Eric Sieverts

Zoeken in Google of in 'echte' bronnen

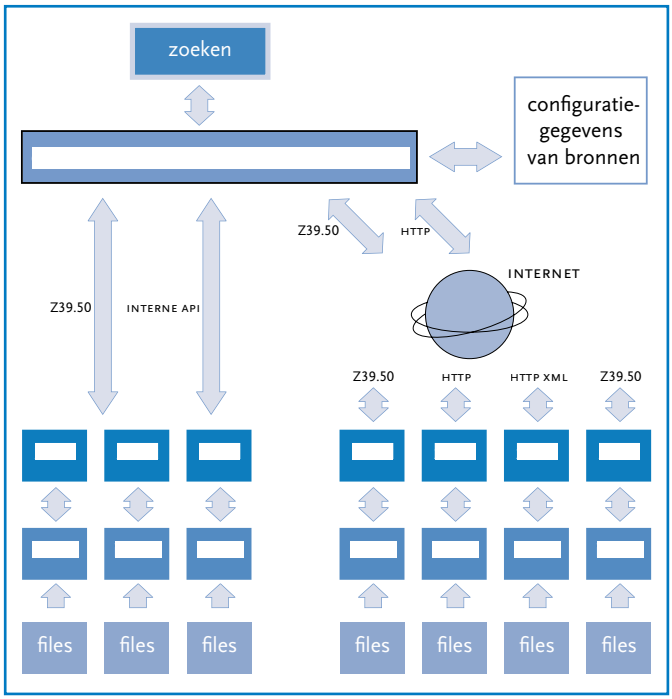
Google is bijna synoniem geworden met het nieuwe zoeken. Wereldwijd is het verreweg het meest gebruikte zoekhulpmiddel. Steeds meer mensen komen ook niet verder meer dan Google. Wat je met Google niet kunt vinden, lijkt niet meer de moeite waard te zijn. De toekomst van professionele databases en het professionele zoeken lijken zo ten onder te gaan aan het grote succes van het zoeken. Eric Sieverts kijkt naar mogelijke oorzaken, hoe erg het is en wat er aan gedaan zou kunnen worden.

Zoekmachines op het web – en vooral Google – hebben een enorme impact gehad op het zoekgedrag van mensen. Iedereen kan tegenwoordig zoeken. En steeds meer mensen maken ook voor hun professionele informatiebehoeften gebruik van webzoekmachines. Die vormen inderdaad een prima aanvulling op de klassieke zoektools en de klassieke online bronnen die ons al veel langer ten dienste stonden. Ook steeds meer informatie die voorheen helemaal niet digitaal beschikbaar was, kan nu op die manier op het web gevonden worden.
Voor sommige groepen gebruikers leidt dat ertoe dat ze alleen nog maar Google gebruiken; zelfs sommige leidinggevenden schijnen te denken dat alle voor hun organisatie belangrijke informatie gratis – via zoekmachines – op het web te vinden is, zodat geen bibliotheek of informatiecentrum meer nodig is.

Waar komt het grote succes van die zoekmachines vandaan? Daar is een aantal redenen voor:
• ze zijn zo makkelijk te gebruiken;
• door nieuwe, vaak taaltechnologische retrieval-technieken geven ze zulke goede zoekresultaten;
• dat uit zich vooral in verbeterde methoden van relevantieordening;
• er zit zo veel in.

Dat gebruiksgemak, die nieuwe retrieval-technieken en die goede relevantieordening zijn zaken die direct met elkaar te maken hebben en die in dit blad ook al regelmatig aan de orde zijn gekomen. Er hoeven geen commando's en geen ingewikkelde Booleaanse combinaties van zoektermen te worden ingetikt. Een paar woorden in een enkel zoekvenstertje leveren meteen een aanklikbare lijst met resultaten. Die resultaten zijn geordend op grond van hun mate van overeenkomst met het ingetikte rijtje zoekwoorden en ook steeds vaker op grond van het automatisch berekende belang van de betreffende webpagina's of websites. Dat is die zogenaamde relevantieordening.
Steeds vaker worden ook automatisch suggesties gedaan voor een iets afwijkende spelling van een zoekterm, die meer zou opleveren (Google: 'agression: did you mean aggression?') of worden juist – op basis van statistiek – suggesties gedaan om een zoekvraag in te perken op een bepaalde betekenis of context van gebruikte zoektermen (bij Teoma, Wisenut, All-the-web, AltaVista).

