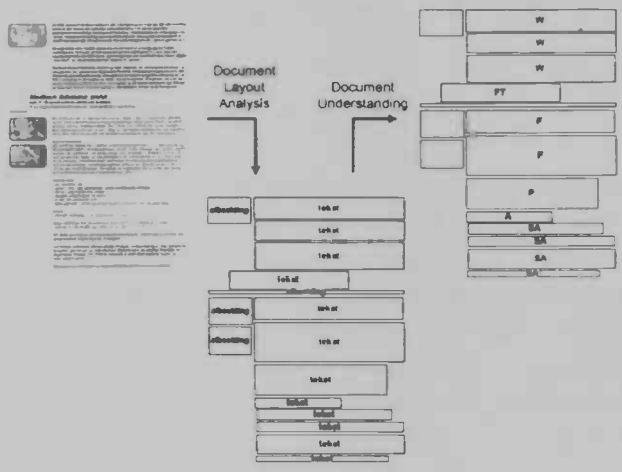
WORDT
NIET UITGELEEND

# Het geautomatiseerd labelen van personeelsadvertenties

MATCH care

## een toepassing van Document Understanding




Erwin Bouwman

augustus 2006

Rijksuniversiteit Groningen

Afstudeerbegeleider: Michael Wilkinson

# Samenvatting

MatchCare Information Services is een grote speler op de Nederlandse vacaturemarkt. Haar voornaamste activiteit is het verzamelen, verrijken en verkopen van vacature-informatie en het verrichten van onderzoek naar trends op de arbeidsmarkt. Deze vacature-informatie wordt uit diverse gedrukte media verkregen door het invoeren van de gepubliceerde personeelsadvertenties in een door MatchCare ontwikkelde applicatie. Dit proces, het vacatureverwerkingsproces, is verreweg het meest arbeidsintensieve proces binnen de organisatie en daarom wordt er veel tijd en aandacht besteed aan het efficiënter inrichten van het proces.
Met name wordt, door middel van diverse projecten, getracht subprocessen van het vacatureverwerkings-proces te automatiseren.

Het subproces waar dit afstudeerproject zich op richt is de labeling van vacaturetekst: deze eerste stap bij het invoeren van de tekst van een personeelsadvertentie in de applicatie bestaat uit het opdelen van de vacaturetekst in een zestal onderdelen (de labels), te weten Werkgeversomschrijving, Functietitel, Functieomschrijving, Profiel, Arbeidsvoorwaarden en Sollicitatie-informatie.
Het vermoeden lijkt gewettigd dat er bepaalde wetmatigheden m.b.t. de opbouw van de advertenties bestaan die het automatiseren van dit labelproces mogelijk maken.

Het voortraject van dit afstudeerproject richt zich op het concreet maken van deze wetmatigheden d.m.v. een kwantitatieve analyse van het domein van de personeelsadvertentie. Door middel van een uitgebreid literatuuronderzoek wordt de bruikbaarheid van verschillende Document Understanding-methodieken voor het automatiseren van de labeling onderzocht. Het hoofddoel van het onderzoek is het ontwerpen van een methodiek (en een bijbehorende informatiearchitectuur) dat van de gevonden wetmatigheden gebruik maakt bij het verrichten van de geautomatiseerde labeling. De ontwikkelde labelmethodiek maakt gebruik van een nieuwe representatie van de layout van een vacature waarbij de nadruk ligt op de wit-ruimte tussen tekstblokken: het *aandachtsprofiel*.

# Voorwoord

Deze afstudeerscriptie bevat een verslag van het afstudeeronderzoek dat in de periode van december 2001 tot maart 2003 is uitgevoerd bij MatchCare Information Systems.

Het onderzoek werd uitgevoerd om de haalbaarheid van een geautomatiseerde labelmethodiek als vervanger voor het handmatige labelproces van MatchCare te onderzoeken. Alhoewel de resultaten veelbelovend zijn zal de voorgestelde labelmethodiek nooit in de praktijk geïmplementeerd worden. Dat komt doordat MatchCare in het voorjaar van 2003 failliet is gegaan. Het faillissement werd grotendeels veroorzaakt door de economische recessie na de aanslagen van 9/11 en door een vernieuwde wetgeving die het copyright op personeels-advertenties veel strenger formuleerde waardoor het voor MatchCare veel moelijker werd vacaturegegevens te publiceren en te verkopen.

Door het faillissement van MatchCare en het vertrek van mijn oorspronkelijke afstudeerdocument, professor Spaanenburg, naar Zweden heb ik in het voorjaar van 2003 besloten fulltime te gaan werken. Een paar maanden geleden heb ik toch besloten mijn afstudeeronderzoek af te ronden.

Mijn dank gaat allereerst uit naar mijn afstudeerdocent Michael Wilkinson en als tweede naar alle mensen die al die jaren het vertrouwen en de overtuiging hebben gehouden dat ik ooit mijn Informatica-studie zou afronden.

# Inhoud


| | | |
|---|---|---|
| | Samenvatting | iii |
| | Voorwoord | iv |
| | Inhoud | v |
| **1** | **Inleiding** | **1** |
| 1.1 | De opdrachtgever | 2 |
| 1.2 | Het vacatureverwerkingsproces | 4 |
| 1.3 | Procesinnovatie | 8 |
| 1.3.1 | .....Marco Nijmeijer | 9 |
| 1.3.2 | .....Maryam Wilhelm | 16 |
| 1.4 | Probleemdefinitie | 18 |
| 1.5 | Hypothese | 19 |
| 1.6 | Beperkingen | 22 |
| 1.7 | Indeling van de scriptie | 22 |
| **2** | **Domeinverkenning** | **23** |
| 2.1 | Tekstuele wetmatigheden | 25 |
| 2.2 | Layoutgerichte wetmatigheden | 29 |
| 2.3 | Tekstopmaakgerichte wetmatigheden | 36 |
| 2.4 | Statistische methoden | 37 |

## 3 Document Processing 39

3.1 Document Analysis 39
3.2 Image Preprocessing 42
3.3 Document Layout Analysis 44
3.3.1 .....Segmentation 45
3.3.2 .....Block Type Classification 46
3.3.3 .....Documentmodellen 48

## 4 Document Understanding 51

4.1 Methodieken gebaseerd op rules 52
4.2 Methodieken gebaseerd op een documentmodel 53
4.2.1 .....Template Matching 54
4.2.2 .....Graph Matching 55
4.2.3 .....Language Models 55
4.2.4 .....Machine Learning 57
4.3 Text Mining 59

## 5 De labelingsmethodiek 61

5.1 Functionele beschrijving 61
5.2 Afwegingen bij de methodiekkeuze 62
5.3 Soorten eigenschappen 63
5.3.1 .....Geometrische eigenschappen 63
5.3.2 .....Tekstuele eigenschappen 64
5.3.3 .....Tekstopmaakgerichte eigenschappen 64
5.4 Soorten rules 65
5.5 Labelmethodiek in pseudo-code 65

## 6 Conclusies 69

**R**eferenties 71

**A**ppendices 77

A Lijst van gebruikte figuren 77
A Lijst van gebruikte tabellen 79
B Keywordlijsten 80
C Qualitative Spatial Relations 81

# 1 Inleiding

*"The most important weapon in the war of economic supremacy in the 21st century will be the organization of knowledge"*
***Alvin Toffler, Powershift (1990)***

De wereldberoemde en vaak geciteerde trendwatcher en toekomstanalist Alvin Toffler voorspelde al in 1990 wat de invloed van de verregaande informatisering op de economie zou zijn. Deze voorspellende woorden geven aan dat informatie in de huidige maatschappij het belangrijkste machtsmiddel is: vroeger telde alleen het bezit van geld, maar tegenwoordig kan je door informatie te bezitten geld verdienen. Volgens Toffler zullen veel machthebbers hun positie verliezen, omdat ze zich onvoldoende realiseren hoe belangrijk het machtsmiddel *informatie* is, en omdat ze de bevolking nog steeds min of meer als een soort proletariaat beschouwen en niet als een *cognitariaat*.

MatchCare Information Systems, de opdrachtgever van dit afstudeeronderzoek, speelt goed in op deze gedachte met vacature-informatie als kernproduct: de vacature-informatie wordt verzameld, verwerkt, verrijkt en verkocht. MatchCare geeft dus een goed voorbeeld van het omzetten van informatie in geld.

De laatste jaren van MatchCare staan vooral in het teken van uitbreiding van de klantenportefeuille. Door deze uitbreiding krijgt MatchCare steeds meer vacatures te verwerken. Dit werd initieel bewerkstelligd door het inzetten van meer personeel, maar zodra het aantal medewerkers (o.a. door ruimtegebrek) haar bovengrens had bereikt kwam het besef dat automatisering van het vacatureverwerkingsproces de juiste methode was om de vergrootte vacature-input te kunnen afhandelen.
Vanaf oktober 2000 is er door diverse medewerkers van MatchCare en door afstudeerders aan de RuG tijd en aandacht besteed aan het steeds verder automatiseren van het vacatureverwerkingsproces. Mijn afstudeeronderzoek maakt daar ook deel vanuit.

In dit inleidende hoofdstuk wordt achtereenvolgens dieper ingegaan op de opdrachtgever MatchCare, het vacatureverwerkingsproces dat het kernproces van MatchCare vormt, de inhoud en resultaten van voorgaande afstudeeronderzoeken bij MatchCare door Marco Nijmeijer en Maryam Wilhelm, de specifieke aanleiding voor mijn afstudeeronderzoek en de formulering van mijn onderzoekshypothese.

## 1.1 De opdrachtgever

*Matchcare is een jonge organisatie gespecialiseerd in het leveren van hoogwaardige, arbeidsgerelateerde informatie aan partijen die zakelijk actief zijn op de arbeidsmarkt. Matchcare levert vacaturedata, maar verzorgt ook rapportages over bijvoorbeeld trends in de arbeidsmarkt of over arbeidsmogelijkheden van cliëntpopulaties. Kortom, Matchcare maakt voor haar cliënten de arbeidsmarkt transparanter. Cliënten van Matchcare zijn mediabedrijven, gemeenten, uitvoerende instellingen, reïntegratiebedrijven, arbodiensten, jobsites en flexorganisaties. Bijvoorbeeld de jobsite Zoomnet.nl en de overheidsinstantie Centrum voor Werk en Inkomen (vacaturezuilen).*

***Werkgeversomschrijving uit een vacature van MatchCare, Nieuwsblad v/h Noorden (9 februari 2002)***

MatchCare Information Systems (hierna te noemen: MatchCare) is ontstaan uit de uitgeverij IntraMedium. Deze uitgeverij werd in 1995 gesticht door dhr. Keupink (nu General Manager van de productielocatie van MatchCare te Groningen) en hield zich bezig met het uitgeven van "De Sollicitatiekrant". Deze wekelijks uitgegeven krant bevatte ongeveer 2000 tot 3000 samenvattingen van vacatures voor middel en hoger opgeleiden die in de week ervoor waren gepubliceerd in de landelijke en regionale dagbladen, in diverse huis-aan-huisbladen en in diverse vakbladen. De samenvattingen van de verzamelde vacatures werden gemaakt door ruim 100 medewerkers die, verspreid door het hele land, de samenvattingen via modem verstuurden naar het kantoor in Groningen. Op het kantoor werden de samenvattingen vervolgens gecontroleerd op inhoud en volledigheid, ingedeeld op rubriek en opgeslagen in een centrale vacaturedatabase. "De Sollicitatiekrant" kende een wekelijkse oplage van een paar duizend exemplaren.
Na een overname door VNU Business Publications ging IntraMedium in mei 1997 verder onder de naam KeyMark Services. De activiteiten werden uitgebreid met het verzamelen en verkopen van arbeidsmarktgerelateerde informatie. Dit was een voordehand liggende stap omdat uit de centrale database veel informatie over de arbeidsmarkt afgeleid kon worden. De kernactiviteit bleef echter het verzamelen en publiceren van samenvattingen van zo veel mogelijk vacatures. "De Sollicitatiekrant" werd vanaf 1999 niet meer gepubliceerd omdat het wekelijkse aantal samenvattingen te groot werd om middels de krant te publiceren.
In juni 2000 is KeyMark Services overgenomen door AMSE (Advanced Matching Systems Europe), een holding van de participatiemaatschappij APM (Advanced Participation Management). KeyMark Services kreeg door

deze overname de merknaam "MatchCare" van AMSE mee en werd daarmee omgedoopt tot het huidige MatchCare Information Services.

Na de overname door AMSE is MatchCare zich gaan richten op het uitbreiden van haar klantenportefeuille. De belangrijkste bedrijfsdoelstelling was dan ook het binnenhalen van alle grote Nederlandse uitgevers als klant. Als eerste nieuwe, grote klant is de uitgeverij "De Telegraaf" binnengehaald. MatchCare verkoopt sinds 1 augustus 2000 wekelijks een vacaturedatabase met *fulltext* vacatures aan de Telegraaf: hierbij wordt de volledige tekst van een vacature opgenomen in de database in tegenstelling tot bij samenvattingen waarbij alleen een samenvatting van de tekstvelden wordt gemaakt. De tekst van een fulltext vacature wordt zeer gestructureerd en gesegmenteerd in de centrale vacaturedatabase opgeslagen: één vacaturerecord bevat bijna 150 tekst-, selectie- en boolean velden.
Samen met de Telegraaf is MatchCare de joint-venture Zoomnet gestart. De website www.zoomnet.nl maakte (de site is inmiddels om diverse redenen weer uit de lucht) het werkzoekenden en werkgevers mogelijk anoniem contact met elkaar te leggen. De werkzoekende kon op de site zijn opleiding, werkervaring en carrièrewensen vastleggen. Een door MatchCare ontwikkeld matchingalgoritme zocht de best passende openstaande vacatures bij de geregistreerde gegevens. Het matchen was mogelijk doordat gebruik werd gemaakt van gestandaardiseerde waarden voor de functie, opleiding en branche. Het toekennen van deze gestandaardiseerde waarden aan vacatures, de zogenaamde *verrijking*, maakt deel uit van het vacatureverwerkingsproces.

Na "De Telegraaf" zijn geleidelijk alle grote Nederlandse uitgevers van landelijke en regionale dagbladen als klant binnengehaald. De belangrijkste bedrijfsdoelstelling,

> *"als leverancier van vacatures, gegevens en rapportages over de huidige arbeidsmarkt de marktleider van Nederland worden en een zo groot mogelijk dekkingsgebied van informatiebronnen over de professionele arbeidsmarkt bereiken"*

was daarmee gerealiseerd. Deze uitgevers (Telegraaf, PCM, Wegener, VNU, en de groep Regio-i (waarin een aantal kleinere regionale uitgevers vertegenwoordigd zijn)) vormen, samen met on-line vacaturebanken, de belangrijkste afnemers van de geproduceerde vacature-informatie.
Het binnenhalen van deze uitgevers heeft niet alleen voor een stijgende omzet gezorgd, maar heeft er ook voor gezorgd dat MatchCare geleidelijk toegang kreeg tot alle Nederlandse landelijke en regionale dagbladen. Hierdoor kreeg MatchCare een steeds vollediger beeld van de Nederlandse arbeidsmarkt en konden de rapportages over de arbeidsmarkt steeds beter de daadwerkelijke situatie weergeven. Afnemers van deze rapportages zijn ondere anderen overheidsinstellingen.

Naast het hoofdkantoor van MatchCare, gevestigd in Zaltbommel, kent het bedrijf 2 productielocaties: de primaire productielocatie in Groningen, gevestigd in het Zernike Science Park, en een secundaire locatie in Deventer. Het hoofdkantoor in Zaltbommel vormt de thuisbasis voor het management en de administratieve, financiële en commerciële afdelingen. De locatie in Deventer huist een tiental medewerkers die alleen de vacatures uit de regionale dagbladen en enkele huis-aan-huisbladen in de buurt van Deventer en Utrecht verwerken.

De productielocatie in Groningen telt ongeveer 10 vaste medewerkers, meer dan 150 productiemedewerkers met flexibele werktijden. Naast deze medewerkers zijn er nog ongeveer 15 thuiswerkers die vanuit hun huis, verspreid door heel Nederland, vacatures verwerken. Kortgezegd houdt de productielocatie in Groningen zich bezig met het verzamelen, verwerken, verrijken en beheren van vacature-informatie uit de gedrukte media. Het is dus simpelgezegd een dataproductiebedrijf.

## 1.2 Het vacatureverwerkingsproces

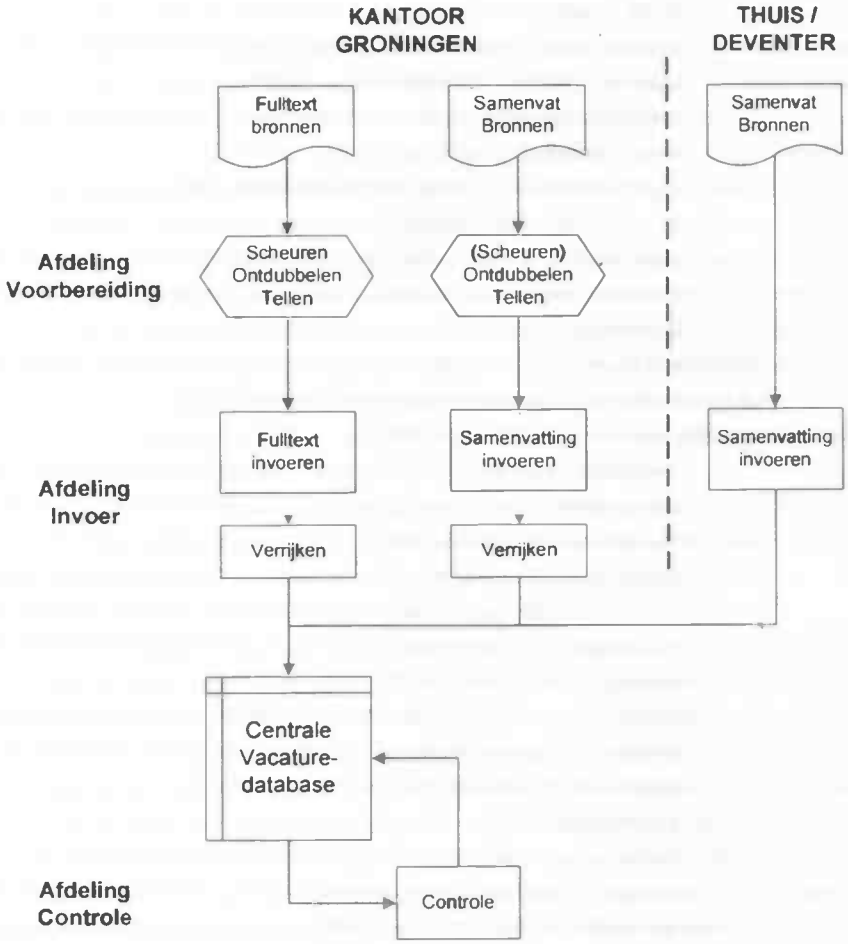
Het belangrijkste proces binnen de productielocatie te Groningen is het vacatureverwerkingsproces, dat vaak verkort het vacatureproces genoemd wordt. Het vacatureproces zoals het eruit zag voordat het deels geautomatiseerd werd wordt weergegeven in Figuur 1.1.

In het vacatureproces wordt onderscheid gemaakt tussen fulltextbronnen en samenvatbronnen. Fulltextbronnen zijn alle *bronnen* (landelijke en regionale dagbladen en vakbladen, maar geen huis-aan-huisbladen) waarvoor MatchCare een contract heeft met de klant om de complete vacaturetekst in het eindproduct, een wekelijks op te leveren vacaturedatabase, te verwerken. MatchCare heeft dergelijke contracten met alle grote Nederlandse uitgevers. Voorbeelden van fulltextbronnen zijn Dagblad v/h Noorden, de Telegraaf, de Volkskrant, het Algemeen Dagblad, de Alkmaarsche Courant, Dagblad de Limburger, het Brabants Dagblad, het Utrechts Dagblad en het Rotterdams Dagblad. Zoals in Figuur 1 te zien is worden de fulltextbronnen alleen op de productielocatie in Groningen ("kantoor Groningen") verwerkt.
Samenvatbronnen zijn alle bronnen waarvoor MatchCare de vacaturetekst alleen maar hoeft samen te vatten. MatchCare zou de tekst van deze vacatures trouwens niet eens fulltext mogen verwerken omdat de tekst van een vacature auteursrechtelijk beschermd is. Voorbeelden van samenvatbronnen zijn Adformatie, Binnenlands Bestuur, Textilia en de UK. Afnemers van de wekelijks op te leveren vacaturedatabase met deze samengevatte vacatures zijn met name on-line vacaturebanken.

Het vacatureproces bestaat uit de volgende onderdelen:

- **Scheuren, ontdubbelen en tellen:** op de productielocatie in Groningen wordt het invoeren van de vacatures voorafgegaan door een voorbereidingsproces uitgevoerd door de afdeling Voorbereiding. De invoermedewerkers die thuis hun vacatures invoeren verrichten deze werkzaamheden, voorzover nodig, zelf.
  Het voorbereidingsproces begint met het verzamelen van alle vacaturepagina´s die in de bron voorkomen, het zogenaamde *scheuren*. Dit wordt gedaan om de bron in porties met een klein aantal vacatures, meestal één vacaturepagina, te kunnen verdelen over de invoerders. Sommige samenvatbronnen bevatten zo weinig vacatures dat het scheuren niet nodig is.
  Bij de volgende stap, het *ontdubbelen*, wordt gekeken of er vacatures zijn die in meerdere bronnen voorkomen. Deze vacatures worden dubbele vacatures genoemd. Het komt veelvuldig voor dat in verschillende edities van een dagblad (bijvoorbeeld Brabants Dablad editie Tilburg en editie Den Bosch) dubbele vacatures voorkomen. Een dubbele vacature hoeft maar één keer in de invoerapplicatie

ingevoerd te worden en kan dan simpelweg gekopieerd worden. De dubbele vacature moet namelijk wel dubbel (met beide bronvermeldingen) in de vacaturedatabase opgenomen worden.
De laatste voorbereidingsstap betreft het *tellen* van het aantal vacatures dat voorkomt op een vacaturepagina of in een niet-gescheurde samenvatbron. Door het tellen wordt een beeld verkregen van het totale aantal te verwerken vacatures zodat de productieplanning daarop afgestemd kan worden.




**Figuur 1.1 - het vacatureproces van MatchCare.**

- **Fulltext/samenvatting invoeren**: voor het invoeren van fulltext vacatures en samenvattingen van vacatures wordt gebruik gemaakt van dezelfde invoerapplicatie, genaamd BEN. BEN bevat bijna 150 tekst-, selectie- en boolean velden waarin vacature-informatie verwerkt kan worden. Een vacaturerecord in BEN bevat een hoofdinvoergedeelte waarin de invoerder de volgende gegevens dient in te vullen: bron, datum van de bron, functietitel en gevraagde opleiding. Verder is BEN opgedeeld in een zestal tabbladen:
    - **Werkgever** – bevat de genoemde informatie over de werkgever (bijvoorbeeld over de afdeling waarbinnen de functie valt en in welke branche de werkgever werkzaam is);
    - **Functie** – bevat de functie-omschrijving (bijvoorbeeld welke verantwoordelijkheden bij de functie horen);
    - **Profiel** – bevat de gestelde eisen (bijvoorbeeld het aantal jaren werkervaring, ervaring met bepaalde applicaties of bepaalde karaktereigenschappen);
    - **Arbeidsvoorwaarden** – bevat de genoemde arbeidsvoorwaarden (bijvoorbeeld "auto van de zaak", "winstuitkering" of "PC privéplan");
    - **Sollicitatie** – bevat informatie over de sollicitatieprocedure (bijvoorbeeld "reageren binnen 14 dagen" of "brief in het Engels");
    - **Adverteerder** – bevat de adresgegevens van de adverteerder (in de meeste gevallen de werkgever, maar de adverteerder kan ook een uitzend-, detacherings- of werving & selectiebureau zijn).

  De opdeling in deze 6 zogenaamde *informatiecategorieën* is gekozen omdat de meeste vacatures zich ook houden aan deze zelfde logische indeling van de vacaturetekst.

  Er zijn drie soorten invoervelden binnen BEN:
    - Tekstvelden – deze velden bevatten stukken vrije tekst. Alle tabbladen met uitzondering van het tabblad Adverteerder bevatten een tekstveld en het tabblad Werkgever heeft er zelfs twee (één voor de werkgeversomschrijving en één voor de eventueel aanwezige afdelingsomschrijving). De invoerder probeert eerst alle informatie in het vacaturerecord op te nemen door middel van de selectie- en boolean velden, maar alle informatie die overblijft kan de invoerder volledig (in het geval van een fulltext vacature) of middels een samenvatting opnemen in het bijbehorende tekstveld;
    - Selectievelden – deze velden bevatten een waarde die uit een lijst van mogelijke waarden geselecteerd moet worden (bijvoorbeeld het opleidingsniveau, het aantal jaren ervaring dat vereist wordt of de branche van de werkgever). De meeste van deze selectievelden zijn verrijkingsvelden;
    - Boolean velden – met behulp van deze velden in de vorm van een checkbox wordt door de invoerder aangegeven of de vacaturetekst bepaalde veel voorkomende informatie bevat. Voorbeelden van veel voorkomende informatie zijn dat de functie recht geeft op een auto van zaak, dat de sollicitatiebrief in het Engels geschreven moet worden en dat de vacature de zin "acquisitie n.a.v. deze vacature wordt niet op prijs gesteld" bevat. De boolean velden versnellen het fulltext invoerproces doordat de invoerder deze veel voorkomende zinnen met behulp van een checkbox aan het vacaturerecord toevoegt in plaats van door het intypen ervan.

- **Verrijken**: Het *verrijken* van de vacature is een onderdeel van het invoerproces waarbij de invoerder informatie aan het vacature-record toevoegt die niet letterlijk in de vacaturetekst terug te vinden is. Het verrijken vereist dus een goede interpretatie van de vacature-informatie. Met de toegevoegde informatie wordt de vacature-informatie aan de ene kant geclassificeerd en aan de andere kant gestandaardiseerd. Het classificeren gebeurt door het selecteren van de sector waarin de functie valt (bijvoorbeeld "Juridisch", "Informatie en Communicatietechnologie" of "Commercieel"), de branche waarin de werkgever valt (bijvoorbeeld "Zakelijke dienstverlening", "Marketing en Communicatie" of "Overheid") en het opleidingsniveau dat van de sollicitant gevraagd wordt (bijvoorbeeld "WO", "HBO" of "MBO/HBO"). Door het classificeren wordt het mogelijk in de vacaturedatabase te zoeken op functies en/of werkgevers die in een bepaalde sector cq. branche vallen of op functies die een bepaald opleidingsniveau vereisen. Het standaardiseren gebeurt door de functietitel en de gevraagde opleiding te matchen met een functietitel en een opleiding uit een gestandaardiseerde lijst. Voor één en dezelfde functie kunnen namelijk verschillende functietitels bestaan (de functietitels "Schoonmaker" en "Objectverzorger" bijvoorbeeld duiden dezelfde functie aan evenals de functietitels "Programmeur C++" en "Applicatieprogrammeur"). Voor de gevraagde opleiding geldt ook dat er verschillende varianten en schrijfwijzen bestaan die één en dezelfde opleiding aanduiden. Het standaardiseren zorgt ervoor dat een zoekactie in de vacaturedatabase niet alleen de vacatures oplevert die de gezochte functietitel letterlijk bevatten, maar ook de vacatures met een afwijkende functietitel die dezelfde functie beschrijven.
  Zowel het classificeren als het standaardiseren zijn nodig om in de praktijk de gegevens in een on-line vacaturebank enigszins zoekbaar te maken. Ook vereenvoudigt het verrijkingsproces hierdoor het maken van bepaalde rapportages over bepaalde trends op de arbeidsmarkt (bijvoorbeeld over het wekelijkse aantal vacatures per sector, werkgeversbranche en opleidingsniveau).
  Na het invoeren en verrijken van de vacature slaat de invoerder het vacaturerecord op in de centrale vacaturedatabase.
- **Controle**: de afdeling Controle bestaat uit een klein aantal medewerkers, de controleurs, die steekproefsgewijs de vacaturerecords controleren op correcte invoer, volledigheid en juiste verrijking. De vacaturerecords worden zonodig door de controleurs gecorrigeerd en aangevuld.

Het verwerken van de vacatures vindt hoofdzakelijk plaats op vrijdag en op zaterdag. Dat komt doordat verreweg de meeste vacatures (ruim 90% van het totaal) verschijnen in de zaterdagedities van de bronnen. Dat er op vrijdag ook al vacatures verwerkt kunnen worden komt doordat MatchCare op de vrijdagnamiddag al een aantal katernen van de zaterdagedities van de Telegraaf, de Volkskrant, het Parool en het Dagblad van het Noorden ophaalt bij de distributiecentra en de vacaturepagina's van Dagblad de Limburger digitaal aangeleverd worden via een FTP-site. In een meting over het eerste kwartaal van 2002 kwam naar de volgende spreiding van de aanlevering van vacatures naar voren: op vrijdag ca. 28%, op zaterdag ca. 64% en op de rest van de dagen samengenomen ca. 8%. Deze onregelmatige spreiding zorgt ervoor dat de verwerking van fulltextvacatures geconcentreerd is op de vrijdagavond en de zaterdag.

Toen het fulltextproces opgezet werd (met de Telegraaf als enige fulltextbron) werden er wekelijks ca. 100 vacatures fulltext verwerkt. In diezelfde periode werden er wekelijks ca. 10.000 samengevat. Door het geleidelijk binnenhalen van alle grote Nederlandse uitgevers is het accent in het vacatureproces verschoven van het samenvatten naar de fulltextverwerking. Tegenwoordig worden er wekelijks ca. 5.000 fulltextvacatures en ca. 5.000 samengevatte vacatures aan de centrale vacaturedatabase toegevoegd.

Het proces dat de meeste aandacht krijgt is het fulltextproces. Dat heeft drie oorzaken:

- het is het proces dat de meeste tijd in beslag neemt: op vrijdag en zaterdag is er bijna geen aandacht voor het samenvatproces omdat de invoerders alle beschikbare tijd nodig hebben voor de verwerking van de fulltextvacatures. Het is zelfs lange tijd niet gelukt om alle fulltextvacatures op vrijdag en zaterdag te verwerken zodat er op maandag nog een restant verwerkt moest worden en er een nalevering van de vacaturedatabase plaats moest vinden;
- het is het proces dat dat het meeste verbeterd kan worden (dit wordt uitgelegd in sectie 1.3);
- het is het proces dat voor de meeste opbrengst zorgt.

## 1.3 Procesinnovatie

Een paar maanden nadat begonnen was met de fulltextverwerking van de Telegraaf werd duidelijk dat, wilde MatchCare meer bronnen op de fulltextmanier gaan verwerken, het nodig was óf meer invoerders in dienst te nemen óf het fulltextproces te versnellen. De bedrijfsleiding koos ervoor het aantal invoerders aanzienlijk uit te breiden maar tegelijkertijd ook te gaan werken aan een langetermijnoplossing: het versnellen van het fulltextproces door het (gedeeltelijk) automatiseren ervan.

