Hybride classificatie maakt ongestructureerde informatie beheersbaar

Overload wordt onderscheid

Organisaties hebben te maken met een snelle groei van de hoeveelheid ongestructureerde informatie. Gevolg is een verminderde productiviteit omdat medewerkers worden gedwongen om door een groeiend aantal documenten te zoeken voordat ze de benodigde informatie vinden. Ongestructureerde informatie blijft dus door de organisatie zweven en de hierin opgeslagen kennis wordt niet benut omdat niemand weet waar ze te vinden is. Iets wat concurrentievoordeel moet opleveren, wordt al met al een zware last. Hybride classificatie biedt hier uitkomst.

Organisaties moeten hun 'overload' aan informatie zien te veranderen in een concurrentievoordeel. Een manier om dit te bereiken is het onderbrengen van informatie in categorieën (classificeren). Documenten moeten daarbij worden geordend in categorieën waarmee gebruikers intuïtief kunnen navigeren naar individuele bestanden. Hierdoor zijn medewerkers minder tijd kwijt met het zoeken naar informatie, en kunnen ze meer tijd besteden aan het gebruik van informatie (ongeacht de hoeveelheid informatie of hoe snel het groeit). Ondernemingen die zich niet bezighouden met classificatie zullen onvermijdelijk de nadelen hiervan onder ogen zien; werknemers moeten langer zoeken naar documenten en zullen vaker werken met verouderde data. Het is de paradox van het informatietijdperk: hoe rijker een onderneming wordt in termen van informatie-assets, hoe meer nadelen er ontstaan door het niet managen van deze assets. Dit risico kan worden geëlimineerd door het inzetten van classificatiemethoden.

Typen classificatie

Niet alle classificatiemethoden kunnen goed uit de voeten met structuren zoals de onderneming die vereist. Sommige automatische methoden creëren eigen categorieën en taxonomieën die zijn gebaseerd op vastgestelde algoritmen, terwijl volledig handmatige systemen weer arbeidsintensief en subjectief zijn. Intelligente classificatie biedt een balans tussen deze twee uitersten doordat zij verder gaat dan het organiseren van informatie in categorieën alleen. Computers doen het werk, maar mensen die de business kennen, voeren het beheer.

Classificatiesystemen worden in drie typen verdeeld: handmatige systemen, zelflerende systemen en gecontroleerd lerende systemen. Elk heeft zijn eigen voor- en nadelen (zie ook afbeelding). Handmatige systemen staan aan het ene uiterste van het spectrum. Dit systeem vertrouwt uitsluitend op mensen om informatie te classificeren. Aan de andere kant staan zelflerende systemen, waarbij naast de documenten die moeten worden geclassificeerd (verder) geen menselijk input nodig is. In het midden staan de gecontroleerd lerende systemen, die het menselijke intellect combineren met de machinale efficiëntie.

Handmatige classificatiesystemen. Handmatige systemen vertrouwen op experts of ontologen om elk document in het systeem te inspecteren en handmatig te classificeren. Handmatige classificatie is aantrekkelijk omdat de experts betekenisvolle categorieën kunnen creëren en een hoge nauwkeurigheid kunnen bereiken. Zij kunnen documenten beter in categorieën indelen dan een machine, omdat zij de inhoud, context en nuances van de bedrijfsinformatie begrijpen. Bovendien kunnen zij categorieën creëren en verwijderen als de business dat voorschrijft. Hoewel handmatige classificatie een hoog niveau van nauwkeurigheid biedt, bestaat er een aantal nadelen. De arbeidskosten voor het bouwen en onderhouden van handmatige systemen zijn hoog, evenals het potentieel aan menselijke fouten wanneer meer dan één persoon bij het proces betrokken is.

