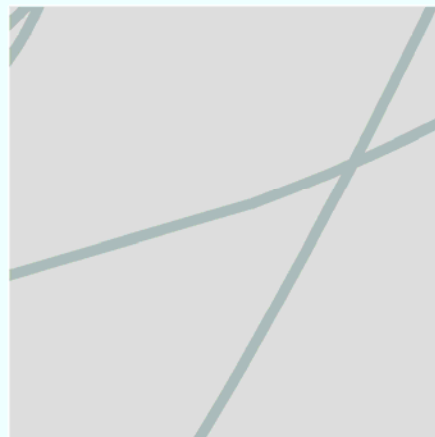


AquaBrowser

Y.Buma

Handleiding AquaBrowser van MediaLab

26 juni 2003





Colofon

AquaBrowser

Handleiding AquaBrowser van MediaLab

Stichting Digitale Universiteit
Nijenoord 1, 3552 AS Utrecht
Postbus 182, 3500 AD Utrecht
Telefoon 030 - 238 8671
Fax 030 - 238 8673
e-mail buro@diguni.nl

Auteurs

Y.Buma

Copyright

Stichting Digitale Universiteit

Deze uitgave is binnen het consortium van de Digitale Universiteit vrijelijk te gebruiken, mits voorzien van adequate bronvermelding. Niets uit deze uitgave mag buiten het consortium openbaar worden gemaakt, verspreid en/of verveelvoudigd door middel van internet, druk, fotokopie, microfilm of op welke andere wijze dan ook zonder voorafgaande schriftelijke toestemming van het bureau van de Digitale Universiteit.

Datum

26 juni 2003

Kenmerk

97120



Inhoudsopgave

1	Inleiding	5
1.1	Behoeftte van de gebruiker centraal	5
1.2	De AquaBrowser	5
1.3	Neem een kijkje	6
2	Booleaans versus associatief zoeken in het onderwijs	7
2.1	De boomstructuur van de traditionele zoekmachine	7
2.2	De associatieve zoekmachine	7
3	De AquaBrowser	9
	Prioriteit zoekwoorden	9
3.2	De associaties	9
3.3	Voorbeelden van zoekwegen en resultaten	10
3.4	Specifieke eigenschappen van de AquaBrowser	10
3.5	Wat is er mogelijk met de Standaard AquaBrowser?	10
4	AquaBrowser in het onderwijs	12
4.1	De Kennisbank	12
4.1.1	De content	12
4.1.2	Ervaringen van studenten	12
4.2	De AquaBrowser	13
4.2.1	Introductie van de AquaBrowser	13
4.3	Samenvattend	14
5	Enige technische informatie	15
5.1	Minimale PC en client eisen	15
5.2	Wat kun je zelf spideren en waarvoor is maatwerk nodig	15
	Hoe spiderde de AquaBrowser	15
	Statische sites	15
	Dynamische sites	15
	Javascript	16
5.3	Eendagscursus 'werken met de AquaBrowser'	16
6	Samenwerking met MediaLab	17
6.1	De licentie van DU	17
6.2	Samenwerking met MediaLab	17
6.3	Reacties uit de beroepspraktijk	18
	Bijlage 1	19
	Beschrijving van verschillende extra functionaliteiten AquaBrowser	19
	Locatiemodule	19
	Treeview Table	19
	Translation Language	19
	Search accent	19
	Subject recognition	19
	Boolean Search	19
	User statistics	19
	Sort / Ranking mechanisme	20
	Aqua site Bobby proof (t.b.v. visueel gehandicapten)	20
	Spideren van PowerPoint en Excel	20
	Zoeken in mandjes	20
	Bijlage 2 Gedachtenwolk communicatieproject	20

Bijlage 3 Hoe bouwt de AquaBrowser zijn thesaurus op?	21
De AquaBrowser zelf kan als Java applet opgenomen worden in de webbrowser, maar hij kan ook andere vormen aannemen.	22
Bijlage 4: Verschillende associatieve zoekmachines onderzocht	23
Keuze voor AquaBrowser	24
Bijlage 5: Nadere Informatie over het spideren van sites	25
Inleiding	25
1. Statische en dynamische sites	25
Versleutelde adressen	26
Crawler pagina's en sitemaps	26
Javascript	27
Bijlage 6: Korte inhoud introductie cursus AquaBrowser bij MediaLab	28



1 Inleiding

De AquaBrowser is een onderdeel van Com2Know, een kennisnetwerk voor het communicatieonderwijs en de communicatieberoepspraktijk. Dit kennisnetwerk is ontwikkeld door twee HBO-opleidingen communicatie: de Fontys Hogeschool Communicatie uit Eindhoven en de School voor Communicatiemanagement uit Utrecht.

Het hoger beroepsonderwijs is sterk praktijkgericht. Steeds grotere delen van de curricula worden thematisch of projectmatig aangeboden. Aan de hand van praktijkgerichte situaties en problemen wordt gezocht naar divergerende oplossingen. Zeker in het communicatievak bestaan vaak geen eenduidige oplossingen. Deze handleiding is geschreven vanuit het perspectief van twee HBO communicatie opleidingen. Het is dan ook mogelijk dat niet alles wat hierin gesteld wordt opgaat voor alle andere opleidingen.

1.1 Behoeftte van de gebruiker centraal

Al werkend aan het project leerden we hoe ook de behoefte van de gebruiker invloed heeft op de keuze van de techniek. We hebben gemerkt dat studenten op hun zoektocht naar informatie en kennis vaak niet precies (kunnen) omschrijven waar ze naar op zoek zijn. Al dan niet expliciet gebruiken ze dan vormen van creatief denken, die in booleaanse zoeksystemen ontmoedigd worden. De boomstructuur maakt dat een zoeker die niet precies weet wat hij zoekt, gemakkelijk de verkeerde weg inslaat en op dood spoor raakt. Het kan ook voorkomen dat studenten juist een te beperkte kijk hebben op de problematiek, (zij hebben immers nog niet alle noodzakelijke kennis) waardoor zij mogelijkheden over het hoofd zien. Ook dan zal de boomstructuur hen op een verkeerd been zetten.

Daarom zijn we, na eerste gewerkt te hebben met een booleaans systeem dat een vrij uitgebreide invoer van metadata vergde, op zoek gegaan naar systemen die deze bezwaren kunnen ondervangen. In de AquaBrowser vonden we de oplossing¹ en bij de Digitale Universiteit de financiering die deze duurdere oplossing haalbaar maakte.

1.2 De AquaBrowser

De AquaBrowser werkt via een full-text zoekstelsel, waardoor het niet nodig is data te koppelen aan uitgebreide metadata. Dit scheelt veel handmatige indexering. Bovendien zoekt de AquaBrowser associatief, waardoor de zoeker voortdurend op nieuwe ideeën wordt gebracht. Daarbij worden diverse soorten relaties gelegd. Ook laat de techniek van de AquaBrowser typefouten toe. Tenslotte presenteert de browser deze verwijzingen in een visueel aantrekkelijk spinnenweb, dat de zoeker stimuleert tot verder zoeken. Al met al een browser die zowel voldoet aan de onderwijskundige als beheersmatige wensen van ons project.

¹ Onderzochte alternatieven als Autonomy, Sharepoint Portal en Muscat vielen af wegens gebruiksmogelijkheden, kosten of flexibiliteit. (zie ook bijlage 4)



1.3 Neem een kijkje

Als u na het lezen van deze handleiding nieuwsgierig geworden naar de AquaBrowser en hoe die functioneert bij Com2Know, dan is het wellicht interessant om even een kijkje te nemen op onze site: www.com2know.com. U ziet dan de AquaBrowser zoals die bij ons functioneert.

Fontys Hogeschool Communicatie
Hogeschool van Utrecht, School voor Communicatiemanagement



2 Booleaans versus associatief zoeken in het onderwijs

Studenten hebben meestal een nog beperkte algemene ontwikkeling en een beperkt relatienetwerk. Daardoor realiseren zij zich vaak niet dat een problematiek complexer is dan deze hij op het eerste oog lijkt. Tevens missen ze, mede door het beperkte netwerk, de mogelijkheid snel informatie te verzamelen. Hierdoor (en door het tijdgebrek binnen het lesprogramma) moet soms genoeg worden genomen met minder goede analyses en daardoor met een minder goed advies. Maar vooral: een minder optimaal leertraject.

