

Arnhem – Enschede – Nijmegen



een vergelijking in zoeksystemen

maart 2007

1. Inleiding

Wat onderscheidt de zoektechnologie van Aarhusportaal van andere zoekmachines? Om dat goed in beeld te brengen hebben we in dit rapport drie technieken naast elkaar geplaatst: een traditionele zoekmachine (exact zoeken met gemeente Arnhem), een moderne zoekmachine (zoeken met taaltechnologie met gemeente Enschede) en die van Aarhusportaal (zoeken met nog meer taaltechnologie met met documenten van gemeente Nijmegen).

2. Inhoud

3. Beginsituatie	2
4. Test.....	4
5. Waarnemingen	11
6. Conclusie	13



3. Beginsituatie

Eerst leggen de de situatie vast van waaruit we de diverse tests hebben afgenomen.

	Arnhem	Enschede	Nijmegen
Ontsluitings techniek	<u>Exact zoeken.</u> Deze techniek gaat uit van de door de gebruiker ingetypte karakters. Bekendste voorbeeld is Google. Resultaat: je vindt enkel wat letterlijk in de tekst staat. Geef je <i>vervuiling</i> op dan vind je alleen <i>vervuiling</i> , en bijvoorbeeld geen <i>verontreiniging</i> of <i>vuil</i> .	<u>Geëxpandeerd en genormaliseerd zoeken.</u> Dat is dus zoeken met taaltechnologie. De zoekvraag wordt uitgebreid met gerelateerde termen en er wordt gezocht op onderliggende concepten. Resultaat: je vindt veel meer relevante teksten maar ook meer ruis.	<u>Geëxpandeerd en genormaliseerd zoeken na disambiguering en analyse.</u> Door gebruik van extra technieken zoals semantische netwerken wordt rekening gehouden met betekenisverschillen in de zoekvraag en met de context van het antwoord. Resultaat: je vindt alles, terwijl de ruis toch minimaal is.
Lokatie	http://www.arnhem.nl/sites/internet_nieuw/secure/zoeken/	http://cms3.enschede.nl/ , zie zoekvenster rechtsboven. Er is nog een zoekvenster, onder <i>Loketten</i> , maar dit betreft de zelfde functionaliteit, maar dan toegepast op een deelbestand.	Er zijn zoekapplicaties onderzocht: – Zoekdienst: hier wordt gezocht op trefwoorden. http://aarhus-1.irion.nl/aarhusSearch/web/checks.jsp?database=14&queryLg=nl . – Dialoogdienst: hier wordt gezocht door hele vragen op te geven. http://aarhus-1.irion.nl/kenniswijk/web/init.do?database=null&queryLg=nl
Fabrikanten	Onbekend	Carp Technologies BV uit Enschede	Irion Technologies BV uit Delft



Corpus	Een vergelijking van zoekmachines behoort eigenlijk uit te gaan van hetzelfde corpus. Wij zijn echter niet in staat om onafhankelijk van de zoekmachines gebruik te maken en ze te voeden met hetzelfde corpus.		
	Alle webpagina's van de gemeentelijke site, plus Digitaal Loket (de gemeentelijke producten, circa 500 documenten over gemeentelijke onderwerpen), allemaal te vinden op http://www.arnhem.nl/inter/index.html .	Alle webpagina's van de gemeentelijke site plus Digitaal Loket (de gemeentelijke producten, circa 500 documenten over gemeentelijke onderwerpen als <i>wonen voor ouderen</i> , <i>huldiging sportkampioenen</i> en <i>kinderopvang</i>) en OndernemersLoket, allemaal te vinden op http://cms3.enschede.nl/ .	334 milieudocumenten die op de site van van gemeente Nijmegen staan http://www2.nijmegen.nl/ en van het gemeentelijke afvalbedrijf DAR, http://www.dar.nl/ .
Formaten	HTML, PDF, MSWord	HTML, PDF, MSWord	HTML, PDF, MSWord
Tijd	5-9 maart 2007	5-9 maart 2007	5-9 maart 2007



4. Test

We hebben aan drie zoekmachines een aantal zoekvragen gesteld. De zoekvragen hebben steeds betrekking op een bepaald taalonderwerp. Op deze manier wordt onderzocht hoe de zoekmachines met taalvarianten omgaan. Hieraan ligt de veronderstelling ten grondslag dat zoekmachines krachtiger zijn wanneer ze meer grip hebben op de natuurlijke taal waarin de teksten zijn gesteld.