Bediening van die zoekmachines hoef je dus nauwelijks te leren. Dat neemt overigens niet weg dat een aantal van die zoekmachines, als een gebruiker niet schrikt van wat ingewikkelder zoekacties, ook nog allerlei mogelijkheden biedt om meer gerichte en complexere zoekacties te doen. Google, AltaVista, Hotbot, All-the-web en enkele andere bieden desgewenst wel Booleaanse combinaties, gebruik van zoekvelden, 'citatie'-zoeken, inperkingen op 'formele' kenmerken, zoeken naar plaatjes en dergelijke. Toch is het volgens mij vooral het adagium 'tik maar wat in en je vindt altijd wel wat', dat zoekmachines zo populair heeft gemaakt.



Figuur 1. Metasearch/gateway-oplossing

Daarnaast geeft het gebruikers natuurlijk ook vertrouwen dat je potentieel zo 'veel' kunt vinden. Ondanks frustraties van onhanteerbaar grote zoekresultaten wordt steeds meer belang gehecht aan de grootte van zoekmachines. Met tussen de 1,5 en 2,5 miljard webpagina's staan Google, All-the-web en Wisenut daarbij duidelijk bovenaan. Daarmee is mijn vroegere waarschuwing, dat bij online hosts vele malen meer informatie te vinden is dan op het web – in kwantitatief opzicht – intussen aardig achterhaald.

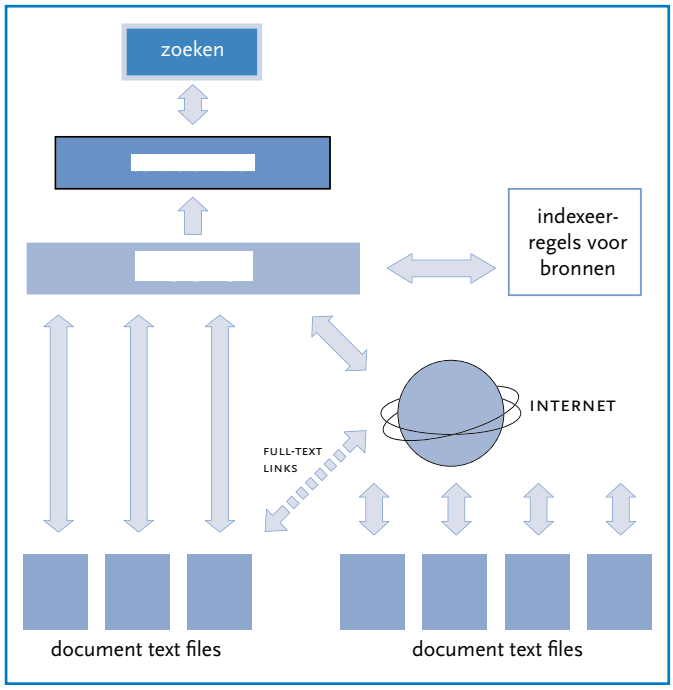
Wat zoekmachines missen

Ondanks die kwantiteit, wordt door zoekmachines toch ook een heleboel gemist. En daarbij zit vaak juist heel veel kwaliteit. In de eerste plaats natuurlijk al die professioneel verzamelde, betrouwbare en vaak al van oudsher in goed ontsloten databases opgenomen informatie, die alleen tegen betaling bij online hosts of op afgeschermde lokale intranetten van allerlei organisaties beschikbaar is. Op die informatie en die databases kom ik vanuit een ander gezichtspunt nog nader terug.

