Zoeken op WWW: Lycos revisited, Lycos NlightN'd en de WWW Worm

Wie Weet Welke Wellicht Waardevolle Wijsheid Waar op het Web op ons Wacht deel 2

Eric Sieverts

De snelle veranderingen bij Lycos geven alweer aanleiding in de serie Wie Weet ... nog even op dat systeem terug te komen alvorens www Worm te bespreken. Het overzicht van zoekmachines is geactualiseerd en uitgebreid.

In de eerste bijdrage aan deze serie heb ik het verschijnsel www-zoekmachine geïntroduceerd. Daarnaast ben ik uitgebreider ingegaan op één van die zoekmachines, namelijk LYCOS, en heb ik aan de hand daarvan verteld op welke manier veel van die machines via een soort robots hun informatie verzamelen. Bovendien gaf ik in die aflevering al een overzicht van adressen van andere zoekmachines die ik tot dat moment had achterhaald en die hopelijk later nog eens aan de orde zouden komen. Hopelijk kunnen we zo uiteindelijk tot een evaluerende vergelijking met plus- en minpunten van al die systemen komen.
Uit deze tweede bijdrage blijkt (voor de regelmatige Internet-gebruiker niet verrassend) wat een snel evoluerend en veranderend fenomeen het World Wide Web is. Recente veranderingen bij Lycos geven namelijk al weer aanleiding nog even op dat systeem terug te komen. Verder blijkt een aantal adressen (URL's) uit het vorige overzicht van zoekmachines intussen ook veranderd te zijn. Bovendien kan nog een viertal nieuwe machines aan de lijst toegevoegd worden. Al met al een goede reden om de bijgewerkte lijst in deze aflevering opnieuw in zijn geheel op te nemen. Overigens begin ik me wel af te vragen of deze serie ooit zal eindigen, als bij elke aflevering waarin niet meer dan drie of vier zoekmachines uitgebreid aan de orde kunnen komen, ook weer vier nieuwe aan de verzamellijst toegevoegd worden!
De andere deze keer te bespreken systemen zijn een 'nieuwe' en een 'oude'. De nieuwe, NlightN, wordt deze keer maar meteen opgenomen omdat daarmee twee ontwikkelingen geïllustreerd worden. Dit is in de eerste plaats een voorbeeld van een voor gratis gebruik opgezet zoeksysteem (het al besproken Lycos; helemaal nieuw is hij dus ook weer niet) dat nu door een commercile organisatie, deels tegen betaling,

Dr. E.G. Sieverts is docent aan de Opleiding Informatiedienstverlening en -management (IDM) van de Hogeschool van Amsterdam. E-mail: e.g.sieverts@fei.hva.nl.

wordt toegepast. In de tweede plaats wordt daarbij niet alleen informatie van www-pagina's aangeboden, maar worden ook al veel langer bestaande commerciële databases in één keer toegankelijk gemaakt. De oude zoekmachine die dit keer besproken wordt is de intussen al klassieke www Worm. Ook die heeft echter al weer een belangrijke face-lift ondergaan. Daarbij wordt meteen geïllustreerd hoe bij het zoeken in www soms gebruik gemaakt kan worden van de structuur die HTML-documenten hebben, als alternatief voor in databases gebruikelijke veldstructuren. Wellicht ten overvloede zij nog vermeld dat ik bij mijn beschrijvingen uitga van de mogelijkheden die bij gebruik van een grafische www-browser als Mosaic of Netscape geboden worden. Zelf gebruik ik Netscape.

Nogmaals Lycos

Allereerst kan gemeld worden dat de Lycos database intussen (begin november 1995) al ca. 11 miljoen Webpagina's bijna full-text geïndexeerd heeft. En dat zou (naar eigen zeggen) ongeveer 91% van het web zijn, ongeveer vier keer zo veel als '*de nummer twee*'. Het woord 'bijna' uit de eerste regel vereist nog enige nadere uitleg. In tegenstelling tot wat ik vorige keer schreef, blijkt Lycos toch niet echt de volledige tekst te indexeren, maar alleen de eerste circa 300 woorden (ongeveer 20 regels) uit elk www-document, plus aanvullend nog maximaal 100 andere woorden die het programma op statistische gronden voor dat document het meest kenmerkend acht. In de meeste gevallen is dat geen ernstige beperking. Alleen als echte artikelen en boeken full-text worden aangeboden, zal niet de volledige tekst afzoekbaar zijn. Dat betekent dat je bij een zoekactie op '*bolshevism* AND *socrates*' niet het vierde hoofdstuk van Lady Chatterley's Lover (waarvan elk hoofdstuk een afzonderlijk www-document vormt) zult terugvinden, hoewel die beide woorden daarin wel voorkomen.[1]

