Zoekmachines op Internet rommelen maar wat aan

ACHTERGROND

6 APRIL 2000

Dat zoekmachines geen weet hebben van alle documenten op het Internet, is bekend. Nieuw is dat ze ook onbetrouwbaar zijn in hun omgang met de documenten die ze wél kennen. Dat blijkt uit een gezamenlijk onderzoek van bibliotheken van bedrijven, universiteiten en hoge scholen. Ze presenteerden afgelopen dinsdag hun bevindingen tijdens de 'Online Conferentie Nederland'.

Rolf Zaal

Wie een ooit op Internet aangetroffen document niet kan terugvinden, legt de schuld algauw bij zichzelf. Er kunnen echter ook andere oorzaken een rol spelen. „Zoekmachines op Internet, zoals Webcrawler, Excite, Hotbot, MSN en Vindex doen hun werk slecht. De meeste zoekmachines vertonen variabel gedrag. De resultaten vertonen daardoor storende en vervelende fluctuaties; zodanig dat deze zoekmachines onbetrouwbaar kunnen worden genoemd", zegt Wouter Mettrop van het Amsterdamse Centrum voor Wiskunde en Informatica.

„Documenten die de ene keer wel worden gevonden, blijken een tijdje later onvindbaar. Ook komt het voor dat een document dat met een bepaalde zoekopdracht wel wordt gevonden, met een andere zoekopdracht niet boven water komt. Een paar dagen later kan het weer anders zijn. Dan werkt een andere zoekopdracht weer niet of komt een aanvankelijk 'verdwenen' document opeens weer in beeld. En dan praat ik uiteraard over fluctuaties in het gedrag van de zoekmachines, die dus niet voortkomen uit het weghalen of verplaatsen van de documenten zelf."

Wie een ooit op Excite gevonden document terugzoekt, heeft bij een latere zoekactie 34 procent kans dat het niet meer is te vinden. Op Snap is die kans 14 procent. Een via MSN aangetroffen document is in 30 procent van de gevallen met andere trefwoorden via diezelfde zoeksite niet te vinden. Vindex en Hotbot missen bij bepaalde zoekopdrachten 23 procent van documenten die ze wel kennen. Dat blijkt uit het feit dat de betreffende documenten via andere zoekopdrachten wel worden aangetroffen. Geen van de bekende zoekmachines, van Altavista tot Webcrawler, is vrij van dit onvoorspelbaar gedrag.

Mettrop is als wiskundige verbonden aan de bibliotheek van het Amsterdamse Centrum voor Wiskunde en Informatica. Samen met collega's van onder meer het Unilever Research Laboratorium, de Hoge School Amsterdam en de universiteiten in Brussel, Leiden en Utrecht heeft hij zitting in de werkgroep Information Retrieval Tools. Eigenlijk was het de bedoeling dat de werkgroep zich zou richten op het zoekbereik van de zoekmachines.

documenten werden nu eens wel en dan weer eens niet teruggevonden. We schrokken daar nogal van."

„Als ervaren gebruikers van online catalogi, bibliotheeksystemen en publicatie-databases ga je er van uit dat een document wel of niet voorkomt. Maar 'misschien wel', 'misschien niet' of 'soms wel', dat is een probleem. In de eerste plaats voor gebruikers van die zoekindexen, maar ook voor ons als onderzoekers. Toen bleek dat de volledigheid van de indexering niet op een reproduceerbare wijze te onderzoeken zou zijn, hebben we ons onderzoek gericht op die fluctuaties. Dat leek ons een probleem waarvan het in kaart brengen minstens zo belangrijk is. Voor zover we kunnen na gaan is dat nog door niemand anders opgepakt."

Waarom is de uniformiteit in de verwerking van zoekopdrachten belangrijk? Mettrop, met enige terughoudendheid: „We zijn geen sociologen. We hebben dus niet onderzocht hoe mensen Internet-zoekmachines gebruiken en wat ze ervan verwachten. Maar uit eigen observatie weten we wel dat een zogeheten 'known item search' veel voorkomt. In ons vakgebied misschien wel vaker nog dan de vraag 'zou er wellicht ooit iets geschreven zijn over ....'.
Bij de voorbereiding van een patentaanvraag of in het kader van een dissertatie-onderzoek kun je je gewoon niet permitteren een publicatie over het hoofd te zien. Eigenlijk betekent de nu bestaande situatie dat je aan de uitkomsten van een zoektocht op Internet geen enkele harde conclusie kunt verbinden.
Andersom is het natuurlijk ook zo dat degene die iets op het Internet zet dat als regel doet met de bedoeling dat anderen het daar weer kunnen vinden. Dat geldt voor wetenschappers en voor bedrijven."