Het bedrijf koos ervoor het fulltextproces en niet het samenvatproces te innoveren om de volgende redenen:

- het invoeren van een fulltextvacature kostte 2 à 3 keer zoveel tijd als het samenvatten van een vacature zodat er bij het fulltextproces de meeste tijdswinst geboekt zou kunnen worden. Het verwerken van één fulltextvacature kostte een invoerder gemiddeld ruim 10 minuten zodat een invoerder maar 4 à 6 vacatures per uur kon verwerken. Dit in tegenstelling tot het samenvatten van een vacature waarmee een invoerder gemiddeld nog geen 5 minuten bezig was (gemiddeld 15 vacatures per uur). Verreweg het grootste gedeelte van de tijd bij het invoeren van een fulltextvacature was de invoerder bezig met het letterlijk overtypen van de vacaturetekst;
- de bedrijfsleiding voelde intuïtief aan dat het samenvatten van een vacature moeilijker te automatiseren zou zijn dan het digitaal klaarzetten van de letterlijke vacaturetekst voor de invoerder en daarmee zou al een aanzienlijke tijdswinst geboekt kunnen worden;
- de bedrijfsleiding had vanaf het begin van het fulltextproces de verwachting dat het zou uitgroeien tot het grootste en belangrijkste proces van MatchCare;

In oktober en november 2000 zijn respectievelijk Marco Nijmeijer en Maryam Wilhelm een onderzoek begonnen naar de automatisering van het fulltextproces als afstudeeronderzoek aan de Rijksuniversiteit Groningen. Marco Nijmeijer ontwikkelde een nieuw fulltextproces waarbij de vacaturetekst digitaal wordt klaargezet door de afdeling Voorbereiding waarna de invoerder in de invoerapplicatie BEN alleen nog maar de selectievelden en boolean velden hoeft bij te werken [1]. Maryam Wilhelm onderzocht de mogelijkheid van het toepassen van text-mining op het fulltextproces [2]. Daarbij onderzocht ze specifiek de mogelijkheid om de informatiecategorie Profiel geautomatiseerd te detecteren aan de hand van de vacaturetekst. In de volgende twee paragrafen worden de onderzoeken nader besproken.

## 1.3.1 Marco Nijmeijer

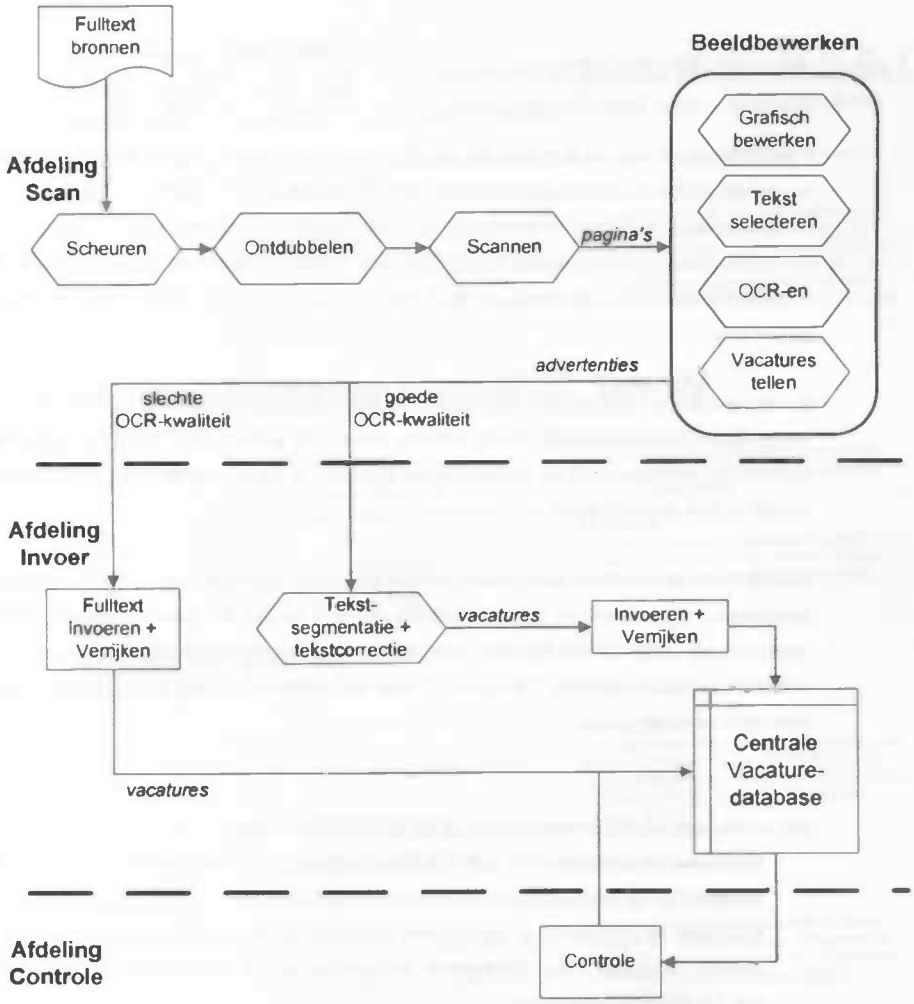
Marco Nijmeijer is van oktober 2000 tot juni 2002 als afstudeerder bij MatchCare betrokken geweest. Zijn onderzoek richtte zich op het automatiseren van een nieuwe verwerkingsmanier van fulltextvacatures. Zijn doelstelling was het "tikken" te vervangen door het "klikken", met andere woorden: het letterlijk intypen van de vacaturetekst moest vervangen worden door een werkwijze waarbij de verwerking grotendeels met de muis (het "klikken") kon gebeuren. Het door Marco ontworpen nieuwe fulltext vacatureproces is te zien in Figuur 1.2.

In het vernieuwde fulltextproces zorgt de Scanafdeling (voorheen de afdeling Voorbereiding) ervoor dat de vacaturetekst klaargezet wordt voor de afdeling Invoer. Dat gebeurt door de stappen scheuren, ontdubbelen, scannen en beeldbewerken. De afdeling Invoer zorgt ervoor dat de opgeleverde vacaturetekst op de juiste manier in BEN verwerkt wordt.

Voordat het vernieuwde vacatureproces beschreven kan worden moet eerst het verschil tussen een personeelsadvertentie en een vacature duidelijk gemaakt worden. Een personeelsadvertentie is de gehele communicatie-uiting van één bepaalde adverteerder. In een personeelsadvertentie kan de adverteerder meerdere vacatures opnemen. Zie Figuur 1.3 voor een voorbeeld van een personeelsadvertentie met meerdere vacatures.

Het vernieuwde vacatureproces bestaat uit de volgende onderdelen:

- **Scheuren en ontdubbelen**: deze voorbereidingsstappen zijn onveranderd gebleven. Het tellen van de vacatures op de vacaturepagina's is echter iets naar achteren in het proces geschoven;
- **Scannen**: de gescheurde en ontdubbelde vacaturepagina's worden gescand met een scanner die speciaal ontwikkeld is voor het scannen van krantenpagina's (A2-formaat). Per vacaturepagina wordt een TIF-bestand aangemaakt;
- **Beeldbewerken**: het beeldbewerken is de processtap waarin de gescande vacaturepagina in TIF-formaat wordt omgezet naar een aantal RTF-bestanden (één bestand per personeelsadvertentie) waarin de ge-OCR-de tekst wordt geplaatst. Het beeldbewerken wordt uitgevoerd in de door Marco ontwikkelde applicatie LimCut. Marco heeft deze naam gekozen omdat de applicatie in eerste instantie ontwikkeld was om de gedownloade vacaturepagina's van Dagblad de Limburger op te knippen in tekstsegmenten. Dagblad de Limburger begon al enige maanden voor de aanschaf van de scanner zelf de vacaturepagina´s te scannen en voor MatchCare klaar te zetten op een FTP-site. Daarmee was Dagblad de Limburger de eerste bron die digitaal verwerkt werd.




**Figuur 1.2 - resultaat van het afstudeeronderzoek van Marco Nijmeijer: het nieuwe fulltextproces.**

Het beeldbewerken bestaat uit 4 onderdelen:

1. **Grafisch bewerken:** allereerst kan de beeldbewerker zonodig gedeeltes van de gescande vacaturepagina bewerken met een aantal grafische operaties. Het doel van het grafisch bewerken is de gescande vacaturepagina zodanig te bewerken dat elk tekstsegment (als afbeelding) zo optimaal mogelijk aan de OCR-applicatie aangeboden wordt. Optimaal in dit verband houdt in dat de vacaturetekst zo zwart mogelijk op een zo wit mogelijke achtergrond staat. De operaties worden steeds uitgevoerd op een door de beeldbewerker gemaakte selectie op de gescande vacaturepagina.

De beschikbare operaties zijn:

- Thresholden: met deze operatie stelt de beeldbewerker een threshold in zodat de grijswaarden lager dan de threshold worden omgezet in 0 (volledig wit) en de grijswaarden gelijk aan en groter dan de threshold worden omgezet in 1 (volledig zwart). Het doel van deze operatie is het tekstsegment (als afbeelding) met een zo groot mogelijk kleurverschil tussen de tekst en de achtergrond aan te kunnen bieden aan de OCR-applicatie;
- Inverteren: met deze operatie inverteert de beeldbewerker de grijswaarden. Dit is nodig als er witte tekst op een donkere achtergrond geplaatst is;
- Vullen met wit: met deze operatie maakt de beeldbewerker de selectie helemaal wit. Dit kan handig zijn als er midden in een tekstsegment kleine afbeeldingen staan. Door het wissen van dat soort afbeeldingen verklein je de kans dat de OCR-applicatie de afbeelding probeert om te zetten in tekst.

2. **Tekst selecteren**: na het grafisch bewerken geeft de beeldbewerker de leesvolgorde binnen de personeelsadvertentie aan door tekstsegmenten te selecteren (door er met de muis een blok omheen te trekken) en deze segmenten per personeelsadvertentie met elkaar te verbinden in de leesvolgorde. In deze stap wordt dus voor het eerst onderscheid gemaakt tussen de verschillende personeelsadvertenties op de vacaturepagina;
3. **OCR-en**: na het selecteren van de tekstsegmenten worden deze aangeboden aan de OCR-applicatie. De afkorting OCR staat voor Optical Character Recognition en staat voor een methodiek waarmee een afbeelding die tekst bevat wordt omgezet in digitale tekst door het herkennen van de vormen van de karakters uit de afbeelding. Elke personeelsadvertentie wordt in de vorm van een aantal tekstsegmenten (als afbeelding) aangeboden aan de OCR-applicatie. Als resultaat wordt per personeelsadvertentie één RTF-bestand met daarin de ge-OCR-de tekst opgeleverd. Als de OCR-applicatie een acceptabel resultaat oplevert gaat de personeelsadvertentie verder het proces in bij de stap "Tekstsegmentatie + tekstcorrectie". Zo niet, dan moet de vacature op de oude manier fulltext ingevoerd worden in BEN (dus door het letterlijk overtypen van de complete vacaturetekst);
4. **Vacatures tellen**: als laatste handeling telt de beeldbewerker het aantal vacatures op de vacaturepagina. Dat wordt gedaan om een beeld van het totale aantal te verwerken vacatures te verkrijgen zodat de productieplanning daarop afgestemd kan worden.

Na het beeldbewerken worden de personeelsadvertenties in de vorm van RTF-bestanden overgedragen aan de afdeling Invoer voor de verdere verwerking ervan. Tijdens het gehele proces hebben de beeldbewerkers en invoerders de originele vacaturepagina bij de hand zodat ze daarop kunnen terugvallen in het geval van onduidelijkheden;

- **Tekstsegmentatie en tekstcorrectie**: in deze stap splitst de invoerder de tekst van de personeelsadvertentie op in de 6 informatiecategorieën die MatchCare hanteert: Werkgeversomschrijving, Functie-omschrijving, Profiel, Arbeidsvoorwaarden, Sollicitatie-informatie en Adverteerdersgegevens. Er wordt een extra segment aangemaakt voor de functietitel. Het segmenteren is als extra stap vóór het invoeren in BEN opgenomen in het proces omdat de tekst van een personeelsadvertentie, afgaande op intuïtie, snel is op te delen in de 6 categorieën zonder de tekst in zijn geheel te hoeven lezen. Deze verwerkingsmethode is intuïtief sneller dan de methode waarin de gehele tekst in één stap wordt opgedeeld in alle mogelijke invoervelden. De praktijk heeft inmiddels uitgewezen dat deze theorie valide is.

**De Gemeentelijke Geneeskundige en Gezondheidsdienst van Amsterdam**

vraagt voor het Streeklaboratorium voor de Volksgezondheid en Rijksinstituut voor de Virologie te Amsterdam zo spoedig mogelijk een microbiologisch analist en een analist

**Algemene informatie**
Het Streeklaboratorium van de GG&GD te Amsterdam verricht microbiologisch onderzoek voor verschillende instellingen, zoals de Polikliniek voor Geslachtsziektebestrijding en de Afdeling Infectieziekten van de GG&GD, verpleeghuizen en huisartsen in Amsterdam. Virologisch onderzoek wordt uitgevoerd voor een aantal ziekenhuizen in Noord-Holland en de Flevolandse ziekenhuizen, voor welke ook het klinisch bacteriologisch onderzoek wordt gedaan. Daarnaast wordt onderzoek verricht op het gebied van geslachtsziekten en de epidemiologie van infectieziekten. Het Streeklaboratorium werkt samen met verschillende Nederlandse en buitenlandse instituten.

# microbiologisch analist *(m/v)*

voor 16 uur per week

**Taakomschrijving**
Het uitvoeren van dagelijkse werkzaamheden als analist op de afdelingen: klinische bacteriologie, TBC, parasitologie, serologie en moleculaire routinediagnostiek.

**Functie-eisen en voorwaarden**
- HBO-A of HLO medische microbiologie dan wel een gelijkwaardige opleiding
- u dient goed in teamverband te kunnen werken;
- u wordt geacht in weekend- en bereikbaarheidsdiensten te participeren.

**Aanstelling**
Geschiedt aanvankelijk op basis van een tijdelijke aanstelling voor bepaalde tijd van 7 maanden - bij goed functioneren - gevolgd door een vaste aanstelling.

# analist *(m/v)*

voor 24-32 uur per week

**Taakomschrijving**
Het uitvoeren van dagelijkse werkzaamheden ter behoeve van de virologische en serologische diagnostiek.

**Functie-eisen en voorwaarden**
HBO-A of HLO, dan wel gelijkwaardige opleiding, aantoonbare ervaring als analist op bovengenoemde terreinen strekt tot aanbeveling.

**Aanstelling**
Geschiedt aanvankelijk op basis van een tijdelijke aanstelling voor bepaalde tijd van 7 maanden gevolgd - bij goed functioneren - door een tijdelijke aanstelling voor onbepaalde tijd. De totale aanstellingsduur bedraagt vier jaar.

*Het gemeentelijk beleid is erop gericht dat de samenstelling van het multidisciplinaire personeel een afspiegeling wordt van de bevolkingssamenstelling. Daarom worden vrouwen, gehandicapten en leden van etnische minderheidsgroepen uitdrukkelijk voor sollicitatie uitgenodigd. Bij het vinden van de geschikte vrouw geeft zij de voorkeur. Bezwaren naar aanleiding van deze aanbeveling worden niet op prijs gesteld.*

**Salaris**
Voor beide functies bedraagt het salaris afhankelijk van werkervaring minimaal € 1.822,- en maximaal € 2.654,- dienstijdelijkloon (salarisgroep 8) brutto per maand bij een volledige werkweek.

**Informatie**
Voor verdere informatie omtrent de functie microbiologisch analist kunt u contact opnemen met: mevrouw I. L. mde, hoofdanalyse bacteriologie telefoonnummer (020) 5555 301 of met de heer Dr. P.G.H. Peerboom, microbiolog, telefoonnummer (020) 5555 224.
Voor verdere informatie omtrent de functie analist kunt u contact opnemen met mevrouw K. Adama, hoofdanalyse virologisch serologie, telefoonnummer (020) 5555 304 of met mevrouw Drs. J. Spaargaren, hoofd Streeklaboratorium, telefoonnummer (020) 5555 364.

**Sollicitatie**
Uw schriftelijke sollicitatie kunt u - voor 31 mei 2002 - richten aan de afdeling Personeel, Organisatie & Opleiding van de GG&GD, t.a.v. mevrouw A.J.M. van Rijswijk, P&O-adviseur, Nieuwe Achtergracht 100, 1018 WT Amsterdam. Ook kunt u uw brief met c.v. als bijlage (in MS-Word) bij een e-mailbericht voegen via adres vacatures@ggd.amsterdam.nl

**Figuur 1.3 - een voorbeeld van een personeelsadvertentie met meerdere vacatures.**

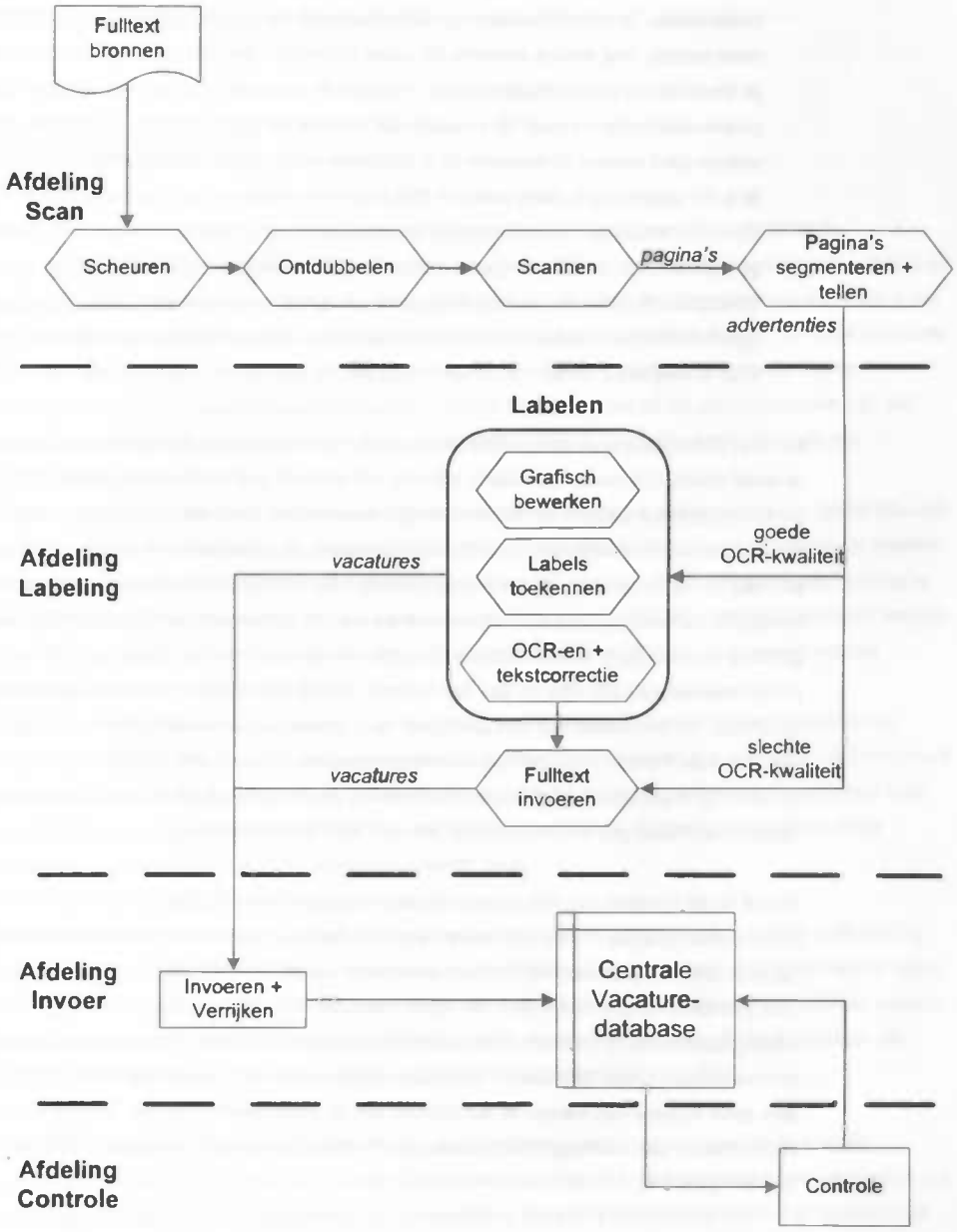
De segmentatie wordt uitgevoerd in de applicatie YAN, hetgeen een afkorting is voor Yet Another NewLimi (wederom een naam die gebaseerd is op Dagblad de Limburger), ontwikkeld door Maryam Wilhelm. YAN is opgedeeld in twee verticale schermhelften: in de linkerhelft staat de platte tekst (dus zonder opmaak) zoals die is opgeleverd door de OCR-applicatie en in de rechterhelft staan de 6 tabbladen Werkgever, Functie, Profiel, Arbeidsvoorwaarden, Sollicitatie en Adverteerder (per

informatiecategorie een tabblad) en een extra veld voor de functietitel. De invoerder selecteert in de linkerhelft een stuk tekst en opent vervolgens door met de rechtermuisknop op de selectie te klikken het contextmenu. In het contextmenu geeft de invoerder aan naar welk tabblad het stuk tekst verplaatst moet worden, met andere woorden, tot welke informatiecategorie het stuk tekst behoort. In YAN moet de invoerder de personeelsadvertentie opsplitsen in vacatures: per vacature die deel uitmaakt van de personeelsadvertentie zorgt de invoerder dat de tekst die bij de vacature hoort in de rechterhelft van het scherm staat waarna de invoerder de 6 tabbladen en de functietitel exporteert naar de invoerapplicatie BEN. De opsplitsing is nodig omdat in BEN alleen met vacatures, en niet met personeelsadvertenties die meerdere vacatures kunnen bevatten, gewerkt kan worden. In deze stap moet de invoerder tevens de ge-OCR-de tekst zonodig corrigeren zodat de tekst in BEN geen OCR-fouten bevat;

- **Invoeren**: de invoerstap is in zoverre gewijzigd dat de invoerder geen vacaturetekst meer hoeft over te typen in BEN. De complete vacaturetekst staat in het nieuwe fulltextproces opgedeeld in de 6 tabbladen klaar in BEN zodat de invoerder alleen nog maar de selectie- en boolean velden hoeft bij te werken.

Het is in één oogopslag te zien dat het invoeren van fulltextvacatures door dit nieuwe fulltextproces is versneld doordat de meest tijdrovende activiteit, het letterlijk overtypen van de vacaturetekst, is weggevallen. De voorbereidende stappen zijn echter meer tijd gaan vergen. De invoertijd die een invoerder nodig had voor het invoeren van een fulltextvacature in BEN (dus zonder de voorbereidende stappen) op de oude manier bedroeg ca. 10-15 minuten. Bij het nieuwe fulltextproces moeten er drie deelprocessen gemeten worden (de benodigde tijd voor het scannen is verwaarloosbaar en de scheur- en ontdubbelprocessen worden nooit geklokt): de benodigde tijd voor het beeldbewerken in LimCut, voor het opdelen van de personeelsadvertentie in YAN en voor het invoeren in BEN. Een meting over 10 weken gaf de volgende resultaten: het beeldbewerken kost gemiddeld ca. 1 minuut per personeelsadvertentie, het opdelen kost gemiddeld ca. 4 minuten per personeelsadvertentie en het invoeren kost gemiddeld 3 à 4 minuten per vacature. Er is dus een duidelijke tijdswinst behaald, want de benodigde tijd voor het invoeren van een vacature is omlaag gebracht van ruim 10 minuten naar 8 à 9 minuten.

Al snel na de invoering van het nieuwe fulltextproces werd door mij opgemerkt dat het opdelen van de personeelsadvertentie in de 6 informatiecategorieën beter en efficiënter gedaan zou kunnen worden tijdens het beeldbewerkingsproces. De beeldbewerker heeft namelijk de afbeelding van de personeelsadvertentie voor zich op het beeldscherm staan terwijl de invoerder in YAN alleen maar de platte ge-OCR-de tekst als uitgangspunt heeft. In YAN kan de invoerder dus geen gebruik maken van de layout (opmaak) van de personeelsadvertentie. Hoofdstuk 2 (Domeinverkenning personeelsadvertenties) laat zien dat de layout een zeer grote bijdrage kan leveren bij het opdelen van de personeelsadvertentie. De bedrijfsleiding en Marco waren hiervan ook overtuigd zodat Marco het fulltextproces nogmaals aanpaste. Het resultaat daarvan is te zien in Figuur 1.4.




**Figuur 1.4 - het vernieuwde fulltextproces nadat het opdelen van de personeelsadvertentie is verplaatst van YAN naar LeO, de nieuwe labelapplicatie.**

Dit zijn de verschillen tussen het 'oude' vernieuwde fulltextproces en het 'nieuwe' vernieuwde fulltextproces:

- **Pagina's segmenteren en tellen**: na het scannen van de vacaturepagina's wordt het aangemaakte TIF-bestand eerst in de applicatie LimCut gesegmenteerd in losse personeelsadvertenties. Door de segmentatiestap wordt er voor elke afzonderlijke personeelsadvertentie een TIF-bestand aagemaakt. Deze stap is ingevoegd om de volgende stap, het labelen, overzichtelijker te maken;
- **Labelen**: het *labelen* van een personeelsadvertenties is de nieuwe term voor het segmenteren van de personeelsadvertentie. Het labelen wordt uitgevoerd in de nieuwe applicatie LeO, hetgeen staat voor Labelen en OCR-en. De applicatie LeO is een kruising tussen LimCut en YAN, want in LeO wordt zowel het grafisch bewerken, het OCR-en als het segmenteren van de personeelsadvertentie uitgevoerd. Het grafisch bewerken blijft overanderd maar wordt uitgevoerd op een TIF-bestand waarin maar één personeelsadvertentie is opgenomen. Na het grafisch bewerken kent de labelmedewerker labels toe aan de tekstsegmenten waaruit de personeelsadvertentie bestaat. In LeO wordt een iets andere segmentatie gemaakt dan in YAN gebeurde: in LeO worden de informatiecategorieën Sollicitatie-informatie en Adverteerdersgegevens in hetzelfde segment opgenomen. Dat wordt gedaan omdat in bijna 95% van alle personeelsadvertenties deze twee informatiecategorieën niet strikt gescheiden worden en dus lastig en heel bewerkelijk op een grafische manier te segmenteren zijn. In YAN lukte het segmenteren wel omdat het segmenteren van ge-OCR-de tekst eventueel zelfs op woordniveau gedaan kan worden. Een ander verschil van het labelen in LeO ten opzichte van het segmenteren in YAN is het feit dat bij het labelen ook de functietitel een apart label krijgt, alhoewel de functietitel geen informatiecategorie is. In LeO worden dus de volgens labels gebruikt: W – Werkgeversomschrijving, FT – FunctieTitel, F – Functie-omschrijving, P – Profiel, A – Arbeidsvoorwaarden en SA – Sollicitatie-informatie/Adverteerdersgegevens. Net als in YAN moet de personeelsadvertentie bij het labelen worden opgesplitst in afzonderlijke vacatures. De labelmedewerker labelt een vacature door de tekstsegmenten die bij één vacature horen te selecteren (door er met de muis een blok omheen te trekken), met de rechtermuisknop het contextmenu te openen en daaruit het benodigde label te selecteren. Een voorbeeld van een gelabelde vacature is te zien in Figuur 1.5. Na het toekennen van de labels zet de labelmedewerker de OCR in werking. Het OCR-resultaat wordt getoond in een tekstvenster waarna de labelmedewerker de tekst zonodig corrigeert. Na de tekstcorrectie wordt de vacature geëxporteerd naar BEN. In het geval van een personeelsadvertentie met meerdere vacatures wist de labelmedewerker de selecties die slechts bij één vacature horen en selecteert de tekstsegmenten die bij de volgende vacature horen;
- **Fulltext invoeren**: Als na het uitvoeren van de OCR blijkt dat de tekstkwaliteit te slecht is om te gaan corrigeren, of als al voor het labelen de verwachting bestaat dat een bepaalde personeelsadvertentie slechte OCR-kwaliteit gaat opleveren (bijvoorbeeld doordat er een raster op de achtergrond is geplaatst waardoor de OCR-applicatie de tekst slecht kan herkennen), dan wordt besloten de vacaturetekst fulltext in te voeren. Dat wordt door de labelmedewerkers in LeO gedaan zodat de invoerafdeling altijd een opgedeelde vacaturetekst zonder fouten als uitgangspunt in BEN heeft staan. Zo wordt ten allen tijde, dus ook bij het fulltext invoeren, het labelen van de personeelsadvertentie uitgevoerd door de afdeling Labeling en het invoeren in BEN door de afdeling Invoer;
- **Invoeren**: de enige wijziging in deze stap is dat de tekst van de informatiecategorieën Sollicitatie-informatie en Adverteerdersgegevens als één segment wordt gelabeld. De invoerder zal dus zelf nog de tekst moeten verdelen over de twee tabbladen.

Ook deze vernieuwing bracht een duidelijke tijdswinst met zich mee: het labelen (voorheen het beeldbewerken samen met het opdelen) kost nu 3 à 4 minuten in plaats van ca. 5 minuten. Het invoeren in BEN is qua benodigde tijd onveranderd gebleven.

## 1.3.2 Maryam Wilhelm

Maryam Wilhelm is haar onderzoek bij MatchCare in november 2000 begonnen. Ze onderzocht de mogelijkheid om text-mining in het nieuwe vacatureproces te integreren en ontwikkelde daar een architectuur voor, genaamd PressAnalyzer. Volgens de onderzoekshypothese kan door het toepassen van text-mining het menselijke interpreteren van vacaturetekst benaderd worden en kan zo een gedeelte van het vacatureproces geautomatiseerd worden. Nauwkeuriger gezegd probeerde ze in haar onderzoek een eerste stap te ontwikkelen voor het geautomatiseerd herkennen van het profielsegment uit de gehele vacaturetekst. Het onderzoek vormde dus een eerste stap in het geautomatiseerd labelen van de gehele tekst van een personeelsadvertentie.
Het profielsegment kan – evenals de andere segmenten waaruit een vacature is opgemaakt – geïdentificeerd worden door de dichtheid waarmee bepaalde karakteristieke woorden in dat segment voorkomen te bepalen. Het is dus noodzakelijk van tevoren door statistische analyse van een testset van personeelsadvertenties te achterhalen welke woorden karakteristiek zijn voor het profielsegment. Deze karakteristieke woorden worden keywords genoemd. Om een voor de hand liggend voorbeeld te geven: één van de keywords van het profielsegment is het woord "ervaring".
Ook de dichtheid waarin deze keywords in een bepaald tekstsegment voorkomen wordt gebruikt bij het labelen van het segment, omdat het voor kan komen dat één van de keywords ook buiten het profielsegment wordt gebruikt. Het keyword wordt dan wel geteld, maar de dichtheid is laag omdat het keyword niet vaak voorkomt in dat segment. Door te zoeken naar het tekstsegment waarin de profielkeywords een hoge dichtheid vertonen wordt het profielsegment geïdentificeerd.

De theoretische onderbouwing van PressAnalyzer concentreert zich met name op het RAPIER-systeem [3]. Het RAPIER-systeem kan, met als basis een aantal geannoteerde voorbeelden, een ongestructureerde keten van woorden omzetten in een gestructureerde database. De RAPIER-methodiek kan tevens gebruikt worden om stukken tekst met een bepaalde inhoud met een zekere precisie te lokaliseren. Deze eigenschap zou gebruikt kunnen worden bij het geautomatiseerd labelen van personeelsadvertenties.

Maryam heeft een pilotversie van PressAnalyzer geïmplementeerd om te onderzoeken of de RAPIER-methodiek met succes toegepast kan worden op personeelsadvertenties. Met deze pilotversie werd alleen het profielsegment van de personeelsadvertentie gedetecteerd. Maryam's bevindingen waren dat de pilotversie een tijdswinst van 200% opleverde (PressAnalyzer verwerkte de personeelsadvertenties dus drie keer zo snel als handmatig gedaan kon worden) maar slechts in 30% van de personeelsadvertenties het profielsegment correct detecteerde. De twee belangrijkste oorzaken van het niet-correct detecteren waren:

1. Het profielsegment van de personeelsadvertenties bevatte geen van de gezochte keywords. In dat geval werd er geen segment gedetecteerd;
2. De overgang tussen verschillende segmenten en de grootte van het profelsegment werd niet juist gedetecteerd. Hierdoor werd er meer dan alleen het daadwerkelijke proifielsegment gelabeld als "Profiel".

