HO
GENT

Faculteit Bedrijf en Organisatie

Automatische transformatie van ingescande tabellen naar gestructureerde digitale data

Milad Nazari

Scriptie voorgedragen tot het bekomen van de graad van
professionele bachelor in de toegepaste informatica

Promotor:
Martijn Saelens
Co-promotor:
Bram Vandewalle

Instelling: Into Care by Predictive NV

Academiejaar: 2019-2020

Derde examenperiode

Faculteit Bedrijf en Organisatie

Automatische transformatie van ingescande tabellen naar gestructureerde digitale data

Milad Nazari

Scriptie voorgedragen tot het bekomen van de graad van
professionele bachelor in de toegepaste informatica

Promotor:
Martijn Saelens
Co-promotor:
Bram Vandewalle

Instelling: Into Care by Predictive NV

Academiejaar: 2019-2020

Derde examenperiode

Woord vooraf

Ik zou graag meneer Vandewalle willen bedanken voor enerzijds deze bachelorproefonderwerp en anderzijds voor de inhoudelijke ondersteuning en hulp die hij aangeboden en gegeven heeft. Hiernaast wil ik eveneens meneer Saelens bedanken voor de feedback en opvolging van mijn bachelorproef.

Samenvatting

Alhoewel meer en meer processen wereldwijd volledig digitaal plaatsvinden, worden toch nog een grote deel van procedures en data opslag uitgevoerd op niet-digitale manieren. Zo krijgen de meeste mensen hun factures nog steeds per brief, kassatickets worden nog steeds afgedrukt op papier, notities nemen op papier blijft de populaire keuze hoewel er tal van notitie-apps bestaan, etc. Dit heeft tot gevolg dat essentiële data nog massaal op een niet-digitale media bewaard wordt, namelijk op papier.

Tot enkele jaren geleden was dit probleem niet zo beduidend maar nu meer digitale platformen voor dataverwerking gebruikt worden, is het omzetten van data op papier naar digitale data, m.a.w. het digitalisatieproces steeds belangrijker geworden.

Hierdoor werden tal van digitalisatiesofwareproducten ontwikkeld, zoals Abby FineReader en Adobe Acrobat Pro DC. Hoewel deze software producten veel features hebben, zijn ze betalend en closed source. Toch hebben enkele bedrijven enkele van hun digitalisatie oplossingen open source gemaakt, zoals Google met diens bekende OCR-software, Tesseract OCR, die door iedereen gebruikt kan worden om tekst in foto's om te zetten in tekstdata.

Inhoudsopgave

1 Inleiding 19

1.1 Probleemstelling 19

1.2 Onderzoeksvraag 22

1.3 Onderzoeksdoelstelling 22

1.4 Opzet van deze bachelorproef 22

2 Stand van zaken 25

2.1 Inleiding 26

2.2 Tabulair data 26

2.2.1 Layouts 26

2.2.2 Digitale representaties 26

2.3 Tabeldetectie 26

2.3.1 Regelgebaseerde technieken 26

2.3.2 Datagedreven technieken 26
2.3.3 Performantieberekening 26
2.4 Tabelanalyse 26
2.5 End-to-end-systemen 26
3 Methodologie 27
3.1 Systeemvereisten 27
3.1.1 Goals 27
3.1.2 Non-goals 27
3.1.3 Vereisten 27
3.2 Selectie technologieën 27
3.2.1 Programmeertaal 27
3.2.2 Interne tabelmodel 27
3.2.3 Tabeldetectie en Tabelstructuuranalyse 27
3.2.4 OCR 27
3.2.5 Back end server 27
3.2.6 Front end 27
3.3 Scoresysteem 27
4 Proof of concept 29
5 Resultaten 31
6 Optimalisatiemogelijkheden 33
6.1 Domeinkennis 33
6.2 Natural Language Processing 33

6.3 Anomaliedetectie 33

7 Conclusie 35

A Onderzoeksvoorstel 37

A.1 Introductie 37

A.2 State-of-the-art 38

A.3 Methodologie 38

A.4 Verwachte resultaten 39

A.5 Verwachte conclusies 39

Bibliografie 41

Lijst van figuren

1.1 Voorbeeld van een tabelafbeelding. Bron: support.microsoft.com 20
1.2 Voorbeeld medicatieschema. Bron: apotheektsjoen.be 21

Lijst van tabellen

Woordenlijst

OCR Optical Character Recognition, optische tekenherkenning, is de transformatie van afbeeldingtekst in bewerkbare, digitale tekst. 19, 20, 22

Acroniemen

GUI Graphical User Interface. 22

1. Inleiding

In deze sectie wordt de context en achtergrond rond deze bachelorproef meegedeeld. Alsook wordt de probleemstelling, de onderzoeksvragen en onderzoeksdoelstellingen uitgelegd. Daarbovenop wordt de opzet van de bachelorproef verduidelijkt.

