



Zoeken op WWW met Lycos

Wie Weet Welke Wellicht Waardevolle Wijsheid Waar op het Web op ons Wacht *deel 1*

Eric Sieverts

Een inleiding in het zoeken naar informatie op World Wide Web, gevolgd door een bespreking van het retrievalstelsel Lycos

Tot voor kort was Gopher één van de belangrijkste hulpmiddelen om je weg te vinden in de vele diensten en informatiebronnen die op het Internet aanwezig zijn en om daar bovendien meteen toegang toe te krijgen. Toch vormen menusystemen met hun soms eindeloze keuzelijsten niet zo'n snel en handig hulpmiddel om gericht informatie te zoeken, zeker niet als menu's vaak weinig systematisch, kennelijk door niet-informatiedeskundigen, zijn opgezet. Voor de verknochte Internetgebruiker was de komst van het Veronica-zoekstelsel daarom een hele verademing. Wie aan professionele online information retrieval systemen gewend is, ziet echter al snel dat Veronica bijzonder weinig geavanceerd is en maar zeer beperkte mogelijkheden biedt.

Veronica doet immers niet meer dan een woordindex aanleggen van alle tekst die in Gopher-menu's voorkomt. In een enkele opdrachtregel kan de gebruiker vervolgens een Booleaanse combinatie met AND's en OR's formuleren. Een goede kant van Veronica is natuurlijk dat alle Gopher-menu's, waar ook ter wereld, regelmatig automatisch afgeschuimd worden om zo volledig en zo up-to-date mogelijke informatie te kunnen bieden. De kwaliteit van die informatie is echter niet beter dan die van de oorspronkelijke (vaak korte, weinig descriptieve) menubeschrijvingen. Je kunt nooit verwachten meer te vinden dan daarin staat. Het is zoiets als literatuur zoeken op alleen maar titels (en dan nog heel korte). Er zijn geen trefwoorden en geen uitgebreider beschrijvingen voor handen. Dat de achterliggende informatie al helemaal geen ingangen voor deze index levert, is Veronica echter nauwelijks aan te rekenen. Uiteindelijk mondt het merendeel van de interessante Gopher-keuzes immers uit in automatisch opgezette Telnet-sessies waarmee achter-

liggende bibliotheekcatalogi of lokale retrieval-systemen opgestart worden, waar niet eenvoudig, automatisch, van buiten af, indexeerbare tekstinformatie uit afgeleid kan worden.

Nu Gopher in korte tijd al weer hopeloos ouderwets geworden lijkt te zijn en iedereen zich verdringt om zijn eigen World Wide Web pagina op Het Net te brengen, ziet de wereld voor de informatiezoeker er ineens weer heel anders uit. In de eerste plaats biedt www met zijn op hyperlinks gebaseerde structuur natuurlijk een veel flexibeler manier om alles met alles te kunnen doorverbinden dan de starre Gopher-menu's. (Al blijf je als gebruiker natuurlijk nog steeds afhankelijk van de systematiek en de kwaliteit van de Hyperlinks die de bouwers van www-pagina's in hun teksten ingebouwd hebben.) Een tweede voordeel is dat in World-Wide-Webpagina's veel uitgebreider beschrijvingen met veel meer tekst gebruikt worden en dat die teksten zelf vaak al een belangrijk deel van de geboden informatie bevatten. Dat biedt in principe de mogelijkheid wat zinvoller en bruikbaarere zoeksystemen te ontwikkelen. Zijn die er echter ook al? Het antwoord op die vraag is een nog enigszins aarzelend maar toch ook al onmiskenbaar JA. Ook voor die systemen geldt echter, zoals bij zoveel op het Internet, dat je er wel eerst moet zien achter te komen waar welke zoekmogelijkheden zich bevinden. Het zijn namelijk door Internetgebruikers zelf opgezette, meest uit lokale informatienood geboren, systemen. Gelukkig zijn er ook al weer heel wat mensen en instituten die informatie over die zoeksystemen verzameld hebben, en die overzichten daarvan in hun eigen World Wide Web-pagina's opgenomen hebben. Ik heb een eerste beperkte inventarisatie van www-retrieval-systemen gemaakt. Ik heb daarbij gebruik gemaakt van enkele van zulke, op informele wijze verzamelde overzichten (van een familielid in Bristol, van een spreker van City University Dublin op het recente StinfoN-

