Informatietechnologie en Gegevensanalyse

Johan Blomme

1999

www.johanblomme.net

Inhoud

Inleiding

Tegenwoordig functioneren organisaties in een zeer dynamische omgeving waarbij marktposities snel kunnen veranderen. Klanten worden kritischer, producten lijken steeds meer op elkaar en technologische ontwikkelingen volgen elkaar snel op. Het tijdig herkennen van en het snel kunnen aanpassen aan deze veranderingen is van cruciaal belang.

Als bedrijven zich moeilijk kunnen diversifiëren op basis van producten alleen, wordt het probleemoplossend vermogen een doorslaggevend verkoopsargument. Meer en meer wordt de klant het middelpunt van elke bedrijfsactiviteit. Het besef groeit dat naast verkoop ook service, informatie en advies belangrijk zijn. Klantenservice wordt een zelfstandig marketinginstrument, ja zelfs als merk naar voren geschoven. Hierdoor kunnen ondernemingen een concurrentievoordeel opbouwen of behouden. De vaststelling dat meer en meer de nadruk komt te liggen op bedrijfsactiviteiten die een concurrentievoordeel opleveren, heeft tot gevolg dat informatietechnologie (die voorheen vooral werd aangewend voor de automatisering van (deel)processen in de administratie en de productie), gebruikt wordt om relevante bedrijfsinformatie te genereren. Het succes van relationele databases heeft geleid tot een overweldigende hoeveelheid operationele en historische data waarover bedrijven kunnen beschikken. Echter het ontsluiten van informatie op basis van operationele databases levert allerlei problemen op, o.m. op het vlak van prestaties en veiligheid. Het is op dit punt dat de nieuwste ontwikkelingen op het gebied van informatietechnologie kunnen ingezet worden om de juiste informatie over markten en consumenten ter beschikking te stellen en om nieuwe vormen van klantgericht denken en handelen te ondersteunen.

Tegen de achtergrond van het toenemend belang van kennis als economische productiefactor, schetsen we in hetgeen volgt de betekenis van informatietechnologie voor gegevensanalyse. In het bijzonder blijven wij stilstaan bij de ontwikkelingen die geleid

hebben tot het gebruik van toepassingen die gegroepeerd worden onder de noemer *business intelligence*. Met deze term wordt verwezen naar de processen en technieken om strategisch bruikbare informatie en kennis aan gegevens te onttrekken. Door het inrichten van gegevenspakhuizen (*datawarehousing*) worden data uit diverse bronnen samengebracht in een centrale gegevensencyclopedie die voor de gebruiker snel en gemakkelijk toegankelijk is. M.b.t. de betekenis van databasetechnologie voor beslissingsondersteuning wordt de aandacht gericht op gebruikersgestuurde OLAP-technieken en het zoekproces dat aangeduid wordt als *data mining*.

1. Verandering als uitdaging

De afgelopen decennia ontwikkelen Westerse maatschappijen zich gestadig naar op dienstverlening gerichte samenlevingen. De wereld is in economisch opzicht één grote markt geworden. "The market place" wordt steeds meer "the market space". Kenmerkend voor de overgang van arbeidsintensieve naar kennisintensieve economieën is het steeds toenemend belang van kennis als economische productiefactor. De explosieve ontwikkeling van de informatie- en communicatietechnologie fungeert in dit opzicht als een katalyserende factor. Veranderingen die reeds op gang waren gebracht, voltrekken zich door de stuwende kracht van nieuwe technologieën in een hoger tempo.

De veranderingen die lijken samen te komen in een beweging die leidt tot een in hoge mate door technologie gedomineerde toekomst staan niet op zichzelf en zijn verstrengeld met andere maatschappelijke transformatieprocessen. Niet alleen op economisch vlak, ook op cultureel vlak vallen grenzen weg. De toegenomen welvaart heeft zich o.m. vertaald in een culturele diversificatie met een pluriformiteit aan levensstijlen, culturele

uitingen, vrijetijds- en consumptiepatronen. Deze verscheidenheid tekent zich niet alleen af op maatschappelijk niveau. Culturele verscheidenheid leidt eveneens tot een grotere verscheidenheid aan individuele identiteiten. Door het proces van individualisering zien individuen zichzelf minder als onderdeel van vaststaande collectiviteiten en meer als zelfstandige individuen met zelfgekozen sociale verbindingen en netwerken. De hier genoemde processen, en in het bijzonder de vaststelling dat economieën steeds meer gedragen worden door kennis, blijven niet zonder gevolgen voor de bedrijfsvoering.

Hoe is, onder invloed van de verschillende transformaties die onder de noemer van de kennissamenleving schuilgaan, het vak marketing het laatste decennium veranderd ? Aan het begin van de jaren negentig werd door de *Gartner Group* vooropgesteld dat het overwinnen van veranderingen de hoofdtaak van de bedrijfsvoering zal uitmaken. Kortere productlevenscyclussen, een toenemende concurrentiedynamiek en de vaststelling dat veel markten veranderd zijn van een verkoopmarkt in een koopmarkt, zijn slechts enkele factoren die aangeven dat de marktsituatie voor veel bedrijven een stuk moeilijker geworden is. De afgelopen jaren is veel geschreven over de gewijzigde omgeving waaraan de marketing zich moet aanpassen. Individualisering, demassificatie, marktversplintering, fragmentatie, graascultuur, e.a. zijn vaak gebruikte termen om de hedendaagse consument te omschrijven. In een samenleving met vrijwel onbeperkte keuzemogelijkheden en onvoorspelbare varianten wordt doelgroepselectie veel minder eenvoudig. De ongrijpbare consument. Producten gaan steeds meer op elkaar lijken (en worden gemakkelijker gekopieerd) terwijl de concurrentie heviger wordt.

Bepalend voor de veranderingen in het vak marketing is de afnemende rol van de traditionele marketing mix-instrumenten (product, prijs, promotie, plaats). Deze blijken onvoldoende houvast te bieden in snel veranderende markten. Bedrijven moeten zich in toenemende mate concentreren op de wijze waarop zij hun dienstverlening organiseren. Een goed functionerende organisatie

is immers veel moeilijker te kopiëren. Kwaliteit zal in de hele (klanten)benadering kwantiteit als hoofdprioriteit verdringen. In deze context heeft Schultz (1990 : 12) het over de noodzaak van een nieuw type marketing :

"In the past, traditional marketeers, in an effort to organize their marketing activities, tried to aggregate customer wants and needs. In other words, they tried to find common products and concepts and areas which would appeal to large numbers of people who could be served economically and efficiently with a product or service. Because customers and their social norms were rather homogeneous, that management approach resulted in the development of the mass market for consumer products. ... Today, there's a new look at the marketplace. A new look to consumers. ... Today, and increasingly tomorrow, the market for many consumer products is and will be driven by time, not money. It is also driven by a desire by people to be different and unique, not more like others. This dramatic upheaval in the marketplace has resulted in an equally dramatic change in marketing. ... we're moving from what was historically mass marketing to a new form of specialized marketing".

Wat uit het bovenstaande kan afgeleid worden, is dat het meer dan ooit belangrijk wordt om de markt en de ondernemingsmissie duidelijk te definiëren. Schultz (1990 : 19) heeft het in dit opzicht over het belang van strategisch management :

"Two basic principles guide the development of any strategic marketing plan. 1. The organization's planning view and orientation must be external, that is, toward customers and markets, not internal or toward what has been done or what facilities or technologies are now available. 2. Strategic marketing plans must be based on identifying and making use of some sustainable competitive advantage".

De nieuwe eisen overstijgen de combinatie van product-prijs-promotie-plaats. Om in te spelen op de marktontwikkelingen moeten bedrijven zich met de hele organisatie richten op de markt. Bedrijfsprocessen dienen de klant als vertrekpunt te nemen. Het managen van klantrelaties wordt daarmee een essentieel onderdeel van marketing. Massamarketing is daarmee niet verdwenen, maar wél van sterk afnemend belang geworden. Het ultieme doel van de marketeer wordt het op individuele basis benaderen en bedienen van de klant. Dit brengt een fundamentele

verandering teweeg in de marktbewerking. Het gaat niet langer (alleen) om marktaandeel (*market share*) maar (in toenemende mate) om aandeel in de bestedingen van klanten (*share of wallet*). Dit betekent dat niet zozeer beoogd wordt om zoveel mogelijk producten aan eenieder te verkopen of af te zetten binnen een bepaald marktsegment, maar om een relatie met een klant zodanig te ontwikkelen dat deze zoveel mogelijk producten van een bepaalde leverancier afneemt. Hiermee verschuift het accent van de marketingactiviteiten meer van het winnen van nieuwe klanten naar het behouden van bestaande klanten (*retention marketing*). Aangezien het vervullen van behoeften van individuele klanten centraal staat i.p.v. het verkopen van producten is het, met het oog op het kennen van de *lifetime value* van een klant, belangrijk diens behoeften nauwkeurig in kaart te brengen. Een klant kan immers beschouwd worden als een generator van een aantal transacties gedurende een bepaalde periode, en het komt erop aan de behoeften van de klant te kennen en inzicht te verwerven omtrent welke klanten (potentieel) winstgevend zijn of niet.

2. Informatietechnologie

De veranderingen aan de vraagzijde van de markt resulteren de afgelopen jaren in een steeds nadrukkelijker wordende noodzaak tot klantgerichtheid. Hierdoor zijn ook de eisen t.a.v. bedrijfsinformatiesystemen veranderd. Werden deze laatste vooral in verband gebracht met het automatiseren van bedrijfsprocessen, waarbij automatisering rond de bestaande processen werd gedrapeerd, dan wordt informatie meer en meer als een productiefactor beschouwd. In die zin krijgt informatie een strategische betekenis aangezien het een middel wordt tot ondersteuning van de besluitvorming. Het snel anticiperen op de wisselende behoeften van consumenten waardoor een bedrijf een concurrentievoordeel kan verwerven, is afhankelijk van informatie.

De revoluties die zich voltrekken in de informatietechnologie, en daarmee samenhangend in de media, grijpen diep in in het marketingproces. Het is daarbij niet dat de informatietechnologie voorschrijft wat wel en niet kan. Veeleer moet informatietechnologie gezien worden als een *enabling technology* waardoor bedrijfsprocessen (zoals de klantenadministratie) aansluiten op hetgeen vanuit klantenperspectief wenselijk is (en niet omgekeerd) en besluitvormingsprocessen door informatie ondersteund worden.

De voorbije decennia werd automatisering vooral aangewend voor de ondersteuning van eerst de administratieve en daarna de overige bedrijfsprocessen. Het resultaat ervan was de snelle verwerking van gegevens en het gebruik van deze laatste voor het aansturen van de verschillende onderdelen van het productie- en distributieproces. Door de technologie kregen bedrijfsprocessen en de daaruit resulterende producten en diensten een informatiecomponent waardoor 'waarde' werd toegevoegd aan de fysieke keten. Geheel in overeenstemming met de kenmerken van een aanbodgerichte marktstrategie was de toegevoegde waarde van de informatiecomponent intern gericht en werd deze voornamelijk aangewend voor het kwantitatief en kwalitatief optimaliseren van bestaande producten en diensten om uiteindelijk het marktaandeel te vergroten.