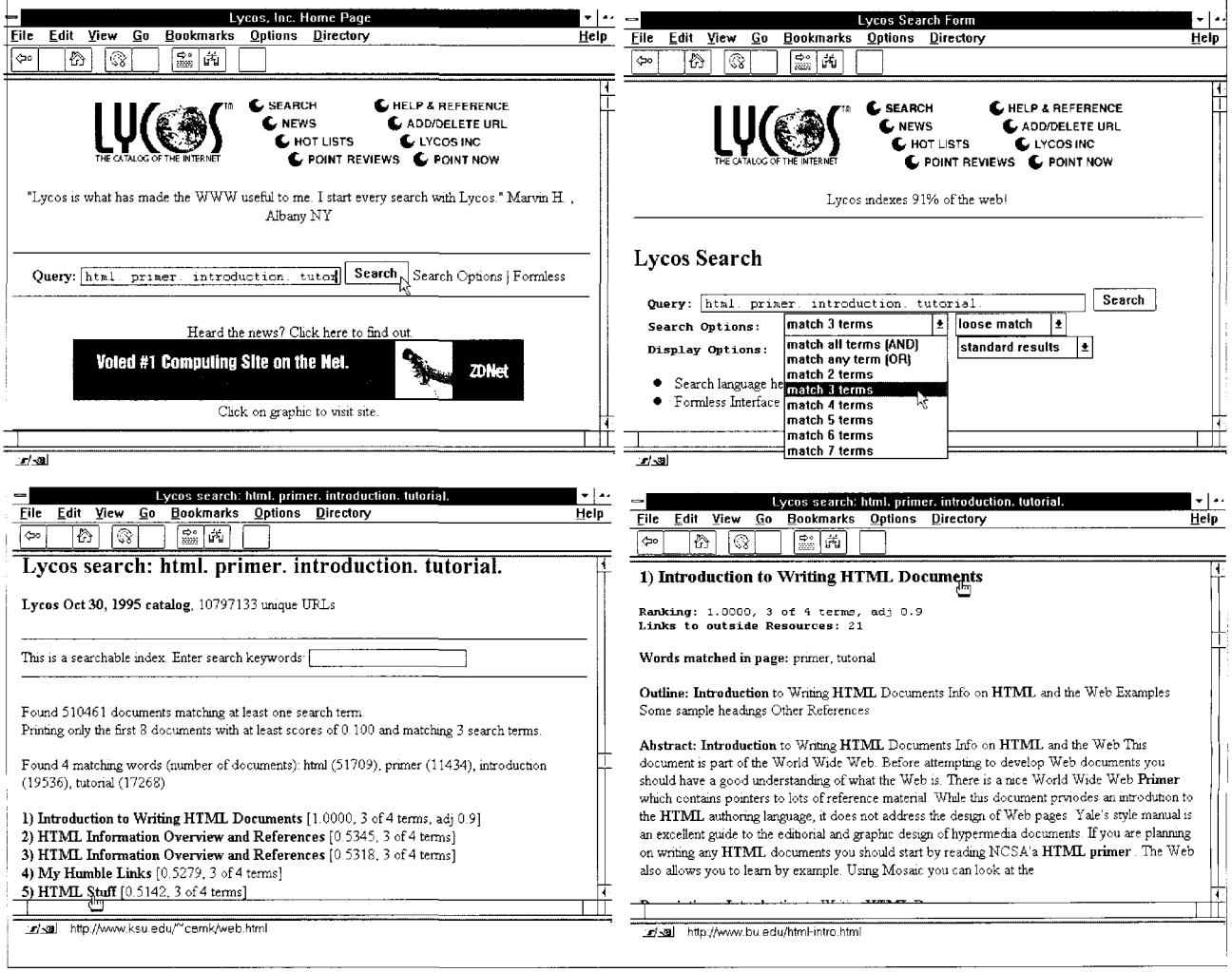
De keuze tussen de verschillende beschikbare Lycos-servers, die in de vorige aflevering gemeld werd, is intussen voor de gebruiker aan het oog onttrokken; de verdeling van de zoekcapaciteit over de ter beschikking staande computers vindt nu automatisch plaats. De gebruiker kan verder nog altijd kiezen tussen een simpele zoekregel, zonder verdere poespas, en een formulier waarin je meer kunt specificeren. Sinds de vorige bespreking is aan de simpele zoekregel niets veranderd, afgezien van een wijziging van de grafische vormgeving - er is geen enge loopspin meer te zien, wel een regelmatig wisselende advertentie. Het zoekformulier heeft echter al weer twee maal ingrijpende wijzigingen en uitbreidingen ondergaan, overigens zonder de onderliggende zoektechniek te beïnvloeden. Het wordt opgevraagd vanuit het beginscherm door te klikken op 'Search Options' (http://www.lycos.com/cgi-bin/nph-randurl/cgi-bin/largehostform1.html). In dat scherm kunnen met pull-down menu's vier keuzes ingesteld worden; twee voor 'Search Options' en twee voor 'Display Options'.

Vooral de eerste Search Option is een interessante uitbreiding. Daarbij kun je nu aangeven hoeveel van de als zoekvraag ingetikte termen minimaal in de gevonden documenten moeten voorkomen. Als in de zoekregel een reeks van 5 zoektermen is ingetikt, kun je zo bijvoorbeeld eisen dat in de te vinden documenten tenminste 3 van die termen moeten voorkomen. Binnen een best-match zoekmechanisme zoals dat van Lycos, dat in feite altijd een OR-relatie toepast voordat de relevantiescores berekend en de resultaten in volgorde gezet worden, biedt dat de mogelijkheid om toch een soort globale AND-eis op te leggen. Naast de keuze dat minstens 2, 3, 4, 5, 6 of 7 woorden uit de reeks ingetikte zoektermen moeten voorkomen, kan ook worden aangegeven dat ze allemaal moeten voorkomen (*match all terms*, dus een volledige AND-relatie) of dat minstens één term moet voorkomen (*match any term*, dus een OR-relatie). In al deze gevallen blijft de best-match berekening voor bepaling van de relevantie volgorde gewoon uitgevoerd worden. Alleen wordt bij een strengere eis de reeks te tonen documenten eerder afgekapt en daarmee het aantal door de computer te verwerken documenten te voren al ingeperkt. Wat deze laatste uitbreiding nog niet biedt, is de mogelijkheid aan te geven dat één bepaald woord uit de reeks zoektermen beslist moet voorkomen.

