Wouter Mettrop en Hans van der Laan

WWW-zoekmachines onderzocht

Naast incompetent zelfs onbetrouwbaar

WWW-zoekmachines zijn incompetent: ze bestrijken ieder slechts een klein deel van het web. De Werkgroep IRT onderzocht in hoeverre ze dan tenminste daar hun werk foutloos doen en bevond ze onbetrouwbaar.

NEE, DEZE BIBLIOTHEEK bezit *De ontdekking van de Hemel* van Harry Mulisch niet. U ziet dat boek immers niet in de catalogus. Maar het staat wel op de plank bij diens andere werken…? O, dat heeft de catalogus wel eens meer, zegt de uitleen. Meestal op vrijdagochtend, straks herinnert hij het zich wel weer. En ja hoor, maar nu ontbreekt er iets bij W.F. Hermans. Ondenkbaar? De Werkgroep IRT (zie kader) heeft in een langlopend onderzoek geconstateerd dat search engines op het web dat soort gedrag vertonen.
Zoekmachines worden vaak met elkaar vergeleken door te kijken naar hun grootte: hoeveel webpagina's hebben ze geïndexeerd. Die grootte zou significant zijn als een grote engine de zoeker ook altijd een groot aantal relevante documenten zou leveren. Maar naast omvang zijn er ook andere factoren die meespelen. Factoren die te maken hebben met de manier waarop zoekmachines hun verzameling webpagina's indexeren, en met de effectiviteit van hun retrieval-mechanisme. Wij onderzoeken hier twee van zulke factoren: het *indexeergedrag* en *retrieval-gedrag:*

- indexeergedrag: welke elementen van een webpagina of document indexeert een zoekmachine eigenlijk. Diens vermogen om documenten terug te vinden hangt immers mede af van het aantal elementen (en het daaraan toegekende gewicht) in de index. Is dit indexeergedrag over lange tijd constant?
- retrieval-gedrag: vindt of toont een zoekmachine eigenlijk wel alle documenten die hij zou moeten vinden op een zoekvraag? En verandert dat 'vinden' in de loop van de tijd? Maakt hij kort of langer optredende fouten?

Opzet onderzoek

Wij hebben dertien search engines vergeleken in de periode oktober 1998 tot september 1999. Tien internationale: AltaVista, Euroferret, Excite, HotBot, InfoSeek (tegenwoordig Go Net), Lycos, MSN, NorthernLight, Snap en WebCrawler. Drie regionale Nederlandse: Ilse, Search.nl en Vindex. In het kader van het onderzoek naar het indexeergedrag hebben wij ze laten zoeken naar 32 verschillende html-elementen in een testdocument.
Dat testdocument bevat naast tekst (een deel van het boek *The Secret Garden* van F. Burnett) html-elementen zoals plaatjes en geluidsfragmenten in verschillende formaten, JAVA-scripts en een formulier om een CGI-script mee aan te roepen. Bij elkaar 32 elementen (zie kader op pagina 19), een keuze dus want html kent er meer.
Van dit testdocument hebben we zestien identieke exemplaren geplaatst op verschillende servers. Dat heeft zo zijn redenen. Zoekmachines vinden of indexeren namelijk niet alle documenten op het web, of doen dat pas na lange tijd. Als ze een document wel vinden en indexeren dan zijn ze daar gekomen via links of directories, of na aanmelding door de eigenaar. Een deel van de zestien testdocumenten is door ons aangemeld, naar een ander deel zijn links gelegd vanuit andere – al bij zoekmachines bekende – documenten. In totaal zeven testdocumenten zijn aangemeld én gelinkt.
Voor het onderzoek hebben we een serie van 32 zoekvragen geformuleerd, voor ieder html-element één. Vuren we die af op een zoekmachine, dan zouden we in het *ideale* geval (alle testdocumenten zijn gevonden en geïndexeerd en de zoekmachine indexeert inderdaad alle 32 elementen) bij iedere vraag als resultaat een lijst van alle zestien testdocumenten krijgen, met bij ieder dezelfde annotatie.
Nu is niets ideaal, het ligt voor de hand dat er twee afwijkingen kunnen zijn. De eerste is al genoemd: de zoekmachine heeft niet alle testdocumenten gevonden of bewaard (ontdubbelen?), dus het resultaat bevat minder dan zestien treffers. En verder indexeert de zoekmachine niet alle gezochte elementen, dus op bepaalde vragen komen er helemaal geen treffers.
Dit zouden we *normaal* gedrag kunnen noemen als tenminste dit gedrag door de tijd (korte en lange termijn) gezien constant blijft. Maar ook dat is te veel gevergd. Niets belet (het baasje van) een zoekmachine om te besluiten voortaan een bepaald html-element alsnog te gaan indexeren, of helemaal niet meer. Een testvraag die altijd hetzelfde aantal treffers gaf levert dan opeens niets

