De toekomst van de zoekmachines

Wie zoekt die vindt

Soms wordt het web vergeleken met een gigantische bibliotheek waar iemand alle boeken door elkaar heeft gegooid en waar je geblindoekt op zoek mag naar de juiste informatie. Zoekmachines als Google mogen dan al een flinke hulp zijn, vaak zijn ze toch niet opgewassen tegen de chaos die er heerst op het wereldwijde web. Blijft dat zo? Computer Magazine organiseerde een eigen kleine zoektocht en wierp een blik op de toekomst van de zoekmachines.

Het W3C, oftewel het World Wide Web Consortium, de organisatie die instaat voor de ontwikkeling en de evolutie van het wereldwijde web is zich alvast bewust van de ongebreidelde groei van het wereldwijde web en alle problemen die dat met zich meebrengt. Maar zij werken al enige tijd aan een oplossing voor het probleem. Die oplossing heet semantic web en moet de opvolger worden van het huidige wereldwijde web. Wij vroegen aan Frank Van Harmelen, professor aan de Vrije Universiteit van Amsterdam en één van de onderzoekers die van dichtbij betrokken is bij de ontwikkeling van het semantic web, om uitleg.

Laten we beginnen met de meest voor de hand liggende vraag: wat is het semantic web precies?

Van Harmelen: “Om dat duidelijk uit te leggen, kunnen we best kijken naar de beperkingen van het huidige wereldwijde web. Dat is enerzijds een groot succes – niemand had tien jaar geleden kunnen voorspellen dat het wereldwijde web zoveel invloed zou gaan hebben op ons dagelijks leven – maar heeft tegelijkertijd ook een aantal grote beperkingen. Het web is op dit moment heel erg nuttig als je Engels of een andere taal kan lezen en beeldjes en foto’s kan begrijpen. Dat is iets wat mensen heel erg goed kunnen, maar voor computers is dat andere koek. Die kunnen totaal niet overweg met het huidige web, althans niet wat inhoud betreft. En dat maakt dat we momenteel maar heel weinig aan computers hebben als we op zoek gaan naar informatie op het web. De ondersteuning die we van computers krijgen is heel erg beperkt. Het enige wat je dure pc eigenlijk doet, is informatie op één plek ophalen, die naar een andere plek brengen en daarna die informatie op je scherm zetten. Maar die informatie begrijpen, combineren, interpreteren, selecteren, beoordelen, enz. dat wordt volledig aan ons overgelaten. De computer kan daar niet bij helpen omdat hij simpelweg niet begrijpt wat er precies op die pagina’s staat. Hoe linkt dat nu naar semantic web? Wel, het idee achter semantic web is dat we het huidige web proberen uit te breiden met extra informatie die ervoor zorgt dat computers die inhoud van webpagina’s wél kunnen begrijpen. Dat wil niet zeggen dat we het huidige web vergeten of dumpen: het semantic web is een uitbreiding, een extra laag bovenop het al bestaande wereldwijde web. Concreet betekent dat dat we een deel van de inhoud van de huidige pagina’s zo moeten aanpassen dat computers ze ook begrijpen. En daar zijn we intussen druk mee bezig.”

Hoe moeten we ons dat voorstellen?

Van Harmelen: “Er werden een aantal specifieke talen ontwikkeld die computers kunnen begrijpen en die op een steeds rijkere manier aan de computer uitleggen waarover een bepaalde webpagina precies gaat. Je kunt bijvoorbeeld in zo’n taal noteren dat er zoiets bestaat als de Vrije Universiteit in Amsterdam, dat er zo iemand is als Frank Van Harmelen en dat er tussen die twee een relatie is als ‘werkt voor’. Verder kan je definiëren dat er ook een ‘gebouw’ is en dat dat gebouw ‘onderdeel is van’ de Vrije Universiteit en dat Frank Van Harmelen ‘werkt in’ dat gebouw. Die relaties moet je uiteraard ook gaan definiëren. De relatie tussen mij en het gebouw is immers fundamenteel anders dan de relatie tussen de universiteit en het gebouw. Maar als je dat alles definieert in één van die talen en je gaat daarna zoeken naar mensen die ‘werken voor’de Vrije Universiteit of ‘werken in’ dat gebouw, begrijpt de computer wat je zoekt, omdat je het vooraf hebt geleerd. Op die manier zou de computer al veel meer ondersteuning kunnen geven bij je zoektocht naar informatie.”

