Knowledge Discovery

De zoektocht naar verhulde en onthulde kennis

S.J.B.A. (Stijn) Hoppenbrouwers en H.A. (Erik) Proper
ID Research
Groningenweg 6
2803 PV Gouda
E.Proper@acm.org

"I don't know what I'm looking for, but I'll know when I find it"
P.D. Bruza [1]

Gepubliceerd als:

S.J.B.A. Hoppenbrouwers and H.A. Proper. Knowledge discovery - De zoektocht naar verhulde en onthulde kennis. *DB/Magazine*, 10(7):21--25, November 1999.

1 Inleiding

Menigeen zal bij het woord knowledge discovery wellicht denken aan technologieën zoals search engines, agent technologie, mining tools, meta-data standaarden, query talen/protocollen, etc. Naar onze mening bestaat knowledge discovery, net als kennismanagement, echter uit beduidend meer dan de onderliggende technologie. In dit artikel benaderen we knowledge discovery om deze redenen dan ook juist vanuit een conceptueel perspectief.

Het eerste doel van dit artikel is het verkrijgen van een beter begrip van knowledge discovery. *Wat is het precies en wat kan ik ermee?* Dit zullen we doen door referentie modellen op te stellen (in termen van een paradigma) die de essentiële mechanismen achter knowledge discovery op een conceptueel niveau weergeven. Deze referentiemodellen kunnen vervolgens tevens gebruikt worden om verschillende voor knowledge discovery relevante technologieën te positioneren: *Waar past wat?*
Op basis van de referentiemodellen en de link naar de onderliggende technologie, kan een applicatiearchitect vervolgens knowledge discovery toepassingen ontwerpen en inpassen in bestaande applicatiearchitecturen.

2 Wat willen we ontdekken?

Voordat we kunnen inzoomen op knowledge discovery moeten we eigenlijk eerst een beter gevoel hebben voor datgene wat we zouden willen 'ontdekken'. Met andere woorden, wat bedoelen we precies met 'kennis'. Nu zouden we hier een diepgaande, filosofische discussie kunnen gaan voeren over wat kennis precies is, maar dat zal ons niet veel verder helpen in onze begripsvorming rond knowledge discovery. We zullen het hier dus beknopt houden.

Voor we nader op het begrip 'kennis' ingaan eerst nog even dit: (expliciete) kennis is 'verpakt' in informatie, terwijl informatie op zichzelf weer wordt 'gedragen' door data (expressies in een symbolentaal). Kennis is een vrij lastig begrip; in onze ogen is de belangrijkste eigenschap van kennisdragende informatie dat deze 'sturend' is, met andere woorden beslissingen mogelijk maakt.

Door Nonaka & Takeuchi [2] wordt een onderscheid gemaakt tussen impliciete kennis en expliciete kennis. *Expliciete kennis* is de kennis die expliciet gemaakt is of expliciet gemaakt kan worden. Denk hierbij aan kennis die uitgedrukt is (of kan worden) in termen van feiten, regels, specificaties, of gewoon tekstuele beschrijvingen. Met *impliciete kennis* wordt bedoeld de kennis die impliciet in de hoofden van mensen aanwezig is. Men moet hierbij o.a. denken aan vaardigheden die men moeilijk expliciet kan maken. Bijvoorbeeld de *precieze* handelingen die een jongleur verricht om zes kegels te jongleren, of de manier waarop een top-manager in een bedrijf een strategische beslissing neemt. Impliciete kennis zit dus dicht aan tegen wat we doorgaans ervaren als intuïtie. Het onderscheid tussen impliciete en expliciete kennis is relevant in de context van knowledge discovery en ICT ondersteuning. Het ontdekken van impliciete kennis zal andere vormen van ICT ondersteuning vergen dan het ontdekken van expliciete kennis.

Naast het verschil tussen expliciete en impliciete kennis is er nog een ander relevant onderscheid te maken. Soms zal het zo zijn dat kennis in potentie aanwezig is, terwijl men zich van die kennis niet bewust is. Dit loopt uiteen van verborgen vaardigheden van medewerkers (verborgen voor het individu of voor de organisatie), via kennis die in documenten aanwezig is maar niet goed is geïndexeerd, tot aan kennis die verborgen zit in de nog onontdekte patronen in datacollecties (de basis voor datamining). Dit is het onderscheid tussen *onthulde* en *verhulde* kennis. In tegenstelling tot het onderscheid tussen impliciete en expliciete kennis betreft het hier dus de *status* van de kennis in tegenstelling tot het fundamentele *type* van de kennis.

