

De volgende slag om de rangorde

De nieuwste stap in het ordenen van zoekmachineresultaten is dat de volgorde wordt bepaald door de voorkeuren van de individuele gebruiker. Wij worden op onze wenken bediend. De keerzijde is vervolgens de vraag, wat diezelfde zoekmachines van ons weten.

Wouter Gerritsma

Nog niet zo lang geleden speelden trefwoorden de belangrijkste rol voor zoekmachines om de resultaten te presenteren. Daar werd volop misbruik van gemaakt. Googles grote kracht was de geheel nieuwe technologie om linkpopulariteit mee te laten wegen in de rangorde waarmee zoekresultaten gepresenteerd werden. Deze methode dreigt nu ook weer vast te lopen door de grootschalige inzet van *link farms*, *doorway pages* en wat dies meer zij. De volgende stap is het ordenen van zoekmachineresultaten volgens de voorkeuren van de individuele gebruiker. Wij worden op onze wenken bediend. De keerzijde is vervolgens de vraag, wat diezelfde zoekmachines van ons weten.

Te kust en te keur

Zoekmachines zijn er te kust en te keur. Ze hebben een aantal principes gemeen. Zo hebben ze allemaal een *spider* of een *bot* die het web afstruikt op zoek naar nieuwe en vernieuwde websites. Ze hebben allemaal hun database waarin de gevonden data worden opgeslagen. Verder een gebruikersinterface waarmee mensen hun zoekvragen kunnen stellen die uiteindelijk in een resultatenpagina resulteren.

‘Zoekmachines beschikken over een schat aan informatie over hun gebruikers’

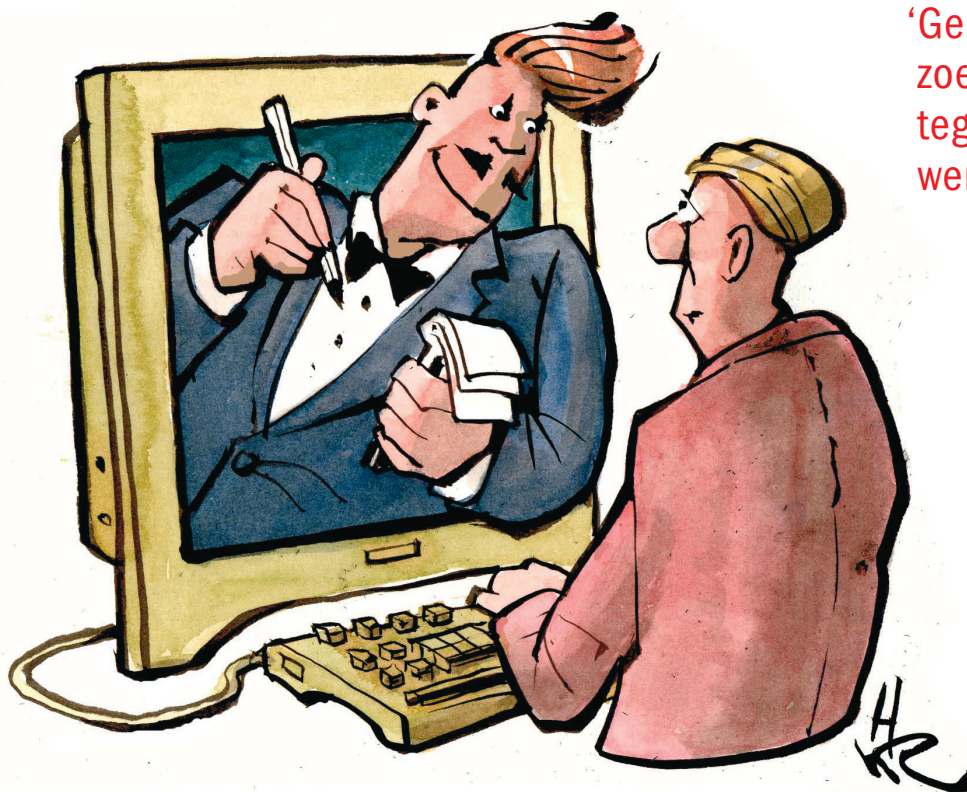
De resultatenpagina's verschillen aanzienlijk tussen de zoekmachines, niet alleen in de advertenties die links of rechts worden gegeven, of verfijningsmogelijkheden op de zoekresultaten die er worden geboden, maar vooral ook door de rangorde van de gepresenteerde zoekresultaten. Een aardig hulpmiddel om dat te visualiseren is *Thumbshots Ranking*. Hiermee kunnen de posities van de eerste honderd zoekmachineresultaten van zeven verschillende zoekmachines met elkaar worden vergeleken.

