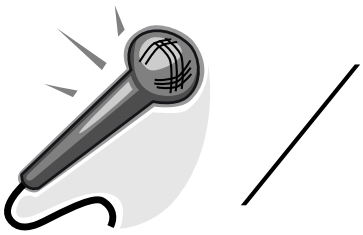
Frank den Hollander / f.j.den.hollander@rug.nl

Fotografie / Gerhard Lugard

PurpleSearch: de nieuwe slimme zoekmachine



Na een lange ontwikkelperiode komt rond het verschijnen van dit nummer van Pictogram de innovatieve zoekmachine PurpleSearch beschikbaar.

V.l.n.r. Ane van der Leij, André Keyzer en Bart Alewijnse

PurpleSearch is een zoekmachine die zo'n 120 wetenschappelijke literatuurbestanden in één keer doorzoekt. Door slimme zoektechnieken toont PurpleSearch de zoekresultaten uit de bestanden die het meest relevant zijn voor een bepaalde zoekterm. PurpleSearch is ontwikkeld bij de Universiteitsbibliotheek Groningen.
Een viergesprek met de ontwikkelaars.

Wie zijn jullie:

Bart Alewijnse: Ik ben nu net klaar met mijn studie Informatiekunde en werk als assistent-programmeur bij de afdeling Digitale Bibliotheekvoorzieningen van de UB.

André Keyzer: Ik ben bij dezelfde afdeling senior applicatieontwikkelaar.

Ane van der Leij: En ik werk als vakreferent Geschiedenis en Scandinavische talen bij de afdeling Informatievoorziening en Collectievorming van de UB.

Wat is PurpleSearch eigenlijk? Weer een combine harvester?

Keyzer: Een *smart combine harvester* dan op zijn minst. PurpleSearch is het slimme broertje (of zusje?) van RUG Combine.

Van der Leij: Net als RUG Combine is PurpleSearch een federatieve zoekmachine die het mogelijk maakt dat een hele trits bestanden gelijktijdig worden doorzocht bij het zoeken naar literatuur over een bepaald onderwerp. De zoekresultaten worden vanuit al die bestanden opgehaald en gepresenteerd in de interface van PurpleSearch. Daarin verschilt PurpleSearch niet van Combine. Maar Combine blijkt als vreselijk ingewikkeld te worden ervaren. André bedacht dat dat beter kon.

Worstelen met informatie

PurpleSearch gaat op een slimme manier om met de geoogste metadata. Die worden vastgehouden en geïndexeerd, zodat er een kennisnetwerk ontstaat over de inhoud van alle bestanden. Dat leidt tot een *recommender*-functie: het systeem kan aan de hand van de zoekterm die de gebruiker intikt, bepalen welke bestanden het meest relevant zijn

voor die zoekvraag. PurpleSearch toont dan ook de zoekresultaten uit díe bestanden.

Keyzer: En daarnaast ook zeer veel meer: een spellingchecker, plaatjes, allerlei toegevoegde informatie uit vele bronnen. En uiteraard een link naar full-textartikelen die tevoorschijn komen, en mogelijkheden om de resultaten te bewaren, te exporteren en te bewerken.

Hoe is het project PurpleSearch ontstaan?

Keyzer: Bij toeval eigenlijk. Uit verzamelde statistiek komt naar voren dat gebruikers van literatuurbestanden, zoals catalogi en bibliografische bestanden, nogal worstelen met het vinden van informatie. Al die verschillende interfaces en zoeksystemen maken het ook niet gemakkelijk en het gaat helaas dan ook vaak mis.

Alewijnse: Bovendien moeten ook onervaren gebruikers eerst al weten welke bestanden van belang zijn voor hun vakgebied of voor hun zoekvraag.

Keyzer: We kregen het idee dat we onze gebruikers daarbij zouden kunnen helpen. Dus zijn we gaan nadenken hoe we zouden kunnen helpen bij het zoeken en vinden van informatie. Dat is uiteindelijk PurpleSearch geworden.

Wie is er mee begonnen?

Keyzer: Het oorspronkelijke idee is van mij. Bart is er al in het begin bij gekomen als ontwikkelaar en meedenker natuurlijk. Ane was al betrokken bij RUG Combine en we hadden de inbreng van de vakreferenten ook nodig.