Veel studenten zoeken hun informatie al dan niet bewust op een associatieve manier als zij met een project bezig zijn. Deze associaties samen kunnen worden weergegeven in een gedachtenwolk (mindmap) waarin allerlei zaken die te maken hebben met het project zijn verwerkt. Deze gedachtenwolk is in het begin nog niet gestructureerd, maar ontstaat al werkend en wordt stapsgewijs ingevuld. Dit was een van de resultaten van het Com2Know project. (In bijlage 2 is een voorbeeld van een dergelijke gedachtenwolk te vinden.)

Als een project in werkelijkheid associatief wordt opgebouwd en als een student nog geen grote kennis over het vakgebied heeft, dan is het de dus vraag of de traditionele booleaanse zoekstructuur wel het meest geschikt is voor deze groep gebruikers?

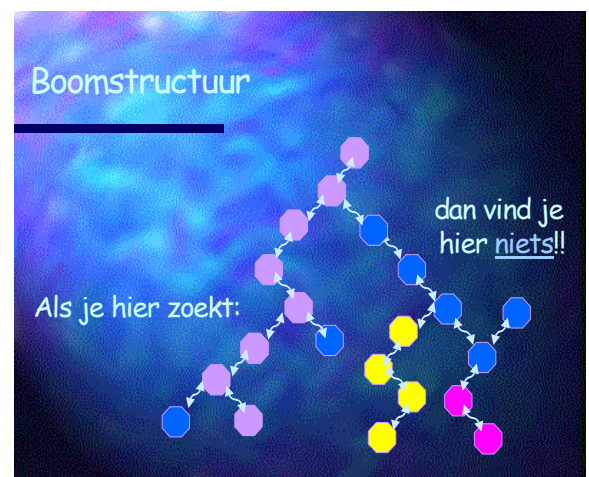
2.1 De boomstructuur van de traditionele zoekmachine

Bij de booleaanse zoekwijze is het eigenlijk een vereiste dat je goed weet wat je zoekt en waar je het zoeken moet. Kortom, dat je een duidelijk beeld hebt van de te verwachten informatie.

Er wordt immers gewerkt met een boomstructuur die je maar op één manier kunt doorlopen. Een eenmaal ingeslagen weg 'moet' tot het einde worden afgelopen. Ben je ergens verkeerd afgeslagen dan is de kans dat je het gezochte alsnog vindt vrijwel nihil (zie bijgaand plaatje).

Het is natuurlijk altijd mogelijk dat je per ongeluk iets vindt dat je niet verwachtte, maar de kans is klein. Daarmee is ook de kans klein dat een student bij het zoeken naar informatie aanloopt tegen invalshoeken waarmee hij het probleem kan benaderen waar hij nog niet aan gedacht had. En dat is nu juist één van die processen die in het onderwijs plaats moeten vinden.

Wij kwamen dus tot de conclusie dat een boomstructuur niet de meest geschikte manier is om studenten te helpen bij het zoeken naar informatie.



2.2 De associatieve zoekmachine

Een associatieve zoekmachine werkt precies omgekeerd. Documenten worden doorzocht en er wordt een woordenlijst opgebouwd van woorden die regelmatig in relatie tot het gezochte woord

worden aangetroffen. Deze worden door de AquaBrowser in een gedachtenwolk weergegeven en tegelijk met de zoekresultaten getoond.



Op deze manier wordt de student niet alleen geholpen aan relevante documenten, maar ook aan termen om het zoekproces mee voort te zetten. Hem worden dus associaties geleverd die hem op nieuwe ideeën brengen en daarmee ook zijn kennis en inzicht vergroten.

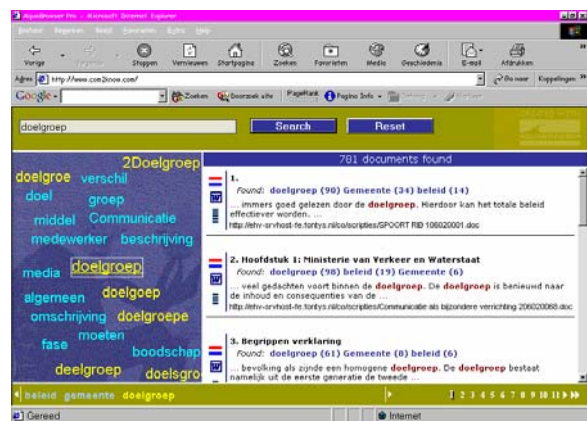
3 De AquaBrowser

De Aquabrowser is een associatieve zoekmachine die ontwikkeld is door het Nederlandse bedrijf MediaLab. De AquaBrowser doorzoekt aangeboden content full text en bouwt daarmee een eigen woordenlijst² op die gebruikt wordt om de associaties aan te leveren. Deze associaties worden gepresenteerd in een gedachtenwolk.

De AquaBrowser 'kopieert' de gespiderde content, die in een aparte database wordt opgeslagen. Dit heeft als voordeel dat de content sneller wordt geproduceerd dan als men naar de locatie waar de content is opgeslagen zou moeten gaan. In het resultaat wordt echter ook een link naar de originele vindplaats opgenomen zodat de zoeker ook daar kan kijken. Om te zien hoe dit in de praktijk werkt kunt een kijkje nemen bij www.com2know.com.

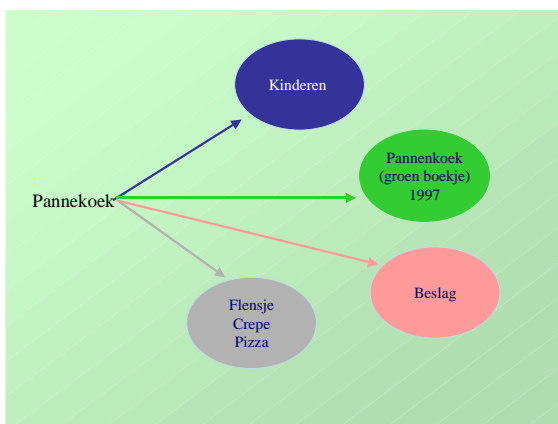
3.1 Prioriteit zoekwoorden

Een sterk punt van de AquaBrowser is de wijze waarop het zoekresultaat wordt opgebouwd. Niet (zoals in de traditionele systemen) door steeds het eerste zoekwoord als hoofdwoord te nemen en dan steeds fijner toe te spitsen op later toegevoegde woorden. Nee, het laatste zoekwoord (de laatste associatie van de student) is het uitgangspunt. Vervolgens wordt gekeken in hoeverre eerdere zoektermen ook in het document terug te vinden zijn. Steeds met een minder zware weging naarmate de zoektermen eerder in het zoekproces zijn gebruikt.



3.2 De associaties

De AquaBrowser presenteert in haar gedachtenwolk verschillende soorten associaties. De meest interessante voor een niet al te precieze student is misschien wel dat ook als er een spelfout in het zoekwoord zit, de AquaBrowser ook het woord in de juiste spelling zal tonen. Je hoeft dan niet helemaal opnieuw te beginnen of te denken dat er niets te vinden is.



In het totaal levert de AquaBrowser de volgende alternatieven:

- fuzzy alternatieven (het lijkt erop, dus ook een woord met een typefout)
- vertalingen
- associaties (aan de context gerelateerde woorden)
- woorden die vaak samen voorkomen
- synoniemen

² Voor uitleg over de wijze waarop de AquaBrowser haar woordenlijst opbouwt zie bijlage 3: Hoe bouwt de AquaBrowser zijn thesaurus op?



3.3 Voorbeelden van zoekwegen en resultaten

Onderstaand worden enkele willekeurige voorbeelden van een associatieve zoekweg getoond. Een zoekweg die is uitgevoerd op de AquaBrowser van Com2Know. De content bestaat dus vooral uit voor het vak communicatie relevante content.

