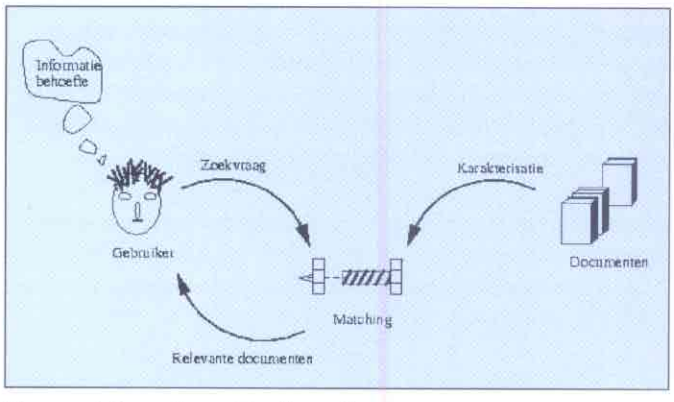
Bernd Wondergem, Patrick van Bommel en Theo van der Weide

Information retrieval: een uitdagend onderzoeksgebied

Recente wetenschappelijke ontwikkelingen in information retrieval (IR) worden in dit artikel door Wondergem, Van Bommel en Van der Weide aan de orde gesteld. De drie belangrijkste modellen voor IR worden besproken en vergeleken, alsmede het doel en de gebruikte technieken van een aantal IR-projecten in Nederland. Met tot besluit: waarom een perfect IR-systeem onmogelijk is.

HET DOEL VAN information retrieval (IR) is eindgebruikers eenvoudige en effectieve toegang te geven tot grote hoeveelheden veelal ongestructureerde informatie. Deze informatie was oorspronkelijk alleen tekstueel maar is tegenwoordig vaak multimediaal.
De gebruiker heeft een bepaalde informatiebehoefte, die vervuld kan worden door relevante documenten. Dit gebeurt in een aantal stappen. De inhoud van de aanwezige documenten wordt beschreven. Dit wordt karakteriseren of indexeren van documenten genoemd. Per document hoeft dit slechts eenmaal te gebeuren. Vervolgens formuleert de gebruiker zijn informatiebehoefte in een zoekvraag (*query*). Dan worden de zoekvraag en de karakteriseringen van de documenten vergeleken (*matching*). Documenten die een grote gelijkenis hebben met de zoekvraag, worden opgeleverd als relevante documenten en gepresenteerd aan de gebruiker.
Als reactie op de aangeboden documenten kan de gebruiker soms expliciet aan het systeem duidelijk maken welke (niet) relevant zijn. Dit heet *relevance feedback* en wordt door het systeem gebruikt om vervolgens betere resultaten (meer relevante, minder irrelevante documenten) aan te bieden. Dit gebeurt door de zoekvraag op basis van de feedback aan te passen. Het is natuurlijk van belang hoe automatisch die aanpassing plaatsvindt. In huidige IR-systemen gebeurt dit vaak (semi-)automatisch. De nieuwe zoekvraag levert weer een aantal documenten op, waarop de gebruiker opnieuw relevance feedback kan geven. Zo wordt de zoekvraag dus interactief en iteratief ge(her)formuleerd.



Figuur 1. Information retrieval-paradigma

Geschiedenis

Ondanks de inzet van Hugo de St. Caro begint geautomatiseerde ontsluiting van informatie pas na de Tweede Wereldoorlog. Vanaf het verschijnen in 1945 van Vannevar Bush's visionaire artikel "As we may think" (Bush, 1945) is veel theoretisch en praktisch werk verricht naar het opslaan en ontsluiten van informatie. Dr. Vannevar Bush (1890-1974), wetenschappelijk adviseur van president Roosevelt, kijkt de toekomst in en ziet onder andere Memex, een persoonlijke, geautomatiseerde uitbreiding op het menselijk geheugen. Memex ondersteunt associatieve selectie, adaptatie van gegevens aan persoonlijke wensen en combinatie van verschillende teksten.

