Business Intelligence vanuit een tekstuele grondslag

Textual Analytics

Bill Inmon

Analytics bestaan al sinds het eerste computerprogramma werd geschreven. Begon een bedrijf eenmaal data te genereren, dan verschenen financiële analisten, verkoopanalisten, marketinganalisten etcetera, die stonden te trappelen om die data op een innovatieve en creatieve manier te gebruiken.

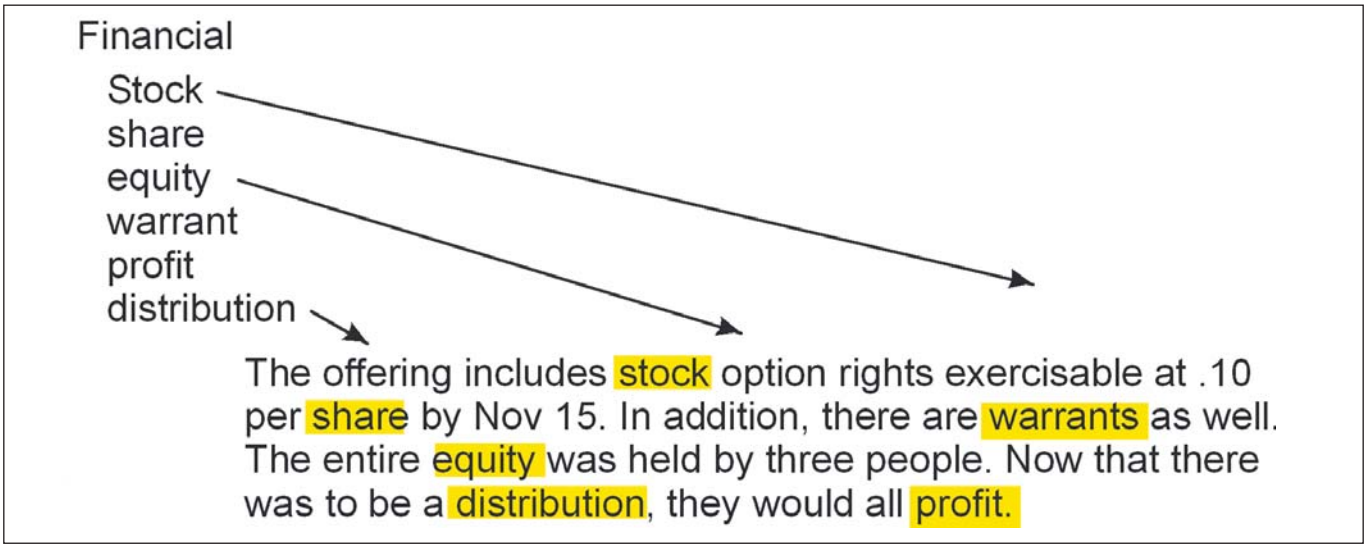
In het begin was het lastig om bij de data in applicaties te komen en de tools die de analisten gebruikten om de data te benaderen en te analyseren waren tamelijk primitief. Naarmate de tijd verstreek en de hoeveelheid data groeide, groeiden ook de toepassingsmogelijkheden van analytics als wapen in de concurrentiestrijd. Uiteindelijk werd het datawarehouse uitgevonden als fundament voor analytische processen. Het datawarehouse bevatte data die geïntegreerd, historisch en granulair waren, verzameld vanuit vele legacy-systemen. Het datawarehouse bewees een ideale basis voor de analyse van data te zijn. Data in het datawarehouse waren voorspelbaar en gemakkelijk te benaderen; omdat de data bovendien granulair waren, konden ze worden omgevormd voor vele doeleinden.