Samen werken aan de zorg voor onze bewoners waarbij de wens van onze bewoners centraal staat. Dat is waar wij, "Aeneas" verpleeg- en kloosterverzorgingshuis, naar streven. "Aeneas" is stark in ontwikkeling. De 350 medewerkers zijn werkzaam op drie locaties in Breda: de hoofdlocatie bevindt zich aan de Zaart, het kloosterverzorgingshuis Liesbosch aan de Liesbosstraat en de verpleeghuisdependance in Heksenwiel (Haagse Beemden).

W

Voor de sector Behandelzaken zijn wij op zoek naar twee

**psychologen m/v** FT

in een dienstverband van **onbepaalde tijd**, waarbij het **aantal uren bespreekbaar** is.

**Functie-inhoud:** F

U maakt deel uit van het multidisciplinaire behandelteam. U verricht psychologisch onderzoek, observatie en u behandelt en begeleidt bewoners. U geeft behandel- en begeleidingsadviezen en u begeleidt afdelingsteams in het omgaan met gedragsproblematiek. De zorg wordt zowel binnen het verpleeghuis als daarbuiten in substitutieprojecten geboden.

**Functie-eisen:** P

U heeft een universitaire opleiding psychologie afgerond, afgestudeerd in de klinische psychologie, psychogerontologie of neuro- en revalidatie-psychologie, met ervaring in het werken met psychogeriatrische bewoners. U bent een geringstreerd gezondheidszorgpsycholoog of in de overgangsregeling tot gezondheidszorgpsycholoog. U heeft ervaring met psychodiagnostiek en behandelmethoden bij ouderen. U bent gericht op multidisciplinaire samenwerking en daarnaast bent u in staat om zelfstandig te werken.

**Arbeidsvoorwaarden:** A

De arbeidsvoorwaarden zijn conform de CAO Verpleeg- en Verzorgingshuizen. Salarisring volgens FWG functiegroep 60 (minimaal € 2.230,- en maximaal € 3.477,- op basis van een 36-urige werkweek). Indeling volgens FWG 3.0 zal nog plaatsvinden.

**Informatie en sollicitatie:** SA

Voor meer informatie kunt u contact opnemen met Monique Nuijten, psychologe, te bereiken via het secretariaat Behandelzaken, tel. 076-5266600. Uw schriftelijke sollicitatie kunt u tot **28 mei 2002** richten aan Verpleeghuis Aeneas, t.a.v. Gudrun Billet, afdeling P&O, Zaart 35, 4819 ED Breda.

Aeneas

Figuur 1.5 - een gelabelde personeelsadvertentie. Met de afkortingen W, FT, F, P, A en SA worden achtereenvolgens de labels Werkgeversomschrijving, FunctieTitel, Functie-omschrijving, Profiel, Arbeidsvoorwaarden en Sollicitatie-informatie/Adverteerdersgegevens aangeduid.

Maryam's conclusies waren als volgt:

- De meest effectieve en succesvolle benadering van de menselijke interpretatie van tekst lijkt gedaan te kunnen worden met een rule-based benadering. Echter, de implementatie is complex en tijdrovend en vereist een brede kennis van logisch programmeren en machine learning methoden;
- Het geautomatiseerd labelen wordt nauwkeuriger als het algoritme probeert alle labels in de vacaturetekst te vinden en niet slechts één label (in het geval van PressAnalyzer het profiellabel). Op die manier kan namelijk ook kennis over andere labels gebruikt worden bij het labelen van een bepaald tekstsegment. Daardoor kan voorkomen worden dat een gedetecteerd segment overlapt met een ander segment;
- Het lijkt zinvol om bij het labelen niet alleen gebruik te maken van de vacaturetekst, maar ook van eigenschappen van de gescande afbeelding. Bijvoorbeeld de layout van de personeelsadvertentie zou kunnen bijdragen aan het bepalen van de labels. Een ander argument om bij het labelen niet alleen te kijken naar de vacaturetekst is het feit dat tijdens het OCR-proces herkenningsfouten gemaakt kunnen worden zodat de opgeleverde tekst fouten bevat. Het labelen zou daardoor beïnvloed kunnen worden.

Maryam's onderzoek heeft zich beperkt tot een voorstudie naar de bijdrage die text-mining zou kunnen leveren aan het fulltextproces. Het ontwikkelde prototype van PressAnalyzer is nooit in het fulltextproces opgenomen.

## 1.4 Probleemdefinitie

De situatie bij MatchCare na de afstudeeronderzoeken van Marco en Maryam kan als volgt geschetst worden:

- In zijn onderzoek heeft Marco heeft het bestaande fulltextproces gedeeltelijk geautomatiseerd en heeft zodoende een nieuw fulltextproces opgeleverd waarin de verwerkingstijd van een vacature aanzienlijk is afgenomen. Hij heeft software ontwikkeld waarmee een personeelsadvertentie handmatig in segmenten opgedeeld kan worden en waarmee de segmenten gelabeld kunnen worden;
- Het onderzoek van Maryam heeft aangetoond dat op basis van slechts de ge-OCR-de tekst van een gescande personeelsadvertentie al d.m.v. text-mining een uitspraak gedaan kan worden over de labeling van een bepaald tekstsegment.

Na het onderzoek van Maryam is de bedrijfsleiding erg geïnteresseerd geraakt in de mogelijkheden om het handmatig labelen van personeelsadvertenties te automatiseren. Het onderzoek van Maryam heeft dan wel niet in directe zin een positief resultaat gehad (slechts 30% van alle profielsegmenten werden correct gedetecteerd), wel heeft ze duidelijke oorzaken daarvoor kunnen aangeven evenals een aantal verbeterpunten waardoor het geautomatiseerd labelen meer succes zou kunnen hebben. De bedrijfsleiding heeft daarom besloten het onderzoek naar het geautomatiseerd labelen van personeelsadvertenties door te zetten.

Het geautomatiseerd labelen moet MatchCare de volgende verbeteringen opleveren:

- De verwerkingstijd van vacatures moet nog verder afnemen om het steigende aantal vacatures te kunnen blijven verwerken. Het is aangetoond dat door eentonige, zich herhalende werkzaamheden zoals het segmenteren van een personeelsadvertentie en het labelen van de segmenten de snelheid en nauwkeurigheid van de medewerkers daalt door concentratieverlies. Een bijkomend voordeel van automatiseren is dat de verwerkingssnelheid van computers steeds maar toeneemt terwijl de verwerkingssnelheid van mensen na een paar weken trainen de limiet heeft bereikt;
- De productiecapaciteit moet steigen zodat er minder medewerkers nodig zijn voor de verwerking van de vacatures. Hierdoor kan MatchCare personeelskosten besparen. MatchCare geeft hierdoor invulling aan de langetermijnvisie: de productiecapaciteit laten toenemen door te automatiseren in plaats van door meer medewerkers aan te nemen (wat door ruimtegebrek niet eens zou kunnen);
- Het verwerken van vacatures op onmenselijke werktijden mogelijk maken: het geautomatiseerd labelen zou, als de productieplanning dat toelaat, eventueel 's nachts kunnen draaien;
- Het voorkomen van RSI: door veel zich herhalende bewegingen bij het segmenteren en labelen is een verhoogde kans aanwezig dat de medewerkers last krijgen van RSI. Door het automatiseren daalt deze kans aanzienlijk;
- Het verhogen van de nauwkeurigheid en betrouwbaarheid van de labeling: een geautomatiseerd proces kan zich vaak herhalende bewerkingen nauwkeuriger en betrouwbaarder uitvoeren dan een mens dat kan.

Naast de eis dat de te ontwerpen geautomatiseerde labelmethodiek bovengenoemde verbeteringen moet gaan opleveren geldt nog een aantal eisen:

- Het moet de nauwkeurigheid van het handmatig labelen benaderen;
- Het moet de bestaande labelcomponent kunnen vervangen zonder dat daarvoor de rest van het bestaande fulltextproces gewijzigd moet worden;
- Het onderhoud moet zo eenvoudig mogelijk zijn;
- Verbeteringen moeten zo eenvoudig mogelijk door te voeren zijn. Bij voorkeur moet de geautomatiseerde labelmethodiek modulair opgebouwd worden zodat elk onderdeel vervangen of herschreven kan worden zonder dat de gehele methodiek aangepast moet worden;
- De kosten die de implementatie met zich meebrengt mogen de verwachte kostenreductie niet overtreffen.

## 1.5 Hypothese

Eén van de verbeterpunten uit het onderzoek van Maryam was dat er bij het geautomatiseerd labelen ook gebruikt gemaakt zou kunnen worden van eigenschappen van de gescande afbeelding van een personeelsadvertentie. In mijn eigen jarenlange ervaring bij MatchCare (eerst al invoerder, later als procescoördinator van het fulltextproces) heb ik gezien dat er een groot aantal wetmatigheden bestaat in het domein van de personeelsadvertenties. Deze wetmatigheden kunnen worden onderverdeeld in tekstuele, layoutgerichte en tekstopmaakgerichte wetmatigheden. Een mens ontdekt deze wetmatigheden geleidelijk

tijdens het handmatig labelen en zal daarom na een aantal weken ervaring opgedaan te hebben, gebruikmakend van de aangeleerde wetmatigheden, sneller kunnen labelen dan de eerste keer. De gevonden wetmatigheden geven echter geen totale zekerheden, want op elke wetmatigheid komen wel uitzonderingen voor. Het zijn echter wel zeer bruikbare richtlijnen. De oorsprong van dit onderzoek ligt in de gedachte dat het mogelijk is, uitgaande van de wetmatigheden, het labelingproces te automatiseren.

De tekstuele wetmatigheden betreffen eigenschappen van de ge-OCR-de tekst van de personeelsadvertentie. Er zijn drie tekstuele wetmatigheden die waarschijnlijk kunnen bijdragen aan het geautomatiseerd labelen:

1. In bijna alle personeelsadvertenties is de tekst optimaal geclusterd. Dat wil zeggen dat alle tekst behorend bij dezelfde informatiecategorie dicht bij elkaar staat en niet doorsneden wordt door de tekst die bij tot andere categorie behoort. Door deze eigenschap kan de tekst gesegmenteerd worden in sequentiële stukken tekst die elk één informatiecategorie bevatten. De enige uitzondering hierop vormen de informatiecategorieën Sollicitatie-informatie en Adverteerdersgegevens die in bijna alle personeelsadvertenties samen en vermengd in één alinea opgenomen zijn. Daarom worden ze in LeO samengenomen in één label en zullen ze in de rest van dit onderzoek ook samengenomen worden;
2. Voor elke informatiecategorie kan een lijst van keywords worden opgesteld die de betreffende informatiecategorie uniek bepalen. Dit is de wetmatigheid die Maryam heeft onderzocht;
3. De informatiecategorieën binnen een vacature staan bijna altijd in een vaste volgorde. Als de gebruikelijke leesvolgorde wordt gehanteerd (van links naar rechts en van boven naar beneden, maar wel met inachtname van een eventueel aanwezige kolomstructuur), dan staan de informatiecategorieën bijna altijd in dezelfde volgorde als de volgorde die MatchCare hanteert: Werkgeversomschrijving, Functie-omschrijving, Profiel, Arbeidsvoorwaarden, Sollicitatie-informatie/Adverteerdersgegevens.

De layoutgerichte wetmatigheden betreffen eigenschappen van de fysieke layout, dus van de geometrische indeling van de tekstsegmenten waaruit de personeelsadvertentie bestaat. Er zijn drie belangrijke layoutgerichte wetmatigheden waarvan onderzocht moet worden of ze bij kunnen dragen aan het geautomatiseerd labelen:

1. Twee gegeven informatiecategorieën hebben vaak binnen een personeelsadvertentie dezelfde relatieve positie ten opzichte van elkaar. Het profielsegment staat bijvoorbeeld vaak boven het arbeidsvoorwaardensegment;
2. Naast de relatieve positie kan ook de absolute positie van een tekstsegment een aanwijzing vormen over de informatiecategorie van het tekstsegment. Zo zal het tekstsegment met de functietitel nooit rechtsonder en het tekstsegment met de arbeidsvoorwaarden nooit linksboven in de advertentie staan;
3. De tekstsegmenten worden vaak gescheiden door wit-regels. In de gevallen dat ze niet worden gescheiden door wit-regels worden ze bijna altijd wel gescheiden door alinea-overgangen.

De tekstopmaakgerichte wetmatigheden concentreren zich op eigenschappen van de opmaak van de vacaturetekst, zoals lettergrootte en wel of niet vetgedrukt, cursief gedrukt of onderstreept.
Er zijn drie tekstopmaakgerichte wetmatigheden waarvan de bruikbaarheid onderzocht moet worden:

1. De functietitel wordt bijna altijd weergegeven in het grootste, en daardoor meest opvallende, lettertype;
2. Als in de tekst een kopje met een afwijkende opmaak is opgenomen (bijvoorbeeld "Je verantwoordelijkheden zijn:"), dan geeft dat in bijna alle gevallen het begin van een nieuwe informatiecategorie aan. Deze kopjes kunnen ook zeer goed tekstueel worden gebruikt, want ze bevatten woorden (zoals "verantwoordelijkheden" in het voorbeeld) die de informatiecategorie uniek bepalen;

3. Het komt regelmatig voor dat de informatiecategorie Werkgeversomschrijving in een afzonderlijke kolom, dus buiten het hoofdgedeelte van de personeelsadvertentie, wordt geplaatst. In bijna alle gevallen verschilt de tekstopmaak van die afzonderlijke kolom van de tekstopmaak dat wordt gebruikt in het hoofdgedeelte.

Om te kunnen onderzoeken in welke mate deze wetmatigheden gelden in het domein van de personeelsadvertenties en om hun bruikbaarheid bij het geautomatiseerd labelen te onderzoeken worden de bovengenoemde wetmatigheden geformuleerd als deelhypotheses.

Deelhypothese tekstueel1:

*De tekst van personeelsadvertenties is optimaal geclusterd.*

Deelhypothese tekstueel2:

*Voor elke informatiecategorie kan een lijst van keywords opgesteld worden die de betreffende informatiecategorie uniek bepalen.*

Deelhypothese tekstueel3:

*De informatiecategorieën staan binnen een vacature in de standaardvolgorde van Werkgeversomschrijving, Functie-omschrijving, Profiel, Arbeidsvoorwaarden en Sollicitatie-informatie/Adverteerdersgegevens.*

Deelhypothese layout1:

*Twee gegeven informatiecategorieën hebben een vaste relatieve positie ten opzichte van elkaar.*

Deelhypothese layout2:

*De absolute positie van een tekstsegment vormt een aanwijzing over het toe te kennen label.*

Deelhypothese layout3:

*De tekstsegmenten worden gescheiden door wit-regels en als dat niet het geval is door alinea-overgangen.*

Deelhypothese tekstopmaak1:

*De functietitel wordt weergegeven in het meest opvallende lettertype.*

Deelhypothese tekstopmaak2:

*Kopjes, vaak met een afwijkende opmaak, geven het begin van een nieuwe informatiecategorie aan.*

Deelhypothese tekstopmaak3:

*Als de Werkgeversomschrijving in een afzonderlijke kolom is geplaatst, dan heeft die kolom een afwijkende tekstopmaak.*

In hoofdstuk 2 worden deze deelhypotheses door middel van een statistische domeinverkenning besproken en gekwantificeerd.

Alle deelhypotheses samengenomen geven aanleiding tot de volgende onderzoekshypothese:

*Het is mogelijk een geautomatiseerde methodiek te ontwikkelen die,*
*gebruikmakend van tekstuele, layoutgerichte en tekstopmaakgerichte wetmatigheden,*
*in staat is de meerderheid van alle personeelsadvertenties uit de gedrukte media te labelen.*

De hypothese wordt onderzocht door middel van een uitgebreid literatuuronderzoek. Het literatuuronderzoek heeft als doel bestaande relevante methodieken en algoritmen in kaart te brengen en te beoordelen op hun toepasbaarheid voor het geautomatiseerd labelen. De verwachting bestaat dat de labelmethodiek kan worden samengesteld uit bestaande methodieken en algoritmen zodat er geen nieuwe methodieken en algoritmen ontwikkeld hoeven worden.

## 1.6 Beperkingen

Om het onderzoek niet onnodig gecompliceerd te maken wordt eerst alleen onderzoek gedaan naar het labelen van de 'basis' personeelsadvertentie. Dit houdt in dat de personeelsadvertenties die worden meegenomen in het onderzoek aan de volgende eisen moeten voldoen:

- de personeelsadvertentie bevat slechts één vacature;
- de tekst van de personeelsadvertentie moet op te delen zijn in de genoemde informatiecategorieën. Het komt namelijk voor dat personeelsadvertenties alleen maar een algemeen beeld van de werksfeer bij het bedrijf schetsen en niet ingaan op de details van de functie. Zo'n personeelsadvertentie is zelfs handmatig niet goed te labelen en zal daarom niet worden meegenomen in dit onderzoek.

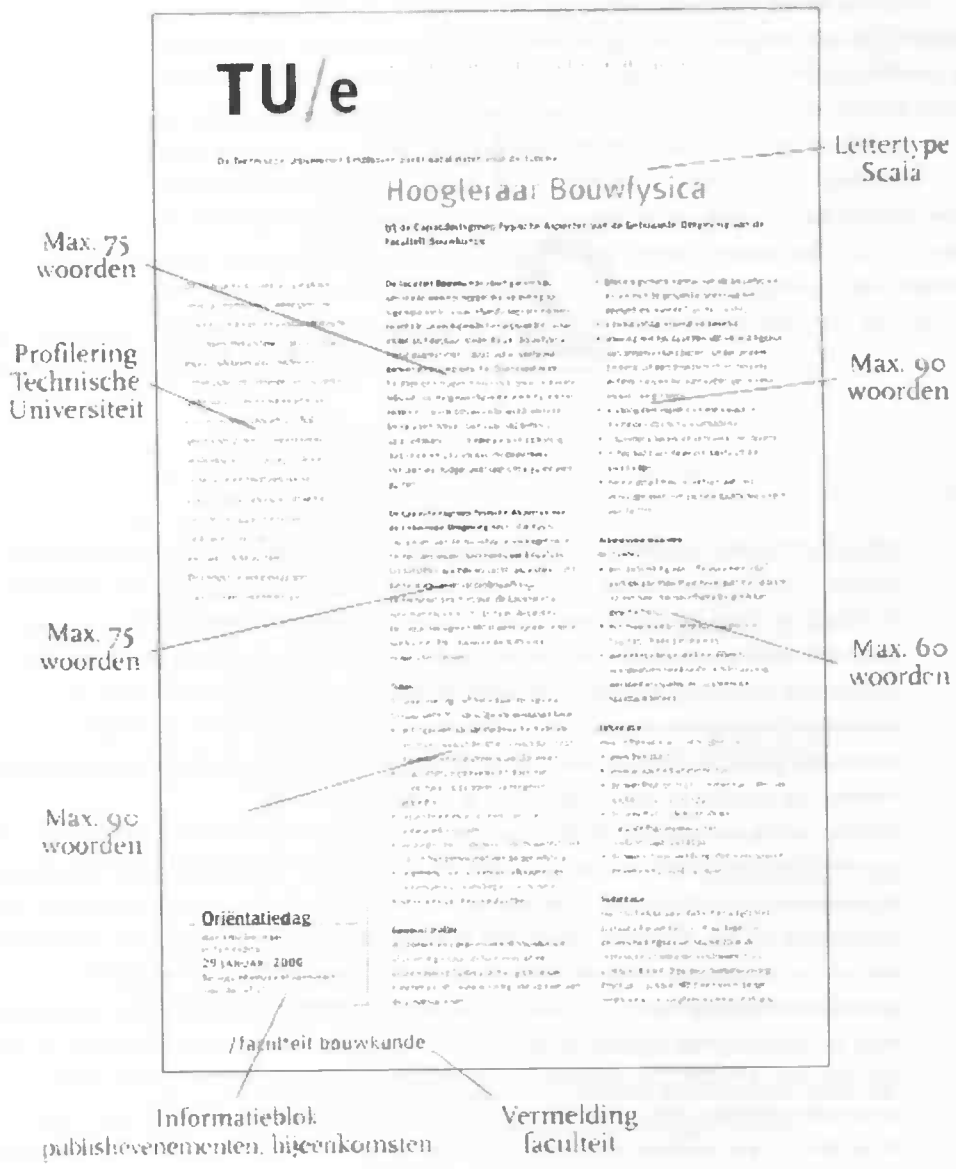
## 1.7 Indeling van de scriptie

In Hoofdstuk 2 wordt het domein van de personeelsadvertenties verkent door middel van een statistische analyse van de geformuleerde deelhypotheses. Hoofdstuk 3 geeft een overzicht van het onderzoeksterrein Document Image Understanding, een korte bespreking van image preprocessing technieken en een korte literatuurbespreking van segmentatietechnieken. Hoofdstuk 4 bevat een uitgebreid literatuuronderzoek naar de logische analyse van documenten. De gevonden literatuur wordt besproken en tevens wordt de toepasbaarheid bij het geautomatiseerd labelen van personeelsadvertenties besproken. In Hoofdstuk 5 wordt vervolgens een ontwerp van de geautomatiseerde labelmethodiek weergegeven. In Hoofdstuk 6 worden, ten slotte, de conclusies gegeven en worden de mogelijkheden voor toekomstig onderzoek besproken.

# 2 Domeinverkenning

Een gedrukte personeelsadvertentie is een communicatiemiddel waarmee een adverteerder aan zoveel mogelijk geschikte werkzoekenden zijn vacature bekend wil maken in de hoop de perfecte medewerker voor de vacature te vinden. De meeste adverteerders proberen hun personeelsadvertenties zo te ontwerpen dat de advertentie opvalt en daardoor de aandacht van de lezers getrokken wordt. Dat kan een adverteerder bereiken door bijvoorbeeld afbeeldingen of hele grote lettertypes te gebruiken. Ook proberen de adverteerders de lezers aan te spreken door een vaste layout te gebruiken waarin de huisstijl van de adverteerder is terug te vinden. Iedere adverteerder probeert op die manier in de vaste layout zijn eigen identiteit vast te leggen en daarmee herkenbaar te worden voor de lezers.
Ondanks de berg aan verschillen die dit oplevert zie je toch dat bijna alle personeelsadvertenties een aantal gezamelijke kenmerken vertonen. Die gezamelijke kenmerken, oftewel wetmatigheden, zijn bijna allemaal algemeen geaccepteerde middelen op het vlak van het overzichtelijk en gestructureerd presenteren van de vacaturetekst om deze zo duidelijk mogelijk over te brengen aan de lezers. In Figuur 2.1 is een voorbeeld te zien van hoe een bepaalde werkgever zich houdt aan bepaalde wetmatigheden. De afgebeelde personeelsadvertentie in Figuur 2.1 is afkomstig van de site van de Technische Universiteit Eindhoven [4] en bevat een sjabloonachtige instructie die aangeeft hoe een personeelsadvertentie van de TU/e er uit dient te zien. Dat zorgt niet alleen voor herkenbaarheid onder de lezers maar het zorgt er ook voor dat de personeelsadvertenties altijd de identiteit van de TU/e uitstralen. Deze quote, afkomstig van de site van de TU/e, dienst als een voorbeeld dat er binnen een bedrijf vaste afspraken bestaan over het opstelen van een personeelsadvertentie:

> *"De nieuwe advertentielijn is naast de profieltekst over de TUE en koptekst opgebouwd uit zeven tekstblokken. Het stukje met een algemeen verhaal over de TUE staat vast en hierin mogen geen wijzigingen worden aangebracht. De blokken zijn achtereenvolgens Beheerseenheid (max. 75 woorden), Werkonderdeel (max. 75 woorden), Taken (max. 90 woorden), Gewenst profiel (max. 90 woorden), Arbeidsvoorwaarden (max. 60 woorden), Informatie en Sollicitatie."* [4]




**Figuur 2.1 - voorbeeld van het gebruik van een huisstijl bij het opstellen van een personeelsadvertentie.**

Door mijn jarenlange werkervaring bij MatchCare heb ik een duidelijk beeld opgebouwd van de verschillende wetmatigheden die voorkomen in het domein van de personeelsadvertenties. Het formuleren van de wetmatigheden is een kwestie van nagaan welke aspecten van de personeelsadvertentie de labelaar gebruikt bij het labelen van een personeelsadvertentie en in welke mate de verschillende aspecten een rol spelen bij de bepaling van de labels. De aspecten die in de ruime meerderheid van de personeelsadvertenties een nuttige bijdrage leveren kunnen geformuleerd worden als wetmatigheden. Elke wetmatigheid kan een aanwijzing vormen voor de labeling van een personeelsadvertentie, maar ze vormen niet allemaal even sterke aanwijzingen. Dit hoofdstuk bespreekt en kwantificeert de geldigheid van de wetmatigheden aan de hand van de resultaten van een statistische domeinverkenning en maakt daarmee aannemelijk dat een aantal van de wetmatigheden zeer bruikbaar zijn bij het geautomatiseerd labelen. De statistische domeinverkenning is uitgevoerd over 723 personeelsadvertenties die verschenen in de Volkskrant van 18 mei, 25 mei en 1 juni 2002. Een beschrijving van de gebruikte methoden bij de statistische domeinverkenning is terug te vinden in sectie 2.4.

## 2.1 Tekstuele wetmatigheden

Het belangrijkste doel van een gedrukte advertentie is het overbrengen van de boodschap van de adverteerder aan de lezer. In het geval van personeelsadvertenties wil de adverteerder bekend maken dat er een vacature is en dat de adverteerder op zoek is naar geschikte kandidaten om de vacature te vervullen. Verreweg het belangrijkste onderdeel van de boodschap bestaat uit tekst. Met de tekst wordt immers aan de lezers duidelijk gemaakt wat de activiteiten van het bedrijf zijn, wat de inhoud van de functie is, welke eisen er gesteld worden aan sollicitanten, etc. De tekst moet ervoor zorgen dat geschikte kandidaten geïnteresseerd raken en de overige lezers doorlezen.
Bij de grotere personeelsadvertenties worden ook vaak grafische onderdelen gebruikt, zoals een bedrijfslogo of een afbeelding die de sfeer op de werkvloer weergeeft. De grafische onderdelen staan echter altijd in dienst van de tekst en worden vooral gebruikt om de aandacht van de lezer te trekken en een bepaald sfeerbeeld op te roepen bij de lezer. Deze grafische onderdelen dragen nooit bij aan de labeling van een personeelsadvertentie en zullen daarom geheel buiten beschouwing gelaten worden in dit onderzoek.

Om de boodschap zo duidelijk mogelijk over te brengen volgen bijna alle adverteerders een aantal wetmatigheden die, nader bekeken, heel vanzelfsprekend zijn. De drie tekstuele wetmatigheden komen duidelijk naar voren in de personeelsadvertentie in Figuur 2.2:

1. De vacaturetekst is optimaal geclusterd. Een cluster is in deze context een tekstsegment waarvoor geldt dat de woorden in het tekstsegment een sterke onderlinge samenhang vertonen (doordat ze tot dezelfde informatiecategorie behoren) en de clusters onderling weinig samenhang vertonen (doordat alle clusters tot een andere informatiecategorie behoren). Optimaal geclusterde vacaturetekst duidt op een stricte scheiding van alle informatiecategorieën, terwijl een niet optimaal geclusterde vacaturetekst tekstsegmenten bevat waarin de informatiecategorieën gemengd zijn. Een optimaal geclusterde vacaturetekst zorgt voor structuur en overzichtelijkheid.

De personeelsadvertentie in Figuur 2.2 laat tevens de reden zien waarom MatchCare de twee informatiecategorieën Sollicitatie-informatie en Adverteerdersgegevens samengenomen heeft: de twee categorieën zijn samengenomen in één tekstsegment en zijn daardoor moeilijker te scheiden;

2. De personeelsadvertentie in Figuur 2.2 bevat een groot aantal woorden die met zeer grote waarschijnlijkheid aangeven tot welke informatiecategorie het tekstsegment waar ze in voorkomen behoort. Dit zijn de zogenaamde keywords uit het afstudeeronderzoek van Maryam Wilhelm. Enkele voorbeelden uit de personeelsadvertentie in Figuur 2.2 zijn de woorden "Taken" (Functie-omschrijving), "eisen", "werkervaring" en "stressbestendigheid" (Profiel), "jaarcontract" en "schaal" (Arbeidsvoorwaarden) en "sollicitaties", "CV", "schriftelijk", "e-mail" en "t.a.v." (Sollicitatie-informatie/Adverteerdersgegevens). Het is ook mogelijk dat een groep woorden tezamen hiervoor gebruikt kunnen worden, zoals in dit voorbeeld het woordenpaar "Functie eisen". Dat het mogelijk is keywords te gebruiken bij het labelen komt doordat elke informatiecategorie een heel specifiek onderdeel van de vacature beschrijft en er woorden te vinden zijn die alleen maar gebruikt worden bij dat ene specifieke onderdeel;

Het Advies- en Meldpunt Kindermishandeling (AMK) Noord Holland is een laagdrempelige organisatie waarbij iedereen kan vragen om advies en consult of een melding kan doen van vermoedens van kindermishandeling. Het AMK ziet zich ten doel kindermishandeling zo vroeg mogelijk te stoppen en stimuleert preventie van kindermishandeling. Het AMK werkt met maximale opienheid

Het AMK Noord Holland zoekt voor haar vestiging in Haarlem op korte termijn een

Maatschappelijk Werkende m/v
voor 32-36 uur per week

Taken
- aannemen van meldingen m.b.t. vermoedens van kindermishandeling
- registratie van de meldingen
- voeren van gesprekken met ouders en kinderen
- het op gang brengen van hulpverlening
- geven van advies en consult
- geven van voorlichting
- deelname aan relevante overlegstructuren

Functie eisen
- opleiding HBO maatschappelijk werk bij voorkeur aangevuld met vervolgopleiding
- werkervaring relevant voor de functie
- goede communicatieve eigenschappen
- stressbestendigheid
- goede mondelinge en schriftelijke uitdrukkingsvaardigheden
- kunnen werken in teamverband
- bekend zijn met het werken in Word

Algemeen
Het dienstverband vindt voorlopig plaats op basis van een jaarcontract. Arbeidsvoorwaarden en salarisring zijn volgens de CAO-GGZ, schaal FWG 55

Sollicitaties
Sollicitaties met een uitgebreid CV kunt u zo spoedig mogelijk schriftelijk of per e-mail sturen aan
Advies en Meldpunt Kindermishandeling
Rubenslaan 2, 1816 MB Alkmaar
t.a.v. mevr. E.J.M. Kipbeek
Email: amkalkmaar@wxs.nl

**Figuur 2.2 - een personeelsadvertentie waarin de drie tekstuele wetmatigheden duidelijk naar voren komen.**

3. Bij de gebruikelijke leesvolgorde (van links naar rechts en van boven naar beneden, met inachtname van de eventueel aanwezige kolomstructuur) staan de informatiecategorieën in de personeelsadvertentie in Figuur 7 in de volgorde die MatchCare hanteert: Werkgeversomschrijving, Functie-omschrijving, Profiel, Arbeidsvoorwaarden en Sollicitatie-informatie/Adverteerdersgegevens. De praktijk wijst uit dat dat de meest logische volgorde is om de verschillende categorieën waaruit de vacaturetekst bestaat in te plaatsen. Deze wetmatigheid is binnen het domein van de personeelsadvertenties een soort conventie geworden waar maar weinig adverteerders van afwijken.

