Interview

Zoekmachines zijn een handig hulpmiddel om informatie te verzamelen, maar ook na het zien van de resultaten blijft het een hele klus de juiste informatie uit een zoekopdracht te filteren. Met behulp van text mining is het mogelijk om een stap verder te gaan en ook patronen en relaties in je data bloot te leggen om zo meer controle te hebben over de enorme hoeveelheid aan informatie die tegenwoordig beschikbaar is.

Text mining: verder in zoektechnologie

Alles vinden, zelfs als je het niet zoekt

Deze techniek wordt al veelvuldig toegepast door bijvoorbeeld opsporingsdiensten en toezichthouders, maar ook in het bedrijfsleven doet text mining steeds meer haar intrede om data te structuren en te analyseren en vervolgens eenvoudig te visualizeren.

De informatie die beschikbaar is, zal alleen maar blijven groeien. Volgens de Wetten van Moore, één van de oprichters van Intel en uitvinder van de computerchip, verdubbelt de reken- en opslagcapaciteit van een computer iedere achttien maanden. Ook de hoeveelheid informatie die men produceert en verzamelt verdubbelt elke achttien maanden. Dit heeft als gevolg dat er een enorme toename aan beschikbare informatie en data is en dat het steeds moeilijker wordt om die doorzoekbaar te maken en te archiveren. Zeker omdat het overgrote deel daarvan (meer dan negentig procent) ook nog eens ongestructureerd is.

Data mining is een techniek die alleen uit de voeten kan met gestructureerde data die is ondergebracht in relationele databases. Via deze methode kan bijvoorbeeld aan creditcardbetalingen of pintransacties een aantal kenmerken worden toegevoegd zoals datum, leeftijd van de pashouder of salaris om zo patronen van interesse en gedrag te herkennen.

Het overgrote deel van de beschikbare informatie is echter ongestructureerd en om daarvan gebruik te kunnen maken, moet deze eerst gestructureerd worden. Text mining is een techniek die ervoor zorgt dat ongestructeerde data gestructureerd wordt en zo door mens en computer verwerkt kan worden met behulp van methodes waar we al bekend mee zijn.

Jan Scholtes is directeur van ZyLAB, een bedrijf dat geavanceerde zoeksoftware levert. “Onze software wordt veel gebruikt door inlichtingen- en opsporingsdiensten, maar ook bijvoorbeeld door oorlogstribunalen. Alle VN oorlogstribunalen in de wereld maken gebruik van onze software. Verder wordt het onder andere gebruikt door de Belastingdienst, de FBI en het Amerikaanse Witte Huis”, schetst hij.
Sinds begin dit jaar is Scholtes tevens bijzonder hoogleraar aan de Universiteit van Maastricht in de text mining. “De techniek bestaat nu ongeveer 25 jaar en is in het begin van de jaren tachtig ontwikkeld binnen veiligheids- en inlichtingendiensten. In die periode was het alleen mogelijk om te zoeken op trefwoorden die van tevoren waren ingebracht. Denk bijvoorbeeld aan de kaartenbak van de bibliothecaresse, maar dan geautomatiseerd. Wij hebben in die periode full text zoeken ontwikkeld, zodat je ook op de informatie binnen de

Arjen van den Berg
is eindredacteur van Software Release Magazine

‘Text mining is erg gerelateerd aan zoeken. Alleen weet je niet altijd van tevoren waarnaar je op zoek bent.’

documenten kunt zoeken. Handmatig grote hoeveelheden ongestructureerde informatie doorbladeren is immers geen doen,” legt hij uit. “Momenteel zijn de resultaten die je terug krijgt bij de huidige hoeveelheden informatie gigantisch en zijn er technieken zoals text mining nodig om deze op voorhand van structuur te voorzien. De structuur bestaat vaak uit entiteiten en feiten die uit de tekst gehaald zijn en kunnen gebruikt worden om de informatie verder te analyseren.”

Internetzoekmachines tonen alleen de beste of meest populaire antwoorden op een zoekvraag. Ook is het zo dat iedereen zijn best doet om bij een zoekopdracht boven in de lijst met gevonden resultaten te komen. Fraudeonderzoekers, opsporingsdiensten en toezichthouders willen echter alle informatie kunnen vinden. Fraudeurs en criminelen doen hun best om juist niet bovenaan de resultatenlijst te komen. Door het gebruik van synoniemen en termen die weinig voorkomen in hun documenten blijven ze onvindbaar. Text mining is een techniek die uitkomst biedt.