Fundamentele beperking

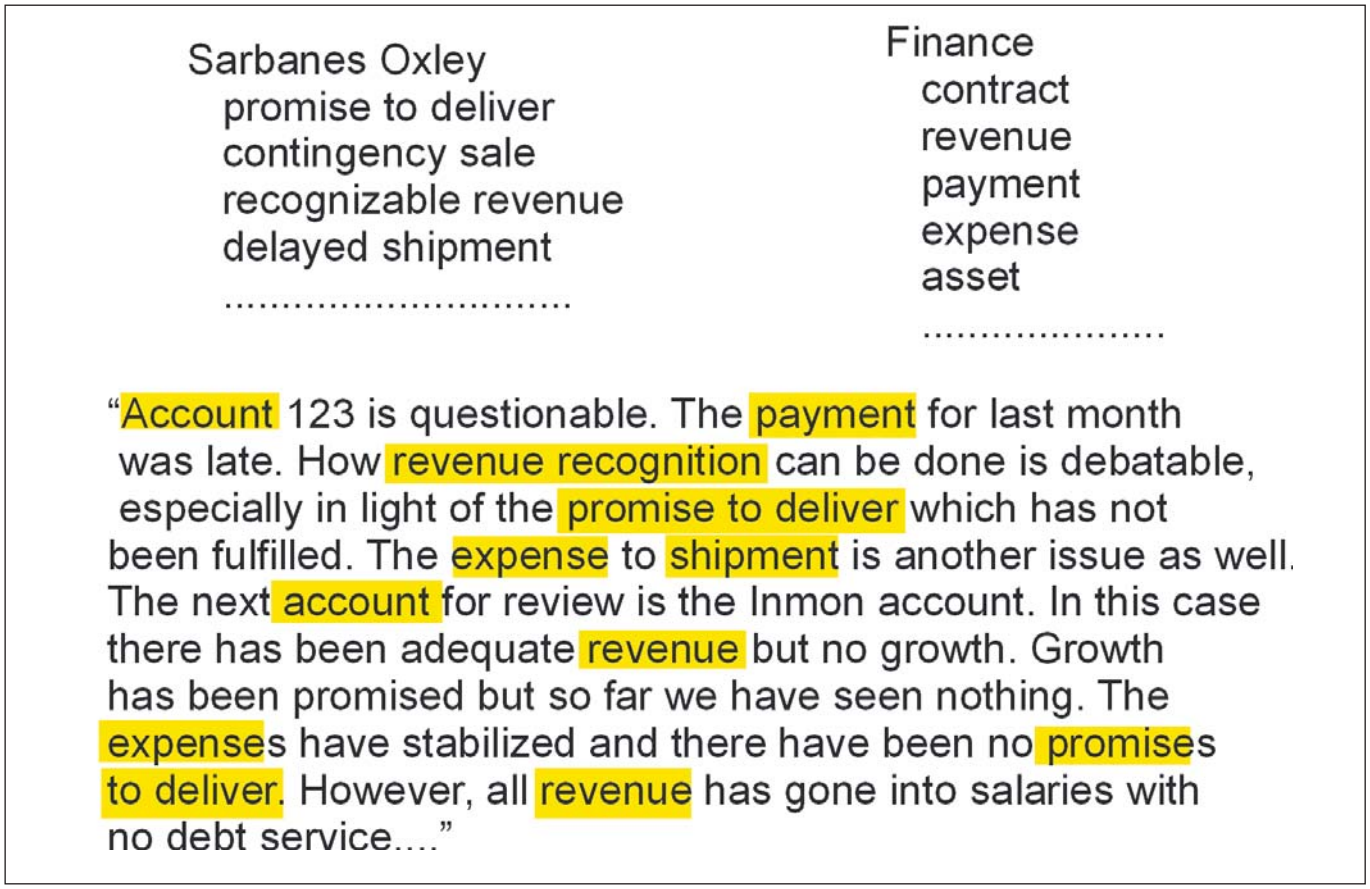
In de loop der tijd werd duidelijk dat aan analytics een fundamentele beperking kleefde, namelijk dat alleen numerieke data konden worden geanalyseerd. Dat was op zichzelf best nuttig, maar een bedrijf beschikt tegenwoordig over enorme hoeveelheden data die geen numerieke vorm hebben: ongestructureerde tekstuele data uit e-mails, medische rapporten, contracten, garanties, rapportages, call centers en noem maar op. De meeste schattingen geven aan dat 80 procent van de data binnen een bedrijf bestaat uit tekst en niet uit getallen.

Die ongestructureerde tekstuele data herbergen een schat aan informatie, maar zijn niet zo netjes geordend en toegankelijk als de numerieke data. Ze lenen zich dus niet zo gemakkelijk voor analyse, omdat de in gebruik zijnde software en technologie bijna helemaal zijn afgestemd op het werken met numerieke gestructureerde data. De grote mate van ongeordendheid van de tekstuele data maakt het vrijwel onmogelijk om de data op een zinvolle manier te benaderen en te analyseren. Inmon Data Systems (IDS) ontwikkelde een nieuwe technologie die gemaakt is voor tekstuele analyse.

Als het onderwerp 'tekstuele analyse' ter tafel komt, is het logisch dat de gedachten direct uitgaan naar search engines zoals bijvoorbeeld Google en Yahoo. Het uitvoeren van eenvoudige zoekopdrachten op ruwe tekst kan worden beschouwd als een primitieve vorm van tekstuele analytics, want er zitten vele beperkingen aan vast.



Afbeelding 1: Query op dataklasse.



Afbeelding 2: Proximity-analyse.

