

Informatiesporen in de computer

Spreektaal en schrijftaal zijn niet exact en hoeven dat in het dagelijks verkeer ook niet te zijn. Iedereen begrijpt de vraag: 'Wilt U een beetje opzij gaan?', afstand en richting hoeven niet opgegeven te worden, maar volgen uit de situatie. Wie zegt: 'Wilt U zich 65 cm in noord-oostelijke richting verplaatsen?' beoogt of een humoristisch effect, of geeft te kennen dat hij verschrikkelijk pedant, zo niet enigszins gek is.

Slechts bij zeer speciale toepassingen stuit de onnauwkeurigheid van de gewone taal op bezwaren. Wiskunde en logica zijn één voorbeeld, hier heeft men steeds meer de gewone taal vervangen door symbolische notaties en symbooltalen waarin ieder symbool een eenduidige betekenis heeft. Een veel nieuwer gebied waar de onnauwkeurigheid van de gewone taal moeilijkheden oplevert is 'information retrieval' met behulp van de computer. Hier werkt men met zeer grote bestanden – bijvoorbeeld titels en samenvattingen van publikaties – die in het geheugen van de computer zijn opgeslagen. De gebruiker wil die publikaties uit het bestand lichten die hij voor zijn werk nodig heeft. Dan krijgt men te maken met de tegenstelling tussen de onnauwkeurigheid van de taal en de grote mate van exactheid waarmee de computer te werk gaat. Voor het zoeken in een bestand werkt men in vrijwel alle systemen met trefwoorden, die een bepaald onderwerp karakteriseren. De opgave van de trefwoorden moet exact gebeuren omdat de meeste systemen niet tolerant zijn voor afwijkende spellingen en tikfouten. Is het trefwoord GELUIDSHINDER en tikt de gebruiker in GELUIDSHINDER, dan kunnen de meeste systemen dit niet als hetzelfde begrip herkennen.

In vele gevallen zal men niet met één trefwoord kunnen volstaan, maar er meer nodig hebben om een bepaalde vraag te karakteriseren. Dan moet de gebruiker weten hoe hij het bestand moet

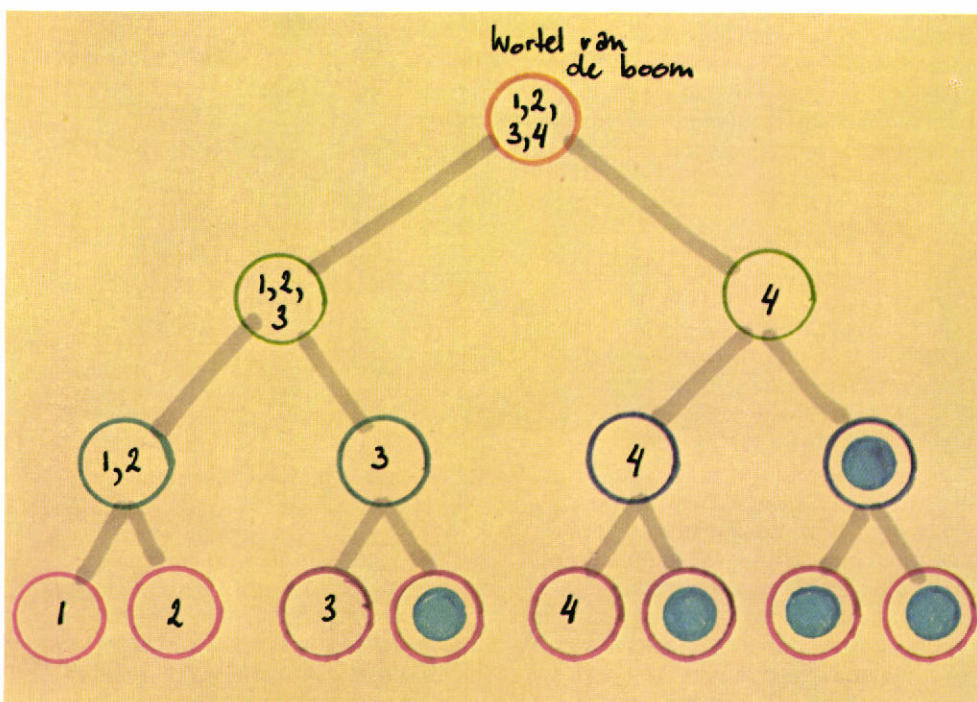
ondervragen en welke combinatie van trefwoorden hij moet toepassen om de grootste kans op een goed antwoord te hebben. Dit is in nog sterkere mate het geval als niet één, maar verschillende bestanden doorzocht moeten worden, want die lopen qua indeling van de informatie vaak enorm uiteen. Terminals maken het de vragensteller mogelijk door middel van een dialoog met de computer zijn vraag nader te preciseren als de antwoorden hem niet bevallen, maar zij nemen het bedieningsprobleem niet weg. Zelfs met een terminal kan men de hulp van een vakman, die de systemen door en door kent, meestal niet ontberen. De gebruiker moet samen met de vakman de beste formulering uitwerken voordat de vraag op het bestand in de computer losgelaten wordt. Maar ideaal zou zijn als de gebruiker, zonder vooropleiding, zelf achter de terminal kon gaan zitten om zijn vraag te stellen in een formulering die hem het beste voorkomt en dan als antwoord krijgt: 'Publikatie X komt het meest overeen met Uw vraag; is dat wat U zoekt?' Dan is de computer aan de mens en de onnauwkeurigheid van de menselijke taal aangepast, in plaats van dat de mens zich moet conformeren aan de exactheid die de computer eist.