Maar het hele systeem valt dus met de manier waarop je die informatie aan je computer ter beschikking stelt?

Van Harmelen: “Dat klopt ja. Een ander voorbeeld: als je zoekt naar het werkadres van Frank Van Harmelen dan zal je dat misschien niet vinden omdat ik te lui of te dom ben geweest om dat op mijn website te zetten. Maar als de computer weet dat ik werk voor de Vrije Universiteit en hij kent dat adres, kan hij aan de hand van die informatie misschien zelf afleiden welk adres hij moet doorspelen. Maar dan moet je de computer dus eerst de nodige achtergrondinformatie. Alles zal dus afhangen van de kwaliteit van wat wij de ontologie noemen. Dat is een verzameling van definities die begrippen en de relaties tussen die begrippen verklaart.”

De kwaliteit van de ontologie bepaalt met andere woorden de kwaliteit van de hulp die je computer kan bieden.

Van Harmelen: “Precies. Je kan een ontologie zien als een gestructureerde manier om de betekenis van woorden weer te geven voor een bepaald domein. Om terug te komen op dat voorbeeld van daarnet: daar moet je de computer gaan uitleggen wat een universiteit is, wat een werknemer is, wat de relatie tussen die twee is en hoe die relatie vorm heeft in de echte wereld, enz. Dat alles noemen we metadata. En dat is wat we nodig hebben om het semantic web te creëren. Zonder die metadata komt er geen semantic web.”

En waar komt al die metadata dan vandaan?

Van Harmelen: “Dat is de vraag die me tijdens presentaties het vaakst gesteld wordt (glimlacht). Wel, die komt niet uit de pen van individuele gebruikers. Als we kijken naar de oorsprong van het web – de eerste honderdduizend pagina’s of zo – dan kunnen we zeggen dat die nog met de hand geschreven zijn door mensen die achter hun computer zaten en webpagina’s in HTML maakten. Maar dat is allang niet meer zo. We hebben niet zomaar ruim 3 miljard pagina’s op het wereldwijde web. Die worden gegeneerd uit databases, die worden geschreven met of door specifieke programma’s, enz. Wel, die databases en programma’s zullen in de toekomst niet enkel HTML gaan genereren, maar ook metadata. Een makkelijk te begrijpen voorbeeld op dat vlak is de website van Amazon. Eigenlijk is dat gewoon een omgekeerde database. Alle informatie zit in de database en die database zetten ze om naar HTML-pagina’s, zodat wij die informatie kunnen lezen en begrijpen. Maar als dat kan, kan Amazon niet alleen maar HTML-pagina’s genereren. Dan kan diezelfde informatie ook in een andere, voor de computer te begrijpen taal, neergezet worden. En op die manier zou mijn persoonlijke *shopping agent* mij bijvoorbeeld kunnen assisteren bij het zoeken naar boeken of muziek die past bij mijn vooraf ingestelde persoonlijke voorkeuren. Dat is al één bron van metadata. Een andere belangrijke bron zijn gespecialiseerde programma’s die op een oppervlakkige manier natuurlijke taal kunnen begrijpen – Engels of Nederlands bijvoorbeeld – en daar metadata uit destilleren. Dat soort programma’s bestaat al en er zijn al bedrijven die daar hun geld mee verdienen. Metadata zal dus voor een groot deel automatisch of semi-automatisch gegenereerd worden.”

Die metadata moet dan natuurlijk wel gestandaardiseerd worden om uitwisselbaar en bruikbaar te zijn.