De ontwikkeling van IR, weergegeven met een tijdsbalk in figuur 2, kan grofweg worden verdeeld in twee fasen.
Ontwikkeling van ideeën en technieken In de eerste fase, lopend van "Bush" tot halverwege de jaren zeventig, worden belangrijke technieken en ideeën ontwikkeld en uitgewerkt in de wetenschappelijke wereld. Hoewel Calvin Mooers de term "Information Retrieval" in 1950 introduceert (Mooers, 1950), kan het vakgebied in die tijd beter worden beschreven als "data retrieval". Dit omdat gewerkt werd met informatie *over* documenten, zoals auteur, titel en plaatscode, in plaats van de informatie *in* documenten. De inhoud van documenten wordt pas op grote schaal meegenomen vanaf het moment dat automatische indexering gebruikt wordt. Dit wordt in de tweede helft van de jaren vijftig bij IBM ontwikkeld. De ICSI-conferentie in Washington (1958) markeert de start van IR zoals we het nu kennen.
Een aantal jaren daarvoor zijn al maten geïntroduceerd om de effectiviteit van IR-systemen uit te drukken. *Recall*, dat

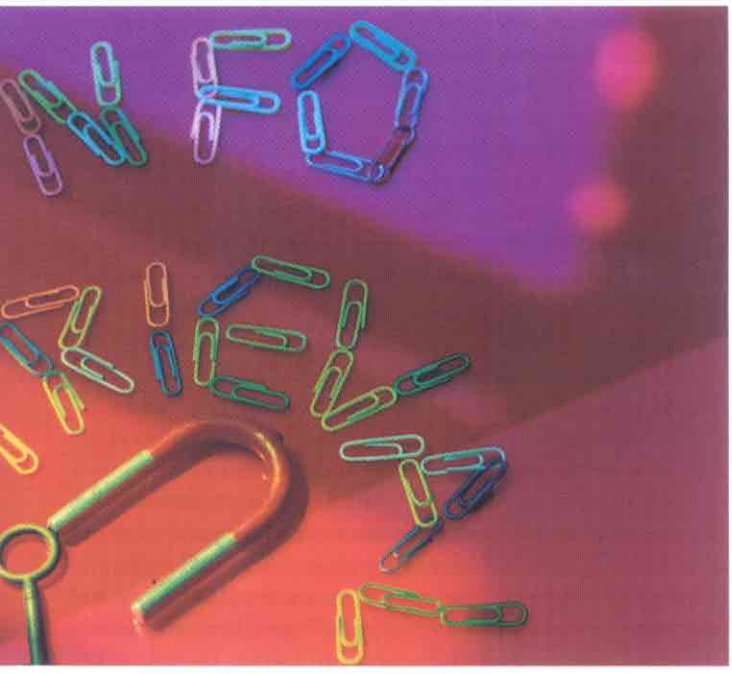
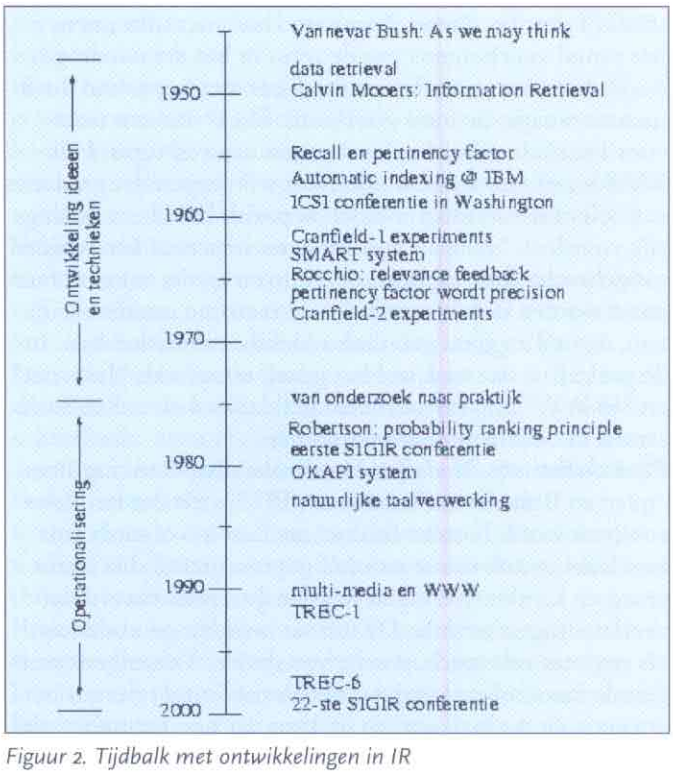