Om de slordigheden in het afspeuren van Internet in kaart te kunnen brengen plaatsten Mettrop en collega's een gedeelte van F.H. Burnetts beroemde kinderboek 'The secret Garden' op zestien verschillende locaties van het Internet. Zeven van deze documenten werden actief bij de zoekmachines aangemeld. Ook werden 'links' aangebracht, vanaf pagina's die bij de zoekmachines al bekend waren.
Om de documenten te traceren werden
32 zoekopdrachten opgesteld. Elk van deze opdrachten refereerde aan een specifiek element van het document, zoals de titel, de hoofdtekst, de auteursnaam, een meta-tag of een bijschrift bij een illustratie. Deze vragen werden herhaaldelijk op de diverse zoekmachines ingevoerd.
In het totaal waren er 43 rondes waarbij telkens 32 zoekvragen aan dertien zoekmachines werden voorgelegd. Eén ronde, waarbij alle machines werden bevraagd, duurde negen dagen. Het experiment heeft dan ook meer dan een jaar in beslag genomen (oktober 1998 tot december 1999).

De bevindingen waren op zijn zachtst gezegd ontluisterend voor de zoekmachines, die zich als regel presenteren als de toegangsportalen tot het Walhalla van informatie. Om de chaos van het vruchteloos zoeken nog enigszins te ordenen, onderscheiden Wouter Mettrop en zijn collega drie categorieën van fluctuaties:

- volstrekte onvindbaarheid van documenten die bij andere gelegenheden wel werden gevonden (documentfluctuaties);
- het slechts vindbaar zijn van documenten op een beperkt deel van de termen (elementfluctuaties) en;

(indexfluctuaties).
Documentfluctuaties zijn het vervelendst. Ze zijn alleen te omzeilen door de zoekacties te verdagen naar een moment met een gunstiger gesternte. Excite spant de kroon in het opzicht van documentfluctuaties. Gemiddeld bedraagt de kans een in principe bekend document te missen daar
34 procent. Via Webcrawler wordt 17 procent van de documenten op die manier gemist. Dat is fors boven het gemiddelde dat op bijna 9 procent uitkomt. Alleen Ilse.nl en Search.nl bleken in het onderzoek vrij van documentfluctuaties.
Elementfluctuaties zijn hinderlijk omdat de gebruiker bij het niet vinden van een document blijft zitten met de twijfel of een anders geformuleerde zoekopdracht niet toch het gezochte document boven tafel zou brengen. De kans dat een bepaalde zoekopdracht een wel degelijk geïndexeerd document niet identificeert is het grootst bij MSN: 30 procent. Maar ook de zoekprogramma's Vindex en Hotbot hebben reden zich te generen: bij hen wordt bij de gemiddelde zoekopdracht 23 procent de (wel degelijk geïndexeerde) documenten over het hoofd gezien.
Bij de gemiddelde zoeksite wordt door elementfluctuaties 8 procent aan documenten gemist. Alleen Euroferret, Northernlight en Lycos gaven gedurende het onderzoek geen elementfluctuaties te zien. In het geval van Euroferret wordt deze verdienste overigens sterk gerelativeerd door het gegeven dat Euroferret sowieso weinig termen lijkt te indexeren.

Hoe erg is dit nu allemaal? Voor literatuurstudies en wetenschappelijk onderzoek, bijvoorbeeld naar publicatiegedrag, zal het een serieuze handicap zijn. Buiten wetenschappelijke kringen kan het tellen van links naar sites een zinvol onderzoeksinstrument zijn. Bijvoorbeeld als alternatief voor de dikwijls als geflatteerd en onbetrouwbaar aangemerkte bezoekcijfers, die van belang zijn om de waarde van advertentieruimte op sites te beoordelen.
Mettrop: „Het heeft met deze tijd te maken. Zoekmachines zouden ook bruikbaar moeten zijn om ook in andere contexten performance op Internet te meten."
Toch lijken de meeste beheerders van zoeksites niet erg te zitten met de door de IRT-werkgroep gesignaleerde problemen. De eerste informele reacties komen er min of meer op neer dat de zoeksites helemaal niet vergeleken willen worden met orthodoxe databases. Men lijkt een zekere mate van onbetrouwbaarheid wel acceptabel te vinden.

Algemeen directeur Merien ten Houten van Ilse.nl zegt: „Ik ken het probleem. Ik ben blij dat Ilse er verhoudingsgewijs weinig last van heeft. Toch loop ik er zelf ook weleens tegenaan. Je hebt een document niet in je bookmarks gezet en kunt het niet terugvinden via de zoekopdracht waarmee je het eerder wel aantrof. Dat kan buitengewoon vervelend zijn. Aan de andere kant is het zo dat de gebruiker meestal niet naar een bepaald specifiek document zoekt. Gewoonlijk heeft hij geen weet van dat niet-gevonden document en is hij tevreden als hij over het gezochte onderwerp iets heeft gevonden."
Mettrop: „Ja, voor zo'n benadering kun je natuurlijk kiezen. Maar dan is het wel goed als de gebruiker dat ten minste weet."