In de tweede Search Option kun je aangeven of je een *loose match*, een *fair match*, een *good*, een *close* of zelfs een *strong match* wilt hebben. Dit heeft gedeeltelijk hetzelfde effect als de voorgaande optie, zij het dat nu alleen gekeken wordt naar de door het systeem berekende relevantie-scores van gevonden www-documenten. De manier waarop die globaal berekend worden, is in de vorige bijdrage al beschreven. Deze relevantiegraad kan maximaal de waarde 1 hebben voor documenten waarin alle ingetikte termen exact zo in het begin van het document voorkomen. De bij deze optie gemaakte keuze bepaalt bij welke waarde van de relevantiegraad de reeks te tonen documenten wordt afgebroken. Daarmee kun je dus al te lange reeksen, aan het eind weinig relevante zoekresultaten voorkomen.

Met de Diplay Options kan het per 'pagina' (maximaal) te tonen aantal resultaten ingesteld worden op 10, 20, 30 of 40. Overigens kunnen altijd weer vervolgpagina's opgeroepen worden, zolang er nog meer te tonen valt. Daarnaast kun je *summary*, *standard* of *detailed* presentatie van de zoekresultaten kiezen. *Summary* geeft maar één regel (met URL en be-

Figuren 1-4
Voorbeelden van zoek- en resultaatschermen van Lycos. Zowel het eenvoudige zoekscherm als het uitgebreider zoekformulier worden getoond. In de korte resultaat-presentatie wordt een enkele regel per gevonden document getoond, met alleen aanklikbare titels, hun berekende relevantiegraad en hoeveel van de gevraagde zoektermen erin voorkomen. De uitgebreide presentatie bevat bovendien 'outlines' en abstracts van de gevonden pagina's die tegelijk met het indexeren door de computer gegenereerd zijn.



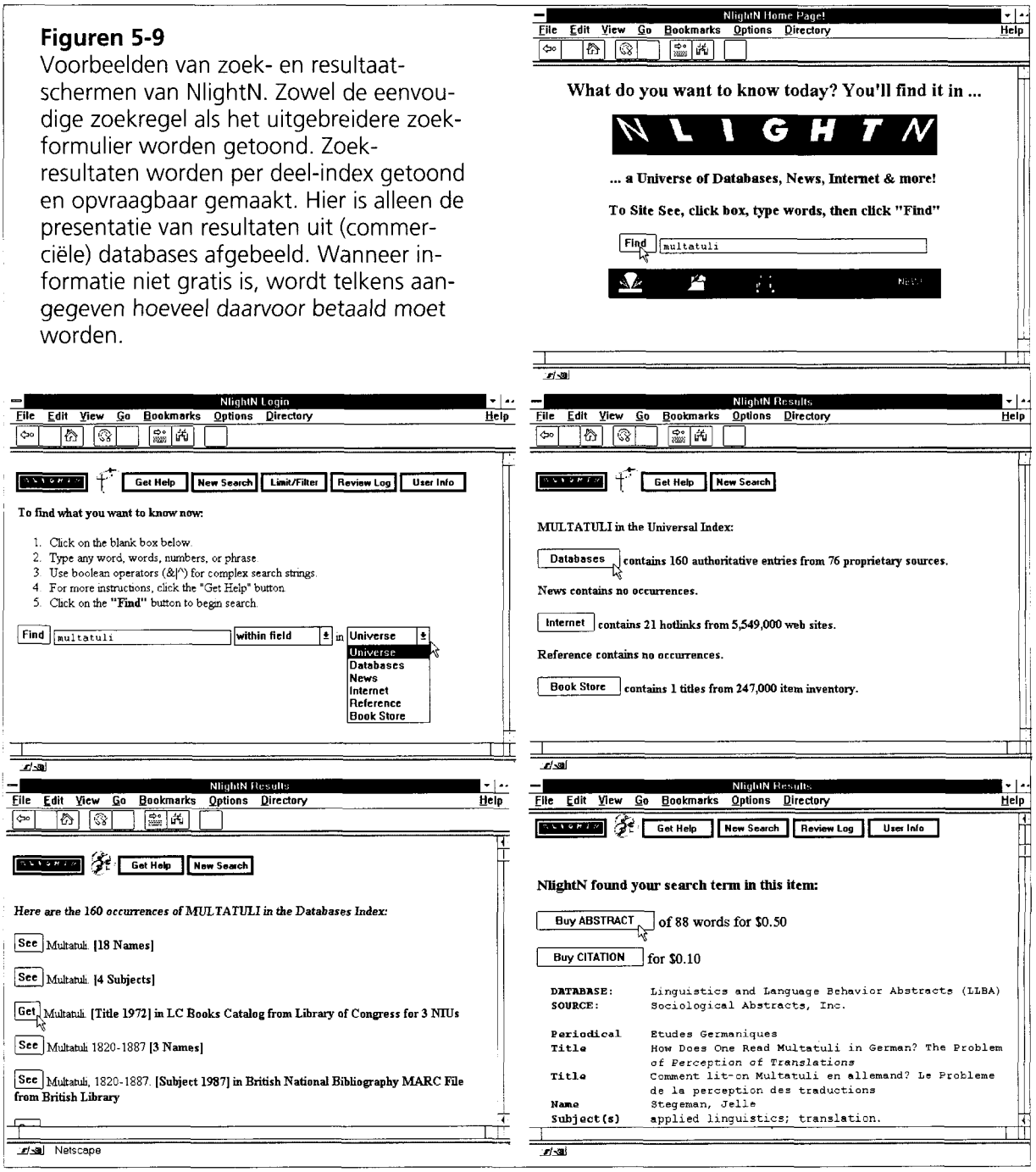
rekende relevantiegraad) per gevonden document, *standard* en *detailed* geven ook een beperkte weergave van de inhoud zelf.
Ook nieuw bij Lycos is tenslotte dat naast de gerichte zoekmogelijkheden, een aantal categorieën met zorgvuldig voorgeselecteerde verwijzingen is opgezet, waarop zonder zoeken direct geklikt kan worden. Daaronder zijn bijvoorbeeld Business, News, Reference and Weather. Op dit verschijnsel, dat we ook bij andere zoekpagina's wel tegenkomen, zal ik in een volgende aflevering wellicht nog terugkomen.
Ten slotte is nog een kritische noot op zijn plaats. In de nogal reclame-achtige teksten die Lycos als antwoorden op Frequently Asked Questions op het net gezet heeft, wordt hoog opgegeven over het voortdurend up-to-date houden van de gegevens in de index. Daarbij wordt trots vermeld dat de Lycos-robot wel 50.000 Web-pagina's per dag bezoekt om ze te indexeren, te herindexeren of uit de index te verwijderen. Als dat getal echt klopt, leert een simpele berekening dat met een huidige omvang van 10 miljoen geïndexeerde pagina's, elke pagina hoogstens eens per 200 dagen opnieuw door de robot bezocht wordt. De nochtans getrokken conclusie '*so the Lycos catalog is never outdated*' mist dan dus elke grond.

