Drie zoekproducten voor ongestructureerde data

Infor Vspaces, Autonomy en FAST strijden om de macht

Robbert Hoeffnagel

Waar veel bedrijven beslissingen baseren op data die vastliggen in gestructureerde database-omgevingen, zit in veel gevallen de meeste kennis juist verstopt in ongestructureerde gegevens. Hoe halen we kennis uit die enorme hoeveelheden e-mails, voice-mails, webpagina's, tekstdocumenten, presentaties, illustraties en dergelijke? Aanbieders als Autonomy, FAST en – opmerkelijk genoeg – Infor doen verwoede pogingen hiervoor zinvolle oplossingen aan te dragen.

Onlangs introduceerde Autonomy een opmerkelijk product. Het gaat om een zogeheten 'informatie governance platform'. Autonomy Information Governance is volgens het bedrijf het eerste governance-platform dat in real-time policy's afdwingt als het gaat om de vraag wat er nu wel of niet met informatie mag gebeuren.

Compliance rules

Interessant genoeg hebben we het hier over programmatuur waarmee ongestructureerde data kunnen worden beheerd. Daarbij speelt de software de opmerkelijke rol dat het door een analyse van de content en de context waarin deze content wordt aangeboden, inzage creëert in de inhoud van e-mails, Word-documenten, presentaties en dergelijke. Op basis van dat inzicht wordt bovendien direct vastgesteld in hoeverre die content onderworpen is aan een of meer 'compliance rules' die bijvoorbeeld aangeven welke functionarissen de informatie mogen inzien, of de content buiten de firewall terecht mag komen, noem maar op.
Autonomy heeft het product opgedeeld in een aantal modules. Deze zijn gericht op compliance, op wat het bedrijf noemt 'enterprise legal hold' en op 'disposition management'. De eerste module is uiteraard gericht op het zoveel mogelijk voorkomen van overtredingen van wet- en regelgeving. De tweede component is gericht op juridisch medewerkers en maakt het mogelijk om de inhoud van in principe iedere repository binnen een onderneming vast te leggen. Hier spelen uiteraard ongestructureerde data een belangrijke rol, omdat veel juridisch relevante informatie nu eenmaal niet in databases vastligt maar in de vorm van e-mail en dergelijke beschikbaar is. De derde en laatste module – 'disposition management' – richt zich op het categoriseren van informatie waarbij het onderscheid met name kan worden gericht op de waarde die informatie voor de organisatie vertegenwoordigt. Ook regelt deze module dat informatie niet – al of niet per ongeluk – kan worden verwijderd.
De software biedt daarmee een combinatie van search in zowel ongestructureerde als gestructureerde informatie met een compliance engine. Autonomy is echter niet de enige partij die zeer actief is met het ontwikkelen van search-technologie waarmee talloze soorten en typen data kunnen worden ontsloten. Daarom kijken we in dit artikel – naast Autonomy – ook naar FAST dat inmiddels door Microsoft is overgenomen en – wellicht onverwacht – Infor.

Black box

Allereerst Autonomy. Opmerkelijk aan deze firma is dat het redelijk open praat over de technologie die het heeft ontwikkeld. In white papers stelt het bedrijf dat het redelijk 'naïef' te noemen is om te denken dat één standaard set aan methodieken en technieken voldoende is om ieder probleem op het gebied van enterprise search op te lossen. Een black box-aanpak is wat Autonomy betreft niet mogelijk.
Daarom probeert men twee benaderingen aan te bieden: de software kiest op basis van een analyse van de context voor de beste set van technieken en hulpmiddelen, maar biedt tegelijkertijd de mogelijkheid aan om met name de relevantie-algoritmes aan te passen.