Het eerste woord is steeds het woord waarmee het zoekproces is gestart. Het tweede woord is één van de woorden die voorkwamen in de gedachtenwolk, daar is weer verder meer gezocht. Het derde woord komt uit de gedachtenwolk die het resultaat is van het eerdere zoekproces, et cetera.

Voorbeeld A: (1) cultuur, (2) gedrag, (3) manieren, (4) toevoegen.

Dit levert bij voorbeeld documenten op over:

- ◆ het gebruiken van kunst bij organisatievraagstukken;
- ◆ het toevoegen van informatie aan kennisbanken en
- ◆ besturen met visie.

Voorbeeld B: (1) communicatietheorie, (2) communicatietermen, (3) communicatiestromen, (4) communicatiepatronen.

Dit levert bij voorbeeld documenten op over:

- ◆ het bereiken van veranderingscommitment;
- ◆ communicatiecultuur en
- ◆ interpersoonlijke communicatie in taak- en werkrelaties.

Voorbeeld C: (1) veranderen, (2) omgeving, (3) vindplaats, (4) gossip, (5) overdracht.

Dit levert bij voorbeeld documenten op over:

- ◆ de rol van roddel in organisaties;
- ◆ overdracht van waarden en normen in het onderwijs en
- ◆ klantvolgsystemen.

Natuurlijk is de content die door de AquaBrowser gespiderd is bepalend voor de resultaten en de gedachtewolk.

3.4 Specifieke eigenschappen van de AquaBrowser

Daarnaast heeft de AquaBrowser een aantal specifieke eigenschappen die, mits bekend, voordeel opleveren bij het werken met deze zoekmachine. Deze zijn:

1. Bij de zoekresultaten vindt men niet alleen een verwijzing naar de vindplaats van het document en de mogelijkheid hiernaar 'door te klikken', maar ook naar het door de AquaBrowser gescande (en op de server opgeslagen) kopie van het document. Deze mogelijkheid wordt bovenin aangegeven bij de zoekresultaten. De verwijzing naar het oorspronkelijke document staat onderin. Een voordeel van 'klikken' op het gescande document is dat het resultaat aanzienlijk sneller op het scherm verschijnt.
2. Zit men in dit document dan verschijnen ook de extra zoekmogelijkheden:
 - ◆ next en previous result voor het snel bekijken van de gevonden documenten en
 - ◆ next en previous hit voor het snel vinden van de woorden die in de zoekterm voorkwamen.
3. Voor een nieuwe zoekactie moet wel gebruik worden gemaakt van de knop 'zoek opnieuw'. Gebeurt dit niet dan wordt het zoekproces voortgezet en niet een nieuwe zoekactie gestart.

3.5 Wat is er mogelijk met de Standaard AquaBrowser?

De standaardfunctionaliteit van de AquaBrowser omvat (naast de eerder genoemde mogelijkheden) het volgende:

- ◆ Er kan een eigen thesaurus worden toegevoegd. Zowel een whitelist (woorden die moeten worden getoond) als een blacklist (woorden die juist niet mogen worden getoond);
- ◆ Standaardformaten die gespiderd kunnen worden zijn:
 - ◆ Websites (*.html en *.htm)
 - ◆ MS work en MS Word (*.doc)
 - ◆ PDF
 - ◆ Tekst (*.txt)
 - ◆ En elk ander document dat in een van deze vormen kan worden geconverteerd;
- ◆ Standaard Engelse, Duitse en Nederlandse woordenlijsten en vertaallijsten zijn standaard inbegrepen;
- ◆ De lay-out van de toolset kan aan de eigen wensen worden aangepast.



4 AquaBrowser in het onderwijs

Zoeken met de AquaBrowser is, zoals in het vorige hoofdstuk uitgelegd, anders dan zoeken met andere gangbare zoekmachines zoals Google. Omdat de werking van deze laatste zoekmachines iets is waar studenten (en docenten) aan gewend zijn is het belangrijk de verschillen en mogelijkheden goed te begrijpen. De belangrijkste zijn:

1. Bij Booleaans zoeken verengt het zoekpad met elke extra zoekterm die wordt toegevoegd. Bij het zoeken met de AquaBrowser wordt de laatste zoekterm als uitgangspunt genomen en worden documenten gerangschikt naar de mate waarin naast deze zoekterm ook de eerder zoektermen zijn gevonden. Dit gebeurt echter in omgekeerde volgorde. Dus eerst de laatste, dan de een na laatste etc.
2. De AquaBrowser levert een gedachtenwolk die kan leiden tot nieuwe zoektermen. Mogelijkheden kunnen zijn o.a. associaties, andere spellingswijzen, andere talen.

Het zal duidelijk zijn dat bij implementatie van de AquaBrowser in het onderwijs dit verschil goed moet worden uitgelegd. Gebeurt dat niet dan zal een deel van de potentiële gebruikers snel afhaken. Bij onze eerste pilots met de AquaBrowser waren er groepen studenten die wel een introductie kregen in de AquaBrowser en groepen die direct aan het werk werden gezet. Deze laatste groep haakte over het algemeen sneller af.

4.1 De Kennisbank

4.1.1 De content

Een zoekmachine kan nog zo geavanceerd zijn, uiteindelijk bepaalt de kwaliteit van de informatie die doorzocht wordt of hij (en daarmee de kennisbank) wordt gebruikt. Vandaar toch even wat opmerkingen over de te spideren content. Althans zoals dat door Com2Know is toegepast. Criteria voor content waren:

- alle content (websites, documenten etc.) moet of informatie bevatten die direct of indirect relevant is voor het vakgebied;
- de informatie moet hoogwaardig zijn;
- websites moeten veel content hebben en regelmatig worden ververs. Een voordeel van dit criterium is dat de content actueel wordt gehouden. Dit betekent wel dat er veel tijd gespendeerd moet worden aan het zoeken naar de juiste websites. Het blijkt dat erg veel websites slechts heel beperkte informatie bevatten of erg commercieel gekleurd zijn;
- voor websites gold tevens het criterium dat de verwachting moet bestaan dat deze over een paar jaar nog steeds bestaat.

Hoewel deze eisen in principe algemeen zijn, zijn ze voor de AquaBrowser extra belangrijk. Er is een behoorlijke hoeveelheid content nodig om te garanderen dat de associaties die de AquaBrowser oplevert ook zinvol zijn. (Het spideren van content is over het algemeen redelijk eenvoudig. In hoofdstuk 5.2. wordt hier nader op ingegaan)

4.1.2 Ervaringen van studenten

In ons project leverde de content wel de volgende ervaringen op met studenten³ die aan de pilots meededen:

³ Over het algemeen waren de reacties van studenten uit de verschillende pilots op onderstaande punt gelijk.

- Studenten verwachten geen uitgebreide en lange documenten in een zoekmachine. Zij gaan in eerste instantie op zoek naar korte (hapklare) brokjes informatie. Uitgebreidere informatie verwachten zij in boeken en dergelijke. Door ons is echter ook veel uitgebreide content aan de kennisbank toegevoegd. Bij voorbeeld: scripties, en omvangrijke nota's afkomstig van overheidssites.
- Studenten verwachten in eerste instantie geen communicatiekundige kennis in de kennisbank. Daar zijn boeken voor. Blijkbaar is er voor hen een verschil in het type informatie dat zij in een bepaald medium verwachten.
- Er is positieve waardering voor het feit dat de AquaBrowser alleen relevante content geeft.
- En een zeer onverwachte reactie: men is ontevreden dat de gevonden PDF-documenten niet 'geknipt en geplakt' kunnen worden. Blijkbaar komen zij met het doorzoeken van sites zelden zo ver dat PDF-documenten gedownload worden.

Het is aan te raden met deze reacties rekening te houden bij de introductie van een kennisbank.

4.2 De AquaBrowser

Uit de door ons uitgevoerde pilots blijkt dat het aan te raden is de AquaBrowser met goede toelichting te introduceren bij studenten. Hoewel de AquaBrowser zeker geen ingewikkelde zoekmachine is, is ze wel erg verschillend van de meest gebruikte zoekmachines.