Het onderscheid tussen *impliciet* en *expliciet* aan de ene kant en *onthuld* en *verhuld* aan de andere kan uitgezet worden in een 2x2 matrix (Figuur 1). De gemaakte onderscheiden bijten elkaar dus niet. We krijgen dan de volgende combinaties:

- **Impliciete, Verhulde kennis**: bijvoorbeeld competenties of expertise die bij een medewerker bestaat, maar waar de organisatie feitelijk geen weet van heeft.
- **Expliciete, Verhulde kennis:** bijvoorbeeld waardevolle inzichten die verborgen zitten in *beschikbare* datacollecties (naar boven te halen via data mining).
- **Impliciete, Onthulde kennis:** bijvoorbeeld bij de organisatie bekende expertise van een medewerker, waarop gericht een beroep gedaan kan worden.
- **Expliciete, Onthulde:** bijvoorbeeld best-practice documentatie, een draaiboek, knowledge bases, leerboeken, wetenschappelijke artikelen, etc.

	Impliciet	**Expliciet**
Verhuld	Onbekende competenties	Onbekende patronen en structuren in data
Onthuld	Bekende competenties	Gedocumenteerde kennis

Figuur 1: Vier Kennistypen

De vier soorten van kennis vragen om verschillende vormen van ICT ondersteuning, variërend van ondersteuning die is gericht op "expertise analyse", of pro-aktieve kennisdeling en Human Resource-gerelateerde informatie- en planningssystemen, tot datamining en standaard vormen van documentatiebeheer. In brede zin kan ICT natuurlijk in alle genoemde gevallen ter verdere ondersteuning ingezet worden. Het is echter wel zo dat ICT van nature beter is toe te passen op expliciete dan op impliciete kennis.

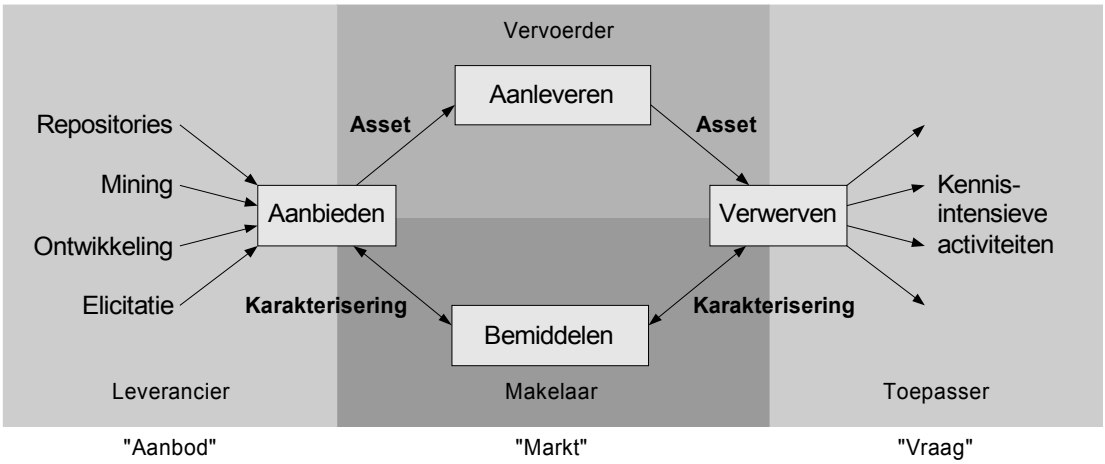
Nu we hebben besproken wat voor kennis er te 'discoveren' valt, kunnen we ons richten op het eigenlijke "ontdekken" hiervan. We ontwikkelen daarvoor een kader waarin de relatie met zowel architectuur als concrete ICT wordt gelegd.

3 Kennismarkt paradigma

We hebben gezien dat er een aantal soorten kennis te onderscheiden valt. In de praktijk is daarbij vooral het verschil tussen *verhulde* en *onthulde* kennis van belang: onthulde kennis is te lokaliseren (zelfs als deze nog impliciet is), verhulde niet (zelfs al is hij ooit ergens expliciet gemaakt). Maar wat valt hier nu concreet mee te doen in een kennismanagement of een knowledge discovery context?