Zelflerende systemen. Om de kosten en tijd van de handmatige classificatiemethoden terug te brengen, besluiten veel ondernemingen over te gaan op zelflerende systemen. Daarbij onderzoekt software alle documen-

ten in het systeem en maakt zogenaamde 'educated guesses' over de wijze waarop documenten moeten worden geordend. Om deze ordening vorm te geven, maakt het systeem gebruik van statistische algoritmen zonder enige input van menselijke experts. Het grootste voordeel van het zelflerende systeem is de verminderde implementatie-inspanning. In theorie kan de software worden geïnstalleerd en gericht op een aantal bestanden, die vervolgens worden onderscheiden in clusters van gerelateerde documenten. Menselijke fouten zijn te voorkomen door automatische classificatie te baseren op een wetenschappelijke methode. En wanneer de eerste clustering is afgerond, heb je er weinig omkijken meer naar. In de praktijk worden deze voordelen veelal weer teniet gedaan. Besparingen op de implementatie wegen vaak niet op tegen de hogere beheerskosten en lagere benuttingsgraad van de geclassificeerde informatie. Omdat documenten zijn georganiseerd op manieren die logisch zijn voor het algoritme van het systeem - en dus niet voor de mensen die de informatie maken en gebruiken - wordt het nog moeilijker om ze terug te vinden.

	Classificatiemethode		
Kenmerk	**Handmatig**	**Zelflerend**	**Hybride**
Eigenschappen			
Nauwkeurigheid	++	+/-	++
Consistentie	+/-	+	++
Tijd om nieuwe documenten te verwerken	-	++	++
Betekenisvolle categorieën	++	-	++
Diepte van de taxonomie	++	-	++
Mogelijkheid onderliggende classificatieprocessen te begrijpen	++	-	++
Mogelijkheid om op maat te maken of aan te passen	++	-	++
Mogelijkheid om aan te passen aan dynamische businessprocessen	+	-	++
Gemak en kosten-effectiviteit van modificatie, groot of klein	+/-	-	++
Totale flexibiliteit van het classificatiesysteem	++	-	++
Implementatie en onderhoud			
Implementatietijd en arbeid	H	L	M
Speciale eisen voor implementatie	L	H	L
Vereiste technische achtergrond voor onderhoud	L	H	M
Hoeveelheid vereist onderhoud	M	H	L
Onderhoudsgemak	M	H	L
Kosten voor implementatie en onderhoud op lange termijn	H	M	L
Tijd voor investering, implementatie en onderhoud op lange termijn	H	M	L
Tevredenheid			
Tevredenheid gebruiker	H	L	H
Tevredenheid beheerder	M	L	H

++	= Uitstekend
+	= Goed
+/-	= Matig
-	= Slecht

H	= Hoog
M	= Medium
L	= Laag

Gecontroleerd lerende systemen. Gecontroleerd lerende systemen verschillen van zelflerende systemen in het feit dat zij input van mensen nodig hebben om hen te 'leren' hoe ze informatie moeten classificeren. Dit classificeren van een verzameling documenten begint met het verzamelen van trainingsdocumenten voor de afzonderlijke categorieën. Net zoals bij zelflerende systemen zijn de lagere opzet-, arbeids- en onderhoudskosten de belangrijkste voordelen. Bovendien is het potentieel aan menselijke fouten, zoals bij het handmatige classificatiesysteem, geminimaliseerd. Het grote verschil tussen de twee lerende systemen is dat het gecontroleerd lerende systeem het voor experts mogelijk maakt om categorieën te creëren en ook enigszins te beïnvloeden welke documenten daarin worden geplaatst. Informatie wordt geclassificeerd rondom de behoeften van de business.

Leren leren

Er zijn verschillende manieren waarop lerende systemen 'leren'. Hiermee varieert ook de nauwkeurigheid waarmee de documenten worden geclassificeerd. Er zijn drie typen die het

meest voorkomen, namelijk statistische analyse, semantische netwerken en neurale netwerken.

Bij statistische analyse 'leert' het classificatiesysteem van de proefdocumenten die door experts zijn gemaakt en die in categorieën worden geplaatst. Het systeem gebruikt een algoritme om de relatie tussen de documenten in iedere categorie te bepalen en classificeert nieuwe documenten tegen die definities. In het algemeen geldt: hoe meer documenten bestaan van waaruit deze systemen leren, hoe nauwkeuriger het resultaat is.

Semantische netwerken maken gebruik van speciale woordenboeken die specifieke terminologieën en definities bevatten. Deze woordenboeken worden, samen met een aantal taalkundige regels, gebruikt om vast te stellen hoe woorden en uitspraken uit documenten overeenkomen met de business-terminologie van een onderneming. Semantische woordenboeken zijn aantrekkelijk omdat ze precies de betekenis van een woord definiëren en documenten context geven.

