Jan Mark Pleijsant, Harriët Hangelbroek en Jan van der Starre

Information retrieval in Luilekkerland: slimme zoeksystemen

Zoeksystemen worden steeds gebruikersvriendelijker en toegankelijker, en bovendien steeds 'slimmer'. Informatieprofessionals lijken hierdoor brodeloos te worden, en voor de professioneel uitgegeven en gestructureerde contentbronnen lijkt het einde nabij. Of toch niet? Pleijsant, Hangelbroek en Van der Starre schetsen de mogelijkheden van geavanceerde zoeksystemen. Ze gaan daarbij in op de effecten die deze systemen hebben op zowel de rol van de informatieprofessional als de gestructureerde contentbronnen.

Wat is 'slim' zoeken? Een ideaal 'slim' zoeksysteem is een systeem dat *begrijpt* naar welke informatie de gebruiker op zoek is en tegelijkertijd de contentbronnen zelf, zowel wat betreft inhoud als opbouw, *begrijpt*. Een 'slim' zoeksysteem biedt optimaal gebruiksgemak tegen een minimum aan inspanning.
Een ideaal zoeksysteem geeft optimale, exact bij de zoekvraag passende resultaten op basis van alle beschikbare contentbronnen, zowel binnen de organisatie als daarbuiten, hetzij gestructureerd, hetzij ongestructureerd, ongeacht de vorm van de content (tekst, afbeeldingen, beeld, geluid, digitale of ingescande documenten, boeken en tijdschriften).

Waar in de traditionele situatie vrijwel elke bron een eigen zoeksysteem of informatiemakelaar vereist, volstaat in de ideale situatie één enkel 'slim' informatieportaal. Dit zal voor vele organisaties efficiencywinst betekenen, door een hogere mate van effectiviteit van het vindproces en de kwaliteit van de beantwoorde informatievraag.

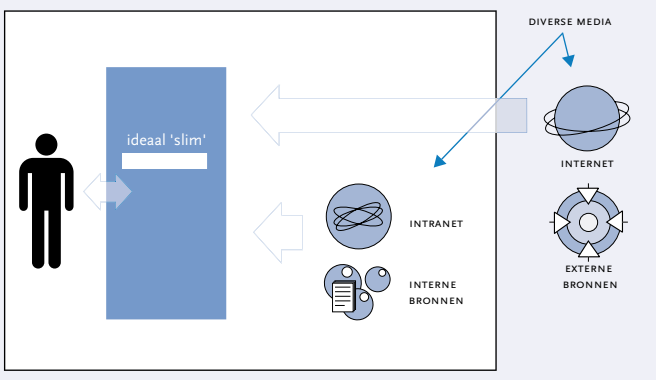
Geavanceerd zoeken: de huidige mogelijkheden

Hebben we deze ideale situatie al bereikt? Laten we eens kijken naar de mogelijkheden van hedendaagse, geavanceerde zoeksystemen. Daarbij kunnen drie deelgebieden worden onderscheiden: bronnen, systeem en gebruikersinterface.

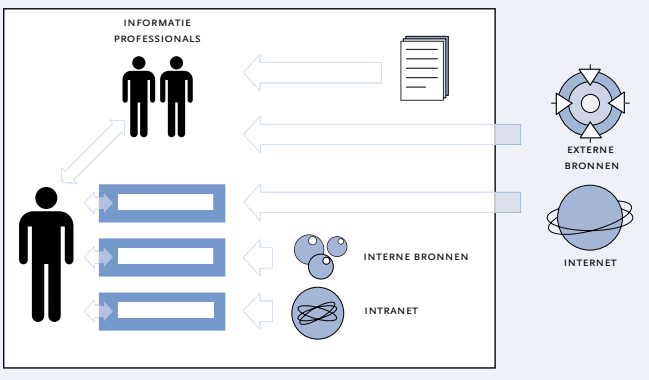
Bronnen

Met behulp van geavanceerde zoeksystemen zijn organisaties in staat om vele verschillende soorten bronnen (zoals een Document Management Systeem, internet en intranet) via één systeem of informatieportaal te ontsluiten. Dit betekent, dat via één zoekvraag vrijwel alle aangesloten digitale bronnen gelijktijdig bevraagd kunnen worden. Papieren bronnen zullen echter nog overwegend op de traditionele manier worden ontsloten via de bibliotheek en de informatieprofessional. Ditzelfde geldt voor professionele bronnen van uitgevers en marketing- en analysebureaus, die veelal uitsluitend via de bijbehorende ontsluitingsmechanismen toegankelijk zijn.