Dat gebeurt met een *workbench* die het voor de applicatiebeheerder mogelijk maakt om de zoekresultaten te optimaliseren. In principe kan het relatieve gewicht van ieder veld worden aangepast en ook kan bijvoorbeeld de impact van bijvoorbeeld tikfouten inzichtelijk worden gemaakt en worden geneutraliseerd. Ook kunnen resultaten 'vastgezet' worden. Autonomy geeft hiervoor het voorbeeld van een query naar een gele Toyota Prius. Een andere gebruiker omschrijft die kleur wellicht als

'goud. Een zoekvraag naar een gele Prius moet dus wellicht gelijkgeschakeld worden met de vraag naar een 'gouden Prius'. Een simpel voorbeeld wellicht, maar het geeft het idee wel goed weer. Zou de beheerder niet in staat zijn om te sleutelen aan de resultaten van een zoekactie, dan kan dit dus een mindere kwaliteit aan informatie opleveren.

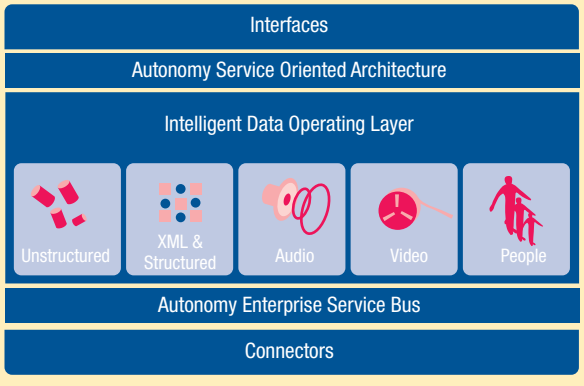
Verder geeft Autonomy aan dat het zich nadrukkelijk baseert op onder andere Bayesiaanse inferentie. Voor wie dat even kwijt is: de klassieke kansberekening stelt dat als een munt honderd keer wordt opgeworpen en de munt 99 maal met de 'kop' omhoog terecht komt, de kans dat bij de honderdste worp de munt met kop dan wel munt omhoog valt desondanks even groot is. De Bayesiaanse benadering zegt dan echter dat de munt in zo'n geval waarschijnlijk niet helemaal zuiver is of – bijvoorbeeld – wel eens aan beide zijden van één en dezelfde munt-beeltenis kon zijn voorzien. Kortom: in dit geval is de kans groot dat de honderdste worp ook een 'kop' zal opleveren.

Daarnaast heeft Autonomy goed gekeken naar het werk van bijvoorbeeld Claude Shannon om te bepalen welke concepten in een document het meest belangrijk of informatief zijn.

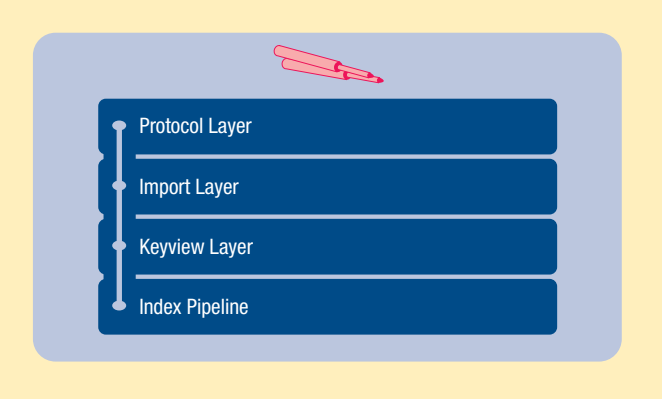
IDOL(s)

Het bedrijf past dit soort theorieën toe om met de hulp van een architectuur die 'Intelligent Data Operating Layer' ofwel IDOL wordt genoemd ongestructureerde data te kunnen ontsluiten. In afbeelding 1 is deze architectuur weergegeven. Hierin valt direct op dat een en dezelfde aanpak in staat is om zowel ongestructureerde als gestructureerde informatie te ontsluiten. Ook is duidelijk dat zowel interne als externe bronsystemen kunnen worden gebruikt. Dat voorkomt dus dat meerdere los van elkaar functionerende ontsluitingsmechanismen 'ergens' weer aan elkaar geknoopt moeten worden.

