ARTIKEL

ZOEKTECHNOLOGIE

Auteur: Edwin Dijk
2011

Laatste update: 04-05-2011

KORTE BESCHRIJVING:

Dit is één van de weinige vrij beschikbare Nederlandstalige artikelen op het gebied van zoek(machine)technologie. Het biedt een inzicht in verschillende soorten zoekmachines, verschillende algoritmes en manieren om informatie te bekijken. De vraag is natuurlijk: hoe bewerkstelligen we het beste resultaat?

Dit artikel is een ideaal beginpunt voor ieder die ambieert een eigen zoekmachine te bouwen.

INHOUDSOPGAVE

INLEIDING

IR, oftewel Information Retrieval, is een vakgebied dat een prominente plaats heeft gekregen in de levens van de moderne mens. Door de uitvinding van het internet is informatie in steeds grotere mate beschikbaar voor het grote publiek. Dat levert natuurlijk interessante situaties op. De enorme hoeveelheid informatie wordt vervolgens ingedeeld, beoordeeld en gekanaliseerd door zoekmachines als Google, Bing en anderen.

Het doel van een zoekmachine is dus in principe 'het voldoen aan de informatiebehoefte van de eindgebruiker'. Hoe dit precies gebeurt en welke theorie hierachter schuil gaat is natuurlijk razend interessant om te bekijken. De vraag is natuurlijk: hoe wordt een zoekmachine op de meest efficiënte manier opgezet?

Zonder informatiebehoefte is er geen markt voor een zoekmachine en zolang een zoekmachine niet aan de informatiebehoefte voldoet heeft deze geen bestaansrecht. Met andere woorden, zullen we moeten bepalen hoe we informatie opslaan, beoordelen en nazoeken. In dit onderzoek wordt dit uitvoerig besproken.

Er staan hier en daar in dit onderzoek "Food for thought" koppen. Dit is puur een gedachtegang die op het moment van schrijven in mij op komt en raakvlakken vertoont met het hoofdonderwerp maar niet strikt noodzakelijk is voor de strekking van het verhaal.

Edwin Dijk

HET OMGEKEERDE: ZOEKMACHINE-OPTIMALISATIE

Bij het maken van een zoekmachine, probeer je in te breken op een markt die al bestaat. Dat betekent dat je twee kanten op kunt: of je gaat mee in de markt, of je scheidt jezelf af en doet het op je eigen manier.

Ik geef je graag de keuze. Het bedenken van een nieuw soort zoeken is een nobel streven en misschien is het daar ook wel eens tijd voor. In eerste instantie gaan we kijken hoe we op dit moment bovenaan kunnen komen in de zoekmachines.

PageRank™
De uitvinding van Google, waardoor de zoekmachine zo succesvol is geworden. Het is een systeem waarbij het aantal links naar je website bepaalt hoe hoog je in de zoekmachine komt te staan. Hoe meer PageRank de website die naar jouw website linkt heeft, hoe meer je website krijgt. Het resultaat is dat hooggewaardeerde websites links verkopen aan webmasters, zodat zij hoger in Google kunnen komen. Het PageRank principe is dus deels verziekt door de commercie. Toch is het een interessante factor om mee te nemen, al dan niet in deze vorm.

Semantische structuur
Dit zijn de HTML-tags h1, h2, h3, bold, strong enzovoort. Het zijn de elementen waarmee je opmaak aan een tekst meegeeft, geïnspireerd door kranten en magazines. Het eerste wat opvalt in een krant, zijn de hoofdkoppen. Die is groot en daarom eigenlijk ook belangrijk. Aan de hand van de hoofdkoppen weet je al een beetje waar het over gaat en kun je vervolgens bepalen het te gaan lezen. Dit principe wordt in HTML ook gebruikt. In de zoekmachine-optimalisatie branche is het optimaliseren van de hoofdkoppen één van de belangrijkste factoren. Daar kunnen we ons voordeel mee doen!

Sitemaps
De sitemap is in feite een pagina met links naar alle pagina's op de website. Er bestaan grofweg twee varianten waar je rekening mee moet houden: de opgemaakte sitemaps (in HTML) en de XML sitemaps. Als je eenmaal zover bent dat je een spider hebt gebouwd, zal deze de sitemaps die in HTML gemaakt zijn meteen meenemen in de database. Of dat voor XML zo is, hangt af van je spider. In feite maakt het niet uit als je spider in eerste instantie geen XML indexeert.