Van Harmelen: “Dat is inderdaad een belangrijk punt. Als ik het bijvoorbeeld heb over ‘werknemer’ en iemand anders heeft het over ‘employee’, dan moet de computer wel weten dat dat hetzelfde begrip is en dat hij, wanneer hij bijvoorbeeld naar ‘werknemer’ zoekt, ook ‘employee’ moet meenemen in die zoekopdracht. Maar dat is nu precies de winst die we hopen te halen met het semantic web. De huidige zoekmachines zijn – al is het een klein beetje overdreven gesteld – toch vooral bezig met character matching. Akkoord, ze zijn wel wat slimmer dan dat, maar de basis blijft toch het simpel vergelijken van cijfertjes en lettertjes. Daar moeten ontologieën dus wat aan gaan doen: er wordt niet langer alleen geschreven dat ik werknemer ben bij de Vrije Universiteit, maar dat begrip ‘werknemer’ en de relatie met andere begrippen zal gedefinieerd worden aan de hand van een achterliggende ontologie. Maar dat wil ook zeggen dat wanneer iemand anders het woord ‘employee’ gebruikt, hij moet verwijzen naar eenzelfde soort ontologie en dat die twee ontologieën ook aan elkaar moeten gelinkt worden. Daarna pas zal de computer weten dat die twee begrippen hetzelfde zijn. Om dat alles te doen heb je inderdaad gestandaardiseerde talen voor metadata nodig.”

Kan de computer die link zelf niet leggen? Is de technologie nog niet ver genoeg gevorderd daarvoor?

Van Harmelen: “Dat is op dit moment een *hot topic* in het onderzoek. Het kan al op experimentele basis in zorgvuldig gekozen testdomeinen, maar het kan nog niet in de wildgroei van het wereldwijde web. Maar ik verwacht wel dat rond die technologie een hele nieuwe commerciële markt zal groeien. Er zijn nu al bedrijven die ontologieën ter beschikking stellen. Ze hebben bijvoorbeeld een grote commerciële ontologie met termen als ‘werkgever’, ‘werknemer’, ‘product’, ‘adres’, ‘prijs’ enz. die aan elkaar gerelateerd zijn en waarnaar je mag linken als je daarvoor betaalt. Bovendien zorgen zij dan dat hun ontologie gelinkt wordt met andere ontologieën. Op die manier krijg je als het ware een soort semantische dienstverlening, die mensen toelaat om op die pagina’s makkelijker en sneller informatie te vinden.”

Gaat de gewone surfer iets merken van al die veranderingen? Het lijkt toch vooral om een back-office operatie te gaan. Wat verandert er concreet voor de gewone internaut?

Van Harmelen: “Het semantic web zal voor een groot deel een succes zijn als het onzichtbaar blijft. Al die technologie waar we het al over gehad hebben, zit inderdaad ‘onder water’. Het enige wat je als surfer zal merken is dat de kwaliteit van de resultaten van je zoekmachine zullen verbeteren. De huidige zoekmachines zijn heel goed in *recall*: alles wat er te vinden is, vinden ze ook. Maar ze halen heel wat minder goede resultaten als het om *precision* gaat: ze vinden niet alleen wat je nodig hebt, maar ook nog een hoop andere dingen die je niet kan gebruiken. Ik maak er nu een beetje een karikatuur van, maar je mag toch stellen dat de precisie nog flink omhoog zou moeten. Ook de manier waarop informatie zal worden aangeboden zal allicht veranderen. Als ik nu bijvoorbeeld mijn eigen naam in een zoekmachine intyp, krijg ik twee soorten resultaten: resultaten die over mij en mijn wetenschappelijk werk gaan, maar ook resultaten over het Nederlandse dorpje Harmelen. Probleem is dat de zoekmachine geen onderscheid maakt en die informatie nog gewoon door elkaar zet. Naarmate het semantic web verder zal evolueren moet de zoekmachine in staat zijn om te oordelen dat er twee verschillende soorten hits zijn en dat die ook apart moeten getoond worden of dat er gevraagd moet worden wat je precies zoekt: de persoon Frank Van Harmelen of het dorpje Harmelen.”