Om tot échte Textual Analytics te komen, moet de ongestructureerde tekst eerst worden geïntegreerd. Als dat niet gebeurd is, zullen de analyses slechts oppervlakkige en twijfelachtige resultaten geven. De eerste stap is dus de integratie van de ruwe tekst:

- er moet rekening gehouden worden met het gebruik van verschillende soorten terminologie om tot consistente resultaten te komen, zelfs als de originele bronteksten verschillend zijn;
- er moet rekening gehouden worden met alternatieve spellingvormen, zelfs met algemeen voorkomende foutieve spellingen;
- woorden moeten worden teruggebracht naar hun Latijnse of Griekse stam;
- enzovoort.

Een search vindt dus plaats op ruwe tekstuele data, analytics kunnen alleen worden toegepast op geïntegreerde data. Een search kan iets eenvoudigs zijn zoals "Vertel me waar de term 'Katherine Heigl' voorkomt". In dat geval gaat de zoekopdracht naar de brondocumenten of een index daarvan en kijkt of en zo ja waar de gevraagde term voorkomt. Een analytische zoekopdracht kan bijvoorbeeld zijn "Geef me alle plaatsen waar termen en informatie die verband houden met Sarbanes Oxley voorkomen". Soms ligt de noodzaak tot integratie van de data helemaal niet zo voor de hand. Bij een medisch dossier bijvoorbeeld. Stel, er moet een Engelstalig medisch dossier worden geanalyseerd. In dat dossier komt de term 'ha' voor. Wordt er nu een search gedaan op de ruwe data, dan levert 'ha' een groot aantal hits op. Maar 'ha' betekent voor een leek weinig of niets. Die search is dus twijfelachtig. Nu worden de data geïntegreerd. Dat maakt het mogelijk de term 'ha' voor alle cardiologen te vertalen naar 'heart attack', voor alle internisten naar 'hepatitis A' en voor verpleegkundigen naar 'head ache'. Door die conversieslag komen er geen vage termen als 'ha' meer voor, en worden patiënten met hartaanvallen, hoofdpijn en Hepatitis A niet meer op één hoop gegooid. Uit dit eenvoudige voorbeeld blijkt dat de integratie van tekst de mogelijkheden tot analyse ontsluit. Maar dat is natuurlijk niet de enige reden om data te integreren.

Zoeken naar tekstcategorieën

Stel, er is een aantal teksten over veehouderij. Een gedeelte van die teksten handelt over paarden. In een aantal gevallen wordt het soort paard besproken, elders wordt ingegaan op de leeftijd en volwassenheid van het dier, een andere passage betreft het geslacht.

Er bestaat de wens om deze teksten te analyseren, vooral waar het om paarden gaat. Met een search-opdracht kan eerst gezocht worden naar veulens, dan naar pony's, dan naar merries en hengsten. De zoeker moet van te voren precies weten wat hij wil zoeken. Vervolgens moet hij alle gevonden informatie samenbundelen. Zoeken naar een breed spectrum aan informatie is op die manier monnikenwerk.

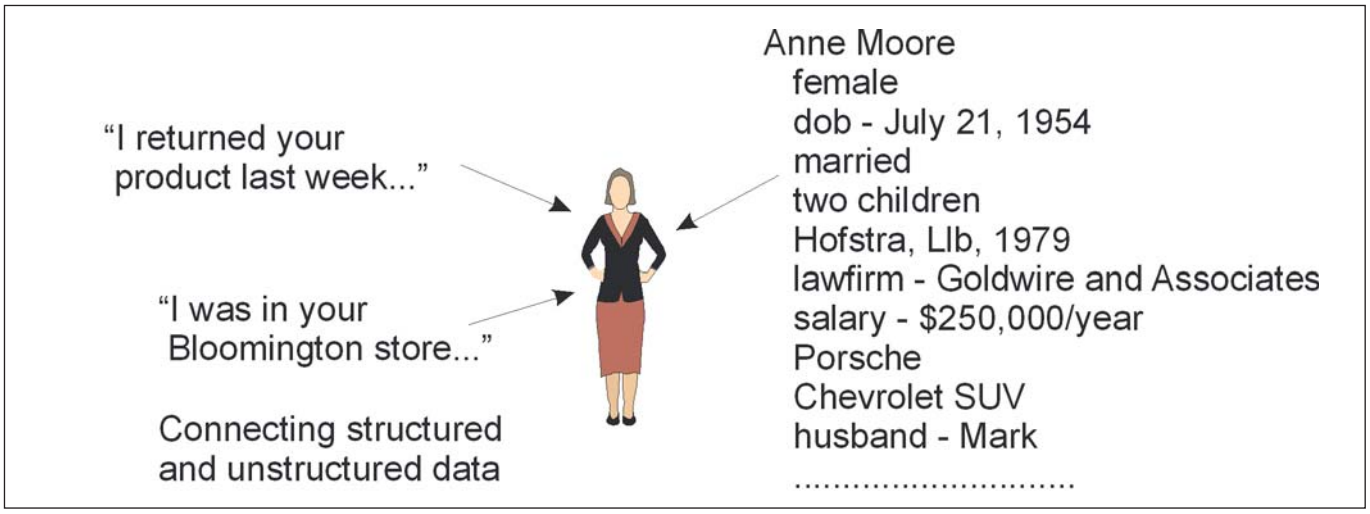
De benadering met geïntegreerde tekst is een betere. Hierbij wordt alle informatie over paarden in een categorie ondergebracht. Het integratieproces identificeert alle passages waarin stukjes informatie over paarden voorkomt. Het resultaat is een overzicht dat bijvoorbeeld ook verwijzingen bevat naar: palomino; dekhengst; merrie; teugels; zadel; hooisoorten; hekken; 'horse whispering'; stallen; races; enzovoort.
Als de tekstuele analist informatie wil hebben over paarden, hoeft hij alleen maar de categorie 'paarden' in te voeren: wat een verschil met de zoekopdracht op ruwe tekst.

