Ondanks verschillen genoeg mogelijkheden voor integratie

OLAP en datamining: niet apart maar samen

Jaap Verhees en Rob Peters

Hoe kan men snel en efficiënt de interessantste combinaties van meetwaarden en dimensies uit een OLAP-kubus halen? Hoe kan men de meest relevante dimensieniveaus voor een OLAP-omgeving bepalen? Wat te doen wanneer resultaten van datamining snel en flexibel moeten worden geverifieerd? Degelijke vragen kunnen worden beantwoord door integratie van OLAP en datamining. In het artikel "Mine your own business" in DB/M nummer 3/2003 is al kort een praktijkvoorbeeld uit de marketing aangestipt, waarin voor bepaling van producten en meest winstgevende doelsegmenten, associatieregels worden gevonden door middel van toepassing van datamining-techniek op OLAP-data. In dit artikel wordt gesteld dat datamining-resultaten verder kunnen worden benut in de OLAP-omgeving maar ook omgekeerd, dat OLAP-resultaten van nut zijn in de datamining-omgeving.

Datamining en OLAP zijn technologieën die beiden worden gebruikt binnen de Business Intelligence-discipline. Tot nu toe worden beide productfamilies gescheiden ingezet, omdat gebruikers tot verschillende groepen behoren. Het onderscheid tussen deze groepen is gebaseerd op afwijkende specialismen.

OLAP nu

Een OLAP-kubus is een multidimensionale kijk op bedrijfsdata. Gangbare dimensies zijn nog steeds Tijd, Product, Markt, Geografie, Verkooporganisatie, Scenario (plan tegenover realisatie) en een meetwaarde-dimensie met meetwaarden zoals Omzet, Kosten van Verkochte goederen, Winst, of ratio's zoals Winstpercentage en ROI. Elke dimensie kan een hiërarchische structuur bevatten. Zo kan de dimensie Tijd opgedeeld zijn in jaren, de jaren in kwartalen, de kwartalen in maanden, enzovoort. Een OLAP-tool maakt een snelle aggregatie van meetwaarden langs de dimensie hiërarchieën mogelijk. Tevens laat deze tool toe dat subkubussen worden geselecteerd via simpele navigatie-operaties zoals *slice and dice*, *pivot* en *drill*.

In de praktijk bestaat de lay-out van een OLAP-kubus uit een hiërarchie die de bedrijfsanalisten al een paar jaar gebruiken in hun rapportages. Zo is het bijvoorbeeld in het bankwezen nog steeds gangbaar dat de dimensie Klant alleen geaggregeerd kan worden langs de klantsegmenten 'particulieren', 'ondernemers', en 'publieke overheid'. Andere attributen van de datawarehouse-dimensie Klant worden dan niet opgenomen in de OLAP-kubus. Bijvoorbeeld, in het verzekeringswezen heeft de datawarehouse-dimensie Klant attributen zoals leeftijd, huwelijksstatus, aantal kinderen, woonwerkafstand, aantal jaren klant. Dergelijke attributen kunnen gerepresenteerd worden of als een additionele dimensie of als een attribuut-dimensie. Het laatste betekent dat een attribuut simpelweg als een label wordt toegevoegd aan de basis-dimensie.

Datamining nu

Tegenwoordig is datamining niet alleen dusdanig data-gestuurd dat het voorheen onbekende of onverwachte patronen en regels vindt in de data in een datawarehouse of een specifieke datamine. Datamining-technologie is nu ook in staat om operationele datasets continu te monitoren en in 'near real-time' een nieuwe observatie of record te scoren (het toepassen van bestaande datamining-modellen zoals patronen of beslisregels of associatieregels op nieuwe data).
Het gaat zelfs nog verder dan 'near real-time scoring' want ook de datamining-modellen zelf kunnen tegenwoordig geautomatiseerd worden herberekend, op basis van de meest recente observaties. Dit is mogelijk dankzij de inzet van datamining-operaties als database extenders, zoals wordt gedaan door de grootste database-leveranciers (IBM, Microsoft, Oracle).
De belangrijkste datamining-operaties die momenteel worden ingezet zijn:
Classificatie, gebruikt historische data om een model te bouwen van een bepaald concept, waarna dat model later wordt gebruikt