Werkgroep IRT

De werkgroep IRT (Internet Retrieval Tools) is begin 1997 opgericht, op initiatief van de werkgroep PAD van de Vogin. Een en ander als vervolg op het onderzoek *Evaluatie van search engines,* door Wouter Mettrop uitgevoerd in september 1996. Uitgebreide gegevens over het onderzoek van IRT zie www.cwi.nl/cwi/projects/IRT/. Theoretische achtergronden, tabellen en lijsten van alle gegevens zijn onverkort weergegeven. Het onderzoek loopt door.
Deelnemers IRT: Louise Beijer, Hans de Bruin, Rudy Dokter, Marten Hofstede, Hans van der Laan, Hans de Man, Wouter Mettrop (coördinator), Paul Nieuwenhuysen, Eric Sieverts, Hanneke Smulders en Ditmer Weertman.

meer op, of vice versa. Laten we zeggen dat dit nog *redelijk* gedrag kan worden genoemd. Op zich al een interessant gegeven: sleutelen de leveranciers vaak aan hun zoekmachines? Ons onderzoek is erop gericht om dergelijke wijzigingen in gedrag te signaleren.

Je zou verwachten dat een zoekmachine een ooit gevonden en geïndexeerd document nooit meer "vergeet", of "weggooit". Nu weten vele eigenaars van een webpage dat dit niet waar is: hun aangemelde pagina is bijvoorbeeld na een paar maanden weer verdwenen. Onderzoek hiernaar leverde verrassende resultaten. Documenten worden vergeten maar een zoekmachine "herinnert" ze zich opeens weer, individuele testdocumenten komen en gaan (terwijl ze door ons natuurlijk steeds zonder onderbreking zijn gehandhaafd). Dit noemen we *vergeetachtig* gedrag: er treden *documentfouten* op.

Algauw merkten wij dat er nog meer veranderingen optreden. Stel dat zoekmachine X in een ronde zes documenten kent, dan valt te verwachten dat sommige testvragen (die immers steeds betrekking hebben op een specifiek element in het te vinden document) zes treffers opleveren en andere geen enkele. Er blijken echter ook antwoorden te worden gegeven die wel iets opleveren maar minder dan de verwachtte zes. In deze gevallen spreken we van een *incomplete* lijst van treffers. We noemen dit *inconsistent* gedrag: er treden elementfouten op. Naar wat er ten grondslag ligt aan deze *elementfouten* kunnen we slechts raden (zie kader).

De 32 onderzochte html-elementen

I. title tag
II. Meta-tag keywords
III. Meta-tag description
IV. Meta-tag author
V. comment tag ()
VI. ALT tag
VII. text of a link to an http document
VIII. URL of a link to an http document
IX. H3 tag
X. table header
XI. text of an internal link
XII. text of a reference anchor
XIII. text of a link to a sound file
XIV. name of an au sound file
XV. name of a wav sound file
XVI. name of an aiff sound file
XVII. name of a ra sound file
XVIII. text of a link to an image
XIX. name of a gif image file
XX. name of a jpg image file
XXI. name of an inline gif image file
XXII. name of an inline jpg image file
XXIII. name of an applet (including extension class)
XXIV. name of an applet (without extension class)
XXV. terms after the first 100 lines in a document
XXVI. terms after the first 200 lines
XXVII. terms after the first 300 lines
XXVIII terms after the first 400 lines
XXIX. terms after the first 500 lines
XXX. terms after the first 600 lines
XXXI. terms after the first 700 lines
XXXII. the URL of a document