Omwerken van tekstuele data

Er zijn vele vormen van tekstuele integratie. Een belangrijke component is de verschillende spelling van woorden, of bijvoorbeeld namen. Een naam als 'Osama bin Laden' kan ook gespeld worden als 'Usama Ben Ladeen' en alle mogelijke tussenvormen. Door bij de integratie rekening te houden met de verschillende mogelijke spellingen, mist de tekstanalist geen enkele vermelding, iets dat bij eenvoudige text search wel het geval zal zijn.
Een andere manier om tekstuele data te integreren is door gebruik te maken van de Latijnse of Griekse woordstammen. Woorden die uit het Latijn afstammen komen in verschillende vormen voor, als vervoegingen en verbuigingen. Een eenvoudige zoekopdracht zal geen relatie leggen tussen 'move' en 'moving', 'moved', 'mover', 'moves', 'remove', 'removed' etcetera. Om tot een effectieve tekstanalyse te komen moet worden vastgesteld dat al deze woorden de woordstam 'mov' gemeen hebben.
Er zijn vele andere mogelijkheden denkbaar, zoals het beoordelen of tekst al dan niet relevant is voor de business, het weghalen van tussenwoordjes – bijvoorbeeld voorzetsels, voegwoorden – en leestekens, hoofdletter(on)gevoeligheid, enzovoort.

Scope van search en analyse

Een van de uitdagingen van een search engine is de scope van het benaderde materiaal en analyse door de query. Een search engine kan zeer grote hoeveelheden bronmateriaal aan, zoals bijvoorbeeld internet. Een tool voor tekstanalyse kan alleen data benaderen die tevoren zijn bewerkt en geïntegreerd – een aanzienlijk kleinere scope dus. Kan een search engine alle data ter wereld aan, de Textual Analytics vinden alleen plaats op data binnen de invloedssfeer van de onderneming. Een search engine kan immers data eenvoudigweg niet integreren of anderszins bewerken, omdat het geen eigenaar is van de brondata.
Naast de standaard search query's die de tekstuele analist zal uitvoeren, zijn er heel wat aanvullende typen query's denkbaar. Een van de eenvoudigste is de query op dataklasse.
In afbeelding 1 heeft de tekstuele analist een query gedaan op de categorie 'financial information'. Deze query omvat heel wat verschillende termen, die elk verband houden met financiën, zoals bijvoorbeeld stock, share, equity, warrant, profit, enzovoort. Als resultaat wordt aangegeven waar de verschillende begrippen in de bevraagde teksten voorkomen. Dit type query wordt soms een indirecte query of categorie-query genoemd.
Een belangrijk query-type is er een die zoekt naar elementaire aanwezigheid van informatie, bijvoorbeeld het woord 'water', of beter 'water' als lettercombinatie. In verschillende teksten wordt dan verwezen naar bijvoorbeeld 'waterstand', 'zeewater', 'bewateren', 'waterig' en 'Waterford'. Nadat deze query is uitgevoerd kan worden gezocht naar tekst voorafgaand of juist volgend op 'water'. Deze tekstuele referenties aan 'water' samen met hun omringende tekst, worden *snippets* genoemd. Aan de hand van elke snippet kan de analist de context gaan bepalen van het gezochte woord. Snippets zijn daar bijzonder handig voor.
Een ander type query is de nabijheids- of proximity search. In een proximity-analyse wordt gezocht naar woorden die in nabijheid van elkaar in de tekst voorkomen. Een proximity query wordt bijvoorbeeld toegepast als in twee of meer documenten moet worden gekeken of twee of meer woorden binnen een tevoren bepaalde afstand van elkaar voorkomen, denk bijvoorbeeld aan 'equity' en 'shares'.
Natuurlijk kan proximity-analyse gedaan worden met zowel lijsten van woorden als met individuele woorden. Afbeelding 2 laat een



Afbeelding 3: Tekstuele data gerelateerd aan gestructureerde data.

dergelijke query zien. Er zijn twee woordenlijsten: één voor termen die verband houden met de Sarbanes Oxley Act en een ander die gerelateerd is aan financiën. De analytische proximity search wordt gedaan aan de hand van woorden uit beide lijsten.