Een neuraal netwerk is een systeem dat bestaat uit computerprogramma's en datastructuren die de menselijke hersenen pogen na te bootsen. De neurale netwerken, die bestaan uit een groot aantal processoren, worden eerst gevoed door een substantieel aantal data om hen te 'trainen'. In theorie kan het getrainde systeem zich automatisch aanpassen aan veranderingen in content die is gebaseerd op herhaling en frequentie. Hierdoor kunnen grote volumes inhoud worden geclassificeerd met slechts weinig begeleiding.

In het algemeen is het gecontroleerd lerende systeem een verbetering ten opzichte van het zelflerende systeem, omdat het de mogelijkheid geeft nauwkeurig informatie te classificeren rondom de behoeften van de business. Door de instructies van een expert biedt deze methode meer nauwkeurige resultaten dan mogelijk is met het zelflerende systeem. Vergeleken met het handmatige classificatiesysteem biedt het gecontroleerd lerende systeem een meer werkbare balans tussen uitvoering en hulpbronnen. Toch bestaan ook hier nadelen. Hoe meer documenten men bijvoorbeeld gebruikt om het automatische classificatiesysteem te 'trainen', hoe beter het resultaat. Dit maakt het selecteren van de documenten om te trainen een zeer moeilijke taak. Semantische netwerken kunnen bovendien een hoog niveau van nauwkeurigheid bereiken wanneer de documenten die zij classificeren gebruikmaken van consistente terminologie. Als documenten een grote variëteit aan woorden gebruiken om hetzelfde concept te uiten, is het moeilijk zo niet onmogelijk om de informatie op één lijn te plaatsen. Het grootste nadeel van het semantische netwerk is de arbeidsintensiviteit die nodig is om de woordenboeken op te zetten en te behouden. Alle business-termen moeten worden toegevoegd en handmatig gedefinieerd. Tot slot bestaat er, zelfs wanneer getraind is met een groot aantal documenten, een limiet aan de nauwkeurigheid van automatische classificatie. Volgens een studie van Microsoft Research zijn meer dan negenduizend documenten nodig om het neurale netwerk te leren nieuwe data te classificeren, met een maximum van tachtig procent nauwkeurigheid op het eerste categorieniveau.

De hybride aanpak

Het lijkt of geen enkele methode geheel geschikt is; alle methoden hebben hun specifieke voor- en nadelen. Maar er is één systeem dat nog niet besproken is, te weten het gecombineerde (hybride) systeem. Hybride systemen bieden intelligente classificatie; zij combineren de voordelen van het menselijke intellect met de efficiënte verwerkingskracht van machines. De hybride benadering erkent dat menselijke betrokkenheid essentieel is bij het classificeren van bedrijfsinformatie. Het menselijke vermogen is noodzakelijk om context te begrijpen en de basisregels te creëren om het classificatieproces te controleren. Met hybride systemen kunnen deze regels handmatig of door middel van de hiervoor beschreven leerprocessen worden aangemaakt. Het grote verschil is echter dat de regels in beide gevallen door de experts zijn aan te passen om het resultaat te verbeteren. Met andere woorden: nadat de categorieën zijn vastgesteld definiëren mensen die de onderneming kennen en begrijpen de regels die bepalen welk soort documenten worden toegewezen aan welke categorie. Deze regels kunnen variëren van simpel tot complex, afhankelijk van de organisatie. Het voordeel van deze methode is dat steeds nieuwe categorieën kunnen worden aangemaakt en dat regels zijn te veranderen. Deze regels kunnen ook rekening houden met gestructureerde data zoals metagegevens (bijvoorbeeld titel, onderwerp, auteur en datum).

Hybride classificatie maakt duidelijk dat het niet praktisch is om mensen in te zetten om al het werk te doen. Als de experts de basisregels hebben vastgesteld zijn alleen computers in staat om de documenten te classificeren met de benodigde snelheid en consistentie. Deze hybride benadering verlangt meer planning vooraf, maar daar tegenover staan nauwkeurigheid en besparing op beheerskosten. En dat komt de terugvindbaarheid van ongestructureerde informatie alleen maar ten goede.

Victor Cohen

Victor Cohen is General Manager bij Verity Benelux B.V., leverancier van search-, classificatie- en taxonomiesoftware. E-mail: vcohen@verity.com.