Dr. E.G. Sieverts is docent aan de Faculteit Economie en Informatie van de Hogeschool van Amsterdam.

In bijgaande tabel heb ik vast de namen en de URL's vermeld (de 'uniform resource locators' oftewel de WWW-adressen waarmee ze aangeroepen kunnen worden) van de systemen die ik tot nu toe heb verzameld. De nieuwsgierige lezer hoeft dan niet op volgende afleveringen te wachten alvorens ze zelf te kunnen uitproberen. Met dat lijstje heb ik overigens geenszins een claim gelegd; wie ervaringen met één van de genoemde systemen op papier wil zetten moet dat zeker niet nalaten. Laat echter wel tevoren even op docsiev@ai.fei.hva.nl weten over welk systeem, zodat dubbel werk voorkomen wordt. Ook aanvullingen op mijn lijstje zijn uiteraard welkom.

Retrieval-systemen voor World-Wide-Webinformatie

<i>naam</i>	<i>aanbieder</i>	<i>URL</i>
AliWeb	Nexor	http://web.nexor.co.uk/public/aliweb/doc/search.html
CUI W3/Catalog	Univ. Genève	http://cui_www.unige.ch/w3catalog
EINet/Galaxy	EINet/MCC	http://galaxy.einet.net/search.html
Harvest	Colorado Univ.	http://harvest.cs.colorado.edu/
JumpStation	Stirling Univ. (Scotland)	http://www.stir.ac.uk/jsbin/js
Lycos	Carnegie Mellon Univ.	http://lycos.cs.cmu.edu
NIKOS*	Rockwell Network Systems/ California Polytechnic	http://www.rns.com/cgi-bin/nikos
W5	Univ. Utrecht	http://pablo.ubu.ruu.nl:8000/
WebCrawler	Univ. of Washington (Seattle)	http://webcrawler.cs.washington.edu/ WebCrawler/WebQuery.html
WWW Worm	Univ. of Colorado	http://www.cs.colorado.edu/home/ mcbryan/WWWWW.html
Yahoo Search	Stanford Univ.	http://akebono.stanford.edu/yahoo/ search.html

* vroegere naam 'WWW Nomad'; ook wel bekend als 'Zorbamatic'

congres, enzovoort), en van met gebruikmaking van die zoekmogelijkheden verder nog gevonden overzichten.

Dat heeft inmiddels al een aardig lijstje opgeleverd. De daarin voorkomende systemen verschillen onderling overigens nogal. In de eerste plaats in de aantallen *www*-pagina's die zo toegankelijk gemaakt zijn. Dat kan variëren van een schamele 12.000 tot het respectabele aantal van bijna 1,5 miljoen! Voor de gebruiker zal het dus nuttig zijn te weten op welke wijze de in een bepaald zoekstelsel geïndexeerde gegevens ge-

selecteerd c.q. op het Net opgespoord zijn. In de tweede plaats worden door sommige systemen de volledige teksten van de ontsloten *www*-pagina's geïndexeerd, terwijl andere dat alleen met de titels, de pagina-headers, de hyperlink-ankers en/of de bestandsnamen doen. Ook dat is voor een gebruiker uiteraard nuttig te weten. Dat alles levert te veel materiaal (en te veel werk) op om meteen alle systemen uit mijn lijstje in een enkel artikel te bespreken. Deze keer bespreek ik Lycos, mijn op dit moment favoriete systeem.

