Vubis Smart en het multidimensionale Caché-dbms

Fuzzy logic en dataverrijking in de bibliotheek

Eric Conderaerts

Bij het behandelen van content management en datamining wordt vooral uitgegaan van informatie die ongestructureerd is. Daarbij wordt vergeten dat er nog oneindig veel gegevens beschikbaar zijn die wel gestructureerd zijn. Een typisch voorbeeld van dergelijke systemen zijn de geautomatiseerde bibliotheekcatalogi die al sinds de jaren zestig en zeventig van de vorige eeuw met grote zorg zijn opgebouwd. Decennia lang bleven die databases geautomatiseerde kaartenbakken, maar vandaag de dag zijn er talloze middelen om de echte waarde van die gestructureerde data tot zijn recht te laten komen en te verrijken.

Het Canadese softwarehuis Geac ontwikkelde in samenwerking met de Technische Universiteit Eindhoven en de Vrije Universiteit Brussel een nieuw informatiesysteem: Vubis Smart, waarin die nieuwe mogelijkheden volop worden benut. Vubis Smart combineert een aantrekkelijk technisch concept (zie kader) met een aantal technieken om data te verrijken. De belangrijkste principes hierbij zijn fuzzy logic en dataverrijking, die gebaseerd zijn op zowel IT-standaarden (als OpenURL) als op statische analyse van gebruikersactiviteiten en van andere data in de database.

GESTRUCTUREERDE INFORMATIE EN METADATA

Eén van de problemen bij het succesvol dataminen van databases is dat de infor-

Informatiesysteem voor musea, bibliotheken en archieven

Vubis Smart is een client/server-systeem gebaseerd op het 'thin client' principe (niet te verwarren met 'thin client solutions' als Windows Based Terminal of Cytrix MetaFrame). Essentieel hierbij is dat alle applicatielogica zich op de server bevindt en dat de client zich alleen bezighoudt met de presentatie en met de koppeling met andere Windows-applicaties. Ook de interfaceobjecten (de definities van forms, lijsten, buttons, en dergelijke) bevinden zich allemaal op de server, wat zelfs de meest ingrijpende wijzigingen van het systeem toelaat zonder dat de client-software moet worden aangepast. Zijn er toch updates van de client nodig, dan kan dat via de zogenaamde 'auto update'-functie, waardoor een client zelf detecteert dat een nieuwe versie beschikbaar is en deze automatisch installeert.

De server kan opereren op Windows-, Linux- en Unix-platforms, de client draait onder Windows. Bij het ontwikkelen wordt vooral gebruik gemaakt van Caché Object Script, HTML, JavaScript, C / C++, Visual Basic en XML. Als database-systeem wordt Caché gebruikt (zie kader 'Het multidimensionale datamodel').
De gebruiker wordt geconfronteerd met maar één, relatief eenvoudige interface. Voor de bibliotheekbezoeker (via het internet of in het fysieke bibliotheekgebouw) is dat een webinterface, voor de bibliotheekmedewerkers een Windows-interface, die toegang biedt tot alle modules (inclusief systeembeheer en een rapportgenerator). Doordat een gebruiker slechts met één interface wordt geconfronteerd (via die interface kan hij ook ingewikkelde query's op de database loslaten zonder daarvoor een query-taal te moeten leren) neemt de totale complexiteit van het systeem sterk af, en bijgevolg ook de 'total cost of ownership'. Functioneel is het systeem naast een zoek- en informatiesysteem ook een op de bibliotheeksector toegesneden ERP-systeem, dat modules biedt voor catalogusbeheer, besteladministratie, tijdschriftenbeheer en uitleenadministratie.