Een andere vorm van tekstuele analyse is die waarbij tekstuele data worden gerelateerd aan gestructureerde data. Afbeelding 3 geeft daarvan een voorbeeld, de demografische data van de klant worden gekoppeld aan de communicatie met de klant. De e-mails die de klant heeft gestuurd worden getoond als attachments aan de klant. Door het samenvoegen van tekstuele informatie en gestructureerde informatie, wordt daadwerkelijk een totaalblik op de klant bereikt.

Textual Visualisation

Een bijzonder waardevolle vorm van tekstuele analyse is die van het visualiseren van tekst. Hierbij wordt geïntegreerde tekst als het ware 'verteerd' en geclusterd om correlaties te vinden, en relaties tussen woorden en zinnen. In afbeelding 4 wordt getoond dat de tekst van een aantal documenten is benaderd en gecombineerd. De tekst uit de documenten is geïntegreerd en vervolgens in een werkgebied getild, waar de tekst geclusterd is in wat aangeduid kan worden als 'thema's'. De thema's worden vervolgens weergegeven zoals gevisualiseerd in afbeelding 4. Deze vorm van visualiseren wordt een SOM genoemd, een Self Organizing Map. De hier getoonde SOM is gebouwd door Raptor International gebaseerd op tekstuele data geïntegreerd met IDS.

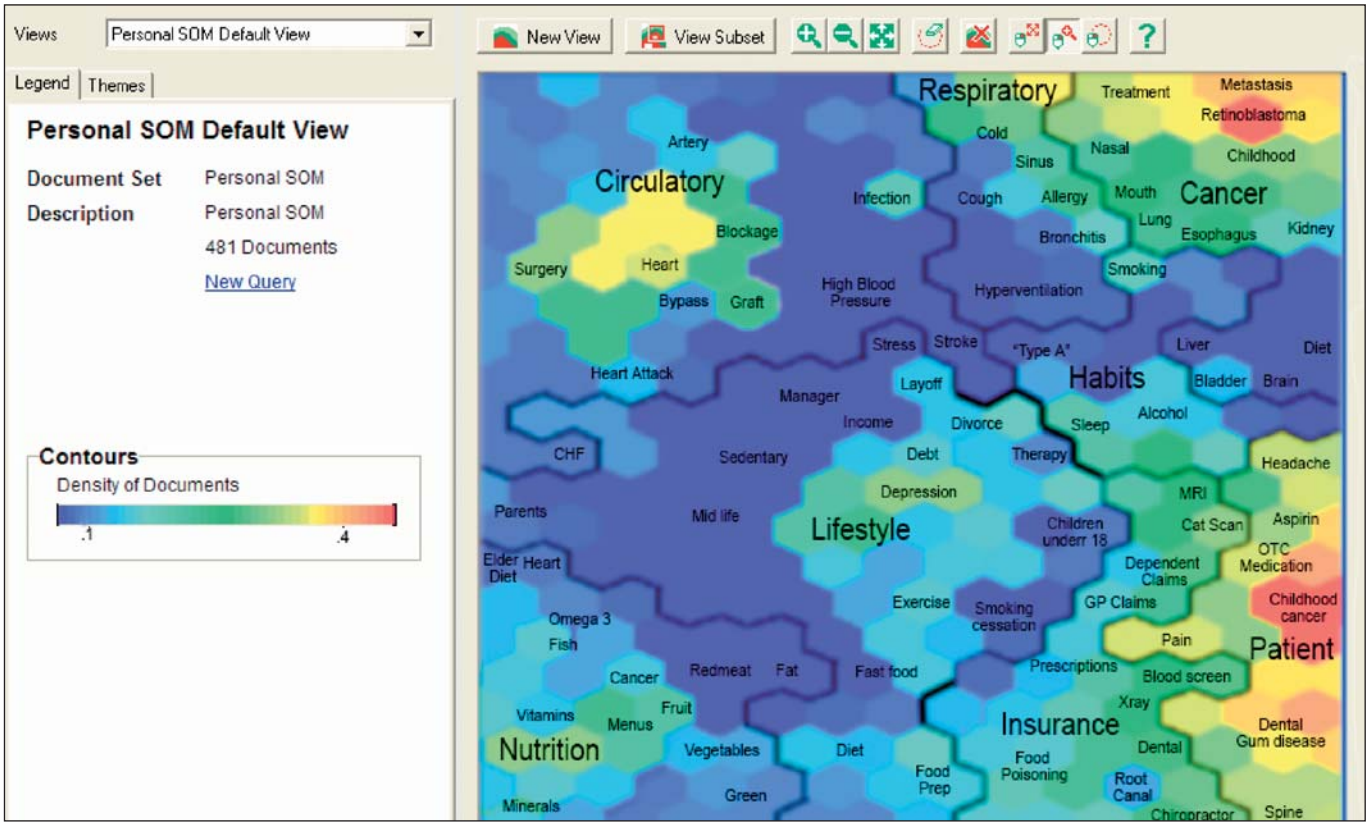
De clustering van data in een SOM heeft veel toepassingen, zoals bijvoorbeeld:

- identificeren van correlaties;
- identificeren van de belangrijkste thema's;
- het zo organiseren van data dat de belangrijkste thema's duidelijk worden.

SOM's kunnen zowel worden gemaakt uit heel grote of juist kleinere hoeveelheden data, en bieden uitzicht op uitgestrekte hoeveelheden informatie – soms 1000 documenten tegelijk.
Het zal duidelijk zijn dat Textual Analytics een geheel ander onderwerp is dan search engine processing.

Overbruggen van de kloof

Een van de bepalende factoren bij het bouwen van een effectieve omgeving voor Textual Analytics, is om ongestructureerde data te benaderen in een gestructureerd formaat. Met andere woorden, als u Business Objects of Cognos wilt gebruiken voor ongestructureerde tekst, dan moeten de data worden aangeboden in een vorm waarmee BO of Cognos iets kan. Dat betekent dat de ongestructureerde data – nadat ze zijn geïntegreerd – moeten worden geherstructureerd in een relationeel formaat. Er bestaat dus de noodzaak om tekstuele informatie in een gestructureerd formaat te plaatsen, waar ze herkenbare relationele velden zijn in een voorspelbaar format.
Afbeelding 5 laat zien dat er aantekeningen bestaan over een doktersbezoek. De notities zijn gemaakt in een format dat handig is voor de arts. Daarna worden ze ingelezen in de software die de



Afbeelding 4: Self Organizing Map.

aantekeningen omzet in een format en structuur die nodig zijn in de wereld van analytic processing, oftewel een relationeel formaat. Enkele kolommen in de relationele tabellen zijn: Patiëntnaam; Datum bezoek; Naam van de arts; Medicatie; Dosering.