Bij het kwantificeren van deze tekstuele wetmatigheden werden de volgende resultaten verkregen:

1. Een optimaal geclusterde vacaturetekst: ruim 93% van de personeelsadvertenties vertoont een optimale clustering. Het merendeel van de overige personeelsadvertenties bevat één informatiecategorie die in twee tekstsegmenten verdeeld is waartussen een andere informatiecategorie geplaatst is (bijvoorbeeld twee profielsegmenten met daartussen het arbeidsvoorwaardensegment).
   Andere statistische bevindingen die met deze wetmatigheid samenhangen zijn:
   - In bijna 95% van de personeelsadvertenties staat de tekst van de informatiecategorieën SI en AG door elkaar. Daardoor is het in LeO onbegonnen werk tekstsegmenten te definiëren zodanig dat de twee informatiecategorieën gescheiden worden en is besloten de twee categorieën samen te nemen bij het labelen;
   - In 10% van de personeelsadvertenties zijn niet alle informatiecategorieën vertegenwoordigd. Het is dan meestal de categorie Arbeidsvoorwaarden die ontbreekt;
   - Het is mogelijk dat er tekst in de personeelsadvertentie voorkomt die tot geen van de gebruikte informatiecategorieën behoort. Zie bijvoorbeeld het tekstsegment over de oriëntatiedag in Figuur 2.1. Dit komt echter zo zelden voor dat deze mogelijkheid in het onderzoek genegeerd zal worden.
2. Aanwezigheid van keywords: om de effectiviteit van keywords te berekenen is voor elke informatiecategorie handmatig een lijst van keywords samengesteld. Deze lijsten van keywords zijn te vinden in Appendix B. Vervolgens is tijdens het labelproces in LeO steekproefsgewijs gekeken in welke mate de keywords uit de samengestelde lijsten voorkwamen in de handmatig gedefinieerde tekstblokken. Een maat voor de effictiviteit van de lijsten van keywords is de *precision*. Als een keyword werd gevonden in de ge-OCR-de tekst van een tekstblok, dan werd gecontroleerd of het door de labelaar handmatig toegekende label overeenkwam met het label dat bij het gevonden keyword hoorde. De precision is het percentage van door de gevonden keywords bepaalde labels die overeenkwamen met de handmatig toegekende labels. Een andere maat voor de effectiviteit is de *recall*. De recall is het percentage tekstblokken dat door het vinden van één of meerdere keywords hetzelfde label toegewezen kreeg als het handmatig toegekende label. Het onderzoek wordt kort beschreven in sectie 2.4.
   De resultaten van het onderzoek zijn te vinden in Tabel 1.

| | precision (%) | recall (%) |
|---|---|---|
| Werkgeversomschrijving | 85 | 57 |
| Functie-omschrijving | 82 | 73 |
| Profiel | 97 | 88 |
| Arbeidsvoorwaarden | 98 | 86 |
| Sollicitatie-inf./Adv.gegevens | 97 | 69 |

**Tabel 1 – de precision en recall van het labelen door middel van keywords.**

De hoge precisionpercentages tonen aan dat de handmatig samengestelde lijsten van keywords bijna alleen maar keywords bevatten die slechts worden gebruikt bij één informatiecategorie. Dat komt doordat de lijsten ook met dat doel voor ogen zijn samengesteld. De precision wordt verlaagd door personeelsadvertenties waarin alinea's tekst voorkomen die meer dan één informatiecategorie bevatten.

In dat geval selecteert de labelaar in LeO de gehele alinea als één tekstsegment en kiest het label dat het grootste gedeelte van de lading dekt.
De recallpercentages laten zien dat er tekstsegmenten zijn waarin geen keyword aangetroffen is. Het lage percentage bij de categorie Werkgeversomschrijving kan verklaard worden door het feit dat er voor die categorie moeilijk keywords te vinden zijn die niet ook regelmatig gebruikt worden voor de categorie Functie-omschrijving. Het recallpercentage wordt in het algemeen verlaagd door OCR-fouten waardoor het voorkomt dat keywords in de vacaturetekst niet herkend worden;

3. Aanwezigheid van een standaardvolgorde van de informatiecategorieën: in 71% van alle personeelsadvertenties staan de informatiecategorieën in de standaardvolgorde Werkgeversomschrijving, Functie-omschrijving, Profiel, Arbeidsvoorwaarden en Sollicitatie-informatie/Adverteerdersgegevens als de gebruikelijke leesvolgorde gevolgd wordt. Als alleen de personeelsadvertenties met slechts één tekstkolom bekeken worden, dan komt het percentage zelfs op 89%.
Er zijn drie varianten van personeelsadvertenties die niet de standaardvolgorde gebruiken:
   a. Personeelsadvertenties waarbij de categorie Werkgeversomschrijving niet de eerste maar juist de laatste categorie is. In die personeelsadvertenties is de categorie Werkgeversomschrijving als een afzonderlijke kolom aan de linkerkant geplaatst. Het komt veel vaker voor dat de Werkgevers-omschrijving als een afzonderlijke kolom aan de rechterkant van de personeelsadvertentie is geplaatst, maar dan verstoort het de standaardvolgorde niet. Zie Sectie 2.3 voor een kwantificatie van dit fenomeen;
   b. Personeelsadvertenties waarbij de clustering niet optimaal is doordat de tekst die tot een bepaalde informatiecategorie behoort doorsneden wordt door tekst die tot een andere informatiecategorie behoort;
   c. Personeelsadvertenties waarin niet alle categorieën vertegenwoordigd zijn. Dit is het geval in ongeveer 13% van alle personeelsadvertenties. Als deze wetmatigheid ook personeelsadvertenties accepteert waarbij alle *aanwezige* informatiecategorieën in de standaardvolgorde staan, dan geldt de wetmatigheid voor 88% van alle personeelsadvertenties.

Er moet wel worden opgemerkt dat deze wetmatigheid alleen maar een uitspraak doet over de volgorde van de informatiecategorieën en dus geen uitspraak doet over de plaats die de functietitel in de volgorde inneemt. De volgorde die de functietitel inneemt is namelijk veel onvoorspelbaarder en zou er dus voor zorgen dat deze wetmatigheid een deel van zijn kracht inlevert. Deze wetmatigheid kan dus alleen worden toegepast als er al een tekstsegment als functietitel is gelabeld en dat segment buiten beschouwing wordt gelaten.
Er kan ook gekeken worden naar het percentage personeelsadvertenties waarbij categorie $x$ voor categorie $y$ staat. Deze onderlinge volgorde tussen twee informatiecategorieën geeft een afgeleide wetmatigheid die stelt dat elk paar categorieën bijna altijd een vaste interne volgorde heeft. Ook deze afgeleide wetmatigheid laat zien hoe sterk er sprake is van een volgorde van de informatiecategorieën. Tabel 2 laat de kwantificatie van die afgeleide wetmatigheid zien. De getallen in Tabel 2 laten zien dat er alleen al op basis van de vacaturetekst met een behoorlijke zekerheid uitspraken gedaan kunnen worden over de labeling van een personeelsadvertentie.

| *Staat voor (%)* | Werkgever | Functie | Profiel | Arbeidsvoorwaarden | Soll./Adv. |
|---|---|---|---|---|---|
| Werkgever | - | 90 | 92 | 94 | 94 |
| Functie | 10 | - | 96 | 100 | 100 |
| Profiel | 8 | 4 | - | 100 | 100 |
| Arbeidsvoorwaarden | 6 | 0 | 0 | - | 95 |
| Soll./Adv. | 6 | 0 | 0 | 5 | - |

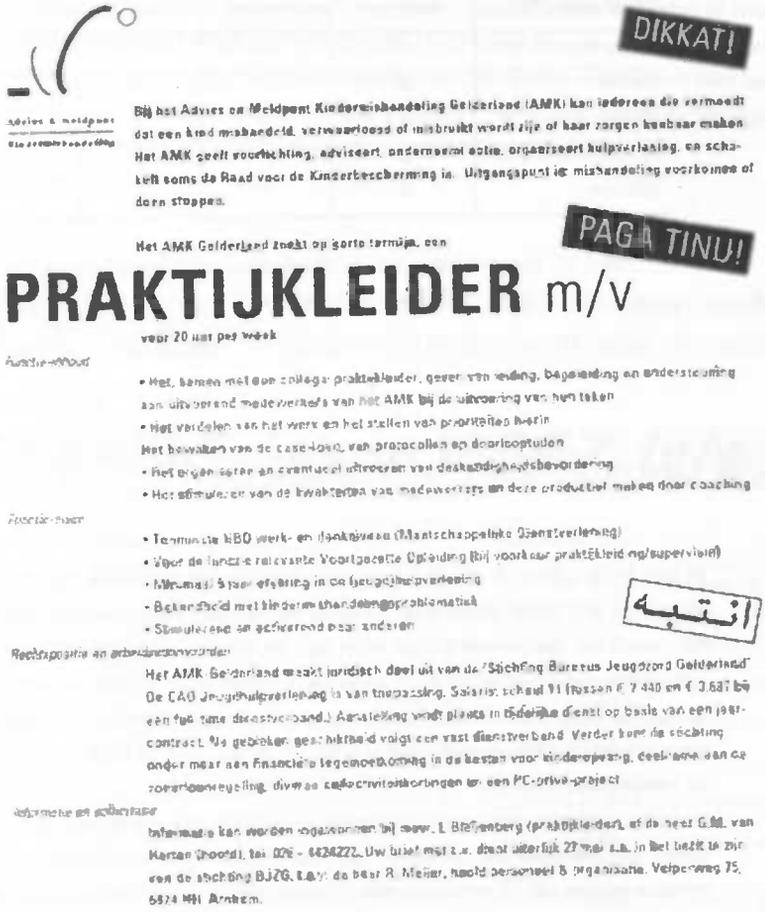
**Tabel 2 – Kwantificatie van de afgeleide wetmatigheid die stelt dat elk paar categorieën bijna altijd een vaste interne volgorde heeft.**

## 2.2 Layoutgerichte wetmatigheden

In deze sectie worden de drie layoutgerichte wetmatigheden besproken. Door het afstudeerproject van Marco Nijmeijer is er al bewijs geleverd dat de layout van een personeelsadvertentie een zeer bruikbare rol speelt bij het labelen van een personeelsadvertentie: toen het labelen werd verplaatst van YAN naar LeO en de labelaar gebruik kon maken van de gescande personeelsadvertentie werd er een duidelijke versnelling van het labelproces gemeten. In YAN had de labelaar slechts de ge-OCR-de tekst ter beschikking, terwijl in een gescande personeeladvertentie naast de tekst óók de layout en de tekstopmaak gebruikt kunnen worden bij de bepaling van de labels.
De layout zorgt met name voor een versnelling van het labelproces doordat de layout de structuur van de personeelsadvertentie veel duidelijker maakt dan de ge-OCR-de tekst kan doen doordat de absolute en relatieve posities van de tekstsegmenten zichtbaar zijn. In ge-OCR-de tekst is er geen zicht op de absolute en relatieve posities omdat alle tekst in een sequentiële volgorde staat.

De personeelsadvertenties in Figuur 2.3 en in Figuur 2.4 kunnen goed gebruikt worden om de drie layoutgerichte wetmatigheden te presenteren:

1. Twee gegeven informatiecategorieën hebben een vaste relatieve positie ten opzichte van elkaar: deze wetmatigheid komt vooral sterk naar voren bij personeelsadvertenties waarbij alle tekst in één kolom staat, zoals bij de personeelsadvertentie in Figuur 8. In Figuur 8 is bijvoorbeeld te zien dat de categorie Functie-omschrijving boven de categorie Profiel staat en dat de categorie Profiel boven de categorie Arbeidsvoorwaarden staat. Dat deze wetmatigheid vooral geldt voor personeelsadvertenties met maar één kolom tekst komt door de wetmatigheid dat er een standaardvolgorde van informatiecategorieën aangehouden wordt (de derde tekstuele wetmatigheid). De relatie `categorie x staat voor categorie y` is dan namelijk identiek aan de relatie `categorie x staat boven categorie y`. De relatie `Profiel staat voor Arbeidsvoorwaarden` is dus equivalent met de relatie `Profiel staat boven Arbeidsvoorwaarden`.



DIKKAT!

Bij het Advies en Meldpunt Kindermishandeling Gelderland (AMK) kan iedereen die vermoedt dat een kind mishandeld, verwaarloosd of misbruikt wordt zijn of haar zorgen kenbaar maken. Het AMK geeft voorlichting, adviseert, ondernemen actie, organiseert hulpverlening, en schakelt soms de Raad voor de Kinderbescherming in. Uitgangspunt is mishandeling voorkomen of daar stoppen.

PAGINA TINIJ!

Het AMK Gelderland zoekt op korte termijn, een

# PRAKTIJKLEIDER m/v

voor 20 uur per week

Functie-inhoud

- Het, samen met een collega-praktijkleider, geven van leiding, begeleiding en ondersteuning aan uitvoerend medewerkers van het AMK bij de uitvoering van hun taken
- Het verdelen van het werk en het stellen van prioriteiten hierin

Het bewaken van de casuïstiek, van protocollen en doorlooptijden

- Het mede sturen en eventueel uitvoeren van deskundigheidsbevordering
- Het stimuleren van de kwaliteiten van medewerkers en deze productief maken door coaching

Functie-eisen

- Tenminste HBO werk- en denkniveau (Maatschappelijke Dienstverlening)
- Voor de functie relevante Voortgezette Opleiding (bij voorkeur praktijkleiding/supervisie)
- Minimaal 5 jaar ervaring in de jeugdhulpverlening
- Bekendheid met kindermishandelingsproblematiek
- Stimulerend en activerend naar anderen

انتبه

Rechtspositie en arbeidsvoorwaarden

Het AMK Gelderland maakt juridisch deel uit van de 'Stichting Bureaus Jeugdzorg Gelderland'. De CAO Jeugdhulpverlening is van toepassing. Salaris: schaal 11 (tussen € 2.440 en € 3.637 bij een fte tons dienstverband.) Aanstelling vindt plaats in tijdelijke dienst op basis van een jaarcontract. Na gebleken geschiktheid volgt een vast dienstverband. Verder biedt de stichting onder meer aan financiële tegemoetkoming in de kosten voor kinderopvang, deelname aan de zorgverzekering, diverse cursuscontributieborgingen en een PC-privé-project.

Informatie en solliciteren

Informatie kan worden ingewonnen bij mevr. L. Bleijenberg (praktijkleiden), of de heer G.M. van Kersen (hoofd), tel. 026 - 4424222. Uw brief met c.v. dient uiterlijk 27 mei e.a. in het bezit te zijn van de stichting BJZG, t.a.v. de heer R. Meijer, hoofd personeel & organisatie, Velperweg 75, 6824 MH Arnhem.

**Figuur 2.3 - Een personeelsadvertentie aan de hand waarvan de eerste en tweede layoutgerichte wetmatigheden duidelijk gemaakt kunnen worden.**

Voor personeelsadvertenties met meer dan één kolom tekst geldt deze wetmatigheid in mindere mate. Bij die personeelsadvertenties zijn de relaties `staat voor` en `staat boven` niet meer identiek en is er op basis van de derde tekstuele wetmatigheid niet met zekerheid af te leiden is hoe de informatiecategorieën over de kolommen verdeeld zijn. Hierdoor zijn de absolute posities van de categorieën minder goed te voorspellen en daarmee ook de relatieve posities. Zie voor een verdere uitleg ook de tweede wetmatigheid;

2. De absolute positie van een tekstsegment vormt een aanwijzing over het toe te kennen label. Ook deze wetigmatigheid is een afgeleide van de wetmatigheid dat er een standaardvolgorde van informatiecategorieën aangehouden wordt. Uit die wetmatigheid zijn namelijk de kansen af te leiden dat een bepaalde informatiecategorie op een bepaalde verticale positie staat (bij personeelsadvertenties met één kolom tekst) en de kansen dat een bepaalde informatiecategorie *in een bepaalde kolom* op een bepaalde verticale positie staat (bij personeelsadvertenties met meer dan één kolom tekst).

Voor personeelsadvertenties met maar één kolom tekst is dit eenvoudig in te zien. De informatiecategorieën zijn in de standaardvolgorde in de kolom terug te vinden van boven naar beneden. Dit is een gevolg van het feit dat een kolom standaard van boven naar beneden gelezen wordt.
Voor personeelsadvertenties met meer dan één kolom tekst moet een andere benadering gebruikt worden. Als de kolommen tekst omgezet zouden worden naar één kolom door de kolommen onder elkaar te plaatsen, dan hebben we dezelfde situatie als bij een personeelsadvertentie met maar één kolom tekst. Er geldt dan wederom dat de informatiecategorieën in de standaardvolgorde in de kolom van boven naar beneden terug te vinden zijn. Als de kolommen weer teruggebracht worden in hun oorspronkelijke staat, dan wordt de kans dat een bepaalde informatiecategorie op een bepaalde verticale positie staat omgezet naar de kans dat een bepaalde informatiecategorie *in een bepaalde kolom* op een bepaalde verticale positie staat. Het omzetten zorgt ervoor dat de absolute posities van de informatiecategorieën en de relatieve posities tussen twee informatiecategorieën veel meer variëren. Zie bijvoorbeeld de informatiecategorie Profiel in de personeelsadvertentie in Figuur 9. Het Profiel staat bovenaan in de derde kolom maar had bij een kortere Werkgeversomschrijving onderaan in de tweede kolom gestaan. Dat maakt voor de absolute positie én de relatieve positie ten opzichte van bijvoorbeeld de Functie-omschrijving een groot verschil. Door dit fenomeen kunnen de absolute en relatieve posities minder sterk als wetmatigheid gebruikt worden bij een personeelsadvertentie met meerdere tekstkolommen;

3. De tekstsegmenten worden gescheiden door wit-regels en als dat niet het geval is door alinea-overgangen. Wit-regels worden in personeelsadvertenties gebruikt om binnen een tekstkolom een duidelijk onderscheid aan te brengen tussen verschillende soorten informatie en zorgen zo voor duidelijk gestructureerde vacaturetekst. De personeelsadvertentie in Figuur 9 laat dit duidelijk zien.
   Er geldt dat elke overgang tussen twee informatiecategorieën aangegeven wordt door een wit-regel (of eventueel door een alinea-einde), maar niet dat elke wit-regel een overgang representeert. In dit voorbeeld is dat te zien bij de informatiecategorieën Werkgeversomschrijving en Arbeidsvoorwaarden. In de categorie Werkgeversomschrijving is een 'extra' wit-regel aangebracht tussen de alinea over het cluster MGGZ en de alinea over de poliklinieken MGGZ. In de categorie Arbeidsvoorwaarden is een 'extra' wit-regel aangebracht tussen de alinea over de geboden aanstelling en de alinea waar het salaris in vermeld wordt. Het gebruik van een 'extra' wit-regel geeft dus aan dat de werkgever één informatiecategorie zoals MatchCare die definieert opsplitst in meerdere onderwerpen.
   Het komt ook voor dat een personeelsadvertentie de overgangen tussen de informatiecategorieën niet aangeeft door wit-regels maar alleen door bij een nieuwe informatiecategorie een nieuwe alinea te beginnen. Het spreekt voor zich dat een overgang door een wit-regel een duidelijkere scheiding is dan een overgang die slechts door een alinea-einde wordt aangegeven.

Bij het kwantificeren van deze layoutgerichte wetmatigheden werden de volgende resultaten verkregen:

1. Twee gegeven informatiecategorieën hebben een vaste relatieve positie ten opzichte van elkaar: voor het representeren van relatieve posities kan gebruik worden gemaakt van *Qualitative Spatial Relations*. De qualitative spatial relations zijn een uitbreiding op Allen's interval relations [5] die hij ontwikkelde voor 1-dimensionale tijdsintervallen. Qualitative Spatial Relations [6] breiden Allen's interval relations uit naar twee dimensies zodat ze te gebruiken zijn voor het representeren van ruimtelijke relaties tussen twee objecten die gedefinieerd zijn door hun *bounding boxes*. Op zowel de X- als de Y-as zijn 13 relaties gedefinieerd: `precedes` (ligt voor), `meets` (sluit aan bij), `overlaps` (overlapt), `starts` (start bij), `during` (binnen), `finishes` (eindigt samen met), `equals` (is gelijk aan) en de inversen van de eerste zes relaties. De relaties geven aan hoe de bounding box-randen, dus de intervallen op de x-

en y-as, van de twee categorieën geplaatst zijn ten opzichte van elkaar. Door de onnauwkeurigheid bij het handmatig definiëren van tekstblokken is het niet zinvol om de relaties te bepalen op basis van de exacte begin- en eindpunten van de twee intervallen. In [7] wordt een aangepaste set relaties besproken waarin niet de exacte begin- en eindpunten gebruikt worden maar waarin twee punten gelijk genoemd worden als de punten binnen een gegeven afstand (een marge die kan worden ingesteld door een parameter) van elkaar liggen: *Thick Boundary Rectangle Relations* (TBRR).

GG&GD De Gemeentelijke Geneeskundige en Gezondheidsdienst van Amsterdam

zoekt zo spoedig mogelijk

**arts poliklinieken MGGZ *(v/m)***

met ervaring in de huisartsgeneeskunde en in de psychiatrie, voor 32 uur per week

**Algemene informatie**
Het cluster MGGZ is een dienstverlenend onderdeel van de GG&GD op het terrein van maatschappelijke en geestelijke gezondheidszorg. Vanuit MGGZ wordt ondersteuning geboden aan individuele burgers en gezinnen bij het vinden van noodzakelijke en passende hulp en "vangnet"-mogelijkheden voor hen die in de reguliere zorg (nog) niet geholpen kunnen worden. Vanuit MGGZ worden hulpverleners en maatschappelijke instellingen ondersteund bij hun dienstverlening aan individuele burgers en gezinnen, vooral bij lastige problemen en bij onvoldoende motivatie voor hulp bij de betreffende cliënt. Daarnaast biedt MGGZ de gemeente Amsterdam ondersteuning bij het ontwikkelen van nieuw beleid op het terrein van de maatschappelijke gezondheidszorg.

*Het gemeentelijk beleid is erop gericht dat de samenstelling van het ambtenarenapparaat een afspiegeling wordt van de Amsterdamse bevolking. Daarom worden vrouwen, geïnteresseerden en leden van etnische minderheidsgroeperingen uitdrukkelijk verzocht te solliciteren. Bij het voldoen aan de gestelde eisen genieten zij de voorkeur.*

*Aequitas non aanleiding van deze advertentie wordt niet op prijs gesteld*

De hulpverlening van de poliklinieken MGGZ vindt plaats vanuit 5 poliklinieken. Op iedere poliklinieken werken artsen, verpleegkundigen en maatschappelijk werkers. Er wordt multidisciplinair gewerkt. De poliklinieken MGGZ bieden medische basiszorg, verslavingszorg, psychiatrische zorg en case management. Dit aanbod wordt ingezet als aanvulling op de reguliere zorg. De teams worden ondersteund door een eerstverantwoordelijk arts en administratieve medewerkers. Er is aandacht voor coaching en scholing van medewerkers. De dagelijkse leiding ligt in handen van de teamleider. Totaal bestaat een poliklinieken uit ongeveer 9 tot 16 fte's.

**Werkzaamheden.**
- medische verantwoordelijkheid voor de patiëntenzorg,
- integrale medische diagnostiek bij intake, monitoring en screening,
- indicering en uitvoering medische behandeling,
- registratie in het patiëntendossier,
- communicatie en consultatie binnen het zorgnetwerk.

**Functievereisten**
- ten minste opleiding tot basisarts,
- is besluitvaardig, neemt initiatief en is stressbestendig,
- heeft overtuigingskracht,
- kan samenwerken, is flexibel en werkt patiëntgericht,
- analyseert complexe situaties en komt tot realistische oplossingen,
- registratie als huisarts of als sociaal geneeskundige is een pre.

**Arbeidsvoorwaarden**
Aanstelling geschiedt in eerste instantie op basis van een aanstelling voor bepaalde tijd voor de duur van een jaar, inclusief proeftijd van 7 maanden.

**Salaris**
Het salaris bedraagt - afhankelijk van opleiding en ervaring - minimaal € 3.170,- en maximaal € 4.366,- (salarisgroep 12) bruto per maand bij een volledige werktijd.

**Informatie**
Nadere informatie over deze functie kunt u krijgen bij de heer G. van Santen (eerstverantwoordelijk arts MGGZ), telefoon 020-555 5473 en/of bij mevrouw H.C. Plas (hoofd poliklinieken MGGZ en methadonbusk), telefoon 020-555 5391.

**Sollicitatieadres**
Schriftelijke reacties kunnen - binnen veertien dagen - worden gericht aan de afdeling Personeel, Organisatie en Opleiding van de GG&GD, ter attentie van de heer W.J.Th. van Rooijmulen, P&O-adviseur, Nieuwe Achtergracht 100, 1018 WT te Amsterdam. Ook kunt u uw brief met c.v. als bijlage (in MS-Word) bij een e-mailbericht voegen: c.v.@ggd.amsterdam.nl

**Figuur 2.4 - Een personeelsadvertentie met meerdere tekstkolommen.**

Als voorbeeld voor de sterkte van deze wetmatigheid wordt de kans op een bepaalde relatieve positie van de informatiecategorie Functie-omschrijving ten opzichte van de informatiecategorie Werkgeversomschrijving weergegeven. Dat gebeurt in Tabel 3 voor personeelsadvertenties met één tekstkolom en in Tabel 4 voor personeelsadvertenties met meerdere tekstkolommen. De relaties worden in de tabellen afgekort met de beginletter van de relatie en krijgen een "`i`" toegevoegd als het gaat om de inverse relatie.

| *Soort relatie (%)* | P | M | O | Fi | Di | S | E | Si | D | F | Oi | Mi | Pi |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| x-intervallen | 0 | 1 | 7 | 7 | 1 | 3 | 55 | 5 | 9 | 4 | 6 | 2 | 0 |
| y-intervallen | 7 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 11 | 77 |

**Tabel 3 – de kansverdeling van mogelijke spatial relations voor de informatiecategorie Functie-omschrijving ten opzichte van de categorie Werkgeversomschrijving bij personeelsadvertenties met één tekstkolom.**

De hoge waarde bij de relatie `equals` tussen de twee x-intervallen geeft aan dat de informatie-categorieën bijna altijd een identiek x-interval hebben en dus zo goed als exact onder elkaar geplaatst zijn. De hoge waarden bij de relaties `meets (inverted)` en `precedes (inverted)` laten zien dat de categorie Functie-omschrijving bijna altijd onder de categorie Werkgeversomschrijving geplaatst is (NB: de oorsprong wordt gevormd door de linkerbovenhoek van de personeelsadvertentie). De relaties die niet voorkomen geven aan dat de twee y-intervallen elkaar niet kunnen overlappen, hetgeen natuurlijk voor zich spreekt bij een personeelsadvertentie met slechts één tekstkolom.

| *Soort relatie (%)* | P | M | O | Fi | Di | S | E | Si | D | F | Oi | Mi | Pi |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| x-intervallen | 9 | 2 | 14 | 1 | 4 | 1 | 29 | 2 | 4 | 2 | 12 | 1 | 19 |
| y-intervallen | 1 | 3 | 17 | 1 | 0 | 1 | 0 | 3 | 18 | 0 | 12 | 12 | 32 |

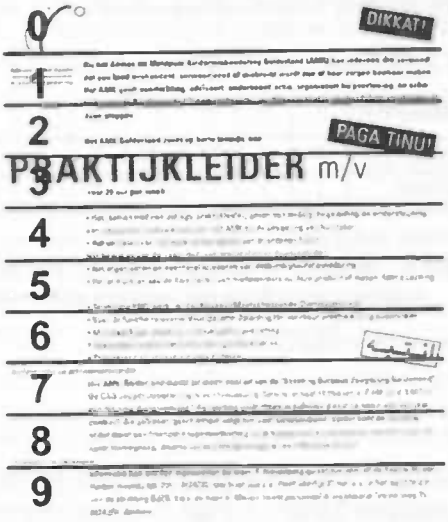
**Tabel 4 – de kansverdeling van mogelijke spatial relations voor de informatiecategorie Functie-omschrijving ten opzichte van de categorie Werkgeversomschrijving bij personeelsadvertenties met meerdere tekstkolommen.**

Tabel 4 laat zien dat er zowel voor de x- als voor de y-intervallen veel minder sterke relaties gelden en de wetmatigheid voor personeelsadvertenties met meerdere tekstkolommen dus minder krachtig is. De meest voorkomende relatie, de relatie `Pi` voor de y-intervallen, zegt dat de categorie Functie-omschrijving onder de categorie Werkgeversomschrijving geplaatst is. Helaas bevat maar 36% van alle personeelsadvertenties slechts één tekstkolom.
De onderzoekmethode wordt beschreven in Sectie 2.4. De kansverdelingen voor enkele andere paren van informatiecategorieën zijn te vinden in Appendix C.

2. De absolute positie van een tekstsegment vormt een aanwijzing over het toe te kennen label: ook bij deze wetmatigheid is het nuttig onderscheid te maken tussen personeelsadvertenties die slechts één tekstkolom hebben en personeelsadvertenties die meerdere tekstkolommen hebben. Bij de laatste soort personeelsadvertenties is namelijk, zoals eerder uitgelegd, veel meer variatie aanwezig in de absolute posities van de informatiecategorieën.
De sterkte van deze wetmatigheid wordt voor personeelsadvertenties met één tekstkolom berekend door de personeelsadvertenties te segmenteren in 10 horizontale zones met identieke hoogte, zoals aangegeven in Figuur 2.5, en vervolgens te berekenen in hoeveel procent van de personeelsadvertenties een bepaalde informatiecategorie in een bepaalde zone aangetroffen wordt. De resultaten zijn te zien in Tabel 5.

| ***Categorie in zone (%)*** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Werkgeversomschr. | 67 | 77 | 83 | 80 | 36 | 5 | 1 | 0 | 0 | 0 |
| Functie-omschr. | 0 | 6 | 15 | 43 | 67 | 60 | 58 | 12 | 0 | 0 |
| Profiel | 0 | 0 | 0 | 17 | 54 | 40 | 57 | 16 | 0 | 0 |
| Arbeidsvoorwaarden | 0 | 0 | 0 | 0 | 1 | 6 | 67 | 63 | 12 | 0 |
| Soll./Adv. | 0 | 0 | 0 | 0 | 1 | 0 | 15 | 57 | 68 | 73 |

**Tabel 5 – De kans dat in een personeelsadvertentie informatiecategorie *x* aangetroffen wordt in zone *y*.**




**Figuur 2.5 - Een personeelsadvertentie opgedeeld in 10 horizontale zones. Aan de linkerkant staan de zonenummers.**

De som van de percentages voor een bepaalde zone kan de 100% overschrijden. Dat komt doordat er in een zone meerdere informatiecategorieën tegelijk kunnen voorkomen. Als de som van de percentages voor een bepaalde zone kleiner is dan 100%, dan geeft dat aan dat het ook voorkomt dat de betreffende zone geen tekst bevat. De som van de percentages voor iedere informatiecategorie is groter dan 100%. Dat betekent dat iedere informatiecategorie gemiddeld in meer dan één zone staat.
Uit Tabel 5 valt af te leiden dat er een zeer sterke onderlinge volgorde tussen de informatiecategorieën aanwezig is.
Voor de personeelsadvertenties met meer dan één tekstkolom kan een zelfde soort benadering gebruikt worden, maar dan uitgebreid naar twee dimensies. De verkregen resultaten zijn echter zo onbruikbaar als aanwijzing voor het labelen dat ze niet verder besproken zullen worden.