URL: http://www.cwi.nl/cwi/projects/IRT
Oorzaken diffuus

Weltrop en zijn collega's van de werkgroep Information Retrieval Tools zorgvuldig vermeden: „Daar hebben we per se niets over willen zeggen. Het gaat erom dat je als gebruiker weet dat fluctuaties bestaan. Hoe het komt, doet er voor het zoeken op Internet ook niet toe."

Een kleine rondvraag onder 'information retrieval'-specialisten en aanbieders van zoekdiensten op Internet wijst uit dat de mensen die het zouden moeten weten geenszins gelijkluidend oordelen over de mechanismen die debet zijn aan het vreemde gedrag van zoekmachines.

Jan Prins van de Internet Society Nederland: „Tja. Het is bijna menselijk he? De term 'onbetrouwbaar gaat me wat ver, maar 'onvoorspelbaar' zijn ze zeker. Dat heeft te maken met allerhande hightech-oplossingen in de zoekalgoritmen van de machines. Vroeger werkten die gewoon op 'full text', maar door het groter worden van het web voldoet dat niet meer.

Op eenvoudige zoekopdrachten van gewone gebrui-

kers kwamen onhanteerbare hoeveelheden documenten bovendrijven. De oplossingen die daarop zijn gevonden, verschillen per zoekmachine. De meeste oplossingen gaan in de richting van wat 'conceptual search' wordt genoemd. Dat zijn 'fuzzy search'-achtige algoritmen, die de context van een document of een trefwoord betrekken in de beslissing om het al dan niet in de index op te nemen.

Zo zijn er sites die vaak opgevraagde documenten een hogere prioriteit geven. Dat kan tot veranderingen leiden, die je zou kunnen aanmerken als inconsequent gedrag. Nou, dat is dan de prijs die men betaalt voor het ontsluiten van Internet op een manier die voor de meerderheid van de gebruikers beter werkt dan het recht-toe-recht-aan indexeren."

De ontwikkeling die Prins beschrijft doet zich inderdaad voor. Toch biedt dat hooguit een deel van de verklaring. De onderzoekers werkten immers met identieke documenten. De context was dus onveranderlijk. Niettemin werden de identieke documenten door de zoekmachines op verschillende tijdstippen anders behandeld. Bovendien waren de zoekopdrachten zodanig gedefinieerd dat er buiten de testdocumenten slechts weinig 'vreemde' documenten in de resultaatset belandden.

Ook éminence grise van de full text retrieval in Nederland, Thijs Chanowsky, denkt niet dat de verklaring voor de inconsequenties van zoekmachines moet worden gezocht in de ingebouwde 'intelligentie' van zoekmachines: „Dan zouden ze alleen maar beter moeten werken, en dat is duidelijk niet het geval. Ik erger me ook vaak groen en geel. Ik denk dat het onder meer te maken heeft met allerhande trucs, die ze toepassen om de efficiëntie te verhogen.

Er vinden dagelijks miljoenen mutaties plaats op het Internet. Real-time zijn die niet meer bij te houden. Wat sommige zoekmachines daarop hebben gevonden, is dat ze nieuwe documenten in 'slagen' analyseren. Eerst worden bijvoorbeeld alleen de eerste honderd woorden geïndexeerd. Als het wat minder druk is, worden er nog eens duizend bij gepakt en uiteindelijk volgt ooit de rest. Wellicht herhaalt die hele procedure zich nog eens in het kader van technieken die de retrieval moeten verbeteren, zoals het omgaan met synoniemen en variaties in de schrijfwijze."

Als Chanowsky's verklaring klopt, dan zouden de zoekresultaten in de loop van het proces een tendens tot verbetering moeten vertonen. En dat was niet het geval. In het onderzoek wisselen verbetering en verslechtering elkaar af volgens een grillig patroon. Terwijl bij een bepaalde zoekopdracht het ene document uit de resultaatset wegvalt, duikt soms een ander exemplaar weer op.

De verlossende uitleg komt, hoe kan het ook anders, van een van de mensen achter de oudste en meest geraadpleegde zoekmachine op Internet: Altavista.com. Vice President Research Andrei Broder: „We zijn

Altavista door er alles aan om het zoekvermogen te beperken. Hoe is echter niet zo duidelijk waar de oorzaak is.