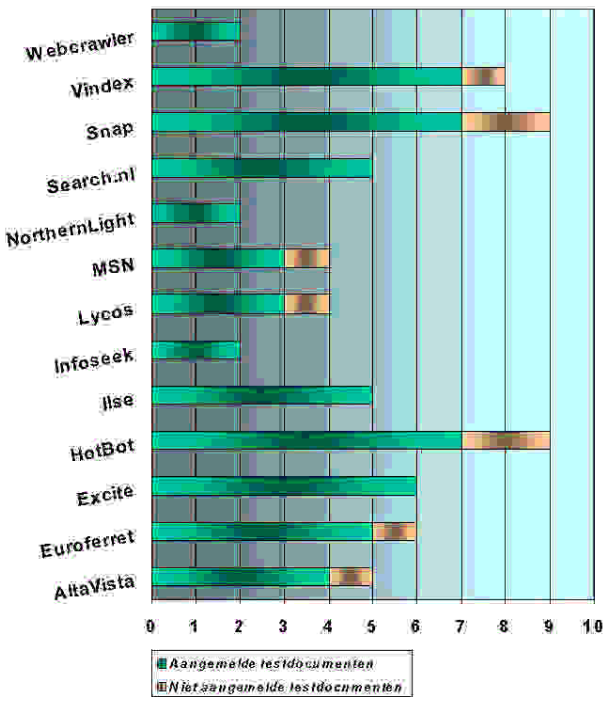
In de tijd

De eerste fase van het onderzoek was gericht op het ontwerpen van het testdocument, het plaatsen op verschillende locaties, het aanmelden van een aantal daarvan aan de te onderzoeken zoekmachines, en het opstellen van 32 zoekvragen die elk een element van het testdocument voor hun rekening nemen. In deze fase werden de zoekvragen "met de hand" aan de zoekmachines gesteld.

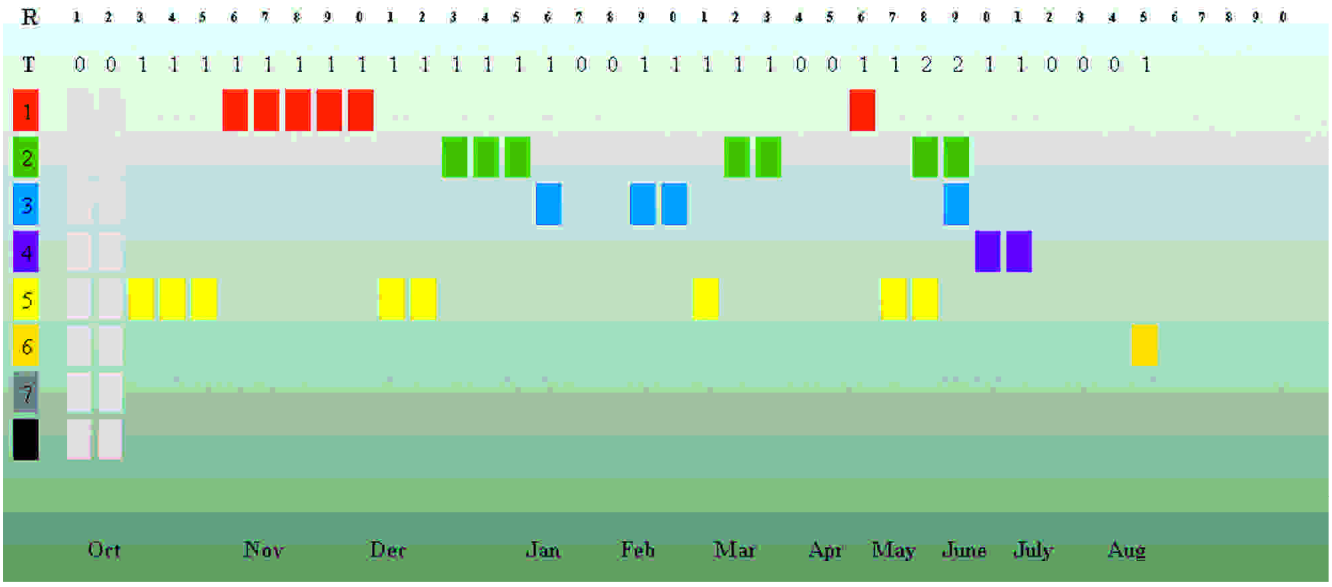
Om het gedrag van de zoekmachines in de tijd te kunnen volgen werd het afvuren van de vragen geautomatiseerd. Dat gebeurt sequentieel en – om geen overlast te veroorzaken – niet in staccatotempo. Na 29 minuten wordt de volgende vraag verstuurd, zodat de 32 vragen aan één zoekmachine een kleine 16 uur in beslag nemen. Daarna komt de volgende zoekmachine aan de beurt. Na ruwweg negen dagen start de hele cyclus opnieuw, in totaal hebben er in de testperiode 35 ronden plaatsgevonden.