3. De tekstsegmenten worden gescheiden door wit-regels en als dat niet het geval is door alinea-overgangen: in maarliefst 85% van alle personeelsadvertenties worden wit-regels gebruikt om de grens tussen twee informatiecategorieën aan te geven. In de overige 15% wordt de grens aangegeven doordat er een nieuwe alinea begonnen wordt. Slechts zeer zelden komt er een personeelsadvertentie voor waarin alle tekst samengenomen is in één alinea. Dat komt waarschijnlijk zo zelden voor omdat zo'n personeelsadvertentie veel moeilijker te lezen is en de lezer dat als minpunt van de adverteerder kan ervaren. Het kost de lezer namelijk veel meer moeite om de informatie op een rijtje te krijgen omdat er geen structuur aanwezig is.
De term "wit-regel" wordt in deze context iets ruimer opgevat dan normaal. Onder de term "wit-regel" worden namelijk niet alleen tekstloze regels tussen twee alinea's bedoeld, maar ook de overgangen van onderste alinea van een kolom naar de bovenste alinea van de volgende kolom. Met de term worden dus alle overgangen bedoeld waarin twee alinea's niet direct op elkaar aansluiten in fysieke zin.
De informatiecategorieën Sollicitatie-Informatie en Adverteerdersgegevens worden in de meeste personeelsadvertenties vermengd en zonder duidelijke overgang weergegeven. Daarom is door MatchCare besloten die twee categorieën in het labelproces te beschouwen als één gezamelijke informatiecategorie.
Er geldt dat elke overgang tussen twee informatiecategorieën aangegeven wordt door een wit-regel (of eventueel door een alinea-einde), maar niet dat elke wit-regel een overgang representeert. Het komt dus voor dat bepaalde informatiecategorieën meerdere door wit-regels omgeven alinea's omvatten. In Tabel 6 wordt weergegeven wat het gemiddelde aantal alinea's per informatiecategorie is en tevens in hoeveel procent van de personeelsadvertenties er voor een bepaalde categorie slechts één alinea gebruikt wordt.

| | gem. aantal alinea's | % (aantal=1) |
|---|---|---|
| Werkgever | 2,4 | 27% |
| Functie | 1,1 | 79% |
| Profiel | 1,1 | 81% |
| Arbeidsvoorw. | 1,4 | 52% |
| Soll./Adv. | 2,8 | 12% |

**Tabel 6 – Twee relevante statistieken met betrekking tot de hoeveelheid alinea's die voor een informatiecategorie gebruikt worden.**

Uit deze getallen kan geconcludeerd worden dat de layoutgerichte wetmatigheden wel bijdragen aan de labeling, maar niet allemaal zo sterk als de tekstuele wetmatigheden. De layoutgerichte wetmatigheden zijn alleen sterk te noemen als het gaat om personeelsadvertenties die maar één tekstkolom bevatten. Wetmatigheid drie is zeer bruikbaar bij de bepaling van de grenzen van een informatiecategorie.

## 2.3 Tekstopmaakgerichte wetmatigheden

Een personeelsadvertentie maakt voornamelijk gebruik van variatie in de tekstopmaak om op bepaalde stukken tekst de aandacht te vestigen. De adverteerder kan variëren door verschillende lettergroottes of lettertypes te gebruiken of door tekst wel of niet vetgedrukt, cursief gedrukt of onderstreept weer te geven. Het duidelijkste voorbeeld daarvan wordt gegeven door de eerste wetmatigheid:

1. De functietitel wordt weergegeven in het meest opvallende lettertype: dit geldt voor álle personeelsadvertenties. Omdat de functietitel het belangrijkste stuk informatie van een personeelsadvertentie is, wordt daaraan het meest opvallende lettertype gegeven. Deze wetmatigheid geeft een zeer eenvoudige manier om de functietitel te detecteren;
2. Kopjes, vaak met een afwijkende opmaak, geven het begin van een nieuwe informatiecategorie aan: kopjes boven een alinea worden gebruikt om de lezer te vertellen wat de inhoud van de alinea onder het kopje is. Het kopje vormt dus eigenlijk al een labeling van de alinea eronder. In 79% van de personeelsadvertenties wordt gebruik gemaakt van kopjes. De personeelsadvertentie in Figuur 9 is daar een voorbeeld van. Bij die advertenties kan er gelabeld worden door de kopjes te detecteren en tekstueel te interpreteren. Het interpreteren van de kopjes kan het eenvoudigst gedaan worden door zo veel mogelijk voorkomende kopjes op te nemen in de lijsten van keywords. Er komen maar een klein aantal verschillende kopjes voor per informatiecategorie, dus de lijsten van keywords worden niet onnodig lang daardoor;
3. Als de Werkgeversomschrijving in een afzonderlijke kolom is geplaatst, dan heeft die kolom een afwijkende tekstopmaak: in een aantal voorgaande wetmatigheden is verteld dat er personeelsadvertenties zijn waarin de Werkgeversomschrijving in een afzonderlijke kolom is geplaatst. Dat komt voor bij 32% van de personeelsadvertenties. In Figuur 9 is een voorbeeld te zien van zo'n personeelsadvertentie. In ruim 95% van alle personeelsadvertenties waarin een afwijkende kolom met Werkgeversomschrijving voorkomt heeft die kolom een afwijkende tekstopmaak (meestal is de tekst cursief gedrukt). Deze wetmatigheid kan dus met een grote zekerheid toegepast worden als er tijdens het labelen een afzonderlijk geplaatste kolom met een afwijkende opmaak gevonden wordt.

Kort samengevat: door te kijken naar de tekstopmaak en de variatie in de tekstopmaak kunnen sterke aanwijzingen gevonden worden voor de toe te kennen labels. De best bruikbare aanwijzingen worden gegeven door kopjes die boven alinea's geplaatst zijn, waardoor het label van de alinea onder het kopje met grote zekerheid vastgesteld kan worden.

## 2.4 Statistische methoden

Dit afstudeeronderzoek maakt gebruik van een groot aantal statistieken. De statistieken worden in Hoofdstuk 2 gebruikt om de wetmatigheden die voorkomen in het domein van de personeelsadvertenties aannemelijk te maken en tevens aan te geven welke wetmatigheden het sterkst zijn. Door dat te bepalen is duidelijk geworden welke matigheden het beste gebruikt kunnen worden in de rule base van de geautomatiseerde labelmethodiek.

De statistische domeinverkenning is uitgevoerd over 723 personeelsadvertenties die verschenen in de Volkskrant van 18 mei, 25 mei en 1 juni 2002. Een uitgebreide beschrijving van de domeinverkenning is terug te vinden in Hoofdstuk 6.

Er zijn drie soorten statistische methoden gebruikt voor het verkrijgen van de statistieken:

1. Het simpelweg tellen hoe vaak een bepaalde eigenschap voorkomt in een personeelsadvertentie. Voorbeelden van statistieken die door deze methode bepaald zijn:
   - de aanwezigheid van wit-regels;
   - de aanwezigheid van kopjes;
   - het percentage personeelsadvertenties waarbij de informatiecategorieën Sollicitatie-informatie en Adverteerdersgegevens vermengd zijn;
   - het percentage personeelsadvertenties waarbij de apart geplaatste kolom Werkgeversomschrijving een afwijkende tekstopmaak heeft;
   - het percentage personeelsadvertenties waarbij de functietitel het grootste lettertype heeft.
2. Door berekeningen los te laten op de handmatige labelresultaten uit LeO. Deze statistieken zijn verkregen door een statistische module, VacStat, te implementeren (in Delphi) en die op te nemen in LeO, de labelapplicatie. VacStat maakt gebruik van de tekstsegmenten en de toegekende labels die door de labelmedewerkers handmatig zijn vastgelegd in LeO. VacStat doet dus niets met de in de personeelsadvertentie aanwezige grafische afbeeldingen (zoals bedrijfslogo's), want de labelmedewerkers leggen die onderdelen niet vast als segment. Voorbeelden van de statistieken die op deze manier verkregen zijn:
   - het gebruik van een vaste volgorde van de informatiecategorieën;
   - het aantal segmenten per informatiecategorie;
   - het percentage personeelsadvertenties waarbij alle categorieën vertegenwoordigd zijn;
   - de kansverdeling van de Qualitative Spatial Relations (de relatieve posities van de tekstsegmenten ten opzichte van elkaar). Ik maak gebruik van Qualitative Spatial Relations met inachtname van Thick Boundaries. Uitgangspunt zijn de handmatig gedefinieerde tekstblokken in LeO. Ik kijk naar elk gedefinieerd blok, dus als er 2 blokken W en 2 blokken F zijn, dan worden er 4 relaties uitgerekend;
   - de verdeling van de absolute posities van de tekstsegmenten

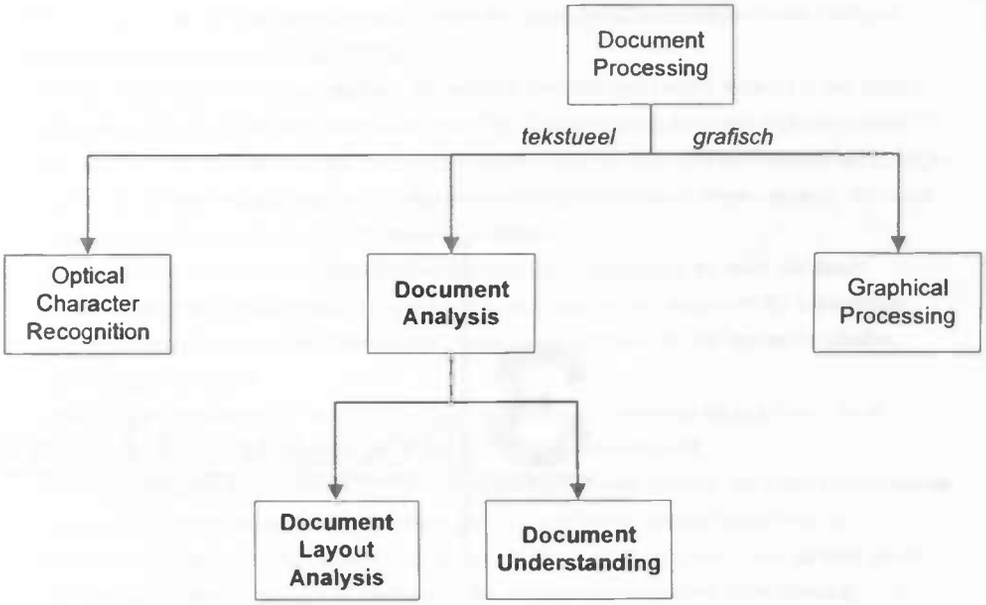
3. De statistische methode voor de bepaling van de effectiviteit van keywordlijsten maakt gebruik van de OCR-resultaten die werden verkregen uit LeO. Als OCR-applicatie werd gebruik gemaakt van het commerciële OCR-pakket OmniPage. De OCR-resultaten bevatten de vanuit de gescande afbeelding herkende tekst en enkele tekstopmaakeigenschappen zoals het lettertype en het wel of niet vetgedrukt, cursief gedrukt of onderstreept zijn van de tekst. Ook deze resultaten werden verkregen uit VacStat. VacStat berekent de effectiviteit van de keywordlijsten door simpelweg te tellen hoevaak elk keyword uit elke keywordlijst voorkomt in de ge-OCR-de tekst van de tekstsegmenten. Er wordt bijgehouden hoe vaak een gevonden keyword in het *juiste* tekstsegment voorkomt (de precision) en hoe vaak in een tekstsegment een keyword uit de juiste keywordlijst, dus de lijst die het label van het tekstsegment moet herkennen, bevat (de recall).
De keywordlijsten zijn handmatig samengesteld door middel van domeinkennis.

# 3 Document Processing

Hoofdstuk 3 bespreekt in het kort de onderzoeksgebieden die onder Document Processing vallen en gaat vervolgens dieper in op het onderzoeksgebied Document Analysis. Document Analysis houdt zich in het algemeen bezig met het omzetten van een gedrukt tekstdocument naar een digitale, symbolische representatie. Binnen MatchCare is dat het proces van het omzetten van een personeelsadvertentie op een gedrukte vacaturepagina naar een record in de centrale vacaturedatabase, oftewel het vacatureverwerkings-proces.
Het onderzoeksgebied Document Image Understanding, een deelgebied van Document Analysis, speelt in dit afstudeeronderzoek een centrale rol. Document Image Understanding houdt zich namelijk bezig het omzetten van de geometrische representatie van een document (de layout-representatie) naar een semantische representatie van het document. Dat is voor MatchCare te vertalen naar het toekennen van labels aan tekstblokken binnen een personeelsadvertentie. Document Image Understanding wordt uitvoerig beschreven in Hoofdstuk 4.

## 3.1 Document Analysis

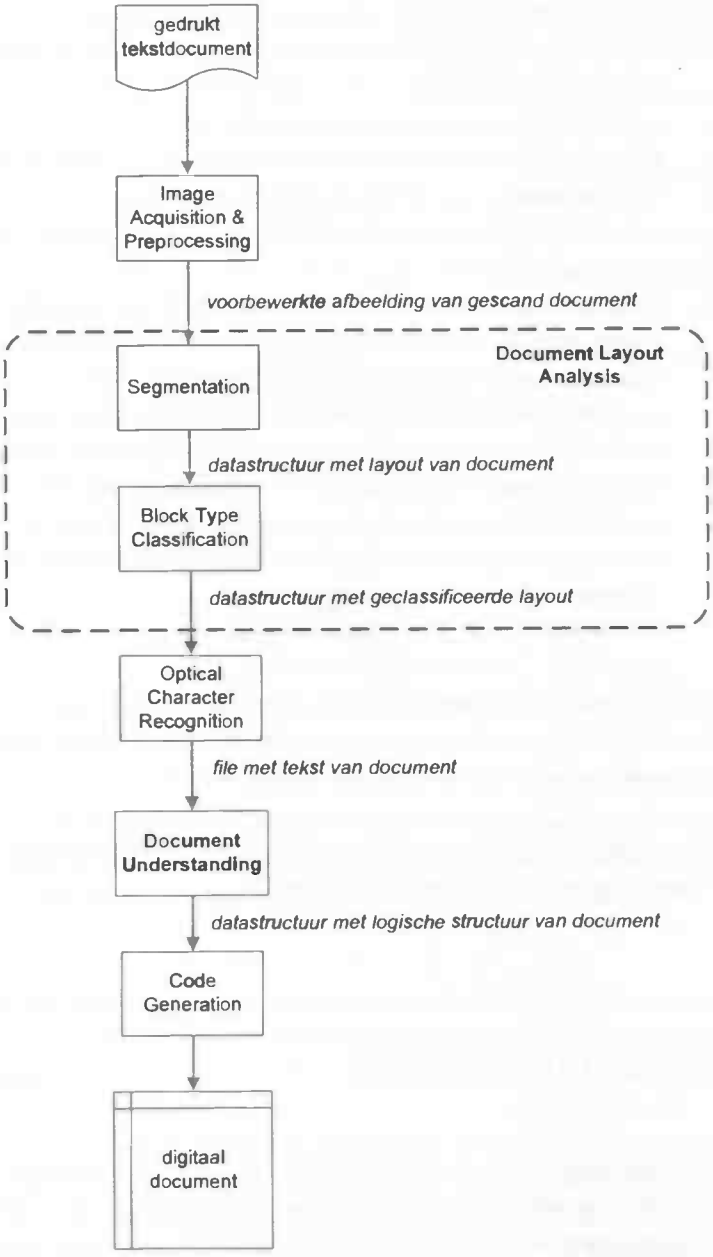
Het onderzoeksterrein Document Processing verricht onderzoek naar het afleiden van een digitale representatie uit een gedrukt document. Figuur 3.1 laat de verschillende deelgebieden van Document Processing zien.
Document Processing is op basis van het soort gedrukt document te delen in twee grote deelgebieden: Textual processing en Graphical processing. Textual Processing houdt zich bezig met het gedrukte documenten die voor het grootste gedeelte uit tekst bestaan, zoals kranten, wetenschappelijke artikelen en formulieren. Graphical Processing houdt zich daarentegen bezig met documenten die voor het overgrote deel een grafische content hebben. Voorbeelden daarvan zijn diagrammen van elektronische schakelingen, topografische en wegenkaarten, muziekpartituren en bouwontwerpen.




**Figuur 3.1 – Overzicht van onderzoeksgebieden die onder Document Processing vallen. De voor mijn afstudeeronderzoek relevante gebieden zijn vetgedrukt.**

Binnen de wetenschap bestaat er geen eenduidige naamgeving voor deze onderzoeksgebieden. Zo wordt Document Analysis ook wel Document (Image) Understanding genoemd (naar het uiteindelijke doel van het proces) en wordt Document Layout Analysis ook wel Document Analysis genoemd. Dat kan voor verwarring zorgen bij een literatuuronderzoek. In dit literatuuronderzoek wordt echter een constante en consistente naamgeving (de naamgeving uit Figuur 3.1) gehanteerd om verwarring te voorkomen.

Document Analysis is het proces dat een gedrukt document dat hoofdzakelijk bestaat uit tekst omzet naar een digitale, symbolische representatie van het document. Document Analysis is in feite is het precies het omgekeerde van het schrijven van een document, want de doelstelling is dezelfde documentrepresentatie te verkrijgen als die gebruikt werd bij het creëren van het document. MatchCare kent een proces dat precies aan deze omschrijving voldoet: het vacatureverwerkingsproces. In het vacatureverwerkingsproces, vaak verkort het vacatureproces genoemd, wordt een personeelsadvertentie op een gedrukte vacaturepagina omgezet naar een record in de centrale vacaturedatabase. Naast dit toepassingsgebied kan Document Analysis onder andere gebruikt worden voor het detecteren en lezen van adresgegevens op een envelop en voor het categoriseren en archiveren van documenten waardoor het mogelijk wordt eenvoudig gegevens terug te vinden in grote documentcollecties.
Document Analysis is een onderzoeksgebied waarin vele disciplines van de informatica bijeenkomen, waaronder image processing, pattern recognition, natural language processing, artificial intelligence en database systemen. Dat maakt het een zeer interessant onderzoeksgebied.




**Figuur 3.2 – het stroomdiagram van het Document Analysis-proces.**

In Figuur 3.2 is te zien dat een gedrukt tekstdocument een groot aantal subprocessen moet doorlopen voordat het omgezet is in een digitaal document:

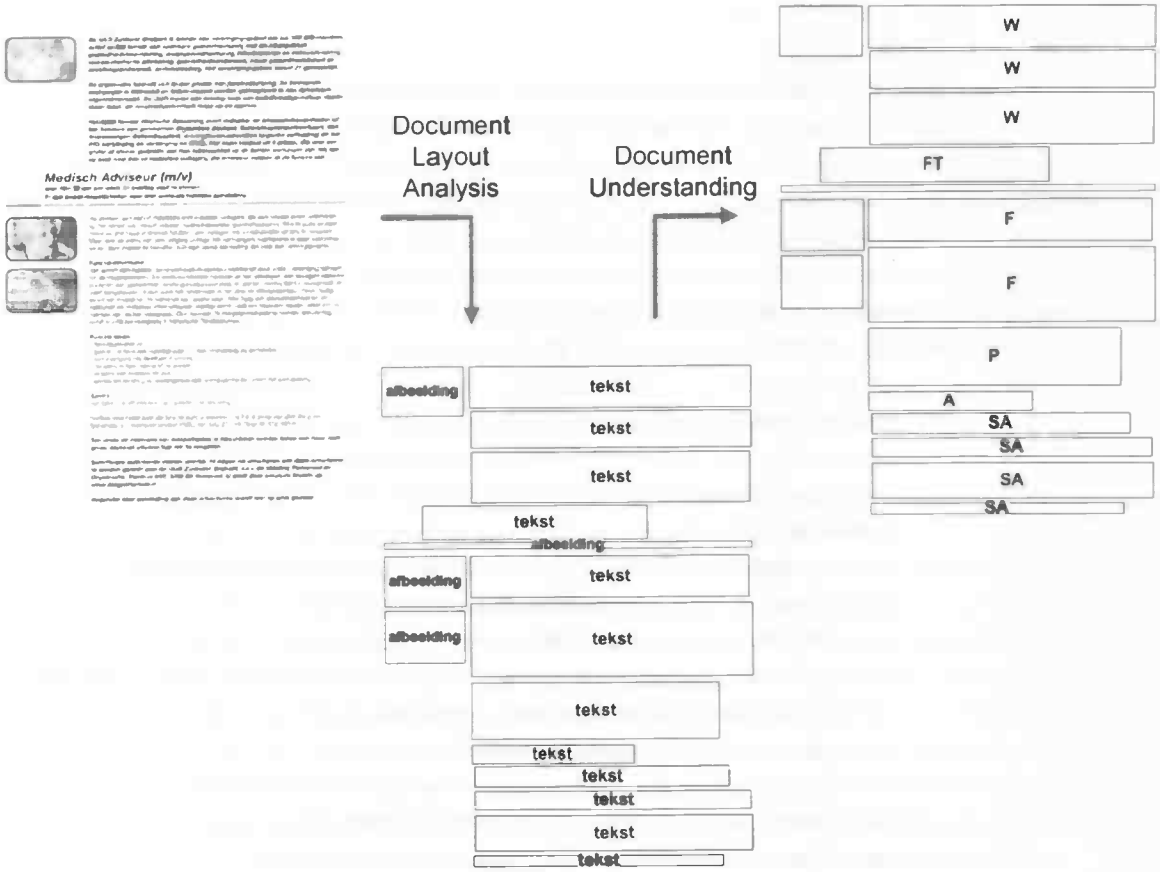
- **Image Acquisition & Preprocessing**: het gedrukte tekstdocument wordt omgezet in een digitale afbeelding (Image Acquisition). Dat gebeurt door het scannen van het document met een scanner of kopieerapparaat. Vervolgens wordt de digitale afbeelding van het document voorbewerkt om te zorgen dat het zo optimaal mogelijk aan de vervolgprocessen aangeboden wordt (Preprocessing). Dat wordt gedaan om de kans op fouten zo klein mogelijk te maken;
- **Segmentation**: de gescande afbeelding wordt opgedeeld in blokken die elk maar één soort informatie bevatten, resulterend in de geometrische structuur van het document. De belangrijkste soorten informatie zijn tekst en afbeeldingen. Overige soorten informatie zijn grafieken, tabellen, vergelijkingen en lijsten;
- **Block Type Classification**: dit subproces bepaalt welke soort informatie elk blok bevat. Dit is relevant omdat elke soort informatie een eigen verwerkingsmethode heeft;
- **Optical Character Recognition**: OCR zet de afbeelding van tekst om naar een digitale representatie van de tekst, bijvoorbeeld de ASCII-representatie. OCR kan tevens gebruikt worden om de tekstopmaak (lettertype, lettergrootte, vet/cursief/schuingedrukt) te bepalen. Zowel de tekst als de tekstopmaak kunnen nuttig gebruikt worden in de volgende stap, Document Understanding;
- **Document Understanding**: de gevonden blokken worden voorzien van een semantisch label. Met andere woorden, dit subproces bepaalt de betekenis van de gevonden blokken en bepaalt tevens de onderlinge relaties tussen de blokken;
- **Code Generation**: de semantische structuur van het document wordt omgezet naar een digitaal document.

De subprocessen Segmentation en Block Type Classification vormen samen de Document Layout Analysis. Figuur 3.3 laat de relatie zien tussen een gescande personeelsadvertentie, de gesegmenteerde personeelsadvertentie en de gelabelde personeelsadvertentie.

In de secties 3.2 en 3.3 komen korte literatuurstudies van achtereenvolgens de subprocessen Image Preprocessing en Document Layout Analysis aan bod.

## 3.2 Image Preprocessing

Het digitaliseren van een gedrukt tekstdocument door het scannen ervan geeft nooit een perfecte digitale afbeelding. De digitale afbeelding bevat altijd wel wat ruis (door ruis op het gedrukte document en door een niet-perfecte scanner of kopieerapparaat) en zal ook vaak een beetje scheef staan (doordat het gedrukte document niet helemaal recht op de scanner is gelegd). Om te zorgen dat de digitale afbeelding zo goed mogelijk verwerkt wordt is het nodig de gescande afbeelding te optimaliseren door te proberen deze 'verminkingen' ongedaan te maken.
Daarnaast is het verstandig de gescande afbeelding om te zetten van een kleuren- of gray scale afbeelding (de meeste scanners produceren een 8-bit gray scale afbeelding) naar een zwart-wit afbeelding.




**Figuur 3.3 – De relatie tussen een gescande personeelsadvertentie, de gesegmenteerde personeelsadvertentie (na het Document Layout Analysis-proces) en de gelabelde personeelsadvertentie (na het Document Understanding-proces).**

Deze drie afbeelding-naar-afbeelding transformaties worden tezamen Image Preprocessing genoemd. In de wetenschappelijke literatuur spreekt men respectievelijk over Noise Reduction, Skew Correction en Binarization.

**Binarization** transformeert de gescande afbeelding van een kleuren- of gray scale afbeelding naar een zwart-wit afbeelding. Dit is nuttig omdat het OCR- en het segmentatieproces eenvoudiger randen kunnen detecteren in een zwart-wit afbeelding dan in een kleuren- of gray scale afbeelding. Meestal wordt in het Binarization-subbproces een threshold gebruikt. Kleurwaarden gelijk aan en boven de threshold worden omgezet in de kleurwaarde wit en kleurwaarden onder de threshold worden omgezet in zwart.
Een adaptieve aanpak wordt gegeven in [8] en een aanpak die om kan gaan met tekst op een achtergrond die een texture bevat (een bijzonder moeilijk binarizationprobleem) wordt gegeven in [9].

**Noise Reduction** probeert de gescande afbeelding schoon te krijgen van ruis. Vaak wordt daarvoor image morphology (erosion en dilation achter elkaar toegepast) gebruikt.

**Skew Correction** zorgt ervoor dat de gescande afbeelding recht komt te staan. Het komt vaak voor de pagina's niet helemaal recht onder de scanner of het kopieerapparaat worden gelegd en er dus een afbeelding ontstaat waarop de tekst scheef staat. Dat kan hinderlijk zijn voor met name het segmentatieproces. Er zijn namelijk segmentatie-algoritmes waarbij ruis tussen twee tekstkolommen ervoor kan zorgen dat de twee kolommen als één kolom gezien worden.
Skew Correction-algoritmes bepalen eerst hoeveel graden de tekst uit de normale tekstrichting staat en roteren de gescande afbeelding vervolgens om de *skew* ongedaan te maken. Er zijn twee Skew Correction-algoritmes die gebruik maken van het feit dat in vrijwel alle tekstdocumenten de tekstregels langs parallelle rechte lijnen lopen:

1. **Het Hough-transformatie algoritme [10, 11]**: elke zwarte pixel $(x, y)$ wordt door de Hough-transformatie $\rho = x \cos(\theta) + y \sin(\theta)$ getransformeerd naar de parameterruimte $(\rho, \theta)$, de Hough-ruimte. Zwarte pixels die op een rechte lijn liggen geven pieken in de Hough-ruimte en geven daarmee de gezochte hoek aan;
2. **Het projectieprofiel algoritme [12]**: onder een aantal vooraf gedefinieerde hoeken wordt een projectieprofiel gemaakt. Een projectieprofiel is een afbeelding van de zwarte pixels op een projectie-as die onder een bepaalde hoek ten opzichte van de y-as staat. In het projectieprofiel worden de tekstregels (lange rijen zwarte pixels) weergegeven als pieken en de ruimte tussen de tekstregels (geen of zeer weinig zwarte pixels) als dalen. Deze aanpak bepaalt onder welke hoek de pieken en dalen optimaal zijn. Die hoek is de gezochte hoek.

Er zijn ook diverse aanpakken die gebruik maken van de geometrische structuur van een gescande afbeelding, maar die kunnen daardoor gebruikt worden in het Preprocessing-subproces. De geometrische structuur is immers tijdens het Image Preprocessing-proces nog niet bekend.

## 3.3 Document Layout Analysis

Het doel van Document Layout Analysis is de gescande afbeelding op te delen in blokken die elk maar één soort informatie bevatten. Het opdelen in blokken, oftewel segmenteren, wordt gedaan in het subproces Segmentation. Het classificeren van de gevonden blokken op basis van het soort informatie dat ze bevatten wordt gedaan in het subproces Block Type Classification. Deze subprocessen worden respectievelijk in de paragrafen 3.3.1 en 3.3.2 kort besproken. De output van Document Layout Analysis is de geometrische structuur van het document, weergegeven in een bepaald model. Paragraaf 3.3.3 geeft een kort overzicht van een aantal veelvoorkomende documentmodellen.
Document Layout Analysis bestaat uit twee deelprocessen, Segmentation en Block Type Classification, die hieronder worden besproken.

## 3.3.1 Segmentation

De verschillende algoritmes om de gescande afbeeling op te delen in blokken kunnen worden ingedeeld in drie categorieën: top-down algoritmes, bottom-up algoritmes en hybride algoritmes. De hybride algoritmes vormen een combinatie van top-down en bottom-up algoritmes.
Top-down algoritmes nemen de gehele afbeelding als uitgangspunt en splitsen de blokken op een interatieve manier steeds in kleinere blokken totdat het gewenste niveau is bereikt. Bottom-up algoritmes beginnen op pixelniveau, clusteren de pixels in connected components en clusteren vervolgens verder totdat het gewenste niveau is bereikt. Door in de algoritmes parameters te wijzigen is mogelijk het gewenste segmentatieniveau in te stellen. Er kan daardoor gesegmenteerd worden op bijvoorbeeld woordniveau, regelniveau of alineaniveau.

Een aantal veelgebruikte top-down algoritmes voor het segmenteren zijn:

- **Recursive X-Y Cuts** [13,14]: het Recursive X-Y Cut algoritme, ook wel RXYC genoemd, segmenteert de afbeelding op een recursieve manier in kleinere rechthoekige blokken. Bij elke stap van de recursie wordt, om en om, het horizontale of het verticale projectieprofiel van een blok berekend. Het blok wordt gesegmenteerd bij elk dal in het projectieprofiel waarvan de grootte boven een gegeven threshold ligt. Een dal in het projectieprofiel representeert een horizontale of verticale leegte in het blok. Door de threshold te variëren kan bepaald worden hoe groot een leegte mag zijn om tot een segmentatie van het blok te leiden. Deze techniek is zeer geschikt voor documenten waarvan bekend is dat ze zich houden aan de *Manhattan-layout*: alle blokken kunnen, na Skew Correction, geïsoleerd worden door een set van horizontale en verticale lijnsegmenten die door witte ruimte getrokken worden;
- **Run-Length Smearing Algorithm (RLSA)** [15]: dit algoritme 'smeert' de zwarte pixels in de gescande afbeelding uit in de x- en y-richting en maakt op die manier van elk blok één groot connected component. Het 'uitsmeren' wordt gedaan door voor de x- en y-richting een afzonderlijke threshold in te stellen en twee naburige zwarte pixels te verbinden met zwarte pixels als ze minder ver dan de ingestelde threshold in de te smeren richting uit elkaar liggen. De twee uitgesmeerde afbeeldingen worden vervolgens door middel van de pixelwise OR-operatie samengevoegd. Vervolgens kunnen de blokken gevonden worden door middel van het Connected Components-algoritme;
- **Shape-Directed Covers** [16]: dit algoritme segmenteert het document door zo groot mogelijke rechthoeken die alle wit-ruimte bedekken te vinden. Alle rechthoekige gebieden in het document die daardoor niet bedekt worden vormen de gezochte blokken.

Een paar voorkomende bottom-up algoritmes voor het segmenteren zijn:

- **Connected Components (CC)** [17]: het CC-algoritme bekijkt voor elke zwarte pixel alle 8 omliggende pixel en berekent op basis van de kleuren van de omliggende pixels welke kleur de zwarte pixel moet krijgen. Hierdoor worden de naburige pixels, die de afzonderlijke letters in het document omvatten, geclusterd. Door de initiële clusters vervolgens te gebruiken als basis voor de volgende clustering kunnen grotere componenten, zoals woorden en alinea's, verkregen worden. Connected Components kunnen ook berekend worden door een morphologische closing uit te voeren;
- **Docstrum** [18]: Dit algoritme berekent de *k*-nearest neighbors voor elke zwarte connected component in het document. Vervolgens wordt voor elk paar nearest-neighbors de afstand en hoek tussen de nearest-neighbors berekend. Door de connected components te clusteren met

gebruikname van de afstand- en hoekkenmerken kan het document gesegmenteerd worden. Door te werken met k-NN wordt het aantal paren dat moet worden onderzocht beperkt;

- **Segmentatie met behulp van Area Voronoi Diagrams** [19]: een Area Voronoi Diagram is een generalisatie van een normaal Point Voronoi Diagram. Een Area Voronoi Diagram wordt gegenereerd uit een set van niet-overlappende figuren (de connected components) in plaats van uit set van punten zoals bij het Point Voronoi Diagram. Een Area Voronoi Diagram wordt gegenereerd als volgt gegenereerd: eerst wordt een Point Voronoi Diagram geconstrueerd uit een set van sample points (zwarte pixels) op de contouren van de connected components. Vervolgens wordt het Area Voronoi Diagram verkregen door uit het Point Voronoi Diagram alle edges te verwijderen die zijn gegenereerd door een paar van punten die afkomstig zijn uit hetzelfde connected component. De afbeelding kan vervolgens gesegmenteerd worden door uit het Area Voronoi Diagram de juiste edges te selecteren.

