Automatisch classificeren

J. van Aalten

1 Inleiding

Classificeren, oftewel het ordenen van dingen in overeenkomstige categorieën, is iets wat iedereen dagelijks bewust of onbewust doet. Om informatie gemakkelijk en snel terug te kunnen vinden, sla je documenten op volgens de mappenstructuur van je C-schijf of archiveer je je e-mails naar onderwerp.
Bibliotheken zijn van oudsher de deskundigen op het gebied van het classificeren van boeken en andere gedrukte informatie. Sinds de jaren 60 van de vorige eeuw is informatie echter steeds meer in digitale vorm beschikbaar. Door de explosieve groei van informatie en met de komst van internet, nam het terugvindbaar maken van informatie een andere wending. Met behulp van grote internetzoekmachines zoals AltaVista waande iedereen zich een informatiespecialist en was het zoeken naar informatie niet langer het alleenrecht van bibliothecarissen. Met een paar trefwoorden en een freetextzoekmachine zoals Google heeft iedereen de informatie die hij of zij nodig heeft binnen handbereik.
Het classificeren van informatie leek hiermee overbodig te gaan worden. Leek, want een verdere toename van informatie bemoeilijkt ook het zoeken. Het gebruik van één of meerdere zoekwoorden levert namelijk, zelfs in combinatie met Booleaanse operatoren of door te zoeken in specifieke velden, niet altijd het gewenste resultaat op.
Handmatig toekennen van metadata aan digitale documenten was een goede manier om informatie toegankelijk te maken. Alleen bleef ook binnen organisaties de digitalisering van documenten verder stijgen, waarmee het handmatig toekennen van metadata steeds arbeidsintensiever werd. Té arbeidsintensief. Er ontstond een patstelling: handmatig toekennen is te arbeidsintensief, maar zonder deze metadata worden documenten niet of

slecht teruggevonden. Een antwoord bleek te liggen in het automatisch classificeren van documenten.
In dit artikel wordt automatisch classificeren beschreven vanuit een bibliotheekvisie, dus zonder 'lastige' algoritmen en statistische formules. De aandacht ligt met name op de verschillende methoden voor automatische classificatie en op de rol die 'traditionele' ontsluitingsmethodieken zoals taxonomieën en thesauri hierbij (kunnen) spelen.

2 Wat is automatisch classificeren?

2.1 Twee manieren

Het automatisch classificeren van informatie is het indelen van documenten in categorieën met behulp van classificatiesoftware. Dit kan grofweg op twee manieren:

1. Met behulp van vooraf gedefinieerde classificatieregels.

Informatieprofessionals maken classificatieregels op basis waarvan een document in een categorie ingedeeld kan worden. De meest eenvoudige vorm van een classificatieregel is een simpele query, zoals 'IR or information retrieval'. De classificatiesoftware kan vervolgens op basis van deze en soortgelijke regels de documenten die aan de zoekvragen voldoen in de juiste categorieën classificeren.

2. Op basis van kenmerken uit documenten.

Voor iedere categorie wordt een trainingset met relevante en eventueel niet-relevante documenten aangedragen. Op basis van kenmerken van deze documenten maakt het systeem zelf classificatieregels, waartegen andere documenten weer geclassificeerd kunnen worden.

De categorieën waarin de documenten ingedeeld worden, kunnen zelf gemaakt of gekocht worden. Deze categorieën worden hiërarchisch gerangschikt, verreweg in de meeste gevallen in een taxonomie (zie paragraaf 4). Sommige software kan ook automatisch categorieën aandragen. Beide methoden sluiten elkaar niet per definitie uit en kunnen aanvullend aan elkaar gebruikt worden wanneer de classificatiesoftware dat ondersteunt.

2.2 Voor- en nadelen

Het grote voordeel van automatisch classificeren is de tijdwinst. Handmatig kunnen volgens onderzoeksbureau Gartner 20 tot 100 documenten per dag geclassificeerd worden, afhankelijk van de lengte van de documenten.