Door de toepassing van direct marketing technieken bleek het mogelijk om de informatiecomponent ook afzonderlijk te exploiteren. Door te sturen vanuit speciaal daartoe gebouwde marketing databases, nemen de mogelijkheden tot effectieve marktbewerking aanzienlijk toe. Aan de hand van de in een database opgeslagen informatie over prospecten en klanten, is het mogelijk om in grote markten een individuele benadering toe te passen. Aanvankelijk ging het om gepersonaliseerde massacommunicatie, waarbij eenzelfde boodschap wordt verstuurd aan individueel geadresseerde personen. Maar ook massale persoonlijke communicatie vindt thans op grote schaal plaats. Hierbij wordt de boodschap individueel afgestemd op de

kenmerken die van de ontvanger bekend zijn. Met de mogelijkheden die de informatietechnologie vandaag biedt, kunnen prospecten en klanten niet alleen individueel benaderd worden, maar wordt het ook mogelijk om op massale schaal maatwerk te leveren. Daardoor kan massamarketing in consumentenmarkten vervangen worden door één-op-één marketing[1].

De individuele relatie op basis van data in marketing databases kan niet alleen tot stand gebracht worden omdat technologische mogelijkheden grootschalig datagebruik mogelijk maken, maar ook omdat deze een nieuwe dimensie scheppen voor de communicatie tussen aanbieder en afnemer. Werd in de tijd van de overheersende massamarketing een medium vooral als advertentiemedium gebruikt, dan worden media in toenemende mate gebruikt om één-op-één relaties te leggen met prospects en klanten. In de jaren negentig zijn het vooral de communicatiemogelijkheden die tot een verandering leiden[2]. De massamedia zullen in de (direct) marketing een plaatsje moeten opschuiven voor de intussen steeds drukker bereden informatiesnelweg. Niet in het minst omdat het een bij uitstek interactief medium is, waardoor de gewenste direct response vanzelfsprekend wordt. Multimedia, interactieve media, elektronische snelwegen en interactieve marketing zijn termen die steeds vaker gebruikt worden om aan te geven dat de informatietechnologie nu gebruikt wordt als een middel om op een nieuwe manier te communiceren[3]. De consument krijgt meer vormen en methoden van communicatie tot zijn beschikking en ontwikkelt zo, op basis van zijn behoeften en wensen, een eigen wijze van interactie met de omgeving. Afstand, plaats en tijd vormen geen belemmering meer waardoor de informatiecomponent een eigen leven gaat leiden parallel aan en in interactie met de fysieke waardeketen[4]. Hierdoor ook beweegt de informatiecomponent van *back office*- naar *front office*-processen. Aan de voorkant van de organisatie komt een laagdrempelige klanteninterface te staan. De oriëntatie verschuift van een product- naar een klantoriëntatie. Bedrijven moeten zich meer en meer richten op het reactieve van de consument. Zij moeten zo

bereikbaar mogelijk zijn, zowel in medium als in tijd. Telematica-instrumenten zorgen voor een betere interactie met de consument en zijn daarom belangrijke instrumenten in de (elektronische) distributieketen. Door callcenters en internet ontstaan virtuele marktplaatsen waardoor gegevens over prospecten en klanten ingewonnen worden. De uitgebreide communicatie-, databeheer- en analysemogelijkheden laten toe een dieper inzicht in de markt en de eigen organisatie te verkrijgen. Dit verklaart het toegenomen belang van de klantendatabase waarop informatietechnologie met succes kan toegepast worden voor zowel het aanboren van informatie en kennis over klanten als het aansturen van marketingacties.

3. Gegevensanalyse

De besproken ontwikkelingen blijven inderdaad niet zonder gevolgen voor de databasetechnologie. De sterk gestegen aandacht voor datawarehousing (gegevenspakhuizen) en data mining (gegevensdelving) houdt verband met het toegenomen belang van informatie ter ondersteuning van beslissingen (*decision support*).

Zowel het feit dat veel gegevens van prospecten en klanten geautomatiseerd verzameld worden als de vaststelling dat operationele systemen heel wat (verborgen) informatie in zich houden, verklaren de belangstelling voor nieuwe technologieën die toelaten gegevens te ontsluiten en hieraan informatie te onttrekken die een gerichte en directe benadering van bestaande en nieuwe klanten toelaten. De ontwikkelingen op het vlak van databasetechnologie voor beslissingsondersteuning markeren in dit opzicht een omslag, een technologische verschuiving ten opzichte van het relationele model dat in de jaren tachtig opgang maakte.

Het relationele model is ontwikkeld in een periode dat databanken een andere functie in de informatievoorziening van bedrijven hadden dan tegenwoordig het geval is. Relationele database management systemen (RDBMS) richten zich bij uitstek op de ondersteuning van administratieve processen *(On Line Transaction Processing*, OLTP). Sterk samenhangend met de administratie van transactionele gegevens is ook de betekenis van normalisatieprocessen die de gegevensintegriteit van relationele databases moeten verzekeren enerzijds en de structuur van de zoektaal *SQL* anderzijds.

Het belang van informatie ter ondersteuning van beslissingen voert evenwel tot de conclusie dat een operationele database niet meer volstaat. Vooral het feit dat beslissingsondersteuning de toegang tot een brede verzameling gegevens vereist, is hiervan de oorzaak. Om die reden wordt de ontwikkeling van gegevenspakhuizen (datawarehousing) door velen beschouwd als de spil van de hedendaagse IT-architectuur. Terwijl transactionele gegevens (OLTP) zodanig georganiseerd zijn dat ze snel opgeslagen en opgeroepen kunnen worden, zijn de gegevens in een gegevenspakhuis georganiseerd op een wijze die uiteenlopende analyses mogelijk maken.

Het multidimensioneel modelleren van gegevens dat kenmerkend is voor OLAP *(On Line Analytical Processing*) maakt het mogelijk door aggregatie en samenvatting snel toegang te krijgen tot gegevens en deze vanuit verschillende gezichtspunten te benaderen. Naast de meer traditioneel op hypothesentoetsing geënte verificatie-technieken bieden met name explorerende technieken voor kennisontdekking (data mining) de mogelijkheid tot extractie van verborgen patronen in gegevens.

3.1. Datawarehousing

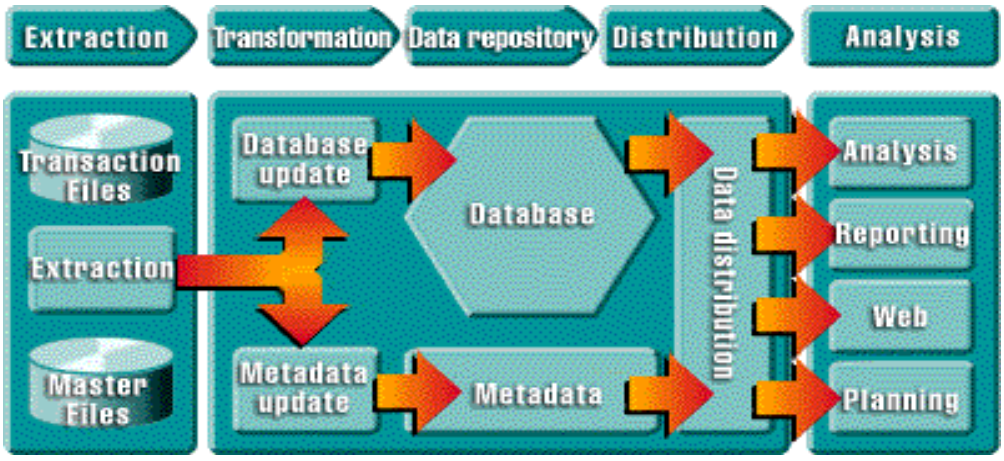
De afgelopen jaren sterk gestegen belangstelling voor datawarehousing wordt onderbouwd door de groeiende oriëntatie op klanten en, daarmee samenhangend, het strategisch belang van informatie voor besluitvorming. Als een continu proces voorziet datawarehousing in het ter beschikking stellen van gegevens om informatie voor beslissingsondersteuning te leveren. Datawarehousing is bedoeld om productie-georiënteerde gegevensbestanden voor analysedoeleinden beschikbaar te maken. De gegevens waarmee een datawarehouse wordt gevuld, zijn grotendeels afkomstig van de in de organisatie geëxploiteerde operationele/transactionele systemen, de bronsystemen. Daarnaast worden ook externe gegevens (bv. marktonderzoekgegevens en gegevens uit externe databanken) in een datawarehouse opgenomen. Eén van de redenen voor het opzetten van een datawarehouse is dat de operationele processen niet of zo weinig mogelijk verstoord worden door activiteiten die gericht zijn op het genereren van managementinformatie. Daarom worden queries voor dit laatste doel bij voorkeur niet rechtstreeks op de gegevens in productie-omgevingen uitgevoerd. Deze gegevens worden daarom gekopieerd naar een omgeving waar zij zonder storende werking kunnen worden geanalyseerd. Dit heeft (zoals o.m. het geval is met multidimensionele databases) het bijkomende voordeel dat de gegevens op een wijze worden opgeslagen die meer geschikt is voor dit analysewerk.

In zijn veel geciteerde studie, *Building the Data Warehouse* (1996) , bakent W.H. Inmon een gegevenspakhuis af van een relationale database door een gegevenspakhuis te definiëren als een onderwerp gerichte, gëintegreerde, tijdsgebonden en statische verzameling van gegevens ter ondersteuning van besluitvorming. Terwijl de gegevens in een operationele database transactie-georiënteerd zijn, zijn de data in een gegevenspakhuis gericht op de behoeften van de eindgebruiker en als zodanig gemodelleerd. Dit laatste houdt in dat gegevens over eenzelfde onderwerp (bv.

klantgegevens) die in een OLTP-omgeving in verschillende productiesystemen opgeslagen en verwerkt worden, in een gegevenspakhuis samengebracht worden per onderwerp. De data zijn anderzijds geïntegreerd, hetgeen betekent dat gegevens die in OLTP-omgevingen vaak in een verschillend formaat worden beheerd, in een gegevenspakhuis zodanig geconsolideerd worden dat ze op eenduidige wijze te benaderen zijn. Om die reden maakt het aanleggen van gegevensdefinities (metadata, gegevens over gegevens) een cruciaal onderdeel uit van datawarehousing.

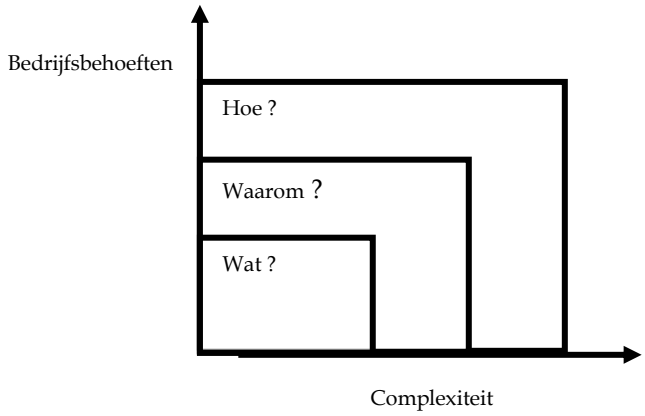
Een ander verschilpunt met gegevens opgeslagen in operationele databases is dat data in een gegevenspakhuis tijdsgebonden zijn en derhalve een historische dimensie hebben. Weerspiegelen de data in OLTP-systemen een momentopname, dan hebben analyses van data in gegevenspakhuizen vaak tot doel verschuivingen en trends op te sporen. Alleen historische data laten toe veranderingen in kaart te brengen.