Sleutelwoorden
Het is belangrijk om een goed idee te krijgen van hoe mensen hun sleutelwoorden onder de aandacht brengen. Een voorbeeld van ranking aan de hand van *keywords* is dat de zogenaamde *keyword-density* mee telt. Ga je echter aan het rekenen, kan dit onmogelijk een factor zijn die zwaar meetelt. Laat je dus niet verleiden. Keyword-density is een factor die wellicht bij SEO (Search Engine Optimization) regelmatig toegepast wordt, toch is het voor een spider geen topprioriteit.

Meta-tags
Hierin kun je de *keywords* en de *description* van gebruiken. Hoewel de keywords natuurlijk door je spider uitgelezen worden, kan de description dienen als prima paginavulling voor bij je zoekresultaten. Het loont dus de moeite om die op te slaan!

Conclusie
Het loont de moeite om goed te kijken naar de structuur van websites. De zichtbare aspecten moeten immers op een bepaalde manier omgezet worden in opmaaktaal. De meeste opmaak gebeurt tegenwoordig netjes via de daarvoor bestemde HTML-tags, dus het advies luidt natuurlijk om deze te gebruiken.

VERSCHILLENDE SOORTEN "QUERIES"

Een query is in principe een ingevoerde vraag. Of dit nu gebruikersinvoer is of een database aanvraag: het blijft een vraag waarop een exact antwoord wordt verwacht. Het herkennen van de aard van een vraag/query is dus een essentieel aspect van het voorzien in de informatiebehoefte. In dit geval gaan we in op gebruikersinvoer. Grofweg kunnen deze queries ingedeeld worden in drie categorieën:

I. **Navigational**
   a. In feite wil een gebruiker graag de zoekmachine gebruiken om op de bestemming terecht te komen die hij of zij al in gedachten heeft.
   Bijvoorbeeld: men zoekt op 'Microsoft' en wil terecht komen op 'www.microsoft.com' of 'www.microsoft.nl'.

II. **Informational**
   a. Hierbij is een gebruiker op zoek naar informatie en hoeft deze niet van tevoren te weten waar deze uitkomt. De betrouwbaarheid en het informatiegehalte zijn hierbij uiteraard van groot belang.

III. **Transactional**
   a. De naam suggereert al ongeveer wat dit inhoudt: een resultaat waar transacties gepleegd kunnen worden. Denk hierbij bijvoorbeeld aan een webshop, maar ook het online telefoonboek of het opzoeken van meer informatie bij een meer gespecificeerde bron. Met andere woorden is er bij een transactional query vaak nog een aanvullende actie nodig om tot het gewenste resultaat te komen.

Om een idee te krijgen van de omvang van elk van deze categorieën verwijs ik graag naar cijfers over deze drie.

Type of query	User Survey	Query Log Analysis
Navigational	24.5%	20%
Informational	?? (estimated 39%)	48%
Transactional	> 22% (estimated 36%)	30%

Andrei Broder (AltaVista), A taxonomy of web search, 2000

De hierboven weergegeven cijfers zijn niet keihard, maar een grove benadering van de hoeveelheden. Mijn persoonlijke verwachting is dat sinds 2000 de informatiebehoefte enigszins verschoven is en het aantal transactional queries omhoog is gegaan en het aantal informational queries juist omlaag. Dit in verband met het feit dat het internet door de jaren des te meer een commercieel medium is geworden. Maar ook dit is slechts een veronderstelling.

Ieder van deze drie behoeft een andere aanpak, maar de vraag is: hoe classificeren wij een query?

CLASSIFICEREN VAN QUERIES

Queries die ingedeeld zijn in de categorieën navigational, informational en transactional hebben allemaal bepaalde kenmerken waardoor zij mogelijk als zijnde een bepaald type query herkend kunnen worden. Laten we een aantal van deze zaken op een rijtje zetten.