FOTO: EGON VEERE

aangeeft hoe goed het systeem de relevante documenten gevonden heeft, en de *pertinency factor*, duidend op hoe goed het systeem irrelevante documenten heeft weggelaten. Rond 1965 wordt de term pertinency factor vervangen door *precisie*, zoals het nu nog steeds wordt gebruikt.
De meeste IR-systemen uit de begintijd gebruiken Booleaanse operatoren (AND, OR en NOT) om zoekvragen te structureren. In het begin van de jaren zestig wordt het Vector Space Model geïntroduceerd. Dit model heeft een geometrisch in plaats van een logisch uitgangspunt. Dit wordt later uitgelegd. Het bekendste systeem dat gebaseerd is op dit model is het succesvolle SMART-systeem (Salton en McGill, 1983). In de jaren zestig worden door wijlen Gerard Salton en zijn groep uitgebreide experimenten gedaan met dit systeem. Rond dezelfde tijd beginnen de eerste statistische benaderingen van IR. Karen Sparck Jones en Steven Robertson vervullen hierin een belangrijke rol. Het zou echter tot 1977 duren voordat Robertson zijn befaamde *"Probability Ranking Principle"* zal poneren (Robertson, 1977). In een volgende paragraaf wordt beschreven hoe dit principe in IR wordt toegepast.
Ondertussen worden grootschalige experimenten met IR-systemen opgezet. De Cranfield-testen worden gezien als mijlpalen in het evalueren van IR-systemen. Daarbij wordt nog handmatige indexering toegepast. Als Rocchio in 1965 het idee van relevance feedback poneert, is de basis voor IR compleet.

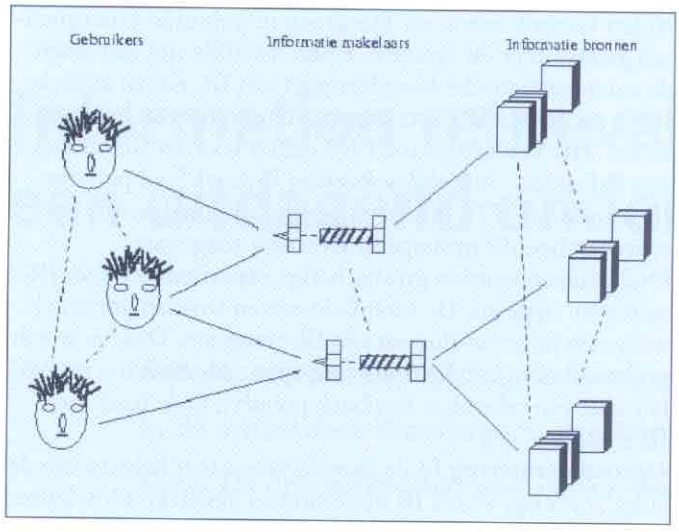


Figuur 2. Tijdbalk met ontwikkelingen in IR

Operationalisering In de tweede fase, vanaf halverwege de jaren zeventig, wordt IR operationeel gemaakt. Ook buiten de wetenschappelijke wereld wordt IR als een belangrijk vakgebied gezien. Hoewel afkomstig uit de bibliotheekwereld, wordt het aandeel informatica allengs groter. Het onderzoek naar IR resulteert in de eerste SIGIR (Special Interest Group on Information Retrieval) conferentie in 1978. Deze conferenties zullen uitgroeien tot de belangrijkste op het vakgebied. Een van de belangrijkste systemen die op basis van het Probability Ranking-principe van Robertson (1977) zijn ontwikkeld, is het OKAPI-systeem (begin jaren tachtig).
In de jaren tachtig worden tevens enkele nieuwe ideeën ontwikkeld. Een van de belangrijkste is het gebruik van natuurlijke taalverwerking: *natural language processing* (NLP). Het doel van NLP in IR betreft het makkelijker, duidelijker en contextafhankelijk formuleren van zoekvragen en een betere retrieval door normalisatie van structuur en betekenis. Twee belangrijke kwesties hierin zijn grammaticale analyse en lexicale analyse. Door capaciteitsproblemen (trage computers, weinig geheugen) en het relatieve succes van andere benaderingen, was dit type onderzoek eerder niet van de grond gekomen.
In de jaren negentig neemt het onderzoek naar IR een grote vlucht. Dit gebeurt mede door de introductie van het world wide web. Multimediale informatie, zoals geluid, beeld en video, begint deel uit te maken van documenten. Informatie wordt veelal gedistribueerd (in een netwerk) aangeboden en is qua inhoud en vorm heterogeen. Het IR-paradigma is toe aan vernieuwing. Een gedistribueerd IR-paradigma wordt beschreven in Wondergem (1998). Het bevat meerdere gebruikers en bronnen binnen een netwerkomgeving (zie figuur 3). Een belangrijke rol hierin is weggelegd voor informatiemakelaars, tussenpersonen tussen gebruikers en bronnen. In 1992 wordt de eerste TREC (Text REtrieval Conference) gehouden, een competitie tussen IR-systemen. Deze jaarlijks terugkerende test verwerft veel aanzien.

