Matchcare Information Services
Groningen

RETRIEVING INFORMATION FROM JOB POSTINGS

Author : A.K. Nijmeijer
Date : 26-08-02
Supervisor : L. Spaanenburg

Matchcare Information Services
Groningen

Retrieving Information from Job Postings

Author : A.K. Nijmeijer
Date : 26-08-02
Supervisor : L. Spaanenburg
Begeleider : M. Keupink
A. Bruggeman

Voorwoord

Als onderdeel van de studierichting Technische Informatica aan de Rijksuniversiteit Groningen heb ik een afstudeeropdracht uitgevoerd bij Matchcare Information Services. Matchcare is een dataproductiebedrijf gevestigd op het Zernike Science Park in Groningen.

Voor u ligt het resultaat van enkele maanden werk en vijf jaar studie aan de Rijksuniversiteit Groningen. De keuze voor het onderwerp van deze scriptie werd meteen duidelijk toen ik met mijn professor sprak over mijn afstuderen. De opdracht heeft een sterk pragmatisch karakter en bovendien is de theorie vrij nieuw van aard. Het zal de lezer niet ontgaan dat deze scriptie niet alleen gericht is op het vinden van een praktische oplossing voor een geconstateerd probleem. Minstens zo belangrijk is de theoretisch onderbouwing voor vervolg trajecten die mogelijkerwijs in de toekomst plaatsvinden.

Een woord van dank gaat uit naar de volgende personen:

- Prof.dr.ir. L. Spaanenburg; begeleider Rijksuniversiteit Groningen
- M. Keupink en A. Bruggeman; respectievelijk General Manager en directeur bij Matchcare Information Services.
- P. Visser, H. Noordhof en M. Wilhelm; directe collega's bij Matchcare Information Services.
- S. L. Geerts; steun en toeverlaat.

Groningen, augustus 2002
A.K.Nijmeijer

Samenvatting

De explosieve groei in computer toepassingen heeft papier niet overbodig gemaakt. Papier blijft als informatiedrager een wezenlijke rol in de maatschappij vervullen en is eerder in belang toe dan afgenomen. De vraag rijst dan hoe de computer gebruikt kan worden om informatie op papier beter te verwerken. Dit speelt o.a. een rol op de personeel bemiddelingsmarkt. De personeelsadvertentie in gedrukte media bereikt een groot en vaak anoniem publiek op een goedkope en eenvoudig toegankelijke wijze en zonder eisen aan de leeslocatie. In de verdere waardeketen speelt echter de computer weer een overheersende rol. Uitgevers van specialistische vacature tijdschriften hebben behoefte aan momentane strategische marktinformatie, de sollicitant zoekt aanvullende informatie over het adverterende bedrijf; zulke zoek processen zijn pas goed van de grond gekomen bij de invoering van datamarkten met een door strakke automatisering vereenvoudigd zoekmechanisme.

De scriptie richt zich op de omzetting van informatie in personeelsadvertenties naar een vacature database, die een verdere computer ondersteuning van andere activiteiten in de waardeketen mogelijk maakt. Het onderzoek begint vanuit de situatie waarin teksten door een leger hulpkrachten in de database gebracht worden en onderbouwt de strategische en tactische keuzes die gemaakt zijn om langs wegen der geleidelijkheid dit proces te verbeteren, te versnellen en te stroomlijnen. Het onderzoek is uitgevoerd bij de firma Matchcare Information Services.

Abstract

Despite the expectations, computerized automation has not made paper work superfluous. Paper has stayed an important carrier of information within the computerized society. On the contrary, new services are being built around paper work, leading to a more intimate coupling of paper and computer to obtain more and better knowledge. A typical example is in job advertisements. The printed job posting in daily and weekly journals reaches a large and often anonymous audience in a cheap and easy manner at random locations. This has always been true, but computerized services have been developed to allow the applicant to obtain more and contemporary information, while the publisher needs to review his market share for a daily feedback on strategic issues. Datamarts on companies and vacancies have come into use, where a fast access to immense datasets are coupled to easy access mechanisms. In other words, the print/machine interface has become a crucial part of a newly developing value network.

The thesis handles on the transformation of information hidden in printed advertisements into information structured in a computer database. Such a high quality database will serves to support other services, elsewhere in the value chain. The research has started at a moment, where legions of low-skilled people were retyping advertisements into the computer database. It recounts the strategic and tactical decisions, that were necessary to improve, accelerate and streamline the process without disturbing the on-going operation. The research has been performed at the Groningen location of Matchcare Information Services.

Inhoud

Hoofdstuk 1 De opdrachtgever

Om een beeld te geven van de onderneming waarbinnen de afstudeeropdracht is uitgevoerd, gaat dit hoofdstuk in op het ontstaan van Matchcare Information Services, de organisatiestructuur, de productiestructuur en het eigen product Zoomnet.

1.1 Het ontstaan van Matchcare Information Services

In mei 1995 stichtte de heer Keupink de uitgeverij IntraMedium. De hoofdactiviteit van dit bedrijf was de uitgave van de Sollicitatiekrant. In de krant stonden verkorte versies van vacatures die in de week daarvoor in verschillende landelijke en regionale dagbladen en in een groot aantal vakbladen verschenen. De vacatures werden geselecteerd op rubriek en een jaar later werd hier de indeling in provincies aan toegevoegd, zodat men ook in de eigen regio gericht naar een baan kon zoeken. Tevens bood de krant informatie in de vorm van artikelen over de arbeidsmarkt en het bedrijfsleven. De krant, die wekelijks verscheen, was bedoeld voor middelbaar en hoger opgeleiden en kende een prijs van fl. 3,95. (€ 1,80) In mei 1997 werd IntraMedium overgenomen door VNU Business Publications. IntraMedium ging toen verder onder een andere naam: Keymark Services, een joint-venture van VNU Business Publications en Pelican Services. Vanaf dat moment waren de bedrijfsactiviteiten niet alleen meer gericht op de Sollicitatiekrant, maar ging Keymark Services zich op meerdere manieren richten op de arbeidsmarkt. Maar de basis, het verzamelen van vacatures, bleef intact. Na verloop van tijd werd de Sollicitatiekrant opgedoekt, omdat het aantal vacatures te groot werd om te publiceren. Tevens ging voor Keymark Services het Internet een veel grotere rol spelen. De joint-venture was echter geen lang leven beschoren. VNU stapte er al snel uit, omdat ze te weinig mogelijkheden zagen zitten. De naam Keymark Services bleef ongewijzigd. In juni 2000 werd Keymark Services overgenomen door Matchcare, waardoor het bedrijf de naam Matchcare Information Services kreeg.

1.2 Huidige organisatiestructuren

De investeringsmaatschappij APM (Advanced Participation Management) heeft verschillende participaties. De holding AMSE (Advanced Matching Systems Europe), met de merknaam Matchcare, is één van die participanten. Keymark Services is in juni 2000 door deze holding overgenomen. De holding bestaat uit 3 werkmaatschappijen. Zoomnet is één van deze werkmaatschappijen.

1.3 Productie structuur

Matchcare Information Services is op dit moment verdeeld over 3 vestigingen: Het hoofdkantoor in Zaltbommel, de primaire productieomgeving in Groningen en de secundaire productieomgeving in Deventer. De hoofdactiviteit van Matchcare Information Services is het verzamelen van vacatures via diverse bronnen en het digitaal verkopen van deze informatie met behulp van applicaties. De vacatures worden uit verschillende bronnen gehaald. De gedrukte media en het Internet zijn hiervan de belangrijkste. Om de informatie op de digitale markt te verkopen is een bewerkingsproces nodig, dat speciaal op de doelgroep is toegesneden. Binnen Matchcare Information Services lopen op dit moment twee van zulke processen: het samenvatproces en het fulltextproces. Deze processen worden in de subparagrafen hieronder besproken.

1.3.1 Samenvatproces

In Groningen wordt het samenvatproces gecoördineerd. Dit proces behelst het verzamelen, het verdelen onder de medewerkers alsmede het daadwerkelijke samenvatten van de vacatures uit diverse gedrukte media en van het internet. De gedrukte media zijn landelijke en regionale dag- en weekbladen, evenals vakbladen.

Gedurende de week worden vacatures in vak- en weekbladen geteld. Na het tellen worden de vacatures verdeeld onder de deelnemers in het samenvatproces. In het geval van thuiswerkers worden de vacatures per post verstuurd. Op zaterdag gaat het voornamelijk om de landelijke dagbladen.

In het geval van regionale dag- en weekbladen gaat het anders in zijn werk. Hierbij wordt gewerkt op basis van goede en harde afspraken. Thuiswerkers tellen de vacatures in de bladen die bij hem of haar in de regio verschijnen en voeren deze vacatures zelf in.