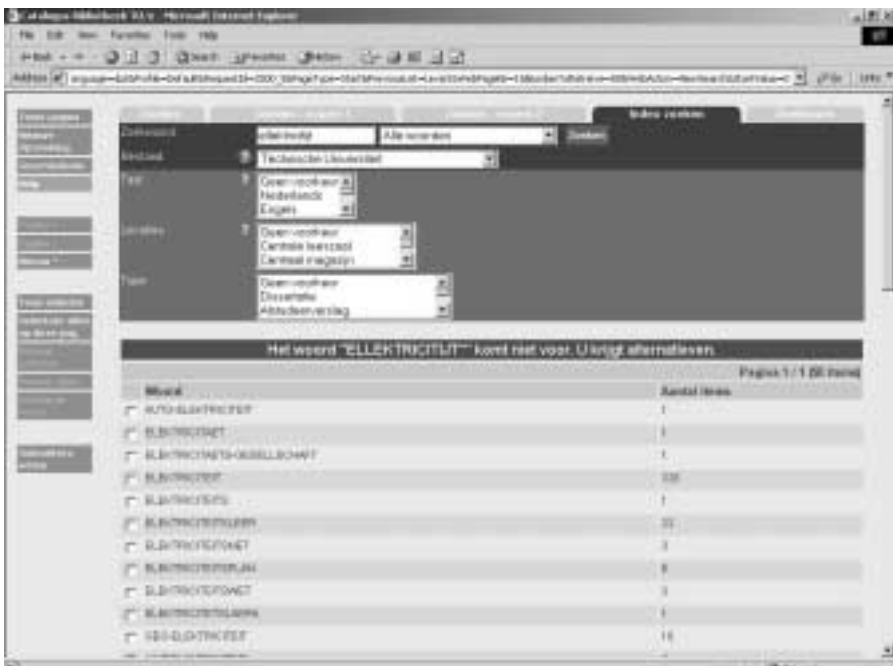
Vubis Smart is geïnstalleerd in bibliotheken, documentatiecentra, archieven en musea. De eerste bèta-site werd geïnstalleerd in juli 2001. Sinds april 2002 is het systeem officieel beschikbaar. Vubis Smart is eigendom van de Vrije Universiteit Brussel en de Technische Universiteit Eindhoven; de commercialisering gebeurt door Geac, een Canadees softwarehuis met vestigingen in onder andere Nederland, België, Frankrijk, het Verenigd Koninkrijk, Canada, de Verenigde Staten en Australië. Gebruikers van Vubis Smart zijn de Bibliotheek Breda, de Technische Universiteit Eindhoven en het Koninklijk Museum voor Schone Kunsten in Brussel.

matie in die databases te 'arm' is. "Het 'dataminen' van databases op zoek naar verborgen patronen is tot nu toe nooit serieus doorgebroken en zal ook nooit een serieuze trend worden. Administratieve databases worden zo ontworpen en gebruikt, dat er doorgaans niet veel onverwachte en onbedoelde maar niettemin nuttige patronen te herkennen zijn. Wie toch dergelijke patronen wil zoeken moet veel, heel veel kennis aan een systeem toevoegen - zoveel kennis, dat het sop de kool vrijwel nooit waard zal zijn." (*Kunstmatige intelligentie en expertsystemen,* René Veldwijk. - Database Magazine, 1999/06, pag. 33).

Nu is één van de typische kenmerken van bibliotheekcatalogi dat de toevoeging van kennis aan data van oudsher al standaard praktijk was. Beschrijvingen van documenten (boeken, muziek, kaarten, foto's, museumobjecten, websites,...) bevatten niet alleen een volledige bibliografische identificatie (titel, auteur, uitgever, editie, enzovoort), maar ook uitgebreide inhoudelijke informatie (trefwoorden, samenvattingen, coderingen uit standaard-classificatieschema's, enzovoort). Bovendien heeft de sector uitgebreid geïnvesteerd in de creatie en het onderhoud van thesauri met goedgekeurde 'authorities' voor allerlei vakgebieden. Hierdoor is de informatie in de database al uitgebreid met veel 'kennis', en is het ontdekken van die 'onverwachte en onbedoelde, maar toch interessante' patronen wel degelijk goed mogelijk.

Voor de opslag van gegevens stelt Vubis Smart de gebruikers in staat metadata te definiëren; hiermee kan de bibliotheek informatie over de eigenlijke informatie in de database vastleggen. Omdat de bibliotheek volledige controle heeft over zijn bestandsformaten en daardoor ook op het gedrag van het systeem (hoe verloopt de invoer?, hoe verloopt de controle op de data?, hoe verloopt het indexeren?, hoe verloopt het opzoeken?) kan de meest diverse - gestructureerde - informatie worden opgeslagen: zowel beschrijvingen van boeken, als van websites, als van museumobjecten, als van eigenlijk om het even wat. Het gebruik van de 'multidimensionale' capaciteiten van het onderliggende Caché-DBMS is hierbij essentieel (zie het kader over de Caché-database).