1.1 Probleemstelling

Alhoewel meer en meer processen wereldwijd volledig digitaal plaatsvinden, worden toch nog een grote deel van procedures en data opslag uitgevoerd op niet-digitale manieren. Zo krijgen de meeste mensen hun factures nog steeds per brief. Volgens de Federale Overheidsdienst Economie (2019) blijft het verzenden of ontvangen van facturen op papier een zeer gangbare praktijk. Zo verstuurde 90 % van de bedrijven er en 97 % ontving er in 2017. Daarbovenop worden kassatickets nog steeds afgedrukt op papier, en notities nemen op papier blijft de populaire keuze hoewel er tal van notitie-apps bestaan. Deze voorbeelden tonen aan dat essentiële data nog massaal op een niet-digitale, en dus niet-automatisch verwerkbare media bewaard wordt, namelijk op papier.

Tot enkele jaren geleden was dit probleem niet zo beduidend maar nu meer digitale platformen voor dataverwerking gebruikt worden, is het omzetten van data op papier naar digitale data, m.a.w. het digitalisatieproces steeds belangrijker geworden.

Hierdoor werden tal van digitalisatiesofwareproducten ontwikkeld, zoals Abby FineReader en Adobe Acrobat Pro DC. Hoewel deze software producten veel features hebben, zoals OCR, tabelherkenning, formulierherkenning, etc, zijn ze betalend en closed source. Wat als gevolg heeft dat ze voor bedrijven een merkbare kost met zich meebrengen, naast een privacy- en veiligheidsrisico aangezien het om closed source software gaat.

Sommige bedrijven enkele van hun digitalisatie oplossingen open source gemaakt, zoals Google met diens bekende OCR-software, Tesseract OCR, die door iedereen gebruikt kan worden om tekst in foto's om te zetten in tekstdata. Hoewel OCR op zich zeer belangrijk is voor digitalisatie, is het niet voldoende voor volledige digitalisatie. Zo kan men de relatie tussen verschillende documententiteiten, die normaal gezien grafisch wordt verduidelijkt, enkel met OCR digitaal niet overbrengen. In documenten worden relaties tussen woorden meestal a.d.h.v. een tabel verduidelijkt. Door gebruik te maken van OCR, verkrijgt men wel de tekst binnen een tabel, maar men verliest essentiële informatie rond de woorden, namelijk tot welke rij en kolom ze behoorden. Het valt tenslotte niet onder de verantwoordelijkheid van OCR-engines om naast tekstherkenning, ook nog tabeltransformatie uit te voeren.

	A	B	C	D
1	Product	Kw 1	Kw 2	Eindtotaal
2	Chocolade	€ 744,60	€ 162,56	€ 907,16
3	Gummibarchen	€ 5.079,60	€ 1.249,20	€ 6.328,80
4	Scottish Longbreads	€ 1.267,50	€ 1.062,50	€ 2.330,00
5	Sir Rodney's Scones	€ 1.418,00	€ 756,00	€ 2.174,00
6	Tarte au sucre	€ 4.728,00	€ 4.547,92	€ 9.275,92
7	Chocoladekoekjes	€ 943,89	€ 349,60	€ 1.293,49
8	Totaal	€ 14.181,59	€ 8.127,78	€ 22.309,37