Daarnaast zijn er allerlei soorten informatie die wel voor iedereen gratis toegankelijk zijn, maar die zoekmachines toch meestal missen. Dat zijn onder meer:
• niet-HTML-documenten (flash-files, PDF-bestanden, office-documenten en dergelijke), al indexeert Google een groot deel daarvan intussen ook;
• real-time, of op zijn minst zeer frequent vernieuwde gegevens, al bestaan er wel gespecialiseerde zoekmachines die bepaalde delen van dit soort informatie – zoals het 'echte' nieuws – door zeer frequent herindexeren toch toegankelijk maken;
• in doorzoekbare databases opgeborgen informatie die je pas vindt als je je zoekvraag intikt in de eigen zoekschermen van elk van die honderdduizenden individuele databases waar je via het web bij kunt komen;
• alle informatie op webpagina's waarvan een zoekmachine het bestaan nog niet weet, omdat ze nieuw zijn, omdat de servers waarop ze staan bij de zoekmachine onbekend zijn, of omdat de zoekmachine domweg nog geen tijd gehad heeft hun inhoud te indexeren; zo waren mijn persoonlijke webpagina's vele maanden na aanmelden bij Google eindelijk pas terug te vinden.

Als je je dit realiseert, en je bovendien bedenkt dat die ongeorganiseerde, gefragmenteerde zooi op het web, ook op geen enkele manier consistent ontsloten is, dan moeten zoekacties met webzoekmachines over het algemeen wel een zeer belabberde recall opleveren. Maar veel gebruikers – ook studenten – zullen daar nauwelijks om malen. Als je maar wat gevonden hebt.

Uiteraard zijn er ook een heleboel informatiezoekers die zich wel bewust zijn van het belang van veel niet gratis op het web te vinden informatie. Wetenschappelijk onderzoekers kunnen zich niet permitteren belangrijke nieuwe informatie te missen, ook al moeten ze daar zelf naar op zoek in al die zoeksystemen die hun universiteit haar eindgebruikers aanbiedt. Bij veel, vooral grote bedrijven wordt al evenzeer onderkend dat het van belang is zo volledig en betrouwbaar mogelijk geïnformeerd te zijn. Of het nu om



Figuur 2. Oplossing met lokale centrale index

technische of om bedrijfsinformatie gaat en of men nu zelf zoekt of het aan zoekprofessionals uitbesteedt, daar worden professionele informatiebronnen nog wel degelijk op waarde geschat, zoals ook duidelijk wordt in de artikelen van Smulders, Evers en Stuijvenberg, elders in dit nummer van Informatie Professional.

Databases *of* zoekmachines

Bij heel wat grote organisaties zijn veel van die professionele informatiebronnen wel degelijk beschikbaar. Door het afsluiten van licenties met hosts, cd-romleveranciers, databaseproducenten, uitgevers, tijdschriftagenten en andere 'aggregators' van informatiebronnen, kunnen allerlei databases en gegevensverzamelingen op lokale intranetten beschikbaar zijn. Binnen de organisatie – bedrijf, ministerie, universiteit – kan iedereen daar meestal onbeperkt in zoeken. Maar gebeurt dat ook, of stelt men zich ook dan tevreden met Google?

Ook zonder uitgebreid onderzoek naar zoekgedrag bij verschillende soorten gebruikers bij verschillende soorten organisaties, is makkelijk een aantal redenen te bedenken waarom iemand ook in een dergelijke situatie een sterke voorkeur voor Google kan blijven houden:
• het kale Google-interface met dat ene zoekvakje is zo simpel en overzichtelijk, zoals ik in de eerste paragraaf al betoogde;
• een enkele Google-zoekactie doorzoekt in één keer die ruim 2 miljard webpagina's op het web, ondanks de in de vorige paragraaf genoemde beperkingen – die de meeste gebruikers zich overigens nauwelijks zullen realiseren – en ondanks de zeer wisselende kwaliteit en betrouwbaarheid van de op het web gevonden informatie;
• van die 'echte' professionele informatiebronnen zijn er meestal (te) veel, die vaak allemaal afzonderlijk doorzocht moeten worden: verschillende bibliografische databases, verschillende primaire bronnen, individuele tijdschriften,