Vergeetachtig gedrag

We hebben al gesteld dat het niet aannemelijk is dat iedere zoekmachine ieder testdocument vindt en indexeert. Dat kan een kwestie van beleid zijn. Te diep in de hiërarchie van een server bijvoorbeeld of – zeker in ons geval – door ontdubbeling van identieke documenten. Maar vreemd genoeg vinden we die patronen (vrijwel) niet. Uit de resultaten (figuur 1) blijkt dat geen van de zoekmachines alle zestien testdocumenten heeft gevonden. Maar het gedrag in de tijd is vaak veel wonderlijker dan hier is af te lezen. Als voorbeeld geven wij het verloop bij Excite



Figuur 1. Cumulatief tijdens de testperiode gevonden testdocumenten. Excite bijvoorbeeld heeft zes documenten gevonden, maar uitsluitend aangemelde.



Figuur 2. Door Excite gevonden testdocumenten. Horizontaal de testronden (R), verticaal de verschillende exemplaren van het testdocument. Excite heeft dus cumulatief zes documenten gevonden, maar nooit meer dan twee tegelijk.

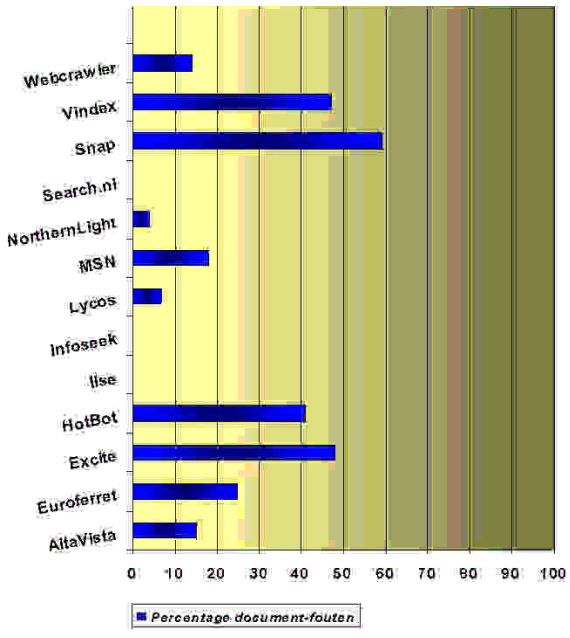
(figuur 2). In figuur 1 zien we dat deze zoekmachine in de loop van het onderzoek in totaal zes verschillende testdocumenten heeft gevonden. Maar uit figuur 2 blijkt dat Excite er nooit meer dan twee tegelijk kent, vaak slechts één en soms zelfs geen. En de werkelijkheid is nog gekker: niet alle gevonden testdocumenten kregen daarbij dezelfde annotatie. Nu is Excite wel de vreemdste. Ilse bijvoorbeeld is zeer constant, en NorthernLight geeft lang de indruk te ontdubbelen (maar in het eind van de testperiode gaat er toch nog iets mis).
We hebben deze *documentfouten* samengevat in figuur 3.

- *Tussen de verschillende zoekmachines bestaan verschillen in aantallen gevonden testdocumenten.*
- *Zoekmachines die als "groot" bekend staan (NorthernLight, AltaVista, HotBot) vinden in een set van aangemelde en gelinkte testdocumenten niet noodzakelijkerwijs een groter aantal testdocumenten.*
- *De geteste zoekmachines vinden gezamenlijk over de hele periode van het onderzoek slechts tien van de zestien testdocumenten. Dit bevestigt het beeld uit andere onderzoekingen dat alle zoekmachines samen nog lang niet het hele web bestrijken.*
- *Vele zoekmachines vertonen in hoge mate een vergeetachtig gedrag. Testdocumenten verschijnen en verdwijnen zonder dat daar regelmaat in valt te ontdekken. Zoekmachines lijden dus aan documentfouten.*

