Eric Sieverts

Keith van Rijsbergen: *'Het Boleaanse model is natuurlijk vreselijk'*

Als buitenlandse spreker op de IP-lezing was professor Keith van Rijsbergen, autoriteit op het terrein van information retrieval, onze gast. In een gesprek vroegen we zijn mening over een aantal ontwikkelingen op dit gebied en kwamen we ook wat meer over hem persoonlijk te weten.

Keith van Rijsbergen is van zeer Hollandse afkomst: hij stamt uit een Sliedrechtse familie van – hoe kan het anders – baggeraars. De naam Keith is van het al even Hollandse Kees afgeleid. En zijn achternaam werd destijds in Australië ineens zonder vragen correct gespeld toen achterneef Wim – voor Nederland – aan het WK-voetbal deelnam, een feit dat Keith in herinnering werd gebracht door de aanwezigheid van het huidige Nederlands elftal dat zich in dezelfde hotel-lounge, al klaverjassend en GSM-end, op de oefenwedstrijd tegen Brazilië voorbereidde.

Uit uw CV weten we dat u weliswaar in Nederland bent geboren, maar dat u al vanaf uw jeugd een zeer internationale levensloop hebt gehad. Hebt u nog wel in Nederland de middelbare school afgemaakt?
'Ik heb de exacte jaartallen nooit zo heel precies bijgehouden, maar ik denk dat ik tot ongeveer de tweede klas van de HBS in een Nederlandstalige omgeving ben opgegroeid, in Nederland en in Indonesië. In Djakarta was Soekarno zelfs nog een van onze buren. Daarna ben ik verder in Namibië en vooral Australië opgegroeid. Mijn moeder woont nu nog steeds in Australië.'

Nadat u al zo lang in zo veel verschillende landen hebt gewoond, ben ik benieuwd of u toch nog een binding voelt met uw geboorteland.
'O, beslist. Ik voel nog altijd heel sterk dat mijn culturele wortels hier in Nederland liggen. Een van de redenen dat ik uit Australië teruggekomen ben naar Europa, hoewel mijn ouders daar bleven wonen, was dat mijn vrouw en ik – zij is een Engelse – heel sterk het gevoel hadden dat onze wortels hier in Europa lagen. En ook met Nederland heb ik nog een sterke band door het lezen van Nederlandse boeken, regelmatige bezoeken – toch wel zo'n drie keer per jaar – en culturele evenementen. En ook historisch nog altijd. Zo heb ik na het overlijden van mijn vader een interessante collectie foto's in bezit, van de wederopbouw van de Rotterdamse haven na de oorlog, waaraan mijn vader heeft meegewerkt.'

Hebt u het gevoel dat die internationale levensloop een voordeel is geweest voor uw professionele carrière?
'Zeker, maar dat is wel heel persoonlijk. Mijn broer en zuster hadden een soortgelijke levensloop, zonder dat dat zo heeft uitgepakt. Voor mij is het een heel sterke intellectuele stimulans geweest zo rond te trekken en aan verschillende talen te worden blootgesteld. Ik heb Indonesisch, Duits, Engels, Afrikaans leren spreken. En mede doordat ik in een aantal verschillende landen op school ben geweest, ben ik ook in contact gekomen met veel verschillende culturen en het daarbij horende verleden. Dat is een voordeel geweest, te meer omdat dat op tamelijk jeugdige leeftijd was, waarop je snel talen oppikt. Daardoor ben ik ook veel meer "relaxed" met vreemde talen dan veel Engelsen.'

