Ontwikkelingen in information retrieval

Informatie gezocht

De reden dat we information retrieval-systemen gebruiken is dat we informatie of kennis terug willen kunnen vinden. Het heeft immers geen zin om moeite te doen informatie op een of andere ingenieuze manier op te slaan als we verwachten deze niet meer nodig te hebben. Een open deur? Natuurlijk! Maar het blijft noodzakelijk - en niet in het minst uit economische overwegingen - om vast te stellen waarom en hoelang we bepaalde informatie willen bewaren. De consequentie van zo'n vaststelling is dat we de betreffende informatie terugvindbaar moeten maken, moeten ontsluiten. Met information retrieval bijvoorbeeld.

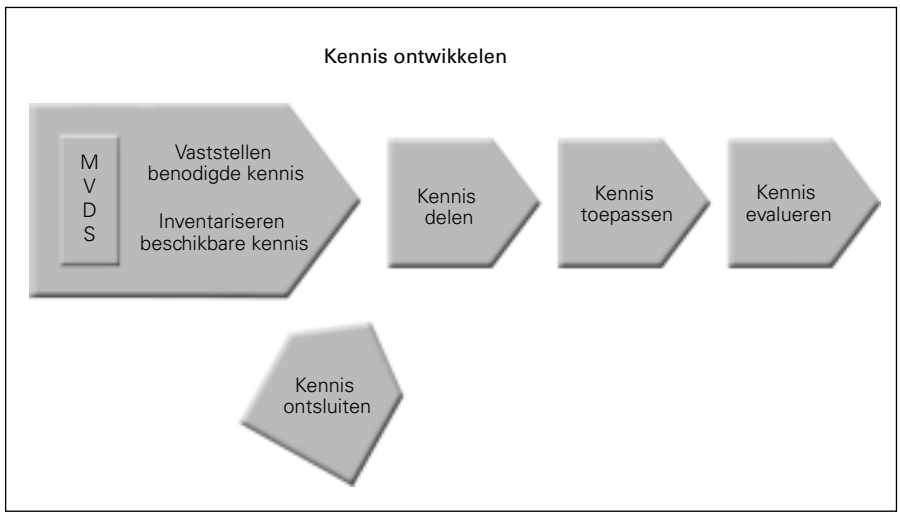
Information retrieval (IR) houdt in brede zin in het opslaan, ontsluiten en terugvinden van informatie. Als we volledig willen zijn is de term dan ook eigenlijk 'information storage and retrieval'. In de praktijk spreken we echter vaak van een IR-systeem. Information retrieval en meer specifiek het ontsluiten kunnen we plaatsen in de kenniswaardeketen van Weggeman (zie afbeelding 1). Termen die nauw aan IR verwant zijn, zijn document retrieval, concept retrieval, knowledge retrieval en fuzzy retrieval. IR vormt een wezenlijk onderdeel in enterprise contentmanagement-, documentmanagement- en documentflowmanagementsystemen.

In dit artikel bespreken we de belangrijkste elementen van information retrieval, namelijk de informatie zelf, de gebruiker en de menselijke en technische intermediairs. Als ontsluitingsmechanismen passeren de conventionele woordsystemen en classificatiesystemen de revue, alsmede de recent populair geworden taxonomieën. Ook geavanceerde software zoals intelligent agents, neurale netwerken en adaptieve hypermedia komen aan bod (zie kader 'De nieuwe IR-systemen').

Elementen van een IR-systeem

De basiselementen van een IR-systeem zijn de informatie zelf, de gebruiker en de intermediaire functie, die zowel een personeel als een technisch karakter kan hebben.

Informatie kunnen we positioneren in de reeks 'gegevens - informatie - kennis' en kan dan worden geformuleerd als 'gegevens in context'. Maar informatie wordt ook wel expliciete kennis genoemd (kennis die is vastgelegd in documenten). Vanwege het overzichtskarakter van dit artikel is het niet relevant het verschil tussen met name kennis en informatie te benadrukken. Wel is het nuttig om stil te staan bij de vraag hoe lang informatie nuttig bruikbaar blijft. Een belangrijk criterium daarvoor is de halfwaardetijd, een begrip dat ontleend is aan de kernfysica. De halfwaardetijd geeft aan hoe snel informatie veroudert. Een handige vuistregel om een indruk te krijgen van de halfwaardetijd is de snelheid van de ontwikkelingen in een bepaald vakgebied in te schatten en als richtlijn te gebruiken. Gaan de ontwikkelingen