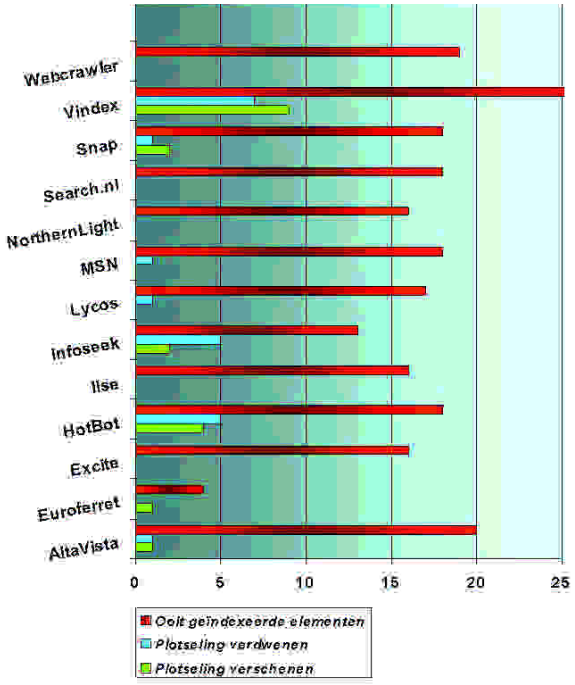
Redelijk gedrag

Het valt niet te verwachten dat iedere zoekmachine alle gezochte html-elementen in het testdocument vindt en indexeert. Dat is vaak te veel werk, het laat de index sterk groeien en – zo is de redenering – op sommige elementen zoekt "men" toch niet. Excite geeft zelfs duidelijk aan dat ze meta-tags niet opnemen "omdat aanbieders die toch voortdurend misbruiken om hun site hoog in de lijst te krijgen". In figuur 4 ziet u dat Euroferret weinig elementen indexeert, Vindex veel, terwijl de andere engines meer en minder het middenveld bezetten. Interessant is ook dat bij verschillende zoekmachines in de loop van het onderzoek langdurige (minstens vier testronden lang) veranderingen werden geconstateerd. Dat zijn (vanwege die blijvendheid) kennelijk veranderingen in indexeergedrag, dus menselijke beslissingen. NorthernLight en Excite bijvoorbeeld zijn hierin heel bescheiden, terwijl HotBot en vooral Vindex flink zitten te sleutelen.
Het is nu interessant om de figuren 1 en 4 te vergelijken en hieruit conclusies te trekken. Euroferret bijvoorbeeld vindt zes van de zestien testdocumenten, maar doet dat kennelijk met slechts vier geïndexeerde elementen. Vindex scoort hoog in zowel testdocumenten als elementen daaruit, maar stelt wel voortdurend de indexeerpolitiek bij.

- *Er blijkt geen samenhang te bestaan tussen het aantal geïndexeerde html-elementen en de prestatie van zoekmachines bij het vinden van testdocumenten.*



Figuur 3. Percentage documentfouten voor de verschillende zoekmachines over de hele testperiode.



Figuur 4. Cumulatief tijdens de testperiode ooit eens gevonden html-elementen. De in die periode plotseling optredende veranderingen in aantal zijn apart aangegeven. (In feite heeft Vindex 27 elementen gevonden.)

- *"Grote" zoekmachines indexeren niet opvallend meer (of minder) elementen dan andere.*
- *Verschillende zoekmachines indexeren in de loop van de tijd bepaalde elementen opeens wel, en andere niet meer. Ander zoekmachines veranderen weinig of niets aan hun indexeergedrag.*