om te voorspellen of een nieuwe observatie voldoende overeenkomt met het concept. Bijvoorbeeld, we kunnen een model bouwen voor minder loyale klanten en dan later voorspellen of bepaalde nieuwe klanten van het type minder loyale klant 'dreigen' te worden;
Database segmentatie of Clustering, verdeelt een database in segmenten die gelijksoortige records bevatten. Dit gebeurt op basis van een aantal gekozen attributen en hun waarden. Bijvoorbeeld, 'moeders met twee of meer kinderen, met een professie en woonachtig in de Randstad' kunnen een interessant segment zijn voor een bank met spaarrekeningen, maar ook bijvoorbeeld voor een fast food-keten of organisatie met vakantiehuisjes;
Associatie, vindt relaties tussen producten, diensten of andere items die klanten neigen – tegelijkertijd of in de loop van de tijd in een bepaald tijdsvak – te kopen. Supermarkten gebruiken een variant van deze operatie, 'market basket analysis', om producten gezamenlijk te tonen in een schap. Boekwinkels op internet, zoals Amazon, gebruiken weer een andere variant van deze operatie en betitelen de resultaten als 'recommendations'.
De belangrijkste leveranciers van datamining-tools zijn op dit moment SAS, IBM, en SPSS met Microsoft en Oracle als meest nabije volgers.

Verschillen

Ter verduidelijking staan in de tabel in afbeelding 1 de belangrijkste verschillen tussen OLAP en datamining opgesomd.
Het eerste item somt het belangrijkste verschil op: OLAP wordt gestuurd door de eindgebruiker; de analist genereert een hypothese voor de bedrijfsproblematiek en gebruikt dan het OLAP-tool om deze hypothese te toetsen. Haaks hierop staat dat in datamining het tool wordt ingezet om uit de beschikbare data een hypothese te genereren. Anders gezegd, bij OLAP-analyse sturen de gebruikers de exploratie terwijl bij datamining de tools de exploratie uitvoeren.

In het geval van OLAP-analyse zullen over het algemeen de data geaggregeerd zijn. Dit is vooral ingegeven doordat de meeste kubus lay-outs zijn gebaseerd op periodieke rapportages, die al eerder bestonden binnen het bedrijf. De preaggregatie is er vooral om de snelheid te bereiken die verwacht wordt vanwege de term Online in OLAP. Datamining veronderstelt juist dat de data op het meest granulaire niveau voorhanden zijn en dat deze (operationele) data worden ontgonnen in de mining-stap van het proces.

Bij OLAP gaat men ervan uit dat de kennis die is opgebouwd in het bedrijf over klanten, producten, regio's, enzovoort, zichtbaar zal zijn in de structuur van de kubussen. Immers, deze komen ook terug in de bestaande periodieke rapportages tussen afdelingen (marketing, financiën, inkoop, verkoop). Hiërarchieën in OLAP-dimensies duiden dan ook regelmatig op bedrijfsregels.

> Ondanks de verschillen is er reden om OLAP en datamining te integreren

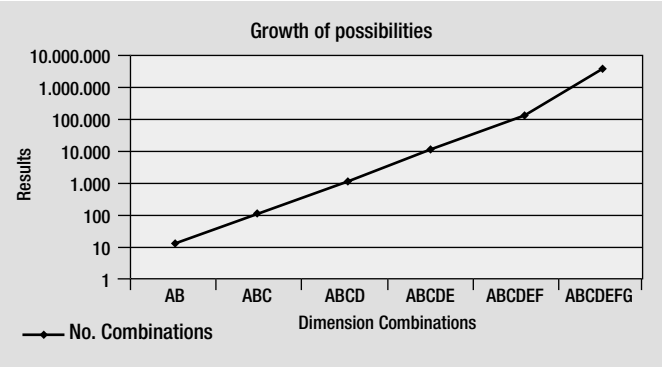
Datamining-technologie daarentegen is beperkt in de opname, weergave en het benutten van bestaande bedrijfsregels. De werking van de bedrijfsregels zit versluierd in de databasestructuur zelf (door middel van foreign keys en constraints), en in de records, datatypen en toegestane waarden van de data zelf.

