Werkgroep IRT (1)

Search engines aan de tand gevoeld

Hans van der Laan en Wouter Mettrop

De Werkgroep Internet Retrieval Tools houdt zich bezig met het hinderlijk volgen van search engines. Die mogen dan wel beweren dat ze de grootste, de slimste, de beste zijn, maar de werkgroep neemt dat niet voetstoots aan. Een eerste onderzoek leert dat search engines lang niet alle meta-data willen aanvaarden. En zoals te verwachten was: er zijn grote verschillen.

Op initiatief van de werkgroep PAD van de Vogin is begin 1997 de werkgroep IRT (Internet Retrieval Tools) opgericht. Een en ander als vervolg op het onderzoek 'Evaluatie van search engines', door Wouter Mettrop in september 1996 uitgevoerd.
De impuls tot die oprichting was duidelijk, de concretisering van het doel lag moeilijker. Iedereen immers ontmoet de hoopvol gestelde vraag: "Welke search engine kan ik het beste gebruiken voor deze concrete zoekvraag?" Daarom rees het idee: kunnen we de Nederlandse zoeker een actueel overzicht verschaffen van de eigenschappen van de verschillende zoekmachines, en hun sterke en zwakke punten in kaart brengen? En dat dan vanzelfsprekend on line aanbieden en - uiteraard - up-to-date houden? Het is allemaal zo gemakkelijk gezegd in het eerste enthousiasme, maar vooral dat onderhouden leek meteen al wat hoog gemikt. En uiteraard is het zinloos om andermans werk te dupliceren. Last-but-not-least: wat kan je met zo'n dozijn vrijwilligers bereiken? Uiteindelijk hebben uitgebreide discussies daarover geleid tot de 'doelstelling van IRT'.

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

Onderzoeksgebied

Search engines verzamelen adressen van webpagina's, indexeren hun inhoud en verzamelen de resultaten daarvan in een grote database. Uiteraard speelt de factor snelheid hier een rol maar van veel meer belang voor het te verwezenlijken doel is de vraag wát en hoe er wordt geïndexeerd. Immers, wat er niet in de database zit kan er niet uitkomen. Leveranciers zijn over het algemeen niet mededeelzaam over deze *indexeerfase*. Tot hoe diep in de vertakkingen van een website worden de bladzijden "meegenomen"; wordt alles ervan geïndexeerd, ook stopwoorden? En zo ja, wat zijn stopwoorden? Worden afstandsgegevens vastgelegd, zodat later met een NEAR kan worden gezocht? En uitermate belangrijk: wat wordt er gedaan met meta-data?
De raadpleger van een search engine krijgt in eerste instantie te maken met de *zoekfase* van het gebruik. Welk arsenaal aan gereedschappen biedt de search engine de gebruiker om zijn zoekvraag zo gedetailleerd mogelijk te beschrijven? Zijn Booleaanse operatoren toegelaten? Worden stopwoorden genegeerd? Deze laatste vraag maakt meteen duidelijk dat indexeer- en zoekgedrag van een search engine nauw met elkaar zijn verweven. Waar stopwoorden niet zijn geïndexeerd levert een vraag naar "to be or not to be" waarschijnlijk niets op. Dit soort overwegingen maakt proefondervindelijk onderzoek naar het indexeergedrag vaak bijzonder moeilijk.
In de *presentatiefase* schotelt de search engine de resultaten van de zoekactie voor aan de zoeker. Uiteraard vormt het *ranking*-mechanisme hier het voornaamste onderwerp van onderzoek. Ook daarover hullen de leveranciers zich meestal in een mythisch stilzwijgen. Daarnaast kunnen vragen als 'wat wordt getoond als samenvatting van een webpage?', en 'vindt er ontdubbeling plaats?', bijdragen aan een beter begrip voor de kwaliteit van een search engine.
De werkgroep heeft ervoor gekozen geen kwantitatieve onderzoeken te verrichten naar vangst en precisie. Die vormen een *mer à boire* waarin zij vreest te zullen verdrinken.