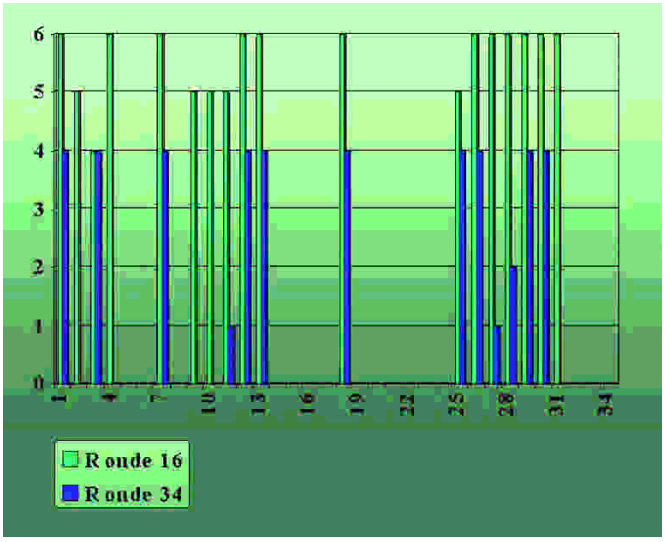
Inconsistent gedrag

Als een testvraag plotseling geen treffers meer oplevert, of als de vraag eerst geen, maar nu opeens alle documenten oplevert die een zoekmachine (gezien de resultaten van andere vragen in de testronde) op dat moment kent, en deze verandering houdt tenminste vier ronden aan, dan nemen we aan dat het gaat om een menselijke beslissing: men heeft het indexeergedrag van de zoekmachine veranderd.
Alles vinden of niets, dat lijkt duidelijk. Intuïtief neemt men aan dat een op één bepaald html-element toegespitste testvraag óf alle aan de zoekmachine bekende testdocumenten in de lijst van treffers oplevert, óf geen enkel.
Tot onze grote verbazing (en ontzetting) signaleerden we echter dat zo nu en dan een lijst van treffers incompleet was: ergens tussen nul en "alles" in. Dit noemen we elementfouten (ze treden op willekeurige momenten op bij willekeurige testvragen, en ze zijn niet blijvend).
Figuur 5 geeft daarvan een treffend voorbeeld. Daarin zijn van HotBot de resultaten van de ronden 16 en 34 weergegeven. In ronde 16 leveren de testvragen 1, 4, 7, … ieder zes treffers op; de testvragen 5, 6, 14, … geven géén treffers. Maar de testvragen 2, 9, 10, … geven vijf treffers, en vraag 3 geeft zelfs maar vier treffers.
Dat is echter nog niet alles. In deze ronde 16 kent HotBot in feite acht testdocumenten. De genoemde "maximale" zes treffers op de testvragen 1, 4, 7, … missen er dus steeds twee van de acht. In ronde 34 eenzelfde beeld. Bij vijf bekende testdocumenten leveren vele vragen vier treffers, maar er zijn er ook van één en twee treffers. Wij hebben de grootte van de elementfouten (6 van 8, of 4 van 8) verder niet beschouwd, alleen het aantal gevallen geteld. In figuur 6 hebben we die als percentages weergegeven. Vindex en MSN springen er sterk uit. HotBot is ook niet zo betrouwbaar. Maar Search.nl, NorthernLight, Lycos en Euroferret gedragen zich in dit opzicht keurig.

- *Meer dan de helft van de onderzochte zoekmachines vertoont inconsistent gedrag en maakt elementfouten.*
- *"Grote" zoekmachines maken niet opvallend meer (of minder) elementfouten dan andere.*
- *Er lijkt niet een direct verband te bestaan tussen het aantal elementfouten en de mate waarin zoekmachines hun indexeergedrag wijzigen.*

En wat nu?

Uit het onderzoek vallen enige conclusies en aanbevelingen af te leiden.
Hoofdconclusie: Zoekmachines zijn onbetrouwbaar!
Voor de praktijk moeten we de elementfouten en de documentfouten vertalen naar inconsistent gedrag op korte en langere termijn. Bij sommige zoekmachines dient men de zoekvraag te herhalen of te modificeren om zo te komen tot een – schijnbaar tegenstrijdig – antwoord. Bij andere zoekmachines dient men de vraag over een periode van een week te herhalen. En bij weer andere dient men zowel het ene als het andere te doen.
Voor een zoekmachine met veel elementfouten geldt dat men hem niet direct moet geloven, er kunnen zich tegenstrijdigheden voordoen. Iemand die uitputtend wil zoeken met dergelijke zoekmachines, zal zijn vraag herhaald, en liefst op verschillende manieren moeten stellen. Er is nooit een moment waarop men zeker weet dat een bepaald document zich niet in de database van de zoekmachine bevindt.



Figuur 5. Aantal treffers dat door HotBot tijdens de ronden 16 en 34 op de verschillende testvragen (horizontaal) is gegeven. HotBot kent tijdens ronde 16 in feite acht testdocumenten, in ronde 34 in feite vijf.