Een goede toelichting van de AquaBrowser zal voorkomen dat een deel van de studenten snel terug zal vallen op de gebruikelijke zoekmachines. Dit gevaar zal overigens kleiner worden naarmate het gebruik van de AquaBrowser normaler is binnen de opleiding en bekend is dat via de zoekmachine snel relevante content te vinden is. Ook dit bleek bij ons bij de latere pilots. Studenten van deze pilots deelden mee ook na afloop van de pilot nog gebruik te maken van de AquaBrowser bij hun verdere studie.

Een ander reden om bij de introductie van de AquaBrowser veel aandacht aan de uitleg te geven is dat men dan optimaal gebruik zal maken van de mogelijkheden. Een aantal mogelijkheden zijn niet opvallend en lopen daarom het risico over het hoofd te worden gezien.

4.2.1 Introductie van de AquaBrowser

Bij introductie van de AquaBrowser is het verstandig de volgende zaken aan de orde te laten komen:

1. Hoe werkt associatief zoeken en welk soort resultaten geeft de AquaBrowser in haar gedachtewolk? (*zie hiervoor elders in deze handleiding*)
2. Hoe wordt het zoekpad en de zoekhistorie opgebouwd? Wat is het verschil met booleaanse zoekmachines? Bij een booleaanse zoekmachine wordt de zoekterm steeds verder verengd, het laatste zoekwoord is een steeds zoekterm. De AquaBrowser doet eigenlijk het omgekeerde. Vanuit de laatst genoemde zoekterm wordt gezocht, waarbij de eerder genoemde zoektermen voor verfijning zorgen.
3. Hoe werkt de zoekhistorie? Onderaan de pagina worden de begrippen waarop gezocht wordt in een historische volgorde weergegeven. Het is mogelijk om op een eerder weer te klikken kan worden teruggegaan naar het resultaat op dat moment. Van daaruit kan dan een andere zoekweg worden gevolgd.
4. Opnieuw zoeken. Het intypen van een nieuwe zoekterm leidt dus alleen tot verfijning van het reeds bestaande zoekresultaat. Wil men met een geheel nieuwe zoekactie aan het werk dan

is het noodzakelijk om op de resetknop rechts boven in het scherm te klikken. De zoekhistorie wordt dan gewist en men kan opnieuw beginnen.

5. Hoe worden de zoekresultaten gepresenteerd? Het schermdeel waar de zoekresultaten gepresenteerd worden laat per gevonden item de volgende zaken zien:
 1. *De titel van het gevonden document. Wordt hier op geklikt dan kom je uit op de in de Aquadatabase opgeslagen versie van het document (hierover later meer). De zoektermen die in het document teruggevonden zijn, met tussen haakjes het aantal keren dat deze term voorkomt in het document.*
 2. *Een korte toelichting op het gevonden document*
 3. *Een link naar het document op zijn oorspronkelijke vindplaats. Je komt dan dus bij voorbeeld terecht op de site waar het document gevonden is en kunt van daaruit de site verder doorzoeken.*
6. Wat is er nog meer te zien? Links naast de presentatie van de resultaten zijn extra gegevens te vinden over het gevonden document. Met een vlaggetje wordt de taal aangegeven en met een icoontje het type document (word, website, pdf et cetera)
 - Als gebruik wordt gemaakt van het opgeslagen document (zie 5.1), dan verschijnt niet alleen het resultaat sneller op het scherm, maar ontstaan ook extra zoekmogelijkheden:
 - *next en previous result voor het snel bekijken van de gevonden documenten en*
 - *next en previous hit voor het snel vinden van de woorden die in de zoekterm voorkwamen.*
 - *De gevonden zoekwoorden worden gearceerd weergegeven.*

4.3 Samenvattend

In onze ervaring zijn de volgende zaken belangrijk bij het succesvol werken met de AquaBrowser in het onderwijs.

- ◆ Zorg dat er voldoende content is zodat de opgeleverde associaties ook zinvol zijn;
- ◆ Geef een gedegen instructie als de AquaBrowser voor het eerst door studenten wordt gebruikt.
- ◆ Wees duidelijk over het wat voor soort documenten de AquaBrowser bevat; zorg daardoor dat de verwachtingen juist zijn.

In essentie komt dit neer op twee zaken: zorg dat wat er gevonden kan worden de moeite waard is en manage de verwachtingen.



5 Enige technische informatie

5.1 Minimale PC en client eisen

PC eisen voor een serversysteem waarop de AquaBrowser is geïnstalleerd:

- | | |
|--------------------|---|
| * Operatingsysteem | - MS Windows 2000 server |
| * Webserver | - Microsoft IIS 5.0 |
| * Webbrowser | - MS Internet Explorer 5.5. of hoger Java 1.1 |
| * Processor | - Intel Pentium III 550 Mhz of sneller |
| * Systeem RAM | - Minimaal 512 MB RAM (content afhankelijk) |
| * Hard disk | - Minimaal 10 GB (content afhankelijk) |
| * CD-rom | - Quad speed of hoger. |

Vereisten voor een Clientsysteem waarop de eindgebruiker werkt met de AquaBrowser:

- | | |
|--------------------|--|
| * Operatingsysteem | - MS Windows 95/98/NT/2000/XP |
| * Webbrowser | - MS Internet Explorer 5.0 en hoger, Java 1.1. |

5.2 Wat kun je zelf spideren en waarvoor is maatwerk nodig ⁴

De AquaBrowser wordt geleverd met een tool waardoor het relatief makkelijk is zelf websites en andere content te spideren. Speciaal voor DU is deze tool zelfs wat verfijnd, zodat ook gemakkelijk kan worden aangegeven dat slechts een deel (en welk) van een website gespiderd moet worden. Deze eenvoudige acties zijn na enige instructie goed zelf te doen door de gebruiker (systeembeheerder).

Hoe spidert de AquaBrowser

Een spider begint vaak met de zogenaamde "homepage" van een website als invoer. Deze homepage pagina wordt door de spider als eerste opgeslagen. Als deze pagina is opgeslagen wordt de pagina doorzocht op links, d.w.z. verwijzingen, naar andere pagina's op de website. Als dat het geval is worden ook deze pagina's opgehaald. Voor elk van deze pagina's wordt dan weer gekeken of hier ook links op voorkomen, etc, etc. Standaard volgt de spider alleen de links binnen de aangegeven website. Eventuele links naar andere websites worden dan niet gevolgd.

Bij het spideren van sites door de AquaBrowser moet onderscheid gemaakt worden tussen het spideren van statische en dynamische sites.

Statische sites

Statische sites zijn sites die een structuur hebben die gelijk is aan de boomstructuur van mappen op de harde schijf. Elk menuonderdeel kan één of meerdere HTML-documenten en submappen (met HTML-documenten) bevatten.

Deze sites zijn zonder problemen zelf te spideren met de AquaBrowser.

Dynamische sites

Naast statische websites zijn er ook dynamische websites. Deze maken geen gebruik van een mappenstructuur zoals hierboven beschreven. Dynamische websites genereren de pagina's op aanvraag. Hiermee wordt bedoeld dat een pagina niet fysiek in een map staat maar bijvoorbeeld in

⁴ Voor een uitgebreide beschrijving van dit onderwerp zie bijlage 5



een database. In deze gevallen wordt geen gebruik gemaakt van HTML pagina's maar van zogenaamde CGI-pagina's⁵. Dit zijn HTML pagina's waarin ook de programmacode staat die door de webserver (de PC waarop de website is geïnstalleerd) wordt uitgevoerd. Met behulp van de CGI pagina's is het mogelijk om informatie uit o.a. databases te halen.

De informatie die ingevuld moet worden op de pagina wordt vaak versleuteld in het adres. De webserver weet aan de hand van de versleutelde informatie welke pagina's moeten worden teruggestuurd. De AquaBrowser kan hierdoor moeilijk de site doorzoeken. Aangezien dynamische websites niet zijn opgebouwd uit mappen is het niet eenvoudig om alleen een specifiek gedeelte van een dynamische website in de AquaBrowser database op te slaan.