Lycos

Onder het motto 'veel is lekker' gaat mijn voorkeur voorlopig uit naar het Lycos-zoeksysteem van Carnegy Mellon University. Daarin zijn namelijk eind december 1994 al bijna 1,5 miljoen www-pagina's geïndexeerd en bovendien worden die ook nog full-text geïndexeerd. De naam van dat systeem is afgeleid van de Lycosidae, een familie van grote grond-spinnen ('spiders') die hun prooi niet in een web vangen maar er hollend achteraan gaan. Ze staan bekend om hun snelheid en hun vooral nachtelijke activiteit. De bouwers van Lycos vinden kennelijk dat die karakteristieken op hun systeem van toepassing zijn. Vragen die je je bij zo'n getal van 1,5 miljoen pagina's kunt stellen zijn natuurlijk:

- hoe komen ze aan de gegevens van al die pagina's?
- hoe weet je of ze daarmee ook 'alles' dekken?

Op die eerste vraag is wel bij benadering een antwoord te geven; op de tweede niet echt. Zoals bij meer van dergelijke zoekprogramma's wordt door Lycos voor het opsporen van alle www-informatie op het Internet gebruik gemaakt van een zogenaamde 'robot'. Je moet je daarbij uiteraard geen Starwars-achtig figuurtje voorstellen, maar een softwareprogramma dat op een slimme manier www-pagina's op het Net inventariseert (en indexeert). Er wordt in dit verband ook wel van 'knowbots' gesproken, hetgeen staat voor 'knowledge acquisition robots'. In eerste instantie gaat het vooral om het inventariseren van welke pagina's er allemaal zijn. Door het feit dat Hyperlinks het plaksel vormen dat het Web bijeenhoudt, inclusief directe links tussen pagina's op Web-servers in heel verschillende uithoeken van de wereld, kan zo'n robotprogramma domweg hyperlinks die hij tegenkomt volgen om nieuwe pagina's te vinden, in de hoop zo het hele Web in kaart te brengen.

het Web afschuimen

Dat volgen van die links zal echter wel op een slimme manier moeten gebeuren, want hoe werk je, op een bepaalde plek in het Web begonnen, op een systematische wijze een zich steeds verder vertakkend net van verwijzingen naar andere Web-pagina's af? Wie ooit wel eens geprobeerd heeft in een eenvoudige hypertext het hele net van doorverwijzingen in beeld te brengen, zal gemerkt hebben dat je zelfs met niet meer dan 100 schermen al heel snel volledig de draad kunt kwijt raken. Nu is een computer natuurlijk veel consciëntieuzer in dit soort zaken, maar er blijft toch altijd nog een goed uitgekiende strategie en boekhouding nodig om deze taak aan te pakken wanneer het om meer dan een miljoen pagina's of documenten gaat. Daarbij wordt soms fundamenteel onderscheid

gemaakt tussen 'depth first' en 'breadth first' strategieën. Enigszins simplificerend kun je stellen dat bij methodes van de eerste soort telkens de eerste de beste link die de robot tegenkomt gevolgd wordt, steeds 'dieper' van pagina naar pagina springend tot hij ten slotte niet verder meer kan, daarbij voortdurend administratie bijhoudend van alle al tegengekomen schermen. Bij de tweede methode werkt hij eerst systematisch alle links vanuit de eerste pagina af, vervolgens alle links vanuit die pagina's enzovoort. Welke methode precies bij Lycos toegepast wordt ben ik niet nagegaan (en dat is in feite voor ons als gebruikers ook niet zo belangrijk).