Tenslotte zijn de data in een gegevenspakhuis duurzaam (statisch). Er worden, in tegenstelling tot databases voor operationeel gebruik, geen gegevens veranderd, noch verwijderd ; er worden enkel gegevens toegevoegd. In een operationele database fungeren normalisatie- en integriteitsregels als condities voor het wijzigen, toevoegen en verwijderen van gegevens op recordniveau. Tegenover de eisen voor update-optimalisatie in RDBMS, staat query-optimalisatie centraal in een datawarehouse-omgeving.



Figuur 1 : Datawarehousing

Ten aanzien van de analyses die op de gegevens in een datawarehouse plaatsvinden, kan een tweedeling gemaakt worden tussen op verificatie gerichte analyses en analyses gericht op het ontdekken van nieuwe kennis. Het onderscheid tussen beide benaderingen kan verduidelijkt worden door de behoeften op het vlak van beslissingsondersteuning weer te geven. In de eerste plaats is het belangrijk te weten wat er in de markt gebeurt. Dergelijke wat-vragen (bv. Wat is de evolutie van het marktaandeel van de eigen producten t.o.v. concurrerende producten gedurende de afgelopen vijf jaren ?) worden beantwoord aan de hand van vraaggestuurde data-analyses, geïnduceerd door gebruikers die aan de hand van op voorhand geformuleerde hypothesen de gegevens in een datawarehouse benaderen. Zowel traditionele queries als OLAP-technieken ondersteunen deze op verificatie gebaseerde gegevensanalyse. Behalve het antwoord op wat-vragen heeft de bedrijfsvoering ook behoefte aan analyses die een antwoord geven op de vraag waarom ontwikkelingen zich voordoen in de markt en hoe hierop kan gereageerd worden. Het antwoord op waarom- en hoe-vragen vereist inzicht in de markt en kennis van het gedrag van klanten om te kunnen voorspellen hoe deze zich in de toekomst zullen gedragen en hoe de verworven inzichten kunnen vertaald worden in een strategisch voordeel.



Figuur 2 : Vraagstellingen bij beslissingsondersteuning

In tegenstelling tot de gebruikersgerichte aanpak die centraal staat in traditionele querying en OLAP, hebben de kennisontdekkende algoritmen die onder de noemer 'data mining' worden ondergebracht een gegevensgestuurd karakter. Gegevens worden zonder vooraf geformuleerde hypothesen doorzocht op ongekende/onverwachte verbanden en patronen. Kennisontdekkende technieken werken autonoom (zonder begeleiding) en de relaties en patronen die erdoor aan de oppervlakte gebracht worden, leiden tot nieuwe inzichten en het nemen van de daarbij passende beslissingen en acties.

3.2. Multidimensionele gegevensanalyse

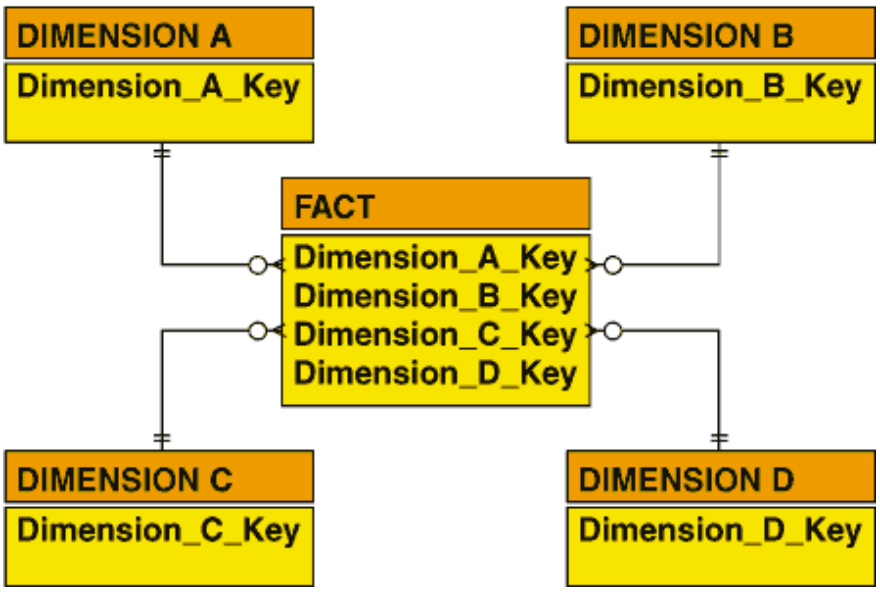
Uit de omschrijving van een gegevenspakhuis kan afgeleid worden dat de organisatie van een datawarehouse verschilt van een operationele database. Aangezien de inhoud van een operationele database door de gebruiker wordt gewijzigd, is het gegevensmodel van een operationele database gericht op het ondersteunen van transacties die vele malen dezelfde bewerkingen op gegevens uitvoeren. Om die reden ligt de nadruk bij de modellering van operationele databases op normalisatie om redundantie en

verstoring van de database-integriteit te voorkomen. De onderliggende genormaliseerde structuur wordt voor de gebruiker verborgen gehouden. De door normalisatie ingebouwde beveiligingsmechanismen schermen de fysieke gegevensstructuur van een operationele database af tegen directe benadering door eindgebruikers. In het geval van het gebruik van een gegevenspakhuis wijzigt de gebruiker geen gegevens en wordt de toegang tot de gegevens gedefinieerd vanuit het gezichtspunt van de gebruiker. De ontwerpfilosofie van een datawarehouse bestaat er dan ook in een open toegang tot gegevens te bieden ter ondersteuning van een breed scala van queries. Bij de analyse van gegevens die in een datawarehouse opgeslagen liggen, neemt de multidimensionele modellering van gegevens een belangrijke plaats in.

De techniek die bekend staat als het multidimensionele stermodel sluit aan bij het uitgangspunt om de gegevens in een datawarehouse te modelleren vanuit het perspectief van de eindgebruiker. In het stermodel worden feiten zoals verkopen, facturen, betalingen, e.d. gekwalificeerd langs meerdere dimensies. Het centrum van de ster wordt de feitentabel (*fact table*) genoemd. In de (gedenormaliseerde) feitentabel worden naast de eigenschappen van het centrale object ook de verwijzende sleutels (*foreign keys*) naar de dimensietabellen bijgehouden. Dimensietabellen bevatten attributen die de dimensiewaarden beschrijven (en worden in SQL-opdrachten vaak als zoekcriteria gebruikt).

In figuur 3 wordt een voorbeeld gegeven van een eenvoudig stermodel[5]. In de feitentabel worden bijvoorbeeld gegevens opgeslagen over de verkoop van producten in een supermarktketen. In de centrale feitentabel worden naast gegevens over aantal verkochte artikelen en omzet ook de verwijzende sleutels bijgehouden naar dimensietabellen (mogelijke dimensies zijn tijd, (winkel)locatie, product en klant).

Een benadering die toelaat een potentiële feitentabel en kandidaten voor dimensies in het datawarehouse te identificeren, is het zgn. Starnet-model. Een dergelijk model geeft aan welke de verschillende dimensies (en de onderdelen ervan) zijn die bij ieder aandachtsgebeid horen. Op die manier vormt het Starnet-model een uitgangspunt voor het modelleren van het datawarehouse. Het Starnet-model is ook vanuit het standpunt van gegevensanalyse een belangrijk hulpmiddel. Het Starnet-model wordt gebruikt om de behoeften te formuleren m.b.t. de samenvoeging van gegevens ten einde het aantal te analyseren dimensies te verminderen en de mate waarin gegevens op een lager dan wel hoger aggregatieniveau geanalyseerd worden.

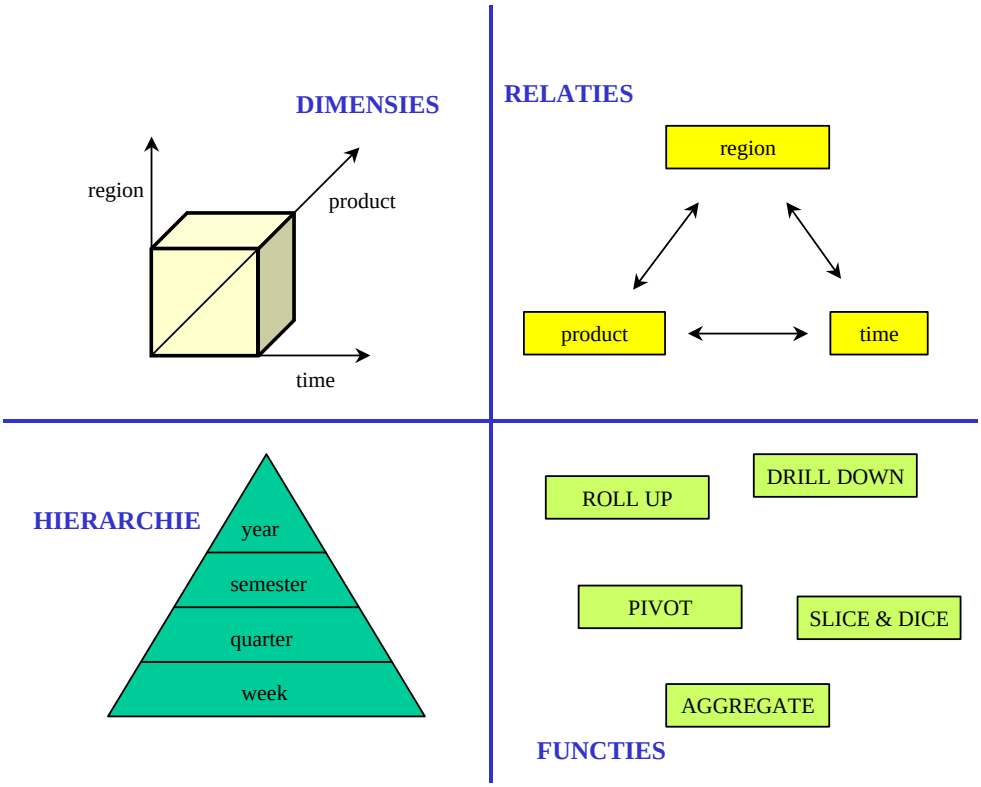


Figuur 3 : Weergave van een sterschema

De methode die bekend staat als *On Line Analytical Processing* (OLAP) is gebaseerd op de multidimensionele analyse van gegevens, onafhankelijk van de wijze waarop de gegevens fysiek zijn opgeslagen. Wat dit laatste betreft, kan een onderscheid worden gemaakt tussen multidimensionele en relationele opslag. In het geval van multidimensionele opslag, die bekend staat onder de naam MOLAP, worden frequent benaderde gegevens uit het datawarehouse door voorberekening, samenvatting en aggregatie

naar meerdere dimensies opgeslagen in de multidimensionele opslagcapaciteit van de OLAP-server. Deze laatste is hierbij dikwijls ingericht als een *data mart*, d.w.z. een datawarehouse met een beperkt functioneel of organisatorisch aandachtsgebied, zoals een business unit. In het geval van relationele OLAP (ROLAP) presenteert de OLAP-server de gegevens die relationeel opgeslagen liggen in een datawarehouse onder de vorm van een multidimensioneel model.