Voor het coördineren van het samenvatproces zijn door de week twee fulltime medewerkers en in het weekend drie tot vijf parttime medewerkers, noodzakelijk. Daarnaast wordt gebruik gemaakt van 60 tot 80 oproepkrachten (thuiswerkers) die zich verspreid over het hele land bevinden. Deze thuiswerkers worden op stuksbasis betaald, hetgeen inhoudt dat thuiswerkers zelf dienen door te geven wat ze hebben gedaan. De stukprijs is afhankelijk van de grootte van een vacature. Dit hangt samen met de hoeveelheid werk en tijd die het kost om een vacature samen te vatten. De thuiswerkers moeten zelf bijhouden hoeveel van de vacatures groot, middelgroot of klein zijn. Dit wordt de 'verdeling' genoemd. Steekproefsgewijs wordt gekeken of deze verdeling niet te positief uitvalt voor de medewerkers.
De vaste medewerkers (coördinatoren) verdelen het werk over de oproepkrachten. Aangezien op zaterdag de grootste hoeveelheid vacatures in de landelijke en regionale dagbladen staat, wordt op zaterdag het meeste werk verricht. Voorafgaand aan de werkdag wordt een taakverdeling gemaakt. 's Morgens wordt alles echt verdeeld.

Alle vacatures worden gedurende de dag ingevoerd (ofwel samengevat) in een speciaal invoerprogramma en vervolgens (vanuit het land) opgestuurd naar Groningen. De gehele zaterdag wordt alles gecontroleerd dat wordt opgestuurd. Op de zaterdagavond vindt de eindcontrole plaats. Na afronding van de controle wordt het bestand verzonden naar de server op het kantoor in Groningen.

1.3.2 Fulltext proces

Het fulltext proces wordt net als het samenvatproces gecoördineerd in Groningen. Het proces is per 1 augustus 2000 gestart. Ten tijde van het uitvoeren van de afstudeeropdracht ontvangen een aantal uitgevers digitale fulltext vacatures retour.
Het proces verloopt als volgt:
De vacatures die in een krant staan worden in hun geheel overgetypt. De vacatures moeten hierbij aan een aantal voorwaarden voldoen. Voldoet een vacature niet aan de voorwaarden, dan zal de vacature niet worden ingevoerd. Eén van de voorwaarden die aan de vacatures worden gesteld, is dat de vacature een bedrijfslogo moet bevatten. De kale informatie in een personeelsadvertentie is vaak onvolledig. De verborgen informatie moet naar boven worden gehaald. In het verrijkingsproces wordt de verborgen informatie naar boven gehaald door bijvoorbeeld aan iedere functie een aantal kenmerken ofwel eigenschappen toe te voegen. Bijvoorbeeld: Een werkgever zoekt naar een telefoniste die een 'enthousiaste teamplayer' is. Dit is weinig veelzeggend. Voor de functie van 'telefoniste' zijn een aantal eigenschappen noodzakelijk, te weten klantvriendelijkheid, heldere stem, enig organisatorisch vermogen, et cetera.

Deze eigenschappen worden toegevoegd aan de functie 'telefoniste'. Dit toevoegen van eigenschappen gebeurt bij alle functies. Een ongeveer zelfde actie wordt verricht bij onder meer branche, opleiding en vakgebied.

Nadat de verrijking heeft plaatsgevonden en het 'product' klaar is, vindt de uitlevering aan de betreffende uitgever plaats. De uitgever plaatst het vervolgens bijvoorbeeld op zijn eigen website.

1.3.3 Zoomnet

Zoomnet is een joint-venture van Matchcare Information Services en De Telegraaf. Zoomnet maakt via een speciale website (www.zoomnet.nl) anoniem contact tussen werkgevers en werkzoekenden mogelijk. De werkzoekende kan via invultabellen op de website een gedetailleerd beeld geven van zijn/haar opleiding, werkervaring en carrièrewensen. Aan de hand van een door Matchcare ontwikkeld softwareprogramma wordt deze informatie 'verrijkt' door het te voorzien van een groot aantal herkenbare codes op het gebied van functie, opleiding en branche. Met de gegevens uit de vacatures gebeurt hetzelfde. Hierdoor ontstaan er zogenaamde matches. Wanneer mensen reageren op een bepaalde vacature krijgen bedrijven automatisch een e-mail voorzien van een wachtwoord. Met het wachtwoord krijgen de werkgevers toegang tot profielschetsen van de kandidaten. Willen de bedrijven daadwerkelijk contact zoeken met de dan nog anonieme werkzoekenden, dan moet er een abonnement bij Zoomnet zijn afgesloten. De website is om een aantal redenen inmiddels uit de lucht. Het principe van dit matchen is echter terug te vinden in een aantal producten van Matchcare zoals bijvoorbeeld CARMA, www.matchcare.nl/webwerving.html.

Hoofdstuk 2 De opdracht

In dit hoofdstuk wordt de opdracht uiteengezet. Het gaat hier om de aanleiding, evenals de daadwerkelijke probleemstelling en deelvragen die hieruit voortvloeien.

2.1 De aanleiding

We leven in het `informatietijdperk´. Dat wil zeggen: informatie heeft in onze samenleving een prominente positie ingenomen. Computers geven ons de mogelijkheid om informatie snel te benaderen, te bewerken en op te slaan, tegen steeds lagere kosten. Als gevolg hiervan treffen we bijvoorbeeld steeds vaker grote `pakhuizen´ van informatie aan [1]; op het internet is het groeiende informatieaanbod niet te stuiten. Gegevens kunnen in een duidelijke structuur opgeslagen zijn. Voorbeelde hiervan is een relationele of object- georiënteerde database, maar ze kunnen ook als tekst, dus in de vorm van een natuurlijke taal of als plaatje zijn opgeslagen. Matchcare Information Services is zo'n `pakhuis´ van informatie. Matchcare Information Services richt zich op het verzamelen, verrijken en distribueren van vacature beschrijvingen, waarbij het begrip `verrijking´ nog nadere invulling behoeft. Momenteel worden de vacatures verzameld uit de landelijke gedrukte media die verkrijgbaar zijn op de publieke markt. Zoals in het vorige hoofdstuk reeds is geschetst, levert Matchcare Information Services fulltext vacatures uit. Het verrijken van vacatures is in feite een interpretatiestap van de invoerder.

> Dhr. Keupink (General Manager): `*Wij beschouwen iedere personeelsadvertentie als een belangrijke informatiedrager. Iedere vacature bevat informatie over de functie, over de werkgever, over het profiel van de gezochte werknemer et cetera. Hoe meer je van deze informatie kunt verwerken, des te bruikbaarder wordt je database. Wekelijks registreren wij zo'n 11.000 vacatures.* ´ [7]

2.2 Opdrachtbeschrijving

Op enige termijn zal de markt voor Matchcare Information Services eerder in vergaand verrijkte dan in ruwe data liggen. Hiervoor zal de mining technologie, zoals die nu nog manueel in het fulltext proces ondersteund is, in vergaande mate geautomatiseerd moeten worden. Ook zal de fulltext verwerking in snelheid moeten toenemen om aan de vraag van de klant te kunnen blijven voldoen. Het domein van dit afstudeerproject zal liggen in de automatisering van het fulltext proces. Toename in invoersnelheid is het voornaamste criterium. Bovendien moet de ondersteuning voor het miningproces hierin worden meegenomen.

2.3 Theoretisch

Het theoretische doel is het bestuderen en adviseren op het gebied van het automatisch vinden en classificeren van één of meerdere advertentie(s) op een krantenpagina. Classificeren gaat er in dit geval vanuit dat de advertentie een vacature is zoals Matchcare Information Services dit heeft gedefinieerd. Deze definitie berust op een aantal kenmerken die intern gehanteerd worden; een kleine schets van de kenmerken is al gegeven in voorgaand hoofdstuk. De kenmerken vormen voor dit verslag een houvast, maar moeten als zeer dynamisch worden beschouwd. Het classificeren dient flexibel geïmplementeerd te worden zodat uitbreiding van kenmerken geen probleem geeft. Het theoretische doel zoals hierboven geschetst zal daarna verder bestudeerd worden om dezelfde technieken toe te passen op één advertentie.