NlightN

De kwaliteiten van het Lycos zoeksysteem blijken nogal wat aandacht getrokken te hebben. Dat blijkt in de eerste plaats uit het feit dat het grote MicroSoft een licentie genomen heeft om de software voor de ontsluiting van zijn MicroSoft-network te gaan gebruiken. Een andere organisatie met een licentie op de Lycos-software is 'The Library Corporation'. Deze

heeft de software (van buitenaf voor de gebruiker niet meer herkenbaar) gebruikt om zijn NlightN systeem op te zetten.
In NlightN worden echter niet alleen www-documenten toegankelijk gemaakt, maar daarnaast ook allerlei andere databases en informatiesystemen. Het is dan ook een commercieel opgezet systeem, waar de gebruiker voor gevonden informatie moet betalen. Dat is voor het www nog een beetje een nieuwe, maar wel steeds gebruikelijker ontwikkeling. Overigens wil dat nog niet zeggen dat zonder een rekening bij NlightN helemaal niet gezocht zou kunnen worden. De enige beperking is dat je zonder betaling niet alle details van de gevonden informatie getoond krijgt, zodat het vaak niet mogelijk is de betreffende documenten - in elektronische of papieren vorm - echt te lokaliseren. Zonder een rekening te hebben, kan ik uit de praktijk dus toch iets over de werking vertellen.
Het principe van NlightN is dat er een soort superindex gemaakt wordt op allerlei informatiebronnen. Bij het zoeken kan die hele index doorzocht worden, maar wordt vervolgens een per deel-index gespecificeerd zoekresultaat vermeld. Zo kan de gebruiker zelf bepalen uit welke groepen informatiebronnen hij de gegevens ook echt wil zien. Dit zelfde principe wordt overigens ook al enige tijd, zij het nog op veel kleinere schaal, toegepast in het in de vorige tabel al genoemde Utrechtse W5 systeem, waarop ik in een volgende aflevering nog eens hoop terug te komen. Een duidelijk voorbeeld van onafhankelijke parallelle ontwikkelingen.
De 'index op alles' of 'Universal Index' van NlightN bevat op dit moment vijf deelindexen voor vijf soorten informatiebronnen, namelijk 'internet', 'databases', 'persberichten', 'desktop reference' en een 'book store'. Het *internet*-deel van de index is gewoon de Lycos www-index, zij het in een wat ander jasje. Het *databases*-deel bevat een groot aantal databases die professionele informatiespecialisten ook al kennen van hun gewone online host-organisaties. Dat zijn enkele honderden bibliografische zowel als full-text bestanden. Daaronder zijn klassiekers als Medline, Psycinfo, ABI-Inform, Metadex, Life Sciences Collection, Pollution Abstracts, US Patents en Findex, maar ook wat populairder bronnen als Showbiz Today en CNN Newsroom. In het *news*-deel vinden we persberichten uit onder meer Business-wire, Deutsche Presse-Agentur, Itar-Tass, US Newswire en Inter Press Service. Over de *desktop reference* en de *book store* wordt online op dit moment nog geen verdere informatie verstrekt.
Als niet-abonnee (je hoort dan nog niet tot de *NlightN'd people*) kun je direct gaan zoeken, maar overal waar voor informatie betaald moet worden, kunnen geen verdere (noodzakelijke) details opgevraagd worden. Je kunt je daarnaast voor gratis testgebruik laten registreren. Dan krijg je een budget van 10 NIU te besteden. Die zijn $0,10 per stuk waard; waar NIU de afkorting voor is wordt evenwel niet duidelijk gemaakt. Dit budget kun je alleen gebruiken voor het 'kopen' van informatie die ook in NIU's geprijsd is. In gewone dollars geprijsde informatie kun je alleen kopen als je echt via creditcard of anderszins een bedrag hebt overgemaakt. Toch kan je ook met die NIU's al een wat beter idee van de aangeboden informatie krijgen.
Bij de standaard (eenvoudige) zoekactie vanuit de NlightN Home-page kan een zoekwoord of een string van zoekwoorden in een zoekregel ingetikt worden. Bij een string wordt dan echt op die woorden in die volgorde, naast elkaar in de tekst gezocht. Het antwoord op de zoekvraag bestaat uit een scherm met de subresultaten (aantallen hits) in elk van de vijf deelindexen. Bij elke categorie die iets oplevert, bevat het scherm een drukknop die het mogelijk maakt die resultaten nader te bekijken. Je krijgt dan een gedetailleerder lijst met resultaten voor die categorie. Hoe het er dan verder uitziet hangt van de gekozen categorie af.
Bij de Internet-gegevens moest je aanvankelijk via een 'hotlink'-drukknop $0,10 betalen om aanklikbare URL's van gevonden web-documenten op het scherm te krijgen. Sinds enige tijd worden, kennelijk uit concurrentieoverwegingen, via 'See'-drukknoppen nu wel gratis aanklikbaar URL's getoond. Dat gaat meestal via een extra tussenstap waarin achter extra 'See'-drukknoppen telkens meer Web-pagina's zijn samengenomen, waarin de gevraagde zoekterm in eenzelfde documenttitel of in eenzelfde woordstring in de tekst voorkomt.
Resultaatlijsten uit de (echte) Databases bevatten rechtstreekse 'Get'-knoppen bij index-ingangen die de zoekterm bevatten en maar één hit opleveren. Indexingangen die meer hits opleveren bevatten 'See'-knoppen waarmee je lijstjes kunt oproepen met de daarbij horende titels en 'Get'-knoppen om die individuele resultaten (één voor één) op te kunnen vragen. Om resultaten achter 'Get'-knoppen te zien te krijgen, moet betaald worden. Voor bedragen van 2 of 3 NIU per hit worden titel, auteurs, tijdschriftnaam en trefwoorden gegeven. Volledige bibliografische gegevens (met jaargang, volume, pagina's e.d.) moeten voor een extra bedrag van meestal $0,10 'gekocht' worden. Eventueel beschikbare abstracts moeten apart betaald worden. Bedragen daarvoor hangen af van de database waaruit het komt. Abstracts uit ABI/Inform kosten bijvoorbeeld $2,00 per stuk.
Bij de deelindex met persberichten krijg je gratis al een lijstje met 'headlines'. Met een 'Get'-drukknop krijg je voor 1 NIU ook de naam van de persdienst en de eerste 2 regels van het bericht te zien. Berichten zelf moeten vervolgens met een 'Buy Citation'-knop