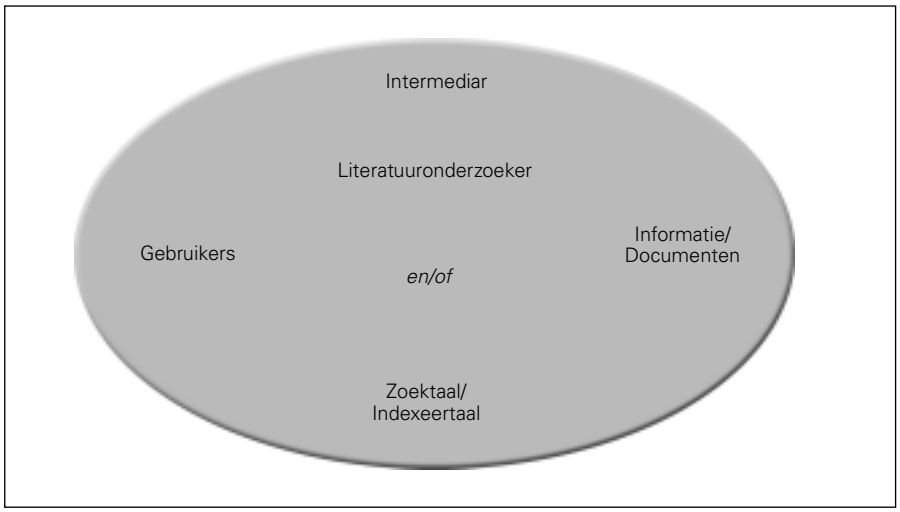


Afbeelding 1. De positie van ontsluiting in de kenniswaardeketen.

snel, dan is er grosso modo sprake van een korte halfwaardetijd. De ontwikkelingen halen elkaar dan links en rechts in, waardoor het belang van bestaande publicaties snel veroudert. Reden genoeg om na te gaan hoe effectief en efficiënt het is om dergelijke informatie te ontsluiten.

De *gebruiker* wil de juiste informatie en zeker niet teveel, op het juiste tijdstip en op de juiste plaats. Afhankelijk van het kennisniveau van de gebruiker en het karakter van de gewenste informatie is een breed scala van zoekstrategieën te formuleren om aan bovenstaande eisen van de gebruiker te voldoen. Dit is één van de redenen waarom ook bij IR-systemen het op het individu gerichte personaliseren en het op groepen gerichte 'mass customisation' van informatie de laatste tijd zo'n aandacht krijgt. Bij IR-systemen spreken we dan over attenderingsdiensten, Selective Dissemination of Information (SDI) of Current Awareness.

Vanwege het complexe karakter van zoektalen zal tussen de (eind-)gebruiker van de informatie en de informatie zelf dikwijls een *intermediair* functioneren (zie ook afbeelding 2). Deze intermediair kan een persoon zijn zoals een literatuuronderzoeker, documentalist of informatiespecialist, maar kan ook op een technische wijze worden ingevuld en krijgt dan gestalte in bijvoorbeeld een zoek- of indexeertaal zoals trefwoordenlijsten, thesauri, classificatieschema's en taxonomieën. Overigens zullen de (eind-)gebruiker en de literatuuronderzoeker c.s. beide gebruikmaken van deze technische middelen, alleen het niveau waarop zal verschillen. Eindgebruikers zijn niet altijd gecharmeerd van toch soms behoorlijk complexe classificatieschema's of thesauri. De mogelijkheden strekken zich uit van een trefwoordenlijst op papier tot actieve thesauri of semantische netwerken, die aan de hand van een eerste vraagformulering suggesties doen voor het weglaten van irrelevante zoektermen of toevoegen van aanvullende relevante zoektermen (de 'relevance feedback').



Afbeelding 2. De relaties tussen de wezenlijke elementen van een IR-systeem.