Door die flexibiliteit is Vubis Smart dan ook niet gebonden aan één van de vele standaarden die in de bibliotheeksector in gebruik zijn voor dataopslag en -uitwisseling (de zogenaamde MARC-standaarden en de daaraan verwante ISO2709-standaard, Dublin Core, e.a.). Integendeel: het systeem stelt de bibliotheek in staat zijn eigen 'flexibele' formaten te definiëren en deze indien nodig door elkaar te gebruiken. Deze metadata is de basis die structuur aanbrengt in de database en die ook de basis vormt voor dataverrijking.



Figuur 1: Fuzzy logic genereert dynamisch zoektermen die lijken op de ingetikte zoekterm.

DATAVERRIJKING

Dataverrijking is het principe om informatie die zich in de database bevindt (dynamisch) te 'verrijken' met andere informatie in die database of met informatie buiten die database.

Globaal gezien wordt data door of binnen Vubis Smart verrijkt op basis van een aantal verwante mechanismen:

1. Door de data-invoerder 'hard' gecodeerde (statische) links tussen documenten (dit kunnen zowel links tussen twee of meer documenten in de database zijn, als links tussen een document in de database en een URL daarbuiten).
2. Dynamisch gegenereerde links tussen documenten op basis van statistische analyse van informatie in de documentbeschrijving (het systeem gaat op basis van de inhoud van bepaalde velden op zoek naar records die identieke of gelijkaardige informatie bevat in die velden; de velden krijgen een weging mee om hun belang te onderstrepen).
3. 'fuzzy logic' relaties op basis van indexen (het systeem reageert actief op zoekacties die geen resultaat opleveren en genereert dynamisch zoektermen die het meest lijken op de ingetikte zoekterm; zie het schermvoorbeeld met de zoekactie naar 'ellektricitijt').
4. Associaties tussen begrippen op basis van metadata in de database (het systeem bouwt automatisch relaties op tussen begrippen in de database; het systeem analyseert de inhoud van een document en groepeert begrippen en documenten op basis van deze inhoud; hierdoor is het systeem in staat verschillende betekenissen van een woord te herkennen en op basis daarvan documenten te groeperen; het begrip 'Java' wordt op die manier gerelateerd

aan begrippen uit diverse vakgebieden, bijvoorbeeld 'internet' en 'programmeertalen', maar ook aan 'aardrijkskunde', (de Indonesische eilanden) 'Lombok' en 'Bali', de 'Indonesische keuken', enzovoort).

5. Relaties op basis van de statistische analyse van transacties (het systeem analyseert uitleentransacties en creëert relaties op basis van het principe 'wie dit object uitleent, leent ook de volgende objecten uit', vergelijkbaar met het principe dat wordt toegepast bij Amazon.com, 'wie dit boek bij ons koopt, koopt ook de volgende boeken').
6. Relaties op basis van de OpenURL-standaard (zie de volgende paragraaf).

VLINK EN OPENURL

Een onderdeel van Vubis Smart is V*link*, een zogenaamde 'OpenURL resolver' of 'link generator'. OpenURL is een protocol dat is gecreëerd om open en contextsensitieve links in webpagina's te genereren. Het protocol is een open standaard die op weg is naar certificatie door de NISO en is gebaseerd op standaardprotocollen als HTTP en de URL-syntax. Een voorbeeld van een OpenURL-query is bijvoorbeeld *http://yourlibrary/vlink/vlink.csp?genre=book&isbn=123456789*. De doelstelling van de link generator is het automatisch genereren van zinvolle links bij een document in een database, bijvoorbeeld links naar Amazon.com of BarnesAndNoble.com voor boeken, naar RollingStone.com of Qmusic.com voor cd's, naar EbscoHost.com of WebOfScience.com voor tijdschriftartikelen, naar Google.com voor alle deze types materialen, enzovoort. Hoe werkt dit? Een internetgebruiker zoekt op een website die OpenURL-compatibel is en vindt daar de beschrijving van een document (document X). In de webpagina die de informatie over dat document toont, is een link opgenomen die naar V*link* verwijst. Als de gebruiker op de link klikt, wordt V*link* geactiveerd en opent zich een nieuw browservenster. Op dit scherm biedt het systeem relevante links aan naar allerlei services (bijvoorbeeld links naar full-text databases, elektronische tijdschriften, 'citation databases', internetzoekmachines, catalogi met bezitsgegevens, online winkels). De links die worden getoond zijn contextsensitief en worden dynamisch gecreëerd op basis van de informatie in de beschrijving van het document (X).