Figuur 1. Ideale vorm van 'slim' zoeken



Figuur 2. Traditionele vorm van zoeken



Systeem

Geavanceerde zoeksystemen kunnen de inhoud van contentbronnen analyseren, indexeren en classificeren door middel van complexe mathematische algoritmes, semantische netwerken, specifiek opgebouwde filters of taxonomieën. Hiermee kunnen deze systemen – in tegenstelling tot eenvoudiger zoeksystemen – inhoudelijke relaties leggen tussen documenten uit diverse bronnen. Daarbij wordt de betekenis van de content in de context gewogen, of wordt een zoekvraag – veelal door middel van een thesaurus of semantisch netwerk – geoptimaliseerd met extra relevante termen. Dit resulteert in soms opmerkelijk goede resultaten uit zowel gestructureerde als ongestructureerde bronnen.
Naast dit bevragen van de beschikbare bronnen zijn dergelijke mechanismen in meer of mindere mate in staat om de inhoud van bronnen automatisch te classificeren en in geschikte categorieën te plaatsen, zoals ook een bibliothecaris dit zou doen. Bovendien beschikken deze systemen doorgaans over *fuzzy search*, waarmee spelfouten of andere schrijfwijzen (bijvoorbeeld bij vreemde namen) door het systeem correct kunnen worden geïnterpreteerd.
De gebruiker krijgt dus de indruk dat het systeem *begrijpt* wat hij wil vinden en dat het systeem de documenten op inhoud kan groeperen. In de praktijk werkt deze inhoudelijke match bij hedendaagse geavanceerde zoekmachines heel aardig. De automatische classificatie werkt echter beter indien de categorieën door een beheerder worden gedefinieerd en minder goed indien deze door het systeem zelf uit de inhoud van de documenten wordt afgeleid.

Gebruikersinterface

Ten slotte kunnen geavanceerde systemen de gebruiker in staat stellen de vraag in natuurlijke taal aan het zoeksysteem op te geven. Met natuurlijk taal kan een gebruiker een gewone zin of zelfs een complete tekst als zoekvraag gebruiken. Dit is van cruciaal belang, omdat lang niet iedereen in staat is om een precieze zoekvraag samen te stellen.
Met name bij traditionele zoeksystemen leidt dit lang niet altijd tot tevredenheid over de opgeleverde resultaten. Veel woorden hebben bovendien synoniemen, die in de zoekvraag buiten beschouwing blijven, of homoniemen (bijvoorbeeld 'bank' in de betekenis van een financiële instelling óf meubelstuk).

Naast vele eenmalige zoekvragen hebben veel mensen permanente interessegebieden en/of een continue informatiebehoefte. De hedendaagse geavanceerde zoeksystemen kunnen aan deze continue informatiebehoefte tegemoet komen door middel van een persoonlijk informatieprofiel. *Agents* zoeken doorlopend naar relevante informatie, waardoor de gebruiker direct – en zonder er zelf naar te moeten zoeken – op de hoogte blijft van actuele ontwikkelingen op zijn interessegebied. Om een agent in staat te stellen goed passende informatie op te leveren, is het wel van belang de agent te trainen. Dit gebeurt doorgaans door middel van relevance feedback, waarbij de gebruiker aan de agent doorgeeft of opgeleverde documenten relevant of juist irrelevant zijn.
Andere soorten agents kunnen 'meekijken' met het werk waarmee de gebruiker bezig is. Als deze bijvoorbeeld een tekstverwerker gebruikt, doet deze *real time* agent in een apart venster suggesties voor gerelateerde informatie. Dit is een vorm van proactieve informatielevering die niet iedereen direct zal aanspreken, maar die toch veel gebruikers enorm behulpzaam kan zijn.