Figuren 5-9
Voorbeelden van zoek- en resultaatschermen van NlightN. Zowel de eenvoudige zoekregel als het uitgebreidere zoekformulier worden getoond. Zoekresultaten worden per deel-index getoond en opvraagbaar gemaakt. Hier is alleen de presentatie van resultaten uit (commerciële) databases afgebeeld. Wanneer informatie niet gratis is, wordt telkens aangegeven hoeveel daarvoor betaald moet worden.



voor $0,10 per stuk gekocht worden. De *reference* collectie levert een lijst met vindplaatsen uit een grote collectie naslagwerken. Teksten (artikelen of definities) kun je voor bedragen van $0,10 of $1,00 kopen. In de *book store*, tenslotte, is de informatie gratis. Die is echter vooral bedoeld om meteen boeken te kunnen bestellen (gewoon de papieren folio-produkten), waarvoor dan (uiteraard) ook betaald moet worden. Geregistreerde gebruikers (ook de gratis test-gebruikers) komen in een zoekformulier met wat uitgebreider mogelijkheden. Ook anderen kunnen daar vanuit de NlightN Home page komen door te kiezen voor *Site* (niet Sight!) *Seeing*. In dat zoekformulier kun je in de zoekregel rechtstreeks Booleaanse combinaties intikken (met & voor AND, | voor OR en ^ voor NOT). Naast de zoekregel is verder een venstertje waarin je kunt aanklikken of het hele 'Universe' (dus alle vijf deelindexen) doorzocht moet worden of dat direct één daarvan gekozen wordt. Een derde venstertje biedt de wat kryptische keuze tussen '*within field*', '*alphabetically*' en '*across fields*'. Dat eerste blijkt te betekenen dat direct op de exacte zoektermen (in alle velden) gezocht wordt. Bij de tweede mogelijkheid wordt ook gezocht op termen die in de index 'alfabetisch' op de ingetikte zoekterm volgen, waarbij je net zo lang door de resultaten (per term) verder kunt bladeren als je wilt. Over de betekenis van de (nieuwe) derde mogelijkheid hult het NlightN hulpscherm zich helaas in stilzwijgen. De praktijk leert dat het (aanzienlijk) minder oplevert.