Om dit probleem op te lossen moet de hulp ingeroepen worden van de webmaster van de betreffende site. Deze kan dan een zogeheten "crawler" pagina aanmaken voor het specifieke onderdeel van de website. Deze crawlerpagina bestaat uit een lijst met links naar alle pagina's van de site of één of meerdere onderdelen daarvan. Dit probleem is overigens niet specifiek voor de AquaBrowser. Ook Google en andere zoekmachines hebben dit probleem en moeten dus gebruik maken van een crawlerpagina. Vaak zal een webmaster deze dus al gereed hebben.

Javascript

Een ander probleem is het gebruik van Javascript in HTML pagina's. Javascript is een programmeertaal die op de browser draait en niet op de server (in tegenstelling tot CGI). Met Javascript kunnen bijvoorbeeld links, menu's of zelfs complete pagina's gegenereerd worden. De spider kan geen Javascript uitvoeren. Alles wat door Javascript wordt gegenereerd wordt door de spider dus niet herkend. Ook dit kan in enkele gevallen worden ondervangen met een crawlerpagina. Het kan echter zijn dat deze sites door MediaLab zelf moeten worden ontsloten. In dat geval zal voor maatwerk van MediaLab moeten worden ingezet.

5.3 Eendagscursus 'werken met de AquaBrowser'

Voor nieuwe gebruikers van de AquaBrowser organiseert MediaLab een speciale introductiecursus die op het hoofdkantoor van MediaLab in Schellinkhout wordt gegeven. (Deze is niet opgenomen in de DU licentie.) Het is een cursus voor de technische mensen die met de AquaBrowser gaan werken. Voor meer informatie zie bijlage 6.

⁵ CGI =Common Gateway Interface, een programma op de webserver dat informatie van de gebruiker verwerkt.



6 Samenwerking met MediaLab

6.1 De licentie van DU⁶

Voor de AquaBrowser Toolset is door DU een licentie afgesloten voor alle bij hen aangesloten instellingen.

Deze licentie is inclusief het gebruiksrecht op de technologie (Liquid Tools) gekoppeld aan en ten behoeve van maatwerk door de bij de DU aangesloten onderwijsinstellingen. Gedurende de looptijd van de licentie ter beschikking komende nieuwe (platform)versies van de programmatuur vallen onder de werking van de licentie.

MediaLab kan maatwerk leveren als:

- de AquaBrowser binnen complexe, meervoudige netwerkstructuren geïmplementeerd moet worden;
- bronnen of databases ontsloten moeten worden die niet rechtstreeks toegankelijk zijn met de spider technologie van de AquaBrowser of anders zijn dan de 'standaard' ondersteunde formaten van de AquaBrowser Toolset;
- additionele functionaliteit toegepast moeten worden zoals omschreven in Bijlage 1.

Maatwerk wordt steeds uitgevoerd op basis van een gedetailleerde projectspecificatie.

Er kan een tweedelijns supportovereenkomst met MediaLab worden afgesproken, waarbij ook na oplevering en het verstrijken van de garantietermijn service wordt verleend

Er kan een implementatie supportovereenkomst worden afgesloten met MediaLab, waarin MediaLab ondersteuning biedt bij reguliere implementatie van de AquaBrowser Toolset op het gebied van *configuration, semantic tuning, design, scheduling en technical training*. Deze service wordt geleverd tegen een eenmalig vast bedrag. De doorlooptijd van een dergelijk traject is over het algemeen twee tot drie weken.

Instellingen uit de categorie SLB Diensten BV en APS IT-Diensten BV die o.m. werken t.b.v. ROC's en basisonderwijs hebben een afwijkende eigen regeling met MediaLab.

6.2 Samenwerking met MediaLab

MediaLab heeft ons getroffen als een flexibel en plezierig bedrijf om mee te werken. Terwijl de onderhandelingen over de aankoop van de AquaBrowser in volle gang waren en er dreigde dat ons project pas lang na de zomervakantie van de AquaBrowser zou kunnen profiteren heeft MediaLab op eigen risico een eerste serie websites en bestanden gespiderd. Gevolg hiervan was dat in augustus 2002 al een demonstratie kon worden gegeven van een werkende AquaBrowser voor het communicatieonderwijs. Weliswaar met een beperkte content, maar toch...

Ook tijdens het verdere proces heeft MediaLab over het algemeen snel gereageerd op onze verzoeken om informatie of om een bestand voor ons te spideren. Bij de oplevering van de AquaBrowser in december 2002 ontstond vertraging doordat de AquaBrowser niet volledig functioneerde. Uiteindelijk werd hier echter een oplossing voor gevonden en werd de

⁶ Aan deze informatie kunnen geen rechten worden ontleend.

opleveringsdatum vastgesteld op de datum dat de AquaBrowser daadwerkelijk actief was voor ons project. Latere onduidelijkheden en problemen werden steeds snel opgelost.

Ook toen een extra functionaliteit voor ons ontwikkeld werd was MediaLab uiterst flexibel. Kortom een prettig bedrijf om mee samen te werken.

6.3 Reacties uit de beroepspraktijk

Hoewel de AquaBrowser speciaal voor het onderwijs is aangeschaft blijkt dat hij ook daarbuiten enthousiaste reacties oplevert. Een concreet voorbeeld hiervan is de samenwerking die ontstaan is tussen Com2Know en een vereniging van beroepsbeoefenaren in de communicatie over het spideren van hun contentbank met communicatieplannen. De vereniging stelt zijn materiaal als leer materiaal beschikbaar en de leden van vereniging kunnen mede gebruik maken van de AquaBrowser.



Bijlage 1 Beschrijving van verschillende extra functionaliteiten AquaBrowser

Het gaat hierbij om functionaliteiten die niet in de standaardlicentie van DU zijn opgenomen.



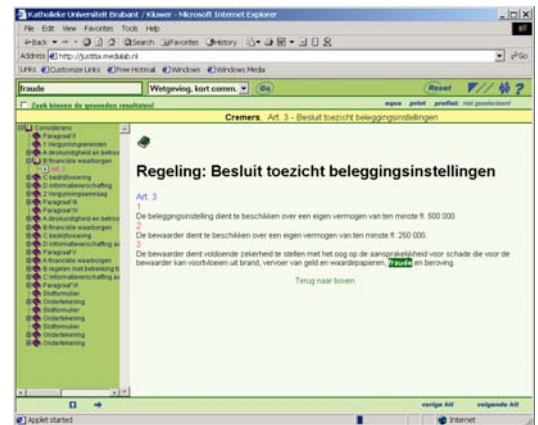
Locatiemodule

De locatiemodule is een optie om een visuele presentatie te koppelen aan items die de AquaBrowser heeft gevonden. Een voorbeeld van deze toepassing is de locatiemodule binnen de bibliotheek systemen. Het systeem wijst automatisch aan waar het gezochte boek staat.

Treeview Table

De treeview table integreert voorgedefinieerde structuren met de AquaBrowser wolk. De integratie van de transparante ontsluiting en bestaande vaste structuren is de meerwaarde. Men kan parallel zowel via de AquaBrowser als de vaste structuur door de informatie surfen. Voorbeeld is de Siso-indeling van de bibliotheken. Dit is de categorisatie die elke bibliotheek gebruikt om hun catalogus en hun bibliotheekruimte in te delen.

Een ander voorbeeld is een toepassing waarbij een gevonden wetsartikel direct 'gelinkt' kan worden naar het onderdeel en plaats in het gehele wetboek.



Translation Language

De AquaBrowser Pro heeft standaard twee talen in haar Core engine: Engels en Duits. Het is mogelijk om dit uit te breiden tot 20 talen.

Search accent

Het is mogelijk om de AquaBrowser te laten zoeken volgens een bepaalde invalshoek. Bij een zoektocht naar een bepaald onderwerp zal een directeur in andere informatie geïnteresseerd zijn dan een jurist van dat bedrijf. De AquaBrowser kijkt als het ware door de bril van de vragsteller.

Subject recognition

De AquaBrowser is in staat om tijdens een query op een intelligente wijze herkennen of de vraag gaat om een persoon, onderwerp etc.. Een voorbeeld is de schrijvers/onderwerp herkenning bij een landelijke bibliotheeksite.

Boolean Search

MediaLab kan parallel aan de AquaBrowser de traditionele boolean search functionaliteit aanbieden.