De hybride algoritmes maken gebruik van zowel een top-down als een bottom-up algoritme bij het segmenteren van een gescande afbeelding. Een typisch voorbeeld van een hybride algoritme wordt gevormd door de combinatie van het RLSA- en het CC-algoritme [71]. Het RLSA-algoritme wordt daarin gebruikt om de aanwezige tekstblokken aan te geven (door de zwarte pixels te smearen) en het CC-algoritme wordt daarna gebruikt om de tekstblokken te lokaliseren. Ook alle Global Optimization-technieken zijn hybride.

De keuze voor een bepaald algoritme hangt sterk af van de eigenschappen van de gescande afbeelding. Factoren als benodigde rekentijd en gevoeligheid voor ruis en *skew* spelen daarbij ook een rol.
Top-down algoritmes zijn vaak sneller en minder gevoelig voor input parameters (de parameters mogen afwijken van de ideale waarde), maar vereisen dat het document een Manhattan-layout (een indeling in rechthoekige blokken die gescheiden worden door wit-ruimte) heeft. Bottom-up en hybride algoritmes zijn vaak nauwkeuriger, maar wel tegen de prijs van een grotere complexiteit (moeilijkere berekeningen en daardoor minder snel) en een grotere gevoeligheid voor input parameters.

## 3.3.2 Block Type Classification

Voor het Document Understandingproces is het essentieel te weten wat voor soort informatie elk gevonden blok binnen het document bevat. Elk soort informatie vereist namelijk een andere verwerkingsmethode. Zo worden tekstblokken vaak doorgezet naar een OCR-applicatie voor het verkrijgen van de tekst en worden grafische blokken verwerkt door een Graphical Processing-component.
Het bepalen van het soort informatie dat een blok bevat wordt gedaan in het subbproces Block Type Classification. In de meeste Document Analysis-methodieken wordt het gezien als een losstaand proces dat na het Segmentationproces wordt uitgevoerd, maar in sommige methodieken worden de segmentatie en het classificeren van de blokken gezien als een ondeelbaar probleem.
In de meeste Block Type Classification-algoritmes wordt alleen onderscheid gemaakt tussen tekstblokken en niet-tekstblokken (vaak grafische blokken). Dat komt doordat de meeste Document Analysis-methodieken alleen tekstblokken gebruiken voor verdere verwerken en de niet-tekstblokken negeren.
In deze paragraaf worden een aantal methodes besproken die gebruikt kunnen worden voor Block Type Classification.

De meeste methodes maken gebruik van het feit dat de texture van een tekstblok significant verschilt van de texture van een afbeelding. Enkele voorbeelden van texture-based methoden zijn:

- **Filter-based recognition** [20, 21]: de gedachte achter de methode in [20] is dat tekst in een gescande afbeelding van een document een unieke texture heeft die eenvoudig gedetecteerd kan worden met een kleine set Gabor-filters. Gabor-filters bezitten namelijk de positieve eigenschap dat ze erg goed gebruikt kunnen worden voor het lokaliseren van bepaalde frequenties. De frequenties in deze toepassing zijn de zwarte pixel-frequenties die voorgekomen in een segment. De set van Gabor-filters is getest op een verzameling van 382 handmatig geclassificeerde gescande krantenpagina's en daarbij werd 97.2 % van de segmenten goed geclassificeerd (als tekst of als niet-tekst).
  In [21] wordt de classificatie bepaald op basis van de karakteristieken van de texture-eigenschappen van de neighborhood van elk pixel. Daarvoor zijn filters ontworpen die gebruik maken van heuristieken en domeingebonden wetmatigheden. Met behulp van een multilayer neural netwerk worden worden de filters geoptimaliseerd;
- **Histogram-based recognition** [22, 72]: in [22] wordt onderscheid gemaakt tussen tekst en afbeeldingen op basis van de (regel)dikte van de connected components. Eerst worden de connected components van de gescande afbeelding bepaald. Vervolgens wordt er een histogram gemaakt van de hoogtes van de connected components. Het histogram vormt een maat voor de (regel)dikte in een connected component. Als die dikte beneden een bepaalde threshold blijft is het connected component een tekstblok. In de andere gevallen is het connected component een afbeelding of een stuk tekst met een grote regeldikte (bijvoorbeeld een krantenkop).
  In [72] worden de diverse features samengevat in een zogenaamde Feature Based Interaction Map (FBIM), die statistische pixel interacties afbeeldt door middel van de ruimtelijke verdeling in een histogram. De performance van dit algoritme is getest aan de hand van 4713 handmatig geclassificeerde segmenten in de image database van de University of Washington, de ground-truth. De ground-truth wordt bij de evaluatie van een algoritme gebruikt als het uitgangspunt van correct geclassificeerde segmenten waarmee de output van het algoritme vergeleken kan worden. De tekstsegmenten werden in 97% van de gevallen correct geclassificeerd, terwijl de niet-tekstsegmenten maar in 89% van de gevallen correct werden geclassificeerd;
- **Feature-based recognition** [23, 25, 26]: in [23] wordt gebruik gemaakt van diverse fysieke eigenschappen van de connected components om het block type te bepalen. De gebruikte eigenschappen van de connected components zijn grootte, hoogte, breedte, dikte, dichtheid, statistieken met betrekking tot maximale lengte van rijen aaneengesloten zwarte pixels, aantal en verdeling van gaten en de grootte van aangrenzende connected components. Het algoritme maakt onderscheid tussen tekst, afbeeldingen, lijnen, tabellen, schuine lijnen en lijnafbeeldingen (bijvoorbeeld grafieken). Er wordt een decision tree classifier gebruikt om de segmenten te classificeren. De decision tree wordt getraind en getest met behulp van de Cross Validation-techniek. Om de performance van het algoritme te kunnen bepalen hebben de auteurs drie ground-truth databases samengesteld: University of Washington Document Image Database UW-I, UW-II en UW-III. De geometrische structuur is in de ground-truth databases vastgelegd in de DAFS-standaard (zie paragraaf 3.3.3). De tekstsegmenten worden in bijna 98% van de gevallen correct herkend.
  In [25] worden ook diverse features van de segmenten gebruikt voor de classificatie, maar in dit artikel worden vier verschillende neural netwerken met elkaar vergeleken. Alle neurale netwerken worden getraind en getest door middel van Cross Validation. Het Back-Propagation Neural Network heeft een performance van 99.35 %, het Radial Basis Function Neural Network een performance van 99.61 %, het Probabilistic Neural Network een performance van 98.18 % en Kohonen's Self-Organizing Feature Maps Neural Network een performance van 99.22 %;

- **Analysis of local properties** [24]: er zijn diverse classification-algoritmes die gebruik maken van de analyse van lokale eigenschappen. Dit artikel bespreekt een tweetal algoritmes die het document eerst opdelen in kleine vierkante overlappende windows waarvoor diverse features worden bepaald. In het eerste algoritme wordt het ruimtelijke Fourier spectrum berekend en gebruikt om de classificatie te verrichten. In het tweede algoritme wordt de classificatie verricht op basis van statistische eigenschappen van de lokale grijswaardenhistogrammen. Er wordt daarbij gebruikt gemaakt van de aanname dat de lichtste grijswaarden voorkomen in de achtergrond van tekstblokken.

### 3.3.3 Documentmodellen

Een documentmodel, ook wel een *generic document* genoemd, wordt gebruikt om de structuurregels voor een klasse van documenten te beschrijven. Het bepaalt dus op een generieke manier hoe de documenten die tot de klasse behoren opgebouwd kunnen worden. Een document heeft in het algemeen twee soorten structuren, te weten de geometrische structuur (ook wel layoutstructuur genoemd) en de semantische structuur. De geometrische structuur is een set van documentobjecten (zoals letters, woorden, alinea's, afbeeldingen) samen met hun attributen (zoals positie, dichtheid en kleur) en de relaties tussen de documentobjecten (de hiërarchie). De geometrische structuur bevat dus kortgezegd een beschrijving van de layout van het document. De semantische structuur geeft een hiërarchische beschrijving van de betekenis van de documentobjecten en van de relaties tussen de documentobjecten (zoals de leesvolgorde).

Onderstaande opsomming bevat een aantal voorkomende documentmodellen:
**X-Y Tree**: dit model wordt gebruikt in combinatie met RXYC-segmentatie. De X-Y Tree geeft een hiërarchische representatie van het gescande document en volgt daarbij de horizontal en verticale opdeling die door het RXYC-algoritme wordt gecreëerd. Bij elke splitsing van een segment (de *cut*) wordt een tree node aangemaakt waaronder de twee nieuwe segmenten worden gehangen. Elke tree leaf representeert uiteindelijk een gesegmenteerd blok inclusief een frame met blokattributen. In een gelabelde tree bevat elke tree leaf tevens de semantische betekenis (het label) van het gesegmenteerde blok. In Figuur 3.4 wordt een voorbeeld van de constructie van een X-Y Tree gegeven.
**Style parameters**: style parameters representeren attributen van gesegmenteerde blokken, zoals de grootte, en de ruimte tussen opeenvolgende blokken. De geometrische structuur wordt echter niet volledig beschreven door de parameters.
**Document Grammar**: Document Grammars bevatten een set deterministische productieregels waarmee de structuur van een documentklasse exact vastgelegd wordt. Zie sectie 4.2 voor een nadere uitleg.
**Graph** [27]: in dit artikel wordt een document beschreven als een graph. De graph bestaat uit layout objecten waartussen edges gedefinieerd zijn. Elke edge representeert de Qualitative Spatial Relation tussen de twee objecten die door de edge verbonden worden.
**Statistical Document Model** [28]: in dit documentmodel wordt de logische boomstructuur van een document gerepresenteerd door kansen op bepaalde lokale tree node patronen. De lokale tree node patronen lijken op n-grams. N-grams is een model dat aanneemt dat alleen de vorige $n-1$ woorden effect hebben op de kansen voor het volgende woord. De kracht van n-grams is dat ze documentobjecten in een incomplete logische boomstructuur kunnen voorspellen, als hun context maar bekend is.

**Area Voronoi Diagram** [19]: zoals beschreven in paragraaf 3.3.1 kunnen geometrische structuren ook gerepresenteerd worden door Area Voronoi Diagrams. De voordelen van Area Voronoi Diagrams zijn dat ze niet afhankelijk zijn van de skew van een afbeelding, niet vereisen dat de segmenten rechthoekig zijn, snel berekend kunnen worden en gebruikt kunnen worden als basis voor het vinden van de connected components in een afbeelding.

Er zijn drie ISO-documentformats gedefinieerd die dienen als internationaal geaccepteerde standaarden voor de representatie van documentinformatie. Door het gebruik van internationale standaarden wordt het eenvoudig uitwisselen van documentinformatie mogelijk gemaakt:

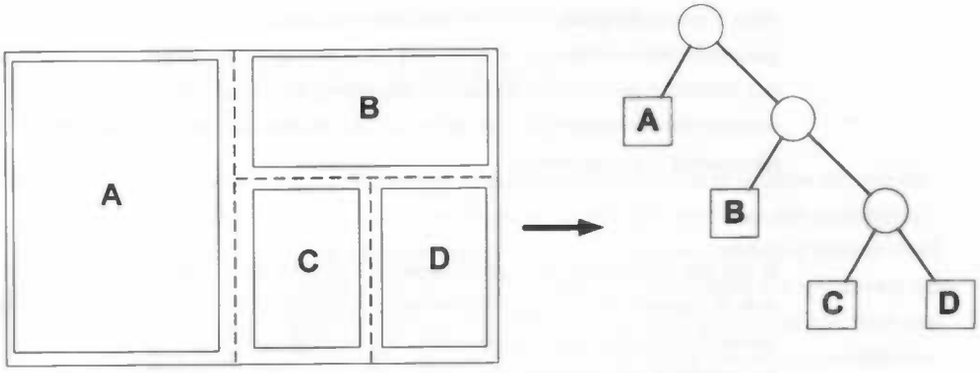
**Open/Office Document Architecture (ODA)** [30]: ODA is in ontwikkeld om de geometrische en semantische structuur van simpele office documenten te representeren. ODA is object-georiënteerd.

**Standard Generalized Markup Language (SGML)** [31]: SGML maakt het mogelijk de geometrische en logische structuur van een document evenals de onderlingen relaties tussen de documentonderdelen te beschrijven in de vorm van reguliere expressies. SGML is een eenvoudigere standaard dan ODA omdat SGML de structuur beschrijft van documenten die tot een voorgedefinieerde klasse behoren. SGML is markup-georiënteerd in tegenstelling tot ODA.

**eXtensible Markup Language (XML)** [32]: XML, voortgekomen uit SGML, is een standaardtaal voor het definiëren van gestructureerde documenten die hoofdzakelijk bestaan uit tekst. XML wordt gebruikt voor data management, data uitwisseling, data conversie en data distributie. XML maakt het mogelijk zowel de geometrische als de semantische structuur van een document vast te leggen. XML maakt het mogelijk het concept van de Document Type Definition (DTD) te gebruiken. DTD heeft een formele set regels waarmee de semantische structuur van een document gedefinieerd kan worden, beschrijft de soorten documentobjecten die gebruikt mogen worden en bepaalt welke relaties er mogelijk zijn tussen de documentobjecten. Omdat het World Wide Web tegenwoordig dienst als standaardplatform voor veel soorten communicatie (bijvoorbeeld voor digitale documentbibliotheken) is XML een veelgebruikt documentformat.

Naast deze ISO-standaarden is nog een andere standaard gedefinieerd:

**Document Attribute Format Specification (DAFS)** [29]: DAFS is een file format, gebaseerd op SGML, waarmee segmentatie-informatie uitgewisseld kan worden tussen applicaties. Het is speciaal ontwikkeld voor het representeren van zowel de geometrische als de semantische informatie die uit een gescand document gehaald kan worden en is daarmee zeer geschikt voor de Document Analysis- en Document Understanding-processen. DAFS beschikt over een format voor het opdelen van documenten in gestandaardiseerde onderdelen, voor het definiëren van attributen van de onderdelen en voor het labelen van de inhoud en van de attributen van de onderdelen. DAFS maakt het mogelijk de hiërarchische structuur van een document weer te geven. Het verschil tussen DAFS en SGML is dat DAFS extra mogelijkheden biedt voor het voor het beheren van opmaakeigenschappen, opslag van gescande afbeeldingen en de informatie-overdracht tussen tussen verschillende DTD's.




**Figuur 3.4 – Een voorbeeld van de constructie van een X-Y Tree uit de segmenten van een gescande personeelsadvertentie.**

# 4 Document Understanding

Hoofdstuk 4 bevat een literatuuroverzicht van vakgebied Document Understanding. Document Understanding wordt gedefinieerd als het proces dat de semantische structuur van een gegeven document bepaalt. Wellicht ten overvloede, de term labelen staat voor het toekennen van semantische labels aan de geometrische objecten die door het DLA-proces in het document zijn gevonden. Met de layout van een document wordt de geometrische structuur van het document bedoeld.
De ontwerpkeuzes voor het geautomatiseerd labelalgoritme, het hoofdonderwerp van dit afstudeeronderzoek, worden gemaakt op basis van dit literatuuroverzicht. In Hoofdstuk 5 wordt het geautomatiseerd labelalgoritme beschreven.

De voorkomende Document Understanding-methodieken kunnen worden geclassificeerd op basis van de karateristieke concepten, de basisgedachten, die de methodes gebruiken om tot een semantische structuur van het document te komen. Het komt ook voor dat algoritmes een aantal concepten combineren. Verreweg de meeste methodieken die voor het Document Understanding-vraagstuk zijn ontwikkeld zijn gebaseerd op *rules*. Dat wil zeggen, de methodieken gebruik diverse soorten rules om de geometrische structuur te vertalen naar de semantische structuur. Een onderverdeling in deze categorie methodieken kan gemaakt worden op basis van de vraag of een methodiek wel of niet gebruik maakt van een documentmodel. Een tweede categorie methodieken maakt gebruik van Text Mining. Text Mining wordt gebruikt om patronen, modellen of rules af te leiden uit ongestructureerde tekst.
Sectie 4.1 geeft een bespreking van de methodieken die zijn gebaseerd op rules, sectie 4.2 bespreekt de methodieken die zijn gebaseerd op rules én gebruik maken van een documentmodel en sectie 4.3 bespreekt als laatste de methodieken die zijn gebaseerd op Text Mining.

## 4.1 Methodieken gebaseerd op rules

Methodieken die zijn gebaseerd op rules gebruiken bijna allemaal de geometrische structuur (oftewel de layout) van een document als uitgangspunt. Die methodieken bepalen door middel van een verzameling van rules (in de vakliteratuur een *rule base* genoemd) hoe elk geometrisch documentobject vertaald moet worden naar een semantisch documentobject. De rule bases bevatten altijd minimaal een verzameling rules die bepalen hoe de geometrische structuur vertaald kan worden naar de semantische structuur, maar vaak ook een verzameling rules die gebruik maken van niet-geometrische eigenschappen van de geometrische documentobjecten, zoals inhoudelijke, tekstuele en tekstopmaakeigenschappen. Rule bases worden in de vakliteratuur ook wel omschreven met de term Formatting Knowledge.
Voorbeelden van geometrische eigenschappen van een documentobject zijn: absolute positie, hoogte, breedte en relatieve positie ten opzichte van de andere objecten. Voorbeelden van inhoudelijke eigenschappen zijn: aantal zwarte pixels, aantal witte pixels en aantal zwart-wit overgangen. Een belangrijke tekstuele eigenschap is het wel of niet aanwezig zijn van voorgedefinieerde keywords. Voorbeelden van tekstopmaakeigenschappen zijn: lettergrootte, lettertype en wel of niet vetgedrukt, schuingedrukt en/of onderstreept.

Het is voor de precisie van de methodieken essentieel dat de geometrische structuur zó is bepaald dat elke object uit de geometrische structuur ook maar één semantisch object bevat. Het is anders niet mogelijk elk semantisch object waaruit het document bestaat te vinden. De geometrische objecten kunnen namelijk niet verder opgesplitst worden in het Document Understanding-proces. Een manier om deze eis te kunnen omzeilen wordt door Ishitani beschreven in [33]. In zijn artikel beschrijft Ishitani hoe het Document Layout Analysis-proces en het Document Understanding-proces tegelijkertijd kunnen worden uitgevoerd en daarbij kunnen samenwerken door geometrische en semantische kennis aan elkaar uit te wisselen. Op die manier is het mogelijk door het vinden van bepaalde informatie alsnog een geometrisch documentobject op te splitsen. Volgens Ishitani is het door deze methodiek mogelijk complexere documentstructuren te analyseren en toch een goede performance te halen.

Er zijn ook methodieken die als basis niet de geometrische structuur gebruiken maar een documentmodel. Die methodieken proberen bijna allemaal tegelijkertijd de geometrische én de semantische structuur van het document te vinden (zie sectie 4.2).

Een voorbeeld van een methodiek die gebruikt maakt van een rule base wordt gegeven door Kuiper en Wieringa in [34]. Zij stellen een fuzzy rule base samen op basis van wetmatigheden die de geometrische relaties tussen de verschillende documentobjecten beschrijven. Een fuzzy rule geeft geen harde beslissingen maar geeft een confidence rate dat een bepaald geometrisch documentobject tot een bepaalde semantische categorie (label) behoort. Bij meerdere mogelijke labels wordt het label met de hoogste confidence rate genomen. Er zijn een aantal verschillende experimenten uitgevoerd. Gemiddeld genomen werd ongeveer 85% van de objecten goed gelabeld.
Een ander voorbeeld van het gebruik van een fuzzy rule base wordt gegeven in [35]. In dit artikel wordt getest met behulp van de University of Washington Document Corpus. Deze document database bevat 979 pagina's van wetenschappelijke artikelen uit diverse journals en over diverse onderwerpen. Elke pagina heeft een segment ground-truth, dat wil zeggen dat elk segment op de pagina gedefinieerd is door een rechthoek en een label. Gemiddeld genomen werd er een performance van 98% behaald.

Een andere methodiek die gebruikt maakt van rules is de Tree Transformation-methodiek. Deze methodiek ziet Document Understanding als een transformatie van de geometrische boomstructuur van een document naar de semantische boomstructuur. In deze methodiek wordt de geometrische structuur zoals gezegd gerepresenteerd door een boomstructuur. De rules in de rule base maken alleen gebruik van de geometrische structuur om te bepalen welk label elk geometrisch documentobject moet krijgen. Er zijn ook rules die bepalen of geometrische documentobjecten in de semantische boomstructuur op een andere plek moeten komen te hangen. Het is namelijk mogelijk dat de semantische structuur afwijkt van de geometrische structuur.
De artikelen [36] en [37] beschrijven een implementatie van deze methodiek. In [37] wordt naast de rule base gebruikt gemaakt van een virtuele field separator-methodiek. Die methodiek houdt in dat er field separators (wit-ruimtes tussen documentobjecten) in de geometrische boomstructuur worden opgenomen als zijnde virtuele geometrische documentobjecten. De rule base maakt gebruik van deze virtuele objecten. Dit lijkt enigszins op het aandachtsprofiel dat in sectie 5.5 geïntroduceerd zal worden: beide methodieken maken gebruik van de aanwezige wit-ruimte tussen de documentobjecten. Het verschil is dat het aandachtsprofiel een weergave is van de bestaande geometrische structuur en dat de virtuele field separator-methodiek de geometrische structuur aanpast door er virtuele objecten in te plaatsen.

## 4.2 Methodieken gebaseerd op een documentmodel

Een documentmodel is een model dat domeinspecifieke kennis representeert. Die domeinspecifieke kennis wordt gebruikt in de rules waarmee een document gelabeld wordt. Iedere toegestane documentklasse wordt hierbij opgevat als een apart domein dat een aantal invariante eigenschappen heeft waarmee het onderscheiden kan worden van de andere documentklasses. Domeinspecifieke kennis is dus invariante kennis over een documentklasse op basis waarvan er onderscheid gemaakt kan worden tussen de verschillende documentklasses. De domeinspecifieke kennis wordt verzameld in een zogenaamde *knowledge base*, een database van kennis. Op basis van de knowledge base kan er vervolgens een aparte rule base aangemaakt worden voor elke documentklasse, maar er kan ook één rule base gedefinieerd worden waarvoor de domeinspecifieke kennis omzet wordt in diverse parameters die gebruikt worden in de rules.
De precisie van deze methodieken is er sterk van afhankelijk of er bij een te labelen document wel het juiste model en daarmee de juiste knowledge base gebruikt wordt. Het herkennen van de documentklasse van een gegeven document is bij deze methodieken dus essentieel.

De methodieken die gebaseerd zijn op domeinspecifieke kennis kunnen worden opgedeeld in methodieken die gebruik maken van handmatig samengestelde verzamelingen van rules en methodieken waarbij de verzameling van rules wordt verkregen door een machine learning-algoritme.

Le et al. beschrijven in [38] een typische methodiek die gebaseerd is op een verzameling van handmatig gedefinieerde rules. Ze beschrijven een methodiek voor het labelen van wetenschappelijke artikelen en maken daarbij gebruik van de aanname dat elk journal waar de artikelen in verschijnen een aantal invariante layouteigenschappen heeft die geformuleerd kunnen worden in een verzameling rules. Dat wordt *domein-specifieke kennis* genoemd. Een database met domeinspecifieke kennis wordt een *knowledge base* genoemd. In deze methodiek wordt gebruik gemaakt van een OCR-systeem om informatie aan te leveren

over de geometrische documentobjecten. De OCR-informatie wordt gebruikt om geometrische en niet-geometrische eigenschappen van de geometrische documentobjecten te berekenen. De rules maken vervolgens gebruik van de berekende eigenschappen om tot een labeling te komen. Als eerste stap in het Document Understanding-proces wordt de naam van het journal waaruit het artikel afkomstig is door (een kleine verzameling) rules bepaald. Het systeem heeft een database met daarin voor elk journal een verzameling relevante karakteristieken. Deze domein-specifieke kennis wordt gebruikt om parameters aan te leveren voor de rules. Het documentmodel in deze methodiek bestaat dus uit een verzameling invariante layouteigenschappen die gelden voor de bepaalde documentklasses. Over een verzameling van 1407 wetenschappelijke artikelen werd een performance van 99.6 % behaald.

Kim et al. beschrijven in [39] een vergelijkbare methodiek. De methodiek is verdeeld in twee submodules: General Label Type Module (GLTM) en Arbitrary Label Type Module (ALTM). Als eerste stap wordt het ISSN-nummer van het journal waar te labelen wetenschappelijk artikel in verscheen opgezocht in een database. De database bevat journalspecifieke kennis die bij de labeling gebruikt kan worden. De journalspecifieke kennis geeft onder andere aan welk layoutype het journal gebruikt. Voor de meest voorkomende layouttypes zijn aparte rule bases gemaakt (in de GLTM). Voor weinig voorkomende layouttypes wordt een algemene (dus niet domeinspecifieke) rule base gebruikt. Over een verzameling van 1149 wetenschappelijke artikelen werd een performance van 97.91 % behaald.

Een ander voorbeeld van het gebruik van een knowledge base wordt gegeven in [40]. De methodiek die beschreven wordt in dit artikel verzamelt voor alle te herkennen documentklasses domeinspecifieke kennis, de geometrische structuur en de corresponderende semantische structuur in een knowledge base. Als een document gelabeld wordt, dan wordt de geometrische structuur van het documen gematcht met de geometrische structuur in de knowledge base door een gegeneraliseerde Hamming-metriek te gebruiken om *confidence rates* te bepalen. Een confidence rate geeft aan met welke waarschijnlijkheid een geometrisch object een bepaalde semantische waarde heeft. De geometrische structuur wordt omgezet naar de semantische structuur waarvoor de confidence rates het hoogst zijn.

Andere methodieken die gebruik maken van document models worden gegeven in [41], [42] (gebruikt een tree als document model) en [43].

De secties 4.2.1 t/m 4.2.4 beschrijven een aantal concepten die gebruikt worden bij het labelen met behulp van een documentmodel. Achtereenvolgens worden Template Matching, Graph Matching, Language Models en Machine Learning beschreven.

### 4.2.1 Template Matching

In [44] wordt een template matching methodiek beschreven. De methodiek maakt gebruik van geometrische en niet-geometrische eigenschappen van de geometrische documentobjecten en definieert tevens een aantal nieuwe eigenschappen waarmee de inhoud van de structuur van de geometrische documentobjecten genormaliseerd wordt. Verschillende documenten bevatten al gauw een verschillend aantal regels bij een bepaalde semantische categorie en daardoor zal bijna elk paar documenten verschillende absolute geometrische eigenschappen (zoals de absolute coördinaten van het documentobject) hebben. De documenten worden genormaliseerd om te voorkomen dat het labelen

hierdoor mislukt. De eigenschappen die het genormaliseerde document representeren worden door een template matching-algoritme vergeleken met een aantal vooraf gedefinieerde genormaliseerde documentstructuren. Het resultaat van deze vergelijking dient samen met de overige eigenschappen als input voor een rule-base waarmee de semantische structuur verkregen wordt.

Summers beschrijft in [45] een methodiek die, gebruikmakend van een template matching algoritme, de afstand van een geometrisch documentobject tot een bepaald vooraf gedefinieerd prototype berekent. Voor elke semantische categorie (label) is een prototype gedefinieerd op basis van geometrische en inhoudelijke eigenschappen van documentobjecten. De gevonden afstand is een maat voor de waarschijnlijkheid dat het documentobject tot de semantische categorie behoort.

### 4.2.2 Graph Matching

Bij de Graph Matching-methodiek wordt de geometrische structuur van een document gerepresenteerd door een graph-structuur. Voor elke documentklasse wordt een graph model gedefinieerd die o.a. gebruik maakt van de geometrische eigenschappen van de documentklasse.
In [46] wordt de geometrische structuur van een document gerepresenteerd door een attributed relational graph waarin alle nodes onderling verbonden zijn. Elke node in de graph correspondeert met een geometrisch documentobject. De attributen van de nodes bevatten geometrische en tekstopmaakgerichte eigenschappen. Een edge tussen twee nodes representeert de relatieve positie van de corresponderende geometrische objecten ten opzichte van elkaar. Het graph model wordt geconstrueerd door middel van een adaptive learning-methode die geïnspireerd is op de back propagation learning-methode die wordt gebruikt bij het trainen van neurale netwerken. De methodiek past steeds het model aan met behulp van error feedback. Het labelen gebeurt door de beste match van het te labelen document met een model te vinden. Dat wordt gedaan door een match cost-functie te minimaliseren. De dataset waarop de methodiek getest werd bestond uit 80 titelpagina's van wetenschappelijke artikelen uit 4 verschillende journals. Er werd een performance van ongeveer 92% behaald.

### 4.2.3 Language models

In language models worden de documentklasses beschreven door een bepaald formeel taalmodel. Het formele taalmodel bevat regels die de structuur van documenten uit de betreffende documentklasse weergeven. Voorbeelden van language models zijn Formal Document Grammars (waaronder Context Free Grammars), N-grams, Hidden Markov Models en stochastische language models.

**Formal Document Grammars**
Formal Document Grammars, vaak weergegeven door Context Free Grammars (CFG), bevatten een set deterministische productieregels waarmee de structuur van een documentklasse exact vastgelegd wordt. De non-terminalproductieregels bepalen de geometrische structuur van de documentklasse. De herkenningsregels, dat zijn de terminalproducties van de grammar, bepalen de mogelijke labels (de set van terminalsymbolen). Grammars zijn zeer geschikt voor documentklasses die een (vooraf bekende) duidelijke structuur hebben, zoals woordenboeken of telefoonboeken.
Bij methodieken die gebruik maken van document grammars wordt het Document Layout Analysis- en het Document Understanding-proces tegelijkertijd verricht door een syntactische analyse van de document

grammar uit te voeren. Document grammars maken dat mogelijk doordat ze regels bevatten voor de geometrische structuur (de non-terminalproductieregels) én voor de semantische structuur (de terminalproductieregels). De syntactische analyse wordt gedaan door een parser die afgeleid wordt uit de document grammar.
In de artikelen [47-52] zijn voorbeelden van deze aanpak te vinden.

**Formal Definition Languages**
Het is ook mogelijk documenten op een formele manier te beschrijven met behulp van Formal Definition Languages. In [53] bijvoorbeeld, wordt gebruik gemaakt van een Lisp-achtige taal die gebaseerd is op rechthoekige objecten. In [54] wordt een ander voorbeeld van een Formal Definition Language (daar Formal Description Language genoemd) gegeven.

