

Zoeken op internet (herziene versie)

Naar aanleiding van een aantal vragen die mijn luisterend oor bereikten, ben ik eens gaan uitzoeken op welke manier je het beste te werk kunt gaan om specifieke informatie op het internet te vinden. Nu ben ik al een aantal jaren vertrouwd met het internet, vanaf 1990, dus zo'n beetje vanaf het moment dat het internet openbaar werd. In het begin zocht je informatie op met Telnet, haalde informatie op met FTP. Al snel ontdekte men dat het wel erg omslachtig was om eerst met de ene opdracht informatie op te zoeken en met een andere opdracht informatie op te halen, dus kwam er Archie die zoeken en ophalen combineerde. Wide Area Information System (WAIS) maakte het mogelijk om op specifieke bestandsnamen te zoeken. Voorwaarde was dan wel dat alle machines een index hadden waarop stond welke informatie op een machine aangeboden werd en in welke directory die informatie zich bevond.

Natuurlijk liet de volgende stap niet lang op zich wachten. In het begin van de jaren negentig was de ontwikkeling van het internet als openbaar medium in volle gang en kwamen ook andere bronnen dan die binnen de universiteit in gebruik. Daarbij hielp Gopher uitstekend. Binnen verschillende menu's kon men kiezen welke informatie men wenste: boeken, wetenschappelijke werken, telefoonnummers, etc. Deze informatie kon men binnenhalen met ingesloten FTP-procedures. Maar de grote ontwikkeling kwam pas met de invoering van de World Wide Web-servers (WWW). Hierdoor werd het mogelijk om ook op inhoudelijk informatie te zoeken. Weliswaar is WWW reeds in 1989 in Genève ontwikkeld door een Belg die hiervoor kortgeleden nog een onderscheiding ontving, maar tijdens een presentatie voor onze club van Jeroen Vanheste in 1995 kwam WWW praktisch niet aan bod. Zijn boekje dat we tijdens die presentatie kochten was van 1994 en was op moment van uitgave reeds niet meer bij de tijd!

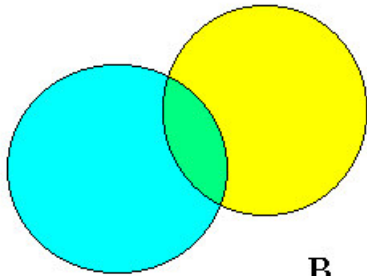
Verfijnen van het zoeken

Om het zoeken te verfijnen kan men gebruik maken van 'Booleaanse' actoren zoals AND, OR, XOR (exclusive OR), NOT en verder werden er later nog ADJ (adjacent) en NEAR aan toegevoegd. AND betekende: "Ik wil alleen die informatie als woord 1 EN woord 2 in te tekst voorkomen". OR wilde zeggen: "Ik wil alleen die informatie die woord 1 OF woord 2 bevatten". XOR is het tegengestelde van AND en wil zeggen alleen die informatie waar OF het woord 1 OF het woord 2 voorkomen maar niet beide. In het bovenstaande staat Booleaanse tussen aanhalingstekens omdat de gebruikte zoektechniek zich niet zuiver laat vertalen naar Boole. Vooral wanneer met in plaats van woorden tekens gebruikt zoals +, -, * dan laten die zich niet rechtstreeks vertalen naar Booleaanse begrippen. De meeste machines gebruiken nu + voor AND, - voor NOT en * als joker. Gebruikt men nu AND dan wordt dit vertaald naar +. Maar er zijn ook veel machines die de Engelse uitdrukkingen niet meer gebruiken en die tekens zelfs als zoektermen gebruiken! (Ilse)

Zoekt men nu informatie in de vorm van XOR (of, of) dan krijgt dit de vorm: +(woord 1 - woord2) +(woord1 -woord2). Bijvoorbeeld, ik heb de woorden 'boom' en 'struik' en ik wil alleen artikelen hebben waarin of het woord boom of het woord struik in voorkomt dan ziet het er als volgt uit: +(boom -struik) +(struik - boom). De trits actoren die door de meeste zoekmachines gebruikt worden zijn nu +, -, *, (,), "", de laatste voor letterlijke aanhalingen zoals bij het zoeken naar "What is" +WAIS. Verder zijn er machines die ook gebruik maken van =, <, >, etc. Je moet zoiets gewoon uitproberen, b.v. +date>2003. Als een zoekmachine een benadering "Geavanceerd" heeft dan kan men beginnen met dit in te vullen en daarna later in het zoekresultaat uitvinden hoe de machine te werk is gegaan, welke zoekactoren zijn toegevoegd.