2.4 Pragmatisch

Niet geheel vanzelfsprekend ligt het eerste doel in het realiseren van een set gereedschappen, waarmee het theoretische doel gerealiseerd kan worden. Echter de automatisering, zal doorgang moeten vinden. Deze gereedschappen hebben tot doel om het proces van `tikken´ te vervangen door `clicken´ om daarmee een snellere en meer efficiënte invoer te behalen. Dit gedeelte dient geleidelijk te worden ingevoerd en met name snel resultaten te boeken. Dit vanwege het feit dat de druk die op het Full-text overtypen van de advertenties ligt een te lage productiviteit heeft die ook niet zal verbeteren. Dat de snelheid in manuele invoer niet zal verbeteren ligt simpelweg aan het feit dat mensen niet meer kunnen typen dan dat ze nu kunnen. Een uitbreiding van het machine arsenaal is waarschijnlijk de enige oplossing, maar dit is een korte termijn visie en een oplossing voor de langere termijn is noodzakelijk.

2.5 Doelstelling

Het doel van de opdracht is in het kort: Het realiseren van een geautomatiseerd invoerproces, waarin het automatisch classificeren van advertenties ingebouwd kan worden. Daarnaast moet de verdere automatisering van de miningtechnologie worden ondersteund. Bovendien dient de techniek opnieuw toegepast te worden op afzonderlijke advertenties.

De complete formulering van de doelstelling van deze scriptie is dan als volgt:

> *Inbouwen van een productiestap die advertenties digitaliseert en de tekstuele informatie segmenteert en classificeert, rekening houdend met de huidige productie.*

Het realiseren van de gereedschappen die de processtappen vervangen heeft een aantal randvoorwaarden waarmee rekening moet worden gehouden.

- Als eerste het feit dat bij implementatie van bepaalde stappen de productie niet in gevaar mag komen. Klanten moeten bediend worden, waarbij een geleidelijke invoer van de automatiseringsstappen zeer gewenst is.
- Ten tweede dienen de gereedschappen dusdanig te worden geïmplementeerd dat een meer algemene inzet mogelijk is. Dit wil zeggen dat in eerste instantie een opzet wordt gemaakt om kranten als bron te nemen, maar een uitbreiding naar internetpagina's of andere media moet eveneens mogelijk zijn.
- Ten derde het kostenaspect. Aanschaf van machines en andere middelen moeten worden afgewogen tegen de kosten in de huidige productie methode.
- Als vierde randvoorwaarde het feit dat er zo snel mogelijk moet worden geautomatiseerd. Een verdeling van de in te voeren stappen is dan ook noodzakelijk om een gefaseerde invoer mogelijk te maken.

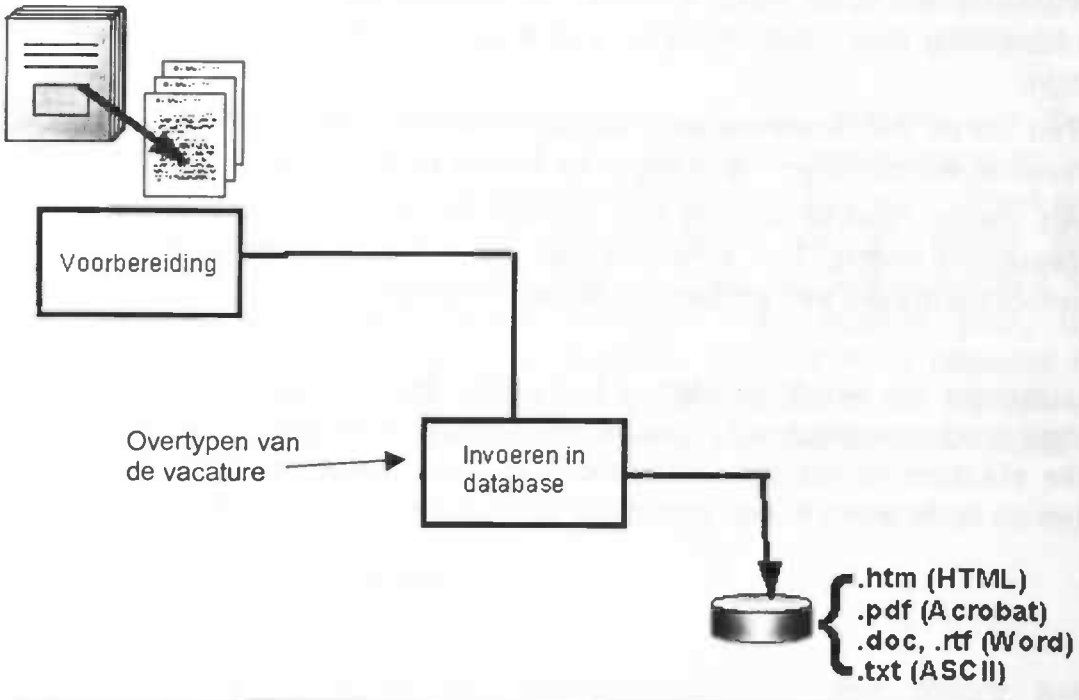
De strategie die wordt gevolgd is ten eerste het in kaart te brengen hoe de huidige productie geschiedt. Daarna zal onderzocht worden welke technieken op welke plaatsen in het productieproces van pas kunnen komen. En als laatste zullen de technieken in een prototype geïmplementeerd worden.

Hoofdstuk 3 Te realiseren workflow

Ter voorbereiding van het automatisering traject zal eerst de huidige workflow in kaart gebracht worden. Daarna zal een toekomstig model van de workflow worden weergegeven.

3.1 Huidige workflow

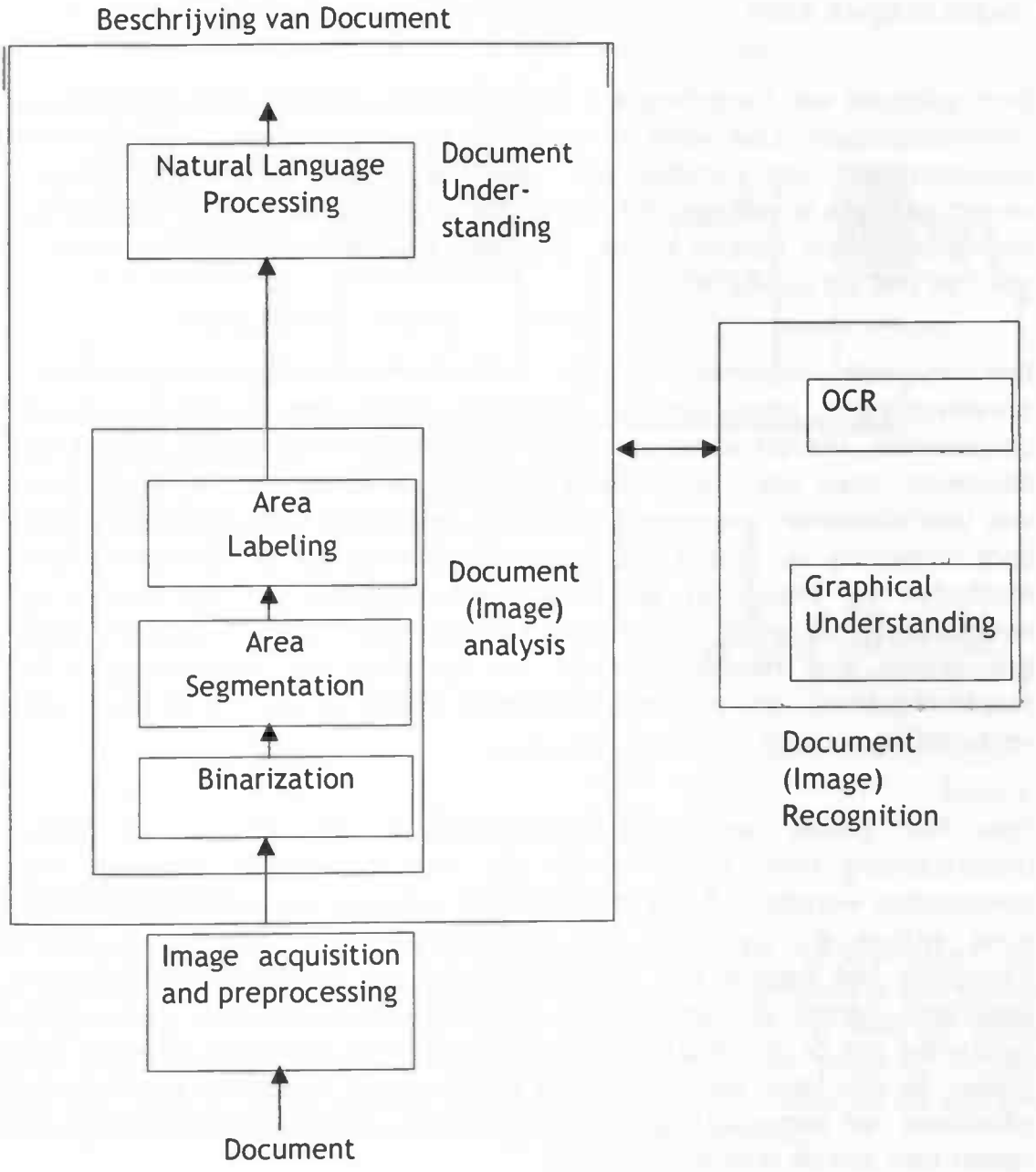
De huidige workflow is gebaseerd op typen. De voorbereiding bestaat uit het tellen en markeren van de voor fulltext invoer in aanmerking komende vacatures. Hoewel de inkomende stroom verdeeld is over samenvatting en



Huidige workflow

Figuur 3-1

fulltext en dan ook nog met verschillende invoer systemen, komt dit figuur 3-1 in functionaliteit wel overeen met de werkelijkheid. Zoals opvalt zijn de handelingen manueel. Een geautomatiseerd systeem is dan ook zeer welkom.