Wanneer gebruik gemaakt wordt van classificatiesoftware, kunnen 20 tot 300 documenten per seconde geclassificeerd worden.
Automatische classificatie wil niet altijd zeggen: beter. Of de kwaliteit van automatische classificatie ten opzichte van handmatige classificatie beter is, is afhankelijk van diverse factoren, zoals de kwaliteit van de categorieën en eventuele classificatieregels. Wanneer aan deze en andere randvoorwaarden wordt voldaan, kan automatische classificatie net zo nauwkeurig of zelfs nauwkeuriger classificeren dan een persoon. Automatische classificatie is in ieder geval consistent, iets wat mensen niet altijd zijn. Wanneer de classificatieregels niet veranderen, zullen de documenten altijd op dezelfde manier ingedeeld worden.
Automatisch classificeren kent ook nadelen. Er is bij automatisch classificeren meer voorbereidingstijd nodig om tot een werkend systeem te komen: het systeem moet ingericht worden en er moeten classificatieregels gebouwd of aangeleerd worden. Want hoewel de term 'automatische classificatie' wellicht anders zou kunnen suggereren, gaat automatische classificatie zeker niet geheel automatisch. Er is vrijwel altijd handwerk nodig, om de hiërarchische structuur en/of de classificatieregels te maken of om een set met trainingsdocumenten te selecteren.
Andere nadelen van automatische classificatie zijn ook softwareafhankelijk: ieder classificatieproduct heeft zijn eigen sterke en minder sterke kanten. Vooral wanneer een product werkt volgens een blackboxmodel, is het voor een gebruiker soms lastig te begrijpen waarom documenten op een bepaalde manier ingedeeld worden. Dit zou kunnen leiden tot een verminderde acceptatie van het systeem of twijfel aan de kwaliteit van de informatie.
Het selecteren van de software voor automatische classificatie is overigens geen sinecure. De enige manier om vast te stellen of de software goed zal werken met documenten uit de organisatie, is door testen te doen met eigen documenten en eigen categorieën. Succesverhalen uit het verleden of bij andere organisaties zijn geen garantie voor succes.

2.3 Randvoorwaarden voor succes

Automatische classificatie is een succes wanneer ten opzichte van handmatige classificatie documenten in minder tijd, maar met vrijwel gelijke of betere kwaliteit geclassificeerd worden.
Wat voldoende kwaliteit van classificatie is, zal per organisatie en collectie verschillen. Is het noodzakelijk dat 100% van de documenten correct ge-

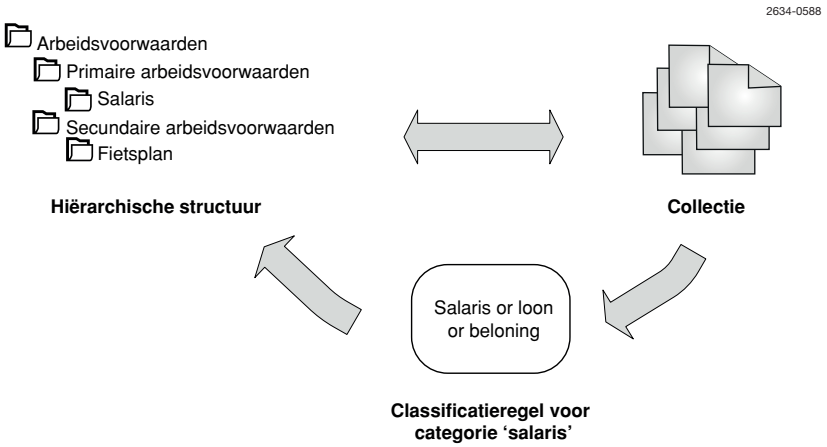
classificeerd worden of is een 80-20-regel acceptabel? Dit is afhankelijk van de eisen en het verwachtingspatroon van toekomstige gebruikers.
Of de kwaliteit van de automatische classificatie voldoende kan zijn, is afhankelijk van:

- kwaliteit van categorieën: zoveel mogelijk consistente, eenduidige en onderscheidende categorieën;
- omvang van de documentcollectie: hoe meer documenten, hoe beter;
- homogeniteit van collectie en domein: hoe beperkter het aantal onderwerpen in een collectie, hoe beter;
- homogeniteit van documenten: hoe minder verschillende onderwerpen binnen een document, hoe beter;
- veranderlijkheid van het domein: hoe minder aan verandering onderhevig, hoe beter;
- kwaliteit van taal in documenten: hoe minder straat-, sms- of gesproken taal, hoe beter;
- soortgelijkheid van documenten: hoe minder verschillende soorten documenten, hoe beter;
- kwaliteit van huidige metadata: hoe meer kwalitatief goede metadata, hoe beter;
- domeinexpertise binnen de organisatie: hoe meer kennis van het domein bij opbouw en onderhoud, hoe beter;
- diversiteit in talen: hoe minder verscheidenheid in talen, hoe beter.