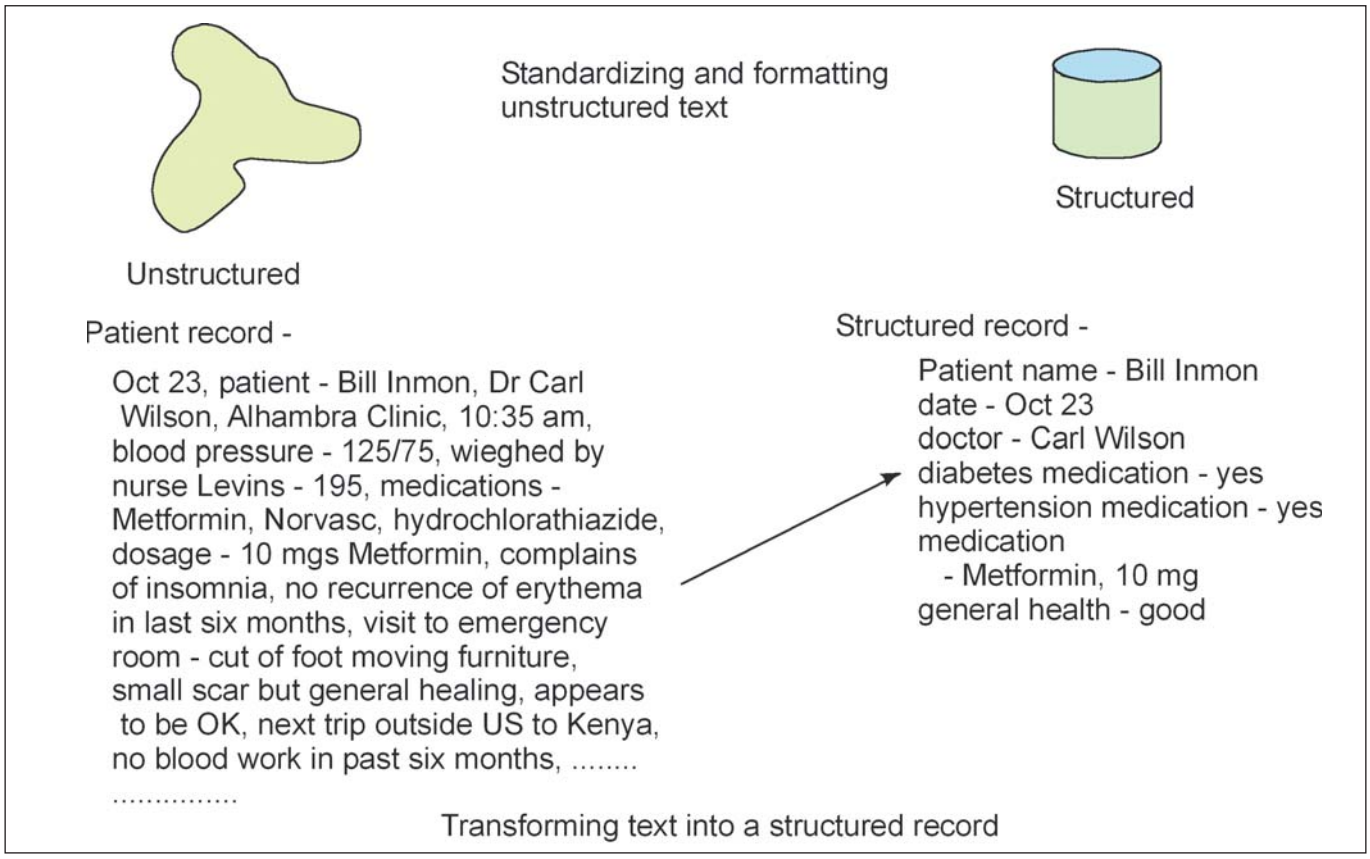
Nu de ongestructureerde data zijn omgezet in een relationeel formaat, kunnen de standaard analytische tools worden toegepast. Er zijn nog enkele belangrijke finesses die niet direct in de afbeelding te zien zijn. Zoals: wat gebeurt er als meer dan één record geconverteerd wordt naar het relationeel formaat?
Afbeelding 6 laat zien dat het medicijn Metformin is voorgeschreven aan Carol Teal. Maar in het ongestructureerde record van deze mevrouw komt dat medicijn niet voor. In plaats daarvan gebruikt zij Glucotrol. De software heeft – onder de hoede van de analist – Glucotrol vertaald naar Metformin, als onderdeel van het transformatieproces. De mogelijkheid om tekst te herkennen en te vertalen is een belangrijke factor bij het prepareren van data voor tekstuele analyse.
Bovendien heeft de analist bepaalde generalisaties of categorisaties op de ruwe tekst gespecificeerd. Zo kan geanalyseerd worden dat een bepaalde patiënt wordt behandeld voor Diabetes II, gebaseerd op de aangetroffen data. Door het vertalen en classificeren van gegevens en ze vervolgens omzetten in een relationeel formaat, kan de eindgebruiker zijn analytische werk op de tekst doen.