	Zoekvraag	Arnhem	Enschede¹	Nijmegen
Typografische varianten (schrijf- of typfouten)	<i>kontainer</i>	0 resultaten, <i>container</i> geeft er 147	13 resultaten, <i>container</i> geeft er 62	Zoekdienst 44 resultaten, <i>container</i> geeft er 47
	<i>gemeentehuis</i>	0 resultaten, <i>gemeentehuis</i> geeft er 88	17 resultaat, <i>gemeentehuis</i> geeft 43 resultaten	Zoekdienst 37 resultaten, bij instelling score 40% is gelijk aan resultaten <i>gemeentehuis</i>
Spellingsvarianten (spelling volgens Van Dale)	<i>projekt</i>	9 resultaten, zoeken op <i>project</i> levert 1792 resultaten op	42 resultaten, terwijl zoeken op <i>project</i> 1541 resultaten geeft	Zoekdienst 40 resultaten, allemaal <i>project</i>
				Dialogdienst idem
	<i>produktie</i>	11 resultaten, <i>productie</i> geeft 181 resultaten	0 resultaten, <i>productie</i> geeft 2 resultaten	Zoekdienst 23 resultaten, <i>productie</i> geeft 33 resultaten.
				Dialogdienst 31 resultaten, <i>productie</i> geeft 34 resultaten

¹ In geval van Enschede komen de gevonden treffers niet direct in beeld. Hierdoor is de relevantie van het antwoord voor de vraag moeilijker vast te stellen. De treffers zijn wel te zien als men kiest voor de optie *Uitgebreid zoeken*.



Morfologische varianten (normalisatie, herleiden tot stamwoord)	<i>bodemsaneren</i>	108 resultaten (<i>bodemsanering</i> levert ook 108 resultaten op)	45 resultaten, waaronder <i>bodemsanering</i> , maar <i>bodemsanering</i> geeft 83 resultaten	Zoekdienst 15 resultaten, waaronder <i>bodemsanering</i> , <i>bodemsanering</i> geeft ook 15 resultaten.
				Dialoogdienst 7 resultaten waaronder <i>bodemsanering</i> , <i>bodemsanering</i> geeft ook 7 resultaten
	<i>container/containers</i>	147 resultaten <i>container</i> 147 resultaten <i>containers</i>	62 resultaten <i>container</i> 85 resultaten <i>containers</i>	Zoekdienst 47 dezelfde resultaten voor zowel <i>container</i> , als <i>containers</i>
				Dialoogdienst 13 resultaten <i>container</i> , 17 resultaten <i>containers</i>
	<i>fietsen/fietsten/gefietst</i>	177 resultaten <i>fietsen</i> 10 resultaten <i>fietsten</i> 6 resultaten <i>gefietst</i>	205 resultaten <i>fietsen</i> 20 resultaten <i>fietsten</i> 49 resultaten <i>gefietst</i>	Zoekdienst 22 resultaten <i>fietsen</i> 17 resultaten <i>fietsten</i> 17 resultaten <i>gefietst</i>
				Dialoogdienst 7 resultaten <i>fietsen</i> 7 resultaten <i>fietsten</i> 0 resultaten <i>gefietst</i>
	<i>kadaster/kadastraal</i>	51 resultaten <i>kadaster</i> 141 resultaten <i>kadastraal</i>	34 <i>kadaster</i> 52 <i>kadastraal</i>	Zoekdienst 43 resultaten <i>kadaster</i> 43 idem <i>kadastraal</i>
				Dialoogdienst 14 <i>kadaster</i> 0 <i>kadastraal</i>



Syntactische varianten (woordverbuigingen en woordvolgorde)	<i>vervuiling grondwater</i>	39 resultaten	64 resultaten	Zoekdienst 46 resultaten, allen relevant: <i>verontreinigende, verontreinigd, verontreiniging, ernstige verontreinigingen van grond en grondwater, verontreiniging oppervlaktewateren, etc.</i>
				Dialoogdienst 46 resultaten, allen relevant.
	<i>grondwater verontreiniging</i>	51 resultaten	36 resultaten	Zoekdienst 48 resultaten
				Dialoogdienst 51 resultaten
Semantische varianten (synoniemen, homoniemen)	<i>vervuiling</i>	153 resultaten (digitaal loket 0 resultaten)	68 resultaten, geen over <i>verontreiniging</i>	Zoekdienst 45 resultaten, ook over <i>verontreinigingen, etc.</i>
	<i>verontreiniging</i>	137 (digitaal loket 1 resultaat)	128 resultaten, geen over <i>vervuiling verontreinigen</i> geeft 22 resultaten	Dialoogdienst 55 resultaten
				Zoekdienst 52 resultaten, ook over <i>vervuiling verontreinigen</i> geeft 50 resultaten ook <i>verontreiniging</i>
<i>gemeentehuis</i>	88 resultaten	42 resultaten, geen over <i>stadhuis</i>	Dialoogdienst 52 resultaten, idem <i>verontreinigen</i>	
				Zoekdienst 37 resultaten, waaronder ook <i>stadhuis</i> en <i>stadskantoor</i>