OLAP stelt de gebruiker in staat eenvoudig data te selecteren en in en uit te zoomen op die data. Daarbij worden de data iedere keer geaggregeerd voor andere combinaties van dimensies en op een ander niveau. Daarentegen bevat datamining een groot scala aan technieken waaronder neurale netwerken, beslisbomen, genetische algoritmen, nearest neigbour, en radiaal basis-functies. En daarnaast kent datamining inmiddels verschillende disciplines gericht op specifieke applicaties. Voorbeelden zijn multimedia-mining, tekst-mining, web-mining.
De OLAP-aanpak maakt tegenwoordig gebruik van het RDBMS, maar zo heeft de variant MOLAP nog steeds een specifieke multidimensionale databasestructuur als argument. De eigenlijke meetwaarden zijn natuurlijk numerieke waarden, die geaggregeerd moeten kunnen worden via slicing en dicing, en categoriale waarden die geordend kunnen worden in geval van navigatie.
De bestaande datamining-technologie werkt over het algemeen op

OLAP	Datamining
'Hypothesis driven' en 'User driven' '	Data driven' en 'Tool driven'
90% geaggregeerd, 10% detailniveau data	90% operationeel daniveau, 10% geaggregeerde en getransformeerde data
Expliciete bedrijfsregels	Impliciete bedrijfsregels
Verschillende operaties en varianten	Verschillende technieken en disciplines
MDDBMS of RBDMS	Flat files of RDBMS

Afbeelding 1: Tabel verschillen tussen OLAP en datamining.

Afbeelding 2: OLAP versus datamining wat betreft kennisopbouw.



Afbeelding 3: Toename van het aantal combinatiemogelijkheden bij toename van het aantal dimensies van twee naar zeven.

een RDBMS waarin de diverse typen data als bekende databaseveldtypen of als Large OBjects (LOB) zijn opgeslagen. Zo is een klant-attribuut als leeftijd, opgeslagen als een veldtype Integer. Maar een AVI-bestand, met een klacht van de klant opgenomen tijdens interactie met het callcenter, kan evenwel dienen voor (multimedia) datamining en is dan opgeslagen als een BLOB (Binary Large OBject), als input voor de datamining-techniek.

Ondanks deze verschillen is er reden om OLAP en datamining te integreren omdat beide technologieën elkaar kunnen complementeren in bedrijfsanalyse. Juist het eerste punt van verschil, het deductieve karakter van OLAP en het inductieve karakter van datamining, is het punt waarop ze elkaar kunnen aanvullen.