Document Understanding

Figuur 3-2

Het abstracte concept voor de extractie van informatie uit gedrukte tekst wordt Document Understanding (DU) genoemd [22]. De taak die we zien voor DU binnen het probleemgebied van de automatische vacature verwerking is op te delen in drie conceptuele gebieden: *document (image) analysis, document*

(image) recognition en Document understanding. De term document understanding is verwarrend omdat het wordt gebruikt voor het hele onderzoeksgebied en voor het laatste proces waar het de document structuur maakt op dezelfde manier als de mens dat zou doen. Een overzicht van DU is te vinden in figuur 3-2.

Deze volgorde van handelingen is niet gebonden aan het gebruik van papier als informatiedrager. Elke vorm van gedrukte tekst is op deze wijze verwerkbaar; een document kan derhalve ook een WEB pagina of een etiket zijn. Het wezenskenmerk is veeleer, dat het totale tekst blok in stukken wordt gehad tot een granulariteit bereikt wordt, waarmee Natural Language Processing (NLP) goed en snel uit de voeten kan.

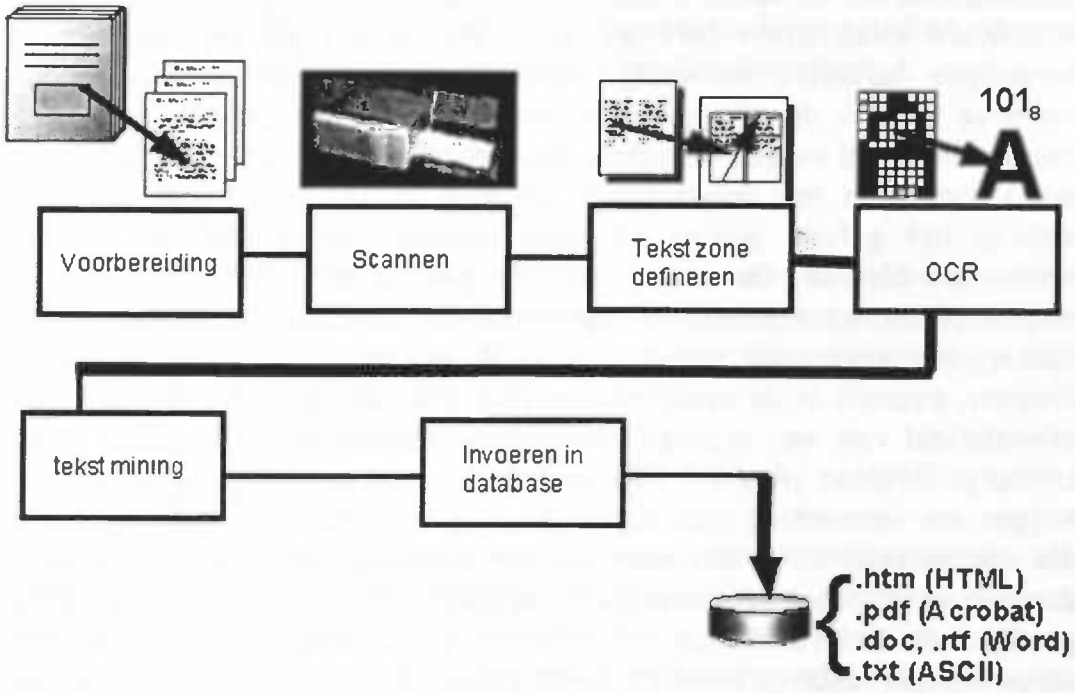
Een verwant probleem is dat van het automatisch genereren van samenvattingen, zowel van één document als van een bundel samenhangende documenten. Hierbij wordt geen gebruik meer gemaakt van de lay-out van het document, maar wordt een nieuwe structuur gebouwd die het relatieve belang van tekstelementen weergeeft. Ofschoon in theorie een generieke bouwwijze door toepassing van statistische hulpmiddelen mogelijk is, zijn met symbolische methoden op basis van geprefabriceerde corpora die het kennis gebied weerspiegelen duidelijk betere resultaten bereikt. Praktisch nadeel hiervan kan het gebrek aan robuustheid zijn ten opzichte van veranderingen in het probleemgebied, en dus een blijvende behoefte aan onderhoud van de verzameling corpora.

Daar het geheel aan vacatureverwerking in de context van Document Understanding past, wordt vanuit die visie een opzet gemaakt voor de toekomstige workflow. Beide procesgangen beginnen met een fulltext origineel, maar wijken daarna af in hun afhankelijkheid van de kennis structuur. Aangezien het bedrijf nog volop bezig is met de bewustwording voor de applicatie kennis en de lopende processen door moeten gaan, lijkt het verstandig om de automatisering te beginnen met het meer robuuste full-text proces. In een later stadium kan dit aangevuld en aangepast worden voor het genereren van samenvattingen. De uitwerking van deze techniek komt daarom verder niet aan de orde in dit verslag.

3.2 Toekomstige workflow

De toekomstige workflow zal gebaseerd zijn op een zekere mate van automatisering en zal dusdanig flexibel van opzet moeten zijn dat de verschillende onderdelen verder geautomatiseerd kunnen worden. Als voorbeeld voor de verdergaande automatisering is het proces Tekst zone definiëring of

tekstmining.
De onderdelen zijn direct terug te voeren op het DU model. De verschillende onderdelen die in het nieuwe geautomatiseerde traject van belang zijn, zijn uitgezet in figuur 3-3. De onderdelen van het proces zullen nader beschreven worden met daarbij de verschillende opties en beargumenteerde keuze. De keuze om deze opzet te gebruiken wordt nu eerst uitgelegd.



Toekomstige workflow

Figuur 3-3

3.3 Uitleg opzet

Er is voor gekozen om op image niveau blokken met tekst te selecteren, deze op te slaan en dan door de OCR te halen. Een andere mogelijkheid was geweest om deze twee stukken parallel aan elkaar te laten lopen en op basis van stukjes image die uit de OCR komen beslissingen te nemen. De opzet is uitvoerig besproken met de direct leidinggevende in het productieproces om tot een goed nieuw proces te komen, zonder al te veel drastische ingrepen.

3.4 Voorbereiding

De voorbereiding is een triviale stap en behelst het letterlijk ophalen en sorteren van pagina's met en zonder advertenties. Bovendien bevat het een tel stap ter voorbereiding aan het controleproces, dat parallel zal lopen met de verwerking. Het tellen van advertenties, c.q. vacatures, is nodig om door het gehele proces heen te weten hoeveel er is verwerkt en een schatting van de hoeveelheid werk dat nog moet worden gedaan te kunnen maken. Ook kunnen

er zo geen advertenties over het hoofd worden gezien of in ieder geval weer worden teruggevonden als dat wel mocht gebeuren.

3.5 Scannen

Het scannen van de krantenpagina's is het begin van het hele productieproces. De kwaliteit van de scans is doorslaggevend voor de rest van het proces. Binnen Matchcare Information Services was niets op het gebied van deze specifiek benodigde hardware aanwezig. Aanschaf van een geavanceerd apparaat ten behoeve van dit doorslaggevende onderdeel van het productieproces zal goed beargumenteerd moeten worden. Om een dergelijke argumentatie tot stand te laten komen, is het noodzakelijk om enig inzicht te verkrijgen in de manier waarop het gehele proces zal gaan draaien, aangevuld met de verwachte resultaten hiervan. De keuze voor de aan te schaffen machine is derhalve uitgesteld tot het moment waarop voldoende resultaten beschikbaar zullen zijn. Echter, het verkrijgen van inzicht in de manier waarop het proces zal gaan draaien, evenals in de kwaliteit van de scans, wordt danig bemoeilijkt door de afwezigheid van een scanner. Bovenstaand probleem is opgelost doordat Het Limburgs Dagblad (Regio-I[1] [4]) de kranten niet snel genoeg in Groningen kon krijgen om verwerking voor de uitlevering te kunnen garanderen. De oplossing die voorgesteld is, is dat men bij het Limburgs Dagblad de pagina's ging in scannen en digitaal versturen richting Groningen. Het feit dat de techniek om goede scans te produceren wel bij Regio-I aanwezig was, komt duidelijk naar voren als hun website bekeken wordt (www.Regio-I.nl). In eerste instantie kon dus worden gewerkt met een duidelijk af te scheiden hoeveelheid werk, namelijk het Limburgs Dagblad. Als met deze krant en de verdere verwerking de kwaliteit en (wel zo belangrijk) de kwantiteit gehaald wordt, kan een duidelijk advies worden uitgebracht om zelf een scanner aan te schaffen. De precieze keuze van welke scanner moet worden aangeschaft, komt later in dit verslag aan bod.