Het zoeken wordt dus eenvoudiger. Zijn er nog andere voordelen?

Van Harmelen: “Een belangrijk thema dat ik nog niet heb aangehaald, is personalisatie. Als ik op dit moment naar een website kijk en jij kijkt naar diezelfde website, zien we allebei hetzelfde. Maar dat is niet de ideale situatie natuurlijk, want jij hebt andere interesses dan ik. Neem Amazon maar weer: die zouden er toch alle voordeel bij hebben ze erin slagen om jou een andere pagina te laten zien dan mij, aan de hand van onze interesses? Je kan de personalisatie zelfs zo ver doordrijven dat je de stroom van informatie flink kan beperken: alles wat jou niet interesseert, hoeft immers niet aan jou aangeboden te worden.”

Hoeveel mensen werken er momenteel aan de ontwikkeling van het semantic web?

Van Harmelen: "Wel, W3C is eigenlijk maar een heel klein clubje. Het ledenaantal ligt heel hoog, maar de omvang van de staf is beperkt: wereldwijd gaat het om een paar dozijn mensen die naast het semantic web ook nog met andere dingen bezig zijn. Het echte werk wordt gedaan door mensen die werken bij de leden van W3C. Bij de werkgroepen rond semantic web vind je bijvoorbeeld mensen van IBM, Hewlett-Packard, Sun, Nokia, enz. Dat mogen niet de meest voor de hand liggende namen zijn, maar heel wat bedrijven hebben baat bij de ontikkeling van het semantic web. Hewlett-Packard ziet bijvoorbeeld mogelijkheden om het semantic web te gaan gebruiken via hun printers. Elke printer zou dan een soort *self describing device* worden: elke printer krijgt een eigen profiel, geschreven in een taal voor het semantic web. En wat gebeurt er dan? Jij loopt een gebouw binnen – een congrescentrum bijvoorbeeld - en alle printers daar maken zich kenbaar aan jouw laptop of pda. Als jij dan later een pagina wil printen, weet je laptop of pda al perfect waar elke printer staat en welke printer het meest geschikt is voor jouw specifieke taak. Een bedrijf als Nokia hoopt dan weer dat allerlei diensten beschikbaar kunnen gemaakt worden via hun mobiele telefoons. Logisch dus dat die bedrijven meewerken aan de ontwikkeling van semantic web. Niemand wil de trein missen."

Hoe zit het met de Googles, Altavistas en Yahoos van deze wereld? In welke mate zijn zij bezig met de ontwikkeling van het semantic web?

Van Harmelen: "Wel, ik heb onlangs nog met mensen van Google gesproken en het verbaasde me dat ze, hoe zal ik het zeggen, beleefd afwachtend waren. Ze waren heel goed op de hoogte van de laatste ontwikkelingen, maar ze vertelden dat ze voorlopig de kat uit de boom keken. Maar tegelijk merk ik wel dat ze toch al experimenteren met bepaalde dingen. Ken je de Open Directory? Dat is een project waarbij vrijwilligers handmatig webpagina's categoriseren. Wel, momenteel heeft Google voor een heel aantal zoekresultaten al koppelingen naar die gigantische database aan gegevens. Onderaan vind je dan een hyperlink naar de categorie waartoe dat specifieke resultaat behoort. Op die manier kan je gaan zoeken naar andere resultaten in die categorie. Dus zonder dat ze het zelf willen toegeven, gebruiken ze toch al semantische hulp. Ze kunnen ook niet anders hé. De populariteit van een zoekmachine wordt nog altijd bepaald door de kwaliteit van de zoekresultaten."

Hoe zit het met andere concrete toepassingen?