Voor wat betreft de www-informatie biedt NlightN dus absoluut niet meer dan wat ook al (gratis) met Lycos gevonden kan worden, en dan bovendien nog veel onslachtiger. De interessante ontwikkeling is echter de integratie met andere informatiebronnen. Gebruikers kunnen zo makkelijk aan enige informatie komen, over welk onderwerp dan ook, waarbij natuurlijk vooral op de eindgebruiker gemikt wordt. Ik zeg hier heel bewust 'enige', want ook voor die eindgebruiker zal het lang niet altijd helder zijn hoe hij echt er (vooral) volledig aan gewenste informatie moet komen. Die zal zich namelijk nauwelijks bewust zijn van hetgeen zo gemist wordt, bijvoorbeeld doordat niet de juiste zoektermen gebruikt worden.
Op zijn beurt zal de echte informatiespecialist al snel alle verfijnde zoekmogelijkheden missen, die gewone online hosts bieden, met veld-specifiek zoeken, online thesauri, limitering, zoom of rank-commando's en wat al niet meer. Dat is namelijk allemaal ingeleverd om het voor de eindgebruiker maar 'makkelijk' te maken. Voordat je eindelijk een redelijk aantal artikelen opgevraagd of bekeken hebt, moet bovendien een enorm aantal keren op knoppen 'geklikt' worden, voor elk afzonderlijk artikel opnieuw, en moet evenzovele keren op het opsturen van een volgend scherm gewacht worden. Vooral daardoor zal een wat diepgravender NlightN zoekactie heel wat langer duren dan een zoekactie bij een klassieke host.
Op dit moment blijkt NlightN nog sterk in ontwikkeling te zijn. Dat is natuurlijk prima, maar het is wel wat verwarrend dat elke paar weken weer deels andere mogelijkheden en bronnen aangeboden worden. Bovendien houdt de online uitleg daar absoluut geen gelijke tred mee. Hoewel deze tekst over NlightN al weer enkele keren aan de actuele situatie is aangepast, dient de geïnteresseerde lezer dus zelf nog goed te kijken hoe het systeem er op dat moment weer uitziet.