Om meer inzicht te verkrijgen in de verschillende partijen die hierbij een rol spelen, en de taken en belangen van deze partijen, hanteren we het "Kennismarkt Paradigma" (Figuur 2). Zoals in elk marktmodel is er een *aanbieder* (de kennis-leverancier) en een *vrager* (de kennis-toepasser), met daartussen eventueel een *makelaar* en altijd een vorm van *vervoerder*. De "koopwaar" binnen het paradigma bestaat uit de kennis-*assets*. Dit zijn handelbare vormen van onthulde kennis, die fysiek getransporteerd worden door de vervoerder. Het *handelbaar* zijn van de assets betekent niet dat het perse moet gaan om expliciete kennis! De impliciete kennis die in het hoofd van een mens aanwezig is, is ook *handelbaar* omdat de persoon met de relevante kennis deze kennis mee kan nemen naar een bepaalde situatie waar deze kennis nodig is. Is dit laatste bijvoorbeeld niet de kern van het werk van een (externe) consultant?

Merk op dat de kennismarkt ook heel goed de algehele kennishuishouding in een organisatie kan weergeven. Er hoeft in het kader van dit paradigma natuurlijk niet noodzakelijkerwijze met geld geschoven te worden tussen de vragende en de aanbiedende partij. Wel roept dit interessante vragen op: hoeveel is kennis eigenlijk waard, en wat is eventueel een bruikbaar betaalmiddel?



Figuur 2: Het Kennismarkt Paradigma

De leverancier wil graag kennis leveren, maar zal daarvoor eerst een afnemer moeten vinden die deze kennis wil toepassen. Dat kan alleen als zij duidelijk kan maken wat er zoal aangeboden wordt, en dus is het van vitaal belang dat de kennis correct wordt beschreven of *gekarakteriseerd*. Dat valt lang niet altijd mee, want kennis is bepaald niet zo "grijpbaar" als een mud steenkool. Terminologieproblemen kunnen hier bijvoorbeeld danig roet in het eten gooien.

Aan de leverancierskant van de kennismarkt kan geput worden uit verschillende bronnen: repositories, mining op data collecties, zelfs het ondervragen van specialisten (elicitatie) of actieve kennisontwikkeling. Maar in ieder geval zal een betrouwbare aanbieder meestal zelf weten wat zij aanbiedt: het gaat hier om *onthulde kennis*, die dus te 'lokaliseren' is, en tevens beschikbaar. Het is heel goed mogelijk om impliciete kennis aan te bieden, bijv. in de vorm van een betrouwbare expert, zolang maar zeker is dat de impliciete kennis afdoende door de toepasser kan worden aangewend (al was het maar in de vorm van een bepaalde *competentie*).

De toepasser is op zoek naar kennis, maar weet niet of die ook te vinden zal zijn. Wat vervelender is: een kennisafnemer weet vaak helemaal niet zo precies waar zij nu eigenlijk op uit is. De kennis is voor de vrager nog verhuld; de vraag is of dat voor de aanbieder ook zo is. Ook hier is karakterisering doorslaggevend, omdat deze de matching "voedt". Dit is feitelijk de meest prominente vorm van Knowledge Discovery: *wat is het precies dat ik zoek?* Dus het ontdekken van de feitelijke behoefte. De makelaar speelt uiteraard een belangrijke rol bij het tot elkaar brengen van vraag en aanbod. Uiteindelijk is Knowledge Discovery dus een proces dat zich overal voordoet waar toepassers van kennis op voorhand niet precies weten welke kennis zij nodig hebben voor hun werkzaamheden.