Voor zoekmachines met veel documentfouten geldt dat er op langere termijn veranderingen optreden. Bij een tegenvallend resultaat verdient het aanbeveling de vraag een dag of tien later (onze testperiode omvatte negen dagen) nog eens te stellen. Een nu gevonden document blijkt wellicht een week later te zijn verdwenen om tien dagen later weer terug te keren.

De hoofdrolspelers

Drie zoekmachines zijn zeer stabiel en geven over praktisch de hele periode dezelfde antwoorden. Eén zoekmachine maakt noch document- noch elementfouten: Search.nl. Verder zijn bijna foutloos: NorthernLight en Lycos. Concreet betekent dit dat deze zoekmachines steeds een compleet antwoord geven en dat dit antwoord over de gehele proefperiode onveranderd is. Een zoekvraag hoeft niet te worden herhaald, noch op korte noch op lange termijn.
Ilse en Infoseek maken geen documentfouten en weinig elementfouten; (weinig incomplete antwoorden en geen variatie op lange termijn), voor Euroferret is dat juist andersom. AltaVista en WebCrawler maken weinig ele-

Hoe zoekmachines werken

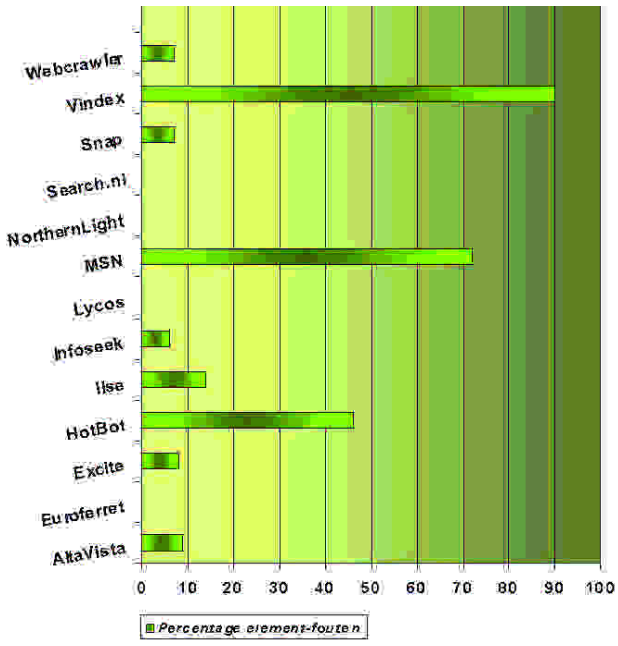
Zoekmachines vervullen op het Internet de functie van intermediair tussen document en gebruiker. Die functie kan worden gezien als een proces waarin de taal van het document en de zoekvraag van de gebruiker worden gematcht. Dit proces bevat een aantal stadia waarin veranderingen en inconsistenties kunnen optreden. Ter illustratie van die stadia een korte beschrijving van hoe zoekmachines werken.

webdocumenten → spider → URL-basket/summarizer → database/indexer → *index* → broker → *resultaat*

Internet-zoekmachines bestaan uit vier onderdelen, die elkaar aanvullen. Zij maken het voor de gebruiker gezamenlijk mogelijk om door het invoeren van zoektermen de documenten op het web te vinden die deze termen bevatten.
Aan het begin vinden we de zogenoemde *spiders of crawlers,* programmaatjes die webpagina's opsporen en downloaden, of andere werkzaamheden ten behoeve van de zoekmachine verrichten (bijvoorbeeld het opsporen van *broken links*). Ze werken doorgaans binnen een groep landen of een domein, gaan uit van een begin-URL en volgen dan de links binnen de bezochte site en vervolgens binnen het hun toegewezen domein, om nieuwe URL's op te sporen. De spiders verzamelen slechts een beperkt aantal URL's per site (en leggen daarbij uiteenlopende criteria aan voor wat ze verzamelen). Dat heeft ten dele te maken met hun eigen capaciteit, maar ook met het feit dat een spider die een site van duizenden pagina's zou moeten downloaden, daar zo lang mee bezig zou zijn dat andere gebruikers van de site daardoor hinder zouden ondervinden. Dit is een mogelijke reden waarom zoekmachines altijd maar een deel van het web dekken. Ze concentreren zich meer op een redelijk volledige en actuele collectie sites dan op de afzonderlijke documenten in die sites. Voorzover ze zich wel op afzonderlijke pagina's concentreren doen ze dat op basis van bekendheid (hoeveel links worden er naar een pagina gelegd en/of is de pagina apart aangemeld). Desondanks liggen ook de magazijnen van de zoekmachines vol met nooit opgevraagde URL's! Wel worden de door een spider gevonden links die niet zijn verwerkt, soms in een zogenoemde *URL-basket* opgeslagen. Sommige zoekmachines kennen dus meer URL's dan ze in hun database ontsluiten.
De door de spider wel gedownloade pagina's worden vervolgens doorgegeven aan een *summarizer*, een programma dat de pagina analyseert en de belangrijkste informatie ervan extraheert (bij AltaVista is dat bijvoorbeeld *Inxight*). De summarizer verricht, in samenwerking met de hieronder genoemde broker, het werk dat de zoekmachine uiteindelijk in staat stelt de gebruiker een "gerankt" resultaat op een zoekactie te presenteren.
Vervolgens worden de gedownloade pagina's, of de door de *summarizer* geproduceerde samenvattingen, in een database opgeslagen. Deze database wordt dan door een *indexer* geïndexeerd. Die bouwt verschillende indexen waarin indextermen met url's als adressen worden verbonden.
Aan het eind van het proces staat dan de *broker,* de gebruikersinterface van de zoekmachine, die het de gebruiker mogelijk maakt een vraag te stellen en daar een antwoord op te krijgen.