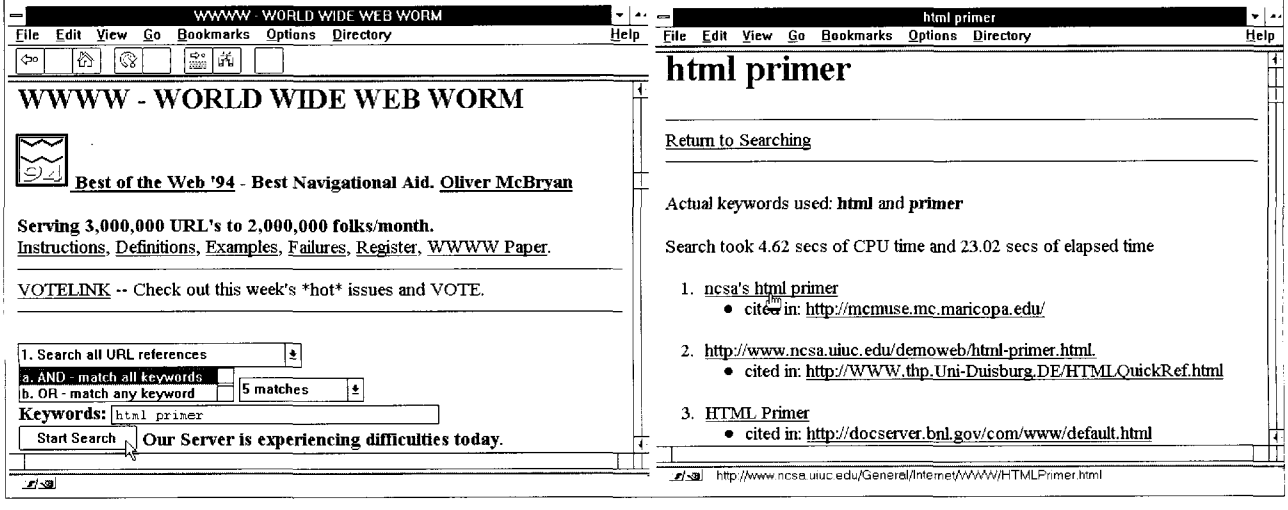
Zoeken met de WWW Worm

De www Worm (ook wel wwww) is één van de klassieke zoekmachines op www. Dit keer geen hardlopende spin (Lycos) die achter de informatie aan gaat, maar een kronkelende worm die zich langzaam maar gestaag door de rijstebrijberg van het www heen eet en alle gegevens die hij daarbij tegenkomt in zijn darmkanaal verteert tot voor ons doorzoekbare brokken. (Tot wat een melige beeldspraak inspireren al die metaforen uit de Internetwereld toch). Volgens de verschafte gegevens waren juni 1995 gegevens van ongeveer 3.000.000 www-pagina's doorzoekbaar, gegevens die begin november nog niet gewijzigd waren. Daarbij beperkt men zich overigens tot alleen die www-documenten waarheen via een hyperlink vanuit andere documenten verwezen wordt. Ook gezien de hierna te bespreken zoekmogelijkheden, zou je de www Worm daarom een soort van citatie-index kunnen noemen. Verder is van al die pagina's maar een heel beperkt deel van de tekst geïndexeerd.
Aanvankelijk bood de Worm weinig gebruikersvriendelijke zoekmogelijkheden, maar daar is nu iets aan gedaan. Zelfs wordt nu een aantal verschillende mogelijkheden geboden. In het beginscherm dat je middels het in de tabel gegeven URL te zien krijgt, kom je in een 'snel'-zoekmogelijkheid terecht, met een aanklikbare keuze tussen zoeken met AND of met OR. Dat houdt in dat tussen de woorden die - gewoon achter elkaar - in de zoekregel ingevuld worden, hetzij een AND-, hetzij een OR-relatie gelegd wordt; een combinatie van die twee in één zoekopdracht is dus niet mogelijk. Zonder dat dat expliciet gezegd wordt, blijkt dat de ingetikte woorden automatisch links én rechts worden getrunceerd. RECHT levert dus ook UTRECHT. Verder kan in een rolluikje aangegeven worden of maximaal 1, 5, 50, 500 of 5000 (!) documenten van het zoekresultaat getoond moeten worden.
Daarnaast zijn langzamer maar flexibeler zoekacties mogelijk met gebruik van zogenaamde 'reguliere expressies', waarvoor de in veel UNIX-systemen standaard aanwezige utility EGREP gebruikt wordt. In je zoekterm kun je daarmee onder meer reeksen tekens met variabele gegevens maar met vaste patronen weergeven. Voor informatici schijnt het werken hiermee gesneden koek te zijn, maar voor argeloze andere gebruikers is dit een zoektaaltje dat eerst nog wel enige oefening vergt. Deze zoekmogelijkheid kom je overigens alleen nog tegen in een vervolgscherm 'wwwWintro.html' waarin een (oude) introductie van het systeem en een aantal illustratieve voorbeelden gegeven worden.
Verder laat de Worm je kiezen in welke onderdelen van www-documenten je wilt zoeken. Die mogelijkheid is er omdat ook www-documenten een soort rudimentaire velden-structuur hebben. Deze structuur is in HTML (de in www gebruikte HyperText Mark-up Language) standaard voorgeschreven en gemarkeerd met vaste HTML begin- en eind-codes. Zo kent elk www-document een titel (begincode *<title>*, eindcode *</title>*) die geen onderdeel uitmaakt van de op het scherm te tonen document-tekst zelf. Wel verschijnt die titel bij gebruik van Netscape in de blauwe vensterbalk bovenaan het scherm. Daarnaast is er een, overigens niet verplichte, opdeling van het document zelf in 'head' (*<head>*) en 'body' (*<body>*). In de praktijk worden de *<head>*-codes meestal genest rondom of binnen het *<title>*-deel gezet, zodat die kop evenmin in de document-tekst zelf terecht komt. De *<body>* is dan de rest van het document. Daarin kunnen eventueel nog speciale elementen aan hun HTML-codes herkend worden. Zo is er een aantal ni-

veaus van hoofdstuk en paragraaf-titels, zogenaamde 'headings' die met *<h1>* tot en met *<h6>* worden aangegeven. Verder zijn uiteraard ook de hyperlinks naar andere documenten te herkennen aan codes in de brontekst. Het adres (URL) waarheen gelinkt wordt, maakt onderdeel van die code zelf uit; tussen de begin- en eindcode van een hyperlink staat het stukje tekst in het document dat op het scherm als link oplicht. Bij gebruik van Mosaic of Netscape wordt die link-tekst in blauw weergegeven. Daarnaast heeft elk www-document natuurlijk ook nog zelf een URL waarop het teruggezocht zou kunnen worden.
Van enkele van deze onderdelen kan bij het zoeken met de www Worm gebruik gemaakt worden. Daartoe biedt de Worm een viertal keuzes. Zo kun je kiezen voor: 'Search only in document titles'. Daarmee zoek je alleen in de (tekst van) de titels van de www-documenten. Alle overige tekst (de 'body') wordt dus niet doorzocht, ook niet eventueel daarin voorkomende hoofdstuk- en paragraaf-koppen. Een andere mogelijkheid is te zoeken op de adressen, de URL's van te vinden documenten of op onderdelen daarvan. Dit wordt aangegeven met 'Search only in document addresses'. Aangezien alle punten en slashes daarin door de Worm als woordscheiders worden beschouwd, kan op deze manier makkelijk naar een www-document gezocht worden waarvan bijvoorbeeld wel de filenaam, maar niet het computer-adres bekend is. Via AND-combinaties biedt dit allerlei mogelijkheden. Een beperking waar je dan wel tegenaan loopt, is dat alleen woorden van minimaal drie letters geïndexeerd worden, zodat je met de standaard zoekmogelijkheid bijvoorbeeld niet op de tweeletterige landencodes uit Internet-adressen kunt zoeken. Echte experts zullen daarom graag van de al genoemde reguliere expressies gebruik maken, waarbij dit wel kan.

Bij beide zoekmogelijkheden is het resultaat een lijst van pagina-titels waaruit een gewenste direct aangeklikt kan worden.
De overige twee keuzes maken het mogelijk specifiek te zoeken in de hyperlinks die in de documenten voorkomen. Met 'Search all URL references' wordt gezocht in de teksten die in hyperlinks gebruikt worden, dus in de op het scherm blauw oplichtende woorden. De gedachte hierachter is dat de hyperlink verwijzingen in een tekst meer over de inhoud van die tekst zeggen dan zo maar willekeurige woorden daaruit. Dat si dus en soortgelijke filosofie als die achter de klassieke citatie-indexen. Het citatie-indexachtige aspect van de Worm manifesteert zich ook door het feit dat het zoekresultaat uit verwijzingen naar paren documenten bestaat, telkens zowel het document van waaruit gelinkt wordt als dat waarheen gelinkt wordt; dus als het ware het citerende en het geciteerde document samen. 'Search all URL adresses' tenslotte biedt de mogelijkheid om op dezelfde wijze te zoeken via de URL's die in de hyperlinks gebruikt worden. Net als eerder bij het directe zoeken, moet je dan dus op zijn minst al een gedeelte van een interessant URL kennen. Overigens bleken zoekresultaten bij deze beide laatste manieren van zoeken ontzettend veel (verwijzingen naar) plaatjes te bevatten. Vermoedelijk komt dat doordat alle links naar plaatjes (die in een grafische browser als Netscape meteen al *in* de tekst worden afgebeeld, maar in feite gelinkt zijn) ook als echte hyperlinks meegeïndexeerd worden. En in www zijn er natuurlijk steeds meer documenten die miegelen van zulke plaatjes, iconen, speciale grafische balkjes, portretten enzovoort. In de praktijk komen er daardoor soms maar weinig 'gewone' links naar tekstdocumenten in je zoekresultaten voor.
In die zoekresultaten wordt van de citerende docu-