Hierna tekeningen die drie zoektermen laten zien, A, B, C. De blauwe cirkel geeft alle gevonden informatie behorende bij zoekterm A. De gele cirkel geeft alle informatie behorende bij zoekterm B. De blauwe en de gele cirkel geven samen de informatie weer die bij A OF bij B hoort. Daar waar de cirkels overlappen (de 'Doorsnee' in de verzamelingenleer), die is groen gekleurd bevat de informatie die bij A EN B hoort. Dat wil zeggen de groene schijf geeft de informatie weer waarin zowel A als B voorkomen. Voegt men nog een zoekterm C toe dan geeft het zwarte schijfje de informatie weer waarin A EN B EN C voorkomen.

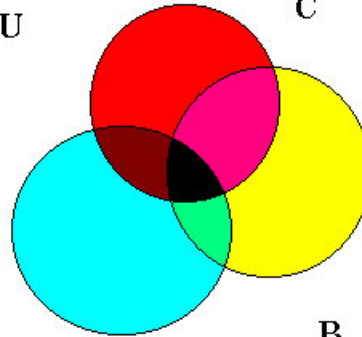
U



B

A

U



B

A

Wat niet, wat wel

Als je geen letterlijke ("") zoektermen hebt, dan moet je zoveel mogelijk woorden vermijden die veel voorkomen zoals lidwoorden, telwoorden, voornaamwoorden. Ook woorden waarvan je kunt verwachten dat ze veel zoekresultaten opleveren zoals water, maan, zon moet je voorkomen. Zelf heb ik het volgende aandachtlijstje:

- * Tussenwoorden (de, het, een, maar, en, etc)
- * Veel gebruikte woorden
- * Heteroniemen (woorden met verschillende betekenissen)
- * Verkeerd hoofdlettergebruik
- * Spelfouten
- * Woorden die niet bij elkaar horen

Wil je iets opzoeken dan schrijf je eerst in een gewone zin wat je wilt weten. Schrap daar alle veelvuldig voorkomende woorden weg en de overige woorden zijn je zoektermen. Het beste is om eerst één zoekterm in te voeren, de fijnste, en dan langzamerhand voeg je de overige grovere zoektermen toe. Het eindresultaat is dan de beste die je kunt krijgen. Onderandere bij Google kun je gewoon je hele vraag invoeren, die minder belangrijke woorden worden dan vanzelf weggelaten bij het zoeken.

In het eindresultaat worden de gevonden pagina's meestal weergegeven in de volgorde van meest voorkomende zoektermen. Er wordt ook opgelet of de veel voorkomende zoektermen vooraan in het document staan of achteraan. Dit verschilt per zoekmachine. Het beste is IXQUICK, deze geeft een resultaat waarbij de zoektermen gekleurd zijn zodat je snel kunt zien of de informatie voor jou van belang is.

In het algemeen is de juiste benadering:

- * Vaststelling van de zoektermen
- * Ordenen zoektermen, fijnste term eerst
- * Vaststellen van de te gebruiken actoren
- * Zoektermen en actoren stapsgewijs uitbreiden
- * Zoektermen aanpassen aan de resultaten
- * Zoektermen eventueel hergroeperen
- * Beoordeel de gevonden resultaten
- * Bij Metamachines kies de machine met gemiddelde aantal gevonden informatie

Waarom dat laatste? Machines met weinig resultaten hebben waarschijnlijk een te kleine gegevensvoorraad, machines met veel resultaten hebben waarschijnlijk veel dubbeltellingen.