3.6 Tekst zone definiëren

Vanwege het feit dat alleen tekstvelden op dit moment van belang zijn, zullen deze dus uit de krantenpagina geëxtraheerd moeten worden. Met behulp van een imaging tool zal van elke advertentie op de pagina aangegeven moeten worden wat de tekst is (zone definiëren, segmentatie). Deze blokken zullen worden gevoed aan de OCR. De opzet om alleen de tekstblokken te definiëren en deze door de OCR te halen is mede bepaald door de keuze van het OCR pakket. Dit is een interessant en uitdagend deel van het hele proces; hier wordt dan ook uitvoerig op ingegaan.

---

[1] Regio-I : internetactiviteiten bundeling van drie regionale dagblad uitgeverijen (www.Regio-I.nl).

3.7 OCR

Met betrekking tot het OCR pakket is ervoor gekozen om in eerste instantie een commercieel product te nemen, zodat er gebruik kan worden gemaakt van de kennis die bedrijven al hebben op het gebied van OCR. Om zelf een OCR-applicatie te ontwikkelen schiet voorbij aan het doel van dit project. Gekozen is voor het standaard pakket '*cuneiForm99*' [3]. De technische beschrijving is te vinden in bijlage A. De term OCR is nu al een aantal keer gevallen, maar wat houdt dit precies in? Deze vraag wordt beantwoord in Hoofdstuk 4.

3.8 Tekstmining

In eerste instantie is er voor gekozen om een proces neer te zetten waarbij het typen weg valt en het clicken (drag and drop) er voor in de plaats komt. Als dit proces staat, kan er gekeken worden naar werkelijke tekstmining. [8]. In Hoofdstuk 6 komt dit onderdeel aan bod.

3.9 Invoeren in database

De database bestaat reeds en dient gevuld te worden op dezelfde manier als bij het fulltext intypen, echter nu automatisch. Uitleveringen worden ook gedaan op basis van deze database. Dit brengt met zich mee dat de ingevoerde gegevens geen fouten mogen bevatten, omdat zulks in directe relatie staat met klanten. De uitleveringen vinden plaats op verschillende manieren waaronder ftp, E-mail, CD-ROM. Momenteel wordt gebruik gemaakt van een statische database, maar er wordt gewerkt aan een nieuw model dat dynamisch van opzet is. Hier zal dan ook rekening mee gehouden moeten worden.

Hoofdstuk 4 Scannen en OCR

Zoals reeds is vermeld, wordt vanuit Limburg in eerste instantie gebruik gemaakt van digitale aanvoer. Echter, vanwege het succes met het prototype van de applicatie en gestimuleerd door ontwikkeling in dit automatiseringstraject is er budget vrijgemaakt voor de aanschaf van een professionele scanner. Dit hoofdstuk zal ingaan op de keuze die is gemaakt, en de specificaties van de aangeschafte scanner.

4.1 Inventarisatie

Na intensief contact met de mensen in Limburg is gebleken dat het vinden van een goede, specifiek voor dit doeleind geschikte scanner niet gemakkelijk is. De eisen op een rijtje:

- De scanner moet A3 formaat kunnen scannen. Dit vanwege het feit dat een gemiddelde krantenpagina dit formaat heeft.
- De scanner moet snel zijn. Dit is een relatief begrip, maar houdt in feite in dat het scannen op zich niet de rest van het proces mag ophouden.
- Triviaal is dat de kwaliteit van de gescande pagina optimaal moet zijn voor OCR.

Uit ervaring bij Matchcare en bij regio-i is gebleken dat een flatbed scanner het beste resultaat geeft, qua invoer.

4.2 Aanschaf

Zoals in de inleiding al is verteld, is er vanwege het succes van het prototype overgegaan tot aanschaf van een professionele scanner. De technische specificaties van de gekochte scanner zijn te vinden in bijlage B.

4.3 OCR introductie

Het acroniem OCR staat voor 'Optical Character Recognition'. Deze cryptische term betekend dat een computer het (vervelende) werk van overtypen van documenten overneemt. Of in andere woorden: door het gebruik van een scanner en speciale OCR software kun je een computer leren lezen.

De eerste stap in alle OCR processen is de optische scan van de tekst met behulp van een scanner. Tijdens dit proces wordt de tekst gedeeld in miljoenen image punten, welke elk een grijswaarde toegekend krijgen. De resolutie van de scanner wordt gemeten in dpi (dots per inch), dit bepaalt in hoeveel beeldpunten een pagina wordt verdeeld. Het resultaat van het scanproces is een plaatje dat met een imaging tool geprint of bewerkt kan worden. Wanneer je de tekst van het plaatje wilt lezen in een wordprocessor moet je een OCR proces toepassen. Welke techniek intern in de OCR applicatie gebruikt wordt, heeft geen invloed; Het probleem blijft bestaan dat op elkaar lijkende karakters wel eens verkeerd geïnterpreteerd kunnen worden. Als voorbeeld de i (van Isaac) en de l (van Leo). Als er bij het scannen bijvoorbeeld een vuiltje op de pagina is bij de i en wel net onder de punt, dan lijkt het karakter wel heel veel op een l. Dit is een groot probleem en hangt sterk af van de resolutie van het plaatje evenals de kwaliteit van het origineel. Hier dient dus serieus rekening mee gehouden te worden, omdat na het OCR proces een extra stap (namelijk OCR fouten corrigeren) niet echt gewenst is. Dit in verband met het kwantiteitsaspect van het gehele project. Een woordenboek of automatische spellingscontrole kan hier eventueel uitkomst bieden.

In het bijzondere geval van verwerking van vacaturetekst is er ook nog de complexiteit van de advertentie lay-out. Deze bestaat namelijk niet alleen uit enkele of meerdere kolommen met tekst, maar ook uit logo's, grafieken achtergrondfiguren en dergelijke. Een ander aspect is de verwerking van het gehele document, wat inhoudt dat ook meta-informatie bewaard dient te worden. Meta-informatie kan bijvoorbeeld de lay-out van een advertentie zijn.

Aanschaf van een goed OCR pakket en de daarbij behorende Developers Kit is een dure zaak. Uitstel op dit gebied, maar wel kunnen aantonen dat het productie proces zoals het is uitgelegd goed werkt, is wel noodzakelijk. Hieronder zal ingegaan worden op de keus voor het Cuneiform pakket.

4.4 Cuneiform OCR pakket

De keuze voor dit OCR pakket is gebaseerd op het feit dat het pakket de

```
[Documents]
Document0=C:\Dagblad De Limburger\01-12-01-Dagblad De Limburger-34-12345-0.rtf
Document1=C:\Dagblad De Limburger\01-12-01-Dagblad De Limburger-34-123-0.rtf
Count=2

[C:\Dagblad De Limburger\01-12-01-Dagblad De Limburger-34-12345-0.rtf]
Item0=C:\Program Files\LimCut\pics\begin0.tif
Item1=C:\Program Files\LimCut\pics\vaccode0.tif
Item2=C:\Program Files\LimCut\pics\vaccode1.tif
Item3=C:\Program Files\LimCut\pics\aanvullend0.tif
Item4=C:\Dagblad De Limburger\01-12-01-Dagblad De Limburger-34-12345-0.tif
Item5=C:\Program Files\LimCut\pics\aanvullend1.tif
Item6=C:\Program Files\LimCut\pics\begin1.tif
Count=7

[C:\Dagblad De Limburger\01-12-01-Dagblad De Limburger-34-123-0.rtf]
Item0=C:\Program Files\limcut\pics\begin0.tif
Item1=C:\Program Files\limcut\pics\vaccode0.tif
Item2=C:\Program Files\limcut\pics\vaccode1.tif
Item3=C:\Program Files\limcut\pics\solinfo0.tif
Item4=C:\Dagblad De Limburger\01-12-01-Dagblad De Limburger-34-123-0.tif
Item5=C:\Program Files\limcut\pics\solinfo1.tif
```

Voorbeeld bestand met OCR acties **Figuur 4-1**

mogelijkheid biedt om een aantal documenten achter elkaar door te verwerken. Bovendien is de opmaak voor het bestand waarin de acties staan beschreven heel duidelijk (zie Figuur 4-1). Natuurlijk speelt de kwaliteit van de OCR een aanzienlijke rol en is het belangrijk dat het pakket verschillende formaten plaatjes aankan.

