Het einde van gestructureerde opslag?

Enterprise Search

Adriaan Hondelink, Anja van der Lans en Victor van der Lelij

De dagelijkse aanwas van e-mail, data, documenten is zo groot dat de gemiddelde medewerker de stroom nauwelijks kan bijhouden. Interessant is te kijken of er mogelijkheden zijn de gebruiker tegemoet te komen; een Enterprise Search Engine kan uitkomst bieden.

Organisaties bieden hun medewerkers steeds meer applicaties aan voor allerhande specifieke toepassingen; databases en platformen in verschillende vormen en met elk hun eigen manier om de data of informatie terug te vinden en vervolgens te verwerken tot interpretaties en nieuwe kennis. Gebruikers raken inmiddels de draad kwijt in de wirwar van applicaties en vragen om uniformiteit, niet alleen om *single sign-on* om het aantal aanlogprocedures (en hinderlijke passwords) te beperken, maar ook om een overzichtelijke en gestandaardiseerde manier voor het bevragen van de applicaties.

Paraplu

Niet alleen het aantal applicaties maar ook de dagelijkse aanwas van e-mails, harde data, elektronische documenten en interessante berichten is zo groot dat de gemiddelde medewerker de stroom nauwelijks kan bijhouden, laat staan nog in staat is deze op een ordentelijke manier te structureren en op te bergen.
Tot voor kort was het noodzakelijk om informatie te structureren met behulp van metadata om het later terug te kunnen vinden. Het dagelijkse gevecht van de gebruiker tussen het uitvoeren van de primaire taak en het toekennen van metadata aan informatie is zelden succesvol voor de metadata verlopen.
Het doel van het toekennen van metadata is om daarna via een folderstructuur, op basis van de toegekende metadata, te kunnen navigeren naar de juiste map waarin de informatie opgeborgen is of via het zoekscherm documenten waarin een term voorkomt te kunnen ophalen. Gebruikers willen het liefst geen metadata toekennen aan documenten en als het dan toch moet, beperken tot een uiterst minimale set. Ervaring leert dat het ingeven van vijf termen het maximale vereist van de gebruiker, daarboven gaat de gemiddelde gebruiker de applicatie mijden en valt terug op vertrouwde gedragspatronen, waarin het toekennen van metadata geen verplichting is.

De combinatie van de grote hoeveelheid te gebruiken applicaties, de dagelijkse aanwas van informatie en de inspanning ten aanzien van het toekennen van metadata maken het interessant te kijken of er mogelijkheden zijn de gebruiker tegemoet te komen in de strijd. Een *Enterprise Search Engine* kan hierin uitkomst bieden.

Een Enterprise Search Engine is te vergelijken met een paraplu die boven de applicaties hangt en waarmee informatie van verschillende aard beschikbaar wordt gesteld aan de gebruiker. De aard, vorm of inhoud en origine van de informatie zijn daarbij ondergeschikt. Dit wil zeggen dat zowel gestructureerde informatie uit databases als ongestructureerde informatie uit internetomgevingen en alle varianten daartussenin, via een Enterprise Search Engine bevraagd kunnen worden. De Enterprise Search Engine slaat elk aangeboden woord of stukje data op in een centrale index. Door het bevragen van de centrale index kan de informatie die ligt opgeslagen in de diverse applicaties worden teruggevonden. Met Enterprise Search is het opslaan van gestructureerde data (en dus voorzien van metadata) in een database geen randvoorwaarde meer om relevante informatie te kunnen terugvinden.
Voor bestaande databases zal dit op termijn kunnen betekenen dat hun toegevoegde waarde voor het besluitvormingsproces minder wordt. Deze databases zullen worden vervangen door meer flexibel in te richten omgevingen, waarin door middel van search vrijelijk over informatie (al dan niet voorzien van een structuur) kan worden beschikt.

Enterprise Search toepassen

Enterprise Search is volgens ons de mogelijkheid om met behulp van elektronische hulpmiddelen data c.q. informatie uit vele verschillende bronsystemen zoals databases, file-systemen,

legacy-systemen, document management-systemen etcetera binnen een organisatie via één standaard centrale zoek-interface toegankelijk te maken. Om te beginnen worden enkele veelgebruikte termen toegelicht.
Bronsystemen: applicaties met de data uit het (legacy)systeem die je met search wilt ontsluiten.
Zoekresultaat: de verzameling van gerelateerde informatie die je gepresenteerd krijgt naar aanleiding van je zoekvraag.
Enterprise Search Engine: het systeem dat wordt gebruikt voor Enterprise Search.