Mooi gereedschap

Foto: Egon Viebre

Voor mensen die associatief willen zoeken, stellen geavanceerde zoeksystemen gebruikers in staat om door categorieën met specifieke onderwerpen te bladeren (browsable directory), die hetzij handmatig, hetzij automatisch door het systeem zijn aangemaakt.

Geavanceerde systemen versus ideale 'slimme' systemen

Hoewel hedendaagse zoeksystemen prima mogelijkheden bieden voor geavanceerd vinden van informatie, bestaat een ideaal 'slim' zoeksysteem, zoals we die in dit artikel hebben gedefinieerd, nog niet. Daarvoor ontbreekt bij deze systemen met name nog de mogelijkheid om de betekenis van de inhoud van multimediabestanden (beeld, geluid) te begrijpen. Ook de integratie van diverse in- en externe bronnen levert soms nog problemen op. Toevoeging van metadata (verrijking) is bij veel bronnen eveneens noodzakelijk om goede zoekresultaten te garanderen. De ontwikkelingen gaan echter zeker in de richting van 'Luilekkerland'.

Introductie en gebruik van 'slim' zoeken

'Met een slim zoeksysteem kunnen we alle informatie die we nodig hebben moeiteloos ontsluiten! Doet u ons maar één keer slim zoeken.' Was het maar zo eenvoudig! Het succes van een 'slim' zoeksysteem binnen een organisatie is afhankelijk van de doelgerichtheid waarmee een organisatie het systeem heeft geselecteerd en ingericht. Voor een succesvolle exploitatie is beleid nodig, een goede communicatie naar de gebruiker en continu beheer.

Selectie van een systeem

De huidige zoeksystemen kunnen getypeerd worden naar de mate waarin en de wijze waarop een beheerder of gebruiker invloed kan uitoefenen op de manier van werken van het systeem en naar de mate waarin ze geschikt zijn voor centraal of decentraal beheerde omgevingen.
De eerste vraag die een organisatie dient te beantwoorden is: hoe werken wij? Oftewel, bij welke processen kan een zoeksysteem effectief worden ingezet.
Bijvoorbeeld: een overheidsorganisatie die zich bezighoudt met het inhoudelijk beheer van documenten, moet bij iedere verandering beoordelen of deze legitiem is. Deze beoordeling geschiedt op basis van precedenten die besproken en vastgelegd zijn in eerdere vergaderverslagen. Alleen ervaren medewerkers weten de besluitvorming bij deze precedenten te vinden in het grote bestand van notulen. Zij kunnen de juiste relaties leggen en bruikbare zoektermen definiëren, of weten uit het hoofd de datum van de betreffende vergadering.
Deze organisatie is niet in eerste instantie gebaat bij een beter zoeksysteem, maar bij een gecategoriseerde vastlegging van precedenten en een beslisboom om minder ervaren medewerkers te leiden naar de juiste categorie. Een zoeksysteem kan daarbij echter wel een aanvullende rol spelen.
De specificatie van de informatiebehoefte leidt tot een vaststelling van het verwachte zoekgedrag: zoeken gebruikers vooral om zich te oriënteren (bijvoorbeeld journalisten), of om eenduidige antwoorden op duidelijk omlijnde vragen te krijgen (bijvoorbeeld bij farmaceutisch onderzoek).

Voorts is het van belang te weten of het kennisdomein van de organisatie stabiel is, of aan sterke veranderingen onderhevig. Met andere woorden: of het beheer van het zoeksysteem centraal of decentraal, omvangrijk of gering zal zijn. Uiteraard spelen verder factoren als kosten, schaalbaarheid en integratie met andere applicaties een belangrijke rol bij de keuze.