Van Harmelen: "Wel, W3C heeft een ontologie ontwikkeld voor het beschrijven van *device capabilities*. Die leert de computer bijvoorbeeld wat een telefoon kan, wat een printer kan, enz. en welke informatie die toestellen onderling kunnen uitwisselen. Dat is een heel concrete toepassing. En er bestaan ook al grote ontologieën voor heel specifieke sectoren. De biomedische sector heeft bijvoorbeeld al heel wat uitgebreide en goede ontologieën met medische termen. Ook de auto-industrie staat op dat vlak al erg ver. Daimler-Chrysler is zelfs een actief lid van de W3C-werkgroepen. Maar dat zijn toepassingen waar je als gewone surfer natuurlijk niet meteen mee in aanraking komt. Voorlopig is het semantic web vooral in de business-to business sector terug te vinden."

Wanneer mogen we de eerste concrete toepassingen voor gewone surfers, consumenten als u wil, verwachten?

Van Harmelen: "Ik denk dat er nu volop semantic web 'eilandjes' aan het ontstaan zijn binnen die specifieke sectoren. Die eilandjes zie ik op termijn naar elkaar groeien en op die manier ga je een écht semantic web krijgen. Pas dan zal de consument er meer van gaan merken. Wat ik wel snel zie gebeuren – en dat is iets waar Philips heel actief in is – is het ontstaan van ontologieën voor het aanbieden van media-inhoud. Een voorbeeld: er zijn nu websites waar de televisiegids aangeboden wordt. Wel, die websites zijn momenteel alleen maar door mensen leesbaar. Met zo'n ontologie kan ook de computer of je pda zo'n pagina lezen, die informatie vergelijken met je interesses en je dan attent maken op interessante programma's in de loop van de week. En zo kan ik me ook ontologieën voorstellen voor muziekgenres bijvoorbeeld, of voor filmgenres. Het enige wat jij dan nog moet doen, is aangeven in welke genres of subgenres je geïnteresseerd bent en de computer doet de rest. Dat zijn toepassingen die ik binnen dit enkele jaren wel verwacht."

Een moeilijke vraag om te eindigen: wanneer verwacht u de grote doorbraak van het semantic web?

Van Harmelen: “Dat is inderdaad een heel moeilijke vraag (grijnst). De toekomst voorspellen valt al niet mee, in de IT-sector is het nog moeilijker. En gebeurtenissen op het wereldwijde web voorspellen is helemaal onmogelijk. Ik heb er met Tim Berners-Lee – de geestelijke vader van het wereldwijde web – over gesproken en hij gebruikte toen de metafoor van een bobslee. Je moet een bobslee eerst in gang duwen, maar eens als hij snelheid begint te maken moet je je nog haasten om erin te springen of hij is weg zonder jou. Wel, we zitten met het semantic web nu nog volop in die duwfase. We moeten met de industrie praten om hen te overtuigen van het nut van het semantic web. Maar dat het semantic web er komt, daar twijfel ik niet aan. En ook Tim Berners-Lee ziet het semantic web als de volgende grote stap voor het wereldwijde web. Maar laat ik wat concreter zijn: ik zou heel teleurgesteld zijn, moesten er over twee of drie jaar geen zichtbare toepassingen zijn voor gewone surfers. En dan denk ik persoonlijk vooral aan e-commerce. Die sector heeft in mijn ogen het meeste voordeel bij personalisatie. Het hele grote web als geheel omvormen tot een semantic web zal nog wel wat meer tijd kosten. Maar het zal gebeuren.”

STREAMERS

“Tim Berners-Lee ziet het semantic web als de volgende grote stap voor het wereldwijde web.”

“Het semantic web zal voor een groot deel een succes zijn als het onzichtbaar blijft.”

“Ik zou heel teleurgesteld zijn, moesten er over twee of drie jaar geen zichtbare toepassingen zijn voor gewone surfers.”

“De overgang naar semantic web betekent niet dat we het huidige web vergeten of dumpen.”