Figuren 10-11
Voorbeelden van schermen van de WWW Worm. Bij het zoeken in hyperlink-teksten uit WWW-pagina's ('URL references' als zoekveld), bevat het zoekresultaat aanklikbare links naar telkens zowel het verwijzende document als dat waarheen verwezen wordt.



menten alleen het URL (direct aanklikbaar) in de lijst met zoekresultaten vermeld. Als het geciteerde document een echt tekst-document is, wordt dat (ook direct aanklikbaar) in de lijst gerepresenteerd door de linktekst. Als een plaatje of icoontje deel uitmaakt van de linktekst, wordt dat daar ook bij getoond (ook al is het - net als de linktekst zelf - in feite uit het citerende document afkomstig). Bij verwijzing naar niet-HTML files (plaatjes, geluid e.d.) is het niet altijd duidelijk wat er in de lijst gezet wordt, soms schijnt dat het laatste stukje van het geciteerde URL - d.w.z. de filenaam - te zijn, soms ook teksten die meer op document-titels of link-teksten lijken.

In de in een introductie-tekst gegeven uitleg van het systeem worden voor de vier hier genoemde zoekmogelijkheden helaas nog de oude benamingen gebruikt, respectievelijk: 'Search only in titles of citing documents', 'Search only in names of citing documents', 'Search all citation hypertext' en 'Search all names of cited URL's'. Ook verder bleek deze hulptekst nog niet overal aan het huidige zoekscherm aangepast te zijn.

In de zoekpraktijk blijkt de Worm redelijk te werken. Toch zal in veel praktijkgevallen het zoeken op alleen document-titels een te sterke beperking zijn. Vaak vindt je niets of krijg je door de automatische truncatie alleen maar ongewenste resultaten. De mogelijkheid specifiek op hyperlinks te zoeken is een aardige aanvulling, zij het ook maar met beperkte toepassing. Mijn standaard zoekvraag naar Multatuli of Max Havelaar (die in Lycos aardig wat opleverde) gaf hier in beide gevallen een nul-resultaat. Een argeloze (eind-) gebruiker zal dan al snel de onterechte conclusie trekken dat er *dus niets* over dat onderwerp in www te vinden is. Kennis van wat de zoekmachine precies doet en waarin hij eigenlijk zoekt, alsmede vergelijkingsmateriaal met andere zoekmachines zijn voor een goed zoekresultaat dus onontbeerlijk.

1. Aanvankelijk bleek Lady Chatterley's Lover (http://www.datatext.co.uk/library/dhl/chat/chapters.htm) overigens helemaal niet geïndexeerd te zijn, hoewel dat met Robinson Crusoe en andere op dezelfde DataText-server aangeboden klassieken uit de wereldliteratuur wel het geval was. Mijn achterdocht dat dit de kop opstekend Amerikaans puritanisme was, bleek gelukkig ongegrond, want intussen zijn ook alle hoofdstukken van Lady Chatterley (op termen uit het begin van die hoofdstukken) in Lycos terug te vinden.

Retrieval-systemen voor World-Wide-Web informatie

naam	*aanbieder*	*URL*
AliWeb	Nexor	http://web.nexor.co.uk/public/aliweb/search/doc/form.html
CUI W3 Catalog	Université Genève	http://cuiwww.unige.ch/w3catalog
EINet Galaxy	EINet/MCC	http://galaxy.einet.net/search.html
Harvest	Colorado University	http://harvest.cs.colorado.edu/
InfoSeek *	InfoSeek Corporation (Cal.)	http://www.infoseek.com:80/Home
JumpStation II ***	Stirling University (Scotland)	http://js.stir.ac.uk/jsbin/jsii
LYCOS	Carnegie Mellon University	http://lycos.cs.cmu.com/
NIKOS **	Rockwell Network Systems / California Polytechnic	http://www.rns.com/cgi-bin/nikos http://www.rns.com/cgi-bin/nomad
NlightN *	The Library Corporation	http://www.nlightn.com/
OpenText Web Index	Open Text Corporation / UUnet (Can.)	http://www.opentext.com/
RBSE	NASA	http://rbse.jsc.nasa.gov/eichmann/urlsearch.html
WS (Zoeken op Internet)	Universiteit Utrecht	http://pablo.ubu.ruu.nl/Ned/Internet.html http://pablo.ubu.ruu.nl/Ned/Zoeken.html
WebCrawler	University of Washington (Seattle)	http://www.webcrawler.com/WebCrawler/WebQuery.html
WWW Worm	University of Colorado	http://www.cs.colorado.edu/home/mcbryan/WWWW.html
Yahoo Search	Stanford University	http://www.yahoo.com/search.html

* voor volledige zoekmogelijkheden moet een rekening geopend worden
** vroegere naam 'WWW Nomad'; ook wel bekend als 'Zorbamatic'
*** wordt eind december 1995 beëindigd