Er bestaan grofweg twee verschillende manieren om Enterprise Search te implementeren. Een eerste mogelijkheid is op basis van zogenaamde *Federated Search* en de tweede manier is op basis van *Integrated Search*. Federated Search is in staat om één zoekvraag aan meerdere bronsystemen tegelijk te stellen.

> De Search Engine indexeert de aangewezen data van het bronsysteem in een centrale index

Bij Federated Search zal de Search Engine per bronsysteem de zoekvraag aanbieden aan de zoekmachine die bij het bronsysteem zelf hoort. De Search Engine verzamelt alle antwoorden van alle bronsystemen en presenteert deze als een gezamenlijk zoekresultaat. Vaak zijn deze resultaten gegroepeerd per bronsysteem en niet te sorteren op relevantie van het resultaat. Het zoekresultaat wordt getoond in de vorm van een hyperlink. Een klik op de hyperlink zal ervoor zorgen dat het betreffende bronsysteem wordt opgestart en het inhoudelijke resultaat vanuit zijn bronsysteem wordt getoond. Voorbeeld van Federated Search is de website www.wieowie.nl. Na het intypen van de voornaam en achternaam worden alle bronsystemen geraadpleegd via hun eigen zoeksysteem. De zoekresultaten worden gepresenteerd per bronsysteem.

Integrated Search werkt anders. Het is net als bij Federated Search mogelijk om één zoekvraag aan meerdere bronsystemen tegelijk te stellen. De Search Engine wordt per bronsysteem zodanig geconfigureerd dat de Search Engine direct toegang heeft tot de brondata. De Search Engine indexeert vervolgens de aangewezen data van het bronsysteem in een centrale index (deze kan dus afhankelijk van het aantal bronnen heel groot worden). Zodra er een zoekvraag wordt gesteld, zal de Search Engine in staat zijn op basis van zijn eigen index razendsnel de zoekresultaten te presenteren aan de gebruiker. Voorbeeld van Integrated Search is www.google.nl. Hierbij is kort geleden bekend geworden dat Google 15.000 servers wereldwijd beschikbaar heeft gesteld voor het tonen van de eerste zoekpagina. Het totale aantal servers dat Google gebruikt is niet bekend.

Verschillen

Kenmerkende verschillen tussen Integrated Search en Federated Search zijn:

- De implementatie van Federated Search is relatief gezien eenvoudiger, want je maakt gebruik van de bestaande zoekfunctionaliteiten van de bronsystemen. De implementatie zal zich focussen op het overzichtelijk tonen van de verschillende zoekresultaten uit de diverse bronsystemen. Het succes van Federated search is zeer afhankelijk van de kwaliteit van de zoekfunctionaliteiten in de diverse bronsystemen en is daarmee kwetsbaar en lang niet altijd bruikbaar;
- Integrated Search zal aanzienlijk sneller zoekresultaten tonen dan Federated Search omdat bij Integrated Search alle data al beschikbaar zijn via de eigen opgebouwde index. Federated Search is zo snel als de langzaamste zoekapplicatie van de bronsystemen;
- Federated Search biedt weliswaar een overzicht van de zoekresultaten, maar is niet in staat om deze data op verschil-

De toekomst van Enterprise Search

Enterprise Search is de laatste jaren in een stroomversnelling gekomen doordat allerlei internet-toepassingen beschikbaar zijn gekomen in de search engines. Denk daarbij aan de categorisering en gebruikersvriendelijke zoek-interfaces. Deze trend zal zich de komende jaren verder doorzetten. De ontwikkelingen waar de search engine leveranciers voor de komende jaren op in zetten zijn:

1. Real-time video indexeren op basis van spraakherkenning. De spraakherkenning wordt in tekstvorm aan het videomateriaal toegevoegd en geïndexeerd. Op basis van deze index kunnen videobeelden snel doorzocht worden.
2. Door middel van analysetoepassingen kunnen de zoekresultaten en de onderlinge relaties van de zoekresultaten ook via slimme visualisatietechnieken zichtbaar worden gemaakt. Denk hierbij aan het visualiseren van relaties tussen verschillende gegevens.
3. Het verrijken en opslaan van zoekresultaten uit (on)gestructureerde bronnen in een semi-gestructureerd systeem, zoals bijvoorbeeld in een document management-systeem. Een dergelijke toepassing is denkbaar in de recherchewereld. Bij rechercheerwerkzaamheden worden data uit gestructureerde en ongestructureerde systemen verzameld. Deze gegevens worden vervolgens gestructureerd opgeslagen, als onderdeel van het 'strafdossier' van een mogelijke dader.
4. Integratie van Business Intelligence/datawarehousing (sturen op getallen, KPI etcetera) in combinatie met Enterprise Search (sturen op kennis). Ook hierin speelt de analyse van de beschikbaar gekomen informatie een belangrijke rol. Met andere woorden: hoe moet ik de informatie interpreteren zodat ik er mijn beleid op kan afstemmen?

Metadata

Het toekennen van metadata gebeurt meestal bij het invoeren van informatie. Met Enterprise Search is het mogelijk om metadata tijdens het zoeken toe te kennen. Een van de belangrijkste redenen om informatie op te slaan in een eigen informatiesysteem is het kunnen terugvinden van en beschikken over de opgeslagen informatie op een later moment. Bij het opzetten van een informatiesysteem wordt er altijd veel aandacht besteed aan het zo gestructureerd mogelijk opslaan van de informatie, door middel van het toekennen van metadata. Het toevoegen van metadata aan informatie heeft als hoofddoel de 'terugvindbaarheid' van de informatie te verhogen. Er zijn vele systemen bedacht die gebruikers er toe dwingen informatie te voorzien van metadata. Een gebruiker die metadata toekent aan een document doet dit op basis van kennis, kunde en ervaring. Als het niveau van één van de drie niet voldoende is kan dit de kwaliteit van de toegekende metadata beïnvloeden. Dit heeft weer direct gevolgen voor de terugvindbaarheid en beschikbaarheid van de informatie.

In plaats van het toekennen van metadata tijdens het invoeren van de informatie in het informatiesysteem, kunnen de metadata ook tijdens het zoeken naar de informatie worden toegekend. Het toekennen van metadata gaat vaak op basis van regels. Deze regels worden via een sleutelwoord (keywords/metadata) toegekend aan de informatie. De regels van het toekennen van metadata kunnen ook worden toegepast tijdens het zoeken in het informatiesysteem. Door het omzetten van de basisregels in zoekvragen en deze zoekvragen te koppelen aan de sleutelwoorden, kan informatie door een zoekmachine worden gecategoriseerd (metadateren). De zoekmachine kent dus de metadata toe in plaats van de gebruiker. Het is echter niet de zoekmachine maar de gebruiker die, via zoekvragen, bepaalt hoe de informatie wordt gecategoriseerd.

Het grote voordeel van het laten categoriseren van de informatie tijdens het zoekproces is dat het op een eenduidige en gestandaardiseerde manier gebeurt. Daarnaast kunnen nieuwe categorieën eenvoudig worden toegevoegd en kunnen bestaande regels, op basis van voortschrijdend inzicht, worden verfijnd zonder dat de informatie in het informatiesysteem hoeft te worden aangepast.

lende manieren inzichtelijk te maken of de resultaten van de verschillende bronnen met elkaar te wegen. Gesteld kan worden dat iedere zoekmachine zijn eigen wegingalgoritme heeft. Het groeperen op relevantie (weging) van resultaatitems uit de meerdere resultaatlijsten, afkomstig van meerdere zoekmachines, levert zodoende een onbeduidend resultaatlijstje op. Integrated Search kan dat wel, omdat er maar één wegingalgoritme wordt gebruikt tijdens het zoeken. Daarnaast biedt een Integrated Search extra functionaliteiten, zoals het gebruik van een overkoepelende taxonomie. Dit biedt een aanzienlijke toegevoegde waarde ten opzichte van de bestaande zoekapplicaties;

