Betuws bedrijf biedt prettig en laagdrempelig starterspakket

Met Collexis instappen in kennismanagement

Paul van der Linden en Izabella Smits-Pukrop

Aandacht voor textmining leidt haast vanzelf tot een uitstapje op het bredere gebied van kennismanagement. Hetzelfde één-tweetje valt ook te constateren bij de software die Tarchon en Collexis op de markt brengen. In een voorgaand artikel stond de intuïtieve software van Tarchon centraal, die duidelijk is ontworpen met het eindgebruik voor ogen. Nu zoomen Paul van der Linden en Izabella Smits in op Collexis, een andere Nederlandse vlaggendrager in textmining en kennismanagement.

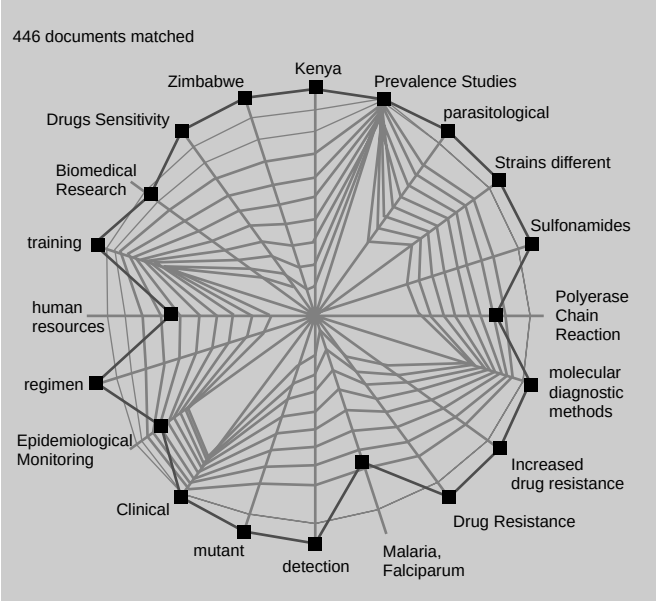
Het Nederlandse bedrijf Collexis is in 1999 ontstaan. Collexis is zowel de naam van de organisatie als van de software. CEO is Peter van Praag, die zo'n twintig jaar IT-ervaring heeft en eerder verbonden was aan Raet, CMG en ICL. Erik van Mulligen is de CTO. Hij houdt zich bezig met de technologie en doet onderzoek voor de Erasmus Universiteit Rotterdam (EUR).

Peter van Praag legt uit dat vijf jaar geleden de Nederlandse en Duitse overheden zich hebben gericht op de chaos die vaak ontstaat rondom ontwikkelingsprojecten. Wetenschappers die bepaalde materialen nodig hebben versturen het verzoek daaromtrent meestal niet slechts één keer, maar meerdere malen. Resultaat hiervan is dat zij dan ook meer dan één keer hetzelfde toegestuurd krijgen, hetgeen niet efficiënt is. De landen die de goederen versturen, wilden graag weten welke verzoeken afkomstig zijn van welke projecten. Om hierachter te komen wilde men de wetenschappers ongestructureerde teksten op hun laptop in laten geven. Door vervolgens de gegevens van deze laptops te synchroniseren kan dan een totaaloverzicht worden verkregen. Deze projectopzet kreeg de naam Shared. Het probleem dat hierbij bestaat is dat doublures kunnen optreden. Daarop werd aan de Erasmus Universiteit gevraagd om te onderzoeken hoe dit het beste kan worden opgelost. De door de EUR ontwikkelde toepassing vormt de basis van het huidige Collexis. De eigendomsrechten hiervan zijn inmiddels door Collexis verworven. Ook de bij de ontwikkeling daarvan betrokken mensen (wetenschappers zoals Erik van Mulligen) zijn nu verbonden aan Collexis. Sinds 1999 is het product commercieel beschikbaar.

Figuur 1: Voorbeeld van een thesaurus in Collexis. Thesauri, die gevalideerde kennis bevatten over specifieke kennisgebieden, vormen een belangrijk onderdeel van de werkwijze die Collexis voor ogen staat.