**N-grams**
N-grams is een populair statistical language model dat veel gebruikt wordt in het vakgebied spraakherkenning. In een N-gram hebben alleen de vorige N – 1 woorden invloed op de kans op een bepaald huidig woord. Normaalgesproken worden N-grams toegepast op lineaire structuren zoals teksten. In [55] beschrijven Brugger et al. een methodiek die gebruikt maakt van Generalized N-grams. Generalized N-grams bevatten een statistische representatie van aanwezige patronen in een documentklasse. In deze methodiek worden N-grams toegepast op tree-structuren. Dat wordt gedaan door niet alleen de normale 'horizontale' lineaire N-grams te gebruiken om naburige tree nodes aan te geven, maar ook 'verticale' n-grams. Het voordeel van de N-gram methodiek is dat het toegepast kan worden op incomplete logische boomstructuren. N-grams kunnen namelijk semantische objecten voorspellen op basis van de context van het object. De voorspellingen worden gedaan op basis van de kansen van het corresponderende tree N-gram.
De tree-structuur van de logische componenten van het document worden gerepresenteerd door de kansen op lokale tree-node-patronen die lijken op N-grams. De semantische tree-structuur wordt opgebouwd vanuit de geometrische objecten in overeenkomst met het gegeven model: door op een bottom-up manier tree nodes toe te voegen uitgaande van de in het model vastgelegde N-grams. Het geeft dus voor een gedeeltelijke semantisch structuur mogelijkheden die elk een bepaalde kans hebben. Er kunnen meerdere mogelijke semantische bomen zijn, maar alleen de boom met de hoogste kans wordt geselecteerd. Het documentmodel wordt geleerd uitgaande van voorbeelden van logische bomen die bijvoorbeeld door de gebruiker gelabelde documenten voorstellen. Bij elke nieuwe trainingboom worden de kansen aangepast door middel van een incremental learning-algoritme.
In het genoemde artikel wordt een trigram methode geïmplementeerd en wordt het model gebruikt om automatisch DTD-regels te genereren in XML.
In het artikel worden helaas geen resultaten besproken.

**Hidden Markov Models (HMM)**
Het Hidden Markov Model wordt gebruikt voor het modeleren van kansverdelingen die gedefinieerd zijn op een set van strings. Het model wordt, net als het N-grams-model, met name toegepast binnen het vakgebied spraakherkenning. Het model wordt algemeen gebruikt voor sequence learning.
In [56] wordt een extensie op het HMM-model gedefinieerd: het Hidden Tree Markov Model. Dit model wordt gedefinieerd voor het modeleren van kansverdelingen over een ruimte van trees. Het model wordt getraind door middel van een Expectation-Maximization-algoritme dat gebruik maakt van gelabelde X-Y Trees. De methodiek is getest op een dataset van 89 facturen die door 9 verschillende bedrijven zijn opgemaakt. Er werd een performance van 99.28 % behaald.

In [57] wordt de hiërarchische layout van een document als een set van attributed Hidden semi Markov Models gedefinieerd. Elk attributed HSMM representeert het projectieprofiel van de bounding boxes in een geometrische documentobject op de X- of de Y-as. De methodiek leert documentmodels door middel van trainingdocumenten waarin de geometrische documentobjecten gelabeld zijn. Een HSMM maakt gebruik van de finite state automaton als onderliggende taal. Het trainingsprobleem wordt gesteld als een Search or Optimization probleem in een Bayesian framework. De projectieprofielen worden omgezet naar input strings voor het Bayesian Learning-algoritme. Het Bayesian Learning-algoritme kan echter alleen ééndimensionale input strings gebruiken om de attributed HSMM te leren. Het Bayesian Learning-algoritme wordt in deze methodiek toegepast op tweedimensionale modellen door het algoritme op een recursieve manier toe te passen op projectieprofiel in de X-richting en het projectieprofiel in de Y-richting.

**Stochastic language model**

Een stochastisch language model verschilt van een deterministisch language model (een document grammar) doordat het stochastisch model een kansverdeling definieert voor elke productie. Daardoor heeft het stochastische model twee belangrijke voordelen ten opzichte van het deterministische model:

- Een deterministisch language model kan resulteren in ambigue parsing-resultaten wanneer de invoer probabilistisch is of wanneer de grammar ambigu is;
- Een deterministisch language model heeft geen mogelijkheid om de relatieve belangrijkheid van rules te leren uit een training dataset.

Voor een stochastisch language model wordt het parsing-resultaat met de grootste kans gekozen als beste resultaat.

Voorbeelden van deze methodiek worden besproken in [58] en [59]. In [59] worden de productieregels gerepresenteerd door een een weighted finite state automaton. Elke state representeert een symbool in de productieregels. Eigenschappen van de geometrische documentobjecten worden omgezet in state duration densities. Het Viterbi-algoritme wordt gebruikt voor het zoeken naar het beste resultaat.

### 4.2.4 Machine Learning

Deze methodieken gebruiken Machine Learning-technieken om document models te leren. Het model van een bepaalde documentklasse wordt getraind met een set van gelabelde documenten die tot de documentklasse behoren. Elk document model representeert een rule base waarmee documenten uit de bijbehorende documentklasse gelabeld kunnen worden. Het grote verschil tussen Machine Learning-methodieken en niet-Machine Learning-methodieken is dat de rule bases verkregen worden door het trainen met gelabelde voorbeelddocumenten en niet handmatig samengesteld worden.

Esposito et al. geven in [60] een voorbeeld van een methodiek die gebruik maakt van Machine Learning. De methodiek bepaalt eerst tot welke documentklasse het document behoort door de geometrische structuur te matchen met modellen van documentklasses. De documentklasses worden dus alleen van elkaar onderscheiden door de geometrische structuur. De document models worden beschreven in een eerste orde logische taal. Daarin worden de eigenschappen van documentobjecten en de ruimtelijke relaties tussen de objecten gebruikt. Het learning algoritme is een separate-and-conquer algoritme dat zowel gebruik maakt van positieve als van negatieve voorbeelddocumenten. Daarbij worden de positieve voorbeelden gebruikt om te leren welke eigenschappen horen bij een bepaald documentmodel en de negatieve voorbeelden om te leren welke eigenschappen juist niet bij een bepaald documentmodel horen. Document Understanding wordt uitgevoerd door de geometrische structuur van het document te matchen tegen models van

semantische categorieën. Daarbij wordt elke semantische categorie beschreven door een verzameling van rules. Er wordt aangenomen dat documenten uit dezelfde documentklasse een verzameling invariante geometrische eigenschappen hebben en dat de semantische categorieën geïdentificeerd kunnen worden door alleen gebruik te maken van geometrische eigenschappen. De rules voor de semantische categorieën worden getraind door middel van een incremental learning-algoritme dat hiërarchische concepten afleidt uit positieve en negatieve voorbeelden. De performance voor het herkennen van de labels loopt uiteen van 92% tot 100%.

**Artificial Neural Networks**

Een veelgebruikte Machine Learning-techniek is het Artificial Neural Network (ANN). Een nadeel van een ANN is dat er training nodig is als voorbereidende stap. Het algoritme moet opnieuw getraind worden als er een nieuw documentklasse verwerkt moet gaan worden. Dit is én een tijdrovend karwei (het kost veel tijd de learning parameters goed in te stellen) én maakt het lastig snel te handelen bij uitzonderlijke situaties die zich opeens voordoen. Deze techniek heeft ook moeite met het gebruiken van de geometrische structuur. Een ander nadeel is dat het moeilijk is verkeerde labelresultaten te analyseren.
Het voordeel van een ANN is dat er niet handmatig een documentmodel of een rule base samengesteld hoeft te worden.

In [61] wordt een methodiek beschreven die gebruik maakt van een *back-propagation neural network* voor het leren van een rule base. Het algoritme maakt gebruik van domein-specifieke informatie (dat is mogelijk doordat eerst het journal waaruit het document afkomstig is bepaald wordt), geometrische en niet-geometrische eigenschappen. Voor elke van de journal-types die herkend moet kunnen worden wordt een afzonderlijk neural network getraind. Er wordt getraind door middel van de *cross-validation* techniek. Het back-propagation neural network wordt geïmplementeerd met een input layer van 16 nodes (eigenschappen van de geometrische documentobjecten), een hidden layer van 8 nodes en een output layer van 5 nodes (de mogelijke labels). In ongeveer 97% van de testdata heeft het neurale netwerk het juiste label bepaald.

In [62] wordt een methodiek beschreven die gebruikt maakt van een domeinspecifiek model en van een universeel model. Het domeinspecifieke model wordt getraind met het doel één bepaalde documentklasse te kunnen labelen. Voor elke documentklasse die herkend moet kunnen worden moet er dus een domeinspecifiek model worden getraind. Het universele model wordt getraind met documentvoorbeelden uit alle documentklasses zodat er een algemeen (universeel) model ontstaat. Het doel van het universele model is een nieuw domeinspecifiek model (als er een documentklasse wordt toegevoegd) te kunnen voorzien van universele kennis zodat die kennis als uitgangspunt gebruikt kan worden bij het trainen van het nieuwe domeinspecifieke model. Alle gebruikte neural networks in deze methodiek zijn multilayer perceptrons die op een feed-forward manier met elkaar zijn verbonden. De methodiek werd getest op 52 krantenpagina's uit 3 verschillende Arabische kranten. De testresultaten voor het universele model liepen uiteen van 91% tot 100% en de resultaten van het domeinspecifieke model liepen uiteen van 95% tot 100%. In beide modellen werd ook getracht de zwarte omkadering van een artikel te detecteren. Dat leverde een performance van respectievelijk 53% en 70%. Deze resultaten liggen behoorlijk onder de performance die behaald werd bij het labelen van de inhoud van de krantenartikelen. De auteurs geven daarvoor als oorzaak dat de zwarte omkadering in de arabische kranten vaak een texture bevat (het is dus niet altijd een simpele, rechte lijn) die herkend wordt als afbeelding.

## 4.3 Text Mining

Text Mining wordt gebruikt om patronen, modellen of rules af te leiden uit ongestructureerde tekst. Eén van de belangrijkste doelen van text mining is het afleiden van patronen in de tekst die dan vervolgens gebruikt kunnen worden in andere methodieken, zoals Text Categorization en Text Clustering.

Text Categorization (ook wel Text Classification of Topic Detection genoemd) is het proces van het toekennen van vooraf gedefinieerde categorieën aan tekstdocumenten. Daarbij wordt gebruik gemaakt van een classification function die getraind wordt door een verzameling gelabelde documenten.
Text Clustering is het proces dat een verzameling tekstdocumenten opdeelt in een aantal groepen. Een clustering is optimaal als de documenten in een groep (cluster) zo gelijk mogelijk zijn en de documenten uit twee verschillende groepen zo verschillend mogelijk zijn. Als tekstdocumenten geclusterd worden op basis van hun tekstuele inhoud, dan corresponderende de gevonden groepen met onderwerpen of thema's.
Alleen de Text Categorization-methodiek is te vertalen naar een Document Understanding-methodiek: Text Categorization moet dan geen tekstdocumenten als input gebruiken, maar afzonderlijke alinea's tekst die afkomstig zijn uit het tekstdocument. Het vinden van een categorie voor elke alinea is gelijk aan het vinden van een label voor elke alinea.
Text Clustering is slecht te gebruiken als Document Understanding-methodiek omdat bij het DU-proces de clusters al vooraf bekend staan. De clusters zijn namelijk de verschillende semantische categorieën (labels) die gedefinieerd zijn. Het is dus onnodig een algoritme te laten zoeken naar een optimale clustering.

Eén van de belangrijke componenten van een text mining-methodiek is Information Extraction (IE). Het doel van Information Extraction is het vinden van specifieke data in natuurlijke tekst. De data die gevonden moet worden wordt in de meeste gevallen weergegeven door een template met daarin een lijst van lege velden die gevuld moeten worden met woorden uit de tekst. Voorbeelden van toepassingen zijn het invullen van een medisch patiëntdossier op basis van ongestructureerde tekst en het destilleren van bepaalde gegevens uit huizenadvertenties, lesroosters en aankondigingen van presentaties.
Een veel gebruikte implementatie van een IE-systeem is RAPIER (Robust Automated Production of Information Extraction Rules) [63]. RAPIER is een relational rule based learner bedoeld om IE-rules af te leiden uit een verzameling trainingsvoorbeelden. De rules in RAPIER maken gebruik van syntactische en semantische informatie uit de tekst. RAPIER leert de IE-rules op een bottom-up manier en de tekstpatronen door een specific-to-general search. Het resultaat van het trainen is een rule base die de rules bevat waarmee de velden in de template gevuld kunnen worden. Als testset werd gebruik gemaakt van 300 ICT-gerelateerde vacatures die op de `austin.jobs` nieuwsgroep waren geplaatst. Al bij 10 trainingsvoorbeelden is de precision 80% en de recall 40%. Bij 250 trainingsvoorbeelden is de precision toegenomen tot 89% en de recall tot 62%.
Voor een IE-systeem geldt over het algemeen dat de precision behoorlijk hoger is dan de recall. Met andere woorden, de kans dat een gevuld veld goed gevuld is is groter dan de kans dat een veld goed gevuld is.

**Text Categorization**

Een classifier bij text categorization definieert voor elke input (een vector van woorden) een functie die de kans geeft dat die input tot een bepaalde klasse behoort. Elk document wordt gerepresenteerd door een vector van de woorden waaruit het document bestaat.

De eenvoudigste aanpak is de k-Nearest Neighbor classifier. Die classifier zoekt de *k* trainingsdocumenten die het dichtst bij het testdocument liggen en kent de categorieën van die trainingsdocumenten toe aan het testdocument.
Een nauwkeurige learning-methode voor classifiers is de Support Vector Machines-methode (SVM) [65]. Als input voor de learning-methode dienen positieve en negatieve trainingsvoorbeelden in de vorm van vectors van woorden. De SVM berekent, op basis van het Structural Risk Minimization principe, een vlak in de vectorruimte die de set van positieve voorbeelden met een zo groot mogelijke marge scheidt van de negatieve voorbeelden. Het vinden van de maximale marge kan gezien worden als een optimization problem. De SVM kan getraind worden door elk willekeurig Quadratic Programming optimalisatie algoritme. Als een maat voor de performance wordt gebruik gemaakt van het Precision/Recall Breakeven Point. Als test corpus wordt gebruik gemaakt van de Reuters-21578 dataset. Voor de tien meest voorkomende categorieën in de dataset heeft de SVM een gemiddeld breakeven point van 86.0.

Andere learning-methoden voor classifiers zijn: decision trees [66], Naïve Bayes classifier [67] en Bayes networks [68].
Een decision tree classifier bestaat uit een tree waarbij de leaves de categorieën representeren, de interne tree nodes bepaalde eigenschappen van de input vertegenwoordigen en de branches een test op een mogelijke waarde van de bijbehorende eigenschap bevatten. Een tekst wordt gelabeld door de boom te doorlopen en bij elke node de waarde die de eigenschap in de tekst heeft de bijbehorende branche te volgen. Een voorbeeld van een decision tree text classifier wordt gevonden in [69].

# 5 De labelingmethodiek

Hoofdstuk 5 beschrijft het ontwerp van de labelingmethodiek die de personeelsadvertenties automatisch moet gaan labelen voor MatchCare. De labelingmethodiek is voortgekomen uit de resultaten van het afstudeeronderzoek die in de vorige Hoofdstukken zijn besproken.

Sectie 5.1 bevat de functionele beschrijving van de labelingmethodiek, sectie 5.2 bevat een korte bespreking van de afwegingen die zijn gemaakt bij het selecteren van het soort methodiek, sectie 5.3 bevat een overzicht van de soorten eigenschappen die gebruikt worden door de methodiek, sectie 5.4 bevat een overzicht van de soorten rules die gebruikt worden door de methodiek en sectie 5.5 tenslotte bevat een weergave van de labelingmethodiek in pseudo-code.

## 5.1 Functionele beschrijving

De functionele beschrijving van de labelingmethodiek bevat een beschrijving van de functionele eisen die er aan de methodiek gesteld worden. Die eisen zijn bepaald door het huidige vacatureproces bij MatchCare. Het is namelijk één van de belangrijkste eisen van het bestuur van MatchCare dat in het bestaande vacatureproces alleen de handmatige labeling van personeelsadvertenties wordt vervangen door een geautomatiseerde labelmodule en dat de aanpassing verder geen effect heeft op de andere deelprocessen van het vacatureproces. Dat betekent dat de input en de output van de labelmodule vastliggen.

Het is nog wel een openstaande vraag of de labelmodule ook verantwoordelijk wordt voor het segmenteren van de personeelsadvertentie of dat de labelmodule alleen verantwoordelijk is voor het toekennen van labels aan de handmatige gedefinieerde tekstsegmenten. Deze labelmethodiek gaat ervanuit dat ook het handmatig definiëren van de tekstsegmenten geautomatiseerd zal worden. Dat is namelijk een voor de hand liggende stap in de automatiseringsgedachte van MatchCare. Dit afstudeeronderzoek zal echter niet de totstandkoming van een geautomatiseerde segmentatiemethodiek bespreken, simpelweg omdat dat buiten de scope van dit afstudeeronderzoek valt.

De input van de labelmodule is een X-Y Tree (zie paragraaf 3.3.3). De X-Y Tree geeft een hiërarchische representatie van een gescande personeelsadvertentie. Elke tree leaf representeert één tekstsegment en heeft een frame waarin alle relevante eigenschappen van het tekstsegment zijn opgenomen. Een tekstsegment bestaat uit een blok tekst dat omgeven is door wit-ruimte. Een tekstsegment kan dus bestaan uit meerdere alinea's tekst als de alinea's niet gescheiden door een wit-ruimte. De relevante eigenschappen van de tekstsegmenten worden genoemd in sectie 5.3. Naast deze eigenschappen bevat het frame ook een tekstpointer naar een tekstbestand waarin het OCR-resultaat van het tekstsegment is opgenomen. De grafische afbeeldingen worden niet als segmenten opgenomen in de X-Y Tree.
Er is gekozen voor de X-Y Tree als datamodel omdat de X-Y Tree simpel en efficiënt te implementeren is en optimaal gebruik maakt van het feit dat de personeelsadvertenties bestaan uit rechthoekige segmenten die van elkaar gescheiden worden door wit-ruimte (de Manhattan-layout).

De output van de labelmodule is een gelabelde X-Y Tree. Deze tree bevat dezelfde hiërarchische structuur als de input tree maar elke tree node bevat tevens één van de 6 mogelijke label (of het lege label als er geen label met een voldoende hoge zekerheid gevonden werd). Als de labels bepaald zijn zorgt een extra subproces ervoor dat er in de Centrale Vacaturedatabase een nieuw vacaturerecord wordt aangemaakt. De gelabelde X-Y Tree wordt vervolgens doorgelopen en daarbij wordt de tekst van elk tekstsegment (te vinden met behulp van de tekstpointer) op de juiste positie in het vacaturerecord geplaatst.

## 5.2 Afwegingen bij de methodiekkeuze

In Hoofdstuk 4 is getoond dat er allerlei concepten beschikbaar zijn waarvan een Document Understanding-methodiek gebruik kan maken. Dat leidt ertoe dat er bij het ontwerpen van een Document Understanding-methodiek een aantal afwegingen gemaakt moeten worden. In deze sectie onderbouw ik kort mijn keuze.

De eerste afweging betreft het wel of niet gebruik maken van een document model: het gebruik van een document model is niet goed mogelijk in het domein van de personeelsadvertenties. Dat komt doordat er moeilijk documentklasses te onderscheiden zijn. Het is niet dus niet goed mogelijk om documentklasses te definiëren die een invariante verzameling geometrische eigenschappen hebben en waarbij die verzameling gebruikt kan worden om de betreffende documentklasse te onderscheiden van de andere documentklasses. Dit komt doordat personeelsadvertenties geen voorspelbare geometrische structuur hebben (in tegenstelling tot bijvoorbeeld een telefoonboek) die omgezet kan worden in een document model (bijvoorbeeld in een document grammar). Personeelsadvertenties hebben een duidelijke waarneembare creatieve component in zich die zich uiten door een grote variatie in de geometrische structuur.
Het is waarschijnlijk wél mogelijk een document model te definiëren op basis van de adverteerder in de personeelsadvertentie. Bij mij bestaat namelijk het vermoeden dat elke adverteerder een vaste volgorde van informatiecategorieën hanteert en tevens voor elke informatiecategorie een redelijk constant aantal tekstsegmenten gebruikt. Dat zou betekenen dat er voor elke adverteerder een document model opgesteld kan worden. De Document Understanding-methodiek begint dan met het vaststellen van de adverteerder. Dat kan gedaan worden met behulp van Information Extraction-rules die de adverteerder destilleren uit de adverteerdersgegevens in de personeelsadvertentie). Vervolgens wordt uit een knowledge base het document model voor de adverteerder gehaald. De rule base kan vervolgens gebruik maken van de invariante eigenschappen. Dit vermoeden moet echter nog getoetst worden aan de praktijk. Een aanwijzing

dat dit vermoeden geldt wordt gevormd door de personeelsadvertentie in Figuur 2.1. Die personeelsadvertentie laat zien dat de adverteerder TU/e gebruikt maakt van een vaste volgorde van de informatiecategorieën en dat er daardoor een document model kan worden samengesteld voor deze adverteerder.

De tweede afweging betreft het wel of niet gebruiken van een neural network om de rule base samen te stellen: er wordt geen gebruik gemaakt van een neural network, ondere andere omdat er geen document models gebruikt gaan worden en een neural network juist uitermate geschikt is voor het leren van document models. Een tweede reden voor het niet gebruiken is dat er voldoende domeinkennis aanwezig is (zie Hoofdstuk 2) om handmatig een rule base samen te stellen. Het is dus niet nodig om met behulp van een neural network de kennis die er al is te vinden. Twee grote bijkomende nadelen van neural networks zijn dat het zeer lastig is verkeerde labelingen van segmenten te analyseren en dat het moeilijk is gebruik te maken van geometrische eigenschappen, de eigenschappen die bij het labelen van personeelsadvertenties heel essentieel zijn.

De derde afweging betreft het wel of niet gebruik maken van een Text Mining-methodiek: er wordt geen gebruik gemaakt van een Text Mining-techniek omdat er slechts één hele simpele tekstuele eigenschap wordt gebruikt in de labelmethodiek. De enige tekstuele eigenschap die wordt gebruikt is de set van 6 waarden die aangeven hoeveel keywords uit de keywordlist voor een bepaald label zijn aangetroffen in de tekst van de tekstsegment.
Information Extraction kan in de toekomst wél gebruikt worden bij het verder opdelen van de tekst, bijvoorbeeld bij het automatisch herkennen van de adresgegevens.

## 5.3 Soorten eigenschappen

De eigenschappen die nodig zijn voor het geautomatiseerd labelen van een personeelsadvertentie kunnen rechtstreeks worden afgeleid uit de wetmatigheden en deelhypotheses die in Hoofdstuk 2 besproken zijn.

De labelmethodiek gebruikt drie soorten eigenschappen. De paragrafen 5.3.1 t/m 5.3.3 bespreken achtereenvolgens de geometrische, tekstuele en tekstopmaakgerichte eigenschappen.

### 5.3.1 Geometrische eigenschappen

Geometrische eigenschappen van de tekstsegmenten kunnen worden onderverdeeld in twee categorieën:

- **Aan zichzelf gerelateerde eigenschappen**: deze eigenschappen hebben alleen betrekking op het betreffende tekstsegment en gebruiken geen informatie over de overige tekstsegmenten. De eigenschappen die in deze categorie vallen zijn:
    - De absolute coördinaten van het tekstsegment. Deze eigenschap wordt gebruikt voor de toepassing van de deelhypothese *layout2* (zie pagina 21);
    - Het volgordenummer van het tekstsegment in de leesvolgorde. Dit is een eigenschap die wordt afgeleid uit de absolute coördinaten van het tekstsegment. Eerst wordt de leesvolgorde van de tekstsegmenten bepaald (een voorbeeld van een algoritme daarvoor

wordt gegeven in [70]) en vervolgens kan voor elk segment het volgordenummer in de leesvolgorde bepaald worden. Deze eigenschap wordt gebruikt voor de toepassing van de deelhypothese *tekstueel3* (zie pagina 21).

- **Aan andere tekstsegmenten gerelateerde eigenschappen**: deze eigenschappen leggen de onderlinge relaties tussen de tekstsegmenten vast. De eigenschappen die in deze categorie vallen zijn:
  - De relatieve posities van de tekstsegmenten ten opzichte van elkaar. De relatieve posities worden geformuleerd met behulp van Qualitative Spatial Relations [6]. Deze eigenschappen worden gebruikt voor de toepassing van de deelhypothese *layout1* (zie pagina 21).

## 5.3.2 Tekstuele eigenschappen

Tekstuele eigenschappen maken gebruik van de tekst in een tekstsegment voor het vastleggen van de eigenschap. Er wordt in de labelmethodiek maar één soort tekstuele eigenschap gebruikt, namelijk de eigenschap of er keywords voorkomen in het tekstsegment. Die eigenschap wordt geïmplementeerd door een set eigenschappen. Voor elke label *x* wordt de eigenschap als volgt bepaald: het aantal keywords uit de keywordlijst voor label *x* dat voorkomt in het tekstsegment. Omdat er 6 mogelijke labels zijn bevat de set 6 eigenschappen. Deze set eigenschappen wordt gebruikt voor de toepassing van de deelhypothese *tekstueel2* (zie pagina 21).

## 5.3.3 Tekstopmaakgerichte eigenschappen

Tekstopmaakeigenschappen maken gebruik van de tekstopmaak van de tekst die voorkomt in een tekstsegment. De volgende tekstopmaakgerelateerde eigenschappen worden gedefinieerd:

- Een verzameling van alle lettertypes die voorkomen in het tekstsegment. De lettertypes worden bepaald door een OCR-component;
- Een verzameling van regelnummers die alle regelnummers bevat waarin *vetgedrukte* tekst voorkomt. Als er geen regels met vetgedrukte tekst voorkomen, dan blijft de verzameling leeg. Er worden soortgelijke eigenschappen gedefinieerd voor de tekstopmaak *schuingedrukt* en *onderstreept*.

Deze eigenschappen worden gebruikt voor de toepassing van de deelhypotheses *tekstopmaak1*, *tekstopmaak2* en *tekstopmaak3* (zie pagina 21).

De deelhypothese *layout3* wordt in de labelmethodiek vertegenwoordigd door een nieuw opgezette eigenschap: het *aandachtsprofiel*. Het aandachtsprofiel is een weergave van de overgangen (wit-regels) tussen de segmenten waaruit de 'sterkte' van de overgangen af te lezen valt. De sterkte komt overeen met de kans dat de overgang de grens tussen twee categorieën vormt. De labelmethodiek maakt gebruik de sterkte van de overgangen omdat voor het voor het correct labelen van een personeelsadvertentie essentieel is te weten waar de grenzen tussen de informatiecategorieën zich bevinden. Het komt veelvuldig voor dat een informatiecategorie meer dan één tekstsegment omvat (zie tabel 6, pagina 35) en als niet in elk van die segmenten een keyword gevonden wordt, dan is niet duidelijk na welk segment de categorie ophoudt en de volgende categorie begint. Dat wordt wel duidelijk als er tussen twee van de 'onduidelijke' segmenten een overgang aanwezig is die met grote zekerheid aangeeft dat er een grens tussen de twee segmenten aanwezig is. Omdat er in de literatuur geen vergelijkbare eigenschap gevonden werd is overgegaan tot het zelf opzetten van deze nieuwe eigenschap.
In sectie 5.5 wordt uitvoerig beschreven hoe het aandachtsprofiel samengesteld wordt.

## 5.4 Soorten rules

Er worden twee soorten rules gebruikt in de labelmethodiek:

- **Knowledge rules**: deze rules bevatten alle kennis die er over het domein van de personeelsadvertenties bestaat (samengevat in de deelhypotheses). De knowledge rules maken daartoe gebruik van de eigenschappen die in sectie 5.3 gedefinieerd zijn.
- **Control rules**: deze rules zijn bedoeld voor het sturen van het labelproces. De controlrules bepalen de volgorde waarin de tekstsegmenten gelabeld moeten worden en bepalen wanneer het labelproces klaar is.

  Control rules sturen de volgorde waarin de tekstsegmenten gelabeld moeten worden omdat de volgorde niet recht-toe-recht-aan is. Het kan namelijk voorkomen dat het voor het labelen van een bepaald tekstsegment nodig is eerst de labels van de naburige tekstsegmenten te bepalen. De control rules bepalen in dat geval de benodigde labelvolgorde.

  Control rules zijn ook nodig als de eerste labelslag (waarin alle tekstsegmenten gelabeld zijn) nog niet een labeling met een voldoende hoge confidence rate heeft opgeleverd. Een tweede labelslag kan in dat geval helpen doordat er in de tweede labelslag voor de labeling van een bepaald tekstsegment misschien gebruik kan worden gemaakt van kennis die er in de eerste labelslag nog niet was, zoals bijvoorbeeld de labels (inclusief confidence rates) van de andere tekstsegmenten. De control rules bepalen in dat geval welke tekstsegmenten opnieuw gelabeld moeten worden.

  De control rules bepalen tevens de stop criteria voor de labelmethodiek. De control rules maken daarvoor gebruik van confidence rates. Per tekstsegment wordt een verzameling van confidence rates bijgehouden. De verzameling bevat per label één confidence rate die aangeeft met welke zekerheid (letterlijk: vertrouwen) het tekstsegment dat label heeft. Het huidige label van een tekstsegment is het label dat de hoogste confidence rate heeft. Eén van de twee stop criteria is dat de huidige labels van alle tekstsegmenten een voldoende hoge confidence rate hebben. Dit wordt een succesvolle labeling genoemd. Het tweede stop criterium is dat alle tekstsegmenten in de laatste labelslag hetzelfde huidige label hebben gehouden. In dat geval kan er geen goede labeling gemaakt worden en is er sprake van een niet-succesvolle labeling.

  Door middel van controlrules is het ook mogelijk verschillende labelstrategieën te implementeren. Een voorbeeld van een simpele labelstrategie is alle tekstsegmenten te labelen met behulp van alle rules. Een voorbeeld van een iets meer geavanceerde labelstrategie is proberen eerst alle tekstsegmenten te labelen met de meest effectieve rules (bijvoorbeeld de rules die zoeken naar keywords in de tekstsegmenten) en pas als dat geen succesvolle labeling geeft de overige rules in te schakelen.

## 5.5 Labelmethodiek in pseudo-code

In de labelmethodiek wordt gebruik gemaakt van een nieuwe representatie van de tekstsegmenten: het zogenaamde *aandachtsprofiel*. Het aandachtsprofiel is een profiel dat aangeeft welke onderdelen van de personeelsadvertentie bij het lezen de aandacht van de lezer trekken. Het gaat uit van de basisgedachte dat grote lettertypes, kolomovergangen, kopjes en wit-regels de aandacht trekken van de lezer bij het bekijken

van de personeelsadvertentie. Het aandachtsprofiel vormt dus een maat voor de opvallendheid van de onderdelen waaruit de personeelsadvertentie bestaat. Het aandachtsprofiel wordt in de rule base gebruikt om de grenzen van de informatiecategorie zo zeker mogelijk vast te stellen. Daarbij geldt de waarneming dat de opvallendheid van een overgang tussen twee tekstsegmenten rechtevenredig is met de zekerheid dat de overgang een overgang tussen twee informatiecategorieën representeert.

De verschillende fasen bij het samenstellen van het aandachtsprofiel worden weergegeven in Figuur 5.1 (a) t/m (f).

Als basis voor het aandachtsprofiel worden de bounding boxes van de tekstsegmenten gebruikt (zie Figuur 5.1 (b)). Om het aandachtsprofiel van de gescande personeelsadvertentie te kunnen berekenen wordt er eerst een afgeleide geometrische structuur bepaald door de tekstsegmenten in een sequentiële verticale volgorde te plaatsen (zie Figuur 5.1 (c)). Dat wordt gedaan op basis van de leesvolgorde. Vervolgens wordt van de afgeleide geometrische structuur een horizontaal breedteprofiel gemaakt door de structuur te projecteren op een verticale lijn (zie Figuur 5.1 (d)). Het aandachtsprofiel wordt uit de afgeleide structuur en het horizontale breedteprofiel afgeleid door middel van een drietal stappen:

- Bereken het aantal pixels dat tussen elk paar bounding boxes van tekstsegmenten aanwezig is;
- Bereken het 'inverse' horizontale breedteprofiel (zie Figuur 5.1 (e)) door alle waarden gelijk aan of boven een gegeven threshold op 0 te zetten en alle waarden beneden de threshold om te zetten naar een waarde die evenredig is met het aantal pixels tussen de twee aangrenzende bounding boxes. Hierdoor worden alle tekstsegmenten op 0 gezet en worden de overgangen tussen de tekstsegmenten omgezet naar een waarde die aangeeft hoe groot de ruimte is tussen de aangrenzende tekstsegmenten is;
- Zet de aanwezigheid van grote lettertypes, kolomovergangen, kopjes en wit-regels door middel van een aandachts-rule base om naar een verhoogde aandachtswaarde. De verhoging moet op een zodanige manier gebeuren dat er een rangorde in de overgangen tussen de tekstsegmenten ontstaat: de overgangen die grenzen aan het tekstsegment met de grootste gemiddelde lettergrootte (bijna altijd de functietitel) krijgen de grootste vermenigvuldigingsfactor, gevolgd door kolomovergangen, de wit-regels en als laatste de alinea-overgangen (zonder wit-regel). Een alinea-overgang kan soms moeilijk te detecteren zijn. Dat is het geval als de laatste tekstregel van een alinea niet een duidelijk kleinere breedte heeft dan de andere tekstregels. De aandachtswaarde van een overgang wordt ook nog eens vermenigvuldigd met een factor als het tekstsegment dat na de overgang staat begint met een kopje.