User statistics

De user statistics module monitort alle relevante processen en kan alle door de klant gedefinieerde informatie meten en rapportages genereren. Additional filters

MediaLab is in staat om voor elke bron een filter/ontsluiting te genereren. In principe is elke bron op elk willekeurig platform te ontsluiten via de AquaBrowser technologie.

Sort / Ranking mechanisme

Bij het gebruik van meerdere bronnen is het mogelijk om de resultaten van een query opnieuw te rangschikken naar het brontype. Voorbeeld is de mogelijkheid om de resultaten te sorteren naar bijvoorbeeld favoriete krant.

Aqua site Bobby proof (t.b.v. visueel gehandicapten)

Parallel aan de AquaBrowser site wordt een geheel tekst georiënteerde gegenereerd die volledig aan de Bobby norm voor visueel gehandicapten voldoet.

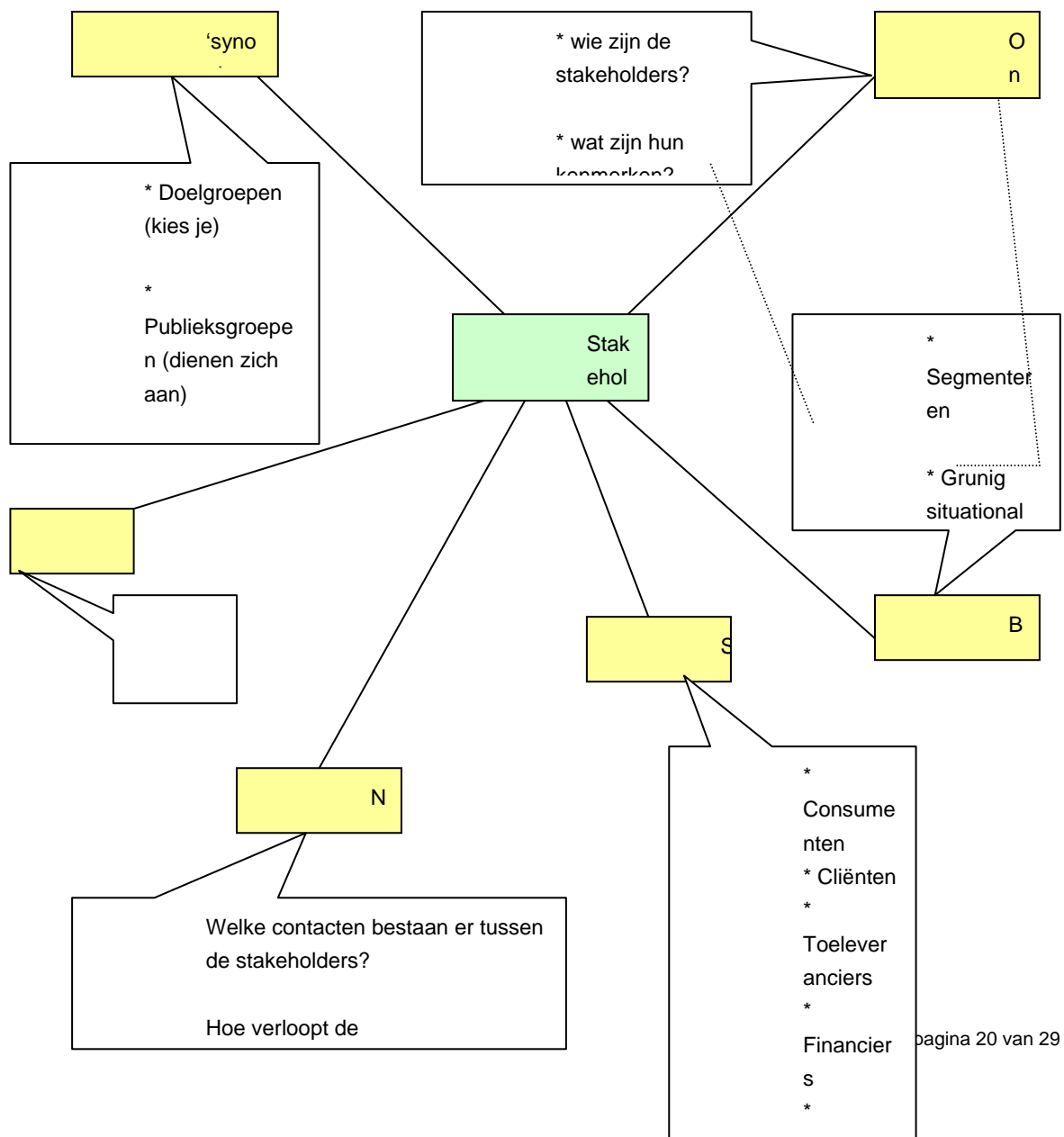
Spideren van PowerPoint en Excel

Deze functionaliteit maakt dat de AquaBrowser ook PowerPoint presentaties en Excel spreadsheets kan doorzoeken

Zoeken in mandjes

Special voor Com2Know is een sort/ranking functionaliteit ontwikkeld waardoor queries gerangschikt kunnen worden naar de vindplaats van de informatie. Bijvoorbeeld alleen scripties.

Bijlage 2 Gedachtenwolk communicatieproject





Bijlage 3 Hoe bouwt de AquaBrowser zijn thesaurus op?

De AquaBrowser is de laatste stap in een reeks van technische bewerkingen van beschikbare informatie. Die technische bewerkingen worden uitgevoerd door een serie instrumenten die de verzamelnaam Liquid Tools dragen. Verschillende onderdelen van deze Liquid Tools vergaren, verwerken, analyseren en verbinden de documenten uit de uiteenlopende bronnen.

Informatie (kennis) ligt opgeslagen in verschillende formaten en op verschillende media; van databases en websites tot tapes en cd-rom's. Voor de Liquid Tools maakt het in principe niet uit hoe de informatiebron eruit ziet, als hij maar digitaal is en bereikbaar.

Liquid Filters halen de informatie uit de verschillende bronnen en geven die door aan de Liquid Knowledge Builder. Het bronmateriaal zelf wordt gecached in de IGOR database. Het grote voordeel hiervan is dat de IGOR-database vele malen sneller is dan conventionele databases en dat de bron verder ongemoeid blijft: deze mag ook op een ander platform staan.

Op het tweede niveau analyseert de **Liquid Knowledge Builder** alle door de filters opgehaalde informatie. Dat gebeurt aan de hand van een ingebouwd woordenboek (bijvoorbeeld de integrale Van Dale), al dan niet uitgebreid met een specifieke thesaurus (met vaktermen). De analyse gaat verder met het maken van mathematische modellen, zoals op basis van woordfrequentie en clustering.

IGOR is de Black Box van het hele systeem. Hier ligt de informatiewolk opgeslagen en worden alle dwarsverbanden gelegd. Het is, simpel gezegd, een razendsnel kennis-opslag en -extractiesysteem. IGOR is ontworpen op basis van COM en is te benaderen vanuit iedere programmeeromgeving. Binnen IGOR zijn op veel punten objecten gedefinieerd die specifieke taken verrichten.

De data-engine is zeer snel. Dat kan omdat hij de data alleen hoeft te cachen. Het formaat van de opgeslagen inhoud is volledig vrij: er zijn meerdere tabellen, variabele records binnen tabellen etc. Deze engine bouwt een index op van alle woorden uit de tekst, zodat ieder woord individueel opzoekbaar is. Standaard levert de engine voor een query een op relevantie gesorteerde lijst van documenten op. Daarbij is de exacte queryterm het belangrijkste. Gestemde varianten komen later. De query 'roos' levert dus eerst 'roos' op, en dan 'rozen', 'klaproos' en 'rozenboom'.

Rankers geven een vaag antwoord op een exacte vraag. Een zoekvraag naar programmeurs uit Amsterdam levert in een conventioneel systeem alleen de programmeurs uit Amsterdam op. Een ranker geeft alle woonplaatsen in een kring om Amsterdam een gewicht (rank). IGOR zal daarom ook programmeurs uit Hoofddorp en Haarlem opleveren, maar bovenaan lijst staan de programmeurs uit Amsterdam (als die er tenminste waren!)

