinkende cijfers vrij over omzet en winst in het derde kwartaal. Is er een bijzondere oorzaak voor de goede resultaten?

Lynch: Informatieprofessionals, maar ook managers zijn de laatste jaren steeds meer het belang van ongestructureerde informatie gaan inzien. Het besef is gegroeid dat zulke informatie belangrijke risico's kan inhouden als hij niet goed doorzoekbaar is, bijvoorbeeld in de juridische sfeer. In de VS moet je als bedrijf verantwoording kunnen afleggen over elke actie die jij of je medewerkers hebben ondernomen. Als de nood aan de man komt, moet je dat ene stuk gewoon vinden en daarvoor heb je gereedschap nodig dat met ongestructureerde informatie overweg kan. In het bedrijfsleven is men dat meer en meer gaan beseffen en Google heeft dat bewustzijn sterk bevorderd.

IP: U heeft over Google ook gezegd dat het de informatiezoekers dom houdt.

Lynch: Google en andere zoekinstrumenten op basis van losse woorden, werken erg grof. Ze komen met een overvloed aan resultaten waar heel vaak toch niet bij zit wat je zoekt, omdat de zoekmachine de betekenis niet snapt van wat je in het zoekvakje hebt ingetikt. En dan krijg je de resultaten ook nog gepresenteerd in volgorde van het aantal links naar een pagina. Maar in een bedrijf is populariteit van een document helemaal geen goed criterium.

IP: Google past toch ook nog wel andere, slimme technieken toe zoals waar zoekwoorden voorkomen en hoever ze uit elkaar staan.

Lynch: Jawel, maar die zijn allemaal gebaseerd op het uitgangspunt van massaal zoeken op losse woorden. Dat is proberen van iets wat slecht is toch nog wat te maken... Als ik een bibliotheek binnenkom en ik loop naar de balie en ik zeg tegen de bibliothecaris: 'Pinguïn'... dan zal die bibliothecaris een paar vragen stellen om erachter te komen wat de betekenis en de context van dat woord 'Pinguïn' is. Een zoekmachine die is gemaakt om betekenis te begrijpen zal ook terugkomen met relevante informatie. Dat is wat wij *meaning based computing* noemen.

IP: Is de overname van Verity door Autonomy een goede stap geweest?

Lynch: Dat was een goede acquisitie. We hebben er goede

nieuwe mensen bij. En omdat wij in de voorafgaande jaren moesten opboksen tegen Verity, dat in de markt al een positie had toen wij net kwamen kijken, hebben we er in die tijd voor gezorgd dat onze producten aansloten op die van hun. Klanten die overstapten hadden dan geen problemen. Daar hebben we nu enorm veel profijt van.

IP: Is er nu dan nog wel genoeg concurrentie om jullie scherp te houden?
Lynch: Bij zo'n tachtig procent van de deals die wij nu sluiten komt geen concurrentie te pas. Op de lange duur is dat misschien geen goede zaak. In het verleden hebben we veel profijt gehad van de invloed van concurrerende bedrijven.

IP: Is het Noorse bedrijf Fast geen serieuze concurrent? Er was de nodige animositeit met dat bedrijf in de afgelopen periode.
Lynch: We komen ze tegen bij 20 procent van de deals die we afsluiten, en in die gevallen hebben we een 90 procent winscore. Waar we wel last van hebben gehad is dat Fast de markt probeerde te beïnvloeden met onjuiste cijfers over hun omzet en klantenaantallen. Ze zijn daarin nu officieel terecht gewezen door de Noorse overheid.

IP: We hebben wel eens de indruk dat de basistechnologie achter het product van Autonomy altijd hetzelfde is gebleven. Zijn er in de komende tijd nog nieuwe dingen te verwachten?
Lynch: Ik ga veel kijken bij nieuwe bedrijfjes en dan zie ik elke keer dat je in wezen drie soorten zoektechnologie hebt. De eerste werkt op basis van zoeken op woorden. Die techniek zit momenteel aan zijn plafond. De tweede geeft op basis van taalkundige regels betekenis aan woorden, maar dat stuit snel op beperkingen. Als ergens staat: 'The dog entered the room. It was furry', dan kun je op basis van grammatica niet weten of 'it' nu slaat op de hond of op de kamer. Technieken die dat proberen, lopen vroeger of later vast. De derde methode, die wij gebruiken, zoekt met kansrekening naar verbanden en beslist op basis daarvan wat de meest waarschijnlijke betekenis is.