“Het is een combinatie van taaltechniek, statistiek en wiskunde. Het is zaak om in teksten bepaalde patronen en verbanden te vinden en die vervolgens om te zetten in getallen, zodat een computer ermee overweg kan. In de text mining speelt patroonherkenning een grote rol. Het is erg gerelateerd aan zoeken. Alleen weet je niet altijd van tevoren waar je naar op zoek bent,” zegt Scholtes.

Text mining richt zich op het ontwikkelen van diverse geavanceerde wiskundige, statistische en taalkundige patroonherkenning in grote hoeveelheden elektronische informatie, zoals tekstdocumenten, emailconversaties en chatsessies. Door die patronen en kenmerken is het mogelijk beter en gerichter te zoeken en sneller inzichten te krijgen die anders verborgen blijven. In plaats van te zoeken op woorden wordt er gezocht op taalkundige patronen van woorden.

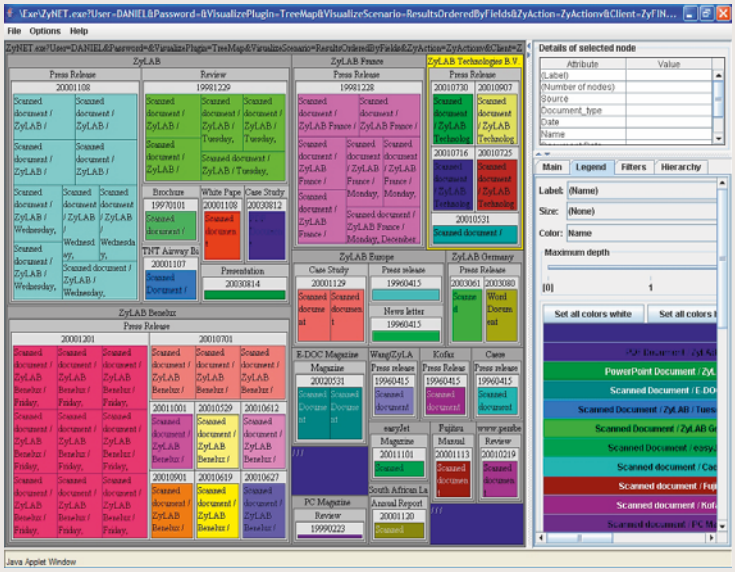
Met behulp van een uitvoerige taalkundige analyse kunnen gegevens aan elkaar gekoppeld worden. Via de taalkundige rollen van woorden en zinsconstructies (onderwerp, persoonsvorm, lijdend voorwerp) kunnen bijvoorbeeld eigennamen geclassificeerd worden en op een gestructureerde manier gepresenteerd worden. “In een emailwisseling wordt iemand niet altijd bij zijn naam genoemd. Hij staat ook vaak anders aangeduid en door deze analyse kun je een patroon ontdekken en alles van die persoon bij elkaar brengen. Vooral in rechtbankdossiers of uitgewerkte gesprekken tussen een verdachte en zijn advocaat is op deze manier veel informatie te vinden,” verduidelijkt Scholtes.

“Het is echt zoeken plus. Je haalt op deze manier 99,9 procent van de resultaten. Mensen zijn niet consistent. Ze halen niet meer dan twintig procent eruit. Deze techniek is goed te gebruiken voor het ‘domme werk’, maar mensen moeten altijd kwaliteitscontroles uitvoeren. Je moet uitkijken dat je geen conclusies trekt, die je niet verifieert. Zo kwam Ted Kennedy ooit eens voor op een lijst van mogelijke terroristen en