Integratie

De integratie van OLAP en datamining heeft twee kanten. Allereerst hebben resultaten van datamining een waarde voor de OLAP-analist. Zo kan datamining de analyse van een OLAP-kubus sturen. Het aantal dimensies en het aantal hiërarchische niveaus per dimensie bepalen het aantal combinatiemogelijkheden. Voor ieder van deze combinaties kunnen de meetwaarden geaggregeerd worden. Zo kunnen drie dimensies met twee, vier en vijf niveaus al 78 combinaties opleveren (2x4 + 2x5 + 4x5 + 2x4x5). De meeste OLAP-kubussen hebben echter meer dimensies. Het is onmogelijk om manueel tijdens iedere analyse alle mogelijk interessante combinaties te onderzoeken (zie afbeelding 3). Bijvoorbeeld, welke combinatie van regio, seizoen, verkoper, producttype en kostprijs heeft de grootste invloed op de marge? Menig gebruiker zal daarom spoedig vervallen in het analyseren van een aantal standaard combinaties die in het verleden interessant bleken, bijvoorbeeld regio en seizoen. Echter, bedrijf en markt veranderen en analyses uit het verleden kunnen hun relevantie verliezen.
Dan is de OLAP-analist gebaat bij een tool die snel en efficiënt de interessante combinaties toont. Dit kan worden gedaan door datamining met de beslisboom-techniek. Met behulp van deze techniek wordt bijvoorbeeld getoond dat de hoogste marges worden behaald in regio noord in het derde kwartaal en in regio zuid door verkoper twee. Dergelijke resultaten kunnen dan worden geanalyseerd in de OLAP-kubus. Door regelmatig deze datamining-analyse uit te voeren kan de OLAP-analyse worden gestuurd.
Ook bij het definiëren van dimensies kan datamining OLAP ondersteunen. Het definiëren van een hiërarchie van klanten is makkelijk te doen in een OLAP-tool, bijvoorbeeld op basis van geografische informatie. Zo is het gebruik van marktregio's in een OLAP-kubus gangbaar. Echter, een hiërarchie die makkelijk is te definiëren, zoals een geografische hiërarchie, hoeft daarmee niet automatisch de meest waardevolle informatie over de bedrijfsvoering te geven.

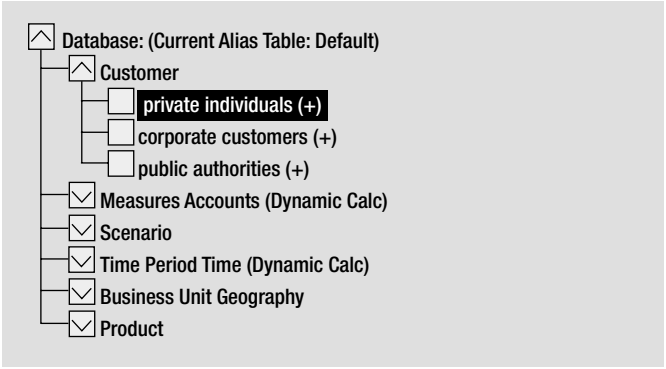
Datamining daarentegen kan een klantsegmentatie opleveren die meer informatie over de klanten meeneemt, zoals gezinsstatus, geschat inkomen, omvang van de woonplaats, en andere demografische data. Zulke segmenten, of zogenaamde clusters, kunnen dan gebruikt worden om een klantdimensie te definiëren in OLAP. Een begripsvolle korte omschrijving van elke afzonderlijke cluster kan worden toegevoegd als een element binnen de hiërarchiestructuur voor de dimensie klant. Dergelijke datamining-resultaten zullen daarom toegevoegd worden aan de OLAP-kubus. De bedrijfsanalist kan zo met de vertrouwde OLAP-analyse tools doorwerken. Waarom zouden we de investering in het gemak van het gebruik van OLAP-tools teniet doen? Dat zou een desinvestering betekenen. Juist deze vorm van integratie maakt ook de investeringen in datamining meer waard voor een grotere gebruikersgroep.