Textmining en kennismanagement

In DB/M 7 van vorig jaar is aandacht besteed aan textmining, waarbij textmining werd getypeerd als het onbekende neefje van datamining. Textmining wint momenteel aan attentiewaarde, omdat een steeds groter deel van de informatie die wij willen benutten zich bevindt in bronnen anders dan goed gestructureerde databases. Denk bijvoorbeeld aan webpagina's, documenten en presentaties. Zie hier de behoefte ook deze informatiebronnen te kunnen doorzoeken.
Maar in het verlengde daarvan willen we ook alle informatie -ongeacht de bron- met elkaar in verband kunnen brengen en ter beschikking stellen van gebruikers. Dat we hierbij rekening wensen te houden met de categorie gebruiker en de individuele gebruiker spreekt welhaast voor zichzelf. En daarmee komen we in het wijdere vaarwater van kennismanagement.



FIGUUR 2: VOORBEELD VAN EEN 'FINGERPRINT'.

MENS

Momenteel telt het in Geldermalsen gevestigde bedrijf 25 medewerkers, een aantal dat volgens Peter van Praag nog maandelijks met één of twee groeit. Ongeveer de helft houdt zich bezig met de verdere ontwikkeling van de technologie. De andere helft doet marketing en sales. De universitaire achtergrond speelt nog steeds een rol. Een deel van omzet en ontwikkelde technologie wordt teruggeploegd naar het Shared-project.

Van Praag stelt dat het juist bij kennismanagement zaak is terug te gaan naar de mens. "Je kunt het niet alleen aan machines overlaten." De leuze die Collexis hierbij hanteert is: "What you seek is what you get".

THESAURI EN FINGERPRINTS

Wat doet Collexis? Het vindt documenten, experts en organisaties voor je. Hierbij wordt gebruik gemaakt van zowel indexen en matching als van thesauri en teksten. Deze teksten kunnen zowel gestructureerd als ongestructureerd zijn. Een thesaurus is een verzameling van concepten die betrekking hebben op een specifiek

Collexis wil zich bezighouden met de onderliggende technologie, niet met het bouwen van applicaties

aandachtsgebied. Het is de conceptuele beschrijving van bijvoorbeeld alle documenten die betrekking hebben op een onderwerp. Daarnaast geeft het ook de relaties weer tussen de verschillende documenten. Inmiddels beschikt Collexis zelf over een aantal thesauri waarvan gebruik gemaakt kan worden, onder andere de Unified Medical Language System (UMLS), Agrovoc (landbouw), Wordnet (algemene thesaurus) en Thesaurus Gezondheidszorg 2000 (in de Nederlandse taal), maar ook Excerpta Informatica, Asis, ACM en Foldoc, die alle vier betrekking hebben op het ICT-domein.

Belangrijk onderdeel van Collexis is de *fingerprint*, zowel een concept als de weegfactoren die daarbij horen. Concepten zijn afkomstig uit thesauri. De weegfactoren geven aan hoe belangrijk het concept is voor de weergave van betekenis van de tekst. Volgens Van Praag zijn twee eigenschappen kenmerkend voor de fingerprints zoals Collexis die kent. "Ze zijn klein (gemiddeld 500 bytes) en ze hebben een uniek karakter."

Hoe werkt het proces? Van de beschikbare inhoud (content, bijvoorbeeld in de vorm van documenten of presentaties) worden fingerprints gemaakt, en ook van de zoekvraag die de gebruiker opgeeft. Het beste werkt dit als die zoekvraag zo groot en uitgebreid mogelijk is. Dit levert immers een 'rijkere' fingerprint op dan als slechts één zoekterm wordt ingegeven. Nadat van documenten en zoekvraag fingerprints zijn gemaakt vindt conceptmatching plaats.