Het Kennismarkt Paradigma komt natuurlijk pas echt tot leven zodra we het toepassen op een alledaagse situatie. Stel u heeft de behoefte aan kennis op het gebied van hogesnelheidstreinen. U gaat met deze behoefte naar een bibliotheek alwaar u de bibliothecaris uw behoefte voorlegt. Deze zal u misschien nog wat additionele vragen stellen, waarna de bibliothecaris een (geautomatiseerde) index raadpleegt. Daarna zal de bibliothecaris u naar de juiste boeken over treinen (en hoge snelheidstreinen in het bijzonder) verwijzen. Hoe is deze situatie gerelateerd aan het kennismarkt paradigma? U bent de kennistoepasser. De bibliothecaris is de kennismakelaar. De schrijvers van de boeken over treinen zijn de kennisleveranciers, terwijl de geschreven boeken zelf de assets zijn. De uitgevers en de boekenhandelaars vormen tenslotte tezamen de "vervoerders(keten)" van de onderhavige kennis.

Boeken zijn natuurlijk een voorbeeld van expliciet gemaakte kennis. Als een voorbeeld van een kennismarkt rond impliciete kennis kunnen we kijken naar een typische situatie die we in veel kantoren kunnen aantreffen. Op elk kantoor is er wel een persoon te vinden die een redelijk goed overzicht heeft van de expertises van de verschillende medewerkers. Deze persoon kunnen we feitelijk ook zien als een kennismakelaar. Stel u heeft een kennisbehoefte om meer te weten te komen over verschillende vormen

van middleware. Met zo'n vraag kunt u nu eerst naar de lokale kennismakelaar. Deze zal u doorverwijzen naar een collega die meer weet over middle-ware. Die collega wordt dan de leverancier van de kennis. En wat is dan de vervoerder? Mocht u een mondeling gesprek aanknopen met de middle-ware expert dan zal de lucht die het geluid van uw woorden overbrengt de vervoerder zijn van de kennis. Als u er voor kiest e-mail te gebruiken om uw vraag te stellen, dan is het e-mail systeem de vervoerder van de kennis.

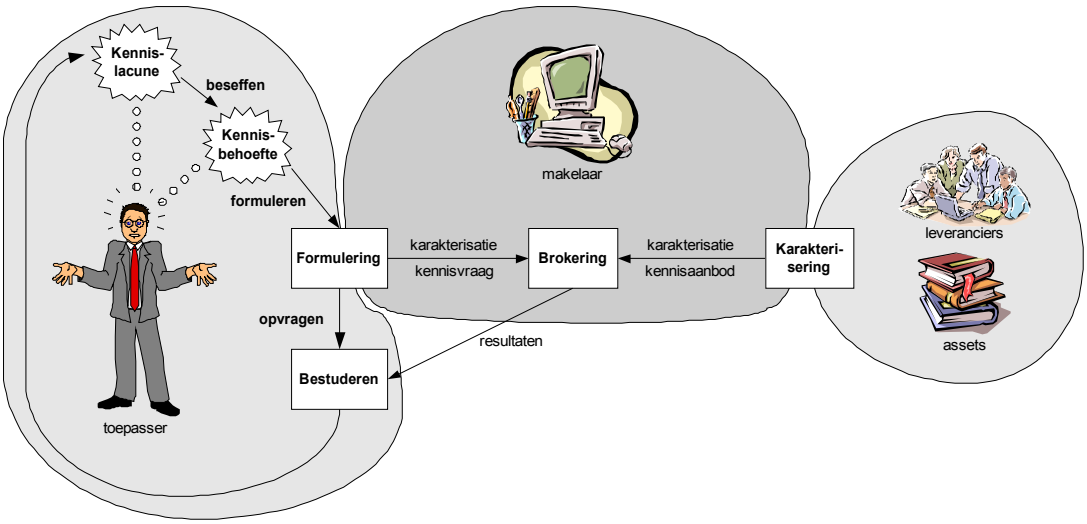
In de volgende twee paragrafen zoomen we in op, respectievelijk, het ontdekken van onthulde en het ontdekken van verhulde kennis. Hierbij maken we ook de koppeling naar de verschillende onderliggende technologieën concreter.

4 Onthulde kennis ontdekken

Knowledge Discovery (van onthulde kennis) is complementair aan een wellicht bekendere tak van sport: *Information Retrieval*. Als alle partijen precies weten welke kennis gevraagd en aangeboden wordt, is het *matchen* en overbrengen (vervoeren) van de data die deze kennis "draagt" een controleerbare, technische kwestie. Hier positioneren wij information retrieval als een ophaalproces waarbij wellicht flink (machinaal of handmatig) gezocht moet worden, maar waarbij het altijd om *onthulde* kennis gaat die wordt vertegenwoordigd door een *expliciete* karakterisering. Als er geen match is, en het gebruikte information retrieval mechanisme zit voldoende degelijk in elkaar, dan is er zekerheid dat de opgevraagde data (de drager van de gezochte kennis) niet beschikbaar is. Hierbij doemt meteen de volgende vraag op: hebben we wel naar het juiste gezocht, met andere woorden, was de query wel een juiste karakterisering van de kennisbehoefte?