Effectieve verkenning

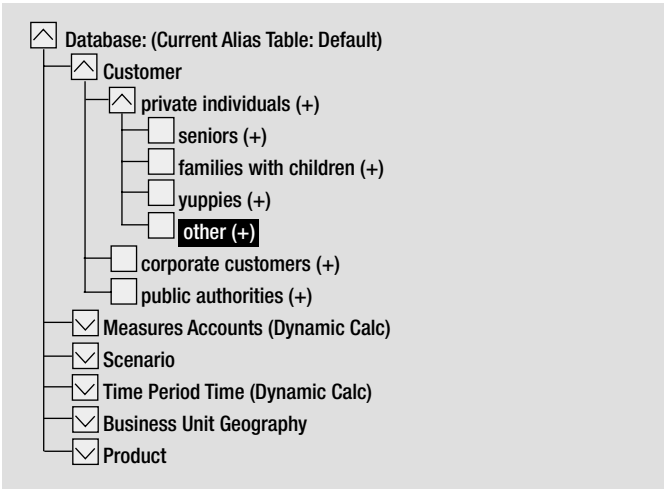
Waarschijnlijk zullen we voor het eigenlijke datamining-proces (definieer bedrijfsdoelstelling, selecteer data, prepareer data, mine the data, inzet van resultaat in bedrijf, actie) nog steeds een datamining-expert inzetten. We bereiken met integratie echter dat de bedrijfsanalist deze datamining-resultaten makkelijker vertaalt naar inzet en acties, omdat ze meer inzichtelijk zijn gemaakt binnen de OLAP-tool. Resultaten worden gepresenteerd in bedrijfsterminologie (zoals klantsegmenten) in plaats van in statistisch en mathematisch jargon.

> OLAP en datamining zijn niet langer verzamelingen stand-alone technieken

Door het creëren van de mogelijkheden om datamining-visualisaties op te starten vanuit de eindgebruikersinterface van de OLAP front-end tools zal de OLAP-analist nieuwe toelichtingen en verklaringen krijgen om de bedrijfsdata beter te begrijpen.
Datamining heeft in traditionele zin een enkele dimensie in het datawarehouse gebruikt naast vele attributen over deze dimensie. Door nu kruis- of multidimensionale analyse van deze



Afbeelding 4: OLAP lay-out van een initieel ruwe hiërarchie van klantgroepen.



Afbeelding 5: Klantsegmenten als dimensies in een OLAP lay-out.

datamining-resultaten toe te laten, krijgen we (extra) inzichten. De tweede kant van de integratie betreft het nut van OLAP voor de datamining-exercities. OLAP kan de eerste stappen in de uitvoering van het datamining-proces verlichten door de datamining-expert te helpen met zijn/haar initiële kennisopbouw omtrent de data. Bijvoorbeeld, door aandacht te richten op belangrijker variabelen (volgens de bedrijfsregels waar de betreffende OLAP-kubus op rust), of door uitersten en uitzonderingen te identificeren of door interacties te vinden. Deze operaties zijn belangrijk om tot zinvolle datamining-resultaten te komen. Want hoe beter we de betekenis van verschillende data begrijpen, hoe effectiever het verkennend datamining-proces zal verlopen.

Bedrijfsregels zijn expliciet zichtbaar in de lay-out van een OLAP-kubus, in de vorm van dimensies en attribuut-dimensies. Deze dimensies en de onderliggende meetwaarden zijn zinvol voor datamining, omdat ze metadata naast data bevatten. Dit is van nut in vooral de eerste drie stappen van het totale datamining-proces. De leveranciers van OLAP-tools pakken deze tweede kant nu op. De OLAP-module Analysis Services binnen Microsoft SQL Server 2000 biedt de mogelijkheid om een datamining-model te baseren op OLAP-kubusinformatie en de resultaten te tonen als een nieuwe dimensie, meetwaarde, of als attribuut-dimensies. Cognos offreert regressie- en extrapolatie-functionaliteit in de nieuwste versie van haar OLAP-tool Powerplay.
In de vakliteratuur worden beide kanten van integratie aangeduid met 'mining then cubing' respectievelijk 'cubing then mining'.