- Integrated Search zal altijd een zoekresultaat tonen, zolang als er data in de index staan die overeenkomen met de opgegeven zoektermen, ook als het achterliggende bronsysteem niet beschikbaar is. De index voorziet namelijk in de data. Dat betekent dat ook referentiebestanden (bestanden met verwijzing naar vindplaatsen van documenten; papier of digitaal) of near-line bronnen ontsloten kunnen worden;
- Kenmerkend verschil tussen Federated Search en Integrated Search is verder de behoefte aan opslagcapaciteit. Federated Search creëert nauwelijks een behoefte aan extra opslagcapaciteit; de data blijven immers in het bronsysteem, terwijl de index van Integrated Search kan leiden tot een verdubbeling van benodigde opslagcapaciteit. Immers, de data zijn ook beschikbaar indien het bronsysteem niet beschikbaar is. Gelukkig is opslagcapaciteit tegenwoordig niet zo'n kostbare zaak meer. Het beheer van de centrale index brengt echter wel specifieke zorgen en kosten met zich mee;
- Een laatste verschil is dat bronsystemen die geen zoekfunctionaliteit hebben, bijvoorbeeld file-systemen, of een hele beperkte zoekfunctionaliteit bieden zoals legacy-applicaties, eigenlijk niet in aanmerking komen voor Federated Search. Het duurt domweg te lang om bijvoorbeeld de G-schijf real-time te doorzoeken om aansluitend het zoekresultaat te tonen;
- Natuurlijk kleven er ook nadelen aan Integrated Search. Als de server of de index uitvalt is geen enkel onderliggend bronsysteem bereikbaar. Bij Federated search kan altijd teruggevallen worden op de eigen 'interne' index van het bronsysteem.

De afweging tussen Integrated Search of Federated Search is een afweging die je mede op basis van bovengenoemde verschillen zal moeten maken. Meestal verdient Integrated Search de voorkeur. Doorslaggevende argumenten hierbij zijn snelheid en de meerwaarde die een taxonomie biedt bij de navigatie door de zoekresultaten.

Ongestructureerd versus gestructureerd

Nu we weten dat er onderscheid is in de wijze van toepassen van search, gaan we ons verder verdiepen in het onderscheid tussen ongestructureerde gegevens en gestructureerde gegevens. Voorbeelden van ongestructureerde gegevens zijn natuurlijk e-mail (inclusief bijlagen), Office-bestanden en bijvoorbeeld intranet-bestanden. Gestructureerde gegevens zijn gegevens die in gestructureerde databases of bestanden (XML) zijn vastgelegd, zoals bijvoorbeeld adresgegevens in een CRM-systeem of de financiële administratie in SAP, Siebel of Oracle databases. Semi-ongestructureerde gegevens zijn bijvoorbeeld gegevens in een document management-systeem waarbij bijvoorbeeld Word-documenten worden opgeslagen. De inhoud van het document wordt niet gestructureerd opgeslagen, maar wel de metadata van zo'n document (zoals datum, auteur, trefwoorden, status en versie).

Uit onderzoeken van Gartner Inc. blijkt dat het merendeel van de organisaties beslissingen neemt op basis van gestructureerde

informatie die tot hen komt in de vorm van rapportages. De gestructureerde informatie beslaat 30 procent van alle data die binnen een organisatie liggen opgeslagen. Algemeen zou je dan dus kunnen concluderen, in termen van management en verantwoording, dat organisaties worden bestuurd, innovaties worden bedacht, investeringen worden gedaan en verantwoording wordt afgelegd op basis van slechts 30 procent van de aanwezige kennis en informatie. Van alle informatie blijft dus 70 procent onbenut.