Werkwijze

Een onderzoek als dit levert zeer veel gegevens, die samen de gebruikelijke omvang van een artikel in *Informatie Professional* verre zouden overschrijden. Bovendien, zoals gezegd, zal de werkgroep die gegevens regelmatig actualiseren. Wat ligt er dan meer voor de hand dan een website <www.cwi.nl/~wouter/search-tools/IRT/>. Hier vindt de lezer voorlopig enkele gegevens, op den duur een compleet overzicht.
Een uitgebreide brainstorming leverde in eerste instantie drie lijsten van testcriteria, voor het indexeer-, het zoek- en het presentatiegedrag. In die lijsten is (vooraf) aangegeven welke criteria zullen worden getest (aangegeven met JA) en welke op een andere manier worden geverifieerd (NEE).

Vervolgens heeft de werkgroep op basis van die testcriteria een omvangrijke webpage ontworpen, waarin zo veel mogelijk tekst, plaatjes, html-codes en meta-data zijn opgenomen. De tekst bestaat uit een aantal hoofdstukken van een klassiek verhaal, in de page zelf aangegeven met: <META NAME="description" CONTENT="This test page, containing a small part of the Secret Garden (by Frances Hodgson Burnett) is part of a larger site about the IRT project. vier, vijf, zes">.

Juist, dat "vier, vijf, zes" is listig in deze meta-tag *description* verstopt om later al dan niet te worden teruggevonden bij het testwerk. Zo zijn er vele onderwaterwerken, die bij weergave van de page via een browser niet direct zichtbaar zijn.

Deze testpage nu is op zeven verschillende locaties (verschillende servers) neergezet, niet *gelinkt* aan andere pages. Daarna zijn alle zeven gemeld aan alle door de werkgroep uitverkoren search engines. Van dat moment af is er regelmatig gekeken of die de pages inmiddels hadden geregistreerd. Hoewel dat geen opzet was, leverde deze periode deels onthutsende ervaringen op, waarover later meer.

Inmiddels had de werkgroep ruim vijftig testvragen ontworpen, die vervolgens aan de gekozen search engines (zij die één of meer van de testpagina's hadden gevonden) werden gesteld. In de praktijk blijkt dan dat er soms toch ambiguïteiten optreden in de interpretatie door de testers, ondanks alle voorzorgen. De vragen zijn daarom hier en daar bijgesteld. De resultaten vormen inmiddels een lijvig dossier. Het aantal voetnoten daarin illustreert de moeilijkheid van interpretatie van - vooral - het indexeergedrag van de onderzochte search engines.

Keuze

De werkgroep had de hoop iets te kunnen zeggen aan de hand van waarnemingen over hoe snel de search engines de zeven test-pages na aanmelding hebben gevonden. Maar het is moeilijk om aan de hand van deze resultaten conclusies te trekken, zeker omdat het geen wetenschappelijk verantwoorde proefopzet was.

Aan de andere kant is het moeilijk om gericht onderzoek te doen naar de snelheid waarmee engines indexeren, en dus lijkt het de werkgroep zinvol om toch iets van deze bevindingen met terughoudendheid te presenteren.

Daar komt bij dat de werkwijze van de robots van de verschillende engines in het algemeen geheim is. Het is heel moeilijk om daar iets over te weten te komen van de engine zelf. Wij denken dat elke informatie welkom is.