TNO heeft onlangs een programma voor information retrieval voltooid dat aan deze wensen vergaand tegemoet komt. Het is FUZZIE gedoopt en men kan er mee werken zoals men zelf wil: zowel trefwoorden als vragen in een gewone taal worden geaccepteerd.

Dit nieuwe programma is gebaseerd op twee ideeën. Het eerste kan men beschouwen als het grondbeginsel van ieder systeem van information retrieval, nl. dat er tussen een vraag en de goede antwoorden daarop een zekere mate van overeenkomst zal zijn. Wie publikaties zoekt over 'Droes bij paarden' kan met grote waarschijnlijkheid verwachten dat in de titels en samenvattingen van de

goede antwoorden de woorden 'droes' en 'paard' voorkomen. De in gebruik zijnde systemen werken alle volgens dit principe. Maar daar kiest de gebruiker eerst de trefwoorden waarvan hij verwacht dat zij ook in de gezochte publikaties zullen voorkomen en voert die in de computer in. De ontwerpers van FUZZIE meenden dat ook deze stap door de computer gedaan moest worden. Dit leidde direct tot de vraag hoe de computer kan bepalen wat de

Het adres van informatie-record nr. 1 is <LLL>
Het adres van informatie-record nr. 2 is <LLr>
Het adres van informatie-record nr. 3 is <LrL>
Het adres van informatie-record nr. 4 is <rLL>
Als voorbeeld voor de informatie-records zelf kan genomen worden
informatie-record nr. 1 is 'AMSTERDAM'
informatie-record nr. 2 is 'HAMSTER'
informatie-record nr. 3 is 'ADAM'
informatie-record nr. 4 is 'WOORD'
De informatie-records nr. 1 en nr. 2 hebben de meeste 3-strings gemeen.



overeenkomst is tussen de vraag en het goede antwoord. Voor zover bekend gebruiken alle systemen voor information retrieval gehele woorden als basiseenheid waarmee wordt gewerkt. Men is er altijd stilzwijgend en ten onrechte vanuit gegaan dat het mechanisme onze taal als het ware moet leren 'begrijpen'. Er wordt verwacht dat het mechanisme opereert op 'begrippen' en die worden in de taal gedragen door woorden. Maar een mechanisme kan niet 'begrijpen', het kan alleen constateren dat een bepaalde combinatie van letters in het geheugen is opgeslagen. Woorden zijn noodzakelijk bij de invoer en de uitvoer, maar niet binnen in de computer.

Deze gedachtengang heeft er toe geleid dat in het TNO-programma kleinere eenheden dan woorden als basiseenheid worden gebruikt; iedere zin wordt automatisch opgesplitst in groepen van drie letters, de '3-strings'. Het woord FUZZIE bijvoorbeeld wordt door de computer omgezet in de basiseenheden FUZ, UZZ, ZZI, ZIE. Complete teksten worden op dezelfde manier opgesplitst. Voor het opbergen van teksten in het geheugen van de computer maakt men gebruik van een boomstructuur (zie illustratie). Iedere afzonderlijke publikatie wandelt als het ware langs één bepaalde tak in de boomstructuur en laat daarbij een spoor van 3-strings achter. Ook elke vraag wordt op deze manier omgezet in een informatiespoor van 3-strings.

Voor information retrieval vergelijkt de computer het informatiespoor van de vraag met de informatiesporen van de publikaties die in het geheugen zijn opgeslagen, waarbij die informatiesporen worden geselecteerd die het best met het informatiespoor van de vraag overeenkomen. De mate van overeenkomst – de 'importantiewaarde' – van het antwoord – wordt door de computer berekend en aan de gebruiker opgegeven, zodat die direct kan zien of vraag en antwoord redelijk goed met elkaar overeenkomen of niet.