Alewijnse: Dat idee van simultaan zoeken is wel goed, maar het kan beter. We maken voor PurpleSearch dan ook gebruik van dezelfde search engine als Combine, de X-server van Metalib. Maar we hebben er van alles omheen geprogrammeerd. Een grote extra component naast het federatief zoeken is het maken van lokale indexen op de zoekresultaten, wat Ane zonet al noemde. En allemaal trucjes. Het project heette eerst ook LiveTrix, tricks dus, rond de X-server.

Is zo'n systeem er niet al?

Keyzer: In deze vorm is het er zeker nog niet. Er zijn natuurlijk wel systemen die gebruikers behulpzaam zijn bij een aantal onderdelen bij het zoeken, maar een systeem dat het hele zoekproces op deze manier van begin tot het eind begeleidt en gebruik maakt van databases van verschillende leveranciers en interfaces is uniek. Dit systeem trekt ook zeker de aandacht in binnen- en buitenland.

Jullie zeiden net dat het systeem een zoekterm herkent en dan de meest relevante bestanden doorzoekt. Maar wat als een zoekterm nog nooit eerder is gebruikt?

Alewijnse: PurpleSearch bekijkt op de achtergrond steeds àlle bestanden, waarbij het van zoekresultaten leert wat er in een bestand veel voorkomt. De beslissing om een gebruiker ergens naar toe te sturen, wordt dus gemaakt op grond van de inhoud van het bestand, niet van de zoekterm.

Maar als een zoekterm helemaal nog niet in de indexen voorkomt, tonen we de resultaten uit een dwarsdoorsnede van de grootste en belangrijkste bestanden. Het moet al heel raar lopen als daar niet toch bruikbare resultaten tussen zitten. En PurpleSearch leert er onmiddellijk weer wat bij.

Kun je zo'n zoekmachine niet gewoon kopen? Google bijvoorbeeld?

Alewijnse: Nee, een aantal leveranciers biedt wel zoekproducten aan die iets soortgelijks beogen, maar daar kleeft een (hoog) prijskaartje aan en ze voldoen gewoon niet aan alle wensen die we hebben. Daarom zijn we zelf maar begonnen.

Google? Een prima zoekmachine natuurlijk, met zeer veel informatie. Maar ook weer lang niet alles en verder 'helpt' Google gewoon niet genoeg om andere informatie te vinden.

Welke technologie zit erachter? Wat heb jij, Bart allemaal moeten leren/ontwikkelen?

Alewijnse: Ik spreek op het moment de zoekfunctionaliteit van het bestaande Metalib aan vanuit een webserver, met de Python-taal omdat

'PurpleSearch leert er telkens weer wat bij'

dat het bij het ontwikkelen makkelijker maakt om de steeds wat veranderende opzet en ideeën achterna te zitten. Er zit een database achter, die onder andere de relevantiedata voor bestanden opslaat. De *klinkt-als*-functionaliteit en wat andere functionaliteiten waren ooit begonnen als technische experimenten die sindsdien verder uitgewerkt zijn.

Web 2.0

Ik moest nogal wat over bibliotheekformaten en -standaarden bijleren, vooral over metadata. Redelijk wat code handelt dan ook gevallen af waar bestanden zich daar niet helemaal aan houden. Verder heb ik het nogal aan de stok gehad met verschillende browsers.

Welke databases worden er doorzocht?
Alewijnse: Heel veel.

Van der Leij: PurpleSearch heeft een lijstje van ruim 120 verschillende databases en catalogi waaruit informatie gehaald kan worden. Bij die 120 zitten natuurlijk de belangrijkste bronnen zoals *Web of Science* en *Pubmed* maar ook uit minder bekende databases zoals het economische bestand *RePec* wordt informatie opgeslagen. Het zijn in eerste instantie de bibliografische én full-textbestanden waarvoor de RUG-bibliotheken licenties hebben afgesloten. We zien natuurlijk graag dat die literatuurbestanden, waarvoor jaarlijks veel geld wordt betaald, goed door studenten en medewerkers worden gebruikt. Overigens zitten er ook wel andersoortige bestanden bij, zoals databases met bedrijfsgegevens, astronomische data of beelddatabanken voor kunsthistorici en dergelijke.

Wat zoekt PurpleSearch zelf uit?
Keyzer: Welke bronnen de beste zijn om informatie te vinden, dat hoeft een gebruiker niet meer zelf te beslissen. Dat doet PurpleSearch gewoon zelf. Deze feature is overigens volstrekt uniek; geen enkel ander systeem kan dat op dit moment.