Het is in ieder geval onze ervaring dat Infoseek en HotBot hier het beste scoren, althans bij elk één van de pages. Maar terwijl HotBot ze alle zeven op den duur vindt, bereikt Infoseek er maar twee. OpenText en Go2 hadden na maanden nog niets gevonden en zijn daarmee uit de test verdwenen. Dat geldt ook voor Zoek.NL, die onmiddellijk na de aanmelding de ontvangst (van die melding) bevestigde maar verder niet meer reageerde. Excite vertoonde het meest vreemde gedrag: sommige gevonden pages werden weer "vergeten", soms later "teruggevonden" maar nooit blijvend. Excite bereikte enige malen een maximum van vier pages, maar dat waren niet steeds dezelfde. Daarnaast traden er in de presentatie ook merkwaardigheden op, die voor nader onderzoek voorlopig zijn gerangschikt onder de post "Inconsequent Gedrag". In het vervolgonderzoek van de werkgroep wordt hier aan gewerkt.

De search engines die op grond van deze aanloopperikelen zijn uitverkoren voor de eigenlijke test zijn vermeld in bijgaand kader.

Indexeergedrag

Bij onze proefnemingen is gebleken dat vrijwel elke search engine zo hier en daar inconsequent gedrag vertoont, dat waarschijnlijk is terug te voeren naar de indexeerfase. Zoals al opgemerkt zal in het vervolgonderzoek dit gedrag nader worden onderzocht. In dit stadium willen we nog geen beschuldigende vingers uitsteken.

In dit stadium van het onderzoek is veel aandacht besteed aan de manier waarop search engines blijken om te gaan met de *tags* die in de test-page zijn opgenomen. We onderscheiden *header-tags*, zoals *meta-tags*, en *body-tags*. Hier volgen de eerste conclusies.

Header

• De tag *title* wordt door alle onderzochte search engines geïndexeerd. Dat ligt voor de hand, maar moet toch worden opgemerkt.

Doelstelling van de Werkgroep Internet Retrieval Tools (IRT)

De Werkgroep IRT houdt zich bezig met de orthodoxe, geheel geautomatiseerde, *echte* search engines (geen subject tree of gespecialiseerde directory). Althans de belangrijkste/bekendste daarvan, en in ieder geval de Nederlandstalige. Wij streven ernaar gebruikers een voortdurend geactualiseerde vergelijking tussen search engines aan te bieden, met als doel de eindgebruiker ten dienste te zijn bij het zoeken van informatie op het Internet, door te adviseren in het geavanceerd gebruiken van search engines. De Werkgroep verkrijgt haar gegevens zo veel mogelijk door eigen waarneming en onderzoek, gegevens van de leverancier worden in het uiterste geval slechts als aanvulling gebruikt. IRT blijft de conclusies voortdurend actualiseren. Onze proeven zijn herhaalbaar en worden steeds herhaald. Een nieuwe search engine kan zonder problemen in het overzicht worden opgenomen.

• Alleen HotBot, AltaVista, InfoSeek, Ilse en Vindex ondersteunen meta-tags (Excite vermeldt expliciet dat meta-tags *niet* worden meegenomen omdat ze zouden worden misbruikt). InfoSeek, AltaVista, HotBot, Ilse en Vindex kijken naar de meta-tag *keywords*; InfoSeek, AltaVista, HotBot en Ilse kijken ook naar *description*; HotBot kijkt bovendien naar *author*.
• De tag *link* (een van de langer bestaande html-tags) wordt door geen enkele search engine ondersteund. Deze tag is bedoeld om verbanden tussen verschillende documenten aan te geven. Jammer, want die tag zou kunnen bijdragen aan opties als "more like this" bij de presentatie van de zoekresultaten, en aan het ontsluiten van achterliggende databases. Vaak wordt de toegang tot zeer relevante informatie gevormd door een simpel database-interface met betrekkelijk weinig termen. Je moet dat interface maar zien te vinden om bij de informatie te kunnen komen. *Links* naar documenten in de database zouden dit euvel kunnen verhelpen.