3 De methodieken

3.1 Op basis van classificatieregels

Op basis van een hiërarchische structuur en bijbehorende classificatieregels worden documenten in de juiste categorie geclassificeerd. Zowel de hiërarchische structuur (dikwijls in de vorm van een taxonomie) als de classificatieregels kunnen door een organisatie zelf gemaakt worden, of gekocht bij een leverancier.
Figuur 1 verduidelijkt hoe de classificatieregels en de hiërarchische structuur samenwerken. Links de hiërarchische structuur, opgebouwd uit verschillende categorieën. Voor iedere categorie zijn classificatieregels gedefinieerd. De documenten uit de collectie worden tegen deze classificatieregels aangehouden en in de juiste categorie(ën) geclassificeerd.



Figuur 1.

Wanneer we bovenstaande classificatieregel koppelen aan de categorie Salaris, dan zullen de documenten met salaris of loon of beloning in de tekst ingedeeld worden in deze categorie. In werkelijkheid zal een dergelijke regel niet voldoende zijn, omdat we documenten met de zinsnede "het was zijn verdiende loon" niet in deze categorie geclassificeerd willen zien worden.

Classificatieregels kunnen zeer nauwkeurig gedefinieerd worden en kunnen daarmee accurate classificatieresultaten opleveren. Wanneer de regels echter te nauw geformuleerd worden, bestaat het gevaar dat de recall te laag wordt.

Door te werken met classificatieregels is het voor informatie- en domeinspecialisten inzichtelijk waarom documenten op een bepaalde manier geclassificeerd worden. En wanneer dit duidelijk is, is het ook mogelijk om de classificering aan te passen en bij te sturen.

Het zelf bouwen en onderhouden van classificatieregels is zeer tijdrovend. Als categorieën uit de hiërarchische structuur veranderen, moeten ook de classificatieregels aangepast worden. Hiervoor is zowel expertise van domeinspecialisten als informatiespecialisten nodig.

3.1.1 Zelf maken

Een organisatie kan ervoor kiezen de classificatieregels voor een automatisch classificatiesysteem geheel handmatig te maken. Hierbij kunnen bestaande indexeertalen (zoals een taxonomie) die binnen de organisatie gebruikt worden als uitgangspunt dienen. Veel classificatiesoftware onder-

steunt het importeren van bepaalde trefwoordenlijsten. Maar ook als een organisatie geen taxonomie, thesaurus, classificatiesysteem of iets soortgelijks heeft, zijn er manieren waarop de taxonomie en classificatieregels voor de automatische classificatie gebaseerd kunnen worden. Denk hierbij bijvoorbeeld aan de hiërarchische mappenstructuur van de gezamenlijke netwerkschijf.
Organisaties die zelf classificatieregels zullen gaan bouwen zijn bijvoorbeeld organisaties die actief zijn op een bepaald gebied waarvoor geen kant-en-klare taxonomieën met classificatieregels bestaan en die vanwege uiteenlopende redenen afzien van classificatie op basis van concepten.
Het voordeel van het zelf opbouwen van classificatieregels is dat deze geheel toegespitst kunnen worden op de eigen organisatie. Ze kunnen zodanig opgebouwd worden dat deze zo dicht mogelijk aansluiten bij de perceptie van de gebruiker en organisatiespecifiek taalgebruik.

3.1.2 Kopen van een leverancier

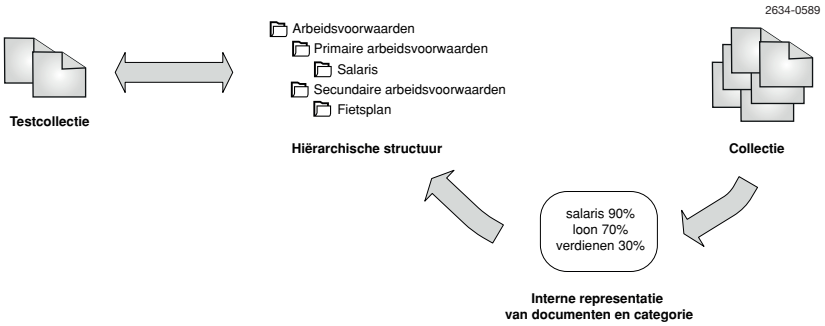
Voor wie het zelf maken van classificatieregels geen optie is, kan ook op zoek naar een leverancier die deze aanbiedt. Bij partijen zoals Lexis-Nexis en Factiva is het mogelijk deze te kopen. Zij leveren taxonomieën en classificatieregels voor diverse vakgebieden en industrieën. Nadeel hiervan is dat deze gericht is op Engelstalige content.
Ook in Nederland bieden leveranciers hun taxonomieën en thesauri aan. De drie kennisinstituten Vilans, MOVISIE en NJi leveren en onderhouden al jaren de Thesaurus Zorg en Welzijn. Kluwer biedt sinds een tijd hun taxonomie, thesaurus en bijbehorende classificatieregels aan voor de classificatie van Nederlandstalige, juridische informatie. Er zal echter dikwijls nog een slag gemaakt moeten worden om ze werkend te krijgen met specifieke classificatiesoftware en organisatiespecifieke content. Zo zijn organisatiespecifieke termen uiteraard niet opgenomen. Indien de organisatie dit wel wil, zullen de regels aangepast moeten worden. Dit is ook noodzakelijk wanneer uit de testen blijkt dat het classificeren van de content nog niet optimaal is.
Wanneer overwogen wordt classificatieregels van een externe partij aan te schaffen, moet de organisatie zich met betrekking tot de classificatieregels in ieder geval het volgende afvragen:

– Op welke manier zijn ze tot stand gekomen?
– Ontbreken er categorieën of regels?
– Kloppen ze inhoudelijk?
– Hoe verloopt de classificatie van eigen organisatiespecifieke content?

– Is het mogelijk ze zelf aan te passen en uit te breiden?
– Hoe zullen onderhoud en aanpassingen plaatsvinden?

3.2 Op basis van kenmerken

Het is ook mogelijk het systeem te leren hoe te classificeren. In dat geval worden voor iedere taxonomiecategorie relevante en eventueel niet-relevante documenten aan het systeem aangeboden. Op basis van de inhoudelijke kenmerken en gelijke patronen van deze documenten maakt het systeem een interne representatie. Op basis van deze kenmerken kunnen andere documenten op dezelfde wijze geclassificeerd worden.



Figuur 2.

Softwareleveranciers gebruiken diverse methoden om documenten aan categorieën toe te kennen. Deze methoden zijn gebaseerd op diverse wiskundige technieken. Zo gebruikt Autonomy de Bayesian-methode, Inxight en Verity K2 (tegenwoordig onder de vlag van Autonomy) gebruiken de k-nearest neighbor-methode. Microsoft gebruikt Support vector machines. De voorbereiding bij deze methode bestaat uit het verzamelen en categoriseren van documenten die representatief zijn voor een bepaalde categorie. Voor deze zogenaamde trainingsdocumenten geldt vaak: hoe meer, hoe beter. Hoe meer documenten er aangeboden worden aan het systeem, hoe beter de bijbehorende classificatieregels die het systeem maakt, zullen zijn. Hoeveel documenten er minimaal aangeboden zullen moeten worden verschilt per systeem, maar reken op minimaal twintig relevante documenten per categorie.

Het selecteren van de relevante documenten moet zorgvuldig gebeuren. Hoe breder van onderwerp de documenten zullen zijn, hoe breder de classificatieregels worden en hoe kleiner de kans dat documenten gemist zullen worden. Wanneer echter te homogene documenten aangeboden wor-

den, kan het zijn dat de classificatieregels te beperkt zijn en andere, wel relevante, documenten worden gemist.
De door het systeem toegekende kenmerken zijn, afhankelijk van het gekozen classificatiepakket, zichtbaar en eventueel aanpasbaar óf onzichtbaar en daarmee niet controleerbaar. Als het systeem niet inzichtelijk maakt hoe de classificatieregels werken, is het voor de classificatiespecialist lastig bij te sturen. Als tijdens de testfase blijkt dat het classificeren niet goed gaat, is voor hem de enige manier om het systeem te sturen door meer trainingsdocumenten aan het systeem aan te bieden. Het nauwkeurig beïnvloeden van het systeem zal echter lastig zijn.

4 De rol van taxonomieën en thesauri

4.1 Thesauri

Om informatie makkelijker terug te vinden kunnen we trefwoorden toekennen aan documenten, om ze hier vervolgens weer op terug te kunnen vinden. Om vast te leggen welke trefwoorden we wel en niet gebruiken en wat hun onderlinge relatie is, maken we een gecontroleerde trefwoordenlijst of thesaurus. Een thesaurus is een alfabetische lijst van voorkeurstermen en hun onderlinge hiërarchische en semantische relaties, zoals synoniemen en homoniemen. Bij het classificeren is een thesaurus een leidraad voor degene die trefwoorden dient toe te kennen. Voor synoniemen wordt gestandaardiseerd welke termen de voorkeur hebben om te gebruiken bij de ontsluiting van de documenten.
Thesauri worden niet alleen ingezet bij het beschrijven van informatie; ze zijn uiteindelijk bedoeld om gebruikers als hulpmiddel te dienen bij het terugzoeken van deze informatie. Bij het uitkiezen van de trefwoorden waarop een gebruiker zoekt, wordt hij verwezen naar woorden waarop hij moet zoeken (de USE-termen) en/of gewezen op andere woorden die hem verder kunnen helpen bij zijn zoekvraag (ruimere begrippen - BT, specifiekere begrippen - NT en verwante begrippen - RT).