Modellen

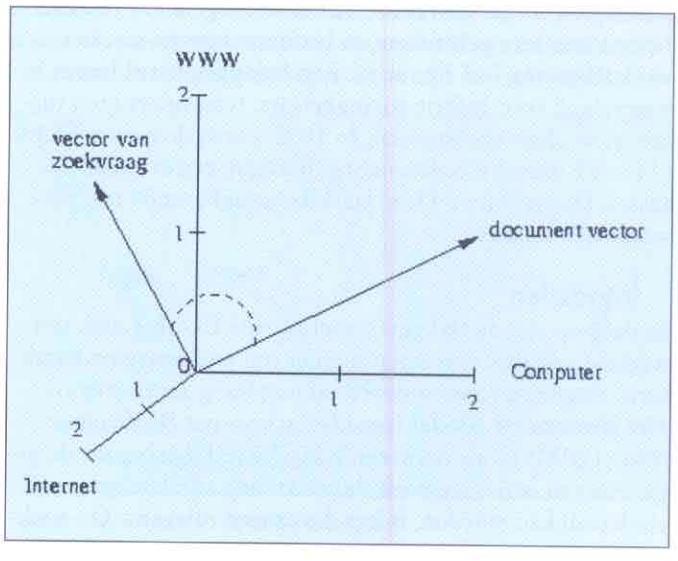
In de loop van de tijd zijn verschillende IR-modellen ontwikkeld, elk met een eigen manier om zoekvraag en karakteriseringen te representeren en matching uit te voeren.
Het Booleaanse Model Het idee achter het Booleaanse Model (BM) is: als door een logische redenering uit de gegevens van een document de zoekvraag afgeleid (geconcludeerd) kan worden, is het document relevant. De zoek-



Figuur 3. Nieuw IR-paradigma

vraag en de karakteriseringen van documenten worden hiertoe gerepresenteerd door logische expressies. Deze bestaan uit termen en Booleaanse operatoren. Karakteriseringen van documenten bevatten vaak alleen de AND operator. Als de zoekvraag bijvoorbeeld "conferentie AND (retrieval OR agents)" is en het document gekarakteriseerd is door "retrieval AND conferentie", kan de zoekvraag uit de karakterisering worden afgeleid en wordt het document relevant bevonden voor de zoekvraag. Afleidingsregels uit de logica worden gebruikt om de redenering te maken. Logische expressies en afleidingsregels kunnen zeer efficiënt worden geïmplementeerd door bitrijen (nullen en enen) en snelle operaties voor conjunctie (AND), disjunctie (OR) en negatie (NOT). Daarom is het BM vaak gebruikt voor praktische toepassingen. Voor gebruikers die precies weten wat ze zoeken en hoe ze het moeten formuleren is het BM zeker geschikt.
Het BM heeft ook nadelen. Zoekvragen zijn vaak moeilijk te formuleren, omdat het resultaat van complex geneste Booleaanse operatoren moeilijk te begrijpen is. Het vereist bijvoorbeeld enig puzzelwerk om vast te stellen dat een document met karakterisering "Internet AND WWW" relevant is voor de zoekvraag "WWW AND NOT(hardware AND NOT(computer OR Internet))". Daarnaast is het niet mogelijk om een ranking van documenten (in volgorde van relevantie) te maken, omdat niet met gewichten de mate van belangrijkheid per term kan worden aangegeven. Booleaanse afleidingen leveren binaire relevantiebepalingen op: relevant of niet-relevant. Daarom wordt het BM een *exact-match* model genoemd. Wel zijn niet-binaire afgeleiden van het BM ontwikkeld, maar die hebben een deel van de strikt logische interpretatie verloren.