Van hieruit gezien lijkt Glasgow, waar u nu aan de universiteit verbonden bent, een beetje een uithoek van Groot-Brittannië, en – in Schotland – ook nog een natte uithoek.
'O nee, beslist niet. In academisch opzicht behoort het tot de paar belangrijkste informatica-opleidingen in het land. Maar ook als stad is Glasgow bijzonder interessant. In de vorige eeuw was het dankzij de scheepsbouw meteen na Londen een van de belangrijke steden van het land. Uit die tijd, maar ook uit de tijd daarvoor, heeft het een prachtig historisch centrum overgehouden. Er is weliswaar een periode van economische neergang geweest, maar Glasgow zit nu alweer vele jaren in de lift. Het is misschien wat minder bekend dan Edinburgh met zijn festival, maar het centrum is beslist groter en zeker zo mooi. Ons eigen instituut is gevestigd in een rij historische herenhuizen. Als beschermd stadsgezicht mag er aan de buitenkant niets veranderd worden, zodat ze alleen van binnen zijn doorgebroken.
En wat de regen betreft. Daar zijn heel grote lokale verschillen in. Zelf vind ik het Schotse landschap zo ongeveer het allermooiste wat je op de wereld kunt vinden. En voor die begroeiing is nu eenmaal vocht nodig. Ik houd erg van wandelen. Daarbij moet je inderdaad voor geschikte uitrusting zorgen. Soms is het dan wel heel bewolkt en mistig, maar als je op een bergtop komt, kan plotseling de lucht helemaal openbreken en krijg je de meest schitterende vergezichten.'

U bent nu alweer dertien jaar verbonden aan de Universiteit van Glasgow. Hebt u daar in die tijd een toonaangevende onderzoeksgroep kunnen opbouwen?
'Op dit moment heb ik een groep van ongeveer tien medewerkers. Ik denk dat het daarmee wel ongeveer de grootste onderzoeksgroep op dit terrein in Groot-Brittannië is. Wat goed is aan deze groep – ook mede door mijn eigen belangstelling voor het continent, de rest van Europa – is dat we veel projecten hebben met partners in andere Europese landen. En dat betekent dat veel jonge onderzoekers uit Nederland, Frankrijk, Denemarken, Duitsland, Italië naar Glasgow komen om enige tijd in mijn groep te werken. Ook door studenten die hier zijn gepromoveerd, bestaat er een heel internationaal netwerk, zoals bijvoorbeeld met Bruce Croft die nu in Amherst Massachusetts de grootste IR-onderzoeksgroep in de Verenigde Staten leidt. Daar hebben ze ook het Inquery-systeem gebouwd. Doordat ik al vanaf de allereerste afleveringen bij de SIGIR Information Retrieval-conferenties betrokken ben geweest en daar al begin jaren zeventig – toen nog vanuit Cambridge – met mijn studenten heen ging, blijf ik dat hele netwerk van mensen ontmoeten, dat nu over de hele wereld verspreid zit.'

Zelf heb ik de indruk dat fundamenteel onderzoek op het gebied van information retrieval heel lang weinig impact heeft gehad op praktische toepassingen. Dat we pas de laatste vijf of hooguit tien jaar toepassingen zien die verder gaan dan het klassieke Booleaanse inverted-file model.
'De afstand tussen fundamenteel onderzoek en toepassingen is inderdaad heel sterk verminderd. En dat komt niet omdat de systeembouwers plotseling veel vriendelijker geworden zijn jegens de mensen aan de universiteiten, maar vooral omdat veel grote bedrijven, zoals MicroSoft, maar ook kleinere bedrijven, zelf zijn begonnen onderzoekers op het gebied van IR in dienst te nemen. Zo heb je bij MicroSoft nu een heel goede IR-onderzoeksgroep. Hoe groot de werkelijke impact van fundamenteel onderzoek op praktische toepassingen is, is natuurlijk veel moeilijker te meten. Zo weten maar betrekkelijk weinig mensen dat de huidige Muscat retrieval-software heel sterk berust op fundamenteel onderzoek dat al in de jaren zeventig is gedaan. In dat geval heeft daar dus wel een heel grote tijdspanne tussen gezeten. Dat die tijdsduur nu sterk aan het verminderen is geloof ik wel. Je ziet bijvoorbeeld ook dat ik nu grote moeite heb om mensen in mijn groep vast te houden, omdat ze voortdurend onderzoeksbanen in het bedrijfsleven aangeboden krijgen – waar die afstand natuurlijk per definitie veel kleiner is. En dát is echt pas een ontwikkeling van de laatste paar jaar.'


FOTO: EGON VIEBRE

Heeft de komst van het web ook een belangrijke rol gespeeld om die afstand te verkleinen?
'Door die grotere digitale omgeving is er een veel sterkere bewustwording gekomen van wat IR eigenlijk inhoudt. Dat heeft zeker bijgedragen.'