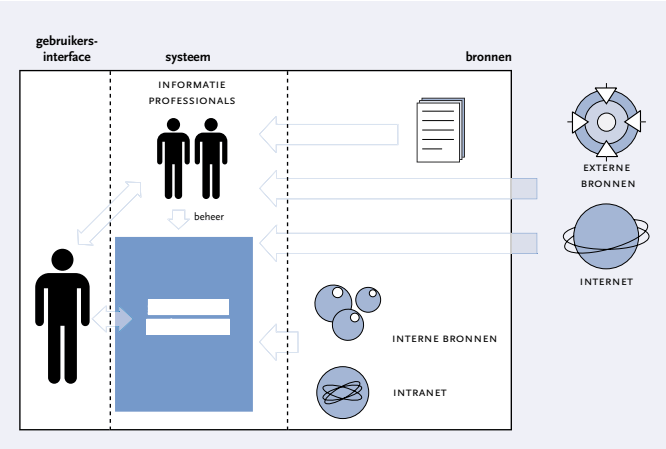
Inrichting en beheer van een systeem

De inrichting en het beheer zullen zich concentreren op een aantal deelaspecten. We noemen hier de belangrijkste.
Het toevoegen van metagegevens is een bewerkelijke taak. De claim is soms dat full-text zoeken deze arbeid overbodig zou maken. De bronnen bestaan echter lang niet altijd uit alleen full-text bestanden. Daarbij is categorisering van zoekresultaten eveneens noodzakelijk: een beperking tot bepaalde typen documenten, informatie over of geschreven door een persoon, documenten binnen een bepaalde datumrange, etcetera. Daarvoor zijn metagegevens nodig. Het toekennen daarvan kan wel met behulp van het zoeksysteem gebeuren: documenten kunnen met behulp van filters of agents automatisch geclassificeerd worden. De filters moeten echter zorgvuldig gemaakt worden en een handmatige controle en verfijnde afstemming blijft daarbij nodig.
Ook het selecteren van de bronnen dient zorgvuldig te gebeuren. Een kwalitatief slechte bron zal zoekresultaten in grote mate beïnvloeden. De selectie is niet een eenmalige handeling, maar een belangrijk onderdeel van het beheer van een zoeksysteem.
Het bouwen en onderhouden van filters en agents en van taxonomieën kan een lastige zaak zijn waar veel systeemkennis voor nodig is. De beslissing voor een centraal of decentraal georganiseerd beheer wordt genomen op basis van de verwachte informatiebehoefte, de dynamiek van het kennisdomein en van de te gebruiken bronnen.
Hoewel geavanceerde zoeksystemen bevraging in natuurlijke taal toestaan, bestaan er daarnaast nog andere, meer gecontroleerde wijzen van zoeken, die een beter resultaat kunnen opleveren (zoals uitgebreid zoeken met behulp van specifieke metagegevens). Het juiste gebruik van die manieren van zoeken vereist een training en voortdurende assistentie van gebruikers.

Het zal duidelijk zijn dat de verwachting ‘doe mij maar een zoeksysteem en alle problemen zijn opgelost’ niet zonder meer bewaarheid kan worden. Gebruikers zullen assistentie nodig hebben, zowel wat betreft de initiële inrichting als wat betreft het beheer en onderhoud van hun persoonlijke agents en de selectie van bronnen.

Veranderende rollen

Er is dus met de huidige geavanceerde zoeksystemen en toekomstige ideale ‘slimme’ systemen nog wel degelijk plaats voor de informatieprofessional en gestructureerde bronnen. Al zullen veranderingen even onvermijdelijk als uitdagend zijn.



Figuur 3. Huidige mogelijkheden met geavanceerd zoeken

Bronnen

Belangrijk is de *democratisering van bronnen*, waardoor medewerkers van een organisatie toegang krijgen tot alle voor hen relevante bronnen in plaats van via een intermediair. Dit geldt voor zowel in- als externe bronnen.

De rol van gestructureerde bronnen – en daarmee de taak van zorgvuldige selectie en beheer – blijft daarbij om twee redenen bestaan. In de eerste plaats is betrouwbare, actuele, authentieke, kwalitatief hoogwaardige en vaak gespecialiseerde content voor vele organisaties van eminent belang. Gestructureerde bronnen als een Document Management Systeem en professioneel uitgegeven contentbronnen bieden in principe deze kwaliteit (bijvoorbeeld met behulp van metagegevens). Want hoewel de huidige, geavanceerde zoeksystemen uitstekend zijn uitgerust om informatie te ontlenen aan ongestructureerde – en veelal ongecontroleerde – content (bijvoorbeeld vanaf het world wide web), staat of valt de uiteindelijke kwaliteit van de opgeleverde informatie met de kwaliteit van de beschikbare content (rubbish in, rubbish out).
In de tweede plaats verschaft de vastgelegde structuur van dergelijke bronnen het zoeksysteem belangrijke contextuele informatie voor het classificatie- en indexeringsproces. Ook dit heeft een positief effect op de kwaliteit van de beantwoording van de zoekvraag in vergelijking tot volledig ongestructureerde content.