Toepassingen

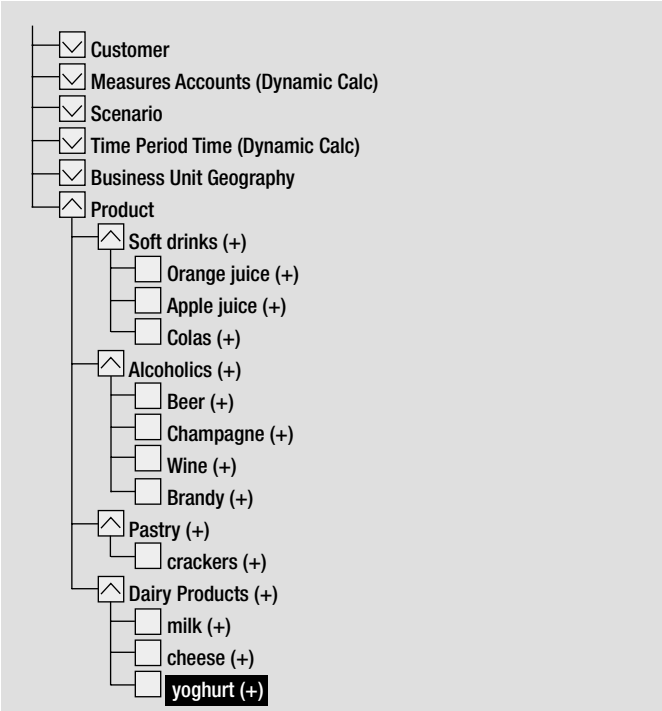
Het beschikbaar stellen van datamining-resultaten aan de bedrijfsanalist die multidimensionale OLAP front-end tools gebruikt, geeft nieuwe inzichten om bedrijfsdoelstellingen op te lossen zoals:

- Het vinden van winstgevende segmenten;
- Het begrijpen van koopgedrag;
- Het optimaliseren van product-portfolio's;
- Het behouden van klanten door inzicht te krijgen in de loyaliteitfactoren.

Deze doelstellingen zijn traditioneel altijd al aangepakt met datamining-tools. Door het integreren van patronen en regels die zijn gevonden door middel van mining-technieken, met de krachtige multidimensionale analyse-mogelijkheden van OLAP, krijgt de bedrijfsanalist nieuwe inzichten en verklaringen voor bijvoorbeeld het waargenomen klantgedrag.

Drie voorbeelden beschrijven hoe datamining en OLAP kunnen integreren. Datamining kan gebruikt worden om informatie te vinden die moeilijk of zelfs onmogelijk is te vinden via OLAP. OLAP kan echter een geschikt tool zijn om datamining-resultaten makkelijk in de bedrijfscontext te plaatsen. Men kan bijvoorbeeld een klantsegmentatie met behulp van datamining bepalen/analyseren en dan vias OLAP kijken naar de werkelijke verkoopcijfers van een specifiek segment voor een specifieke groep van producten. Door de groep gebruikers van mining-resultaten te verbreden wordt de opbrengst van de investeringen in datamining-technologie vergroot.

Voorbeeld 1. Segmentatie/cluster van klanten van een bank.
De klantdimensie zoals de bank die initieel heeft opgezet en in gebruik heeft in haar OLAP-kubus, toont een lay-out met een vrije grove hiërarchie van klantgroepen, zie afbeelding 4.
De focusgroep van de bank bestaat uit private klanten en het datawarehouse bevat veel attributen zoals leeftijd, gezinsstatus, naast gemiddeld jaarinkomen. Een clustermining-techniek met alle attributen vindt de volgende groepen klanten: Senioren; Families met kinderen; Yuppies; Anders.
Deze klantsegmenten worden geladen in de OLAP-kubus, als een extra niveau in de customer-dimensie, zie afbeelding 5. Zo wordt dan dimensionale analyse van elk toegevoegd klantsegment via geografie, product en tijd ondersteund. Zodoende kan de OLAP-analist een schijf bekijken die alleen het klantsegment "families with children" toont om verdere analyse uit te voeren op dit specifieke deel van de kubus, dat nu door middel van dimensies op basis van datamining is opgebouwd.



Afbeelding 6: Producthiërarchie in OLAP lay-out.

Afbeelding 7: Associaties als separate items in de productdimensie.