4.2 Thesauri bij automatische classificatie

Met het automatisch classificeren van informatie staan thesauri weer in de spotlight. En hoewel ze bij het automatisch classificeren van informatie op een andere wijze ingezet worden, is hun primaire rol gelijk gebleven: het toegankelijk maken en classificeren van informatie. Organisaties die besluiten handmatig classificatieregels te gaan maken voor het automatisch

classificeren van hun content, kunnen hierbij gebruikmaken van een eventuele aanwezige thesaurus. Met name de semantische relaties tussen termen zijn waardevol bij het opstellen van classificatieregels.
Het grote verschil bij de inzet van thesaurustermen bij handmatige en automatische classificatie is het feit, dat bij handmatige classificatie de documenten centraal staan en dat bij automatische classificatie dit de classificatieregels zijn. Omdat dit toch een wezenlijk andere insteek is, zijn de thesauri voor handmatige classificatie niet een-op-een geschikt voor automatische classificatie. De vraag is namelijk niet langer: 'met welke woorden omschrijf ik dit document zo goed mogelijk?', maar 'op basis van welke woorden zorg ik dat een document in de juiste categorie ingedeeld wordt?'. Dit houdt in de praktijk in dat woorden zoals 'management' of 'beleid' bij handmatige classificatie logisch kunnen zijn, maar in classificatieregels geen nut hebben.
Wil je een thesaurus vertalen naar classificatieregels, dan is het verstandig rekening te houden met de mogelijkheden die classificatieregels bieden:
– Voor het herkennen of een tekst over een bepaald onderwerp gaat, beperkt men zich niet tot voorkeurstermen, doel is juist zoveel mogelijk synoniemen en spellingvarianten op te nemen
– Daarbij zijn ook hele zinsdelen of combinaties van woorden mogelijk.

Daarnaast bieden classificatieregels dikwijls de volgende mogelijkheden:
– het toekennen van weging aan kenmerkende woorden;
– het vastleggen van relaties tussen kenmerkende woorden door middel van booleaanse of proximity operatoren;
– het gebruik van truncatietekens;
– het specificeren dat een kenmerkend woord in een bepaald documentdeel moet voorkomen.

Of een thesaurus bij een zelflerend systeem ingezet kan worden, is afhankelijk van de gekozen software. Als dat mogelijk is, wordt de thesaurus ingelezen en gebruikt bij het bepalen van de kenmerken.

4.3 Taxonomieën

Er zijn verschillende technieken en hulpmiddelen voor het toegankelijk maken van informatie. Zij hebben alle hun specifieke kenmerken, maar kennen ook overlappingen. In dit artikel worden kort die methodieken genoemd die ook bij het automatisch classificeren ingezet kunnen worden. Zo worden boeken binnen bibliotheken altijd geordend volgens een vooraf

opgezet classificatiesysteem. Zo weet een gebruiker niet alleen over welke onderwerpen een boek gaat, ook staan boeken over hetzelfde onderwerp fysiek bij elkaar geordend. Een classificatiesysteem kan hiërarchisch ingedeeld worden en dan spreken we ook wel van een taxonomie. Een taxonomie kent zijn oorsprong in de biologie, en vanuit deze opzet hebben de onderwerpen een klasse-soortrelatie. Dit gegeven werd buiten de biologie steeds vaker losgelaten, zodat in een taxonomie ook andere hiërarchische relaties worden toegepast. Bijvoorbeeld de geheel-deelrelatie, zoals hoofd-neus.

5 Automatisch classificeren stap voor stap

5.1 Voorbereiden en analyseren

Een goede voorbereiding is essentieel voor een succesvolle implementatie. Als eerste stap is het daarom noodzakelijk om te onderzoeken wat de wensen en eisen van de toekomstige gebruikers en de stakeholders zijn en te kijken naar de content die je wilt laten classificeren. Bepaal vervolgens op basis hiervan welk pakket voor jouw specifieke situatie geschikt is.