Het bepalen van de volgorde waarin de resultaten aan de gebruiker gepresenteerd moeten worden, is voor zoekmachines bepaald geen sinecure. Het succes of falen van een zoekactie staat of valt met de presentatie van de zoekresultaten. De meeste gebruikers kijken niet verder dan eerste twee pagina's aan zoekresultaten (die hooguit tien of twintig resultaten bevat-

ten). Daarnaast worden de indexen van de zoekmachines steeds groter en zijn er per zoekactie meestal vele honderden zo niet duizenden relevante resultaten voorhanden. Een heel wezenlijke vraag voor zoekmachinebouwers is daarom: welke resultaten moeten vooraan komen en welke komen lager in de rangorde op de resultatenpagina.

Woordfrequenties

De eerste generatie zoekmachines zoals Excite, Lycos en AltaVista gebruikten in de tweede helft van de jaren negentig voornamelijk de trefwoorden in de zoekvraag, en die in de resultaten, om de volgorde van de zoekresultaten op de resultatenpagina te bepalen. Daarvoor werden verschillende wiskundige modellen gehanteerd die gebruik maakten van woordfrequenties, afstanden tussen woorden op een pagina en alle mogelijke andere kenmerken van de trefwoorden. Het probleem was echter dat het verschijnen van bepaalde zoekwoorden op een webpagina heel gemakkelijk manipuleerbaar bleek. De eerste vormen van zoekmachinemaniipulatie dienden zich aan. Webpagina's werden volgestopt met populaire zoektermen, al dan niet zichtbaar voor



**‘Gebruiker
zoekmachines
tegenwoordig op z’n
wenken bediend’**

het oog, om de pagina’s hoog in de rangorde van de zoekresultaten te laten eindigen. De eerste vorm van zoekmachine-spamming door middel van trefwoord-spamming was geboren. Eind jaren negentig was er bij de toen populaire pagina’s geen ontsnappen meer aan. De zoekmachine-indexen waren stevig verziekt door deze schijnbaar primitieve spamtechnieken.

Dat vormde een gouden kans voor Google om door te breken. De presentatie van zoekresultaten bij Google werd niet meer alleen bepaald door woordfrequenties van de zoektermen, maar het aantal links naar een pagina was van doorslaggevende invloed. De formule hiervoor werd door Larry Page – een van de twee oprichters van Google – uitgevonden. De linkpopulariteit van een pagina kan worden uitgedrukt met de zogenaamde *Page-rank*. Google werd door deze nieuwe wijze van presenteren van zoekresultaten ongekend populair. Daarnaast spelen andere factoren mee in het succes van Google, zoals het snel kunnen laten groeien van de database van de zoekmachine en het effectief uitschakelen van *dead links*, een fenomeen dat eind jaren negentig ook tot zeer veel frustratie bij zoekers kon leiden. Het gebruik van linkpopulariteit om

zoekmachineresultaten te presenteren werd door de grondlegger van Teoma, de wiskundige Apostolos Gerasoulis, nog een stap verder gebracht. Bij Teoma en nu bij Ask, wordt de linkpopulariteit van een pagina bepaald aan de hand van de links die over het gevraagde onderwerp gaan, de zogenaamde onderwerpsspecifieke linkpopulariteit. Dit is wiskundig gezien een zeer complexe zaak om even snel op te lossen, maar Ask lijkt er zijn voordeel mee te doen.