In feite past Autonomy een Service Oriented Architecture-aanpak toe. Of zoals men het zelf formuleert: het bedrijf biedt een reeks van computing-functies aan die zich onderscheiden naar betekenis. Net als bij meer traditionele SOA-projecten draait het allemaal om granulariteit (hoe groot of klein wordt een individuele service gekozen), modulariteit, hergebruik, het



Afbeelding 1: De 'Intelligent Data Operating Layer' van Autonomy.



Afbeelding 2: Een connector van Autonomy extraheert content uit een repository, importeert deze als IDX of XML en indexeert de data in de IDOL-server.

opdelen in componenten en dergelijke. Verder worden modules toegepast waarvan de beschikbaarheid automatisch kan worden vastgesteld, waarna deze op basis van SOAP met elkaar kunnen 'praten'. Een centrale servicebus – hier 'enterprise messaging bus' geheten – zorgt ervoor dat alle services met elkaar kunnen communiceren.

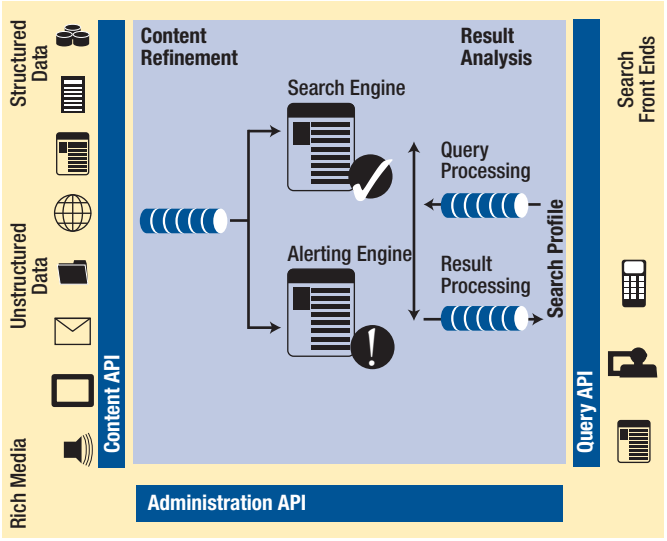
Daarnaast spelen connectoren een hoofdrol in de aanpak van Autonomy. Deze zijn namelijk verantwoordelijk voor het ophalen van content uit een lokaal of op afstand geplaatste repository. De data worden in de vorm van een IDX- of XML-bestand geïmporteerd en vervolgens geïndexeerd. Het aantal connectoren is groot: meer dan vierhonderd. Hieronder bevinden zich interfaces naar bekende omgevingen als SharePoint, Documentum, Exchange, Notes, Oracle en dergelijke. Iedere connector extraheert de complete inhoud van een bestand, hetgeen volgens Autonomy de nauwkeurigheid ten goede komt.

Parameters

In afbeelding 2 is weergegeven hoe zo'n importactie plaatsvindt. Hierbij wordt gebruik gemaakt van Autonomy's KeyView-technologie zodat alle metadata en tekst in meer dan duizend dataformaten in de IDOL-omgeving kunnen worden gebracht. Daarnaast is voorzien in een zogeheten KeyView Filter, waardoor direct tijdens het importeren al gefilterd kan worden. De feitelijke import is in hoge mate instelbaar. Dit gebeurt met de hulp van ruim driehonderd parameters die door een applicatiebeheerder aangepast kunnen worden. Ook bestaat de mogelijkheid om extra velden of values te creëren. Ten aanzien van te importeren mail is het bovendien mogelijk om op separate servers aanwezige mails aan het elkaar te koppelen.