Organisatorische voorbereiding

De organisatorische voorbereiding begint met het benaderen van verschillende personen uit de organisatie. Personen die in ieder geval betrokken moeten worden, zijn: informatiespecialisten, inhoudelijke vakspecialisten, eindgebruikers en technische specialisten. Deze betrokkenheid is niet alleen om een breed draagvlak te creëren binnen de organisatie, de diverse specialisten zijn ook onmisbaar vanwege hun inhoudelijke kennis.

Bepaal wie de stakeholders zijn en probeer erachter te komen wat hun belangen en wensen zijn. Bepaal het beschikbare budget, deadlines, communicatiemomenten en leg alles vast in een projectplan.

Vergeet ook niet de scope vast te stellen: welke soorten informatie en bronnen moeten automatisch ontsloten worden? Leg ook dit expliciet vast. Kijk hierbij uit voor de valkuil om alle informatie die binnen een organisatie aanwezig is, te willen ontsluiten.

Probeer minimaal de volgende vragen te beantwoorden:

– Welke doelgroepen moeten bediend worden?
– Wie zijn de stakeholders?
– Wat zijn hun eisen en wensen?
– Hoeveel budget is er beschikbaar?

– Wat is de projectplanning?
– Wat worden de deliverables?
– Welke specialisten en afdelingen worden betrokken?
– Welke inhoudelijke en technische kennis is er binnen de organisatie?
– Wie wordt verantwoordelijk voor het inrichten en onderhoud van het systeem?
– Hoeveel menskracht is hiervoor beschikbaar?
– Welke (interne en externe) bronnen moeten ontsloten worden?
– Moet er integratie plaatsvinden met andere aanwezige systemen, zoals een dms of cms?

Wensen en eisen

Inventariseer de wensen en eisen van de eindgebruikers enerzijds en degene die met het automatische classificatiesysteem moeten gaan werken anderzijds. Workshops of interviews zijn hier goede methoden voor. Voor eindgebruikers zal de focus van de wensen en eisen liggen op de voorkant van het geautomatiseerde classificatiesysteem. Hoe wil men zoeken, navigeren of een combinatie van beiden; weegt de recall of precisie van zoekresultaten zwaarder; welke eisen worden er gesteld aan de kwaliteit van classificatie; welke zoekfunctionaliteiten en -opties moet het systeem aanbieden?

Probeer minimaal de volgende vragen te beantwoorden:
– Welke wensen en eisen zijn er met betrekking tot gebruik?
– Welke wensen en eisen zijn er met betrekking tot onderhoud?
– Willen de gebruikers liever een hogere recall of een hogere precisie?

Aanwezige ontsluitingssystemen

Vervolgens is het van belang een goed en volledig beeld te krijgen van de ontsluitingssystemen die binnen de organisatie gebruikt worden. Bekijk welke classificatiesystemen of woordsystemen de bibliotheek of informatieafdeling gebruikt. Wordt er al metadata toegekend aan documenten? Zijn er al taxonomieën, thesauri of andere vormen van controlled vocabularies in gebruik of beschikbaar? Maar denk ook aan andere ontsluitingssystemen in de breedste zin van het woord, zoals navigatiestructuren van een intranet of gezamenlijke netwerkschijven.

Probeer minimaal de volgende vragen te beantwoorden:
– Welke ontsluitingsystemen zijn er binnen de organisatie?
– Op welke wijze zouden deze ingezet kunnen worden?

Contentanalyse
Analyseer de content die automatisch geclassificeerd moet gaan worden en raak vertrouwd met de content die ontsloten moet gaan worden. Probeer de kenmerken van de verschillende bronnen vast te leggen.
Beantwoord minimaal de volgende vragen:
– Welke soorten content moeten geclassificeerd worden?
– Hoe homogeen of heterogeen is de content?
– Over welke bestandsformaten praten we?
– Wat is de omvang van de collectie?
– Wat is de lengte van de documenten?
– Zijn documenten al voorzien van metadata?

Pakketselectie
Doe vervolgens onderzoek naar de verschillende soorten methodieken, zoals in het hoofdstuk hiervoor beschreven. De keuze voor een methodiek is onlosmakelijk verbonden met een softwarepakket en bijbehorende leverancier. Bepaal op basis van de inventarisatie zoals die hiervoor beschreven is, welke mogelijkheden er voor de organisatie zijn en welke optie het beste aansluit bij de eisen, wensen, de aanwezige content en ontsluitingssystemen. Een goede oriëntatie op de beschikbare software is noodzakelijk; leveranciers zijn graag bereid te laten zien wat hun product kan. Vergeet niet naast leveranciers ook mogelijke implementatiepartners te benaderen en bepaal wat hun rol in het traject zal zijn.
Classificatiesoftware wordt overigens zelden als standaloneapplicatie aangeboden. Onder andere Autonomy en FAST bieden automatische classificatie aan als onderdeel van hun enterprisesearchoplossingen. Ook zijn er bijvoorbeeld voor Microsoft SharePoint add-ons beschikbaar die automatische classificatie mogelijk maken.
Probeer minimaal de volgende vragen te beantwoorden:
– Welke methodiek is het meest geschikt?
– Welk softwarepakket en leverancier is het meest geschikt?
– Welke ervaringen hebben andere organisaties met software en leveranciers?