Verder zoekt PurpleSearch zelf wel uit hoe gezocht moet worden in de geselecteerde bestanden, gebruikers hoeven niet te gaan uitzoeken wat de mogelijkheden en eigenaardigheden nu weer zijn van een database. Via een slim systeem wordt een database doorzocht en alle resultaten worden ook op identieke wijze gepresenteerd. Geen verschillende interfaces dus meer: één interface om te zoeken en één interface voor presenteren.

'One searches all' is daarom ook de slogan van Purplesearch.

Je zei zonet: dan leert PurpleSearch er weer iets bij?
Van der Leij: Ja. PurpleSearch is een zelflerend systeem. Bij iedere zoekactie werkt PurpleSearch de informatie over de zoekactie opnieuw bij. Dat gebeurt zowel bij termen die het systeem al kent, maar ook bij volstrekt nieuwe termen. Daarmee leert het systeem steeds meer over de gezochte informatie en breidt het kennisnetwerk zich steeds verder uit.

Keyzer: In dat opzicht is PurpleSearch ook een echt web 2.0-systeem. De gebruiker bepaalt door een zoekactie zelf op welke terreinen het systeem meer moet gaan leren. De gebruiker aan het stuur dus.

Bijkomend voordeel is daarbij ook, dat PurpleSearch daarmee toegesneden is op de gebieden die bij de RUG belangrijk zijn. Wanneer PurpleSearch bij een andere universiteit zou worden ingezet, zoals bijvoorbeeld nu al in Wageningen, zou er heel andere informatie opgeslagen worden. Het lijkt ons leuk om eens te gaan kijken hoe verschillend die sets zullen zijn (en om te kijken of we die sets ook kunnen koppelen als een soort gezamenlijke set).

Zijn alle bestanden relevant? Hoe weet je dat?
Van der Leij: Eigenlijk kan alleen de gebruiker beoordelen of een zoekresultaat relevant is. Maar we proberen zo goed mogelijke informatie te leveren, dat wil zeggen resultaten te geven uit die bronnen die door PurpleSearch als relevant voor de zoekvraag zijn beoordeeld.

Natuurlijk kan dat de nodige ruis opleveren, maar dat krijg je uiteraard minstens net zo erg als je zelf databases selecteert en doorzoekt. Lastig wordt het vooral bij ambigue termen die in verschillende vakgebieden kunnen opduiken.

Buitengewoon handig

We laten een gebruiker in ieder geval niet in de kou staan. Elk bestand is voorzien van een korte beschrijving, zodat een gebruiker gericht de keuze kan maken: nee, ik hoef de resultaten van dat informaticabestand niet, maar juist wél de resultaten uit een taalkundebestand, wanneer je hebt gezocht met *semantics* bijvoorbeeld.

Keyzer: PurpleSearch probeert ook altijd mee te denken en een alternatief aan te bieden. Daarbij moet je denken aan alternatieve of verwante zoektermen, die een gerichter resultaat kunnen opleveren.

Daarnaast kijken we kritisch naar zoekvragen die geen treffers opleveren. Daaruit valt veel te leren voor ons en die kennis wordt dan weer gebruikt om PurpleSearch verder te ontwikkelen.

Hoe doe je zoiets: alternatieve zoektermen aanbieden, spelling corrigeren etc.?
Alewijnse: Omdat PurpleSearch al veel geleerd

vervolg van pagina 5 Interview

heeft in de afgelopen tijd, kent het al heel veel woorden. Als een gebruiker iets invoert dat het systeem niet kent, wordt in de beschikbare woordenlijst gekeken of er iets is dat erop lijkt. Spelfouten kun je zo ook signaleren.

Daarnaast worden bij ieder woord of term de relaties met andere termen opgeslagen. Daarbij moet je denken aan relaties zoals: hoe vaak komt het ene woord voor in combinatie met een ander woord. Via deze relaties kunnen dan weer alternatieve zoekwoorden gesuggereerd worden.