De co-occurrence generator analyseert de tekst uit de database, en legt statistisch en semantisch relaties tussen termen uit de tekst. Deze relaties worden opgeslagen in een semantisch netwerk en komen tevoorschijn in de AquaBrowser. Dit proces is volledig automatisch en vereist geen onderhoud. Wel kan men de eigen gegevens verrijken met associaties.

De vertaaleenheid wordt gebruikt om termen automatisch te vertalen naar andere talen, jargon, Oudnederlands, of wat dan ook. Zo kan het gebeuren dat een query op CO₂ beantwoord wordt met een document waarin het woord kooldioxide voorkomt.

De fuzzy alternatieven generator spoort varianten op een woord op. En zo verschijnen zelfs spelfouten in oorspronkelijke teksten in beeld. Een conventioneel systeem zou die nooit vinden. Deze generator werkt net zo snel als de standaard zoekactie.



Stemmers zijn instrumenten die woorden ontleden tot hun stam, om daarmee een zoektocht naar synoniemen mogelijk te maken. Een stemmer reduceert de woorden 'roosje' en 'rozen' tot de stam 'roos', om van daaruit woorden als rozenboom en stokroosje (en zelfs Doornroosje!) te vinden.

De Liquid Context Builder is de bovenste laag van het systeem. Deze verrijkt de vraag van de gebruiker en onthoudt het afgelegde zoekpad. Dit systeem genereert de lijst met gevonden documenten, maar vooral ook de associatieve begrippenwolk die de gebruiker aanwijst langs welke lijnen meer relevante informatie te vinden is. De gevonden documenten en de begrippenwolk worden zichtbaar gemaakt in de AquaBrowser.

De AquaBrowser zelf kan als Java applet opgenomen worden in de webbrowser, maar hij kan ook andere vormen aannemen.



Bijlage 4: Verschillende associatieve zoekmachines onderzocht

In 2001 begon de zoektocht van Com2Know naar een associatieve zoekmachine die ook in het beheer gebruiksvriendelijk zou zijn en op zo veel mogelijk platforms moest kunnen draaien.

Om een zo verantwoord mogelijke keuze te kunnen maken zijn vier associatieve zoekmachines onderzocht, te weten:

- SharePoint Portal van Microsoft.
- Autonomy van Autonomy.
- Muscat van Smartlogik
- AquaBrowser van MediaLab

Hieronder volgt een korte beschrijving van deze zoekmachines zoals ze toen waren.

Sharepoint Portal

Microsoft® SharePoint™ Portal Server 2001 vergroot de mogelijkheden van Microsoft Windows® and Microsoft Office door het aanbieden van een krachtige nieuwe methodiek om informatie te vinden, te organiseren en te delen.

Kenmerken zijn:

- Een portal waarbij documenten gemakkelijk kunnen worden geopend, bewerkt en opgeslagen.
- Een documentbeheer systeem met check-in en check-out schermen, waarbij de gebruiker d.m.v. een aantal kenmerken een profielschets van het document kan geven. Overigens moeten documenten hier ruim zien worden: Word, Excel, PowerPoint etc.
- Een Digital Dashboard waarbij een agenda, aantekeningen, afspraken, evenementen etc. in één oogopslag kan worden gerealiseerd. Uiteraard aan te passen aan de wensen van iedere individuele gebruiker.
- Een opslag van gegevens waarbij toegang kan worden gegeven of geweigerd, voor de hele hogeschool, voor bepaalde afdelingen, voor externe gebruikers (internet) etc..
- Discussion Board voor het opzetten van discussies m.b.t. de content.

Autonomy

De kern van Autonomy is een zoekstelsel (search engine), gebaseerd op een unieke patronen herkenningmethode, die het mogelijk maakt om grote hoeveelheden ongestructureerde informatie op Internet en Intranet te kunnen bevragen. Het verschil met de bekende zoeksystemen op Internet zoals AltaVista e.d. is dat je bij Autonomy zelf bepaalt wat de bronnen zijn waarop gezocht wordt, terwijl je daar bij AltaVista geen inspraak in hebt. Hiermee heb je zelf greep op de kwaliteit van de uitkomsten van de zoekopdrachten. Verder is het zo dat het met behulp van Autonomy mogelijk is om zoekopdrachten te geven in natuurlijke taal in plaats van de veel lastiger te formuleren queries met Booleaanse operatoren (AND, OR enz.). Het zoekstelsel van Autonomy extraheert een concept (meest significante onderdelen) uit de zoekopdracht en gebruikt dit om te zoeken in de bronnen.

Muscat

Deze zoekmachine is ontwikkeld door Smartlogik. De modules bieden een zoekmachine, spider en terugvindprocedure, een thesaurus en een regelgenerator. Deze zoekmachine wordt momenteel b.v. gebruikt door De Universiteitsbibliotheek van Utrecht en door de BBC.



Aqua Browser

De Aqua Browser wordt door MediaLab in Schellinkhout ontwikkeld naar de wensen van de opdrachtgever. (Huidige gebruikers zijn o.m.: Gemeente Utrecht, samenwerkingsverband van overheden over de Waddenzee, Vereniging Digitaal Erfgoed Nederland)

Met de Liquid Tools van MediaLab kunnen documenten uit verschillende bronnen (websites, databases, cd-rom's, e.d.) worden vergaard, verwerkt, geanalyseerd en verbonden. De AquaBrowser legt verbanden tussen betekenissen, contexten en directe relaties. De gevonden informatie wordt op een speelse manier gepresenteerd. Het zoeken naar data wordt met Liquid Tools gekoppeld aan:

- semantische betekenissen
- fuzziness (de gegevens niet volgens nauwkeurig omschreven kenmerken ingedeeld)
- permissiveness (er is geen vastomlijnde zoekstructuur, vrije zoekroutes zijn mogelijk).
- context (de context waarin een begrip gebruikt kan worden)

Door het aanleveren van een 'woordenlijst' kan dit systeem ook wat meer gestuurd worden.

Keuze voor AquaBrowser

Uiteindelijk viel de keuze op de AquaBrowser. De redenen hiervoor waren dat deze aanzienlijk gebruiksvriendelijker was dan de andere onderzochte producten. De overige producten waren in het beheer zeer veel bewerkelijker dan de AquaBrowser. Daarnaast bestond in die periode de indruk dat de AquaBrowser platformonafhankelijk was. Hij was tenslotte ook in Java geproduceerd. Hoewel later bleek dat de AquaBrowser (helaas) toch afhankelijk is van een Windows omgeving veranderde dat niet de uiteindelijke keus. De AquaBrowser bleef het meest gebruiksvriendelijk en bood een aantal interessante mogelijkheden.



Bijlage 5: Nadere Informatie over het spideren van sites⁷

Inleiding

Dit document beschrijft op een niet-technische wijze de manier waarop de spider informatie van het Internet haalt en de problemen die men daarbij kan tegenkomen.

De spider is een proces op de computer waarbij (web)documenten kunnen worden ingelezen. In dit document kijken we hoe de spider webdocumenten via het intranet of internet ophaalt.

Voor het begrijpen van de spider is het prettig om de websites waarin zich de webdocumenten bevinden onder te verdelen in de volgende twee categorieën:

- ◆ Statische websites
- ◆ Dynamische websites

Deze categorieën worden in paragraaf 1 besproken. Na de uitleg van de statische en dynamische websites volgt in paragraaf 2 de beschrijving van de werking van de spider.

In paragraaf 3 wordt beschreven hoe informatie uit databases, zoals Oracle en SQL, met behulp van import technologie in de AquaBrowser database kan worden ingelezen. Deze technologie komt aan de orde als het niet mogelijk is gebleken om een website met de “gewone” spider technologie in de AquaBrowser database in te lezen.

Deze laatst genoemde technologie staat los van de spider technologie maar geeft samen een compleet beeld van hoe MediaLab Solutions BV gegevens kan inlezen uit diverse type bronnen (websites, databases, netwerkschijven).

