

# Information retrieval belicht

De IP-Lezing was met ruim 160 bezoekers een succes. Voor hen die op 7 oktober niet aanwezig konden zijn volgt hier een samenvatting van de voordrachten.



Ontspannen verteltrant van Keith van Rijsbergen

FOTO: DR. ECON VIERBE

ALS EERSTE SPREKER wees Thijs Chanowski ons op de onvolkomenheden van veel zoeksystemen, of liever van de wijze van interactie tussen systeem en gebruiker. Alleen als je exact de goede vraag stelt, krijg je ook het goede antwoord. Of dat nu bij een gestructureerde database is of bij een retrieval-systeem. Als je niet precies de goede vraag stelt, wat veel waarschijnlijker is, zou je toch ook op zijn minst een soort van antwoord moeten krijgen. Daar zijn oplossingen voor te bedenken. Zulke mede door hem bedachte en door zijn Medialab ontwikkelde oplossingen noemt Chanowski 'liquid solutions'.

Dat 'vloeibare' van Chanowski biedt onder meer allerlei varianten op wat anderen ook wel 'fuzzy matching' noemen. Maar dan wel intelligente fuzzy-matching. Als antwoord op je vraag moet je ook gegevens krijgen die alleen maar enigszins lijken op wat je vroeg; en dat moet zelfs gebeuren als er wel een exact antwoord is. Het hangt dan af van de context van je vraag of van het soort informatiesysteem

waarmee wordt gewerkt, hoe dat 'lijken op' moet worden geïnterpreteerd. Als je vraag een geografische component heeft, zou een systeem ook gegevens over een geografisch dichtbijgelegen plaats moeten opleveren of zou het bij een niet correct gespelde plaatsnaam meteen moeten reageren met een lijstje van alleen plaatsnamen die op het ingetikte lijken.

In het geval van een bibliotheekcatalogus zou het systeem kunnen komen met verschillende lijstjes alternatieve of aanvullende zoektermen. Woorden die voorkomen in titels of beschrijvingen die met het door jou ingetikte woord gevonden worden – zoals sommige zoekmachines op het web, maar dan in wat minder aantrekkelijke vormgeving, ook wel doen – en tegelijk ook woorden die op grond van een aanwezig woordsysteem een inhoudelijke relatie hebben met wat jij intikte en ook nog woorden waarvan de spelling meer of minder lijkt op wat je hebt ingetikt. Ook dat soort hulp wordt wel vaker vertoond,

alleen nooit allemaal gelijktijdig, zoals – heel fraai gevisualiseerd – in het Aquabrowser-interface op bibliotheek-OPACS en retrieval-systemen.

Wat in de lezing wat onderbelicht bleef en pas na afloop, door vragen uit de zaal, aan de orde kwam, was hoe je al die extra kennis in je systeem krijgt. Het merendeel daarvan zul je er toch inderdaad tevoren moeten instoppen. En daarvoor blijken beschikbare thesauri,

semantische netwerken, synoniemlijsten, al met trefwoorden ontsloten documenten, geografische kennissystemen en dergelijke toch onontbeerlijke input te vormen.

Hoewel het voor sommige aanwezigen misschien een bekend verhaal was, blijft het een genoegzaam naar Chanowski's betoog te luisteren.



Het complete forum o.l.v. IP-redacteur Eric Sieverts

**D**E TWEEDE BIJDRAGE werd verzorgd door Peter Geytenbeek en Benoit Goes van het Belgische taaltechnologiebedrijf Lernout & Hauspie. Toen zij werden uitgenodigd, was nog niet bekend dat het te tonen systeem voor spraakaansturing van Microsoft-producten, VoiceXpress, een al zover gecommmercialiseerd massaproduct zou zijn, dat we het intussen bijna dagelijks op de radio krijgen aangeprezen. Dat neemt niet weg dat het indrukwekkend is te zien hoe een tekstverwerker met gesproken opdrachten en gedicteerde tekst hands-free bediend wordt.

Dat bij deze live demonstratie niet alles foutloos verliep, kon hopelijk op rekening geschreven worden van de wat onnatuurlijke situatie, waarin de microfoon de ingesproken tekst ook nog een keer met minieme vertraging (en met enige vervorming) uit de geluidsinstallatie opving, waar de voorbereidende training op de stem van de spreker niet tegen opgewassen bleek. Overigens begint het wel te intrigeren of over enkele jaren inderdaad hele kantoorruimten zullen zijn, vol met door elkaar heen - en daarvoor met toenemende stemverheffing - tegen hun pc sprekende beambten.

Een andere ontwikkeling waar het Belgische bedrijf zich mee bezighoudt, betreft het gebruik van automatische vertaalsystemen, onder meer om informatiesystemen in



*Altijd een genoeg naar Thijs Chanowski te luisteren*

verschillende talen te kunnen samenvoegen. Dat is bijvoorbeeld van belang bij het samengaan van bedrijven met vestigingen in verschillende Europese landen. Als bij een vraag in de taal van het ene land, niet ook gezocht wordt in informatiesystemen van de partner, in documenten in de taal van dat andere land, en als zo gevonden documenten niet op zijn minst een automatische ruwe vertaling ondergaan, zal bijzonder weinig van elkaars kennis en informatie geprofitteerd kunnen worden. Dit onderwerp lag veel dichterbij het centrale thema van de middag, maar oogde natuurlijk wat minder spectaculair dan het voorgaande.



*Peter Geytenbeek dicteert de computer hands-free*

**A**LS SLOTSPREKER VERSCHIEEN tenslotte professor Keith van Rijsbergen. Als enige gebruikte hij - op suggestie van de voorzitter - een gewone overheadprojector. Maar hij verklaarde wel, voor het geval die projector het zou begeven, als back-up een powerpointpresentatie bij zich te hebben. Hoewel Nederlander van origine, kon hij, gezien zijn optreden, intussen makkelijk voor een Engelsman doorgaan. In ontspannen verteltrant vatte hij de ontwikkelingen op het gebied van information retrieval van de laatste veertig jaar samen.

Dat de theoretische achtergronden van veel achterliggend fundamenteel onderzoek niet voor alle toehoorders gesneden koek waren, deed eigenlijk niet zo veel ter zake. Het was al heel inspirerend om eens iemand te horen die we in onze kring van vooral praktijkgebruikers anders eigenlijk nooit tegenkomen, maar die wel aan de wieg heeft gestaan van allerlei - soms pas recent in praktijk gebrachte - ontwikkelingen op retrieval-gebied.

Uiteraard kwamen de drie basismethodieken voor information retrieval - het Booleaanse model, het vectormodel en de probabilistische methode - aan de orde. Daarna werd echter ook ingegaan op verdere uitdagingen, zoals het in systemen modelleren van gebruikersinteractie. Op een aantal van deze punten wordt uitgebreider ingegaan in het interview met Keith van Rijsbergen dat elders in deze aflevering van Informatie Professional is afgedrukt. In de loop van zijn lezing bleek dat hij uiteindelijk ook genoeg stof gehad zou hebben, om nog een tweede drie kwartier te vullen. Omdat toehoorders nog vragen moesten kunnen stellen en de altijd belangrijke wandelganggesprekken tussen die toehoorders - onder het genot van een drankje - niet gefrustreerd mochten worden, konden we dat helaas niet meer toestaan. Maar wie weet komt er ook in onze kring nog eens een vervolg op zijn verhaal, al dan niet in een IP-Lezing ergens in de volgende eeuw.