Navigational:

- Mogelijk commercieel (ca. 50% kans)
- Grotendeels of geheel gelijk aan een bijbehorende domeinnaam (bijvoorbeeld microsoft.com)
- Bestaat uit een bedrijfsnaam, naam van de organisatie of naam van de website

Informational:

- Mogelijk commercieel (ca. 25% kans)
- Bestaat uit één of meerdere begrippen
- Begrippen bestaan mogelijk uit twee of meer samengevoegde woorden

Transactional:

- Mogelijk commercieel (ca. 75% kans)
- Bevat vaak werkwoorden die een actie vertegenwoordigen (bv. 'auto kopen')
- Bevat vaak zowel een zelfstandig naamwoord als een werkwoord

Van alle drie categorieën is uiteraard de navigerende categorie gebruikers het makkelijkst te bedienen. In feite is dit de meest tastbare categorie, omdat het resultaat waar de gebruiker om vraagt naar grote waarschijnlijkheid ook in zeer herkenbare vorm bestaat.

De twee andere categorieën hebben wat minder strikte eigenschappen. Dit maakt de aard van de informatiebehoefte dus minder duidelijk. In het geval dat de eerder genoemde cijfers nog steeds min of meer kloppen, kan met een goede afhandeling van navigational queries al een vijfde of meer van het totaal aantal queries met succes geserveerd worden.

Prioriteit is dus om het serveren van navigational queries in goede banen te laten lopen, omdat wij hiermee gemakkelijk een groot deel van onze gebruikers kunnen helpen. Hiermee winnen wij immers bestaansrecht, omdat wij in die gevallen succesvol aan de informatiebehoefte voldoen.

Wat de *informational* en *transactional queries* betreft, kunnen we deze samenvoegen tot één kernbegrip. In feite komen ze op hetzelfde neer, maar is het uiteindelijke doel voor de gebruiker anders. In eerste instantie hoeven wij daar in ons zoekalgoritme geen rekening mee te houden. Wat we wel benaderd hebben, is dat ongeveer een kwart van de informational queries en driekwart van de transactional queries commercieel is. Ik stel dus een vernieuwde structuur voor, die het verschil tussen 2000 (waar het internet nog niet volledig commercieel benut werd) en 2011 (waar het internet in zeer veel gevallen wel commercieel benut wordt) kan overbruggen:

- Navigational
- Informational
    - Commercial
    - Non-commercial

Waarbij we commercial en non-commercial dus kunnen laten voor wat het is, omdat we in eerste instantie te maken hebben met informational queries.

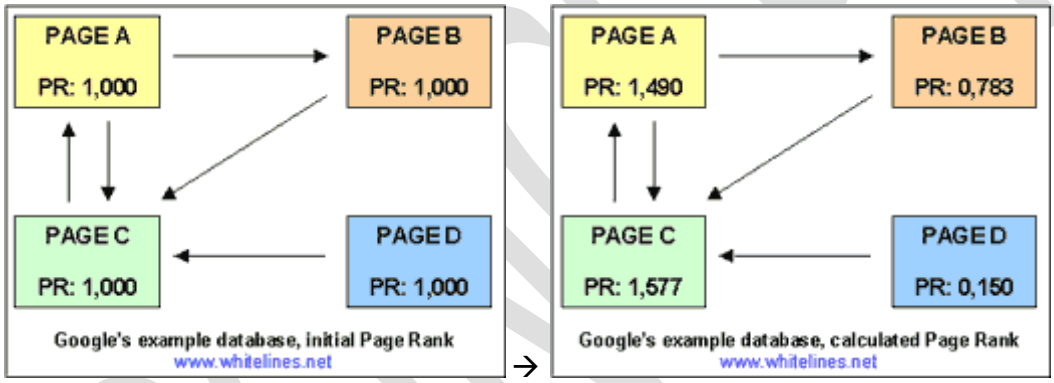
HOE WERKT PAGERANK™

PageRank is één van de meest baanbrekende ontwikkelingen geweest op zoekmachinegebied. De twee uitvinders van Google, Sergey Brin en Lawrence Page, hebben op een slimme manier bepaald welke waarde een pagina heeft aan de hand van de verwijzingen van andere pagina's.