Waarde (semi)ongestructureerde informatie

Het lijkt er vaak op dat alleen de gestructureerde informatie binnen de organisatie status heeft, gecontroleerd is en dat je daarmee een volledig beeld beschikbaar hebt om de organisatie mee te kunnen sturen. Iedereen weet dat het in werkelijkheid anders is. We mogen concluderen dat ook de (semi)ongestructureerde kennis onontbeerlijk is voor de besluitvorming. Moeten we nu alle (semi)ongestructureerde informatie gaan structureren? Nee, maar men moet als organisatie wel nadenken hoe men deze blijkbaar toch belangrijke informatie wil ontsluiten voor de medewerkers. Houd je gestructureerde en ongestructureerde informatie gescheiden, of bied je zoekresultaten in combinatie aan? Er is veel mogelijk, maar het vereist wel een grondige oriëntatie op de techniek van ontsluiting en te verwachten resultaten. Daarnaast moet men zich bewust worden van welke gegevens welke waarde vertegenwoordigen en welke 'relevantie' gegevens hebben. Enterprise Search biedt een scala aan extra mogelijkheden om zoekresultaten te ordenen en hiermee nog sneller de juiste informatie te vinden die nodig is voor beslissingen. Ook is het mogelijk om zowel gestructureerde als (semi)ongestructureerde gegevens via dezelfde Enterprise Search oplossing te ontsluiten. Hiermee creëer je een krachtig middel waarmee je potentieel 100 procent van de bedrijfsinformatie kan ontsluiten zonder hiervoor een kostbare bedrijfsbrede gestructureerde gegevensapplicatie te moeten implementeren.

Conclusie

Betekent Enterprise Search het einde van de gestructureerde opslag? Nee, want ook gestructureerde opslag is van grote waarde. Dit artikel wil wel de noodzaak tot opslag van data in een gestructureerde vorm ter discussie stellen. Dat is een intensieve bezigheid, inclusief de bijkomende kosten van arbeidsintensieve invoer en bouw, onderhoud en classificatie van het datasysteem, terwijl slechts een beperkt deel van de informatie beschikbaar komt. Als organisaties aan de slag gaan met het ontsluiten van de overige 70 procent aan informatie, geeft dat ongekende extra mogelijkheden voor de ontwikkeling van kennis binnen de organisatie.
Het einde van gestructureerde opslag is nog lang niet in zicht, maar met de huidige technologie in de meeste geavanceerde Enterprise Search Engines is het heel goed mogelijk af te stappen van de nadruk op gestructureerde opslag in databases en datawarehouses, ten faveure van een overkoepelende zoekschil.

Adriaan Hondelink, Anja van der Lans en Victor van der Lelij
Adriaan Hondelink is Managing Director van Content Strategy. Anja van der Lans is managing consultant ECM bij Capgemini. Victor van der Lelij is senior consultant bij Content Strategy.

Update

Autonomy introduceert intelligent platform voor informatie-governance

Autonomy Information Governance is volgens Autonomy het eerste informatie governance platform dat policy management real-time automatiseert door het vormen van een conceptueel en contextueel inzicht in alle bedrijfsinformatie. In plaats van het kwalificeren aan de hand van metadata kan dit platform informatie begrijpen en vervolgens kwalificeren.
Het is nu mogelijk om de verbanden tussen verschillende soorten regelgeving en het informatiebeleid in kaart te brengen en hier de bijbehorende acties aan te verbinden.
Om policy management te automatiseren biedt Autonomy Information Governance een dynamische real-time omgeving, waarin de beleidsinformatie van organisaties gevisualiseerd en gecontroleerd kan worden. Dit is mogelijk dankzij de neutrale infrastructuur, waarin gebruik wordt gemaakt van meer dan vierhonderd out-of-the-box data repository connectoren. Deze maken het mogelijk om lokaal opgeslagen informatie zoals e-mails, documenten, audio- of video-informatie te beheren en terug te vinden.

Autonomy is door Logica verkozen als preferred partner op het gebied van enterprise search en high-end informatieverwerking. Het partnership houdt in dat de Autonomy technologie snel wordt ingevoerd bij die relaties van Logica die hoge eisen stellen ten aanzien van kennismanagement, compliance en informatiebeveiliging.

De basis van deze aankondiging is al jaren geleden gelegd door succesvol afgeronde projecten bij een aantal internationale klanten van Logica, waarbij Logica en Autonomy samen betrokken waren. Beide organisaties verwachten dat de samenwerking een grote toegevoegde waarde genereert. Logica kan nu innovatieve oplossingen voor klanten ontwikkelen op basis van Autonomy's geavanceerde IDOL (Intelligent Data Operating Layer) platform. Autonomy profiteert van het uitgebreide internationale netwerk van Logica.