Figuur 4. Hoekberekening in Vector Space Model



Vector Space Model Het Vector Space Model (VSM) heeft een geometrisch uitgangspunt. De zoekvraag en karakteriseringen worden gerepresenteerd door vectoren van termen. In figuur 4 zijn twee voorbeeldvectoren te zien in een driedimensionale ruimte. Beschouw bijvoorbeeld drie termen: "WWW", "Internet" en "computer". De assen geven het aantal voorkomens (gezwicht) van deze termen weer. Op deze manier wordt een driedimensionale informatieruimte opgespannen. Stel dat in een bepaald document de term "computer" twee keer voorkomt, "WWW" één keer en "Internet" ontbreekt. De inhoud van het document wordt dan gerepresenteerd door de vector (2,1,0). De vector van de zoekvraag in figuur 4 geeft aan dat "WWW" en "Internet" gewicht 2 hebben en dat "computer" niet in de zoekvraag voorkomt. De hoek tussen beide vectoren is een maat voor het verschil tussen beide vectoren.
Het VSM werkt in werkelijkheid met informatieruimten met een veel hogere dimensie. Ook dan wordt de gelijkenis tussen twee vectoren bepaald door hun verschil in richting. De grootte van de hoek tussen een vraag- en een documentvector is omgekeerd evenredig met de verwachte relevantie van het document voor de vraag. Er zijn sterke aanwijzingen dat het VSM voor grote full-text bestanden meestal beter presteert dan het Booleaanse model. De gewichten in de vectoren kunnen ook op andere manieren worden berekend. Een veel gebruikte manier is tf*idf (Term Frequency * Inverse Document Frequency): het aantal voorkomens van de term in het document gedeeld door een getal dat evenredig is met het aantal documenten waarin de term voorkomt. Dit is dus een maat voor het onderscheidend vermogen van een term. Het VSM is een *best-match* model, wat wil zeggen dat gradaties van relevantie kunnen worden bepaald. Dit is een belangrijk voordeel. Nadelen zijn dat geen structuur kan worden aangebracht door logische operatoren en dat aangenomen moet worden dat de termen in de vectoren onafhankelijk zijn, dat wil zeggen: geen inhoudelijke relatie hebben. In de praktijk is dat vaak wel het geval: termen als "Internet" en "WWW" zijn bijvoorbeeld gerelateerd en zullen vaak samen in een document voorkomen.

Probabilistische Modellen Het Probability Ranking Principle van Robertson (Robertson, 1977) stelt dat het beste resultaat wordt bereikt als documenten in volgorde van hun kans op relevantie worden gepresenteerd. De zoekvraag en karakteriseringen worden gerepresenteerd door verzamelingen termen. De termen worden gemodelleerd als zogenaamde stochastische variabelen. Vervolgens worden de kansen berekend dat een document d relevant is, gegeven de zoekvraag q, en de kans dat het document niet relevant is. Voor het berekenen van deze kansen worden

rekenregels uit de kansrekening gebruikt. Als de kans op relevantie groter is dan die op niet-relevantie, wordt het document aan de gebruiker gepresenteerd.
Volgens sommige onderzoekers presteert het probabilistische model beter dan het Booleaanse model maar minder goed dan het VSM. Een nadeel is dat in het probabilistische model, net als in het VSM, wordt aangenomen dat de termen onafhankelijk zijn. Bovendien is er geen methode om de initiële kansen en gewichten te bepalen als nog geen relevante documenten bekend zijn.

De eerste sporen

'De eerste sporen van systemen voor het ontsluiten van informatie gaan in ieder geval terug tot het jaar 1247. In dat jaar werden 500 monniken door Hugo de St. Caro in dienst genomen voor het maken van de eerste concordantie van de bijbel. Hugo's monnik-kracht ecclesiastisch data processing system werd vermoedelijk MK/EDPS-1 genaamd of mogelijk System 1 - de geschiedenis is niet duidelijk over dit punt.' (Naar Swanson, 1988).