Hoe PageRank werkt valt aan de hand van de basisformule uit te leggen.

$$PR(A) = (1 - d) + d * SUM ((PR(I->A)/C(I))$$

PR	PageRank
A	De pagina waarvan de PageRank berekend wordt
d	Damping factor, normaliter op 0,85
C	Het aantal hyperlinks
I	De pagina waar de link naar pagina A vandaan komt
I->A	De PageRank waarde die pagina A van pagina I krijgt
PR(I->A)/C(I)	De PR van pagina A is de PR van pagina I gedeeld door het aantal links op pagina I
SUM (PR(I->A)/C(I))	De som van alle pagina's (I) die naar pagina A verwijzen



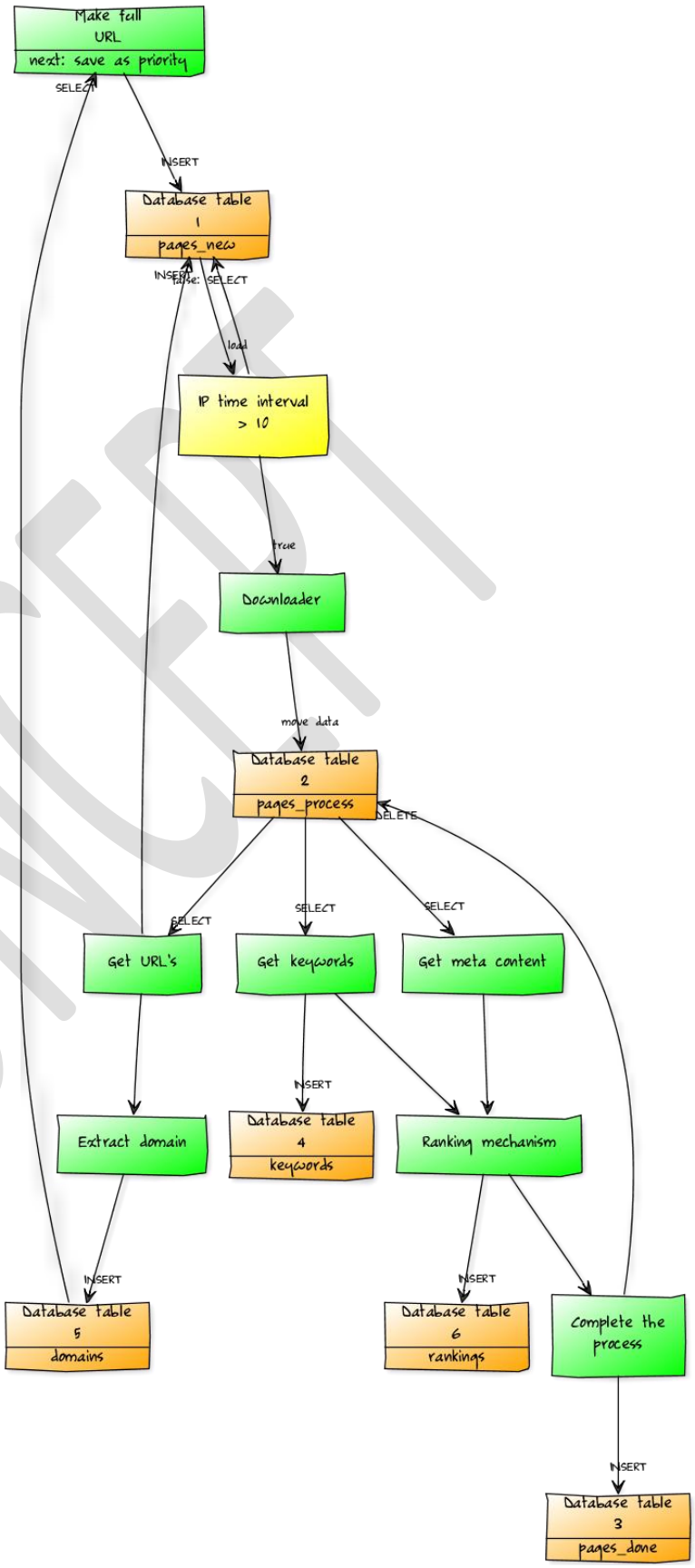
Bron: WhiteLines

Zoals duidelijk zichtbaar is in de grafieken en ook in de bron, kost het 20 iteratieve calculaties om de juiste PageRank aan de websites te hangen. De methode is natuurlijk briljant opgezet, maar reeds gepatenteerd en dus niet bruikbaar in deze vorm. Natuurlijk kunnen we later met het bepalen van onze eigen pagina ranking wel het aantal links in de relevantiebeoordeling opnemen.