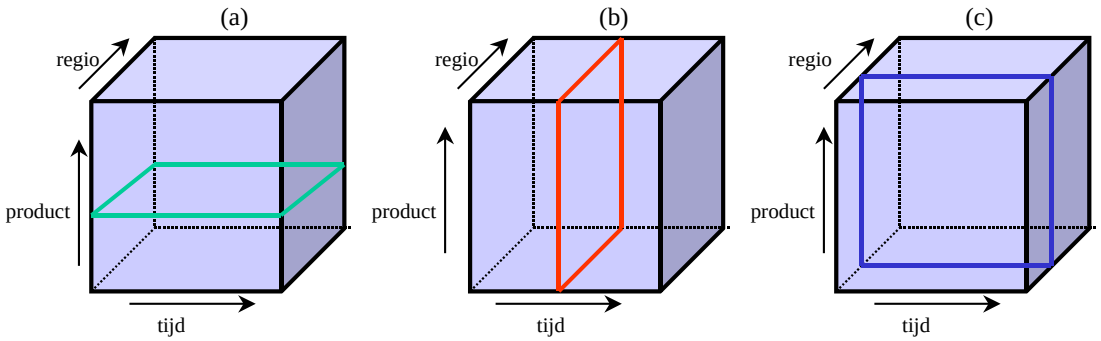
Zoals kan afgeleid worden uit het hierboven beschreven stermodel, worden bij meerdimensionele analyses meerdere gegevensdimensies onderscheiden. Binnen deze laatste kunnen vaak ook nog hiërarchieën aangebracht worden. Illustratief hiervoor is de tijdsdimensie die toelaat gegevens per dag, per maand, per kwartaal, enz. te analyseren. Het is gebruikelijk om in het geval van OLAP gegevens te aggregeren. Detailgegevens verliezen immers hun waarde in de tijd en de naar meerdere dimensies geconsolideerde gegevens bevorderen zowel de efficiënte opslag van data als snelle responstijden.



Figuur 4 : Multidimensionele gegevensanalyse

OLAP-toepassingen voorzien in een aantal navigatietechnieken. Behalve de mogelijkheid om gegevens vanuit meerdere dimensies te analyseren (*pivoting*) en hierbij meerdere gegevensdoorsnijdingen te gebruiken (*slice en dice)* wordt met *drill down* verwezen naar analyses waarbij gegevens (stapsgewijs) vanuit een hoger aggregatieniveau op een lager, meer detaillistisch niveau worden bestudeerd. Het tegenovergestelde wordt *roll up* genoemd. Een OLAP-tool voorziet in de mogelijkheid om snel gegevens te analyseren vanuit meerdere invalshoeken. Stel bijvoorbeeld dat verkoopcijfers worden geanalyseerd naar het soort product, de regio en de tijdsdimensie. Zoals afgebeeld in figuur 5 (a) kunnen verkoopcijfers gerapporteerd worden per product, over alle regio's en alle tijdsdimensies. De afzet kan anderzijds ook beschouwd worden in een bepaalde periode, over alle producten en regio's (b). Tenslotte (c) kan de afzet in een

bepaalde regio, voor alle producten en periodes, gerapporteerd worden.



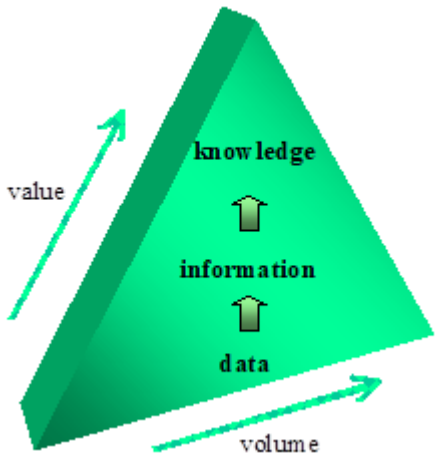
Figuur 5 : OLAP

3.3. Kennisontdekking

3.3.1. Data mining

Vaak in één adem genoemd met OLAP maar niettemin wezenlijk verschillend ervan is het zoekproces dat aangeduid wordt als data mining. In het algemeen kan data mining omschreven worden als het destilleren van onbekende informatie uit grote gegevensbestanden :

"Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques" (Gartner Group).

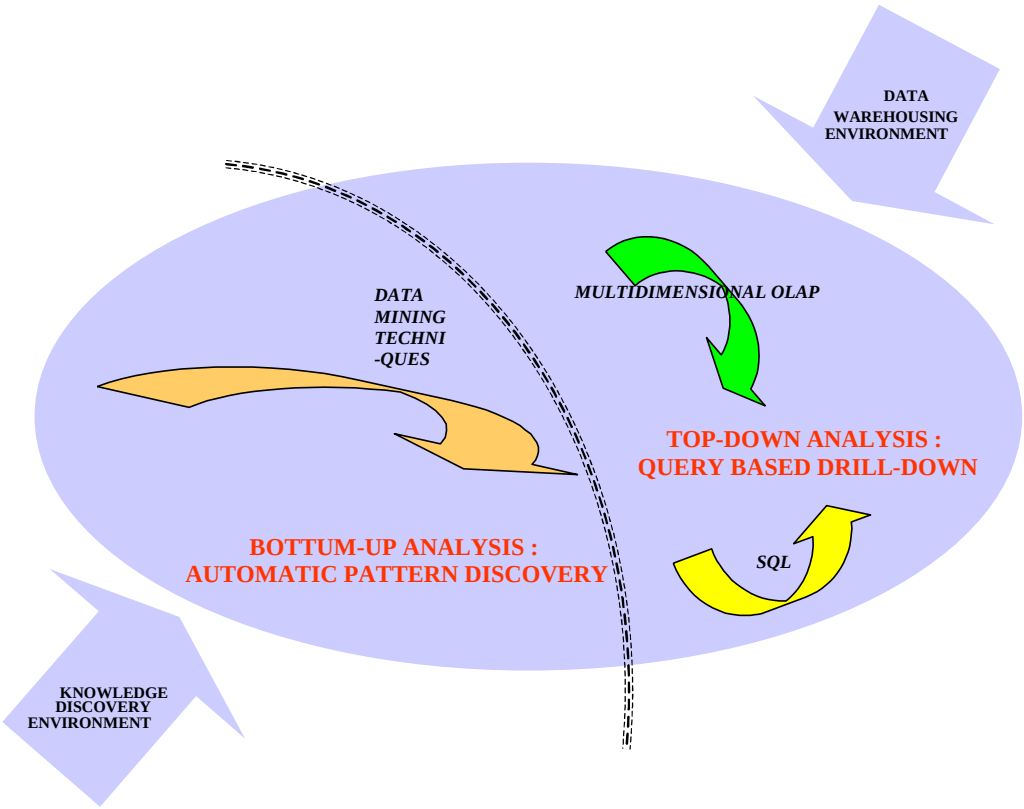


Figuur 6 : Data volume / knowledge value

Dat de aandacht voor data mining en in het bijzonder de mogelijkheden ervan om informatie aan te wenden voor het verwerven van een concurrentieel voordeel steeds meer onderkend worden, is o.m. toe te schrijven aan de convergentie van een drietal technologische ontwikkelingen. In de eerste plaats maakt datawarehousing het mogelijk om op massale basis data op te slaan en deze data toegankelijk te maken voor eindgebruikers. In combinatie met een eveneens sterk toegenomen verwerkingscapaciteit (parallelle databasetechnologie) en het gebruik van kennisontdekkende algoritmen, biedt data mining een geschikte oplossing voor de analyse van grote hoeveelheden gegevens waarvoor querymethoden minder geschikt zijn.

Ook data mining markeert een omslag t.o.v. het relationele model. Het in dit laatste centraal staande begrip 'sleutel' leidde tot het formuleren van verschillende normaalvormen die als leidraad dienen voor het ontwerp van relationele databases. Deze normaalvormen fungeren als *constraints*, o.m. om update- en verwijderanomalieën te vermijden (sleutels identificeren records uniek)[6]. Vanuit het perspectief van data mining gaat de aandacht niet naar het extraheren van unieke records maar naar het aantal keer dat objecten of events voorkomen. Het gaat er niet langer om functionele afhankelijkheden te gebruiken voor het ontwerpen van databases maar om ongekende afhankelijkheden vanuit statistisch

oogpunt te traceren in de gegevens[7]. Dit impliceert eveneens dat in het geval van data mining geen genormaliseerde maar gedenormaliseerde tabellen het startpunt van analyse uitmaken.



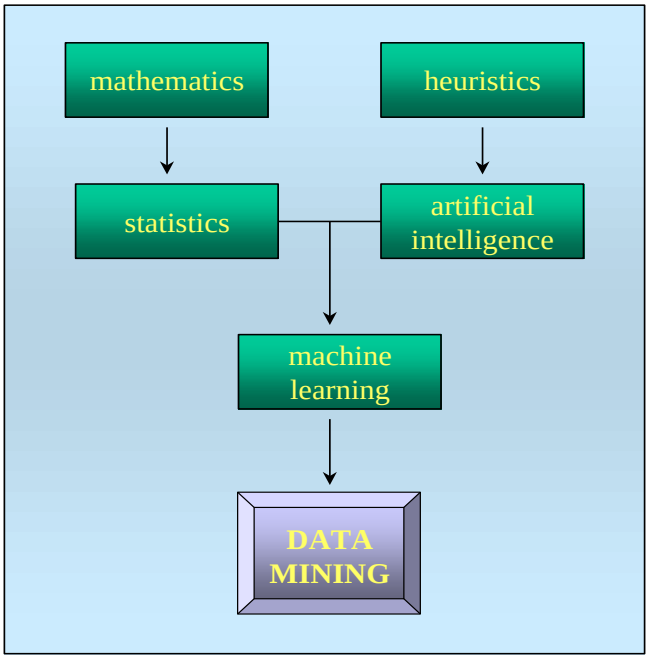
Figuur 7 : KDD versus datawarehousing

3.3.2. Data mining, statistiek en OLAP

Zoals vermeld kan data mining omschreven worden als het op geautomatiseerde wijze zoeken naar patronen en verbanden in grote gegevensverzamelingen. In dit opzicht wordt bij data mining gebruik gemaakt van de inzichten uit de statistiek, de kunstmatige intelligentie en machine-leren. De meeste data mining-technieken ontlenen aan de statistiek het begrippenkader en de methoden aan de hand waarvan variabelen en relaties tussen variabelen geanalyseerd kunnen worden (bv. standaardafwijking, variantie,betrouwbaarheidsintervallen, e.a.). Terwijl steekproefonderzoek de basis vormt voor statistische analyse (*inferences from data*, inferentiële statistiek) worden data mining-technieken aangewend om patronen op te sporen in meestal volledige gegevensverzamelingen die veelvouden van gigabytes groot kunnen zijn (VLDB,*very large databases*). Een hiermee verbonden verschilpunt is dat de analist-gebruiker van statistische technieken verondersteld wordt een idee te hebben van de vorm van het model dat zal gebouwd worden, en dus een inzicht heeft in zowel de variabelen die hiertoe zullen gebruikt worden als de combinaties tussen deze laatste. Aan de basis van de analyse van VLDB ligt daarentegen juist de doelstelling om (langs exploratieve weg) ongekende verbanden en patronen aan de oppervlakte te brengen. Ook omdat traditionele statistische technieken vaak niet goed kunnen omgaan met zeer grote gegevensbestanden, de verwerking van duizenden velden hiermee niet mogelijk is en deze technieken gevoelig zijn voor afwijkende gevallen (*outliers*) in zeer grote databestanden, wordt in het geval van data mining gebruik gemaakt van zgn. adaptieve technieken die – zonder tussenkomst van de gebruiker – toelaten te achterhalen welke variabelen invloedrijk zijn en welke de belangrijke combinaties zijn.