Body

• De commentaar-tag (<!--->) wordt alleen door HotBot geïndexeerd.
• Voor alle search engines geldt dat de *tekst* van een *link* wordt geïndexeerd; of dat nu een link is naar een http-document, naar een intern *anchor*, naar een geluidsfragment of naar een image. Ook de tekst van een referentie-*anchor* wordt geïndexeerd, dat wil zeggen: de tekst die staat op de plek waarnaar wordt verwezen vanuit een ander deel van hetzelfde document. Maar dat is ook niet zo verwonderlijk als men bedenkt dat dit deel uitmaakt van de gewoon zichtbare tekst van het document.
• Delen van URL's in een document (*links*) worden geïndexeerd door HotBot, InfoSeek en AltaVista. Een moeilijkheid in het onderzoek wordt gevormd door het feit dat hier evaluatiecriteria met betrekking tot indexeer- en zoekgedrag door elkaar lopen, al levert dat voor de conclusies geen problemen op. Wel is het hinderlijk dat er een verschil blijkt tussen de praktijk en de leer: sommige search engines pretenderen termen uit URL's te indexeren, maar doen het in de test-situatie niet.
• Full-text - Alle engines indexeren full-text (dus het gehele document) behalve:
- InfoSeek - mist termen (vijf van de zeven) aan het begin en aan het eind van de body.
- SearchNL - mist termen (vijf van de zeven) in het midden van de body.
- Vindex - indexeert zo ongeveer de eerste tweehonderdvijftig regels, en verder niet.
- Euroferret - bijzonder indexeergedrag (wij gaan hier later op in).

• Diacrieten - De meeste engines ondersteunen de volgende zoekmogelijkheden:
- Gebruik diacritische mogelijkheden op het toetsenbord (Me'diterrane'e): (Niet: Ilse, InfoSeek).
- Gebruik ISO8859-1 codes (Me'diterrane'e): (Niet: Lycos, Ilse, AltaVista en Euroferret; Discutabel: HotBot en InfoSeek zoeken op het getal uit de ISO-code).
- Gebruik *Entity Names* (Méditerranée): (Niet: Lycos, Ilse, AltaVista en Euroferret; Discutabel: HotBot en InfoSeek zoeken op de term uit de Entity Name).
- Gebruik normale letters, zonder accenten (Mediterranee): (Niet: HotBot, AltaVista en Vindex).

Het is in alle gevallen aan te raden alle mogelijkheden te proberen bij het zoeken naar diacrieten. Let wel: er kan verschil zijn tussen wat er in de geïndexeerde tekst stond en wat je als zoekletter kunt intikken (door vanaf het toetsenbord ALT-130 in te tikken een document te vinden waarin "é" is gegeven als "é".

• Stopwoorden
- Engines die wel stopwoorden indexeren: Excite, Infoseek, AltaVista, NorthernLight, SearchNL en Vindex.
- Engines die geen stopwoorden indexeren: Lycos, Ilse, HotBot, WebCrawler, Euroferret.

Deze resultaten zijn nog niet definitief. Er moeten nog zaken gedefinieerd worden als: 'wat zijn stopwoorden' en 'over welke taal hebben we het'. Wij komen hierop terug.

• Merkwaardig gedrag: AltaVista en Euroferret zoeken niet op termen in een table header en AltaVista zoekt als enige niet op termen in de tekst van een interne anchor.

In een volgend artikel komen na, een vervolg op het indexeergedrag, ook het zoek- en presentatiegedrag aan de orde. Inmiddels bereidt de werkgroep zich voor op vervolgonderzoek:

- Het in kaart brengen van inconsequent gedrag.
- De werking van ranking-mechanismen.
- Vergelijking van verschillende opties in de presentatiefase.
- Hoe up-to-date zijn de zoekresultaten?

De tot nu geteste search engines

Lycos	www.lycos.com
Ilse	www.ilse.nl
Excite	www.excite.com
Infoseek	guide.infoseek.com
Hotbot	www.hotbot.com
AltaVista	altavista.digital.com
WebCrawler	webcrawler.com
NorthernLight	www.northernlight.com
SearchNL	www.search.nl
Euroferret	www.euroferret.com
Vindex	www.vindex.nl