				Dialogdienst 31 resultaten
	<i>stadhuis</i>	849 resultaten	346 resultaten	Zoekdienst 37 resultaten, waaronder ook <i>gemeentehuis</i> en <i>stadskantoor</i>
				Dialogdienst 34 resultaten
	<i>lantaarnpaal</i>	25 resultaten (digitaal loket 0 resultaten en gemeentegids 0 resultaten)	23 resultaten, in de weblijst het 1 ^e document gaat over <i>paalwoningen</i>	Zoekdienst 1 resultaat: <i>straatverlichting</i>
				Dialogdienst 2 resultaten: <i>Digitale balie straatverlichting + Bel en Herstellijn</i>
	<i>straatverlichting</i>	16 resultaten (digitaal loket 2 resultaten en gemeentegids 1 resultaat)	22 resultaten, zoekmachine gaat ook zoeken op <i>straat</i>	Zoekdienst 1 resultaat: <i>Digitale balie straatverlichting</i>
				Dialogdienst 2 resultaten: <i>Digitale balie straatverlichting + Bel en Herstellijn</i>
Meerwoord uitdrukkingen (worden waar nodig verbonden)	<i>Bouw verordening</i>	413 resultaten op de losse woorden (<i>bouwverordening</i> levert 160 documenten op)	404 resultaten waarbij in de weblijst het 61 ^e document een bouwverordening is	Zoekdienst vindt direct resultaten, in het 1 ^e document met <i>bouwverordening</i>
				Dialogdienst vindt direct resultaten, in het 1 ^e document met <i>bouwverordening</i>



<p>Langere zoekvragen (meerdere zelfstandig-naamwoorden plus voorzetsels en lidwoorden)</p>	<p><i>Waar laat ik mijn oude wasmachine</i></p>	<p>7 resultaten waarvan aantal relevant</p>	<p>10 resultaten, irrelevant: <i>Enschede Overbruggingsregeling ?oude wetters?(2x)</i> en een aantal raadsvoorstellen</p>	<p>Zoekdienst 13 resultaten (niet allemaal relevant), maar het 2^e document is <i>DAR: Afvalinzameling: Wat te doen met..</i> dat is de relevante informatie. Ook documenten 4 en 5 zijn relevant: <i>DAR afvalinzameling grof vuil</i> en <i>Digitale balie: afval van huishoudens</i></p> <p>Dialogdienst Komt uit bij drinkwater besparing</p>
	<p><i>Mijn oude wasmachine is niet opgehaald wat nu</i></p>	<p>1 resultaat wel relevant</p>	<p>23 resultaten, allemaal irrelevant (<i>reclame, bestemmingsplan, bouwvergunning, etc.</i>)</p>	<p>Zoekdienst 2 resultaten, relevant: <i>DAR: Afvalinzameling: Wat te doen met...</i> en <i>DAR afvalinzameling grof vuil</i>.</p> <p>Dialogdienst 1 resultaat: drinkwater besparing, gaat wel over wasmachines maar is niet relevant.</p>
	<p><i>Ik heb last van astma</i></p>	<p>0 resultaten</p>	<p>204 resultaten, allen op de treffer <i>last</i>, niets over astma, 1 resultaat in weblijst: melding woon en leefomgeving</p>	<p>Zoekdienst 5 resultaten, relevant: luchtkwaliteit en gezondheid, aangegeven met pagina's waar het over astma gaat</p> <p>Dialogdienst 13 resultaten waarvan de 1e 8 gaan over luchtkwaliteit en gezondheid</p>