Behalve fingerprints spelen thesauri een belangrijke rol. Een thesaurus bevat gevalideerde kennis. Thesauri kunnen worden aangekocht, maar ook beschikken veel organisaties over hun eigen

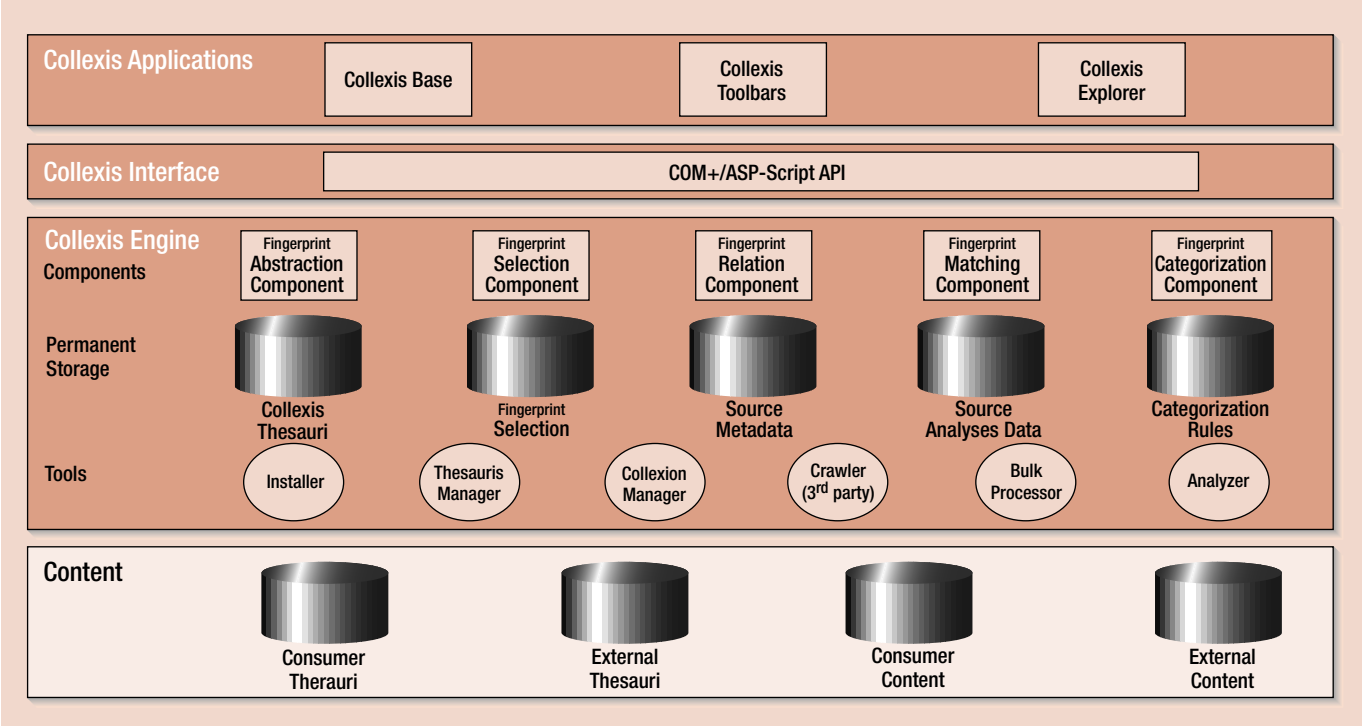
Het 'vector space'-model stelt je in staat bijvoorbeeld de tien best passende documenten op te vragen

thesaurus. Uiteraard kan men van meerdere thesauri gebruik maken. Onderdeel van een standaard project is dat Collexis samen met de klant bepaalt welke soort thesaurus nodig is. Partners kunnen helpen bij het vinden van een thesaurus binnen een verticale sector.

TECHNOLOGIE

Collexis Base is een standaard webapplicatie waarmee eindgebruikers kunnen zoeken in de verschillende collecties. Op basis van tekstuele input kunnen *search-fingerprints* worden gegenereerd. De gegenereerde fingerprints kunnen vervolgens ook weer worden gemanipuleerd. De resultaten van een zoekopdracht zijn documenten. Deze kunnen gegroepeerd worden. Documenten afkomstig van dezelfde auteur zijn desgewenst bij elkaar te vegen. Vervolgens kunnen aan dezelfde organisatie verbonden auteurs weer gegroepeerd worden. Expert en organisatie zijn dus verschillende niveaus waarop je de aanwezige documenten kunt aggregeren. Ook is het mogelijk een persoonlijk interesseprofiel te definiëren.