Het onderzoeken of een idee dat op een researchafdeling is geboren mogelijk al door octrooien van anderen is beschermd, zal altijd door een gespecialiseerde literatuuronderzoeker worden uitgevoerd en nooit door een researchmedewerker! De professionaliteit van de literatuuronderzoeker is hier dikwijls een kritische succesfactor. De ontwikkelingen op het gebied van natuurlijke taalinterfaces en spraakherkenning kunnen er de oorzaak van zijn dat deze technieken op langere termijn een groot deel van deze intermediaire functie over zullen nemen.

Ontsluitingsmechanismen

Een wezenlijke tweedeling in ontsluitingsmechanismen is te maken tussen woordsystemen en classificatiesystemen. Ook kunnen we bij ontsluitingsmechanismen denken aan meer recent populair geworden technieken als taxonomieën en ontologieën. Bij taxonomieën combineert men elementen uit classificatie- en thesaurusbenaderingen om tot betere retrievalopbrengsten te komen. Een ontologie is een soort metabeschrijving van een taxonomie, classificatie of semantisch netwerk en is nuttig in het theoretische kader van het denken over en analyseren van taxonomieën. Vanwege het theoretische karakter worden ontologieën in dit artikel niet verder besproken.

Woordsystemen worden bijna altijd opgebouwd vanuit een bottom-upbenadering: men verzamelt begrippen in de vorm van trefwoorden, termen of concepten, waarna men onderlinge verwijzingen aanbrengt. In de eenvoudige uitvoeringen van trefwoordenlijsten zijn dit de 'zie-' en 'zie ook-'aanduidingen. In thesauri vinden we deze relaties tussen begrippen, aangevuld met onder andere 'broader term' en 'narrower term'. Dit zijn begrippen waarmee men respectievelijk globalere en engere termen definieert die in samenhang met elkaar structuren opleveren die uitmonden in semantische netwerken. Een zogenoemde actieve thesaurus is bijvoorbeeld in staat om ons te attenderen op globalere of engere begrippen wanneer we een zoekvraag formuleren, zodat we die eventueel in de vraagformulering op kunnen nemen. De IR-systemen die op basis van woordsystemen

De nieuwe IR-systemen

Momenteel is een grote verscheidenheid aan IR-systemen verkrijgbaar. Het is dus belangrijk om na te gaan welke functionaliteit gewenst is en een degelijk pakketselectie-onderzoek te doen alvorens tot aanschaf van software over te gaan. In dit kader komen drie groepen IR-systemen aan bod: de intelligent agents, neurale netwerken en adaptieve hypermedia (deze laatste vooral vanwege hun potentieel voor multimediale IR-systemen). Deze driedeling is wel arbitrair, omdat steeds meer onderliggende functies in meer dan een groep voorkomen.

Intelligent agents worden de laatste jaren in groten getale aangeboden. De voor IR-systemen relevante agents of 'bots' staan bekend als 'knowledge bots' of 'search bots'. In mindere mate zijn 'news bots' en 'dataminig bots' interessant. Hun werking is gebaseerd op een breed scala van theoretische retrievalmodellen. Uit de bijgeleverde documentatie wordt meestal niet duidelijk volgens welke principes de agent werkt. Agents die een redelijk uitgebreide functionaliteit claimen, gebruiken meestal combinaties van die modellen. Eisen die men aan een intelligent agent kan stellen zijn: reactief, autonoom, doelgeoriënteerd, adaptief, zelflerend, communicatief en mobiel.

De werking van *neurale netwerken* is gebaseerd op de werking van onze hersenen. Neurale netwerken simuleren de manier waarop onze hersenen informatie verwerken, waardoor ze leren over bepaalde kennisdomeinen. Een standaard neuraal netwerk is opgebouwd uit knooppunten, die invoer ontvangen en uitvoer leveren. De uitvoer van de één is de invoer voor de ander. Op deze manier creëren de impulsen netwerken. Per knooppunt kan sprake zijn van één of meerdere invoersignalen, maar altijd slechts van één uitvoersignaal. Het uitvoersignaal kan wel aan meerdere knooppunten als invoersignaal worden aangeboden. Afhankelijk van wat er in het knooppunt wordt 'berekend' kunnen er vraag- en antwoordpatronen ontstaan die bewaard blijven in het systeem en later weer gebruikt zullen worden wanneer soortgelijke vragen aan het systeem worden gesteld. Door wijzigingen in de vraagstellingen zullen ook de patronen wijzigen en daarmee de antwoorden. Daarmee voldoen neurale netwerken aan de criteria zelflerend en adaptief.