Het is aan te raden om goed in de materie te duiken, omdat de manier van denken zeer goed van pas komt bij het bepalen van de meest effectieve strategie voor het opzetten van een zoekmachine.

Indexeren van pagina's kan op verschillende manieren gebeuren. In feite komt het versimpeld op het volgende neer:

- ➔ Pagina ophalen
- ➔ URL's opslaan
- ➔ Informatie vergaren



Dit kunnen we vervolgens onderverdelen in verschillende taken. Zo moeten we bijvoorbeeld keywords bij kunnen houden, bij voorbaat al een verschil in ranking tussen pagina's bepalen en ervoor zorgen dat alle processen sluitend zijn. Het mag namelijk niet zo zijn dat de zoekmachine eeuwig bezig blijft met berekenen. Hij moet op een zeker punt met een vaste dataset klaar zijn met rekenen en niet meer verder kunnen. (Net als het PageRank algoritme dat na verloop van tijd simpelweg balans bereikt heeft en daarna eigenlijk klaar is met berekenen.)

Organische zoekresultaten

De zoekresultaten die gegenereerd zijn door de zoekalgoritmes zijn de zogenaamde *organische zoekresultaten*. Hier tegenover staan de 'gesponsorde' resultaten oftewel de *betaalde zoekresultaten.* Aangezien het niet bepaald een technische uitdaging is om betaalde zoekresultaten bovenaan de pagina te krijgen, zullen we ons in dit document behoeden van de slechte invloeden van geld. We focussen dus alleen op organische resultaten.

Informational queries

Zoals we eerder hebben besloten, is het belangrijk om in eerste instantie te voldoen aan de wensen van de makkelijkste gebruikers: de *informational queries*. We hebben in het hoofdstuk 'Classificeren van queries' al bepaald dat dit soort

queries een aantal gunstige eigenschappen hebben.

We hebben benaderd dat *informational queries* ongeveer in de helft van de gevallen commercieel zijn. De meeste commerciële partijen zijn tegenwoordig dan ook vertegenwoordigd door een website. In bijna alle gevallen is de bedrijfsnaam dan ook meteen de domeinnaam: dit is nuttige kennis, omdat we deze kunnen gebruiken om onze zoektechniek efficiënter te maken. In eerste instantie kunnen we kijken naar de volgende manier van zoeken:

- ➔ Als er voor de zoekterm een gelijke domeinnaam bestaat
  - Bereken de kans dat de gebruiker deze website bedoelt
    - Is de kans hoger dan (bijvoorbeeld) 80%?
      - Ja: Toon website als eerste resultaat
      - Nee: Toon simpelweg organische resultaten
- ➔ Als er voor de zoekterm **geen** gelijke domeinnaam bestaat
  - Ga door met standaard berekenmethode

Dan komt de volgende kwestie: hoe berekenen we of de gebruiker deze website bedoelt?

De volgende drie vragen komen sowieso aan bod:

1. Hoeveel van de woorden uit de zoekterm staan in de titel?
2. Hoe vaak komt de exacte zoekterm voor op de pagina?
3. Komen er losse woorden uit de zoekterm voor op de pagina?

Op deze manier voorkom je dat er een compleet irrelevant item bovenaan de pagina wordt weergegeven. Het lijkt in eerste instantie wellicht knullig, maar op het moment dat men een bepaald zoekresultaat puur op basis van domeinnaam voortrekt, moet er een realtime check op het resultaat worden uitgevoerd om met zekerheid te kunnen zeggen dat het mogelijk de informatie is waar door de gebruiker om gevraagd wordt. Dit heeft ook nog andere redenen natuurlijk, maar die komen later aan bod.

Informational queries
In tegenstelling tot de navigational queries, zou men kunnen stellen dat informational queries vaak maar weinig te maken hebben met domeinnamen of bedrijfsnamen. Een gebruiker is op zoek naar informatie en de plaats waar die vandaan komt is in die zin irrelevant. Eventueel zou er rekening gehouden moeten worden met Wikipedia.