Met behulp van Collexis Toolbars kan worden gezocht in de



FIGUUR 3: COLLEXIS' ARCHITECTUUR.

verschillende collecties. De Collexis Toolbar is een plug-in-applicatie die onderdeel uitmaakt van Windows en Internet Explorer. Het is mogelijk de Toolbar per eindgebruiker anders in te stellen. Hiermee kan rekening worden gehouden met verschillen in toegangsautorisatie.

Van Praag benadrukt dat Collexis zich wenst bezig te houden met de onderliggende technologie en niet met het bouwen van applicaties. Er is voorzien in een API en derde partijen (partners) kunnen de gewenste toepassingen realiseren. Wel is Collexis bij elk traject betrokken. In de stap thesaurus-recept wordt gezocht naar de één of twee thesauri die het beste bij een specifieke klant passen.

Erik van Mulligen ziet het model van *vector space* als een van de unieke aspecten van de manier waarop Collexis werkt. Een search-fingerprint kan versimpeld worden voorgesteld als een bepaald punt in een driedimensionale ruimte (xyz-as). Vervolgens kan worden bepaald welke document-fingerprints hier het dichtste bij zitten. Hier zit de winst ten opzichte van een Booleaanse aanpak. Zet je bij een Booleaanse aanpak de zoekcriteria te ruim, dan vind je een hele berg resultaten (en dus niets). Zet je de zoekcriteria te smal, dan levert de actie geen resultaten op. Het vector space-model stelt je in staat bijvoorbeeld de tien best passende documenten op te vragen.

De engine is geschreven in C++. De huidige versie (3.5) is een gedistribueerde. Er wordt gebruik gemaakt van de Berkeley-database voor het opslaan van de titels, de document-identifier en andere metadata. Conventionele databases, zoals Oracle en MS SQL Server, leverden de titels niet snel genoeg terug. Berkeley-db wordt vaak gebruikt in echte realtime omgevingen.

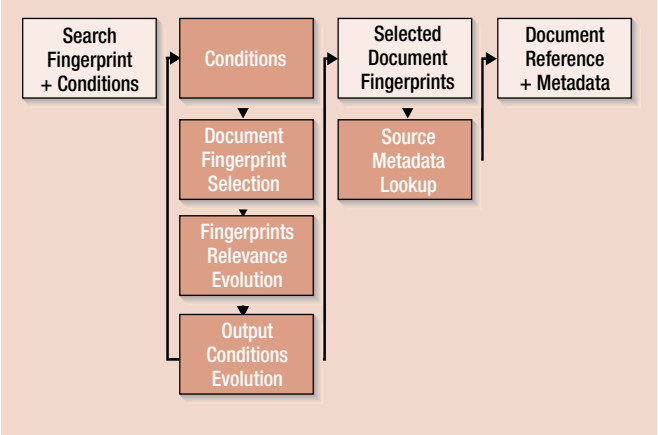
In wezen gaat het om een client/server-opstelling, maar de hele software kan ook op een laptop draaien. Collexis-software is beschikbaar voor Windows NT, 2000 en XP.

ARCHITECTUUR

De architectuur onder Collexis is drielaags. Collexis Applications werken via een interfacelaag op Collexis' engine. De applicatielaag is geconfigureerd en ontwikkeld op basis van de klantspecificaties. Applicatiecomponenten worden vaak toegepast in websites en op webpagina's. Door hiervan gebruik te maken kunnen gebruikers via hun intranet of extranet of via Internet zoeken in de beschikbaar gemaakte collecties. Organisaties kunnen zelf bepalen of zij willen dat de technologie alleen beschikbaar is via hun eigen