sites van uitgevers van plukjes tijdschriften en van aggregators van allerlei types tijdschriften. Zo biedt de website van de Universiteit Utrecht gebruikers binnen de campus toegang tot bijna 200 afzonderlijk te doorzoeken informatiebronnen en kunnen bijna 6000 wetenschappelijke tijdschriften individueel bekeken en (meestal) doorzocht worden (zij het dat het merendeel van die tijdschriften geclusterd beschikbaar is bij 27 grote uitgevers en aggregators, zoals Elsevier, Wiley, Ebsco, HighWire of JStor, en er maar ruim 100 van kleinere individuele aanbieders afkomstig zijn);
• al die afzonderlijke bronnen hebben vrijwel allemaal hun eigen specifieke zoekinterface;
• veel van die bronnen hebben bovendien interfaces die zijn overladen met functionaliteit die beslist uiterst nuttig is om goed te kunnen zoeken (lees het artikel van Marten Hofstede elders in dit nummer daar maar op na), maar die er ook voor zorgen dat de (nog) niet ervaren gebruiker vaak door de bomen het bos niet meer ziet; het interface van SilverPlatter/ERL, met onbetwistbaar superieure zoekmogelijkheden, is daar een goed voorbeeld van.

Wat zou u gebruiken als niemand u het verschil tussen Google en die databases had uitgelegd? Wat zou u gebruiken als u niet wist wat u allemaal miste wanneer u zich alleen tot Google beperkte? Wat zou u gebruiken als u het eigenlijk helemaal niet zo erg vindt een heleboel toch wel relevante informatie te missen? Het verzorgen van (steeds uitgebreider) instructie is nuttig en noodzakelijk, maar zeker niet de panacee voor beter zoekgedrag.

Databases *en* zoekmachines

Zoals ik hiervoor al even aangaf, is het voor veel organisaties om allerlei redenen van belang dat optimaal gebruik wordt gemaakt van beschikbare professionele informatiebronnen. Voorlichting en instructie met betrekking tot dat gebruik zijn voor bibliotheken en informatiecentra intussen al bijna traditionele taken geworden. Dat is echter lang niet altijd voldoende om te zorgen dat individuele medewerkers ook werkelijk optimaal gebruik maken van al die informatieproducten waarvoor vaak duur-betaalde licenties zijn afgesloten.
De bibliotheek heeft dus ook de taak het gebruik van die informatiebronnen bijna even eenvoudig te maken als zoeken met Google. Ze moet er dus naar streven de in de vorige paragraaf genoemde hinderpalen voor het zoeken in professionele bronnen zo veel mogelijk weg te nemen.

Een belangrijke stap daarbij is iets te doen aan het probleem van de vele afzonderlijk te doorzoeken bronnen die ook nog bijna allemaal verschillende interfaces hebben. Op dit moment lijken daarvoor twee veelbelovende aanpakken te bestaan. De meest gebruikte is de metasearch- of gateway-aanpak. Daarbij wordt gebruikgemaakt van een programma dat voor alle opgenomen bronnen één uniform zoekinterface biedt. Het programma zelf doet niet meer dan het gelijktijdig doorsturen van gestelde zoekvragen naar de afzonderlijke zoeksystemen van – desnoods een groot aantal – door de gebruiker geselecteerde bronnen. Daarvan terugontvangen resultaten worden ook eerst weer

door het programma verzameld, zodat ook die in een uniforme presentatie aan de gebruiker worden aangeboden. Voor de communicatie met die verschillende op interne en externe computers draaiende zoeksystemen, wordt nog vooral gebruikgemaakt van het Z39.50-protocol. Maar ook bronnen met een gewoon web-interface (via het http-protocol) of die gebruik maken van nieuwe XML-technieken kunnen in dergelijke gateway-systemen worden opgenomen. Elders in dit nummer is in het artikel van Marten Hofstede een overzicht te vinden van dergelijke systemen die op dit moment op de markt zijn. Tevens geeft hij daarin een analyse van de belangrijkste beperkingen en nadelen die nu nog aan de meeste gateway-systemen kleven. Dat zijn onder meer allerlei beperkingen van de functionaliteit ten opzichte van de oorspronkelijke zoeksystemen en onduidelijkheid welke velden precies worden doorzocht. Veel van de gesignaleerde problemen blijken terug te voeren op beperkingen van het Z39.50-protocol en op het feit dat slechts een grootste gemene deler van de mogelijkheden van de onderliggende systemen wordt aangeboden.