De koppeling met legacy-systemen gebeurt via een zogeheten 'Legacy Compatibility Module' (LCM). Het voordeel hiervan is dat bestaande legacy-systemen kunnen worden vervangen door de IDOL-server, zonder dat de bestaande workflow behoeft te worden aangepast.

Een connector is in staat data te importeren met een snelheid van 60 GB per uur, waarbij iedere connector een overzicht bijhoudt van



Afbeelding 3: FAST ESP kent een gedistribueerde architectuur.

welke bestanden zijn geïmporteerd, wat de security-instellingen zijn, waar zich belangrijke 'data points' bevinden en dergelijke. Hierdoor kan tot een soepele synchronisatie tussen de Autonomy-omgeving en de oorspronkelijke bronsystemen worden gekomen.

Context

Het onlangs door Microsoft overgenomen FAST is een andere bekende naam in de wereld van enterprise search. Het bedrijf biedt met ESP een zogeheten 'Enterprise Search Platform'. Kenmerkend voor de aanpak van de Noren is een zogeheten semantische index. Deze heet officieel 'Contextual Insight' en is bedoeld om de context en de bedoeling van een query te kunnen vaststellen. Dat vergroot de precisie van de zoekactie, meent het bedrijf en biedt de gebruiker bovendien een aantal hulpmiddelen om de vraag verder te verfijnen.

De genoemde semantische index herkent de structuur van een document en optimaliseert de zoekvraag hierop. Normaliter wordt bij een vraag een redelijk grove indeling gebruikt waarbij onderscheid wordt gemaakt in bijvoorbeeld document, webpagina of record. Dat is niet verfijnd genoeg, meent FAST. Hier kiest men voor een verfijning naar ieder willekeurig XML- ofwel gestructureerd element. Bovendien wordt gebruik gemaakt van zogeheten 'entity metadata'. Een entity is hierbij de naam van een persoon, een telefoonnummer, een geografische locatie en dergelijke. FAST onderkent inmiddels veertig van dit soort entiteiten en dat aantal zal de komende jaren alleen nog maar verder toenemen. Bovendien is voorzien in de mogelijkheid om als applicatiebeheerder zelf bedrijfsspecifieke entity's te definiëren. ESP kent een aantal tools die het de gebruiker mogelijk maken om snel en efficiënt door de zoekresultaten te gaan. Zo biedt de zogeheten 'FAST Classifier' mogelijkheden om langs hiërarchische weg door resultaten 'te lopen'. Een vorm van dynamische drill-down maakt het mogelijk om search and navigate-applicaties te ondersteunen.

Voordat de data die naar de search engine worden gebracht formeel worden vastgelegd, is sprake van een slag die 'refinement' wordt genoemd. Hierbij wordt content (gestructureerd, semi-gestructrureerd en ongestructureerd maar ook rich media) onder andere geanalyseerd, getransformeerd en waar nodig of mogelijk verrijkt. Pas daarna vindt het feitelijke indexeren plaats. Het aantal bronsystemen, talen (81) en dataformaten (meer dan vierhonderd) dat wordt ondersteund, is net als bij Autonomy zeer groot. Bovendien is voorzien in tools om bedrijfs- of branche-specifieke formaten die niet out of the box worden ondersteund, alsnog toe te voegen.

De in afbeelding 3 weergegeven architectuur is met name met schaalbaarheid in het achterhoofd ontwikkeld, vertelt men bij FAST. Dat begrip 'schaalbaar' is voor meerdere uitleg vatbaar: het datavolume, het aantal query's per seconde, maar ook als het om het toevoegen van nieuwe data gaat. Verder valt op dat ook bij ESP integratie met andere omgevingen een belangrijke rol speelt. Nieuwe services kunnen zonder problemen worden toegevoegd, waarbij beheer- en operationele API's als Java- of SOAP/WSDL-webservices beschikbaar zijn, terwijl de content- en zoek-API als Java-, C#- (voor .Net) en C++-services aangeroepen kunnen worden.