> **Food for thought**
> Het is frappant om te zien dat er een wisselwerking bestaat tussen het opzoeken van informatie via Wikipedia en het opzoeken van informatie via een zoekmachine beiden een andere insteek hebben. Hoewel ik geen bewijs heb voor deze aanname, denk ik dat er een verschil bestaat tussen de informatiebehoefte bij het gebruik van bijvoorbeeld Google of het gebruik van bijvoorbeeld Wikipedia.
>
> Een encyclopedie als Wikipedia heeft immers anno 2011 nog steeds niet de status van een geverifieerde encyclopedie. Het zoeken via Wikipedia bestaat ook puur uit de informatie van de website zelf en hiermee wordt het zoekveld in die zin beperkt. Het doel van Wikipedia is daarom dus vaak om snel wat informatie op te zoeken en weer weg te gaan. Daar kan menig internetgebruiker zich waarschijnlijk in vinden.
>
> Het gebruik van een volwaardige zoekmachine zoals Google, Bing of Yahoo is de geprefereerde optie als het gaat om wetenschappelijke artikelen met betrouwbare bronnen. Vooral Google is hier een sterke speler in, met name door de enorme zoekindex en het technisch superieure zoekalgoritme. Dat is precies wat wij als zoekmachine-beginners willen bereiken.

Volgens het wetenschappelijke tijdschrift Nature is de kwaliteit van Wikipedia nagenoeg gelijk aan die van Encyclopaedia Brittanica. Op universiteiten en hogescholen wordt Wikipedia echter nog altijd niet als een geldige bron beschouwd en dat is een belangrijk feit als het gaat om informational queries. De teruggegeven resultaten moeten natuurlijk wel van hoge kwaliteit zijn.

Dit is mogelijk een voorbeeld van het niet optimaal functioneren van het PageRank principe, omdat populariteit niet per definitie uit hoeft te monden in betrouwbaarheid of in ieder geval een gevoel van betrouwbaarheid.

RBC
Een ander bezwaar op het PageRank principe is daarbij dat het beginnende websites in de weg staat. Een hele leuke website met een hoop informatie zou bijvoorbeeld niet hoog kunnen ranken, puur om het gegeven dat er geen backlinks voor te vinden zijn. Dat is een probleem, waarbij alleen conventionele techniek kan helpen: Ranking By Content of kortweg RBC.

Deze techniek maakt gebruik van de eigenschappen van HTML. Een webpagina is – indien volgens de webstandaarden – opgemaakt a.d.h.v. een zogenaamde *semantische structuur*. In normale mensentaal betekent dit zoveel als: hoofdkoppen, subkoppen en deelonderwerpen. Aan de hand hiervan kan de opbouw van een tekst bepaald worden en zo kunnen op een gemakkelijke manier de kernwoorden opgehaald worden.

> **Food for thought**
> Jarenlang heb ik het omgekeerde gedaan: de semantische structuur verbeteren om beter te scoren in Google. Eerlijke titels die de lading dekken hebben de hoogste kans van slagen, maar de lading dekken mag prima gebeuren met een extra keyword. Iedereen wil immers gevonden worden, dus om een dergelijke rankingmethode te blokkeren zou op z'n zachtst gezegd niet handig zijn. Zij die gevonden willen worden op een zoekterm bieden hoogstwaarschijnlijk aan wat de bezoeker wil. Op die manier bied je draagvlak onder commerciële partijen, wat het project vervolgens weer levensvatbaarder maakt.
>
> Het is de vraag in hoeverre je uitbuiting door internet marketeers op zijn beloop moet laten. Uiteindelijk komt het erop neer dat de zoekmachine mee moet kunnen met de bestaande markt, omdat er anders geen draagvlak is, maar aan de andere kant wil je voldoen aan een goede informatiebehoefte. Een aspect waar Google sinds medio 2010 in mijn persoonlijke optiek steeds slechter in begint te worden. Op een juiste, eerlijke manier omgaan met content. Dáár gaat het om.

Het onderscheid in content kan worden gemaakt aan de hand van de tags:

TITLE	De browsertitel
H1	Hoofdkop
H2	Subkop
H3	Deelonderwerp
B	Dikgedrukt
STRONG	Dikgedrukt (Populair i.c.m. XHTML)
I	Italisch, schuinschrift
EMPHASIS	Nadruk, schuinschrift (Populair i.c.m. XHTML)
U	Onderlijn
IMG	Het alt attribuut bevat informatie over een afbeelding.

Aan de hand van deze vrij basale HTML-tags kan de globale inhoud van de pagina bepaald worden.