Adaptieve technieken vinden hun oorsprong in het domein van de kunstmatige intelligentie (*artificial intelligence*), een onderzoeksdomein waarin onderzocht wordt hoe de menselijke denkwijze en het leerproces door machines kan gereproduceerd

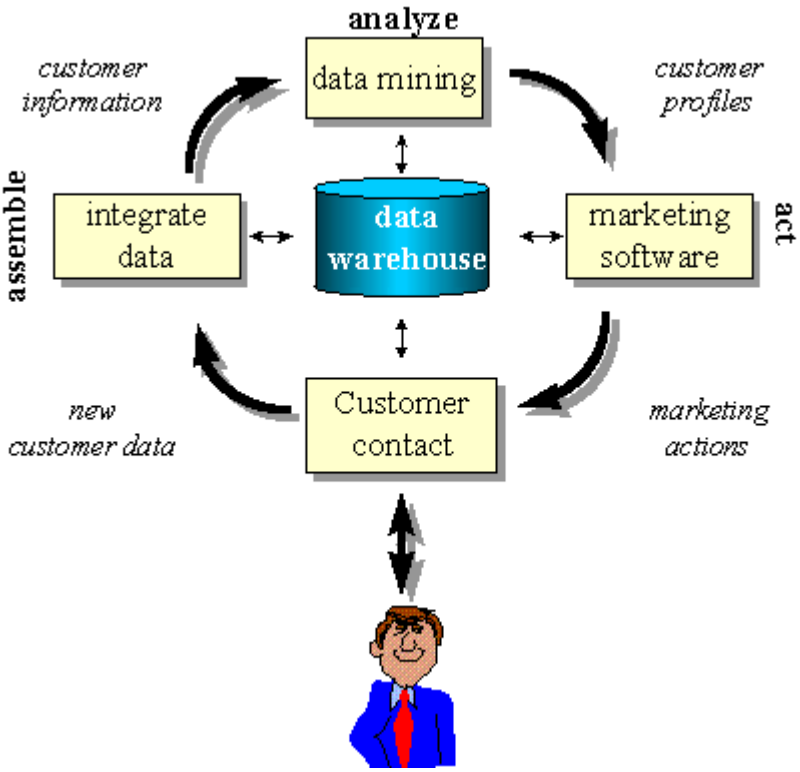
worden. Door de enorme rekenkracht die toepassingen van AI opeisen, bleven commerciële successen ervan uit. Aan het eind van de jaren tachtig, toen de prijs-prestatie verhouding van computers gunstiger werd, werden de technieken die ten grondslag liggen van AI geïmplementeerd binnen het gebied van het machine-leren. Dit laatste kan beschouwd worden als een voortzetting van de inzichten uit het domein van de kunstmatige intelligentie waarbij grondbeginselen uit de statistiek ingebouwd worden in geavanceerde algoritmen voor patroonherkenning.



Figuur 8 : Data mining

Aan de basis van data mining ligt de toepassing van een geheel van kennisontdekkende algoritmen. Om die reden wordt data mining niet zelden vereenzelvigd met *Knowlegde Discovery in Databases* (KDD), waarvan het evenwel een onderdeel vormt (zie figuur 9). KDD is een iteratief proces dat begint met het bepalen van de te beantwoorden vraagstelling. Vervolgens vindt gegevensselectie plaats uit een datawarehouse maar ook databases met transactiegegevens kunnen als bronbestanden dienen (*assemble*). Aangezien de kwaliteit van de gegevens van essentieel

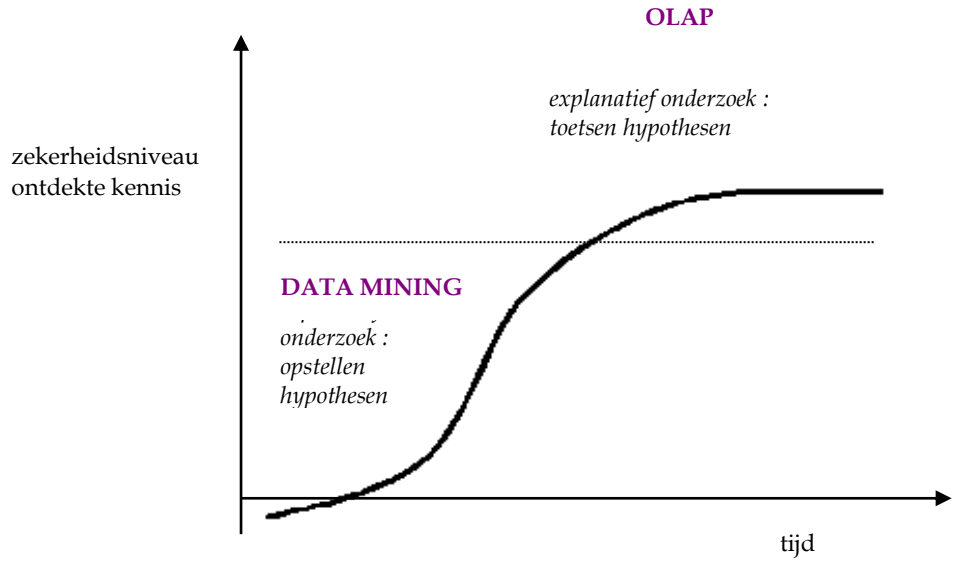
belang is, worden de gegevens veelal in meerdere stappen voorbewerkt (opschonen, transformatie, datatype-conversies, aggregatie, denormalisatie). Nadat de gegevens geconsolideerd zijn, kunnen ze door een (geschikte) techniek onderzocht worden op trends en patronen (*analyze*). Deze dienen eerst geïnterpreteerd en op de bruikbaarheid ervan gevalideerd te worden. Op basis van de nieuw aangeboorde informatie worden o.m. productverbeteringen en klantprofielen ontwikkeld die in een volgende fase tot marketingacties leiden (*act*). Deze laatste resulteren op hun beurt in klantcontacten die nieuwe klantgegevens genereren die terug in de centrale gegevensbank geïntegreerd worden waar ze o.m. gebruikt worden voor het bijsturen van modellen en profielen waarop marketingacties gebaseerd zijn[8].



Figuur 9 : Data mining als een onderdeel van KDD[9]

In vergelijking met OLAP en traditionele statistische technieken waarbij het beantwoorden van tevoren gestelde vragen een zeker inzicht in de gegevensstructuur veronderstelt – en de analyse derhalve gebruikersgestuurd is – wordt het zoekproces naar onderliggende patronen en verbanden in het geval van data mining uitgevoerd zonder veel bemoeienis van de gebruiker. De analyse is gegevensgestuurd. Een belangrijke randvoorwaarde hierbij is dat de gebruiker voldoende domeinkennis heeft van het bedrijfsproces dat geanalyseerd wordt. De gebruiker moet in staat zijn om voor de te beantwoorden vraagstelling de relevante gegevens te verzamelen. Deze gegevens dienen veelal eerst bewerkt te worden om ze geschikt te maken voor analyse. Pas daarna vindt analyse plaats, waarbij voor de interpretatie van de resultaten eveneens deskundigheid op het toepassingsgebied een vereiste is.

Bij data mining wordt een inductieve werkwijze gevolgd die ook bekend staat als exploratieve data-analyse (EDA)[10]. Het verschil met de hypothetisch-deductieve methode die gevolgd wordt voor de toetsing van verbanden kan verduidelijkt worden aan de hand van een empirische cyclus. In een eerste fase van deze cyclus, waartoe ook data mining kan gerekend worden, worden waarnemingen en verbanden tussen waarnemingen onder een gemeenschappelijke noemer geplaatst. Dit proces wordt inductie genoemd. Het resultaat ervan is een theorie, d.w.z. een geheel van uitspraken waarvan het geldigheidskarakter evenwel nog niet vaststaat. Het afleiden van hypothesen uit de abstracte gedeelten van de theorie wordt deductie genoemd. Ondanks de verschillen tussen data mining en OLAP zijn beide ook complementair aan elkaar. De hypothesen die gegenereerd worden uit het data mining- proces kunnen met behulp van OLAP-tools geverifieerd worden.



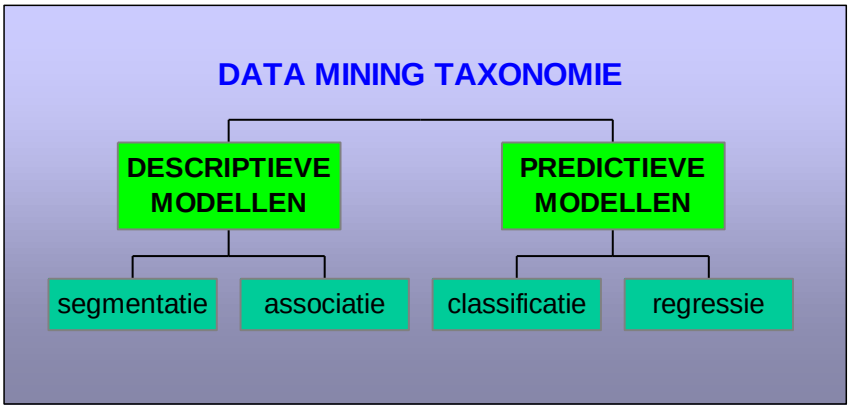
Figuur 10 : Empirische onderzoekscyclus

Een belangrijk verschil tussen data mining en exploratieve data-analyse is dat de doelstelling van data mining er niet in bestaat de interrelaties tussen variabelen inzichtelijk te maken of te verklaren. Veeleer is het de bedoeling op basis van het samenspel van variabelen tot een predictie te komen. Vandaar ook dat heel wat toepassingen van data mining gebaseerd zijn op scoringsmodellen. Scoringsmodellen zijn voorspellende modellen. Het doel ervan is een afhankelijke variabele te voorspellen aan de hand van een aantal onafhankelijke variabelen. Op grond van deze laatste kenmerken kunnen klantscores bepaald worden die een voorspellende waarde hebben. Zo kunnen klantprofielen opgesteld worden die de waarschijnlijkheid weergeven dat een prospect of klant reageert op een mailing of een product aankoopt. Banken en verzekeringsmaatschappijen maken o.m. gebruik van credit-scoringsmodellen om de kredietwaardigheid van hun klanten te beoordelen.

3.3.3. Patroonherkenning

Hoewel data mining wordt toegepast voor het beantwoorden van een breed scala aan vraagstukken, hebben vrijwel alle toepassingen gemeenschappelijk dat gezocht wordt naar patroonherkenning. De methoden die hierbij gebruikt worden, kunnen in een viertal categorieën onderverdeeld worden, nl. segmentatie (clustering), associatie, classificatie en regressie. Het onderscheid tussen deze methoden houdt in de eerste plaats verband met het al of niet aanwezig zijn van een opdeling tussen afhankelijke (te verklaren) en onafhankelijke (verklarende) variabelen.