Een andere mogelijke aanpak is die waarbij een eigen zoekmachine wordt gebruikt. Met zo'n programma wordt op het eigen systeem een lokale centrale index aangemaakt op de tekstgegevens van alle te doorzoeken bronnen. Zo kan niet alleen een uniform zoekinterface worden aangeboden, maar ook een uniforme zoekfunctionaliteit voor alle bronnen. Ze worden immers allemaal met hetzelfde retrieval-programma doorzoekbaar gemaakt. De met een dergelijk programma geïndexeerde gegevens kunnen zowel gestructureerde bibliografische gegevens, als full-text tijdschriftartikelen, als webpagina's zijn. Verder kunnen in principe zowel gegevens die intern op het eigen lokale netwerk staan, als externe gegevens via internet geïndexeerd worden.

Dat dit laatste nog al eens problemen oplevert – en dan vooral licentie-technische – besprak ik enkele jaren geleden al eens in dit blad.[1] Sindsdien is aan die situatie nog niet veel veranderd. Daardoor beperken de praktijkvoorbeelden zich nog vrijwel altijd tot het indexeren van lokaal aanwezige gegevens.
Het op deze techniek gebaseerde zoeksysteem van de Koninklijke Bibliotheek (http://www.kb.nl) doorzoekt gegevens uit de eigen KB-catalogus en uit een flink aantal andere eigen databases, alsmede lokaal aanwezige digitale publicaties. Bij de UB Utrecht worden op deze wijze alleen nog maar tijdschriftartikelen van een viertal leveranciers doorzoekbaar gemaakt; en dan ook alleen de bibliografische gegevens en abstracts, omdat die lokaal geladen zijn. Voor de volledige teksten wordt automatisch naar de versie op de computer van de leverancier doorgelinkt, maar die zijn nog niet full-text doorzoekbaar.
Dit systeem, dat sinds kort onder de naam Omega door het leven gaat, heeft intussen een sterk op Google geïnspireerd gebruikersinterface gekregen.

De toekomst

Uit de voorbeelden in de vorige paragraaf werd duidelijk dat wel al gewerkt wordt aan oplossingen om gebruikers eenvoudiger uniforme interfaces te bieden waarmee – liefst in één keer – alle gewenste professionele bronnen

Wat gebruikers willen

doorzocht kunnen worden. Ook werd duidelijk dat aan dergelijke systemen nog heel wat nadelen kleven. Nieuwe ontwikkelingen gaan echter de goede kant op.
Opvolgers voor Z39.50, de XML-gebaseerde ZING (Z39.50 next generation), SRW (search & retrieve through the web) en SRU (search & retrieve through URL) protocollen, beloven metasearch-oplossingen eenvoudiger en kwalitatief wat beter te kunnen maken. Het met een eigen zoekmachine willen indexeren van externe gegevens waarvoor men een licentie heeft, lijkt stilaan een gebruikelijker wens te worden. Automatische methoden om documenten te classificeren of trefwoorden toe te kennen zijn sterk in ontwikkeling,[2] zodat ook uniforme, gecontroleerde ontsluiting van allerlei verschillende bronnen wellicht tot de mogelijkheden gaat behoren. En ook van de slimme zoeksystemen uit het artikel van Pleijsant (elders in dit nummer) worden al voorlopers genoemd.

Anderzijds worden ook de zoekmachines op het web steeds slimmer en eenvoudiger en komt, bijvoorbeeld door het Open-archive initiatief, ook steeds meer professioneel belangrijke informatie vrij op het web beschikbaar. De strijd om de gunst van de zoeker, tussen de echte bronnen en de Googles van het web, zal dus zeker nog wel even doorgaan.

Noten

1. Eric Sieverts, Onno Mastenbroek, Natalia Grygierczyk – Een uniform retrieval-systeem voor de Universiteit Utrecht – *Informatie Professional*, 3 (1999), 10 (oktober) blz. 34-40.
2. Joop van Gent, Onno Makor – Automatische verrijking in de praktijk – *Informatie Professional*, 6 (2002), 7/8 (juli/augustus) blz. 28-31.

Dr. E.G. Sieverts is werkzaam op de Hogeschool van Amsterdam en bij de Universiteitsbibliotheek Utrecht. Hij is redacteur van Informatie Professional.