Figuur 5.1 (f) geeft het uiteindelijke aandachtsprofiel. In het aandachtsprofiel worden dus zowel geometrische als tekstopmaakgerichte eigenschappen gebruikt.

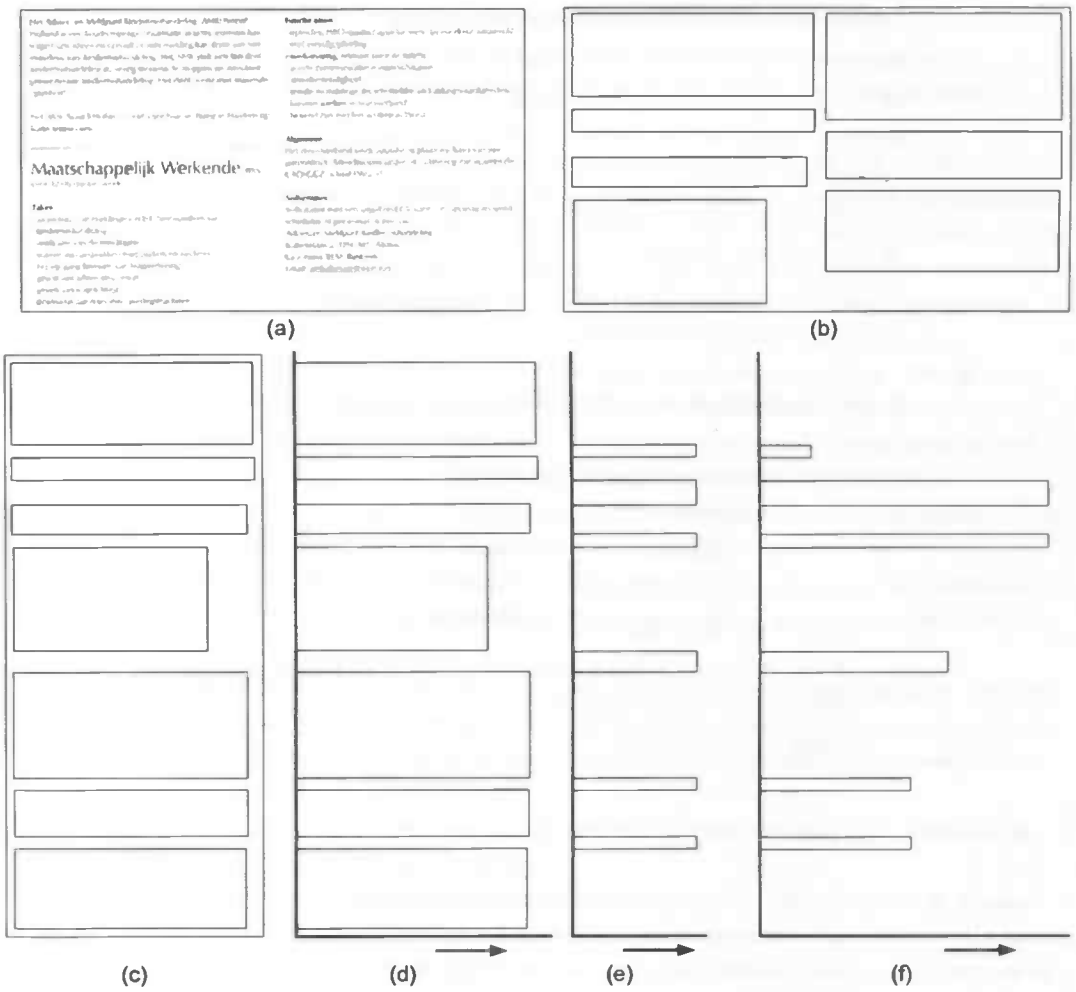
Het aandachtsprofiel dat op deze manier wordt verkregen geeft een structuur waarin de opvallende onderdelen van een personeelsadvertentie duidelijk naar voren komen. De aandachtswaarden geven aan hoe opvallend de overgangen tussen tekstsegmenten zijn. Dat kan in de rule base gebruikt worden om te bepalen waar de grenzen van een informatiecategorie liggen. Het komt namelijk regelmatig voor dat een informatiecategorie verdeeld is over meerdere tekstsegmenten en er dus bepaald moet worden welke tekstsegmenten tot dezelfde informatiecategorie behoren. Hoe opvallender een overgang tussen twee tekstsegmenten, hoe groter de kans dat de overgang ook een overgang tussen twee informatiecategorieën representeert.

De pseudo-code voor de constructie van het aandachtsprofiel kan als volgt worden weergegeven:

```
bepaal leesvolgorde
bepaal sequentiële verticale structuur
bereken horizontaal breedteprofiel van bounding boxes van segmenten
bereken inverse breedteprofiel
bereken gemiddelde lettergrootte per segment
detecteer kopjes
detecteer kolomovergangen
detecteer wit-regels
detecteer alineaovergangen
bereken aandachtsprofiel
```

De pseudo-code voor de labelmethodiek kan als volgt worden weergegeven:

```
bepaal eigenschappen van segmenten
bepaal aandachtsprofiel
zolang niet (labelstatus = klaar) {
    bepaal labelstrategie
    bepaal volgendsegment
    label volgendsegment volgens labelstrategie
    update confidencerates
    bepaal labelstatus
}
```




**Figuur 5.1 – De constructie van het aandachtsprofiel. De verschillende stadia in dit proces zijn (a) de gescande personeelsadvertentie, (b) de bounding boxes van de tekstsegmenten, (c) de sequentiële verticale structuur, (d) het horizontale breedteprofiel, (e) het inverse horizontale breedteprofiel en (f) het aandachtsprofiel.**

# 6 Conclusies

Het afstudeeronderzoek had als belangrijkste doelstelling het ontwikkelen van een geautomatiseerde labelmethodiek die het handmatige labelproces in het vacatureverwerkingsproces van MatchCare, een productiebedrijf op het gebied van vacature-informatie, zou kunnen vervangen.
De belangrijkste redenen voor MatchCare om dit onderzoek te starten waren de noodzaak tot het terugdringen van de benodigde tijd voor het verwerken van een vacature, het daardoor doen toenemen van de productiecapaciteit en het werken op onmenselijke uren mogelijk te maken.

Door middel van een statistisch onderzoek (Hoofdstuk 2) is aangetoond dat er in het domein van de personeelsadvertenties wetmatigheden aanwezig zijn die gebruikt kunnen worden bij het geautomatiseerd labelen van personeelsadvertenties. Er is aangetoond dat deze wetmatigheden voldoende krachtig zijn om gebruikt te kunnen worden voor het samenstellen van een rule base voor het labelen.

Een verkenning van de onderzoeksterreinen Document Layout Analysis (Hoofdstuk 3) en Document Understanding (Hoofdstuk 4) heeft laten zien dat er een grote verscheidenheid aan segmentatie- en labelmethodieken bestaat. Vervolgens hebben een aantal afwegingen geleid tot een ontwerp voor een labelmethodiek die specifiek toegepast kan worden in het domein van de personeelsadvertenties (Hoofdstuk 5).

Het gepresenteerde ontwerp voor de labelmethodiek bevat een nieuw concept, het *aandachtsprofiel*, dat dient als een nieuwe soort eigenschap van personeelsadvertenties die aangeeft welke onderdelen van de personeelsadvertentie de meeste aandacht van de lezer trekken. Het aandachtsprofiel maakt gebruik van de waarnemingen dat de functietitel het opvallendste onderdeel van de personeelsadvertentie is en dat de opvallendheid van een overgang tussen twee tekstsegmenten rechtevenredig is met de zekerheid dat het een overgang tussen twee informatiecategorieën representeert.

Hoofdstuk 5 laat daarmee zien dat de onderzoekshypothese,

*Het is mogelijk een geautomatiseerde methodiek te ontwikkelen die,*
*gebruikmakend van tekstuele, layoutgerichte en tekstopmaakgerichte wetmatigheden,*
*in staat is de meerderheid van alle personeelsadvertenties uit de gedrukte media te labelen,*

in ieder geval op basis van de theorie correct is. Helaas kan de labelmethodiek niet in de praktijk geïmplementeerd en getest worden door het faillissement van MatchCare. De vraag of het geautomatiseerde labelproces sneller was geweest dan het handmatige labelproces blijft daarmee voor altijd onbeantwoord. Op basis van de resultaten van de in Hoofdstuk 4 besproken geïmplementeerde labelmethodieken kan met een behoorlijke zekerheid voorspeld worden dat het geautomatiseerd labelen van personeelsadvertenties de nauwkeurigheid van het handmatig labelen kan benaderen of zelfs kan overtreffen: de besproken methodieken behalen vrijwel zonder uitzondering een nauwkeurigheid die dicht tegen de 100% ligt. Omdat voor het domein van de personeelsadvertenties minstens even sterke wetmatigheden gelden als voor het domein van de wetenschappelijke artikelen of krantenpagina's (de meestgebruikte testdomeinen) kan met behoorlijke zekerheid gesteld worden dat de labelmethodiek voor personeelsadvertenties óók een nauwkeurigheid kan behalen die dicht bij de 100% ligt. Zo'n nauwkeurigheid zou de nauwkeurigheid van het handmatig labelen (die overigens nooit gemeten is) waarschijnlijk zelfs overtreffen.
Aan de eisen van MatchCare dat het labelproces zó ontworpen moet worden dat de overige subprocessen onveranderd kunnen blijven en dat het labelproces te verbeteren moet zijn wordt voldaan door het ontwerp.

De resultaten van dit afstudeeronderzoek kunnen goed gebruikt worden om vervolgonderzoek te verrichten. Het meest voor de hand liggende vervolg van dit onderzoek, de implementatie van de labelmethodiek en het evalueren van de labelmethodiek, zal echter door het faillissement van MatchCare nooit uitgevoerd worden.

Er zijn wel een aantal mogelijkheden voor vervolgonderzoek te noemen, maar die hebben alleen maar een wetenschappelijke doel en geen praktijkwaarde:

- Onderzoeken hoe de labelmethodiek interactief gemaakt kan worden. Interactiviteit met de gebruiker kan zeer nuttig, ondere andere voor het oplossen van ambigue situaties en het maken van run-time beslissingen. Tevens kan de gebruiker in een interactieve methodiek error feedback geven waarmee de parameters uit de rule base aangepast kunnen worden;
- Onderzoeken hoe de labelmethodiek moet worden aangepast om de twee beperkingen uit sectie 1.6 weg te kunnen nemen. Als een personeelsadvertentie meerdere vacatures mag bevatten dan moet herkend worden o.a. herkend worden welke tekstsegmenten bij welke vacature horen. Dat is ingewikkeld omdat sommige tekstsegmenten bij alle vacatures horen (vaak de informatiecategorieën Werkgeveromschrijving en de Sollicitatie-informatie/Adverteerdersgegevens);
- Onderzoeken of het mogelijk is door middel van een Information Extraction-methodiek voor bepaalde velden in de invoerapplicatie een rule base samen te stellen. Daarmee wordt het bijvoorbeeld mogelijk de adresgegevens of de opleidingsgegevens automatisch te herkennen in de vacaturetekst en in het juiste veld van de invoerapplicatie te plaatsen;
- Onderzoeken of het bruikbaar is een document model op te stellen per adverteerder (zie sectie 5.2);
- Onderzoeken of het mogelijk is het verrijkingsproces te automatiseren;
- Onderzoeken of het mogelijk is een vacaturepagina te segmenteren in de personeelsadvertenties waaruit de pagina bestaat.

# Referenties


[1] A.K. Nijmeijer. Retrieving information from job postings. *Master thesis.* Rijksuniversiteit Groningen, The Netherlands, 2001.

[2] M. Wilhelm. PressAnalyzer – print media & text mining (the pilot version). *Master thesis.* Rijksuniversiteit Groningen, The Netherlands, 2001.

[3] M.E. Califf. Relational learning techniques for natural language information extraction. *Dissertation*, University of Texas, Austin, 1998.

[4] Interne Communicatie van de TU/e. Uit *Handwijzer Personeelsadvertenties*, opgeslagen op 29 augustus 2001, http://www.fb.tue.nl/huisstijl/Huisstijl FAQ.htm.

[5] J. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26:832-843, 1983.

[6] H. Walischewski. Automatic knowledge acquisition for spatial document interpretation. In *Proceedings of the 4th ICDAR '97*, pages 243-247, Ulm, Germany, 1997.

[7] M. Aiello, C. Monz, L. Todoran and M. Worring. Document understanding for a broad class of documents. In *International Journal for Document Analysis and Recognition*, vol. 5, no. 1, pages 1-16, 2002.

[8] Y. Liu, R. Fenrich, S.N. Srihari. An object attribute thresholding algorithm for document image binarization. In *Proceedings of the 2nd International Conference on Document Analysis and Recognition*, pages 278-281, Tsukuba Science City, Japan, Oct. 20-22, 1993.

[9] Y. Liu and S.N. Srihari. Document image binarization based on texture features. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pages 540-544, May 1997.

[10] S.N. Srihari, V. Govindaraju. Analysis of textual images using the Hough transform. *Machine Vision and Applications*, 2(3):141-153, 1989.

[11] K. Jain and B. Yu. Document representation and its application to page decomposition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20(3), pages 294-308, March 1998.

[12] Y. Ishitani. Document skew detection based on local region complexity. In *Proceedings of the 2nd International Conference on Document Analysis and Recognition*, pages 49-52, Tsukuba Science City, Japan, Oct. 20-22, 1993.

[13] G. Nagy and S.C. Seth. Hierarchical representation of optically scanned documents. In *Proceedings of the 7th International Conference on Pattern Recognition*, pages 347-349, Montreal, Canada 1984.

[14] G. Nagy, S.C. Seth and S.D. Stoddard. Document analysis with an expert system. In *Proceedings of Pattern Recognition in Practice II*, Amsterdam, The Netherlands, Juni 19-21, 1985.

[15] K.J. Wong, R.G. Casey and F.M. Wahl. Document analysis system. In *IBM Journal of Research and Development*, 26(6):647-656, 1982.

[16] H.S. Baird. Background structure in document images. In *Advances in Structural and Syntactic Pattern Recognition*, pages 253-269, World Scientific, Singapore, 1992.

[17] R.C. Gonzalez and R.E. Woods. *Digital Image Processing*. Addison Wesley, 1992.

[18] L. O'Gorman. The document spectrum for page layout analysis. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1162-1173, 1993.

[19] K. Kise, A. Sato and M. Iwata. Segmentation of page images using the Area Voronoi Diagram. In *Computer Vision and Image Understanding*, vol. 70, pages 370-382, 1998.

[20] A.K. Jain and S. Bhattacharjee. Address block location on envelopes using Gabor filters. In *Pattern Recognition*, vol. 25, no. 12, 1992.

[21] A.K. Jain and Y. Zhong. Page layout segmentation based on texture analysis. In *Pattern Recognition*, 29(5):743-770, 1996.

[22] T. Akiyama and N. Hagita. Automated entry system for printed documents. In *Pattern Recognition*, 23(11):1141-1154, 1990.

[23] J. Liang, J. Ha, R.M. Haralick and I.T. Phillips. Document layout structure extraction using bounding boxes of different entities. In *Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision*, Sarasota FL, pages 278-283, 1996.

[24] W. Scherl, F. Wahl and H. Fuchsberger. Automated separation of text, graphic and picture segments in printed material. In *Pattern Recognition in Practice*, pages 213-221. North-Holland, Amsterdam, 1980.

[25] R. Sivaramaakrishnan, I.T. Phillips, J. Ha, S. Subramanium and R.M. Haralick. Zone classification in a document using the method of feature vector generation. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, pages 541-544, Montreal, Canada, August 1995.

[26] D.X. Le, G.R. Thoma and H. Wechsler. Classification of binary document images into textual or nontextual data blocks using neural network models. In *Machine Vision and Applications*, 8:289-304, 1995.

[27] H. Walischewski. Learning regions of interest in postal automation. In *Proceedings of the 5th International Conference on Document Analysis and Recognition*, Bangalore, India, 1999.

[28] F. Bapst, R. Brugger and R. Ingold. Towards an Interactive Document Structure Recognition System. Internal working paper no. 95-09, Institute of Informatics, March, 1995.

[29] RAF Technology Inc. *DAFS: Document Attribute Format Specification*. 1995.

[30] Text ISO 8613, Information Processing and Office Systems. *Office Document Architecture (ODA) and Interchange Format (ODIF)*. Geneva, 1989.

[31] Text ISO 8879, Information Processing and Office Systems. *Standard Generalized Markup Language (SGML)*. Geneva 1986.

[32] E.R. Harold. *XML Bible*. IDG Books Worldwide Inc. 1999.

[33] Ishitani Y. Document image analysis with cooperative interaction between layout analysis and logical structure analysis. In *Proceedings of DLIA '99*, 1999.

[34] E. Kulper and R. Wieringa. Fuzzy spatial reasoning for document layout. *Master thesis*. Rijksuniversiteit Groningen, The Netherlands, 1999.

[35] S. Klink, A. Dengel and T. Kieninger. Document structure analysis based on layout and textual features. In *Proceedings of the 4th IAPR International Workshop on Document Analysis Systems (DAS2000)*, pages 99-111, Rio de Janeiro, Brazil, 2000.

[36] J.L. Fisher. Logical structure descriptions of segmented document images. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 302-310, Saint-Malo, France, September 1991.

[37] S. Tsujimoto and H. Asada. Understanding multi-articled documents. In *Proceedings of the 10th International Conference on Pattern Recognition*, pages 551-556, 1990.

[38] D.X. Le, J. Kim, G. Pearson G and G.R. Thoma. Automated labeling of zones from scanned documents. In *Proceedings of the Symposium on Document Image Understanding Technology*, pages 219-216, Institute for Advanced Computer Studies, Maryland, April 1999.

[39] J. Kim, D.X. Le and G.R. Thoma. Automated labeling algorithms for biomedical document images. In *Proceedings of the 7th World Multiconference on Systemics, Cybernetics and Informatics*, Vol. V, pages 352-357, Orlando FL, July 2003.

[40] J. Kreich, A. Luhn and G. Maderlechner. An experimental environment for model based document analysis. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 50-58, Saint-Malo, France, September 1991.

[41] A.L. Spitz. Style-directed document recognition. In *Proceedings of the 5th International Conference on Document Analysis and Recognition*, Bangalore, India, 1999.

[42] A. Yamashita, T. Amano, I. Takahashi and K. Toyokawa. A model based layout understanding method for the document recogniton system. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 130-138, Saint-Malo, France, September 1991.

[43] D. Niyogi and S.N. Srihari. Using domain knowledge to derive the logical structure of documents. In *SPIE*, pages 114-125, 1996.

[44] D.X. Le and G.R. Thoma. Page layout classification technique for biomedical documents. In *Proceedings of the World Multiconference on Systems, Cybernetics, and Informatics*. 2000.

[45] K. Summers. Near-wordless document structure classification. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, pages 462-465, Montreal, Canada, August 1995.

[46] J. Liang and D. Doermann. Logical labeling of document images using layout graph matching with adaptive learning. In *Proceedings of the IAPR International Workshop on Document Analysis Systems (DAS2002)*, pages 224-235, 2002.

[47] M. Krishnamoorthy, G. Nagy, S.C. Seth and M. Viswanathan. Syntactic segmentation and labeling of digitized pages from technical journals. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(7):737-747, 1993.

[48] G. Nagy, S.C. Seth and M. Viswanathan. A prototype document image analysis system for technical journals. In *Computer*, 25(7):10-22, 1992.

[49] J. Kim, D.X. Le and G.R. Thoma. Automated labeling in document images. In *Proceedings of SPIE Conference on Document Recognition and Retrieval VIII*, pages 111-122, San Jose CA, January 2001.

[50] R. Ingold and D. Armangil. A top-down document analysis method for logical structure detection. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 41-49, Saint-Malo, France, September 1991.

[51] A. Conway. Page grammars and page parsing: a syntactic approach to document layout recognition. In *Proceedings of the 2nd International Conference on Document Analysis and Recognition*, pages 761-764, Tsukuba Science City, Japan, Oct. 20-22, 1993.

[52] K. Kubota, O. Iwaki and H. Arakawa. Document understanding system. In *Proceedings of the 7th International Conference on Pattern Recognition*, pages 612-614, Montreal, Canada, July 30 – August 2, 1984.

[53] J. Higashino, H. Fujisawa, Y. Nakano and M. Ejiri. A knowledge-based segmentation method for document understanding. In *Proceedings of the 8th International Conference on Pattern Recognition*, pages 745-748, Paris, France, 1986.

[54] H. Fujisawa and Y. Nakano. A top-down approach for the analysis of document images, In *Proceedings of the SSPR '90*, pages 113-122, 1990.

[55] R. Brugger, A. Zramdini and R. Ingold. Modeling documents for structure recognition using generalized n-grams. In *Proceedings of the International Conference on Document Analysis and Recognition*, 1997.

[56] M. Diligenti, P. Frasconi, M. Gori. Image document categorization using Hidden Tree Markov Models and structured representations. In *Proceedings of the International Conference on Applications of Pattern Recognition*, 2001.

[57] S. Mao and G.R. Thoma. Bayesian Learning of 2D Document Layout Models for Automated Preservation Metadata Extraction. In *Proceedings of the 4th IASTED International Conference on Visualization, Imaging and Image Processing (VIIP 2004)*, pages 329-334, September 6-8 2004, Marbella, Spain.

[58] Y. Tateisi and N. Itoh. Using stochastic syntactic analysis for extracting a logical structure from a document image. In *Proceedings of the International Conference on Pattern Recognition*, pages 391-394, Jerusalem, Israel, October 1994.

[59] T. Kanungo and S. Mao. Stochastic Language Model for Style-Directed Physical Layout Analysis of Documents. In *IEEE Transactions on Image Processing*, vol. 12, no. 5, pages 583-596, May 2003.

[60] F. Esposito, D. Malerba, G. Semeraro, S. Ferilli, O. Altamura, T.M.A. Basile, M. Berardi, M. Ceci, N. di Mauro. Machine learning methods for automatically processing historical documents: from paper acquisition to XML transformation. In *Proceedings of the 1st International Workshop on Document Image Analysis for Libraries (DIAL04)*, pages 328–335, IEEE Computer Society, 2004.

[61] D.X. Le, G.R. Thoma and H. Wechsler. Document image analysis using integrated image and neural processing. In *Proceedings of the 3rd ICDAR*, vol. I, pages 327-330, Montreal, Canada, August 14-16, 1995.

[62] K. Hadjar and R. Ingold. Logical labeling of arabic newspapers using artificial neural nets. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 426-431, 2005.

[63] M.E. Califf and R.J. Mooney. Bottom-up relational learning of pattern matching rules for information extraction. In *Journal of Machine Learning Research*, vol. 4, pages 177-210, 2003.

[65] T. Joachims. Text categorization with Support Vector Machines: learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, pages 137-142, 1998.

[66] D.M. Chickering, D. Heckerman and C. Meek. A Bayesian approach for learning Bayesian networks with local structure. In *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence*, 1997.

[67] I.J. Good. *The estimation of probabilities: an essay on modern Bayesian methods.* MIT Press. 1965.

[68] D. Heckerman, D. Geiger and D.M. Chickering. Learning Bayesian networks: the combination of knowledge and statistical data. In *Machine Learning* (20), pages 131-163, 1995.

[69] N. Fuhr and C. Buckley. A probabilistic learning approach for document indexing. In *ACM Transactions on Information Systems 9*, pages 223-248, 1991.

[70] D. Derrien-Peden. Frame-based system for macro-typographical structure analysis in scientific papers. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 311-319, Saint-Malo, France, September 1991.

[71] M. Sharpe, G. Sutcliffe and N. Ahmed. Implementation of an intelligent document understanding and reproduction system. *Technical Report 94/10*, Department of Computer Science, James Cook University, Townsville, Australia,1994.

[72] D. Chetverikov, J. Liang, J. Komuves and R.M. Haralick. Zone Classification Using Texture Features. In *Proceedings of the 13th International Conference on Pattern Recognition (ICPR'96)*, vol. 3, page 676, 1996.

# Appendix A

## Lijst van gebruikte figuren:

Figuur 1.1 pagina 5
Het vacatureproces van MatchCare

Figuur 1.2 pagina 10
Resultaat van het afstudeeronderzoek van Marco Nijmeijer: het nieuwe fulltextproces.

Figuur 1.3 pagina 12
Een voorbeeld van een personeelsadvertentie met meerdere vacatures.

Figuur 1.4 pagina 14
Het vernieuwde fulltextproces nadat het opdelen van de personeelsadvertentie is verplaatst van YAN naar LeO, de nieuwe labelapplicatie.

Figuur 1.5 pagina 17
Een gelabelde personeelsadvertentie. Met de afkortingen W, FT, F, P, A en SA worden achtereenvolgens de labels Werkgeversomschrijving, FunctieTitel, Functie-omschrijving, Profiel, Arbeidsvoorwaarden en Sollicitatie-informatie/Adverteerdersgegevens aangeduid.

Figuur 2.1 pagina 24
Voorbeeld van het gebruik van een huisstijl bij het opstellen van een personeelsadvertentie.

Figuur 2.2 pagina 26
Een personeelsadvertentie waarin de drie tekstuele wetmatigheden duidelijk naar voren komen.

Figuur 2.3 pagina 30
Een personeelsadvertentie aan de hand waarvan de eerste en tweede layoutgerichte wetmatigheden duidelijk gemaakt kunnen worden.

Figuur 2.4 pagina 32
Een personeelsadvertentie met meerdere tekstkolommen.

Figuur 2.5 pagina 34
Een personeelsadvertentie opgedeeld in 10 horizontale zones. Aan de linkerkant staan de zonenummers.

Figuur 3.1 pagina 40
Overzicht van onderzoeksgebieden die onder Document Processing vallen. De voor mijn afstudeeronderzoek relevante gebieden zijn vetgedrukt.

Figuur 3.2 pagina 41
Het stroomdiagram van het Document Analysis-proces.

Figuur 3.3 pagina 43
De relatie tussen een gescande personeelsadvertentie, de gesegmenteerde personeelsadvertentie (na het Document Layout Analysis-proces) en de gelabelde personeelsadvertentie (na het Document Understanding-proces).

Figuur 3.4 pagina 50
Een voorbeeld van de constructie van een X-Y Tree uit de segmenten van een gescande personeelsadvertentie.

## Lijst van gebruikte tabellen:

Tabel 1 pagina 27
De precision en recall van het labelen door middel van keywords.

Tabel 2 pagina 29
Kwantificatie van de afgeleide wetmatigheid die stelt dat elk paar categorieën bijna altijd een vaste interne volgorde heeft.

Tabel 3 pagina 33
De kansverdeling van mogelijke spatial relations voor de informatiecategorie Functie-omschrijving ten opzichte van de categorie Werkgeversomschrijving bij personeelsadvertenties met één tekstkolom.

Tabel 4 pagina 33
De kansverdeling van mogelijke spatial relations voor de informatiecategorie Functie-omschrijving ten opzichte van de categorie Werkgeversomschrijving bij personeelsadvertenties met meerdere tekstkolommen.

Tabel 5 pagina 34
De kans dat in een personeelsadvertentie informatiecategorie *x* aangetroffen wordt in zone *y*.

Tabel 6 pagina 35
Twee relevante statistieken met betrekking tot de hoeveelheid alinea's die voor een informatiecategorie gebruikt worden.

Tabel 7 pagina 81
De kansverdeling van mogelijke spatial relations voor de informatiecategorie Arbeidsvoorwaarden ten opzichte van de categorie Sollicitatie-informatie/Adverteerdersgegevens bij personeelsadvertenties met één tekstkolom.

Tabel 8 pagina 81
De kansverdeling van mogelijke spatial relations voor de informatiecategorie Functie-omschrijving ten opzichte van de categorie Werkgeversomschrijving bij personeelsadvertenties met één tekstkolom.

Tabel 9 pagina 82
De kansverdeling van mogelijke spatial relations voor de informatiecategorie Profiel ten opzichte van de categorie Sollicitatie-informatie/Adverteerdersgegevens bij personeelsadvertenties met één tekstkolom.

# Appendix B

## De keywordlijsten:

**Werkgeversomschrijving**

bedrijf, afdeling, ondememing, faculteit, universiteit, centrum, stichting

**Functie-omschrijving**

taken, verantwoordelijkheden, werkzaamheden, functie-informatie, functie-inhoud

**Profiel**

eisen, werkervaring, ervaring, jaren, stressbestendig, "beschikt over", profiel, kennis, opleiding, "in het bezit van", diploma, enthousiast, flexibel, "wij vragen", vaardigheden

**Arbeidsvoorwaarden**

jaarcontract, maandcontract, schaal, CAO, arbeidsvoorwaarden, salaris, winstuitkering, kinderopvang, FWG, spaarloonregeling, aanbod, "wij bieden"

**Sollicitatie-informatie/Adverteerdersgegevens**

sollicitatie, CV, schriftelijk, e-mail, email, t.a.v., acquisitie, vacaturenummer, telefoonnummer, "binnen 14 dagen", "binnen 2 weken", "afdeling P&O", "afd. P&O", Postbus, sollicitatieprocedure, inlichtingen, geïnteresseerd

# Appendix C

## Voorbeelden van kansverdelingen van qualitative spatial relations:

| *Soort relatie (%)* | P | M | O | Fi | Di | S | E | Si | D | F | Oi | Mi | Pi |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| x-intervallen | 0 | 0 | 5 | 7 | 3 | 6 | 50 | 12 | 5 | 2 | 10 | 0 | 0 |
| y-intervallen | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |

**Tabel 7 – de kansverdeling van mogelijke spatial relations voor de informatiecategorie Arbeidsvoorwaarden ten opzichte van de categorie Sollicitatie-informatie/Adverteerdersgegevens bij personeelsadvertenties met één tekstkolom.**

| *Soort relatie (%)* | P | M | O | Fi | Di | S | E | Si | D | F | Oi | Mi | Pi |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| x-intervallen | 0 | 1 | 7 | 7 | 1 | 3 | 55 | 5 | 9 | 4 | 6 | 2 | 0 |
| y-intervallen | 7 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 11 | 77 |

**Tabel 8 – de kansverdeling van mogelijke spatial relations voor de informatiecategorie Functie-omschrijving ten opzichte van de categorie Werkgeversomschrijving bij personeelsadvertenties met één tekstkolom.**

| *Soort relatie (%)* | P | M | O | Fi | Di | S | E | Si | D | F | Oi | Mi | Pi |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| x-intervallen | 0 | 0 | 7 | 8 | 1 | 2 | 62 | 8 | 3 | 6 | 3 | 0 | 0 |
| y-intervallen | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |

**Tabel 9 – de kansverdeling van mogelijke spatial relations voor de informatiecategorie Profiel ten opzichte van de categorie Sollicitatie-informatie/Adverteerdersgegevens bij personeelsadvertenties met één tekstkolom.**