5.2 Definiëren en bouwen
Na de voorbereidings- en analysefase kan met de daadwerkelijke definitie en bouw gestart worden. Hoe deze fase het best ingevuld kan gaan worden, is afhankelijk van de gekozen methodiek en software. Een lerend systeem

vereist een andere aanpak dan het toepassen van classificatieregels. Daarnaast heeft ieder softwarepakket zijn eigen gebruiksmogelijkheden.
Wanneer de automatische classificatie op basis van classificatieregels gaat plaatsvinden, is nu het moment gekomen om classificatieregels te bouwen of de gekochte classificatieregels te implementeren. Thesauri of andere controlled vocabularies kunnen nu vertaald worden naar classificatieregels. Indien het gekozen softwarepakket hiervoor geschikt is, kunnen deze geïmporteerd worden. Bekijk daarnaast of reeds toegekende (handmatige) metadatering meegenomen kan worden bij de classificatie van documenten en stel de classificatieregels zodanig op, dat er gebruikgemaakt wordt van deze metadata.
Bij een lerend systeem moeten de trainingdocumenten geselecteerd en aan het systeem aangeboden worden. Eerste stap is dan ook een selectie te maken van relevante documenten per categorie uit de taxonomie. Ga hierbij uit van minimaal 20 documenten per categorie. Is de aard van content zeer divers, verhoog dan het aantal documenten tot minimaal 50.

5.3 Testen

Ongeacht voor welke methode of softwarepakket gekozen is, zal na de bouwfase getest moeten worden of de resultaten van automatische classificatie voldoen aan de verwachtingen, of dat er bijgestuurd moet worden. Stel hiervoor een testplan op, met hierin de specificaties over een testset en de eisen die gesteld worden aan de testresultaten.
De testset is een afgebakende collectie documenten die met het nieuwe systeem automatisch geclassificeerd zullen gaan worden. Bij het samenstellen van de testset moet met de volgende aspecten rekening gehouden worden:

- Bepaal hoe de testset het best samengesteld kan worden. Is er wellicht metadatering op basis waarvan een selectie gemaakt kan worden, of moet handmatig een selectie gemaakt worden?
- Bepaal per taxonomiecategorie welke documenten deze categorie goed 'vertegenwoordigen'. Voeg deze documenten toe aan de testset en markeer ze als relevant zijnde voor die categorie. Houd minimaal tien correcte documenten per categorie aan.
- Voeg aan de testset documenten toe die niet-relevant zijn. Markeer voor welke categorie deze niet-relevant zijn. Hierbij kan gedeeltelijk gebruik gemaakt worden van relevante documenten van een andere categorie.
- Laat de testset samenstellen door informatie- en vakspecialisten. Informatiespecialisten hebben oog voor documenten die wellicht moeilijk

geclassificeerd kunnen worden, de vakspecialisten zijn vanwege hun inhoudelijke inzicht van documenten onmisbaar.
– Houd er rekening mee dat documenten in meerdere categorieën ingedeeld kunnen en mogen worden. Voeg bewust documenten toe waarvoor dat het geval is.
– Zorg voor diversiteit in de documenten: varieer zoveel mogelijk in lengte, taalgebruik, bestandsformaten, bronnen, etc.

Leg in het testplan vast wanneer de classificatietest geslaagd genoemd kan worden.
Bepaal ook welke (soorten) documenten in ieder geval goed geclassificeerd moeten worden en welke documenten eventueel gemist kunnen worden. Realiseer je in ieder geval dat 100% recall en precision nooit gehaald kunnen worden en dat het ondoenlijk is hiernaar te streven. In de praktijk zal dus de afweging gemaakt moeten worden tussen:
– Meer documenten geclassificeerd in een categorie, zodat er ook zoveel mogelijk relevante documenten geclassificeerd zijn. De niet-relevante documenten die ook in de betreffende categorie geclassificeerd zijn, worden daarbij voor lief genomen (hogere recall, lagere precision).
– Minder documenten geclassificeerd in een categorie, waarvan er wel zoveel mogelijk relevant zijn. De relevante documenten die niet in de betreffende categorie geclassificeerd zijn, worden daarbij voor lief genomen (lagere recall, hogere precision).