In werkelijkheid zit het nieuwe programma ingewikkelder in elkaar dan hier is voorgesteld. Men werkt met behulp van een 'inverted file' en bij de berekeningen van de importantiewaarde wordt er niet alleen gelet op de syntactische overeenstemming, maar ook op de semantische – dus de inhoud. FUZZIE biedt een aantal voordelen, waarvan de eenvoud van bediening op de eerste plaats komt. De gebruiker stelt zijn vragen in gewone taal en hoeft niets te weten over de interne opbouw van het systeem. Wel moet hij weten welke taal – Nederlands, Engels, etcetera – er in het systeem wordt gehanteerd. Stelt hij een verkeerde vraag, dan blijkt dat direct uit het antwoord en uit de grootte van de importantiewaarde. Nieuwe gebruikers kunnen gemakkelijk met het systeem

spelen en zich in korte tijd de noodzakelijke routine eigen maken. Daarmee komt information retrieval binnen het bereik van allerlei instellingen, zoals middelgrote en kleinere bedrijven, waarvoor tot nu toe de kosten te hoog waren.

Het systeem biedt de mogelijkheid tot het opnemen van complete teksten in het geheugen van de computer. In de huidige versie is de capaciteit erg groot: er kunnen bijna tien miljard documenten mee worden bestreken. Mede hierdoor worden nieuwe terreinen van toepassing voor information retrieval opgelegd. Er zijn gebieden waar een gebruiker weinig of niets heeft aan titels en samenvattingen, maar altijd de volledige tekst tot zijn beschikking wil hebben. Een voor de hand liggend voorbeeld is de jurisprudentie, waar de precieze formulering van wetsartikelen, vonnissen en arresten van eminent belang is. De hoeveelheid jurisprudentie neemt hand over hand toe – ook in Europees verband – en er is al lang behoefte aan geautomatiseerde information retrieval. Met de huidige systemen bleek dit niet realiseerbaar, maar de verwachting is dat het met FUZZIE wel kan. Een Nederlandse uitgever van juridische naslagwerken onderzoekt samen met TNO de mogelijkheden hiervoor.

Een andere mogelijke toepassing is automatisch vertalen. Men kan teksten in één taal en de vertaling daarvan in andere talen volgens de sporenmethode in het geheugen opslaan. Met FUZZIE kan de computer dan een zin in één van de opgenomen talen omzetten in elk van de andere talen. Men mag niet verwachten dat deze automatische vertaling volmaakt is, de mens zal altijd moeten bijschaven.

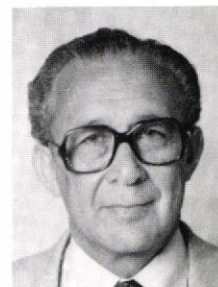
Wellicht kan FUZZIE een bijdrage leveren tot de oplossing van het probleem van de veelaligheid van de informatie-explosie. TNO is er zeker van dat er nog andere toepassingen gevonden zullen worden naarmate dit nieuwe programma voor information retrieval in bredere kringen bekend raakt.

Literatuur

T. de Heer, Quasi Comprehension of Natural Language Simulated by Means of Information Traces; Information Processing & Management, ter perse. T. de Heer, Een nogal slordige zoekstrategie, OPEN, Vol. 10, Sept. 1978.

Voor nadere informatie: Instituut TNO voor Wiskunde, Informatieverwerking en Statistiek, Drs. T. de Heer, Postbus 297, 2501 BD 's-Gravenhage, tel.: (070) 82 41 61

De Koninklijke/Shell Prijs 1978 is toegekend aan dr. F. J. Ritter en dr. ir. C. J. Persoons. Beiden zijn verbonden aan de werkgroep Biochemie van de Hoofdafdeling Technologie TNO; de heer Ritter als leider van de werkgroep en de heer Persoons als wetenschappelijk medewerker.



F. J. Ritter



C. J. Persoons

De TNO'ers kregen de prijs voor hun werk op het gebied van de insectenferomonen. Feromonen zijn stoffen die insecten gebruiken om aan soortgenoten informatie door te geven. Toepassing van deze stoffen maakt het wellicht mogelijk tot zeer gerichte en milieuvriendelijke bestrijding van insectenplagen te komen. In zijn toelichting op het juryrapport zei jurylid prof. dr. J. A. A. Ketelaar ondermeer: 'Beiden hebben op het gebied van het feromonenonderzoek enkele indrukwekkende prestaties geleverd. Het ophelderen van de structuur van het sexferomoon van de Amerikaanse kakkerlak is een prestatie die zowel nationaal als internationaal diep respect heeft afgedwongen, temeer daar andere bekende onderzoekers in de Verenigde Staten en Japan hierin na herhaaldelijk pogen niet waren geslaagd'.

De Koninklijke/Shell Prijs, in 1959 ingesteld, wordt uitgereikt als erkenning voor verricht wetenschappelijk onderzoek op het gebied van de exacte, de maatschappij- of de levenswetenschappen. De toekenning geschiedt overeenkomstig het advies van de Hollandsche Maatschappij der Wetenschappen te Haarlem.