Vreemde eend

Ten slotte een wellicht wat vreemde eend in de bijt: Infor. Dit concern staat toch vooral bekend als de eigenaar van de Baan ERP-producten en andere enterprise-applicaties. Daarnaast is het

Zoekprotocollen in Vspaces

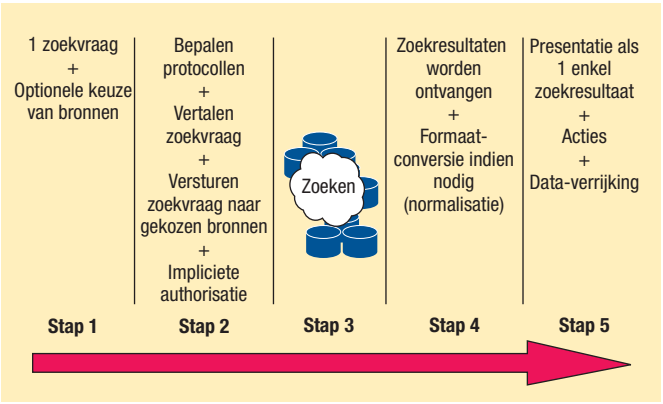
Vspaces ondersteunt een aantal zoekprotocollen:

- Z39.50 – voor bibliografische information retrieval;
- XML-gateways – voor het ondervragen van gateways als PubMed via het HTTP-protocol;
- SRU – search and retrieval via url; in feite de opvolger van Z39.50;
- OpenSearch – een door Amazon.com gelanceerd zoekprotocol dat via HTTP een url van een server verstuurt en als RSS geformatteerde data retourneert;
- MXG – MetaSearch XML Gateway; een gestandaardiseerde XML-gateway;
- HTTP/HTML – voor HTML parsing zodat webpagina's kunnen worden ontleed;
- Vubis Smart XML-gateway – een voor Vubis Smart ontwikkelde gateway die zich gedraagt als het SRU-protocol;
- Vubis Smart Proprietary – een directe koppeling tussen Vspaces en Vubis Smart-databases;
- ODBC/SQL – voor relationele databases. Hiermee kunnen data uit bronsystemen worden gekopieerd en bevraagd. Dit is echter een laatste redmiddel en heeft niet de voorkeur, stelt Infor.

bedrijf echter ook eigenaar van het aloude Geac. En die firma is natuurlijk sinds jaar en dag actief in de markt voor bibliotheeksystemen en bijvoorbeeld zoektechnologie.
Het product van Infor heet Vspaces en vormt een geïntegreerde portal die toegang biedt tot tal van onderliggende informatiebronnen. De software – 'deze omgeving' is wellicht een betere typering – steunt op vijf pijlers: resource discovery, geïntegreerd zoeken, het verrijken van de data, personalisering en diensten die de gebruiker in staat stellen om de in de portal verzamelde informatie nog eens verder te verrijken.
Infor hanteert nadrukkelijk de term 'broadcast search'. Deze vorm van zoeken maakt het mogelijk om via één interface en door het stellen van één zoekvraag simultaan meerdere bronnen te ondervragen. De gevonden resultaten worden als één geïntegreerde lijst van resultaten gepresenteerd. De term 'broadcast' staat hier dus letterlijk voor het uitzenden van de zoekvraag.
Voor de goede orde: in de literatuur komen we voor deze vorm van zoeken ook andere kreten tegen, zoals meta search, federated search, parallel search of single search. Interessant hierbij is dat het in principe niet uitmaakt waar een bron zich bevindt, zolang deze maar via internet bereikbaar is.