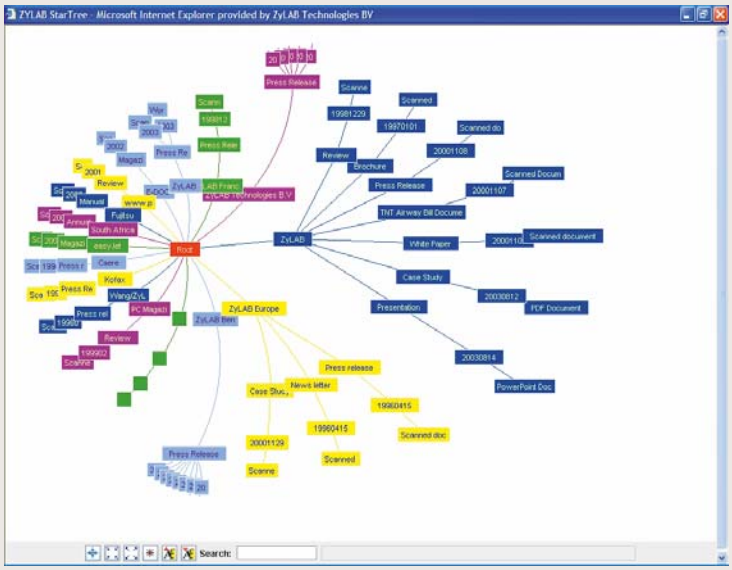
Jan Scholtes is directeur van ZyLAB en sinds begin dit jaar bijzonder hoogleraar in de text mining aan de Universiteit van Maastricht.

‘Deze techniek heeft sinds 9/11 een vlucht genomen. Ook het bedrijfsleven krijgt er steeds meer aandacht voor.’

mocht hij een vliegtuig niet in. Er moet dus altijd een goede controle zijn op de gevonden resultaten,” vervolgt hij.
Wanneer de informatie eenmaal gestructureerd is, volgt de analyse en kunnen de resultaten gevisualiseerd worden. Via een boomstructuur is gevonden informatie goed in beeld te brengen. In het centrum van het diagram staat het onderwerp of persoon met daar omheen de gerelateerde resultaten die weer onderling verbonden zijn, zodat een duidelijk patroon zichtbaar is. Ook een zogenaamde treemap is een veelgebruikte visualisatietechniek. Alle patronen worden geclusterd en krijgen een aparte kleur. Deze visualisatietechnieken zijn vooral erg geschikt om grote collecties email overzichtelijk te maken om vervolgens te analyseren.



Hyperbolic Tree visualisatie van een boomstructuur (Bron: ZyLAB Technologies)



TreeMap visualisatie van een boomstructuur (Bron: ZyLAB Technologies).

“Zo heb je in een oogopslag alle informatie voorhanden. Een beeld zegt vaak meer dan duizend woorden.”

Toepassingen

Zoals al eerder gezegd wordt text mining veelvuldig toegepast op het gebied van fraudeonderzoek, analyse van grote en complexe criminele organisaties en onderzoek door oorlogstribunalen.
Maar er zijn ook andere gebieden waar text mining gebruikt wordt. Denk bijvoorbeeld aan sentiment mining en business intelligence. Bedrijven zijn erg geïnteresseerd in hoe er over hun en hun producten gedacht wordt en geschreven wordt op internet. Met het zoeken op hun bedrijfsnaam verschijnen miljoenen resultaten. Via text mining technieken kan nuttige informatie herkend en geanalyseerd worden om zo een beeld te krijgen en vervolgens de juiste actie te ondernemen. Maar ook de concurrentie kan onder de loep worden genomen. Via text mining technieken is veel informatie over de markt en de concurrentie te verzamelen.
Ook in de farmaceutiche industrie heeft text mining zijn nut bewezen. Bij onderzoek naar nieuwe medicijnen of behandelmethodes is het nuttig uit vele tienduizenden medische observaties patronen te ontdekken. Daarbij is vaak van tevoren niet bekend waar naar gezocht moet worden.
Op commercieel gebied is het analyseren van garantieproblemen van grote waarde. Door vele rapporten van reparaties te analyseren kan een fabrikant informatie halen om het product al vroegtijdig te verbeteren of te veranderen. Ook spamfilters maken veelvuldig gebruik van text mining technieken om voortijdig schadelijke emailberichten te onderscheppen.
Scholtes verwacht dat text mining in de toekomst in steeds meer toepassingen gebruikt wordt. “Sinds 9/11 heeft text mining een enorme vlucht genomen. Naast opsporingsdiensten, toezichthouders, oorlogstribunalen, rechtbanken zien we dat ook het bedrijfsleven steeds meer aandacht heeft voor text mining. Er worden steeds meer toepassingen gebruikt. Automatisch archiveren gaat bij bedrijven echt een groei nemen. Ze willen alles vastleggen en vooral makkelijk terug kunnen vinden.”

Referentie

Inaugurele rede van prof. dr. ir. Jan Scholtes bij zijn aantreden als bijzonder hoogleraar aan de Universiteit van Maastricht in de text mining. «