Hoofdstuk 5 Tekst zone definiëring

In dit hoofdstuk wordt uitgelegd, hoe de tekstzones in eerste instantie zullen worden geëxtraheerd, en hoe dit in de toekomst gaat gebeuren. Onder een tekstzone wordt verstaan: een stuk bij elkaar horende tekst uit een advertentie, bijvoorbeeld een alinea of een titel.

5.1 Bestaande technieken

Bestaande technieken in tekst extractie vanuit images vallen in twee categorieën uiteen, namelijk: Bottom-up en Top-down. Een duidelijk voorbeeld van een top-down benadering is het Run-Length Smoothing Algoritme (RLSA) (het originele algoritme ([19]) bewerkt het plaatje met 4 operatoren.). Een duidelijk voorbeeld van de Bottom-up benadering is het Connected Components (CC) algoritme.

Combinaties van bottom-up en top-down benaderingen worden eveneens gebruikt. Vaak zien we een combinatie van RLSA om het document te 'bevlekken' en CC om blokken te extraheren van het document. [15], [20] en [21].

5.2 RLSA

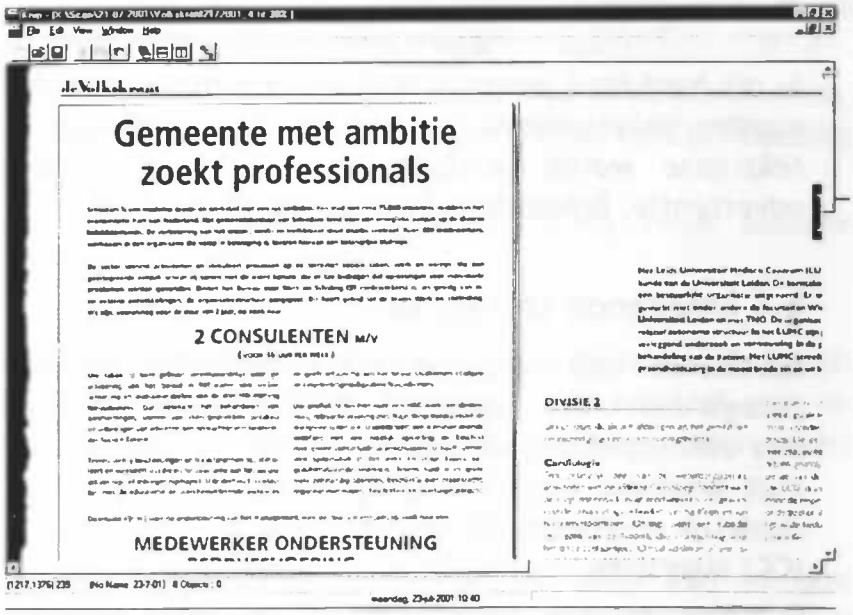
In het RLSA bestaat het segmentatieproces uit het proberen van het maken van continue stromen van pixels door twee zwarte pixels, die niet verder als een bepaalde threshold uit elkaar liggen, te 'mergen'. Deze methode zal eerst rij voor rij en daarna kolom voor kolom toegepast worden, resulterend in twee verschillende bitmaps (plaatjes) Deze twee plaatjes worden door middel van een 'pixelwise' 'OR'-operatie gecombineerd tot één bitmap bestaande uit alleen zwarte en witte pixels. [15] [16] De zo verkregen bitmap heeft een vlek van zwarte pixels waar geprint materiaal (tekst, plaatje of lijn) is gevonden op het origineel. Het moeilijke aan deze methode is het vinden van de juiste horizontale en verticale threshold. Meestal worden deze empirisch bepaald.

5.3 CC

In dit algoritme worden de pixels van een plaatje gecombineerd tot rechthoekige blokken (componenten). Daarna probeert dit algoritme herhaaldelijk om kleinere blokken te combineren tot grotere. In [19] is een voorbeeld van een implementatie van dit algoritme te vinden.

5.4 Imaging tool

In de eerste plaats wordt een imaging tool ontwikkeld die de handmatige tekstzone definiering ondersteunt.
Bovendien dient er rekenschap gehouden te worden met het feit dat dit in de toekomst ge-automatiseerd zal worden. Dit houdt in dat de applicatie flexibel in opzet moet zijn. In Figuur 5-1 is een voorbeeld gegeven van een geladen kranten-pagina van de Volkskrant. Er is ingezoomd zodat er een redelijk deel van één advertentie zichtbaar is. In Figuur 5-2 is te zien hoe verschillende tekst zones zijn gedefinieerd en hoe deze door middel van connectielijnen verbonden zijn, zodat het duidelijk is dat deze blokken bij elkaar horen.

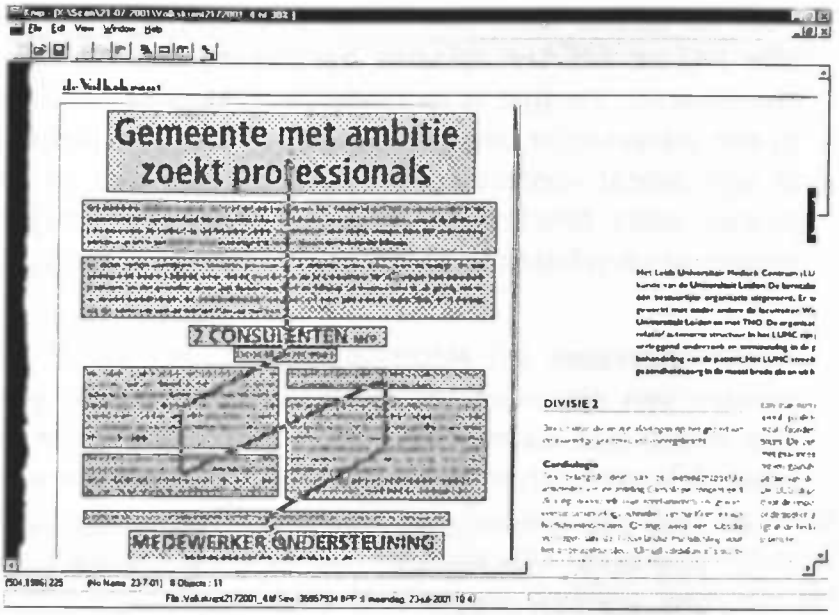


Bron: de Volkskrant 21-07-2001 pagina T11

Figuur 5-1

De keuze om blokken met tekst te selecteren (en niet in één keer de volledige advertentie) heeft een aantal voordelen. Ten eerste het voordeel dat als een logo zich in het midden van de advertentie bevind, er handig omheen gewerkt kan worden. Een logo mag danwel een kenmerk zijn om een advertentie te verwerken, het bevat geen tekstuele informatie. Mocht in de toekomst het logo wel belangrijk worden, wat niet onwaarschijnlijk is als wij in ons achterhoofd houden dat Matchcare Information Services in de toekomst meer informatie wil uitleveren dan tot nu toe, dan kan dit ook geselecteerd worden en wel apart. Een logo van een bepaald bedrijf kan zeer interessante informatie zijn.

Ten tweede kan er per zone worden aangegeven waar de tekst over gaat. Zo hanteert Matchcare een opdeling van een vacature tekst in 6 blokken die mutueel exlusief zijn. Deze kennis is in de loop van de jaren opgebouwd.



Bron: de Volkskrant 21-07-2001 pagina T11

Figuur 5-2

Momenteel wordt deze handeling ondersteund door de gebruiker de mogelijkheid te geven om te kiezen welke informatie een zone bevat. Het onderscheid dat wordt gemaakt tussen logo en tekst kan later worden gebruikt. Deze twee soorten blokken zullen toch afzonderlijk moeten worden behandeld, omdat hier sprake is van verschillende soorten informatie, en dus van verschillende methodes van extractie van deze informatie.