Voorbeeld 2. Market basket-analyse in de detailhandel.
Een typische OLAP lay-out in de detailhandel toont een hiërarchie/taxonomie in de product dimensie zoals in afbeelding 6. Door nu *market basket-analyse* uit te voeren op de individuele transacties (welke producten worden samen in gekocht) wordt bijvoorbeeld de volgende associatieregel gevonden: "Als een klant sinaasappelsap koopt dan koopt deze in 60 procent van alle gevallen ook brandy." Dergelijke regels kunnen dan toegevoegd worden aan de OLAP lay-out. Dat kan op twee manieren, als eerste: voeg alle geassocieerde producten toe als gedeeld lid onder het product waar ze mee zijn geassocieerd. En als tweede manier: voeg alle geassocieerde producten als separaat lid toe aan de productdimensie.
Afbeelding 7 toont het tweede alternatief. In beide gevallen kunnen de productassociaties natuurlijk worden gefilterd en alleen de meest interessante (wat betreft betrouwbaarheid en ondersteuning) worden toegevoegd aan de OLAP lay-out. In beide gevallen kan de analist de implicaties van bedrijfsbeslissingen makkelijk zien.

Via traditionele OLAP-analyse was de analist tot de conclusie gekomen dat sinaasappelsap niet meer op voorraad moet worden genomen omdat het bijvoorbeeld te weinig winst levert. Door de associatie met de zeer winstgevende brandy is de combinatie van beide producten ook winstgevend, zodat het verwijderen van sinaasappelsap uit het productschap tot minder winst kan leiden. Dit alles kan ook worden geanalyseerd in combinatie met andere dimensies, door bijvoorbeeld te kijken naar winst van geassocieerde producten per regio.

Voorbeeld 3. Voorspelde waarden in de OLAP kubus.
Voorspellingen van een neurale predictie als datamining-techniek is van nut om te bepalen of een klant van een bank een hoog of laag kredietrisico gaat vormen, of bijvoorbeeld dat een klant neigt om zijn bankrekening te sluiten en er een bij de concurrent te openen. De voorspelde waarde wordt dan toegevoegd aan de OLAP lay-out, als een extra meetwaarde in een separate kolom. Multidimensionale analyse kan dan bijvoorbeeld leiden tot de conclusie : "Klantverloop is in Amsterdam hoger dan wordt voorspeld". Natuurlijk kunnen de voorspelde waarden ook toegevoegd worden aan de lay-out als discrete dimensies, op dezelfde wijze als de cluster identifiers in het eerste voorbeeld.

Conclusie

Tegenwoordig zijn OLAP en datamining niet langer verzamelingen van stand-alone technieken die alleen gebruikt worden door respectievelijk OLAP-gebruikers en datamining-specialisten. De integratie van datamining en OLAP verhoogt de kwaliteit van OLAP-analyses door vooral sturing, dimensie-uitbreidingen en nieuwe meetwaarden. De kwaliteit van datamining wordt verhoogd door onder andere beter inzicht in data en betere sturing van de exploratie.
De toenadering tussen OLAP en datamining blijkt ook uit de ontwikkeling van die producten. Database-geïntegreerde OLAP-functionaliteit en datamining als extenders in een relationele database-omgeving, maken het eenvoudiger voor beide gebruikersgroepen om data, metadata en informatie met elkaar uit te wisselen. Verder biedt het een nauwere integratie in de uiteindelijke bedrijfsapplicatie, die gevoed zal worden met de kennis waarop dan resultaatgedreven bedrijfsmatige acties ondernomen zullen worden. Kortom, OLAP en datamining zullen steeds minder als aparte tools worden beschouwd en geïntegreerd worden toegepast.

Dr. Jaap Verhees (jaap_verhees@hotmail.com) is adviseur.
Dr. Rob Peters (rob.peters@ordina.nl) is senior consultant Business Intelligence bij Ordina.

Literatuur

John L. Doran, Business Intelligence Building Blocks: The Challenge of Gaining Business Insight. DM Review Online, maart 2003.
Paul F.H. van der Linden, Business Intelligence: drie werelden komen samen. DB/M nummer 6, oktober 2002.
IBM, DB2 OLAP Server, OLAP Miner User's Guide v8.1, IBM, maart 2002.