Een van de belangrijkste oorzaken is dat datgene wat zich aan de gebruiker voordoet als één zoekmachine in feite een samenstelsel is van meer databases. De indexen op die databases worden regelmatig gerepliceerd. Het kan gebeuren dat iemands zoekopdracht de ene keer naar de ene database gaat en een andere keer naar weer een andere. Als de replicatie op dat moment nog niet heeft plaatsgevonden of om een of andere reden niet volledig is geweest, kan dat tot verschillen leiden in de antwoorden.

Een andere oorzaak kan zijn, dat verschillende indexen worden gebruikt, bijvoorbeeld voor verschillende delen van het Internet. Het komt voor dat een van die indexen op enig moment niet bereikbaar is. Dat leidt tijdelijk tot een onvolledige resultaatset. Dat probleem doet zich vooral voor bij zoekmachines die uit veel kleine computersystemen zijn opgebouwd. Altavista werkt met een kleiner aantal grote computers.

Wat ook een rol kan spelen, is het gebruik van 'lastafhankelijke' algoritmen. Dat komt erop neer dat als het druk is, minder intensief wordt gezocht.

Bij Altavista voeren we elk uur een aantal test-zoekopdrachten uit om te zien of het niveau van de fluctuaties nog wel acceptabel is. Zonodig nemen we maatregelen. Het volledig elimineren van het probleem zou echter te hoge kosten met zich meebrengen.

Fluctuaties omzeilen

De kans op het missen van documenten door elementfluctuaties is volgens Wouter Mettrop van het Amsterdamse Centrum voor Wiskunde en Informatica enigszins te verkleinen door diverse zoekopdrachten uit te proberen. Documentfluctuaties zijn enigszins te omzeilen door zoekopdrachten op verschillende tijdstippen, liefst met enkele dagen ertussen, op de zoekmachines los te laten.

Beide problemen zijn gedeeltelijk te ondervangen door verschillende zoekmachines naast elkaar te gebruiken. Maar ook hier geldt, net als bij de twee andere maatregelen dat het de kans op fouten verkleint, maar niet elimineert. Het blijft immers mogelijk dat een bepaalde document- of elementfluctuatie zich op beide machines tegelijk voordoet.

Wel is het zo dat de de kansrekening leert dat dat gevaar drastisch afneemt als een derde en een vierde zoekmachine wordt ingezet. Maar dat is weer bewerkelijk. Tenzij gebruik wordt gemaakt van zogeheten meta-zoekmachines zoals Infind, Metacrawler, Highway61 of Savvysearch, die gelijktijdig verschillende 'gewone' zoeksites bevragen. Ook zijn er zoekprogramma's, zoals Copernic, waarmee de gebruiker zelf een stuk of tien zoeksites tegelijk kunnen bevragen. De resultaten worden, ontdubbeld en ontdaan van dode links, aan de gebruiker gepresenteerd.

Ook aan deze pragmatische oplossing kleven echter bezwaren. Van praktische en morele aard. Door het combineren van verschillende zoekmachines wordt de zoekfunctionaliteit van de gekozen machines teruggebracht tot de kleinste gemene deler: als er één machine tussen zit die een bepaalde opdracht (bijvoorbeeld de NEAR-opdracht) niet kent, dan kan die opdracht dus niet meer worden gebruikt. Bovendien gebruiken meta-zoekmachines informatie van andere sites, die daardoor waardevol Internet-'verkeer' mislopen.

Onderzoekresultaten zijn ontluisterend voor zoekmachine

Je kunt dit artikel lezen nadat je bent ingelogd. Ben je nieuw bij AG Connect, registreer je dan gratis!

INLOGGEN

[redacted]

Hulp bij het inloggen nodig?

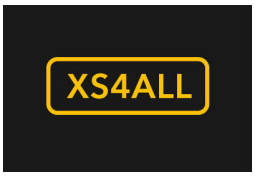
REGISTREREN

- Direct toegang tot AGConnect.nl
- Dagelijks een AGConnect nieuwsbrief
- 30 dagen onbeperkte toegang tot AGConnect.nl

REGISTREREN

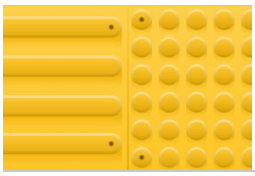
Ben je abonnee, maar heb je nog geen account? Laat de klantenservice je terugbellen!

LEES OOK



KPN drukt XS4ALL-opslokking door

TikTok-eigenaar investeert in sportforum



PARTNERBIJDRAGE

INCLUSIVE DESIGN:...

PARTNERBIJDRAGE

DE TOEKOMST VAN ITSM:...

Microsoft en Oracle slaan brug tussen clouds

‘Privacytoezichthouders te traag met afhandeling klacht tegen Google’

Toezichtsraad: kijk kritisch naar positionering BIT

NIEUWSOVERZICHT ▸