5.4.1 Fuzzy Spatial relations (SR)

Een mogelijkheid om te classificeren welke informatie in welk blok staat is met behulp van SR. Eerder werk heeft aangetoond dat een goede opzet van de SR tot goede resultaten zal leiden. [14] SR staat in het algemeen voor het redeneren over entiteiten die een bepaalde ruimte innemen. De positie van een object wordt met behulp van voorzetsels gekarakteriseerd aan de hand van andere objecten in de ruimte. Vaak wordt van één referentie object gebruikt gemaakt, maar soms worden verschillende objecten gebruikt (bijvoorbeeld het voorzetsel 'tussen'). Er bestaan twee soorten expressies om relaties aan te geven tussen objecten; statisch (bijvoorbeeld op, tussen) en dynamisch (bijvoorbeeld rechts van, richting...). Statische ruimtelijke relaties worden gebruikt om het originele plaatje zonder verlies van informatie weer te geven. Bij dynamische relaties worden van objecten bewegingen gezocht en gevolgd in verschillende plaatjes.

Er zijn twee soorten relaties te onderscheiden:

- Relaties die vergelijkingen aangeven tussen de eigenschappen van objecten (b.v. donkerder, lichter, groter, aantal zwarte beeldpunten ...);
- Relaties die de relatieve positie aangeven van objecten (boven, onder, naast ...).

Voor de laatste soort kan een uitputtende lijst worden gemaakt, in die zin dat alle andere soorten relaties beschreven kunnen worden gebruik makend van combinaties. De lijst is te vinden in [21]. Voor de classificatie van tekstblokken in een advertentie begeven we ons in een 2D ruimte. Deze restrictie elimineert al een aantal voorzetsels om relaties mee aan te geven. Na het segmentatie proces zoals beschreven in 5.4, kunnen ruimtelijke relaties worden gelegd tussen de verschillende objecten, en kan tot classificatie worden overgegaan.

Intern onderzoek bij Matchcare heeft een aantal feiten opgeleverd over de opbouw van een advertentie. [23]. In het ideale geval bevat een advertentie alle informatie categorieën zoals Matchcare deze heeft gedefinieerd. In de praktijk komen echter twee soorten uitzonderingen voor:

- Niet alle informatie categorieën zijn vertegenwoordigd in de advertentie. Het blijkt dat bij 10% van de onderzochte advertenties één of meerder categorieën missen;
- Soms (15%) blijkt dat een categorie gesplitst is in twee blokken waartussen een andere categorie opduikt;

Bij hoge uitzondering blijkt dat de categorieën niet zo mutueel exclusief zijn als verondersteld. Deze uitzondering is zo klein dat deze voor de rest van de verwerking nauwelijks een rol gaat spelen. Deze advertenties zullen apart worden behandeld.

De samensteller van de advertentie besteedt er veel aandacht aan om alle informatie zo overzichtelijk mogelijk aan de consument te presenteren. Door deze aandacht valt een redelijke uitspraak te doen over de opbouw van een advertentie. De categorieën zoals Matchcare deze gedefinieerd heeft komen in 75% van de gevallen in een vaste volgorde voor in de advertenties. Als we de categorieën in volgorde iets veranderen, bepaalde categorieën splitsen over meerdere blokken of een blok weglaten dan zien we dat er nog steeds een vaste volgorde te ontdekken valt. Bij een verdeling van 5 blokken vallen 83% van deze advertenties in een bepaalde volgorde. Bij een verdeling van 6 blokken vallen 90% van de advertenties in een bepaalde volgorde. Bij een verdeling van 7 blokken vallen 67% van de advertenties in een bepaalde volgorde.

Deze cijfers geven ons de mogelijkheid om relaties duidelijk aan te geven. Als we eenmaal een functie titel geclassificeerd hebben dan valt de rest ook te classificeren.

Hoofdstuk 6 Tekstmining

In dit hoofdstuk wordt ingegaan op het onderdeel TextMining. Het eenmaal hebben van de tekst, en de grove beschrijving (categorie) is niet voldoende voor Matchcare. De categorieën verder opsplitsen in kleinere stukken tekst en het extraheren van relevante informatie vanuit deze blokken is noodzakelijk. Intern onderzoek heeft op dit gebied al het een en ander uitgewezen. [8]

6.1 Introductie Tekstmining

Tekstmining is gedefinieerd als: "Het proces van extraheren van niet triviale patronen of kennis uit tekstuele documenten." Deze definitie is analoog aan de definitie van datamining, door vervanging van 'tekstuele documenten' door 'grote hoeveelheden data'. Belangrijk in tekstmining is het feit dat tekstmining een proces is, i.e. tekstmining is niet een magische techniek, maar bestaat uit een compleet proces van verzamelen van informatie, voorbewerken van tekst, mining van de voorbewerkte tekst en het resultaat op een slimme manier gebruiken.

1. Verzamelen → 2. Voorbewerken → 3. Mining → 4. Weergave

Bij TM is de voorbewerking stap van wezenlijk belang.

6.2 Gebruik van Tekstmining

Wetenschappelijk en commercieel onderzoek naar tekstmining heeft de complexiteit van deze systemen aangetoond. De meest succesvolle en meest efficiënte benadering blijkt een rule-based benadering te zijn. Deze benadering is echter tijdrovend en complex. [8]

Tekstmining oplossingen zijn sterk domein afhankelijk. Dit betekent dat de gebruikte Knowledge base - en eventueel door gebruikers gedefinieerde

woordenboeken - moeten worden aangepast voor het verwerken van de verschillende concepten en talen.

6.3 SR versus TM

Het grote verschil tussen tekst mining en Spatial relations ligt in het feit dat Spatial Reasoning plaatsvindt voordat inhoudelijk iets bekend is over de tekst of het logo(het blok). Daar SR alleen over blokken tekst een uitspraak kan doen bestaat altijd een noodzaak naar een oplossing die op woordniveau uitspraken kan doen. Hier komt tekstmining in beeld. Tekstmining bevindt zich, zoals duidelijk mag zijn, na het OCR proces.

Hoofdstuk 7 Een woord ter afsluiting

In de afgelopen jaren hebben zich er bij Matchcare Information Services aanzienlijke veranderingen voltrokken. Aanvankelijk verkeerde het bedrijf in een situatie waarin de inzet van nieuw personeel en meer machines werd gezien als de enige manier om op korte termijn meer vacatures te kunnen verwerken. Maar, hoewel niet altijd vermijdbaar, meer is niet altijd sneller of beter; daarom is voorzichtig een weg van continue verbetering ingeslagen. Het, van oorsprong handmatige, productieproces is geleidelijk vervangen door (semi-) automatische processen.

Het vervangen van het overtypen van vacatures door een semi-automatisch intelligent scan systeem lijkt makkelijker dan het is. Maar de bewandelde weg is niet altijd even duidelijk en efficiënt geweest. De ontwikkeling heeft een aantal fasen gekend, die nogal intuïtief zijn doorlopen als ad-hoc automatiseringstrajecten. Maar ondanks dat visie en inspiratie veelvuldig de kop opstaken, heeft de noodzaak tot handhaving van een gezonde cash-flow altijd de schoenmaker bij de leest gehouden.

Vanuit een historisch perspectief is het logisch dat het handmatig "inkloppen" vervangen is door een OCR applicatie. Maar daarmee staat nog niet alles klaar voor de database. Allereerst is een tussenapplicatie ontworpen waarmee de output van de inleesfase kan worden opgedeeld in 6 tekstblokken. Deze tussenapplicatie diende er tevens voor om de verschillende stappen in dit proces inzichtelijk te maken voor het personeel dat met deze applicatie werkt. Met de analyseapplicatie worden deze tekstblokken verder verwerkt om tot het eindproduct, een verrijkte digitale vacature, te komen. Vervolgens wordt van deze vacatures een samenvatting gemaakt.

Inmiddels is de functionaliteit van de tussenapplicatie verdeeld over de analyse en de beeldbewerking applicatie; daarmee is deze programmatuur na een kort leven overbodig geworden. Daarnaast kan tegelijkertijd een samenvatting van de fulltext vacature worden gemaakt. Op dit moment worden de tekstblokken

automatisch op het plaatje gesegmenteerd en is hiervoor geen manuele actie meer nodig. De eerste stappen richting automatische classificatie zijn gezet, maar momenteel vindt dit nog niet plaats.

De productiviteit van de fulltext afdeling lag voor de invoering van de beschreven stappen op gemiddeld 5 vacatures per uur. Dit aantal is vergroot naar 8,5 vacatures per uur, waarbij bovendien tegelijkertijd een samenvatting wordt geproduceerd. Hierbij dient te worden opgemerkt dat de exacte productieverbetering waarschijnlijk nog hoger ligt, maar moeilijk te meten is in verband met wijzigingen op meerdere punten binnen het systeem. Wel staat vast dat de verhoging van de gemiddelde verwerking, bijna vanzelfsprekend, ook een aanzienlijke kostenbesparing oplevert. Daarnaast worden eveneens opbrengsten uit de gecreëerde samenvattingen gegenereerd.