Kaderstuk 1

Semantic web: de praktijk

Het semantic web mag dan al een prachtig idee zijn, het moet natuurlijk ook nog in de praktijk uitgewerkt worden. De zo noodzakelijke metadata komt er immers niet zomaar. Hoe leer je een computer om het web te begrijpen? Om te beginnen is het nodig om een grote database aan informatie samen te stellen die de computer begrippen en de relaties tussen die begrippen bijbrengt. Daarvoor zijn er een aantal hulpmiddelen beschikbaar: twee daarvan zijn XML en RDF. XML, oftewel eXtensible Markup Language, laat je toe om tags toe te voegen aan je webpagina’s, zodat ze voor de computer gestructureerd worden. XML zegt evenwel niks over de inhoud van de pagina’s. Daarvoor is er RDF, oftewel Resource Description Framework. Die taal maakt gebruik van *triplets* om betekenis uit te drukken. Elk triplet bestaat uit een onderwerp, een lijdend voorwerp en de relatie tussen die twee. Of zoals het ook wel wordt uitgedrukt: een *bron* die een *eigenschap* heeft die een bepaalde *waarde* kan aannemen. Die drie delen worden geidentificeerd door een *Universal Resource Identifier* (URI), die te vergelijken valt met een URL (de meest voorkomende URI). Op die manier kan elke gebruiker van het wereldwijde web dus op een simpele manier een nieuw concept of een nieuw begrip definiëren in RDF en die via het web beschikbaar maken. Maar daar stopt het verhaal natuurlijk niet, want twee databases kunnen nog altijd verschillende URI’s gebruiken voor hetzelfde concept. Daar komen de ontologieën dan ter hulp. Ontologieën zijn verzamelingen van definities die begrippen en de relaties tussen die begrippen verklaren. Heel concreet zal een ontologie dus de link leggen tussen de verschillende URI’s die hetzelfde concept identificeren. Dat alles samen zorgt voor het semantic web. Maar de kracht van het semantic web zal pas ten volle worden gebruikt, als men ook intelligente *agents* gaat ontwikkelen die het zoekwerk kunnen doen en die betrouwbare informatie – eventueel via digitale handtekeningen – kunnen onderscheiden van minder betrouwbare. Je hoort het al: er is nog heel wat werk aan de winkel.

Einde kaderstuk 1

Kaderstuk 2

Taalkeuzes

Om het semantic web volledig te kunnen ontwikkelen, zijn er natuurlijk een aantal gestandaardiseerde programmeertalen nodig. Zo zijn er ondermeer XML, RDF en OWL. XML staat voor eXtensible Markup Language en laat toe om tags te creëren die het definiëren, valideren, verzenden en interpreteren van data tussen verschillende applicaties of organisaties mogelijk maken. RDF is de afkorting van Resource Description Framework, een taal die een kader schept om metadata te benoemen. Dat gebeurt aan de hand van triplets die een concept definiëren. OWL tenslotte, staat voor Web Ontology Language en is de meest recent ontwikkelde taal. Het W3C ontwikkelde OWL omdat RDF een vrij primitieve taal is. Er was nood aan ontologietalen die rijker waren en die ook automatisch logisch redeneren toelieten. Vandaar dus OWL.

Einde kaderstuk 2

Kaderstuk 3

Van goegel tot elgoog

Het semantic web is nog niet voor morgen. We moeten het dus nog een tijdje doen met de bestaande zoekmachines. En dat betekent dan vooral Google, dat de rest zwaar in de schaduw stelt. Maar wist je dat er een pagina bestaat waarop alle verschillende taalversies van Google verzameld zijn? En wist je dat op die pagina ook heel wat Google-parodieën te vinden zijn? Google in alle mogelijke dialecten, Google gespiegeld, enz. Je vindt er zelfs parodieën die door Google zelf op het net gegooid zijn.

J www.chim.be/google/en/sites.php

Einde kaderstuk 3