Het is van belang te beseffen dat de hyperlink die naar V*link* verwijst niet alleen kan worden opgenomen in een Vubis Smart database, maar ook in elke andere database/website die compatibel is met OpenURL. OpenURL ondersteunt twee mechanismen voor het transporteren van metadata over een document. De ene wordt 'by value' genoemd, wat betekent dat de URL alle benodigde informatie bevat om V*link* in staat te stellen de links te genereren. De andere is 'by reference' genaamd, wat inhoudt dat de URL alleen

Het multidimensionale datamodel

Vubis Smart gebruikt Caché als database. Caché is een 'post relational dbms' en één van de belangrijkste 'embedded databases' (databases die worden verkocht 'als onderdeel van' een applicatie). Naast vanzelfsprekende vereisten als betrouwbaarheid, schaalbaarheid en ondersteuning van relevante techniek (Unicode, XML, ODBC, SQL, SOAP) is voor Vubis vooral van belang het feit dat Caché een 'low maintenance' database is. Binnen de bibliotheeksector zijn doorgaans zowel de IT-budgetten als de IT-expertise van het personeel relatief laag, en dan is 'low maintenance' een extra belangrijk gegeven. Caché is een objectdatabase die kan worden geïnstalleerd op alle relevante platforms (Windows en de belangrijkste Linux- en Unix-varianten). Caché biedt een attractief datamodel, waarbij er op drie manieren toegang tot data kan worden verkregen: 'direct', via SQL en via de zogenaamde objectmethode. Dit wordt gerealiseerd door de implementatie van een 'multidimensionaal' datamodel, waarbij er geen vertaalslag nodig is van de 'rijke' objecten naar 'arme' (tweedimensionale) tabellen. Caché wordt geleverd door Intersystems uit Cambridge, MA. Voor meer informatie zie: *http://www.e-dbms.com of http://www.intersys.com.*
(Zie voor meer informatie over Caché ook de twee artikelen die Paul van der Linden heeft geschreven, in DB/M 6 van 1999 (Update, pag. 7) en nr. 3 van 2001 (thema-artikel); red.)

FIGUUR 2: V*LINK* GENEREERT AUTOMATISCH ZINVOLLE LINKS.

een pointer bevat die V*link* in staat stelt de metadata op te vragen via een zogenoemde 'fetch'. Deze is gebaseerd op of maakt gebruik van open protocollen als Z39.50, DOI of OAI, of op basis van een proprietary protocol. Op die manier kan V*link* worden geïntegreerd in elke database die compatibel is met OpenURL en die één van beide transportmechanismen (by value of by reference) ondersteunt. Dit is bijvoorbeeld het geval voor de meeste leveranciers van full-text content (onder andere EBSCOhost, ERL5 (Silver Platter), Web of Science, SwetsnetNavigator, PubMed, ProQuest en andere). Om de V*link* hyperlink in 'andermans' website te integreren wordt het cookiepusher-mechanisme gebruikt dat onderdeel uitmaakt van de OpenURL-standaard; alternatieven hiervoor zijn mechanismen als herkenning op basis van ip-ranges.

DE KWALITEIT VAN DE DATAVERRIJKING

De kwaliteit van dit soort dataverrijking staat of valt met de kwaliteit van de metadata waarop zij is gebaseerd: alleen als de data correct en rijk is, zullen de links correct en rijk zijn. Om de kwaliteit van de links te verhogen, waardoor wordt vermeden dat links worden aangemaakt die toch niet tot een zinvol zoekresultaat zullen leiden, past V*link* regels toe op het moment dat de links worden gegenereerd. Dergelijke regels bepalen bijvoorbeeld dat niet naar PubMed wordt gelinkt voor niet-medische publicaties, en dat niet naar Amazon.com wordt gelinkt voor boeken die gepubliceerd zijn voor 19xx, enzovoort. De sterke punten van dit soort dataverrijking zijn divers. Het systeem biedt toegang tot gerelateerde informatie via één consistente en aanpasbare interface (op basis van CSS), waarbij een onbeperkt aantal profielen kan worden gedefinieerd om de bibliotheek in staat te stellen het aanbod zo precies mogelijk aan te passen aan het type gebruik(er). Alle links worden dynamisch en contextsensitief gegenereerd (dat wil zeggen, ze verschillen voor artikelen, boeken, tijdschriften, films, muziek). De bibliotheek bepaalt zelf naar welke sites gelinkt wordt en welke regels moeten worden toegepast. Ten slotte omvat V*link* ook een oplossing voor het probleem van DRM (digital rights management). Met het bieden van toegang tot informatie op het web steekt ook het probleem van toegangsrechten de kop op; wie heeft toegang tot wat? Om dit probleem op te lossen kan V*link* worden geïntegreerd met third party software als EzProxy, die dan het management van rechten en toegang regelt.