Veel mensen zien het web, met zijn enorme ongestructureerde, gedistribueerde digitale chaos als een interessant testdomein voor nieuwe retrieval-methoden. Anderzijds heb je ook de meer gecontroleerde experimenten van de jaarlijkse TREC-competitie. Welk van beide zorgt naar uw gevoel op dit moment voor de meeste vooruitgang?
'Als je me dat zes jaar geleden had gevraagd, had ik zonder aarzelen "TREC" geantwoord. Maar, als je echt serieuze experimenten doet aan de gedistribueerde informatie op het web, is dat nu waarschijnlijk van meer belang. Ik ben wat dat betreft beslist van mening veranderd. Men wordt zich er ook meer en meer van bewust dat de onderliggende systemen van het web niet goed genoeg zijn voor information retrieval. Maar het biedt wel een geweldige kans om daarmee te experimenteren. Daarbij moet je wel zorgen dat je je experimenten op een redelijk gecontroleerde manier doet. Aan de andere kant blijven nu veel mensen steken in het steeds opnieuw herhalen van hetzelfde experiment. Als iets eenmaal werkt, hoef je die methode niet elk volgend jaar opnieuw in TREC te testen. Mensen zouden wat dapperder moeten zijn om helemaal nieuwe dingen uit te proberen en dat dan ook te doen in die nog weer moeilijker en uitgebreider omgeving van het web.'

Uw naam is verbonden aan een bekend leerboek, "Information Retrieval". De meest recente druk daarvan is nu alweer twintig jaar oud. Toch kun je die nog in PDF van het web halen. Hebt u nooit de neiging gehad daar een nieuwe versie van te schrijven?
'Dat oude boek komt zelfs weer opnieuw uit! Iemand in de Verenigde Staten, Rik Belew van de Universiteit van San Diego, heeft een nieuw leerboek over IR geschreven. Dat heeft heel gedetailleerde links naar passages in mijn boek. Daarom stopt hij een cd-rom met mijn boek achter in zijn boek, zodat zijn lezers meteen al zijn verwijzingen kunnen opzoeken. Maar ik vind dat eigenlijk wat gênant, want dat boek was tien jaar na verschijnen natuurlijk al achterhaald. Alleen is er toen heel lang een "gap" geweest, omdat er nog geen vervangers voor waren. Nu beginnen die wel te komen. Daardoor vind ik eigenlijk dat het nu tijd wordt mijn boek maar liever ergens te begraven!'

En als u zelf een nieuw boek zou moeten schrijven, wat zouden dan de belangrijkste doorbraken van de laatste tien jaar zijn die daar beslist in opgenomen zouden moeten worden?
'In elk geval formele modellen om de interactie met de gebruiker te beschrijven. Ik ben ook van plan om volgend studiejaar een sabbatical year te nemen om een boek te schrijven. Daarin wil ik heel nauwkeurig naar "logica" gaan kijken – logica in een heel brede betekenis. Maar vooral de logische basis voor interactie met de gebruiker. Na mijn lezing vanmiddag was er iemand die vroeg of je, na het stellen van een vraag aan een systeem, enige tijd later, als je die vraag nog een keer zou stellen, niet alweer een heel andere mening bent toegedaan over wat wel of geen interessante antwoorden op je vraag zijn. Dat is een bijzonder belangrijk aspect van IR, waarvoor we moeten proberen modellen te ontwerpen.'

Ondanks de gegroeide impact van fundamenteel IR-onderzoek op praktische toepassingen, lijkt veel van dat onderzoek mij toch nog altijd erg theoretisch of zelfs esoterisch, met een grote afstand tot toepassingen uit de echte wereld. Wat is uw inschatting hoeveel ervan uiteindelijk echt ergens toegepast zal worden?
'Een aantal mensen dat theoretisch werk doet, zoals ikzelf of Bruce Croft, heeft eigenlijk alleen maar belangstelling voor theoretische problemen, als ze – intuïtief – de verwachting hebben dat het van enig praktisch nut zal zijn. In die gevallen is die kans redelijk hoog, ook al kan het vele, vele jaren duren voor het werkelijk wordt toegepast. Aan de andere kant kom je op congressen ook mensen tegen, die daar een lezing komen geven, alleen maar om weer een publicatie aan hun lijst te kunnen toevoegen. Terwijl de kans dat hun werk ooit tot iets praktisch zal leiden vrijwel nihil is. Als ik optimistisch ben, denk ik toch dat zo'n vijftig procent ooit tot toepassingen zal leiden, als ik pessimistisch ben, misschien dertig procent.'