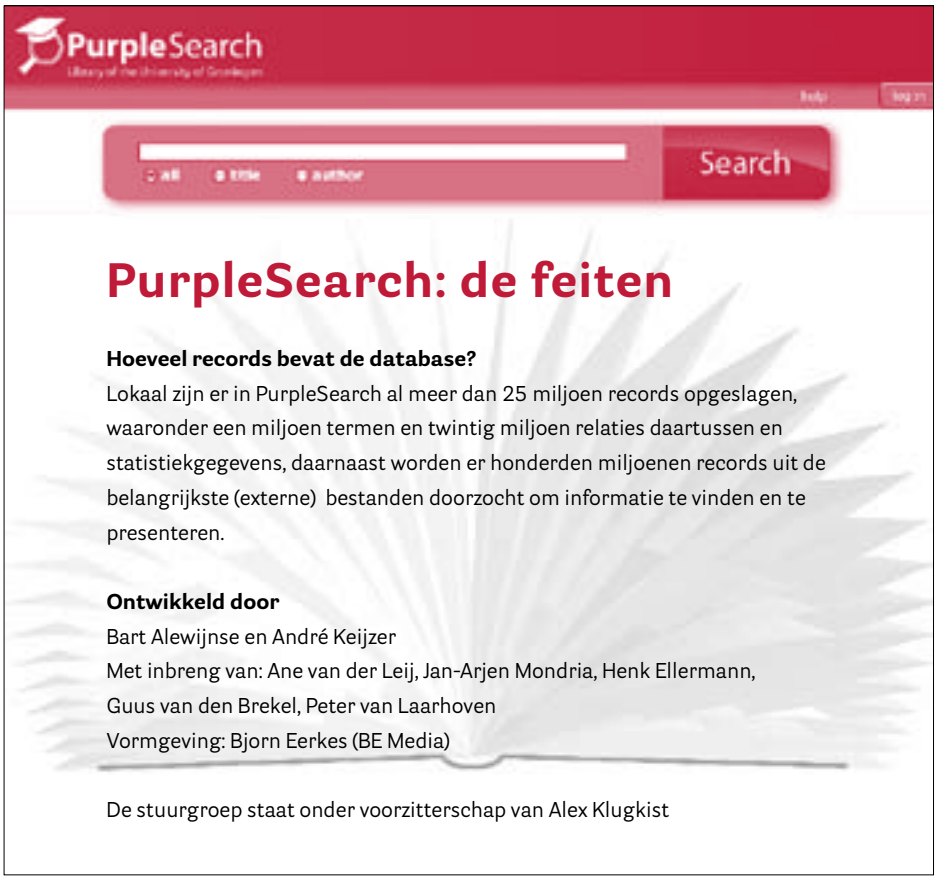
Waarom moet vanaf nu iedereen PurpleSearch gebruiken?

Van der Leij: Nou ja, iedereen... PurpleSearch is er in de eerste plaats voor onervaren gebruikers. Studenten, die nog niet precies weten welke bestanden ze zouden moeten gebruiken voor hun studie of voor hun onderwerp. Maar ik kan me ook goed voorstellen dat meer ervaren gebruikers PurpleSearch willen gebruiken voor een wat ruwere zoekactie, of wanneer men zich oriënteert op een onderwerp dat enigszins buiten het eigen vakgebied ligt. In dat laatste geval is PurpleSearch ook buitengewoon handig voor door de wol geverfde onderzoekers.

Maar als je nou echt een diepgaand en grondig literatuuronderzoek wilt doen over een onderwerp waar je goed in thuis bent, dan zou ik toch verwijzen naar de vakspecifieke bestanden zelf. Dan kun je ook gebruik maken van alle extra's die dergelijke bestanden met hun *native interfaces* bieden. Overigens kun je PurpleSearch dan wel gebruiken als toegang tot dat bestand. Er is altijd een link naar de bestanden zelf aan te klikken.

Geeft PurpleSearch niet teveel treffers?

Keyzer: Misschien wel, maar uiteraard niet meer dan je krijgt als je in de databases zelf zoekt. Omdat zoekresultaten uit meerdere databases in één keer getoond worden, lijkt het vaak veel. We gaan nu proberen om de gevonden resultaten met elkaar te combineren. Dubbele treffers vallen dan af.



PurpleSearch: de feiten

Hoeveel records bevat de database?
Lokaal zijn er in PurpleSearch al meer dan 25 miljoen records opgeslagen, waaronder een miljoen termen en twintig miljoen relaties daartussen en statistiekgegevens, daarnaast worden er honderden miljoenen records uit de belangrijkste (externe) bestanden doorzocht om informatie te vinden en te presenteren.

Ontwikkeld door
Bart Alewijnse en André Keijzer
Met inbreng van: Ane van der Leij, Jan-Arjen Mondria, Henk Ellermann, Guus van den Brekel, Peter van Laarhoven
Vormgeving: Bjorn Eerkes (BE Media)

De stuurgroep staat onder voorzitterschap van Alex Klugkist

• **http://purplesearch.rug.nl**