	<i>Er ligt in de straat een stoeptegel los</i>	9 waarvan aantal wel relevant	35 resultaten, waarvan in weblijst 2 relevant	<p>Zoekdienst 2 resultaten, 1e is zeer relevant: een folder over losse stoeptegels</p> <p>Dialoogdienst 31 resultaten waarvan de 1e meest relevant weer de folder over losse stoeptegels.</p>
	<i>wat een lawaai bij het cafe om de hoek</i>	4 resultaten maar niet relevant	6 resultaten, 1 relevant	<p>Zoekdienst (bij een instelling score 40%) 5^e document relevant</p> <p>Dialoogdienst direct goed doorverwezen naar rubriek <i>geluid</i> met relevante documenten</p>
Zoeken met voorselectie (thema, land, etc.)		Alleen op datum	Alleen op datum	<p>Zoekdienst: onder Uitgebreid zoeken op thema, postcode, informatiesoort, datum</p> <p>Dialoogdienst: zit in de dialoog verstopt.</p>

5. Waarnemingen

Uit de uitgevoerde tests kunnen een aantal waarnemingen worden gedaan. We presenteren hier de belangrijkste:

1. Schrijf- en typfouten

Arnhem kan niet omgaan met schrijf- en typfouten, Nijmegen en Enschede wel. Het grote verschil tussen Nijmegen en Enschede is dat de laatste minder consequent is. Soms doet Enschede het heel goed (*vergunnin* of *verguning*), maar soms ook helemaal niet: (*gemeentehuis* geeft 1 resultaat, namelijk *huis*, en *gemeentehuis* 31). Vermoedelijk is hier gebruik gemaakt van een vaste lijst van typfouten en spellingsvarianten. Bij Nijmegen geeft *gemeentehuis* hetzelfde resultaat als *gemeentehuis*.

2. Spellingsvarianten

Op het gebied van spellingsvarianten komt Arnhem heel slecht en Enschede vrij slecht uit de bus. Dit laatste is vreemd want zoeken op *projekt* geeft wel documenten met de treffer *project* (26), maar veel minder dan wanneer wordt gezocht op *project* (1528). Nijmegen is wel goed in spellingsvarianten en geeft steeds dezelfde identieke resultaten.

3. Morfologische varianten

Qua normalisatie (= herleiden tot stamwoord) doet Arnhem helemaal niks. Enschede gaat soms goed (*vervuilen* levert ook op *vervuilende* en *vervuild*), maar dan ook weer helemaal niet. Bij Nijmegen is de normalisatie veel constanter.

4. Syntactische varianten

Hetzelfde beeld geldt voor de voor syntactische varianten. Nijmegen geeft een constant beeld en Enschede weer een wisselend beeld, bijvoorbeeld bij *vervuiling grondwater* en *grondwater verontreiniging*.

5. Semantische varianten

Op het gebied van semantische varianten (zoals synoniemen en homoniemen) scoort Arnhem het slechts en Nijmegen het best.

6. Meerwoord uitdrukkingen

Qua meerwoord uitdrukkingen doet Arnhem het slecht, Enschede over het algemeen goed, maar weer minder consequent dan Nijmegen.

7. Langere zoekvragen

Met langere zoekvragen blijkt Nijmegen beter om te kunnen gaan, met name de Dialoogdienst. Dat komt door de combinatie van automatische classificatie en zoeken met behulp van een semantisch netwerk.

8. Soorten documenten

Arnhem, Enschede en Nijmegen kunnen allen meerdere soorten formaten aan: HTML, PDF, MSWord.

9. Kwantiteit antwoorden

We kunnen helaas geen uitspraken doen over de “recall” van de verschillende methoden (= het aantal goede antwoorden dat is gevonden ten opzichte van wat aan goede antwoorden in het databestand aanwezig is). Daarvoor is een veel grondiger onderzoek nodig, waarbij alle methoden op eenzelfde corpus worden toegepast. Wel kunnen we iets zeggen over de



“precision” (= de hoeveelheid ruis in de antwoorden). Met Arnhem vinden we de meeste irrelevante antwoorden, met Nijmegen de minste.

10. Pagina niveau

Een belangrijk verschil van Nijmegen boven Enschede is het ontsluiten tot op paginaniveau. De gebruiker zit meteen op de juiste pagina. Bij Enschede moet men nog verder gaan zoeken, hetgeen vaak een hinderlijke klus is.

11. Relevantie

Bij Arnhem en Nijmegen kan de gebruiker snel de relevantie van een antwoord voor de vraag vaststellen. Bij beiden wordt gewerkt met snippets (= de originele stukken tekst waaruit de relevantie blijkt). Enschede komen de snippets enkel bij Uitgebreid zoeken in beeld. Bij Nijmegen wordt bovendien met scoringspercentages gewerkt, hetgeen extra steun biedt: hoe hoger het percentage hoe bruikbaar het antwoord.