Zoveel toch nog?
'Ja, maar dat hangt er natuurlijk vanaf wat je als een toepassing beschouwt. Niet alles zal zo algemeen gebruikt gaan worden als bijvoorbeeld in een AltaVista. Maar als je ook allerlei kleine toepassingen meerekent, die ergens lokaal in een bepaalde bedrijfstoepassing worden ingebouwd, denk ik dat je uiteindelijk aan dergelijke percentages zult komen.
Zelfs onderzoek waarvan je dat helemaal niet zou verwachten, omdat het destijds alleen maar op heel beperkte collecties werd uitgevoerd, blijkt nu toch toepasbaar. Een voorbeeld is het werk aan documentclustering. In de jaren zeventig werd dat gedaan aan heel beperkte sets van gegevens en heel lang had niemand enig idee hoe dat ooit op grote sets toegepast zou kunnen worden. Als je nu kijkt, zie je dat het intussen veel – en heel effectief – in grote systemen wordt toegepast bij de nabewerking van resultaten. Bijvoorbeeld bij persberichten, waar veel bijna identieke berichten binnenkomen en clustering de doublures eruit kan filteren. Aan clustering en visualisatie zie je dat een theoretische aanpak die in een bepaalde context is bedacht, ineens tien jaar later in een iets andere context een essentieel stukje theorie blijkt te zijn geweest.'

FOTO: EGON VIEBRE
Keith van Rijsbergen en Eric Sieverts tijdens de IP-Lezing 1999

Speciaal het probabilistische model scheen mij altijd een hoog theoretisch gehalte te hebben (hoewel we zelf in Utrecht intussen software gebruiken die hierop berust). Een grootheid als "de kans dat iets relevant is voor een vraag, gegeven bepaalde andere voorwaarden", lijkt me erg moeilijk te concretiseren, zolang nog niets over de zoeker bekend is.
'Dat klopt. Dat model gaat er dan ook vanuit dat je het retrieval-proces op een of andere manier aan de gang krijgt. En dat hoeft niet heel moeilijk te zijn. Zodra het systeem dan eenmaal wat informatie van de gebruiker gekregen heeft, is het niet zo moeilijk meer. Het is een zelflerend model. Er zijn nu dan ook al heel wat systemen die dergelijke in de jaren zeventig ontwikkelde eenvoudige modellen gebruiken.'

Mensen willen vaak weten wat het beste systeem in een bepaalde situatie is. Het Booleaanse model, het vector-model of het probabilistische model…
'Het Booleaanse is natuurlijk vreselijk! Dat zoeken op het web vaak niet goed gaat komt door de tekortkomingen van het Booleaanse model. Als je dat op een simpele manier toepast krijg je óf een nul-resultaat, óf alles wat er in het systeem zit. Het probleem is dan altijd hoe je tot een middenweg komt.'

Is dat niet wat overdreven? Er zijn toch systemen waarin je met Booleaanse methoden heel goede resultaten kunt krijgen. Bibliografische databases met goed met trefwoorden gekarakteriseerde documenten…
'O zeker, als je de gegevens heel goed kent en als je zelf heel goed weet wat je moet doen als er niets gebeurt of als juist "alles gebeurt", dan kun je ook daar nog wel goede resultaten mee krijgen. Maar in vrijwel alle gevallen zijn methoden die de resultaten ranken toch veruit superieur. En dan vooral de probabilistische methoden. Maar eigenlijk maakt het niet eens zo veel uit of je nu een vector-model gebruikt en feedback toepast of dat je een helemaal probabilistische methode gebruikt. Theoretisch kun je zelfs de ene methode in termen van de andere definiëren en omgekeerd. Het is dus wat kunstmatig om te proberen tussen die twee modellen een wig te drijven. Alleen het Booleaanse model moet, afgezien van bepaalde zeer speciale toepassingen, beslist als inferieur gezien worden.'

Dr. E.G. Sieverts is werkzaam op de Hogeschool van Amsterdam en bij de Universiteitsbibliotheek Utrecht. Hij is redacteur van Informatie Professional.