Trends in onderzoek

Het wetenschappelijk onderzoek naar IR heeft in de jaren negentig een grote vlucht genomen. Belangrijke trends hierin bespreken we aan de hand van een aantal Nederlandse projecten. In Nederland wordt onderzoek naar IR onder andere verricht aan de universiteiten van Nijmegen, Twente, Utrecht en Tilburg en bij TNO.
Profile - Proaktief Informatiefilter Bij de subfaculteit informatica van de Katholieke Universiteit Nijmegen wordt toegepast onderzoek naar natuurlijke-taalverwerking en automatische classificatie gecombineerd met fundamenteel onderzoek naar hypermedia-modellen en krachtige vraagtalen.
Bij het Profile-project wordt een proaktief informatiefilter ontwikkeld. Interesses van gebruikers worden bewaard in profielen, zodat langetermijninformatiebehoeften kunnen worden behandeld. Deze profielen kunnen worden gebruikt om proaktief, zonder directe opdracht van de gebruiker, relevante documenten aan te bieden. Het Profile-project is een samenwerking van de subfaculteit informatica en het Nijmegen Institute for Cognition and Information (NICI). In het Profile-project wordt agent-technologie gebruikt als implementatie van het IR-paradigma van figuur 3.
Het Profile-systeem kent vier componenten. De eerste, *gebruikersmodellering*, richt zich op het verkrijgen van een beschrijving van de interesses van gebruikers. Ook situationele karakteristieken, zoals het beroep van de gebruiker en de talen die hij beheerst, worden gebruikt om alleen bruikbare informatie aan te bieden.
In de tweede component, de *parsering*, wordt de inhoud van documenten geanalyseerd. Door middel van normalisatie van frasen (zinsdelen) wordt de recall verhoogd. Lexicale normalisatie beeldt verschillende woorden met dezelfde betekenis op elkaar af. Semantische normalisatie achterhaalt bijvoorbeeld de bedoelde betekenis van woorden met meerdere betekenissen. Syntactische normalisatie beeldt frasen met dezelfde betekenis op elkaar af (bijvoorbeeld "de auto is rood" en "de rode auto"). Morfologische normalisatie brengt verschillende woordvormen bij elkaar (bijvoorbeeld huis en huizen).
In de derde component wordt de *matching* tussen zoekvraag en karakterisering onderzocht. Er is een nieuwe krachtige vraagtaal ontwikkeld, zogenaamde Booleaanse Index Expressies, met prettige eigenschappen als hoge expressiviteit, compactheid, begrijpelijkheid en doenlijkheid. Met doenlijkheid wordt bedoeld dat het IR-systeem in staat is de descriptoren te verwerken binnen bepaalde grenzen (bijvoorbeeld aangaande tijd of benodigde hulpbronnen).
De vierde component van het Profile-project draagt zorg voor de *presentatie* van relevante documenten aan de gebruiker. Aspecten uit de cognitieve ergonomie spelen een rol bij een overzichtelijke presentatie.
Twenty One Binnen de leerstoel taaltechnologie aan de Universiteit Twente wordt onderzoek gedaan naar het terugvinden van multimediale en multilinguale (meertalige) informatie. Een belangrijk multimediathema is het vinden van videofragmenten door middel van tekstuele informatie (uit ondertitels of spraakherkenning). Het belangrijkste multilinguale thema is zogenaamde 'cross-language information retrieval', waarbij gebruikers met bijvoorbeeld Nederlandse zoekvragen in verschillende talen tegelijk kunnen zoeken. Het systeem zorgt hierbij voor de automatische vertaling van zoekvragen en documenten.
Het Europese Twenty-One-project resulteerde in het eerste on line 'cross-language'-zoeksysteem in Europa. Bij de jaarlijkse TREC-benchmarkingconferentie voor IR-systemen bewees het Twenty-One-systeem in 1998 tot de beste systemen van de test te horen. Een demoversie van het Twenty-One-systeem is on line beschikbaar bij TNO-TPD.
De projecten PopEye en Olive zijn, eveneens door de EU gefinancierde, vervolgprojecten van Twenty-One, waarbij de nadruk ligt op het terugvinden van videofragmenten. Binnen het dit jaar op het Telematica Instituut gestarte project Druid ontwikkelt Twente sprekeronafhankelijke spraakherkenning voor het terugvinden van Nederlands gesproken informatie.
Hoewel bovengenoemde projecten gericht zijn op directe toepasbaarheid in het bedrijfsleven, is er ook plaats voor fundamenteel wetenschappelijk onderzoek. Een voorbeeld hiervan is het in Twente ontwikkelde "taalkundig gemotiveerde" kansmodel voor IR. Dit model, dat een wiskundige onderbouwing geeft van de al in de jaren zeventig geïntroduceerde tf*idf-weging van zoektermen, presteert uitzonderlijk goed in de TREC-competitie.
Condorcet In Twente wordt ook hard gewerkt aan domein-afhankelijke retrieval met geavanceerde linguïstische technieken. Het Condorcet-project, gefinancierd door STW, richt zich op het automatisch indexeren van wetenschappelijke teksten op het gebied van materiaalkunde en geneeskunde. Condorcet onderscheidt zich van andere IR-projecten door *concepten* in plaats van *termen* te genereren. Door het gebruik van gestructureerde concepten als "cures(aspirin, headache)" en "causes(aspirin, headache)" kunnen documenten subtieler worden geïndexeerd, hetgeen de precisie van het uiteindelijke zoeksysteem verhoogt.
Het toekennen van concepten aan een document gebeurt op basis van intensieve analyse met behulp van taaltechno-

logie en kennistechnologie. Het prototype indexeersysteem is ontwikkeld op basis van achthonderd documenten, en zal worden getest op een corpus van driehonderd documenten. De evaluatie van het project, dat afloopt in september 1999, is op het moment van schrijven nog in volle gang.

Mirror Het Mirror-project aan de Universiteit Twente bestudeert multimedia-databases. Daarbij bleek dat een multimediadatabase niet zomaar een database met multimedia is. Probleem is onder meer hoe de gebruiker moet weergeven waarnaar hij op zoek is. De theoretische basis voor een image retrieval-systeem, gebouwd op basis van het prototype Mirror DBMS, komt uit de hoek van cognitiewetenschappen. Mirror gebruikt hiervoor technieken uit de IR die zijn gegeneraliseerd naar de situatie in multimedia.