In het geval van de toepassing van segmentatie en associatieve technieken wordt geen onderscheid gemaakt tussen afhankelijke en onafhankelijke variabelen. Het doel van de toepassing van interdependentietechnieken zoals deze ook genoemd worden, is het opsporen van relaties tussen variabelen om een mogelijk aanwezige, maar nog niet bekende structuur in de gegevens te ontdekken. Segmentatie-analyses worden aangewend om relaties te identificeren tussen gegevens ten einde groepen te kunnen onderscheiden. Een vaak gebruikte techniek hierbij is clusteranalyse, die tot doel heeft groepen (segmenten) te construeren op basis van kenmerken. Elementen die deel uitmaken van een groep zijn zo homogeen mogelijk ten aanzien van de beschrijvende kenmerken terwijl tussen de groepen een maximale heterogeniteit naar deze kenmerken bestaat. Clusteranalyse is o.m. een geschikte techniek voor het afbakenen van doelgroepen en het opstellen van klantprofielen.



Figuur 11 : Data mining taxonomie

Associatie- of affiniteitsanalyses worden gebruikt om te bepalen welke kenmerken of gebeurtenissen in samenhang voorkomen. Op grond hiervan worden regels opgesteld die deze samenhangen beschrijven. Dergelijke regelinductietechnieken worden veel toegepast in het geval van *market basket*-analyses, waarbij onderzocht wordt welke producten door klanten in samenhang worden afgenomen. Regelinductietechnieken worden ook gebruikt voor *cross-selling* analyses. Aan de hand van gegevens over historisch klantgedrag wordt bepaald welke combinaties van klantkenmerken en producten leiden tot interesses voor andere producten. Deze profielen kunnen o.m. gebruikt worden om tijdens klantcontacten gerichte aanbiedingen te doen.

Een belangrijke reden voor het succes van data mining ligt in het feit dat de patronen die erdoor aan de oppervlakte gebracht worden, voorspellingen toelaten in consumentengedrag. Veel toepassingen van data mining hebben dan ook betrekking op het ontwerpen van predictieve modellen. Dependentietechnieken hebben tot doel de invloed van één of meerdere onafhankelijke variabelen (predictoren) na te gaan op een afhankelijke variabele (criteriumvariabele). Met predicitieve modellen wordt beoogd ofwel het behoren tot een klasse (classificatie) ofwel een waarde te voorspellen (regressie). In het geval van classificatie is de klasse een categoriale variabele die uit twee of meerdere elkaar

uitsluitende categorieën bestaat. Scoringsmodellen die voorzien in de predictie van de respons op een mailing, voorspellen het behoren van prospecten of klanten tot de klasse “ja” of “nee”. Op analoge wijze kunnen klanten in kaart gebracht worden die de grootste kans hebben om binnen een bepaalde termijn te vertrekken (*attrition, churning*). Door de klanten die binnen dit profiel passen extra aandacht te geven kan het verloop teruggedrongen worden. Regressie wordt toegepast in het geval de te verklaren waarde (de afhankelijke variabele) een veelheid aan (numerieke) waarden kan aannemen (continue variabele). Een voorbeeld hiervan is het voorspellen van de beursnotering van aandelen.

3.3.4. Data mining-technieken

Sinds geruime tijd worden statistische technieken gebruikt om patronen in gegevensverzamelingen op te sporen. Veel gebruikte technieken zijn clusteranalyse, factoranalyse en regressie-analyse. Zoals vermeld, wordt clusteranalyse toegepast om segmenten of profielen te construeren aan de hand van kenmerken. Een voorbeeld hiervan is het opdelen van een klantenbestand in groepen en het beschrijven van de profielen van deze klantgroepen in termen van socio-demografische kenmerken, lifestyle-gegevens en aankooppatronen.

Ook factoranalyse is een exploratief-beschrijvende techniek om de dimensionaliteit in de data te analyseren. Het doel van factoranalyse is een hoeveelheid variabelen samen te vatten in een kleiner aantal onderliggende dimensies, die alle een lineaire combinatie zijn van de oorspronkelijke variabelen. Factoranalyse biedt o.m. de mogelijkheid om onderling sterk samenhangende variabelen te reduceren, hetgeen de interpretatie van de onderzoeksbevindingen ten goede komt.

Regressie-analyse is een predictieve techniek die gebruik maakt van het optimaliseringsprincipe dat bekend staat als de methode van de kleinste kwadraten. De bedoeling hiervan is de waarde van een continue variabele te voorspellen aan de hand van een lineaire combinatie van onafhankelijke variabelen. Het verschil tussen de verwachte en geobserveerde waarden van de afhankelijke variabele geldt als criterium voor de beoordeling van het regressiemodel. Een bezwaar dat aan de toepassing van regressie-analyse kleeft, is dat de gegevens die ermee gemodelleerd worden vaak niet voldoen aan de lineariteitsassumptie van de techniek. Het oplossen van de problemen die hiermee gepaard gaan, vereist statistische expertise. Bovendien blijken heel wat gegevens in de marketingpraktijk eerder van categoriale dan van parametrische aard te zijn. Voor het induceren van modellen uit grote gegevensverzamelingen worden om die reden technieken toegepast die voorzien in de niet-lineaire analyse van variabelen[11]. In hetgeen volgt geven we van deze laatste een overzicht.

3.3.4.1. Logistische regressie-analyse

Een tegenwoordig veel gebruikte classificatiemethode is logistische regressie. Deze techniek wordt toegepast voor het voorspellen van categoriale variabelen. Behalve de toepassing ervan bij responsanalyse, wordt de techniek van logistische regressie door bank- en verzekeringsinstellingen vaak aangewend voor het opstellen van modellen voor kredietacceptatie (*credit scoring*). De doelstelling hierbij is het optimaliseren van het percentage geaccepteerde aanvragen zodat het maximaal toegestane infectiepercentage niet wordt overschreden. Aan de hand van klantgegevens (geslacht, leeftijd, beroep, e.d.), het product (doorlopend krediet, persoonlijke lening) en gegevens van in het verleden verstrekte kredieten waarbij voor ieder krediet ook de afloop is vastgelegd, stelt het algoritme van logistische regressie

de gebruiker in staat om profielen te ontdekken met een sterk verlaagde of verhoogde kans op wanbetaling. De kansen op wanbetaling worden afgeleid uit het logistische model dat de vorm aanneemt van een regressievergelijking. Door de waarden van de variabelen die in het model opgenomen zijn, te wegen overeenkomstig de parameters van de regressievergelijking worden logitscores bekomen. De logaritmische transformatie van deze laatste laat toe logitscores te vertalen in waarschijnlijkheden. De resultaten van een logistische regressie-analyse kunnen teruggeschreven worden naar de prospecten- of klantendatabase om deze te verrijken, waardoor voor iedere prospect of klant op basis van zijn/haar karakteristieken een kans op wanbetaling wordt berekend.

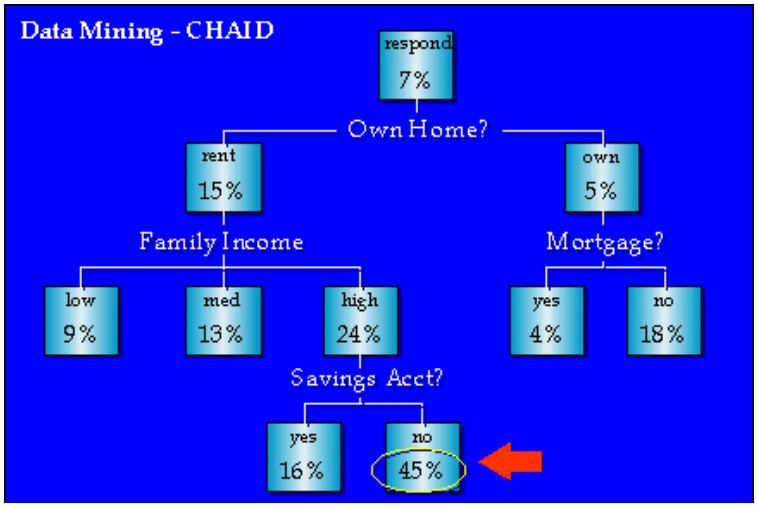
Een nadeel van logistische regressie is dat het opbouwen van een model sterk gebruikersgestuurd is. De kwaliteit van het uiteindelijke model hangt sterk af van de inhoudsdeskundigheid van de onderzoeker. Het is aan de onderzoeker om vast te stellen welke variabelen in aanmerking komen om in het model te worden opgenomen. Zoals dit het geval is bij lineaire regressie-analyse dienen interacties tussen variabelen door de gebruiker opgespoord en in de regressie-vergelijking opgenomen te worden. Een zwak punt is ook de beperkte schaalbaarheid van de techniek. Bij een toenemend aantal predictoren gaat de kwaliteit van de modellen achteruit en wordt het eveneens aan de gebruiker overgelaten om de problemen die daarmee gepaard gaan (bv. multicollineariteit) te ondervangen.

3.3.4.2. Beslissingsbomen

Veel minder gebruikersafhankelijk is het regelinductiemechanisme van beslissingsbomen. Een beslissingsboom is te vergelijken met een scoringsmodel waarbij de inputvariabelen (predictoren) gebruikt worden om waarnemingen te splitsen in verschillende groepen die discrimineren naar de te verklaren variabele (doelvariabele). Het gegevensbestand wordt opgedeeld in elkaar uitsluitende segmenten op basis van een doelvariabele. Deze segmentatie komt tot stand door na te gaan welke variabele en welke samenvoeging van waarden binnen deze variabele het hoogst mogelijk onderscheidend vermogen heeft t.a.v. de doelvariabele. Dit proces wordt iteratief toegepast totdat er geen onderverdeling meer te maken is die voldoende onderscheidend vermogen bezit.

In tegenstelling tot regressie-analyse, worden bij de toepassing van boomstructuren geen vooronderstellingen gemaakt over de functionele vorm van de relatie tussen de onafhankelijke variabelen en de afhankelijke variabele. Een voordeel van beslissingsbomen is dat de meest belangrijke variabelen en de combinaties (en interacties) ertussen voor de voorspelling van de doelvariabele gedetecteerd worden[12]. Op basis van beslissingsbomen worden regels opgesteld die aangeven aan welke voorwaarden een waarneming dient te voldoen om in een bepaald segment terecht te komen. Het segment geeft de voorspellende waarde aan voor de waarnemingen die daarbinnen vallen (te bedenken valt evenwel dat met boomstructuren enkel groepsvoorspellingen en geen individuele prognoses kunnen gemaakt worden). In figuur 12 is een gedeelte van een beslissingsboom afgebeeld. Een voorbeeld van een eenvoudige regel die uit de voorgestelde responsanalyse kan afgeleid worden, is dat terwijl de totale respons in het gegeven voorbeeld 7 % bedraagt, de respons in de gemerkte subgroep stijgt tot 45 %.