Een voorbeeld hiervan heb ik geschreven in het redelijk toegankelijke PHP. Dankzij DOM Document ben je in PHP voorzien van prima handvatten voor het inlezen van informatie in HTML en XML opgemaakte bestanden. Er zijn ook alternatieven beschikbaar, zoals bijvoorbeeld Simple HTML DOM. Het probleem hiermee is echter dat het je code ernstig vertraagt. Het is een prima script voor het opvragen van een enkele pagina zo nu en dan, maar voor een high-performance applicatie is hij nooit bedoeld. Laat je dus niet verleiden!

```
class htmlParser
{
    // Vind de waarde van $tag in $html
    public function find($html = NULL, $tag = NULL)
    {
        $dom = new domDocument;

        // Inladen van de HTML als een DOM Document
        $dom->loadHTML($html);
        $content = $dom->getElementsByTagname($tag);

        $output = NULL;
        foreach($content as $item)
        {
            // Alle keywords met spatie scheiden
            $output .= trim($item->nodeValue) . " ";
        }

        // Gehele string in kleine letters omzetten
        $output = strtolower($output);

        // String omzetten in array, met een spatie als seperator
        $values = explode(" ",$output);

        // Lege arraywaarden weggooien
        foreach($values as $key => $value)
        {
            if($value == "")
            {
                unset($values[$key]);
            }
        }


        // Bij het maken van een nieuwe array worden lege values
        // weggegooid
        $outputArray = $values;

        // De array met keywords uit de bepaalde tag teruggeven
        return $outputArray;
    }
}
```

HOE VERDER

Het belangrijkste aan het ontwerpen van de zoekmachine, is dat je sommige factoren wat meer achterwege laat dan andere. De ervaring leert dat de eerste dagen, een groot deel van je zoekindex geconsumeerd wordt door pornowebsites en andere zaken waar je eigenlijk niet op zit te wachten.

Het voordeel hieraan is echter: op het moment dat er pornosites opduiken bij een zoekopdracht die over iets heel anders gaat, zit je zoekalgoritme niet goed genoeg in elkaar (of is je webindex veel te klein).

Het ontwikkelen van een zoekmachine is een tijdrovend proces. En daarna vergt het ook nog eens heel veel geduld. Het gebeurt niet van de één op de andere dag dat je een webindex van 100 miljoen websites hebt. Je loopt tegen problemen aan als: budget, capaciteit (onlosmakelijk met elkaar verbonden), technische kennis, verschillende (on)mogelijkheden en ga zo maar verder.

Het belangrijkste is echter: niet opgeven.

BRONNEN

Andrei Broder (AltaVista), A taxonomy of web search, 2000

Andrei Broder is een oud-werknemer van AltaVista, een van de zoekmachines die voor de komst van Google een groot aandeel van de zoekmarkt in handen had. Zijn paper is erg interessant om de aard van verschillende zoekopdrachten te achterhalen.

http://www.sigir.org/forum/F2002/broder.pdf

Whitelines Search Engine Marketing & Optimization

Een Engelstalige beschrijving en rekenvoorbeeld van PageRank. Erg helder en duidelijk opgezet. Het is ook aan te raden even naar de Excel sheet te kijken voor wat extra verdieping.

http://www.whitelines.nl/html/google-page-rank.html

Yuml.me

Met deze tool kunnen (redelijk) gemakkelijk UML diagrammen gemaakt worden. In dit document heb ik o.a. deze website gebruikt om afbeeldingen te genereren die het proces dat ik probeer uit te leggen kunnen verduidelijken.

http://www.yuml.me/

Nature 438, 900-901 (15 December 2005) | doi:10.1038/438900a; Published online 14 December 2005

In dit onderzoek wordt aangegeven dat de kwaliteit van Wikipedia en Encyclopaedia Britannica niet eens zo gek ver uit elkaar ligt. Een interessante conclusie, aangezien hogescholen en universiteiten Wikipedia nog niet als wetenschappelijk verantwoorde bron accepteren. Twijfels blijven bestaan over de scope van dit onderzoek: is het wel breed genoeg uitgevoerd om het complete plaatje in beeld te brengen?

Publicatie: http://www.nature.com/nature/journal/v438/n7070/full/438900a.html
Onderzoeksdetails: http://www.nature.com/nature/journal/v438/n7070/extref/438900a-s1.doc

---