Informatieprofessional

De informatieprofessional zal minder vaak in de rol van informatiemakelaar worden benaderd door overige medewerkers binnen de organisatie. Deze kunnen immers uit de voeten met het geavanceerde zoeksysteem. Dit betekent niet dat de informatieprofessional geen voeling meer behoeft te houden met de informatiebehoefte van de medewerkers. Het is nog steeds belangrijk om de vertaling te kunnen maken naar (digitale) dienstverlening. Deze dienstverlening heeft niet alleen betrekking op interne communicatie en voorlichting, maar vormt tevens de basis voor het te volgen beleid met betrekking tot de inrichting van het zoeksysteem.

De professional zal de expertrol gaan vervullen bij de selectie en het inrichten van het geavanceerde zoeksysteem. Afhankelijk van de werking van het systeem, zal de informatieprofessional classificatiefilters en taxonomieën moeten gaan beheren, evenals centrale agents over specifieke onderwerpen die centraal toegankelijk zijn voor de medewerkers. Ook ligt de controle over automatisch gegenereerde onderwerpcategorieën (browsable directories) bij de informatieprofessional, die de mogelijkheid moet hebben hierin wijzigingen te kunnen aanbrengen.

Het onderkennen van de verschillende potentiële in- en externe bronnen en het vaststellen van de te benutten contentbronnen is bij toepassing van een geavanceerd zoeksysteem evenzeer een belangrijke activiteit van de professional. Net als het beheer, classificatie en indexering van de traditionele, papieren collecties (boeken, tijdschriften).
Van de professional mag bovendien een bijdrage worden verwacht in het structureren van bronnen binnen de organisatie. Daarmee krijgt de informatieprofessional een belangrijke centrale rol in het kennismanagement van de organisatie.

‘Slim’ zoeken steeds dichterbij

Ideale, ‘slimme’ zoeksystemen die zowel de gebruiker als de contentbronnen kunnen begrijpen alsof het een professionele zoeker betreft, bestaan eigenlijk nog niet. Wel zijn er systemen die in meer of mindere mate reeds over mogelijkheden van ‘slim’ zoeken beschikken. Voorbeelden hiervan zijn Autonomy, Verity, Knowledge Concepts en Sibylle. Deze systemen verschillen onder meer in de manier waarop de zoekvraag wordt gematched met de beschikbare content en de mate van inrichtingsmogelijkheden van het systeem. Autonomy gebruikt daarvoor complexe mathematische algoritmen, die autonoom aan de gang gaan. Dit beperkt niet alleen het benodigde beheer en onderhoud van het systeem, maar tevens de controle op de inrichting van het systeem. Verity biedt meer controlemogelijkheden, waaronder extended queries en Topics om de zoekvraag met relevante termen uit te breiden. Dit kan leiden tot hoge precision en recall, maar vergt nadrukkelijk beheer en onderhoud. Ditzelfde geldt voor Sibylle dat goede resultaten behaalt op basis van handmatig opgebouwde filters. Deze en andere, vergelijkbare systemen bieden uitstekende zoek- en profieldiensten waarmee echt ‘slim’ zoeken steeds dichterbij komt. Welk systeem voor een organisatie het beste zal functioneren is afhankelijk van doelstellingen, eisen en wensen van de organisatie en de randvoorwaarden die vanuit financieel, technisch en organisatorisch oogpunt aan de inrichting van het systeem gesteld worden.

Jan Mark Pleijsant is senior consultant contentmanagement en information retrieval bij Cap Gemini Ernst & Young, practice Documentair Informatie- en Procesmanagement. Harriët Hangelbroek is senior consultant contentmanagement en information retrieval bij Cap Gemini Ernst &Young, practice Documentair Informatie- en Procesmanagement. Jan van der Starre is senior consultant bij Cap Gemini Ernst &Young, practice Documentair Informatie- en Procesmanagement. Zijn werkterreinen zijn documentaire informatiesystemen, information retrieval en digitale duurzaamheid.