Bovenstaand verhaal wordt inzichtelijker gemaakt in termen van het "Knowledge Discovery Paradigma" in Figuur 3. Dit paradigma is gebaseerd op de Information Retrieval en Knowledge Retrieval paradigma's zoals deze zijn te vinden in [1, 3]. Dit paradigma kan gezien worden als een specialisatie van het kennismarkt paradigma, welke is gericht op het ontdekken van onthulde kennis.

Het in Figuur 3 geschetste traject heeft een lacune in de kennis van een toepasser als startpunt, of liever gezegd: het punt waarop iemand zich realiseert dat er zo'n lacune bestaat en dat er een zekere noodzaak bestaat om die op te vullen. Dit leidt tot het ontstaan van een kennisbehoefte. Bij het lenigen van deze kennisbehoefte komen we bij een cruciale stap: het karakteriseren van de lacune. Dit bestaat uit het *formuleren* van een beschrijving van de kennisbehoefte in termen van een vraag (die de beschrijving bevat maar aan niemand gericht is), of query (die de vraag naar een machine communiceert). Laten we deze vormen in het algemeen omschrijven als de *kennisvraag*. In het ideale geval wordt een kennisvraag in een samenspel tussen de zoekende en de makelaar geformuleerd. In een bibliothecaire context zou dit vergelijkbaar zijn met het gesprek dat zich ontspint tussen iemand die een boek over een bepaald onderwerp zoekt en een bibliothecaris. Door middel van dit samenspel kan de makelaar de zoeker ondersteunen bij het formuleren van de precieze kennisvraag; de cruciale stap in knowledge discovery (van onthulde kennis).



Figuur 3: Het Knowledge Discovery Paradigma

Tegenover de toepasser van de kennis staan de kennisleveranciers en de door hun leverbare kennis assets. Deze assets worden idealiter in een samenwerking tussen de leveranciers en de makelaar beschreven in termen van één of andere karakterisatie. De makelaar kan er hierbij voor zorgen dat dit gebeurt in termen van karakterisaties zoals die ook door de potentiële toepassers van deze kennis gebruikt (zouden) worden in de formulering van hun kennisvraag.
De brokering activiteiten die door de makelaar worden uitgevoerd zijn feitelijk niets anders dan het uitvoeren van een information retrieval proces.

Wat bij knowledge discovery heel belangrijk is, is dat de makelaar een faciliteit aanbiedt die in de toepasser *assisteert* bij het formuleren van de juiste kennisvraag. Dit kan bijvoorbeeld door het inzichtelijk maken van het kennisaanbod (catalogus; index) en het afstemmen van karakterisaties met behulp van synoniemenlijsten of ontologieën. Inzicht in vraag en aanbod (en de manier waarop deze door verschillende partijen verwoord worden) is daarbij van vitaal belang. Wanneer een toepasser van tevoren één of meer mogelijke kennisvragen bundelt tot een *profiel*, wordt het mogelijk voor de makelaar om de toepasser ervan op de hoogte te houden wanneer nieuwe kennis beschikbaar komt. Dit laatste kan men classificeren als *Knowledge Filtering*, of *Information Filtering.*

De makelaar heeft ook de mogelijkheid om, ten behoeve van de effectiviteit en de efficiency, te anticiperen op potentiële kennisvragen. De meest voor de hand liggende vorm van anticipatie is een indexering van de assets op basis van hun karakterisaties. We onderscheiden daarbij een *technisch geoptimaliseerde* index en een *inhoudelijk geoptimaliseerde* index. Technische optimalisatie zal meestal gericht zijn op performance; dat kan bijvoorbeeld inhouden dat een tijdrovend sorteerproces al wordt gedraaid voor een daadwerkelijke query daarvan gebruik maakt. Inhoudelijke optimalisatie echter is meer gericht op de karakterisering: indien de index bijvoorbeeld een kleine set doeltreffend gekozen zoektermen bevat dat aan een asset hangt, vergemakkelijkt dat het brokering proces enorm. De verantwoordelijkheid voor indexering van beide typen kan gespreid zijn. Technische optimalisatie wordt bij voorkeur door het onderliggende DBMS verzorgd, terwijl inhoudelijk doeltreffende indexering vaak typisch mensenwerk is, waarbij in feite een poging gedaan wordt op kennisvragen vooruit te lopen.