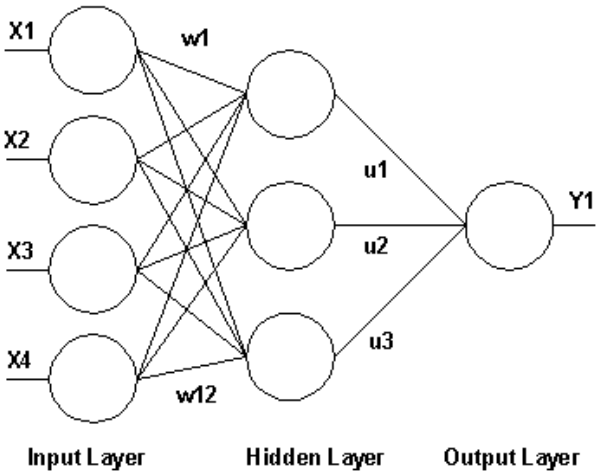
De visualisering van de resultaten en de hoge interpreteerbaarheid van beslissingsbomen hebben ertoe geleid dat deze techniek tegenwoordig veel gebruikt wordt. Zowel voor classificatie als voor regressie is de techniek van beslissingsbomen een aangewezen analysemiddel[13]. Maar ook als hulpmiddel voor het exploreren van gegevens bewijst deze techniek zijn nut, o.m. vanwege de mogelijkheid ervan interacties tussen de verklarende variabelen op te sporen om deze naderhand als input te gebruiken voor logistische regressie-analyse.



Figuur 12 : Beslissingsboom[14]

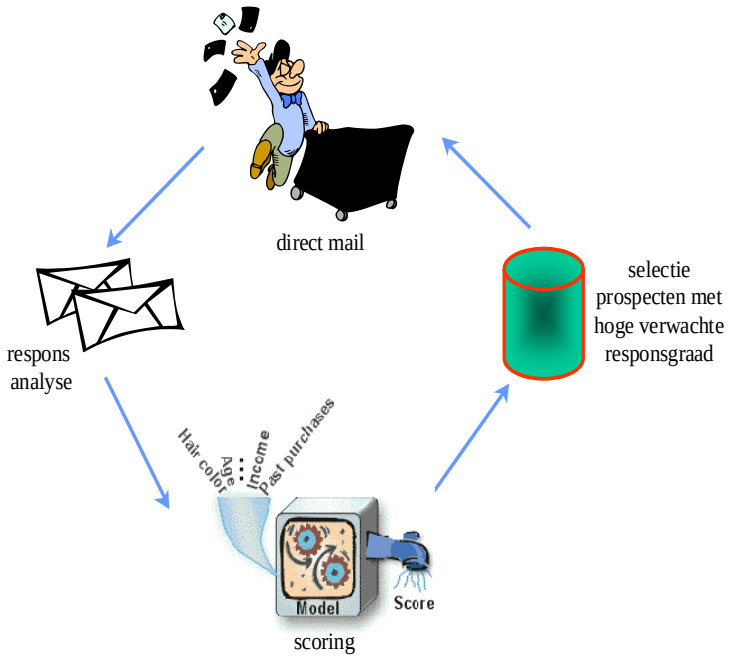
3.3.4.3. Neurale netwerken

Een techniek met een breed toepassingsgebied is neurale netwerken. Omdat neurale netwerken niet-lineaire relaties modelleren en hierbij een veelheid aan variabelen kunnen gebruiken, zijn de voorspellingen via deze methode in de regel accurater dan bij traditionele regressie-analyse. Een neuraal netwerk is een input-output model. De inputvariabelen zijn aan de outputvariabele gekoppeld door één of meer verborgen lagen (*hidden layers*). In elke laag bevinden zich neuronen. De neuronen staan in contact met elkaar via verbindingen. Deze hebben een bepaalde sterkte, ook gewicht genoemd. Het neuraal netwerk berekent voor een gegeven input een voorspelling van de afhankelijke variabele. Voor het trainen van het netwerk wordt de dataset gesplitst in een trainingset en een testset. Tijdens de eerste iteratie worden de waarnemingen uit de trainingset aan het netwerk aangeboden. Per waarneming wordt de netwerk-output vergeleken met de gewenste waarde. Op basis van het verschil worden de gewichten aangepast (*backpropagation*). Bij een volgende iteratie zal het netwerk een kleinere fout maken bij dezelfde input. De training stopt zodra de gemiddelde fout niet meer afneemt. Nadat het model voor de trainingset is ontwikkeld, wordt het verder gevalideerd door de toepassing van de modelparameters op de waarnemingen van de testset[15].



Figuur 13 : Neuraal netwerk

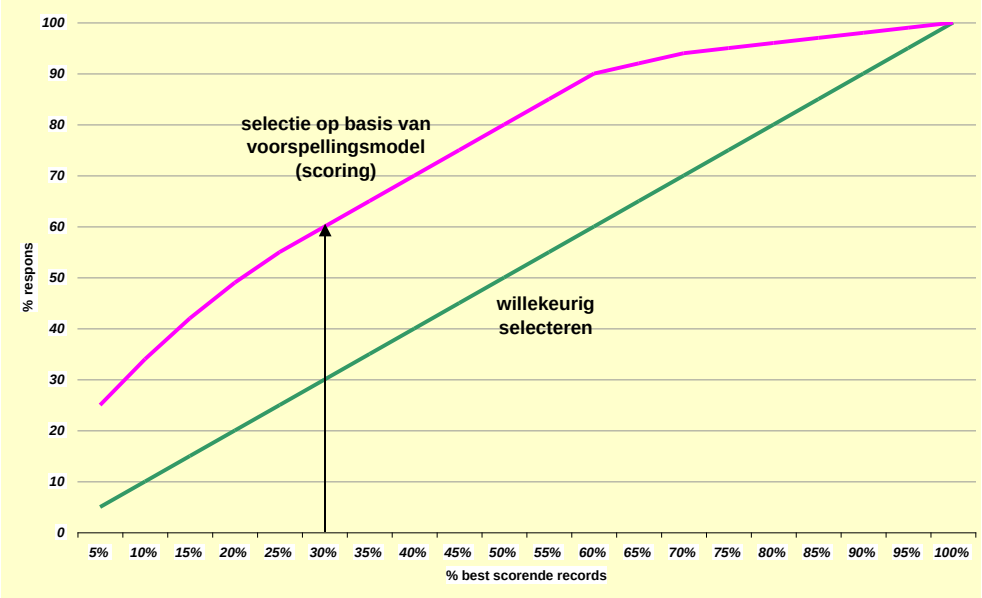
Nemen we als voorbeeld van de toepassing van neurale netwerken het scoren van de responsgraad op een direct mail-actie. Veelal is niet precies bekend welke karakteristieken van klanten van invloed zijn op het responsgedrag. Op basis van de variabelen die zich in de prospecten- of klantendatabase bevinden (klantkenmerken, productkenmerken, responsgegevens voorgaande mailing-acties) is het mogelijk de profielen van respondenten te vergelijken met de (non-) respons. Hierdoor komt een scoremodel tot stand dat de basis vormt voor volgende mailingselecties. Het doel ervan is respondenten te rangschikken in decielen op basis van hun kans op responderen, vertrekkend van de best scorende groep tot de slechtst scorende groep.



Figuur 14 : Respons-analyse en scoring

Aan de hand van de resultaten van de scoringsanalyse is het mogelijk de optimale mailingdiepte te berekenen, d.w.z. het aantal mailings dat aan de best responderende groepen toegestuurd dient te worden zonder daarbij onder het *break-even* niveau te zakken. Het opstellen van een scoringsmodel kan tot een aanzienlijke kostenbesparing leiden : door de respondenten in een mailing te betrekken waarvoor overeenkomstig het scoringsmodel een grotere respons kan verwacht worden, wordt een belangrijke kostenbesparing gerealiseerd. Dit blijkt duidelijk uit de "gains"-grafiek in figuur 15. De diagonaal in deze grafiek geeft de verwachte respons weer in het geval van een toevalsselectie uit de doelgroep. In dit geval neemt de respons lineair toe met de omvang van de doelgroep. De gebogen curve laat de verwachte responsgraad zien in het geval van scoring. Door toepassing van data mining kan door de selectie van de top 30 % meest kansrijke prospects reeds 60 % van het totaal haalbare aantal respondenten binnengehaald worden. Dit veronderstelt evenwel ook dat nadat een model opgesteld is, de database in staat moet zijn om de resultaten van het model te implementeren op prospecten en

klanten ten einde hieruit de geschikte respondenten te extraheren voor een mailingactie.



Figuur 15 : Gains-grafiek

De hier besproken technieken blijken onderling weinig te verschillen in piekprestatie. Verschillen tekenen zich wel af m.b.t. snelheid, gebruiksgemak en inzichtelijkheid.

Gesteld kan worden dat technieken die voorzien in de niet-lineaire analyse van variabelen, en met name technieken van regelinductie en neurale netwerken, doorgaans beter in staat zijn tot patroonherkenning als het gaat om interne bestanden met omvangrijke gegevens over klanten en de historiek van klantbestedingen.

Naarmate de transitie van *list-based* marketing naar *customer-based* marketing zich verder aftekent, zullen niet alleen klantgegevens (verzameling gegevens over klanten) maar vooral klantmodellen (weergaven van kenmerken en behoeften van klanten) centraal komen te staan. Voor het opbouwen en onderhouden van klantmodellen moeten analysetaken met een

hoge frequentie uitgevoerd worden. Het ligt daarom in de lijn van de verwachtingen dat een deel van de analysetaken zal overgelaten worden aan zelflerende, adaptieve technieken[16].

4. Besluiten

In marketing is een ontwikkeling gaande van een massabenadering naar een gesegmenteerde aanpak en verder naar een individuele benadering. De lange tijd dominerende *push*-benadering maakt meer en meer plaats voor een *pull*-benadering waarbij het initiatief niet langer bij de producent maar steeds meer bij de klant komt te liggen. De toenemende invloed van de klant en de intensiever wordende concurrentie eisen dat het aanbod nauwkeurig wordt afgestemd op de wensen en de verwachtingen van de klant. Bedrijfsprocessen dienen zodanig ingericht te worden dat deze zo effectief en zo efficiënt mogelijk gericht zijn op het opbouwen en onderhouden van duurzame relaties met klanten.

De ontwikkeling van een klantgerichte marketingstrategie (*customer relationship management,* CRM) is afhankelijk van de invulling van zowel een interactieve als informatieve dimensie[17]. De omslag van een aanbod- naar een vraaggeoriënteerde benadering impliceert in de eerste plaats dat bedrijven in staat moeten zijn om interactieve media aan te wenden in de communicatie met prospecten en klanten. Het principe van klantgerichtheid houdt in dat de onderneming elke klant individueel benadert en aan de individuele wensen van de consument tegemoet komt met een op maat gesneden aanbod. Dit is alleen mogelijk als de onderneming voldoende inzicht heeft in de voorkeuren van prospecten en klanten. Waar het in toenemende mate over gaat is kennis te verwerven in de zeer dynamische omgeving van de organisatie en deze om te buigen in een competitief voordeel.