Een vraag die dit natuurlijk wel oproept is of een robot op zo'n manier werkelijk in een eindige tijd overall komt. Die eindige tijd zit kennelijk wel goed, want de robot van Lycos blijkt slim genoeg geprogrammeerd te zijn om met 1,5 miljoen pagina's nog altijd uit de voeten te kunnen. De vraag of er geen eilandjes in het Web zitten waarheen vanuit geen enkele andere pagina uit 'de rest' van het Web verwezen wordt, dus waar de robot nooit komt, durf ik echter niet met volledige zekerheid ontkennend te beantwoorden. Die mogelijkheid lijkt immers theoretisch alleszins aanwezig. Anderzijds wordt de kans daarop heel sterk gereduceerd door de tomeloze activiteiten van allerlei lokale organisaties die alle Web-aanbieders in bepaalde geografische regio's trachten te inventariseren en onder overkoepelende menu's en landkaartjes trachten te hangen. Door al die collectieve inspanningen schat ik dat er maar heel weinig geïsoleerde pagina's en eilandjes over zullen blijven. In dat geval is deze manier van data-acquisitie heel wat eenvoudiger dan die welke producenten van bibliografische bestanden zich moeten getroosten om een zo volledig mogelijke dekking van hun onderwerpsgerichte bestanden na te streven. Overigens biedt Lycos Web-aanbieders ook nog de mogelijkheid om actief de URL's van hun eigen pagina's aan het systeem toe te voegen en eventueel al weer opgeheven pagina's te verwijderen.

Overigens blijkt de tijdsfactor wel degelijk een rol te kunnen spelen, want Web-pagina's waar Lycos al eens geweest is, blijken niet op korte termijn opnieuw bezocht te worden voor herindexering. De prioriteit lijkt in eerste instantie te liggen bij het opsporen van nieuwe pagina's. (De wekelijkse updatefrequentie heeft dus kennelijk daarop betrekking). In de praktijk bleek namelijk een pagina waarnaar ik medio januari 1995 op zoek was, geïndexeerd - en dus zoekbaar - te zijn op grond van zijn inhoud van medio november 1994. Deze datum werd in het zoekresultaat gelukkig ook vermeld. De werkelijke inhoud van die pagina bleek er intussen echter al weer anders uit te zien, met stukken tekst waarop Lycos hem nog niet kon terugzoeken; inderdaad bleken de laatste wijzigingen aan de pagina zelf half december aangebracht te zijn.

Zoeken in Lycos

Na deze wat lang uitgevallen beschrijving van de data-acquisitie door Lycos, nu wat meer over de zoekmogelijkheden. De gebruiker die met het www-interface van de programma's Mosaic of Netscape werkt, kan in een zoekvenster van een enkele regel een hele reeks zoektermen intikken. Daarop voert het systeem dan automatisch een 'best-match' zoekactie uit, waarbij het zoekresultaat in 'geranke' volgorde gepresenteerd wordt. Dat houdt dus in dat in principe alle pagina's geselecteerd worden waarin maar minstens één van de ingetikte termen voorkomt. Vervolgens worden die echter wel in zodanige volgorde gepresenteerd dat je het eerst die pagina's te zien krijgt waarin de meeste van je zoekwoorden voorkomen. Daarbinnen wordt bovendien nog verder gesorteerd naar het gewicht van de in de gevonden teksten aanwezige zoektermen, hetgeen berekend wordt uit de positie in de tekst (hoe verder naar voren, hoe zwaarder) en uit de frequentie waarmee die termen in de betreffende tekst voorkomen. Verder worden standaard alle zoektermen automatisch getrunceerd, waarbij een gevonden woord nog weer een lager gewicht krijgt naarmate het meer verschilt van (dus langer is dan) de ingetikte zoekterm. Dat trunceren kan gelukkig wel voorkomen worden door een zoekterm met een punt af te sluiten. Formeel gebruik van AND, OR en afstands-operatoren ontbreekt op dit moment nog aan het systeem, maar er wordt aangekondigd dat daar wel aan gewerkt wordt. Wel kan met een minteken voor een zoekterm aangegeven worden dat Web-pagina's met die term een lagere score in de ranking moeten krijgen. Dat is dus geen echte NOT-operator, maar het zorgt alleen dat de betreffende documenten in de sortering meer naar achteren komen. Als resultaat van een zoekactie worden gegevens over de gevonden pagina's in een opmerkelijk volledige presentatie getoond. Niet alleen de URL en eventuele pagina-headers of -titels, maar ook een flink deel van de oorspronkelijke tekst zelf - de eerste 20 regels daarvan -, alsmede een rijtje van door de computer bepaalde 'zwaarste' woorden uit de tekst, worden in compacte opmaak op het scherm getoond. Op die manier passen niet meer gegevens dan van één of twee gevonden pagina's tegelijk op het scherm. De totale lengte van de lijst - en dus de hoeveelheid vanuit het Lycos-systeem te versturen gegevens - blijft tot een (vooraf in te stellen) maximum aantal hits beperkt. De getoonde URL's van alle gevonden gegevens doen bovendien dienst als directe hyperlink naar de originele Web-pagina's. De gebruiker heeft wel nog de keuze tussen een tamelijk 'kaal' zoekscherm met slechts een enkele korte opdrachtregel (waarin overigens desgewenst een heleboel termen ingetikt kunnen worden aangezien hij horizontaal door-scrollt) en een