In het algemeen kunnen we stellen dat *adaptieve hypermedia* worden ontwikkeld om het leren te vergemakkelijken. Daarom ontwerpt men ze met zelflerende kenmerken. In literatuur over intelligent agents presenteert men (research-)producten als adaptieve hypermedia, waarbij dit zelflerende aspect een belangrijke rol speelt. Het hypermedia-element uit de naamgeving van deze systemen heeft betrekking op de toepassingen in het vakgebied van de multimedia. Patroonherkenning, beeldherkenning en stemherkenning zijn belangrijke onderzoeksgebieden. Het aanbod van commercieel beschikbare software groeit snel [Meerman].

werken, bieden de mogelijkheid om vragen te definiëren met behulp van de Booleaanse operatoren 'and', 'or' en 'not' en met zogenaamde nabijheidsfactoren. In IR-systemen komen deze mogelijkheden in een grote diversiteit voor [Foskett].

Classificatiesystemen ontwerpt men in principe vanuit een top-down-benadering: beginnend bij een overkoepelend concept worden onderliggende begrippen steeds verder gedetailleerd, zodat een boomstructuur ontstaat van generieke naar specifieke deelconcepten. De zo ontstane hiërarchische indeling heeft in zijn oorspronkelijke vorm een erg statisch karakter. Voor de huidige eisen die men aan IR-systemen stelt is dit veelal onvoldoende. De huidige startpagina-achtige websites zijn hiervan een voorbeeld. Zij voldoen slechts in de situatie dat we ons globaal willen oriënteren over een onderwerp. Moeten we gespecialiseerder zoeken, dan hebben we toch echt flexibeler zoekmechanismen nodig (en vaak ook meer inhoudelijk gespecialiseerde databases). Deze vinden we onder andere in indexeersystemen die zijn gebaseerd op complexere classificatiebenaderingen, met als belangrijkste de facetclassificatie. Deze classificatietechniek, uitgevonden door Ranganathan in de dertiger jaren, is weliswaar behoorlijk oud maar mag zich verheugen in hernieuwde aandacht, omdat onder andere de objectgeoriënteerde softwareontwikkelingen het mogelijk maken om het klasseconcept in al zijn facetten uit te bouwen naar IR-toepassingen, in casu facetclassificatiestructuren te ontwerpen. Voorbeelden van gerealiseerde IR-systemen gebaseerd op facetclassificaties zijn Prieto-Diaz en recentelijk onder andere Bailey en Potter [Prieto-Diaz, Bailey, Potter]. Een gedegen inleiding voor het ontwerpen van classificatie-

schema's is te vinden in Foskett en voor classificatiemodellen ten behoeve van kennissystemen in Stefik [Foskett, Stefik].

Ook de aandacht voor het ontwerpen van technisch georiënteerde classificatieschema's, zoals productclassificaties, doet weer opgang. De business-to-consumer-ontwikkelingen van internet maken dat producenten opnieuw moeten nadenken over manieren waarop consumenten zo gemakkelijk mogelijk de gewenste producten kunnen vinden. De producten die de klassieke postorderbedrijven steevast via papieren catalogi presenteerden, moeten nu ook via het scherm zo optimaal mogelijk terugvindbaar worden gemaakt. Hierover nadenken vereist een wezenlijk andere benadering en zal altijd resulteren in een fundamenteel andere opzet van de bijbehorende zoeksystemen. Het combineren van het beste uit de classificatieschema's en uit de hyperlinktechniek zien we om ons heen overal ontstaan [Horner, Chaffy].