Geïntegreerde tekstuele data in een relationele database

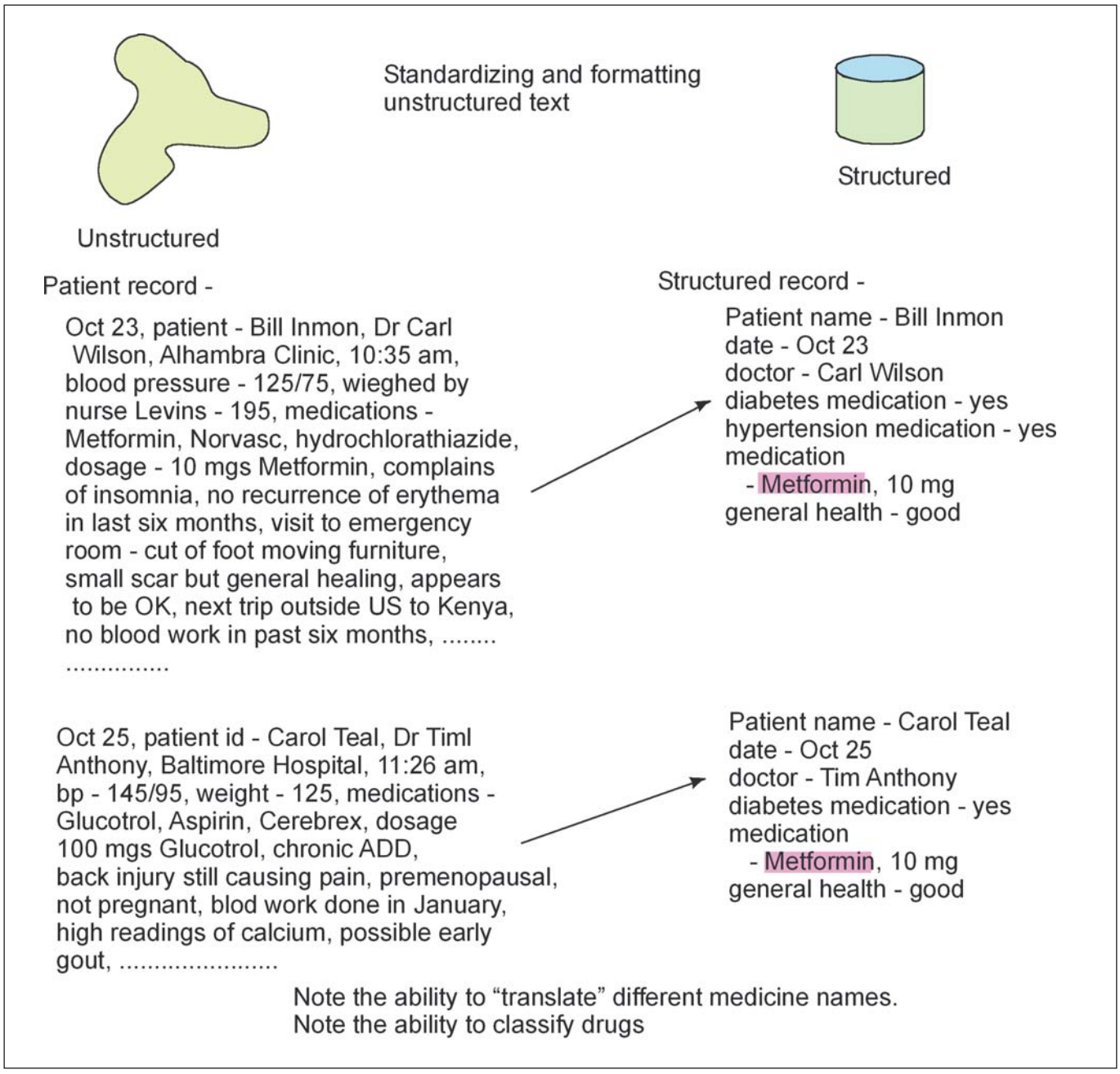
Stel, er is een relationele database die bestaat uit ongestructureerde data, die vervolgens zijn geïntegreerd. De database heeft een relationeel formaat en biedt via standaard SQL toegang aan de bekende analytische tools zoals Business Objects, Cognos en MicroStrategy. Er zijn enkele basismanieren om de data te benaderen:

- Een eenvoudige query. Een woord of zin wordt aan de software gegeven en de database wordt onderzocht. Neem bijvoorbeeld 'water'. Een search van dit type vindt elke lettercombinatie w-a-t-e-r;
- Een eenvoudige query op de context van het woord, een snippet. Dit type search verzamelt de tekst voor en na 'water' om inzicht te krijgen in de context;
- Een indirecte search. Een search die zoekt naar termen die behoren tot een bepaalde dataklasse of datacategorie;
- Proximity search. Bevinden de twee woorden 'water' en 'televisie' zich in bepaalde documenten binnen een afstand van 200 bytes van elkaar?;
- Alternatieve spelling search. Zoek teksten waar 'Osama bin Laden' wordt genoemd in alle mogelijke spellingvarianten.

Dit zijn de de meestvoorkomende vormen van analyse die gedaan kunnen worden op ongestructureerde data die zich in een relationele database bevinden. Textual Analytics kunnen worden



Afbeelding 5: Omzetting tekst in gestructureerd formaat.



Afbeelding 6: Vertaling en specificeren.

uitgevoerd door enorme hoeveelheden documenten te doorzoeken, of juist op één enkel document. Ze kunnen zo eenvoudig zijn als het zoeken naar één woord, of gecompliceerd als het zoeken naar categorieën van woorden en zinnen, of de context van woorden.

De meerwaarde

Hoe dragen deze vormen van Textual Analytics bij aan de business? Het algemene antwoord is dat een infrastructuur van geïntegreerde ongestructureerde data, in een database gezet en benaderd met analytische tools, de onderneming voordelen biedt die het tevoren niet had. Managers kunnen eindelijk antwoorden krijgen op vragen die ze voorheen niet voor mogelijk hielden.

Een gespecificeerd antwoord voor elke individuele onderneming, instelling of organisatie is hier vanzelfsprekend niet mogelijk, maar aan de hand van al het bovenstaande niet moeilijk te bedenken.

Noot

Dit artikel is een bewerkte en vertaalde versie. In geval van discussies geeft de originele Engelstalige tekst van het artikel de doorslag.
Deze tekst is te vinden op onze website www.dbm.nl in het hoofdmenu onder Specials/Extra materiaal.

Bill Inmon

William H. Inmon (binmon@inmondatasystems.com) is oprichter en CEO van Inmon Data Systems, gevestigd in Castle Rock, Colorado.