iets uitgebreider formulier-zoekscherm. In dat laatste scherm kan aangegeven worden dat een ander aantal dan de eerste tien resultaten van een zoekactie getoond moet worden en dat eventueel minder gegevens van de gevonden pagina's getoond moeten worden (alleen url, rank-score, pagina-grootte, aantal aanwezige links en pagina-titel).

Op dit moment zijn er drie zoekmachines naast elkaar actief, waaruit de gebruiker vanuit de Lycos homepage kiezen kan: Lycos1, Lycos2 en Lycos3. Aanvankelijk werd niet duidelijk of daar nog verschillen tussen zijn. Hoewel de interfaces van die drie systemen hetzelfde zijn, werd de indruk gewekt dat er toch verschillen zijn, aangezien Lycos1 de tekst 'lycos small catalog', Lycos2 de tekst 'lycos big catalog' en Lycos3 de tekst 'lycos small database' draagt. In de praktijk bleken echter in alle drie systemen de volle 1,5 miljoen pagina's ter beschikking te staan, en leidden zoekacties tot exact dezelfde resultaten. De verschillende systemen draaien alleen op drie afzonderlijke Sparc machines.

Een bezwaar van het huidige systeem is dat sommige zoektermen zoveel opleveren (zeker als ook nog automatisch getrunceerd wordt), dat de zoekactie niet altijd afgemaakt kan worden. Door de impliciete OR-relatie die bij het best-match zoeken wordt toegepast en het ontbreken van de mogelijkheid AND-relaties toe te passen, is dat probleem niet altijd simpel op te lossen. Een ander probleem vormt de toenemende populariteit van Lycos (waaraan dit stukje helaas nog ook weer wat zou kunnen bijdragen) die maakt dat het systeem niet altijd aan uitvoering van je zoekvraag toekomt, omdat er al te veel andere gebruikers kunnen zijn. (Overigens wordt die bekendheid ook nog weer wat verder bevorderd doordat Lycos, nog geen week nadat ik het zelf 'ontdekt' had, ook al in het Cultureel (!) Supplement van NRC-Handelsblad werd beschreven). Bij zulke capaciteitsproblemen kun je overigens wel altijd nog naar één van beide andere Lycos-servers uitwijken.

Ondanks enkele kleine minpuntjes en hoewel in sommige opzichten wellicht nog altijd wat primitief in de ogen van de doorgewinterde online retrieval specialist, moet ik toch concluderen dat Lycos ongekende zoekmogelijkheden biedt om in een zo ongestructureerd en organisch groeiend geheel als het World Wide Web wellicht waardevolle informatie terug te kunnen vinden.

Afgesloten februari 1995.