Vijf stappen

In afbeelding 4 is weergegeven hoe zo'n broadcast search bij Vspaces in zijn werk gaat. Het zoekproces bestaat bij Vspaces dus uit vijf stappen. In de eerste stap formuleert de gebruiker zijn vraag. Eventueel kan aan de zoekactie een beperking van het aantal bronnen worden meegegeven. Afhankelijk van de gekozen implementatie kunnen bronnen eventueel gegroepeerd worden. Dat kan op grond van het type bron (zoekmachine, bibliografische database en dergelijke), onderwerp (actualiteit, geschiedenis en dergelijke) of een andere zelf gekozen groepering. Per bron kunnen de nodige details worden vastgelegd en aan de gebruiker beschikbaar worden gesteld. Denk aan een beschrijving van de bron, de taal, onderwerpen die behandeld worden en dergelijke. Iedere gebruiker kan een voorkeur voor bepaalde bronnen of groepen van bronnen vastleggen en bewaren.



Afbeelding 4: Het proces van 'broadcast search' zoals Vspaces van Infor dit hanteert.

Interessant aan Vspaces is dat na het ingeven van de zoekvraag de te doorzoeken bronnen eerst opgespoord dienen te worden. Er is dus geen sprake van een permanente koppeling, maar meer van een ad hoc tot stand te brengen interface. Uiteraard zal het hierbij wel nodig zijn dat een toegangsautorisatie plaatsvindt, de zoekvraag vertaald wordt naar een formaat waarmee de bron overweg kan en zal een aantal protocolkeuzes gemaakt moeten worden.
Ook zullen de eenmaal ontvangen resultaten vertaald moeten worden zodat deze aan de gebruiker kunnen worden gepresenteerd. Zoekresultaten kunnen naar het scherm worden gebracht zodra deze binnen zijn. Het is dus niet noodzakelijk om te wachten tot alle te doorzoeken bronnen op de vraag hebben gereageerd. Zodra een gebruiker in deze lijst een keuze maakt voor een specifiek resultaat probeert Vspaces dit resultaat weer te geven in de oorspronkelijke interface. Het handige aan deze presentatiemethode is dat ook de volledige functionaliteit van het bronsysteem beschikbaar komt.

Connectoren

Interessant is verder dat vanuit deze interface ook direct een vervolgvraag kan worden gesteld. Hiertoe heeft Infor Vlink geïmplementeerd, een zogeheten 'link resolver'. Deze is gebaseerd op het OpenURL-protocol, maar wordt alleen als optie aangeboden. Dit is dus geen functionaliteit die standaard beschikbaar is. Een zoekvraag kan ook geen resultaat opleveren. In dat geval is Vspaces in staat om op basis van fuzzy logic-technieken alternatieve zoektermen te genereren en als zoekvraag te broadcasten. Hierbij worden de alternatieven eerst aan de gebruiker getoond zodat deze invloed kan uitoefenen op de herformulering van zijn oorspronkelijke zoekvraag.

Ook Vspaces werkt met connectoren. Deze regelen onder andere het vertalen van de zoekvraag, het selecteren van het zoekprotocol, de autorisatie bij het bronsysteem, het feitelijke versturen van de zoekvraag, het ontvangen van het resultaat en het vertalen van het technische formaat. De software ondersteunt een reeks van zoekprotocollen (zie kader 'Zoekprotocollen in Vspaces'). Vspaces is in principe niet gebonden aan specifieke XML-schema's en ondersteunt onder andere MarcXchange, MarcXML (waaronder Marc21 en UniMarc), Dublin Core XML, generieke XML, RSS en dergelijke. Om de autorisatie en authenticatie goed en eenvoudig te regelen, ondersteunt de software IP-filtering en username/password-combinaties.
Om de eerder genoemde personalisering mogelijk te maken, biedt Vspaces ten slotte het gebruik van profielen. Dit betekent onder andere dat de vormgeving van de zoekpagina kan worden aangepast en dat gekozen kan worden hoe de zoekresultaten worden aangeboden: records kunnen worden gedownload, per mail worden verzonden of desnoods als RSS news feed worden afgenomen.

Robbert Hoeffnagel is freelance journalist.