PARADOXALE SITUATIE

Bibliotheekdatabases (catalogi) zijn te lang veredelde kaartenbakken geweest. Door toepassing van moderne standaarden als OpenURL, OAI en XML en door creatief om te gaan met de data in de database worden allerlei 'verborgen' relaties zichtbaar. Vubis Smart zorgt op deze manier dat de informatie in de database en vooral de waarde ervan toeneemt, zonder dat tegelijk ook de kosten van beheer toenemen. Wel is het zo dat de kwaliteit van alle vermelde verrijkingstechnieken afhangt van de kwaliteit van de informatie die in de database is opgeslagen, ook al kunnen 'fouten' in het zoekproces met technieken als fuzzy logic gedeeltelijk worden gecorrigeerd. En zo tekent zich een paradox af: alhoewel de informatie in de gestructureerde bibliotheekdatabase als te beperkt ervaren wordt en moet worden gelinkt (verrijkt) met informatie buiten die database, is het net de kwaliteit en de rijkdom van de informatie in de bibliotheekdatabase die de kwaliteit van de links en de verrijking bepaalt.

Een goed voorbeeld hiervan is ook de zogenaamde 'trefwoordvoorspelling': de mogelijkheid dat het systeem met beperkte tussenkomst van de mens (die alleen nog controleert) documenten gaat 'rubriceren' (indelen in onderwerpsgebieden). Een dergelijke techniek bouwt verder op de informatie die gekoppeld is aan de al in de database aanwezige documenten en ontleent zijn waarde dan ook sterk aan de kwaliteit van die al aanwezige informatie. Bibliotheken en vergelijkbare organisaties en vooral hun informatietaak staan al enige tijd onder druk (van de informatie op het internet). Toch zijn dit soort organisaties - gezien hun expertise en gezien het feit dat ze beschikken over kwalitatief

Vervolg op pagina 36.

Elke bibliothecaris een eigen acroniem

- DOI: Digital Object Identifier, een systeem voor de identificatie en uitwisseling van intellectuele (digitale) eigendom (zie ook *http://www.doi.org*)
- Dublin Core: protocol voor de uitwisseling van metadata (zie ook *http://dublincore.org*)
- ISO2709: standaard voor het geautomatiseerd uitwisselen van catalogusrecords
- MARC: MAchine Readable Cataloguing, standaard daterend uit de jaren zestig, bedoeld voor het uniform beschrijven van catalogusrecords. De standaard kent vele (nationale en internationale) varianten, waarvan de twee belangrijkste Marc21 en UniMarc zijn. De eerste is vooral in gebruik in de Angelsaksische wereld, de tweede vooral op het Europese vasteland (met name in Frankrijk). Recent verschenen XML-versies van Marc21, genaamd MARC-XML en MODS (zie ook *www.loc.gov*)
- OAI: Open Archive Initiative Protocol for Metadata Harvesting, een applicatieonafhankelijk raamwerk voor het 'harvesten' van metadata, gebaseerd op XML (zie ook *http://www.openarchives.org/OAI/openarchivesprotocol.htm*)
- OpenURL: syntax voor het aanbieden van contextsensitieve services en informatie aan gebruikers op zoek naar informatie. OpenURL is op weg een officiële NISO-standaard te worden. De huidige voorlopige versie is door NISO geaccepteerd als OpenURL versie 0.1 (zie ook *http://library.caltech.edu/openurl/)*
- Z39.50: op het client/server-model gebaseerde standaard voor het zoeken en ophalen van informatie uit databases (zie ook *www.loc.gov*)

Vervolg van pagina 19.

hoogstaande, gestructureerde informatie - bij uitstek geschikt om de 'clicks' en de 'bricks' van de informatie met elkaar te verenigen. Hierbij verwijzen de 'bricks' overigens naar de expertise en de technieken van de bibliotheek en niet naar de bibliotheek als fysiek gebouw. Maar dan is het van groot belang dat die expertise en die gestructureerde informatie ten volle wordt benut. Vubis Smart helpt hen daarbij. ●

Eric Conderaerts (e.conderaerts@geac.nl) is productmanager bij Geac Benelux B.V.

VUBIS SMART IN OPENBARE BIBLIOTHEEK VELDHOVEN.