Deze IR-technieken kunnen niet eenvoudig geïntegreerd worden in een relationeel databasesysteem. Object-georiënteerde databases zijn hiervoor ook niet geschikt, omdat die niet goed schaalbaar zijn. Daarom besteedt het Mirror-project veel aandacht aan het inbedden van de multimedia retrieval-technologie in het databasesysteem. Het resultaat is een DBMS dat basistechnieken aanlevert, waarmee een applicatieontwikkelaar geavanceerde systemen kan ontwikkelen. Een voorbeeld hiervan is de image retrieval demo voor de International Conference on Very Large Databases (VLDB 1999).

UPLIFT UPLIFT is een onderzoeksproject van het Utrechts Instituut voor Linguïstiek OTS, in samenwerking met TNO-TPD. Er wordt onderzocht of het mogelijk is de prestatie van standaard retrieval-systemen te verbeteren door toevoegen van modules die gebruikmaken van taalkundige kennis.

In een eerdere fase van het project is onderzoek gedaan naar het automatisch uitbreiden van zoekvragen met taalkundig verwante woorden (*word-stemming*) ter verbetering van de recall, en het indexeren van teksten met grotere eenheden dan losse woorden (*phrase indexing*) ter verbetering van de precisie. Dit onderzoek heeft zich voornamelijk geconcentreerd op Nederlandstalige teksten. Op dit moment wordt onderzoek gedaan naar het vertalen van zoekvragen of documenten ten behoeve van meertalige retrieval. De verwachte einddatum van het project is eind 2000.

TNO-TPD De afdeling Multimedia Technologie van TNO-TPD bundelt de expertise die binnen TNO is opgebouwd op de gebieden van Taaltechnologie en Multimedia. Het accent van het onderzoek ligt op ontsluiting van ongestructureerde informatie in de vorm van tekst, audio of video en de koppeling met (gestructureerde) kennissystemen.

Er wordt veel samengewerkt met andere kenniscentra. Zo is in samenwerking met de Universiteit Twente en het Duitse DFKI, in de al eerder beschreven EU-projecten TwentyOne, Olive en PopEye, een multimedia-ontsluitingsarchitectuur ontwikkeld die gebaseerd is op een com-

binatie van taalkundige analyse, fuzzy matching en state-of-the-art probabilistische modellen. Momenteel wordt onder andere gewerkt aan cross-language retrieval, automatische toekenning van thesaurus- of rubriekstermen (zie het artikel over ADJUST elders in dit blad), automatische generatie van thesauri en gepersonaliseerde informatievoorziening. Na de eerder vermelde succesvolle deelname met TwentyOne aan TREC 1998, wordt in 1999 voor de derde keer deelgenomen, in de categorieën ad-hoc zoeken, cross-language, filtering en zoeken in audiobestanden. Het onderzoek wordt zo snel mogelijk vertaald in software van productkwaliteit.

URL's van genoemde projecten, congressen e.d.	
BCS-IRSG	http://irsg.eu.org/
Condorcet	http://www.cs.utwente.nl/condorcet/
DORO	http://www.cs.kun.nl/doro/
INN	http://www.sci.kun.nl/infstud/~markuden/
Mirror	http://wwwis.cs.utwente.nl:8080/~arjen/mmdb.html
Mirror-demo	http://wwwis.cs.utwente.nl:8080/~arjen/mmdb/abstracts.html#vldb99
Olive	http://twentyone.tpd.tno.nl/olive/
Popeye	http://twentyone.tpd.tno.nl/popeye/
Profile	http://hwr.nici.kun.nl/~profile/index.html
SIGIR	http://www.acm.org/sigir/
TNO-TPD	http://www.tpd.tno.nl/TPD/smartsite.html
TREC	http://trec.nist.gov/
Twenty-One	http://parlevink.cs.utwente.nl/Projects/twentyone.html
Twenty-One demo	http://twentyone.tpd.tno.nl/21demomooi
Uplift	http://www-uilots.let.ruu.nl/~uplift/

Document Vector Model Op de Katholieke Universiteit Brabant wordt onderzoek gedaan naar de traditionele indeling van IR in een aantal los van elkaar staande modellen. Deze indeling vertoont weinig samenhang. Het is gewoonte geworden om te spreken over het Booleaanse model, het vectorspace-model of het probabilistische model, zonder zich rekenschap te geven van een algemeen model waarop men al deze varianten zou kunnen baseren. Men probeert in deze leemte te voorzien door een nieuw model te introduceren dat een aantal van de bestaande modellen in zich verenigt, of er tenminste de basis voor vormt: het documentvector-model.