De gebruiker zelf

Het nieuwe concept van presenteren van zoekresultaten, gebaseerd op linkpopulariteit, heeft de manipuleerders van zoekmachineresultaten in eerste instantie buitenspel gezet. Dat heeft echter niet lang geduurd. Met het populair worden van Google en het meewegen van linkpopulariteit door andere zoekmachines, werd de noodzaak groter om de zoekresultatenpagina’s gebaseerd op linkpopulariteit te beïnvloeden. De manipuleerders ontwikkelden daarom een arsenaal aan trucs om Google en de andere zoekmachines naar hun hand te zetten. We kregen *link-spamming* van de resultaten. Hiervoor worden *link-farms* ingericht, of er wordt massaal

via blogs gelinkt naar pagina’s die een hogere link-populariteit moeten krijgen. Waar trefwoorden en links geen betrouwbare indicaties meer geven om de rangorde van zoekmachineresultaten te presenteren, zijn de zoekmachines naarstig op zoek naar een volgende methode om de manipuleerders een slag voor te zijn. De sleutel lijkt in handen te liggen van de gebruikers zelf. Zoekmachines, of ze nu Google, Yahoo of MSN heten, beschikken over een schat aan informatie over hun gebruikers. Algemeen surf- en zoekgedrag, maar ook pc-gebonden surf- en zoekgedrag. De toename van ADSL speelt wat dat betreft de zoekmachines in de kaart.

Zoekgedrag

Neem nu het voorbeeld van Google. De zoekmachine weet vanaf welk IP-adres er op welke termen werd gezocht, en welke links er vanaf de resultatenpagina gevolgd werden. Google weet ook wanneer er door iemand vanaf een resultatenpagina op een advertentie, of vanaf een willekeurige webpagina op een *AdSense*-advertentie werd geklikt. Sterker nog Google kan al registreren wanneer er vanaf een pc een pagina met een willekeurige

AdSense-advertentie wordt opgevraagd. Daarnaast is Google sinds maart 2005 in het bezit van Urchin, een populair webstatistiekenprogramma, dat sinds november 2005 gratis ter beschikking wordt gesteld aan beheerders van websites. Daar wordt vrijwel ongemerkt een grote hoeveelheid data aan surfgedrag verzameld. Google weet daardoor waar gebruikers vandaan komen die een bepaalde pagina bezoeken en ook hoe lang dat bezoek duurt. Wanneer een pagina interessant is zal die ook wel langer bezocht worden. Op pc's waarop de Google toolbar, desktop of Google personal zijn geïnstalleerd leert Google nog meer over het surfgedrag op die specifieke pc. Daarnaast beschikt Google over informatie van de gebruikers van GMail en Google Talk. Kortom, Google beschikt over een heel arsenaal aan data over ons zoek- en surfgedrag.

Het wordt een hele kunst voor Google om deze schatten aan gegevens om te zetten in geschikte informatie om de volgende generatie aan zoekresultaten op een relevante wijze aan de gebruikers te presenteren. Dat moet natuurlijk ook nog eens binnen een split-second gebeuren. Het bedrijf experimenteert nu al met gepersonaliseerde zoekresultaten in Google Personal. Het lijkt er echter op dat dit nog maar de eerste voorzichtige zetten zijn op dit terrein. Langzamerhand zullen Google, en zijns gelijken, deze vorm van personalisatie steeds verder gaan beheersen. Maar de korte geschiedenis van zoeken op het web leert ons dat het onherroepelijk is dat er vervolgens ook weer methoden gevonden om de resultaten van gepersonaliseerde zoekresultaten te manipuleren.

De zoekmachines zelf worden tegenwoordig ook met argusogen bekeken. Het vergaren en analyseren van zoveel zoek- en surfdata wordt langzamerhand als bedreigend ervaren. Waar we eerst naar hartelust onze ziel en zaligheid (anoniem) aan het web toevertrouwden klinken er steeds meer kritische geluiden. Wie controleert straks de zoekmachines die alles van ons weten? <

Wouter Gerritsma is informatiespecialist plantenwetenschappen bij Bibliotheek Wageningen UR en blogt over dit soort onderwerpen op www.wouter.nl/blog.