Aan de drie hoofdprocessen uit Figuur 3 (formulering, brokering, en karakterisering) zijn een aantal technologische elementen te koppelen. Ter illustratie volgen hier een aantal voorbeelden:

1. *Search engines*. De meeste webgebaseerde search engines richten zich op het karakteriseren van assets en het brokering proces. De karakterisering van de assets wordt meestal door de search engine zelf uitgevoerd door een keyword gebaseerd indexeringsmechanisme.
   De laatste tijd kunnen we ook zien hoe search engines proberen de zoekenden (de toepassers) te helpen bij het formuleren van hun vraagstelling.
2. *Documentaire informatiesystemen*. Zoals de naam suggereert, richten deze systemen zich in eerste instantie vooral op de kennis die is vervat in documenten. De meeste documentaire informatiesystemen bieden faciliteiten om de opgeslagen documenten te indexeren (soms automatisch, soms handmatig), en op basis hiervan documenten terug te zoeken. Het ondersteunen van gebruikers in hun zoektocht naar relevante kennis wordt alleen in de meest geavanceerde systemen (deels) ondersteund.
3. *Agent technologie*. Deze technologie kan op diverse plekken in de context van het Knowledge Discovery Paradigma ingezet kan worden. Met behulp van agent technologie zou het bijvoorbeeld mogelijk zijn om de toepasser, de makelaar, en de leverancier te laten vertegenwoordigen door agents. In zo'n scenario zou toepasser-agent over het kennisvraag-profiel van de toepasser moeten kunnen beschikken. De toepasser-agent kan vervolgens namens de toepasser 'in gesprek treden' met de makelaar-agent en de resultaten op een gerichte wijze terugkoppelen naar de toepasser. De leverancier-agent kan op eenzelfde manier in contact treden met de makelaar-agent om de aangeboden kennis op een effectieve wijze te karakteriseren.

5 Verhulde kennis ontdekken

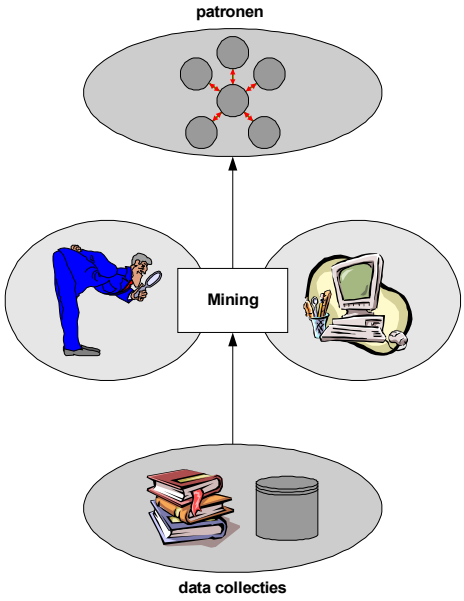
De meeste ICT support voor Knowledge Discovery die is gericht op het ontdekken verhulde kennis legt zich toe op expliciete verhulde kennis. Denk hier aan de diverse vormen van mining, zoals data mining en text mining. In bepaalde kringen wordt Knowledge Discovery overigens zelfs volledig gelijkgesteld aan mining. De twee bekendste soorten van mining zijn (numerieke) data mining en text mining. Bij de eerste wordt gezocht naar patronen (trends, verbanden, etc.) in numerieke data, terwijl de tweede vorm van mining zich richt op patronen zoals die in elektronische teksten zijn te vinden.

Conceptueel gezien kunnen we mining zien als een interactieve (mensgedreven en machineondersteunde) zoektocht naar patronen in verschillende vormen van data. Dit is geïllustreerd in termen van het Mining Paradigma in Figuur 4. Potentieel kan het hierbij gaan om de volgende vormen van data:

1. numeriek gegevens,
2. relationele gegevens,
3. teksten,
4. beelden,
5. audio bestanden,
6. video bestanden.