Conferenties en groepen De belangrijkste gebeurtenis in de wetenschappelijke IR-wereld is de jaarlijkse ACM SIGIR-conferentie. Deze sinds 1978 gehouden conferentie biedt via haar proceedings een goed beeld van de ontwikkelingen en trends in IR. Daarnaast is er het jaarlijkse colloquium van de IR-groep van de British Computer Society, het BCS-IRSG-colloquium. Daar presenteren voornamelijk jonge onderzoekers hun plannen en resultaten.

Onmogelijkheden

Al het onderzoek ten spijt is het een utopie te denken dat ooit een perfect IR-systeem gemaakt zal worden. De reden hiervoor ligt in een aantal fundamentele onmogelijkheden (zie bijvoorbeeld Swanson, 1988):

- De informatiebehoefte kan niet volledig worden uitgedrukt in een zoekvraag. Een eerste oorzaak hiervan is dat niet de gehele context van de informatiebehoefte kan worden meegenomen. De context bestaat uit ontelbare aspecten, zoals de kennis en achtergrond van de zoeker. Een tweede oorzaak is dat bij aanvang iets ontbrekends of onbekends wordt gezocht. De zoekvraag kan niet precies worden geformuleerd als het ontbrekende nog niet is gevonden.
- De relevantie van een document kan niet los gezien worden van die van andere documenten. Als bijvoorbeeld al documenten over een bepaald onderwerp zijn opgeleverd, zal de gebruiker deze en sterk gelijkende niet (nogmaals) willen ontvangen. Dit is de kern van een in Nijmegen ontwikkeld model voor IR: het *incremental satisfaction model* (Weide et al., 1998).
- Het is onmogelijk te controleren of het zoekresultaat alle relevante documenten bevat. Het is namelijk niet mogelijk alle documenten te bekijken. Het algemenere achterliggende idee is dat empirische wetenschap kan worden weerlegd maar niet geverifieerd.
- Het is mogelijk subtiele relevantiebepalingen te maken. Ook is het mogelijk erg effectieve mechanische zoekmethoden te maken. Dit gaat alleen niet samen.

Deze postulaten worden natuurlijk gezien als grote uitdagingen toch zo goed mogelijke IR-systemen te ontwikkelen.

De auteurs danken de volgende personen voor hun bijdrage: Djoerd Hiemstra (Twenty One), Arjen de Vries (MiRRoR), Erik Oltmans (Condorcet), Renee Pohlmann (UPLIFT), Wessel Kraaij (TNO-TPD) en Hans Paijmans (KUB).

Referenties

Bush, V. (1945). "As we may think". In: *The Atlantic Monthly*, Vol. 176, No. 1, p. 101-108. (www.theatlantic.com/unbound/flashbks/computer/bushf.htm)

Mooers, C.N. (1950). "The Theory of Digital Handling of Non-Numerical Information and Its Implications to Machine Economics". In: *Technical Bulleting* No. 48. Cambridge, MA: Zator Co.

Robertson, S.E. (1977). "The probability ranking principle in IR". In: *Journal of Documentation*, Vol. 33, No. 4, p. 294-304.

Salton, G. en McGill, M.J. (1983) *"The SMART and SIRE Experimental Retrieval Systems"*. p. 118-155, New York, McGraw-Hill.

Swanson, D.R. (1988). "Historical Note: Information Retrieval and the Future of an Illusion". In: *Journal of the American Society for Information Science*, Vol. 39, p. 92-98.

Weide, Th.P. van der, Huibers, T.W.C. en Bommel, P. van (1998). "The Incremental Searcher Satisfaction Model for Information Retrieval". In: *The Computer Journal*, Vol. 41, No. 5, p. 311-318.

Wondergem, B.C.M., Bommel, P.van en Weide, Th.P. van der (1998). "Cumulative Duality in Designing Information Brokers". In: Quirchmayr, G., Schweighofer, E. and Bench-Capon, T.J.M., editors, Proceedings of the 9th International Conference on Database and Expert Systems Applications, DEXA'98, p. 125-134, Vienna, Austria (www.cs.kun.nl/~bernd/abstract_duality.html).

Drs. Bernd Wondergem, dr. Patrick van Bommel en dr.ir. Theo van der Weide zijn verbonden aan de IRIS-groep van de subfaculteit informatica van de Katholieke Universiteit Nijmegen.