Taxonomieën

Onder het maken van een taxonomie verstaan we het ontwerpen van een structuur en het benoemen van informatie-items om de locatie van relevante informatie aan te geven. In een internetomgeving betekent dit dat men een breed scala aan metadata moet benoemen, dat regelt dat de informatie systematisch kan worden gemanaged en bewerkt. De mogelijkheden die XML biedt zijn hierbij van groot belang. Startpagina's zijn een hele eenvoudige vorm van een taxonomie (denk bijvoorbeeld aan Yahoo en startpagina.nl). We kunnen met zo'n pagina de overvloed aan informatie enigszins in goede banen leiden en een begin maken met het structureren van de chaos in het 'information overload'-tijdperk. Zeker zo belangrijk is de toepassing in corporate organisaties, wanneer het erom gaat terminologie te standaardiseren. Vooral voor organisaties die complexe terminologie gebruiken is het van wezenlijk belang om die terminologie te structureren en op elkaar af te stemmen (denk aan synoniemen en homoniemen). Dit geldt voor zowel intranetten als extranetten als Communities of Practice [King]. Bij het ontwerpen en bouwen van taxonomieën maakt men altijd intensief gebruik van al bestaande thesauri en classificatieschema's uit het betreffende vakgebied. Op vakgebieden als onder andere chemie, medische zorg, wetgeving, ICT en militaire wetenschappen zijn taxonomieën ontwikkeld. Deze worden zelfs commercieel aangeboden [Snark].

Nadelen die aan taxonomieën kleven zijn onder andere dat het erg arbeidsintensief is om ze te maken en te onderhouden, dat er niet één taxonomie perfect is en dat ze in wat complexere vorm door eindgebruikers als moeilijk in gebruik worden ervaren. Om het nadeel van het arbeidsintensieve karakter te minimaliseren wordt driftig gewerkt aan manieren om taxonomieën te maken door de informatie automatisch te categoriseren met behulp van zogenaamde clusteringtechnieken [Gilchrist, Salton].

Verbindende schakel

De information retrieval-component is een kritieke succesfactor in de nieuwe, complexe generaties informatiesystemen, zoals portals en enterprise contentmanagementsystemen. Deze IR-component zal altijd een ontsluitingsmechanisme bevatten in de vorm van woordsystemen, classificatiesystemen, taxonomieën of combinaties van deze drie. Ze zijn en blijven de verbindende schakel tussen de informatie enerzijds en de gebruikers van die informatie anderzijds. Of zo'n IR-component dan een intelligent agent, een neuraal netwerk of een adatief hypermediasysteem wordt genoemd is van secundair belang [Szuprowicz, White]. Het primaire criterium moet de gewenste functionaliteit zijn.

Leo Meerman

Leo Meerman is directeur Celt Consultancy (lmeerman@celt.nl) en houdt zich bezig met kennis- en documentmanagement.

Literatuur

- Bailey, Samantha*, Developing a taxonomies and information architecture strategy.
- Chaffy, Dave, E-business and e-commerce management : strategy, implementation and practice, Harlow : Pearson, 2002, ISBN 0273 65188 9.
- Foskett, A.C., The subject approach to information, London: Bingley, 1994.
- Gilchrist, Alan*, Mastering the balancing act of taxonomy design considerations.
- Horner, David, Frameworks for technology analysis and classification, In: Journal of information science 18 (1992), blz.57-68.
- King, Cathy*, Taxonomies -Your content management problem solved?
- Meerman, Leo, Intelligent knowledge agents, Lezing op de conferentie Retrieval Day : Over de ontsluiting van informatie. - Rotterdam: Erasmus Universiteit, november 2000.
- Potter, Keith*, Creating, implementing and assessing the effectiveness of a systematic taxonomy.
- Prieto-Diaz, Implementatie van meervoudige classificatie voor het hergebruik van programmatuur, In: Informatie jg. 35, nr.6, blz.403-412.
- Salton, Gerald, Introduction to Modern information retrieval. - Singapore-: - McGrawHill, 1993.
- Snark, The Taxonomy : Creating Knowledge from Chaos, Gevonden: 18 april 2001.
- Stefik, Mark, Introduction to knowledge systems : Classification. - San Francisco, Morgan Kaufmann Publishers, 1995, pp. 541-607, ISBN 1 55860 166 X.
- Szuprowicz, Bohdan, Implementing enterprise portals: integration strategies for intranet, extranet and internet resources. - Charleston : Computer Technology Research Corporation, 2000, ISBN 1 56607 080 5.
- White, Martin, Enterprise information portals, In: The Electronic Library Vol.18 (2000), nr. 5, pp.354-62.

De met * gemerkte verwijzingen zijn lezingen die op de conferentie Knowledge managed with taxonomies zijn gepresenteerd (Londen, 22-24 januari 2002).