De vaststelling dat klantcommunicatie in de toekomst grotendeels elektronisch zal geschieden en het gegeven dat het behoud van een competitief voordeel wel eens zou kunnen afhangen van het kleine verschil in markt- en klantinformatie, bewerkstelligen een toenemende vervlechting tussen informatietechnologie en bedrijfsprocessen. Wat dit laatste betreft, kunnen datawarehousing en data mining aangeduid worden als belangrijke ontwikkelingen ter ondersteuning van de beschikbaarheid van gegevens en de extractie van kennis aan deze laatste.
Mede door de invloed van het Internet en E-commerce die een stimulerend effect hebben op het genereren en opslaan van "clickstream data" zal het belang van data mining in de komende jaren nog toenemen. Tegenwoordig zijn de meeste producten voor data mining nog toegesneden op expertgebruikers, zoals statistici. De resultaten van data mining krijgen pas waarde als er zinvolle actie op ondernomen kan worden. Een belangrijke voorwaarde voor de acceptatie van data mining is het integreren van data mining in zgn. verticale applicaties, d.w.z. toepassingen gericht op een specifieke bedrijfsfunctie. Rekening houdend met het strategisch belang van data mining maar ook met de mogelijkheden om data mining direct toe te passen in het operationele domein, kan in de nabije toekomst verwacht worden dat de technologie van data mining in bedrijfsprocessen zal ingezet worden voor specifieke doelstellingen zoals lening- en polisacceptatie in de bank- en verzekeringswereld, basketanalyse en voorraadbeheer in retail marketing, loyaliteitsanalyses op basis van betaalgedrag in de telecommunicatie, e.a.. Data mining is dan niet langer een proces dat ad hoc wordt uitgevoerd, maar is één van de bouwstenen om *real-time marketing* uit te voeren aan de hand van de combinatie van historische data en gegevens van de *on-line* gebruiker. Om hieraan inhoud te geven, wordt niet alleen gewerkt met klantbestanden (verzamelingen gegevens over klanten) maar ook met klantmodellen. Deze laatste zijn een weergave van de kenmerken, attitutes en mogelijke behoeften van een klant en laten derhalve toe om marketingacties op één-op-één basis aan te sturen. Data mining is hierbij eveneens een hefboom in de ontwikkeling van een lerende organisatie : een hulpmiddel

om alle kennis die in een bedrijf aanwezig is gecontroleerd en efficiënt te gebruiken en om leerprocessen te versterken. Nauw samenhangend met de evolutie van database management naar kennismanagement[18], is het onderwerp van het beheer en de distributie van kennis binnen organisaties via intra/internet, *groupware* en *intelligent agents*.

De hier besproken ontwikkelingen blijven tenslotte ook niet zonder gevolgen voor het marktonderzoek. Met het in de plaats treden van klantconcepten voor productconcepten, zullen ook marktonderzoekers meer te maken krijgen met de analyse van klantgegevens en -modellen. Een hiermee verbonden ontwikkeling is dat het opstellen van klantprofielen en clusters naar aankoop-gedrag samen met analyses van de klantlevenscyclus zullen primeren boven het verzamelen van steekproefgegevens. Het type informatie verschuift van opgegeven informatie naar gedrags-informatie. Anderzijds vervangt informatie op individueel prospect- of klantniveau de meer generieke modellen en veronderstellingen. Klantsegmentaties zullen meer *bottum-up* gebeuren door samenvoeging van vrijwel identieke cases en minder *top-down* vanuit de opdeling van markten in homogene subgroepen. In hun geheel genomen hebben deze ontwikkelingen tot gevolg dat de marktonderzoeksfunctie in toenemende mate integreert met business intelligence en schuift het marktonderzoek op in de richting van informatiemanagement.

NOTEN

[1] Zie hierover Peppers & Rogers, 1993.

[2] Het is evenwel niet alleen de versnelde technologische vooruitgang die het ontwikkelingspad van de nieuwe media bepaalt. Ook de demassificatie van de consumentenmarkt drukt zijn stempel op de ontwikkeling van de nieuwe media. De door Faith Popcorn (1995) beschreven *cocooning* als een belangrijke trend in het hedendaagse consumentengedrag, vormt een gunstige voedingsbodem voor de groei van nieuwe interactieve media. Van 'nieuwe' media in de echte zin van het woord kan niet gesproken worden. De meeste nieuwe interactieve mediavormen zijn ontstaan als uitbouw of integratie van de bestaande communicatievormen : televisie, telefoon en computer. Vandaar dat men de nieuwe media ook wel multimedia noemt. Met nieuwe media wordt dus verwezen naar elk medium dat een directe, interactieve communicatie toelaat tussen de prospect of klant en de onderneming. Interactieve media worden in de eerste plaats gekenmerkt door informatieverstrekking op maat. De gebruiker gaat alleen in op datgene dat hem of haar interesseert. Een ander kenmerk van interactieve media is de vrijheid van timing. De gebruiker wordt niet op ongewenste momenten met een boodschap geconfronteerd. Het gebruik van het medium vindt plaats op het moment dat de gebruiker het zelf wil of op het moment dat daarvoor het meest geschikt is. Een interactief medium kan tenslotte de door gebruikers opgegeven input opslaan. Het medium wordt dan een krachtig instrument om meer te weten te komen over het gedrag van prospecten of klanten.

[3] Lester Wunderman die in de jaren zestig het begrip 'direct marketing' introduceerde als uitbreiding van het direct mail-concept, heeft het in dit opzicht over het gebruik van het wereldwijde internet (De Standaard, 3 mei 1996). De voortschrijdende ontwikkeling van direct marketing , o.m. via internet, is in de visie van Wunderman een veruitwendiging van een nieuw sociaal-economisch model, nl. dat van de informatiemaatschappij waar we opnieuw tot een producent-consument relatie komen zoals we die hadden vóór de industriële revolutie. De markten waren toen klein, zowel in omvang als in geografische uitgestrektheid. Dit maakte dat producenten en verkopers nauwkeurig de koopgewoonten van elke klant kenden. Aan dit op landbouw geënte economisch model kwam na 1840 een einde met de uitvinding van de machines. Die stelden de producent in staat op grote schaal te mechaniseren. Het gevolg was seriewerk, daling van de prijzen, en uiteindelijk de ontwikkeling van grootschalige distributiesystemen zoals supermarkten en winkelketens en de aanwending van de massamedia om het aanbod bij een groot publiek kenbaar te maken.

[4] Voor een heldere uiteenzetting over de betekenis van telematica voor bedrijfsprocessen verwijzen wij naar een bijdrage van R.M. De Wit (1996).

[5] Een uitbreiding van het sterschema is het zgn. sneeuwvlokschema. In een sneeuwvlokschema worden de dimensietabellen die deel uitmaken van een stermodel genormaliseerd. Doordat queries in het geval van een sneeuwvlokschema kunnen uitgevoerd worden op kleinere, genormaliseerde tabellen in plaats van grote ongenormaliseerde tabellen, wordt een betere query-performance bekomen dan dit het geval is met een sterschema (zie hierover Gill & Rao, 1996, hfst. 5).

[6] Zie hierover Blomme (1997).

[7] Zie hierover Adriaans, Knobbe & Van der Hulst, 1996.

[8] Belangrijke producenten van software voor data mining zoals SAS en SPSS ontwikkelen een eigen methodologie die als raamwerk fungeert voor toepassing van data mining-technieken. In het geval van SAS wordt het proces van kennisontdekking opgedeeld in een vijftal stappen die onderdeel uitmaken van de zgn. *SEMMA*-methodologie (*Sample, Explore, Modify, Model, Assess*). Op analoge wijze wordt het KDD-proces door SPSS omschreven met behulp van hetgeen de 5A-benadering wordt genoemd (*Assess, Access, Analyze, Act, Automate*). Verder werd in 1998 door een Special Interest Group van een honderdtal leveranciers en consultants CRISP-DM (Cross Industry Standard Process for Data Mining ; zie hierover www.crisp-dm.org) gelanceerd.

[9] Bron : Holsheimer, M. & Molenaar, C., Marketing and data mining produce a client profile, www.ddi.nl, 1998.

[10] Zie hierover Tukey (1977).

[11] Voor een uitgebreide beschrijving van data mining-technieken verwijzen wij naar Berry & Linoff (1997) en Brand & Gerritsen (1998).
[12] In het geval van onderling sterk samenhangende onafhankelijke variabelen wordt het vertakkingsproces wel instabiel.

[13] Beslissingsbomen die gebruikt worden voor de predictie van categoriale variabelen worden "classification trees" genoemd. In het geval de predictie betrekking heeft op een continue afhankelijke variabele wordt gesproken van "regression trees" (zie hierover Breiman e.a., 1984).

[14] Bron : www.spss.com.

[15] Het "black box"-karakter van neurale netwerken waardoor een verklaring voor de gevonden verbanden achterwege blijft, wordt vaak aangevoerd als een zwak punt van deze techniek.

[16] Zie hierover Den Uyl & Langendoen (1997).

[17] Klantrelaties kunnen derhalve vanuit twee kanten opgebouwd worden, nl. vanuit een mediakant en vanuit een informatiekant. Op de communicatiedimensie worden generieke (thema-)advertenties meer en meer verdrongen door interactieve communicaties. In samenhang hiermee zijn het individuele klantkenmerken die op de informatieve dimensie het startpunt van analyse vormen. Klantsegmenten komen tot stand door individuele klantkenmerken samen te voegen. Naar persoonlijke kenmerken samengestelde doelgroepen of niches bieden meer kansen op slagen voor marketingstrategieën (zie hierover o.m. Postma, 1996 en Molenaar, 1992).

[18] Zie hierover Parsaye, 1999.

GERAADPLEEGDE LITERATUUR

Adriaans, P., Knobbe, A. & Van Der Hulst, M.P., *Data mining and fuzzy databases*, Syllogic (Nl.), 1996.

Berry, M.J. & Linoff, G. , *Data mining techniques for marketing, sales and customer support*, New York, John Wiley & Sons, 1997.

Blomme, J., *Het relationele model en normalisatie,* Damme, 1997.

Brand, E. & Gerritsen, R., Data mining and knowledge discovery, *DBMS*, Vol. 11, nr. 9, 1998, pp. 52-55.

Breiman, L., Friedman, J., Olshen, R. & Stone, C., *Classification and regression trees*, Belmont, CA, Wadsworth, 1994.

Den Uyl, M.J. & Langendoen, E., De inzet van adaptieve analysetechnieken in direct marketing, in A.E. Bronner e.a. (red.), *Recente ontwikkelingen in marktonderzoek*, Jaarboek 1997 van de Nederlandse Vereniging voor Marktonderzoek en Informatiemanagement, Haarlem, Uitgeverij de Vrieseborch, 1997, pp. 107-121.

De Wit, R.M., *Het creëren van toekomst voor uw business met telematica*, Beam'IT Position Paper, 1996 (www.beamit.nl).

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., *From data mining to knowledge discovery in databases*, American Association for Artificial Intelligence, 1996.

Gill, H.S. & Rao, P.C., *De client/server gids voor data warehousing*, Schoonhoven, Academic Service, 1996.

Holsheimer, M. & Molenaar, C., *Marketing and data mining produce a client profile*, 1998 (www.ddi.nl).

Inmon, W.H., *Building the data warehouse*, New York, John Wiley & Sons, 1996.

Michielsen, T., Internet voert ons terug naar preïndustriële tijd, *De Standaard*, 3 mei 1996.

Molenaar, C., *Interactieve marketing. Het einde van de massamarketing*, Brussel, Management Bibliotheek, 1992.

Parsaye, K., From data management to pattern management, *DM review*, januari 1999.

Peppers, D. & Rogers, M., *The one-to-one future. Building business relationships one customer at a time*, London, Piatkus, 1993.

Popcorn, F., *Trends van overmorgen*, Amsterdam, Contact, 1995.

Postma, P., *Het nieuwe marketing tijdperk*, Amsterdam, Contact, 1996.

Schultz, D.E., (1990), *Strategic newpaper marketing*, I.N.M.A., Reston, Virginia.

Thearling, K., (1998), *Increasing customer value by integrating data mining and campaign management software*, Boston, Exchange Applications.

Tukey, J., *Exploratory data analysis*, Cambridge, MA, Addison Wesley, 1977.