1. Statische en dynamische sites

Statische Websites

Een statische website heeft een structuur die vergelijkbaar is met de boomstructuur van mappen op een harde schijf. De webdocumenten staan net als gewone documenten in mappen en submappen. Dit kunnen we zien aan het adres invoerveld van de browser. Een browser is een applicatie waarmee websites kunnen worden bekeken. Voorbeelden van browsers zijn “Internet Explorer”, “Netscape”, “Opera”, etc. Op deze site is een map “AquaBrowser” aangemaakt met daarin de pagina “index.html” (AquaBrowser/index.html).

Elke menuonderdeel kan één of meerdere HTML documenten en/of submappen bevatten. Op deze manier worden alle pagina's verdeeld onder de mappen. Elke pagina wordt dus fysiek apart opgeslagen in de mappenstructuur van de site.

Dynamische WebSites

Naast statische websites zijn er ook dynamische websites. Deze maken geen gebruik van een mappenstructuur zoals hierboven beschreven. Dynamische websites genereren de pagina's op aanvraag. Hiermee wordt bedoeld dat een pagina niet fysiek in een map staat maar bijvoorbeeld in een database. In deze gevallen wordt geen gebruik gemaakt van HTML pagina's maar van zogenaamde CGI-pagina's. Dit zijn HTML pagina's waarin ook programmacode staat die door de webserver (de PC waarop de website is geïnstalleerd) wordt uitgevoerd. Met behulp van de CGI pagina's is het mogelijk om informatie uit o.a. databases te halen. Er zijn legio websites waarbij van

⁷ Deze tekst is afkomstig van MediaLab



deze techniek gebruik wordt gemaakt, zie bijvoorbeeld www.google.com www.minbzk.nl en support.AquaBrowser.com

Alle informatie op deze websites wordt gegenereerd door een klein aantal CGI pagina's. Bij Google worden de resultaten gegenereerd door <http://www.google.nl/search> en bij het Ministerie van Binnenlandse Zaken is alles afkomstig van <http://www.minbzk.nl/asp/get.asp>. Je kunt dit zien in het adres invoerveld van de browser..

Een spider begint vaak met de zogenaamde "homepage" van een website als invoer. We nemen als voorbeeld de homepage van de MediaLab website "<http://www.medialab.nl/>". Deze homepage pagina wordt door de spider als eerste opgeslagen. Als deze pagina is opgeslagen wordt de pagina doorzocht op links, d.w.z. verwijzingen, naar andere pagina's op de website. Als dat het geval is worden deze pagina's opgehaald. Voor ieder van deze pagina's wordt dan weer gekeken of hier ook links op voorkomen, etc, etc.

Standaard volgt de spider alleen de links binnen de aangegeven website. Als er dus links naar andere websites voorkomen dan worden deze niet gevolgd.

2. Problemen bij het spideren

De meeste problemen komen voor bij dynamische websites. We nemen als voorbeeld de website <http://archieff.volkskrant.nl/>. In deze website zijn alle uitgaven van de volkskrant van de afgelopen twee jaar opgeslagen. Als je naar de betreffende pagina kijkt, zie je dat er een aantal selectiecriteria zijn waarbinnen men kan zoeken (categorie en periode).

Vervolgens kan een zoekterm worden opgeven en op de "zoek" knop worden gedrukt. Dit zijn allemaal handelingen die de AquaBrowser spider niet kan uitvoeren. Dit houdt in dat de spider niet verder komt dan deze eerste pagina

Versleutelde adressen

De informatie die ingevuld moet worden op de pagina wordt vaak versleuteld in het adres. Als we bijvoorbeeld op de Volkskrant website zoeken met de term "informatie" dan wordt dit als volgt in het adres versleuteld:

http://archieff.volkskrant.nl/zoek?SORT=presence&FDOC=0&SEC=* &PRD=2y&text=informatie

Dit adres wordt vervolgens opgestuurd naar de webserver. De webserver weet aan de hand van de versleutelde informatie welke pagina's moeten worden teruggestuurd.

Als je precies weet wat er naar de server moet worden opgestuurd kun je dit gebruiken om met de spider meer pagina's uit een dergelijke website te halen. In het adres van de Volkskrant staat waarop gezocht moet gaan worden, namelijk "informatie". In de meeste gevallen wil je met de spider op alles zoeken wat er zich in de database bevindt. Je zou dan als zoekvraag een "*" op kunnen sturen. Bij sommige webserver werkt dit, maar er zijn er ook veel die deze constructie afvangen en geen resultaat teruggeven

Crawler pagina's en sitemaps

Met de AquaBrowser Pro spider is het mogelijk om alleen bepaalde mappen van een website te spideren. Aangezien dynamische websites niet zijn opgebouwd uit mappen is het niet eenvoudig om alleen een specifiek gedeelte van een dynamische website in de AquaBrowser database op te slaan.

Om dit probleem op te lossen zal de hulp ingeroepen moeten worden van de webmaster van de betreffende site. Deze kan dan een zogeheten "crawler" pagina aanmaken voor het specifieke onderdeel van de website. Deze "crawler" pagina bestaat uit een lijst met links naar alle pagina's van de site of één of meerdere onderdelen daarvan.

Een dergelijke lijst met links is vergelijkbaar met een "sitemap". Een sitemap is vrij vertaald een "plattegrond" van de website. Een sitemap geeft de indeling en de hiërarchie weer van alle pagina's waaruit website bestaat. Met een sitemap kan je snel naar een bepaald punt in de website navigeren. Door een sitemap als startpunt voor je spider te gebruiken kan op deze manier de gehele site worden gespidered. Veel websites beschikken over een sitemap

Javascript

Een ander probleem is het gebruik van Javascript in HTML pagina's. Javascript is code die door de browser wordt uitgevoerd. Met Javascript kunnen bijvoorbeeld links, menu's of zelfs complete pagina's gegenereerd worden. De spider kan geen Javascript uitvoeren. Alles wat door Javascript wordt gegenereerd wordt door de spider dus niet herkend. Ook hier is in enkele gevallen weer de oplossing het gebruik van een crawler pagina of een sitemap.

3. Import Technologie

MediaLab heeft een import technologie ontwikkeld voor een aantal databases. De informatie wordt dan niet meer via de spider opgehaald, maar rechtstreeks uit de database (van de website) zelf. Een nadeel van deze methode is dat vaak de opmaak informatie verloren gaat. Deze opmaak informatie wordt meestal door de CGI pagina's gegenereerd.

MediaLab Solutions BV kan de volgende type databases importeren:

- ◆ databases met ODBC interfaces (Access, Dbase, Oracle, SQL, etc.)
- ◆ databases in XML format.

Opmerking:

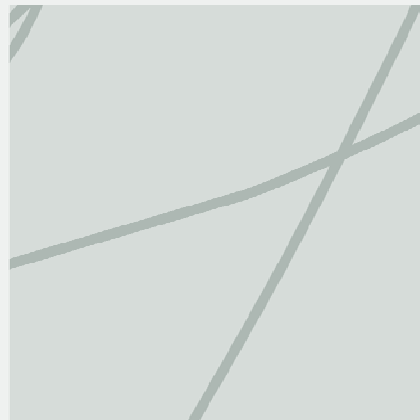
Het importeren van deze databases is maatwerk en kan alleen gedaan worden door professionals van MediaLab Solutions BV.



Bijlage 6: Korte inhoud introductiecursus AquaBrowser bij MediaLab

Tijdens de cursus wordt o.m. ingegaan op:

- * AquaBrowser site
 - Conceptueel
 - Functionality: wat kan de eindgebruiker met de AB?
 - Demo data: www.ode.nl
 - Panes: namen en elementen?
 - Directory structuur
- Hard- and Software specificaties voor de webserver + Installatie
- AquaBrowser Configuration site Part 1 (Content)
 - User Interface
 - Site: Create/Save As/Delete/Start/Stop AquaBrowser website
 - Content: Configure/Schedule spider + Empty database
- Uitstapje naar de settings.xml met spider options
- AquaBrowser Configuration site Part 2 (Design)
 - Default versus User-defined styles
 - Layout Standard Style
 - Maintenance (backup/restore)
- * User-defined styles
 - Directory structuur
 - Componenten
 - Uitgebreid voorbeeld
- * System Files
 - Thesaurus
 - XML settings
 - Stopwords
 - ContentTypesFile



Tekstvak voor beschrijving van de publicatie