Bepaal per taxonomiecategorie hoeveel documenten er geclassificeerd zijn, en hoeveel van deze relevant en niet-relevant zijn. Met behulp van tabel 1 kan vervolgens de recall en precision berekend worden.

Tabel 1.

	Relevant	**Niet-relevant**	**Totaal**
Geclassificeerd	correct	niet-correct	totaal geclassificeerd
Niet-geclassificeerd	niet-correct	correct	totaal niet-geclassificeerd
Totaal	totaal relevant	totaal niet-relevant	

Hierbij is:

$$\text{Recall:} \quad \frac{\text{aantal geclassificeerd en relevant}}{\text{totaal aantal relevant}}$$

$$\text{Precision:} \quad \frac{\text{aantal geclassificeerd en relevant}}{\text{totaal aantal geclassificeerd}}$$

5.4 Verfijnen

Als uit de vorige fase gebleken is dat de recall en precision verbeterd moeten worden, ga je in de verfijnfase de relevante, niet-geclassificeerde documenten en geclassificeerde, niet-relevante documenten opsporen en beoordelen. Hier kunnen verschillende oorzaken voor zijn. Verkeerde classificatie kan een technische oorzaak hebben, bijvoorbeeld een specifiek bestandsformaat waarvan documenten niet geclassificeerd konden worden. Het kan ook zijn dat de classificatieregels niet goed genoeg zijn of dat de trainingsdocumenten niet voldoende zijn in termen van aantallen of kwaliteit.

Als je gebruik maakt van classificatieregels, zijn onderstaande aanpassingen mogelijk.

Mogelijke aanpassingen voor relevante, maar niet-geclassificeerde documenten:

- Voeg termen (synoniemen, afkortingen) toe.
- Maak gebruik van wildcards.
- Geef specifieke termen een hogere waarde.
- Verlaag de drempelwaarde van taxonomiecategorieën.

Mogelijke aanpassingen voor geclassificeerde, maar niet-relevante documenten:

- Verwijder te generieke term.
- Vermijd het gebruik van wildcards.
- Geef generieke termen een lagere waarde.
- Verhoog de drempelwaarde van taxonomiecategorieën.

Bij een lerend systeem dat werkt op basis van kenmerken, kun je aanpassingen maken door meer trainingsdocumenten aan het systeem toe te voegen en te beoordelen welk effect dit heeft en of dit een positief of negatief effect is. Daarnaast is het bij sommige systemen mogelijk om de drempelwaarde van taxonomiecategorieën te verhogen of te verlagen.

5.5 Implementeren

Totdat de automatische classificatie van de testset naar tevredenheid verloopt, zal het traject in de test- en verfijnfase blijven. Na deze fase kan het systeem daadwerkelijk geïmplementeerd gaan worden.
De implementatiefase gebeurt door de IT-afdeling van een organisatie of wordt uitbesteed aan een ingehuurde implementatiepartner.
Zorg ervoor dat je hierbij grip op je project houdt. Bewaak het projectplan en vastgestelde deadlines. Laat je op de hoogte brengen van de stand van zaken en tegenslagen in het traject. Blijf communiceren met je stakeholders en toekomstige eindgebruikers.

5.6 Onderhouden

Met de implementatie ben je er nog niet, onderhoud van een systeem is net zo belangrijk. Het onderhouden van een automatisch classificatiesysteem bestaat uit een regelmatige controle van de classificatieresultaten en hierop volgend het aanpassen van de taxonomie en classificatieregels.
Het periodiek herhalen van classificatietesten is een reactief gebeuren. Het is natuurlijk nog beter als het systeem proactief bijgestuurd kan worden. Het signaleren van nieuwe termen kan door het bijhouden van nieuwe documenten. Het blijft hierbij uiteraard moeilijk te voorspellen welke termen zullen blijven bestaan. Een andere manier om nieuwe termen te ontdekken is het bestuderen van het zoekgedrag van werknemers uit de organisatie. Indien zij op termen gaan zoeken die nog niet in de taxonomie of classificatieregels opgenomen zijn, is het het overwegen waard deze toe te voegen.
Voer ten slotte regelmatig gebruikerstesten uit om te zien of je gebruikers tevreden zijn met het nieuwe systeem.