Traditioneel wordt vooral de mining van numerieke gegevens door ICT ondersteund. Inmiddels bestaan er ook mining systemen die zich toeleggen op relationele gegevens en teksten.

Merk op dat een eenmaal gevonden patroon meteen een onthulde vorm van expliciete kennis wordt (die bijvoorbeeld op elk gewenst moment door een OLAP systeem geëvalueerd kan worden). Knowledge discovery kan dan dus in het Kennismarkt Paradigma uit Figuur 2 gepositioneerd worden aan de kennis-leverancierskant. De door mining gevonden patronen zijn zelf te beschouwen als kennis assets, waar bijvoorbeeld ten behoeve van managementbeslissingen dankbaar gebruik van gemaakt kan worden.



Figuur 4: Mining Paradigma.

In dit artikel gaan we niet nader in op het ontdekken van verhulde impliciete kennis. Momenteel zijn er weliswaar enige vormen van ICT ondersteuning beschikbaar hiervoor, maar dit is vooralsnog een diffuus terrein.

6 Architecturele context

Knowledge discovery, en met name het ontdekken van onthulde kennis, komt binnen een organisatie pas echt tot zijn recht wanneer het wordt geïntegreerd met de bestaande architectuur van de informatievoorziening.

Er zijn diverse manieren om knowledge discovery te integreren met een (bestaande) architectuur voor informatievoorziening. Knowledge discovery is vooral nuttig als ondersteuning van kennisintensieve processen. Als voorbeeld beschouwen we hier kort de mogelijke integratie van knowledge discovery en (kennisintensieve) workflows. Kijkende naar een workflow kunnen we daarin diverse taken terugvinden die door medewerkers uitgevoerd moeten worden. Bepaalde taken kunnen wellicht als kennisintensief gekarakteriseerd worden. In dit geval is de kans groot dat de medewerker die deze taak gaat uitvoeren diverse kennis assets nodig heeft bij het uitvoeren van de taak.

Het kan daarom zinvol zijn om bij het ontwerpen van de totale workflow al van tevoren na te denken over het soort kennis welke door de uitvoerder van die taak nodig is. Dit levert per kennisintensieve taak een initieel kennisprofiel op. Bij het daadwerkelijk uitvoeren van de kennisintensieve taak kan de makelaar uit een knowledge discovery systeem op basis van deze beschrijvingen alvast de benodigde

kennis assets vergaren. Wanneer de uitvoerder meer kennis nodig heeft kan hij/zij deze behoefte eveneens kenbaar maken aan de makelaar door een extra kennisvraag te formuleren. Deze kennisvraag kan vervolgens desgewenst verwerkt worden in het kennisprofiel dat bij deze taak behoort.

7 Afronding

In dit artikel hebben we een geprobeerd een korte beschrijving te geven op conceptueel niveau van wat knowledge discovery is. Dit hebben we gedaan aan de hand van een aantal referentiemodellen. Vanuit deze modellen hebben we ook (kort) een link gelegd naar de onderliggende technologie. Tot slot zijn we nog even kort ingegaan op de positie van knowledge discovery binnen een informatiearchitectuur.

8 Verwijzingen

1. Bruza, P.D., *Stratified Information Disclosure: A Synthesis between Information Retrieval and Hypermedia*. 1993, Nijmegen, The Netherlands: University of Nijmegen.
2. Nonaka, I. and N. Takeuchi, *The Knowledge-Creating Company.* 1995, New York, New York: Oxford University Press. 97-130.
3. Proper, H.A. and P.D. Bruza, *What is Information Discovery About?* Journal of the American Society for Information Science, 1999. **50**(9): p. 737-750.

Over de auteurs:

- S.J.B.A. (Stijn) Hoppenbrouwers is Research Consultant bij ID Research. Zijn expertises zijn met name conceptualisatieprocedures, terminologiebeheer, ontologieën en kennis management.
- H.A. (Erik) Proper is een Senior Research Consultant bij ID Research. Zijn expertises zijn met name informatie architectuur, kennis architectuur, conceptueel modelleren en information retrieval.