Doelgroep

Tot de huidige klantenkring van Collexis behoren onder andere Elsevier, de World Health Organization en Nature. Peter van Praag beschouwt wetenschappelijke uitgevers als eerste belangrijke klantengroep. Hier zijn de thesauri reeds beschikbaar. Een tweede doelgroep is de voedsel- en landbouworganistie van de Verenigde Naties, FAO. Eigenlijk zijn alle organisaties die bezig zijn met het handmatig indexeren potentieel klant. De petrochemie -denk aan patenten en juridische stukken- en onderzoeksafdelingen vormen daarnaast interessante klantgroepen. Veel medewerkers (meer dan honderd) en/of veel documenten zijn de kenmerken van organisaties die Collexis tot de doelgroep rekent.



FIGUUR 4: HET PROCES VAN SELECTIE EN MATCHING.

intranet of extranet danwel dat iedereen de zoekmogelijkheden via Internet kan benutten.

De engine layer of functielaag vormt het hart van Collexis. Hier vindt het abstraheren, matchen, selecteren, relateren en categoriseren van teksten plaats. Al deze activiteiten zijn vertaald in aparte componenten. Tools als Thesaurus Manager, Collexion Manager en Analyzer maken deel uit van de engine, en tevens Collexis

Net als Tarchon richt Collexis zich op de groeiende behoefte orde te scheppen in de informatiechaos

Thesauri, Fingerprint Collexions, Source Metadata, Source Analysis Data en Categorization Rules. Hierbij geldt overigens dat de Collexis Thesauri slechts de verwijzingen bevat naar de teksten. De teksten zelf maken geen onderdeel uit van de engine; deze blijven staan op de oorspronkelijke locaties.

De Collexis Interface-laag maakt het mogelijk via COM+ of ASP-scripts te connecteren met de Collexis-engine.

Collexis is taalonafhankelijk doordat gebruik wordt gemaakt

FIGUUR 5: COLLEXIS' TOOLBAR.

Prijsstelling en marketing

De prijsstelling van de software is gebaseerd op het aantal fingerprints, niet op het aantal gebruikers. Het starterspakket bestaat uit 2500 fingerprints, installatie van de software en training in het gebruik. Dit kost 18.000 euro. Een investering voor grote bedrijven beloopt enkele honderdduizenden dollars. Volgens Peter van Praag doen momenteel vijftien klanten betaald een *proof of principle*. Tien klanten zijn nu *live*. Inkomsten haalt Collexis uit het leveren van licenties (shipping) en onderhoud en support op de engine. Het behouden en uitbouwen van haar technologische voorsprong staat hoog in het vaandel. Daarnaast wil Collexis haar distributienetwerk verder opbouwen.

van nummers. Een vraag is te stellen in het Engels en kan documenten in het Frans, Duits of bijvoorbeeld Italiaans opleveren. Interne koppeling van search- en document-fingerprints aan nummers schakelt het taalafhankelijke element uit.

KENMERKEN

Net als Tarchon richt Collexis zich op de groeiende behoefte orde te scheppen in het heterogene aanbod aan informatie. Het is niet verrassend dat beide bedrijven een aantal kenmerken delen. In de eerste plaats de mogelijkheid gerichter en succesvoller te zoeken. Bij Collexis gebeurt dit met behulp van fingerprints. Maar ook het niet verplaatsen van data ten gunste van verwijzingen en relaties vormt een gemeenschappelijk kenmerk, alsmede de focus op technologie en de band met de wetenschap. Een verschil tussen beide is het gebruik dat Collexis maakt van thesauri. Deze bevatten gevalideerde kennis over specifieke kennisgebieden en vormen daarmee een belangrijk onderdeel van de werkwijze die Collexis voor ogen staat. Met 18.000 Euro voor een starterspakket vormt Collexis een prettige en laagdrempelige manier om in kennismanagement te stappen. ●

Drs. P.F.H. van der Linden (P.vanderLinden@Synergetics.nl) is principal consultant bij Synergetics, The Management Information Company.
Izabella Smits-Pukrop (I.Smits-Pukrop@Synergetics.nl) is senior consultant bij Synergetics.