Wat kan er nu allemaal veranderen of misgaan:
• *webdocumenten* – er verschijnen en verdwijnen documenten op het web (dus in de werkelijkheid). Een zoekmachine houdt het tempo niet bij, of slaat gedeelten van het web over;
• *Spider* – een spider vindt een document niet meer, bijvoorbeeld doordat de er naar wijzende links zijn verdwenen, of doordat de spider niet meer het hele verwijzende document downloadt;
• *Summarizer* – de politiek van samenvatten verandert, en daardoor de geïndexeerde database;
• *Indexer* – het indexeringsbeleid verandert, of er treedt een indexeerfout op;
• *Broker* – door een retrievalfout vindt of toont de broker het document niet, ook al zit het in de index; door een wijziging in de retrieval-politiek vindt of toont de broker nog maar een van verschillende identieke documenten (ontdubbeling).

Zowel indexeerfouten als retrievalfouten leiden in het onderzoek tot het signaleren van *elementfouten.* Vandaar onze lichte aarzeling om hier te spreken van inconsistent gedrag: we kunnen aan de hand van de resultaten de oorzaak niet verifiëren.
In het hele proces dat zich in de zoekmachines afspeelt zijn er stadia waarin identieke documenten mogelijk niet worden gevonden of niet (verder) worden meegenomen. Voor *documentfouten* is het aanwijzen van een oorzaak dus nog moeilijker (behalve daar waar ontdubbeling het beleid is).
Maar eigenlijk is het aanwijzen van oorzaken ook niet zo belangrijk: de gebruiker ondervindt dat de resultaten van zijn zoeken niet optimaal zijn, om het zacht te zeggen. Zoekmachines zijn onbetrouwbaar.



Figuur 6. Percentage elementfouten voor de verschillende zoek machines over de hele testperiode.

mentfouten en weinig documentfouten.
De meeste documentfouten worden gemaakt door Excite en Snap, die echter betrekkelijk weinig elementfouten maken.
Voor Vindex en MSN geldt dat bijna elke vraag incompleet wordt beantwoord (zeer hoog percentage elementfouten), voor HotBot is dat bijna de helft. Deze drie machines maken ook vrij veel documentfouten en zijn dus met recht onbetrouwbaar te noemen.

Het is belangrijk al deze zaken in hun verband te zien, en zich niet vast te bijten in één aspect. Om een voorbeeld te noemen: een zoekmachine als Vindex maakt een zeer onbetrouwbare indruk, maar hij indexeert wel verreweg de meeste html-elementen en bezet de derde plaats waar het gaat om de aantallen gevonden documenten. En uiteraard blijft het bekende adagium gelden: gebruik steeds meer dan één zoekmachine.

Drs. H.R. van der Laan is computerraadsman te Leiderdorp en redactielid van Informatie Professional.
Drs. M.W. Mettrop is medewerker bij de Bibliotheek van het Centrum voor Wiskunde en Informatica (CWI) in Amsterdam.