Zoekmachines

Google: meest favoriete zoekmachine, gooit tussenwoorden weg, gebruikt altijd de + actor, werkt ook met 'klinkt als', waarschuwt voor homoniemen (woorden met verschillende betekenissen), voorkeuringstellingen gelden voor de gehele sessie, geavanceerde zoekmethoden op taal, gebied en manier van zoeken.

Yahoo: zeer geavanceerde zoekmachine, nu in twee versies 'home' en 'search' (search.yahoo.com), geoptimaliseerd volgens Google, standaard + tussen de zoektermen, waarschuwt voor homoniemen (is het dit wat je zoekt), voorkeuren worden per gebruiker vastgelegd, geavanceerde zoekmethoden.

Ilse: verouderde zoekmachine, uitfiltering moeilijk, geeft wisselende resultaten per uur en per dag, gebruik van AND geeft een + in de zoekopdracht, bij gebruik van + moet een spatie tussen de + en de zoekterm staan, geen + wordt uitgelegd als OR, heeft geen voorkeuren, heeft geen geavanceerde benadering.

Exite: goede zoekmachine, standaard + tussen de zoektermen, geeft categorieën waarop verder gezocht kan worden, heeft voorkeuringstellingen, heeft geavanceerde zoekmethoden.

AltaVista: Super goede zoekmachine, standaard + tussen de zoektermen, voorkeuringstellingen, geavanceerde zoekmogelijkheden, doet suggesties om het zoekresultaat te verbeteren.

Lycos: Goede zoekmachine, gebruikt Google gegevens, standaard + tussen de zoektermen, reclame boven de zoekresultaten, doorschakelen naar andere zoekmachines mogelijk, Nederlandse webplek.

Meta-zoekmachines

IXQUICK voorheen Surfboard: Metazoekmachine zoekt via een tiental andere machines, selecteert bronnen op relevantie met *, legio selectiemethoden met voorbeelden in het Nederlands, geavanceerde zoekmethoden, veel reclame tussen de resultaten, test ook op zogenaamde tussenwoorden als de, het, een, van, etc..

MetaSearch: Meta-zoekmachine, na het ingeven van de zoektermen worden de gevonden bronnen in Yahoo, AltaVista en Lycos aangegeven. Daarna kan men verder zoeken in een van de andere machines. Heeft voorkeuringstellingen en geavanceerde benadering.

MetaCrawler: Meta-zoekmachine, geeft resultaten uit Google, Yahoo, AltaVista, Ask Jeeves, About, Ouverture, FindWhat, etc. Groepeert resultaten naar verschillende gebieden b.v. medisch, politiek, geologie, e.d. Kent voorkeuzes en geavanceerde ondersteuning.

Agents

Agents zijn programma's die vanaf de gebruikersmachine zijn zoekwerk op het net uitvoert. Bekende namen op dit moment zijn: WebSeeker, WebSearcher, Firststop. De prijzen van de agents liggen in de grootte van euro 50,- tot euro 100,-.

Resultaten

Door mij zijn twee zoekreeksen samengesteld elk met drie zoektermen. In de eerste reeks (A) heb ik gewoon de termen geschreven met een spatie tussen de termen, in de tweede (B) diezelfde termen met een + teken ervoor, in de derde (C) weer een reeks met drie termen en de vierde (D) is reeks (C) met + voor de zoektermen.

Dit geeft de volgende aantallen bronnen in de genoemde zoekmachines:

Zie volgende bladzijde

Machine	A	B	C	D
Ilse	683.700	90	2.997.987	171
Google	240	240	340	340
Yahoo	183	183	229	229
Exite	33	49	22	38
AltaVista	105	105	162	162
IXQuick	61	49	51	39
Go	138	138	129	129
Lycos	175	175	204	204

Vergelijken wij de kolommen A en B, en C en D, dan zien we dat met uitzondering van Ilse dat de meeste machines gelijke of bijna gelijke resultaten boeken met of zonder extra +'s.
 Exite geeft meer resultaten met gebruik van +'s en IXQUICK geeft minder resultaten met +'s.
 De resultaten met Ilse zijn zo onzeker dat ik de door mij gevonden resultaten bijna niet durf weer te geven.
Zoek het zelf maar eens uit.

+++++JJVV 28 januari/18 februari 2004+++++