Het beschreven automatiseringstraject is geïmplementeerd in de organisatie en toont als huidige werkwijze. Tevens kan deze situatie worden gezien als startsein voor nieuwe ontwikkelingen.
Aandachtspunt is ten eerste het beter bepalen van de standaard lay-out van advertenties. Hiervoor moet data worden verzameld waarop statistische berekeningen kunnen worden losgelaten. Een te onderzoeken optie voor classificatie van blokken is het gebruik van templates.
Een tweede punt is het classificeren van blokken uit de advertentie door middel van Spatial Relations. Ten behoeve hiervan bevat de huidige beeldbewerking applicatie voldoende uitbreidingsmogelijkheden om hierin te voorzien.
Het verdient ten derde de aandacht om textmining toe te passen op advertenties. Textmining toepassen over een hele advertentie heeft niet de gewenste resultaten geleverd. Echter door textmining te zien in combinatie met reeds geclassificeerde blokken, wordt het zoekdomein beter afgebakend. Het inzetten van een combinatie van reeds geclassificeerde blokken tekst, waarmee het zoek domein beter afgebakend is, en textmining is tevens een optie voor verder onderzoek. Het grote voordeel van betere afbakening van het domein is de mogelijkheid om nauwkeuriger te kunnen zoeken naar cruciale woorden in het betreffende stuk tekst.

Vanuit een situatie waarin mensen werden gezien als enige oplossing voor het uitbreiden van de productieomvang, is een situatie gecreëerd waarin de weg is geopend voor vooruitstrevende ontwikkelingen in een nog onontgonnen werkgebied.

Literatuur

[1] W. Amerongen, "Prijsval geeft data warehousing enorme impuls", in "Informatie Management", juni/juli 1997.

[2] D.B. Baarda and M.P.M. de Goede, "Methoden en Technieken" 1998.

[3] Cognitive Enterprises web site http://www.ocr.com

[4] Regio-I : http://www.Regio-I.nl

[5] Infosys Scanners: http://www.infosys-scanner.de

[6] http://www.ocr-systeme.de/englisch/ocrallg.htm

[7] MatchNews. Nr 2. 2001

[8] M.Wilhelm. "PressAnalyzer", Afstudeerverslag Rijksuniversiteit Groningen, Groningen, Nederland, 2001.

[9] T. Lohman. "Marktonderzoek naar vacatures in de gedrukte media", Afstudeerverslag Hanze Hogeschool Groningen, Groningen, Nederland 2000.

[10] A. Jain, B. Yu, "*Document Representation and Its Application to Page Decomposition*", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 3, maart 1998.

[11] P. Visser. "Afstudeerverslag", Afstudeerverslag Rijksuniversiteit Groningen, Groningen, Nederland, 2001.

[12] C.L. Tan, B. Yuan, and Ch.H. Ang. "Agent-based text extraction from pyramid images", International conference on Advances in pattern recognition, Plymouth, UK, 23-25 NOV 1998.

[13] S.M. Smith and J.M. Brady. "SUSAN - a new approach to low level image processing." Int. Journal of Computer Vision, 23(1):45--78, May 1997.

[14] E.Kuiper en R.Wieringa. "Fuzzy Spatial Relations for Document Layout Analysis", Afstudeerverslag Rijksuniversiteit Groningen, Groningen, Nederland, 1999.

[15] M.Sharpe, G. Sutcliffe and N. Ahmed, "Implementation of an intelligent Document Understanding and reproduction system", Internal paper James

Cook University of North Queensland, Department of computer science Townsville, 1994.

[16] P. Chauvet, J. Lopez-Krahe, E. Taflin and H. Maitre, "*System for intelligent office document analysis, recognition and description* ". Signal Processing, Vol. 32 pp.161-190,1993.

[17] D. Olivier and B. Dominique. "A robust and multiscale document image segmentation for block/text line structure extraction." Twelfth international conference on pattern recognition, Jerusalem, pp 306-309, 1994

[18] F.M. Wahl, K.Y. Wong and R.G. Casey. "Block segmentation and text extraction in mixxed text/image documents." Computer Graphics and Image Processing, 20:375-390, 1982.

[19] K.Y. Wong, R.G. Casey, F.M. Wahl. "Document Analysis System." IBM Journal of Research Development 26(6), 647-656, 1982.

[20] K. Jain and B. Yu, "Document Representation and its Application to Page Decomposition," IEEE Transaction on pattern Analysis and Machine Intelligence, Vol. 20(3), pp 294-308, March 1998.

[21] J. Freeman, "the Modeling of Spatial Relations," Computer Graphics and Image Processing, Vol. 4, 1975.

[22] D.Niyoga and S.N. Srihari, "An integrated approach to document Decomposition and structural analysis" State university of New York.

[23] E. Bouwman. Interne communicatie. 2002.

Lijst met figuren

Bijlagen

A CuneiForm`99 Release Notes

Overview

CuneiForm`99 is an Optical Character Recognition (OCR) application designed for Microsoft Windows 95 /98 and Windows NT 4.0 (Intel platform only) or later. With CuneiForm`99 you can convert your printed documents into text files. CuneiForm`99 supports:

- A wide variety of scanners (See Supported Scanners section below for the complete list of supported scanners)
- Direct access to Microsoft Exchange Client to retrieve and recognize fax messages directly from the Inbox (This feature is not supported in Windows NT)
- Interface to Windows shell, which makes recognition of stored images as easy as a mouse click
- Batch recognition and scanning
- Communication with Microsoft Word to upload recognized document directly into your favorite word processor.
- Shortcut bar which gives you simple and fast access to all CuneiForm`99 modules
- CuneiForm`99 Editor to create and edit RTF and HTML documents
- All document types: laser printed, fax, dot matrix, proportional or mono-spaced
- West and East European languages: English, German, French, Italian, Spanish, Portuguese, Dutch, Danish, Swedish, Russian, Ukrainian, Serbian and Croatian (some language add-ons available for an additional charge)
- Text formatting: font face, font styles, alignment, multicolumn text, tables, etc

- Adaptive threshold scan, which automatically adjusts scanner brightness for enhanced recognition capabilities
- Built-in spell checker
- Intel Pentium processor with MMX for faster recognition and image processing

System requirements

- Personal computer with Intel 486DX4/100MHz or higher processor (Pentium 100MHz or higher recommended) running Microsoft Windows 95/98 or Windows NT 4.0 operating system or later.
- 32 MB of memory (64 MB recommended).
- 10 MB of free disk space required for installation and 20 MB recommended for images, text documents and temporary files created during scanning and recognition.
- Scanner is not required but recommended to scan printed documents (See Supported Scanners section below for the complete list of supported scanners).
- VGA or higher-resolution monitor (Super-VGA recommended).
- A Microsoft Mouse or compatible pointing device.
- Internet access and Microsoft Explorer 3.0 or latter or Netscape Navigator 3.0 or later are recommended to access Cognitive Enterprises. Web site http://www.ocr.com for product updates, technical support, online registration, etc.

B InfoSys Scanner Product Information

Technical Data	sirius	vega
Type	Flatbed	Flatbed
Technology	CCD line sensor	CCD line sensor
Grey Scale (internal)	256	256
Document size	420 x 600 mm (DIN A2)	420 x 600 mm (DIN A2)
Resolution	200 or 300 dpi	300 dpi
Scanning Time	6s (200dpi, DIN A2) 8s (300dpi, DIN A2)	8s (300dpi, DIN A2)
Throughput	200 pages per hour at 300dpi, DIN A2	180 pages per hour at 300dpi, DIN A2
Data Specification	bitonal	256 Grey-level
Interface	SCSI 2	SCSI 2
Power Source	115V, 300VA, 50-60Hz 230V, 300VA, 50Hz	115V, 300VA, 50-60Hz 230V, 300VA, 50Hz
Environmental Conditions	15°C to 35°C 20 to 80% rel. humidity (not condensing)	15°C to 35°C 20 to 80% rel. humidity (not condensing)
Dimensions	1.500 x 700 x 420 mm	1.500 x 700 x 420 mm
Weight	ca. 70 kg	ca. 70 kg
Certification according to	CE, UL*, CSA*	CE, UL*, CSA*
Life Expectancy	1.000.000 scans	1.000.000 scans

Technical specifications subject to change

* in preparation

For both scanners, **s i r i u s** and **v e g a**, we offer an option for scanning magazines up to DIN A4 (210 x 297 mm) and up to 450 pages. This enables the user to easily switch from scanning newspapers to scanning magazines.

For more information please contact us.