IP: In het kader van het semantisch web wordt veel nadruk gelegd op het belang van taxonomieën en ontologieën om de computer te laten begrijpen waarover teksten gaan.
Lynch: Tegen mensen die het heil verwachten van semantisch georiënteerde ontologieën zeg ik tegenwoordig gewoon: 'Bel me als je klaar bent'. Het probleem blijft bovendien het gebruik van trefwoorden. Hoe kun je nou een verhaal van tweeduizend woorden in vijf tags stoppen? En waarvoor waren die andere 1995 woorden dan bedoeld?

IP: Wat worden voor uw bedrijf belangrijke nieuwe ontwikkelingen?
Lynch: In het bedrijfsleven is er een enorm emailverkeer en het belang om daarin effectief te kunnen filteren wordt steeds groter. Je moet dus een machine hebben die dat met een hoge trefkans op basis van betekenis kan doen. Als die machine in staat is 95 procent van de documenten goed te benoemen en daarbij die laatste 5 procent apart aflevert voor menselijke beoordeling, dan doet hij het goed. Telefoonverkeer wordt ook steeds belangrijker vanuit juridisch en management oog-

'Een zoekmachine die is gemaakt om betekenis te begrijpen zal ook terugkomen met relevante informatie'

punt. Daar moet je dus met spraakherkenning werken. Opvallend is dat je ook daar weer in de problemen komt als je niet op basis van *meaning* werkt. Luister maar eens in het Engels naar dit: *'recognize speech'*. Kan jij horen of dat over spraakherkenning ging of over het vernietigen van een mooi strand (*'wreck a nice beach'*)? Daar kom je pas achter met behulp van de context.

IP: Gaat de discussie over privacy niet remmend werken op de ontwikkeling van het zoeken in telefoonverkeer?
Lynch: Er is altijd een balans tussen privacy en bepaalde collectieve doelen, zoals misdaadbestrijding of het bedrijfsbelang. Maar mensen houden niet alleen van privacy. Gemak is ook belangrijk. En een belangrijke overweging is dat de machine soms beter het controlerende werk kan doen dan een mens, omdat die machine geen vooroordelen kent en discreet is.

IP: Nog even over websearch. Het aantal te doorzoeken webpagina's blijft steeds verder groeien. Komt er niet een bovengrens waarboven zoeksystemen principieel niet meer kunnen werken. Een grens aan de schaalbaarheid dus.
Lynch: Ik geloof niet dat we daar last van zullen krijgen. Ik denk dat andere nieuwe ontwikkelingen in de sfeer van *consumer search* boeiender zullen zijn. *Implicit query* bijvoorbeeld. Het programma kijkt voortdurend wat er op je scherm staat en biedt uit zichzelf links aan over onderwerpen die daarmee te maken hebben. *Transaction hijacking* is een andere. Je staat op het punt iets te kopen via internet en de computer kijkt nog even of je dat betreffende product niet ergens anders goedkoper kan krijgen. De rol van video en audio wordt ook steeds groter. Alleen is die nog moeilijk doorzoekbaar, maar dat zal niet lang meer duren; daar gaan wij voor zorgen.
Schaalgrootte kan trouwens binnen het bedrijfsleven wel een probleem worden, maar dan vormen toegangsrechten de hindernis. Om een scherm met antwoorden op een zoekvraag op te bouwen moet vaak honderden keren een document worden opgevraagd, voor er eindelijk weer een is waarop de zoeker rechten heeft en dat dus getoond mag worden. Dat is op zich niet erg, maar de computer is al die tijd wel aan het werk. Je moet dus eigenlijk tevoren weten hoe toegangsrechten zijn verdeeld, want anders wordt de belasting voor het netwerk te groot. Autonomy heeft hiervoor een oplossing, genaamd *Mapped Security*. <