12. Meer of minder antwoorden

Bij de Nijmegen-versies kan de gebruiker het aantal antwoorden afstellen door het scorepercentage hoger of lager te zetten. Bij 100% krijgt men de minste antwoorden en de minste kans op ruis. Bij 10% heeft men de meeste antwoorden en ook meer kans op ruis.

13. Voor- en naselectie

Alle methoden kennen voorselectie (= de gebruiker kan een deel van het bestand selecteren om in te zoeken). Bij Arnhem en Enschede op datum. Bij Nijmegen Zoekdienst op thema, postcode, datum en informatiesoort. Bij Nijmegen Dialoogdienst is voorselectie in de dialoog opgenomen. Naselectie (= aantal antwoorden terugbrengen door extra keuzes te bieden) is bij Arnhem en Enschede niet mogelijk, bij de applicatie van Nijmegen wel.

14. Samenvattingen

Bij de Nijmegen-versies wordt bij iedere gevonden pagina een automatisch gegenereerde samenvatting meegeleverd. De samenvattingen zijn opgebouwd uit zinnen van die pagina, en daarom meestal niet perfect. Wel geven ze een prima indruk van de inhoud van de pagina.

15. Ranking

Bij alle systemen lijkt de volgorde waarin de antwoorden worden gepresenteerd te zijn gebaseerd op beste resultaat. Alleen Enschede lijkt daar soms van af te wijken (bij *lantaarnpaal* gaat het 1e antwoord over paalwoningen en het 2^e pas over lantaarnpalen).

16. Talen

Alle applicaties ontsluiten enkel Nederlandse teksten. De Nijmegen-versies zijn in principe ook meertalig in te zetten, de Arnhem-versie vermoedelijk niet, van de Enschede-versies weten we dit niet.

17. Zoektijd

De zoektijd is voor alle applicaties 1 tot 3 seconden. Dit is vooral voor Nijmegen en Enschede een goed resultaat gezien het feit dat bij zoeken met taaltechnologie meer processen moeten worden afgehandeld.

18. Highlighting

Bij de Nijmegen-versies lichten in geval van HTML-documenten de treffers op, zodat de gebruiker deze makkelijk kan vinden. Bij de Arnhem-versie komt geen highlighting voor en bij de Enschede-versie alleen in de snippets.

6. Conclusie

De eerste conclusie van dit onderzoek luidt dat zoeken met taaltechnologie meer oplevert dan exact zoeken. Zowel in Enschede als in Nijmegen is taaltechnologie verwerkt. We zien duidelijk dat daardoor allerlei relevante teksten naar boven komen die anders ongetwijfeld verborgen zouden zijn gebleven. Kortom: met taaltechnologie is de burger beter af.

De tweede conclusie luidt dat deze taaltechnologie klaarblijkelijk niet overal op dezelfde wijze wordt toegepast. In de Nijmegen-versies lijkt dat consistentere te zijn gebeurd dan bij Enschede. Mogelijk speelt de inzet van semantische netwerken hierbij een rol. Ook kan meespelen dat bij Enschede is gekozen voor een “dedicated” oplossing, men heeft de taaltechnologie afgesteld op het typische corpus van Enschede. Bij Nijmegen is voor een generieke oplossing gekozen, men werkt met standaard settings, er is dus niet geoptimaliseerd. Dat kan voor de ene index goed uitpakken en voor de andere minder goed. Mocht dit laatste het geval zijn, dan kan door het wijzigen van instellingen alsnog snel een optimalisatie worden bereikt.

De derde conclusie is dat bij Nijmegen veel meer op de gebruiker is gelet. Er zijn allerlei hulpmiddeltjes die het de gebruiker een stuk makkelijk maken om niet alleen het beste document te vinden, maar ook om het vervolgens te raadplegen. Een heel prettig middel is de pagina-presentatie: je zit meteen bij de goede tekstpassage. Een ander is de dialoogvorm van de Nijmeegse Dialoogdienst: je wordt heel comfortabel naar het antwoord geleid. En ook de automatische samenvattingen worden als ondersteunend ervaren.

Dit onderzoek is uitgevoerd door ir. Karin de Boom en drs. John Verheijden (medewerkers MOOI Informatiebeheer).