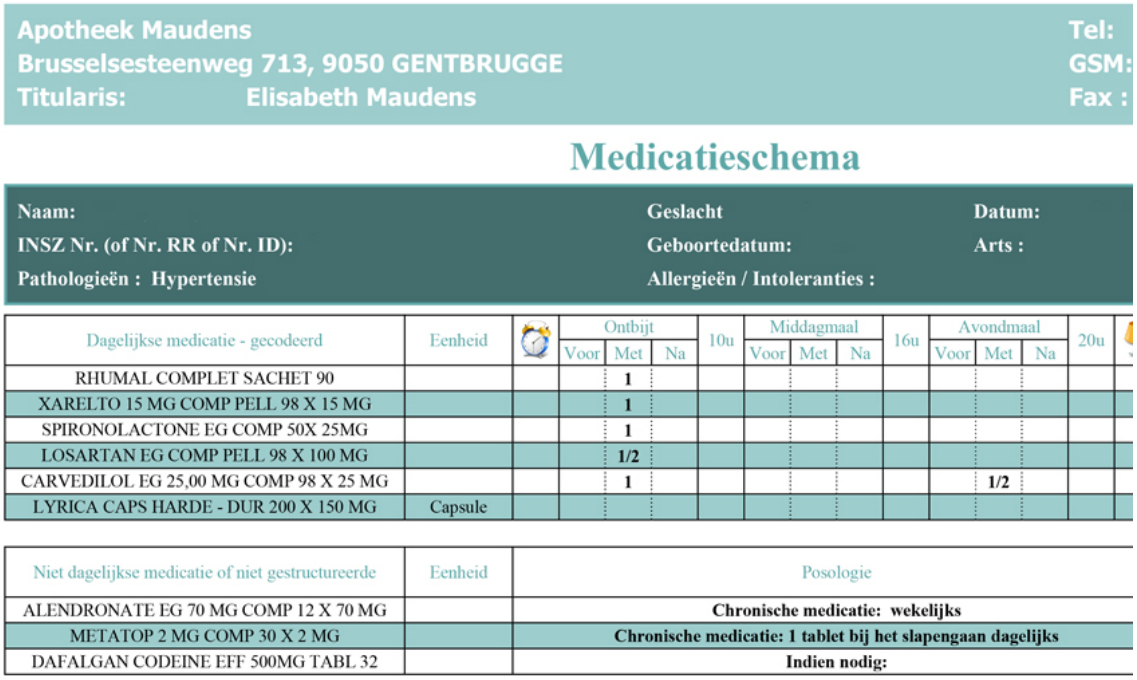
Figuur 1.1: Voorbeeld van een tabelafbeelding. Bron: support.microsoft.com

Indien men bij tabelafbeelding 1.1 enkel OCR voor digitalisatie zou gebruiken, dan verkrijgt men wel de tekst, zoals de tekststukken zoals "Kw 1", "Kw 2", "€744,60", "€ 162,56", en meer, maar men behoudt niet de relatie tussen de tekststukken. Hierdoor zal men enkel met OCR niet te weten komen of de verkoopbedrag van € 744,60 bij de eerste kwartaal behoort, of bij de tweede, wat essentiële informatie is voor verdere financïele analyse.

Tot heden bestaat er geen open source oplossing die tabellen in foto's transformeert naar digitale tabellen, m.a.w. naar digitale structuren waarbij de tekst, evenals de relatie tussen de verschillende teksten getransformeerd wordt. Daarom werd er voor deze bachelorproef beslist om een proof-of-concept van een tabeltransformatiesoftware te creëren die bij een foto automatisch tabellen detecteert en deze tabellen digitaliseert.

Een belangrijke professionele toepassing van digitale tabeltransformatie is het digitaliseren van ingescande medicatieschema's, door technologiebdrijven zoals Into.care die zich bezig houden met digitale gezondheidszorg. Medicatieschema's worden in de gezonheidszorg gebruikt om medicatiedata voor patiënten te bewaren en weer te geven. Volgens de definitie van Apothekersnetwerk (Apothekersnetwerk, 2013) is het medicatieschema een geheel van gestandaardiseerde informatie over de actieve medicatie van een patient, met inbegrip van de identiteit van de geneesmiddelen, hun dosering, indicatie, relevante gebruiksaanwijzingen en bijkomende informatie waar nodig. Het omvat zowel voorgeschreven als niet-voorgeschreven geneesmiddelen en voedingssupplementen.

Deze oplijsting van de actieve medicatie van de patient is niet enkel een essentieel hulpmiddel voor de patient bij de correct inname van medicatie maar ook voor medische professionelen om bv. over- of onderdosering, dubbelmedicatie, en andere geneesmiddelgebonden problemen te voorkomen. Ook wordt het gebruikt bij de communicatie tussen zorgverstrekkers. Het medicatieschema wordt eveneens door verpleegsters geraadpleegd voor het klaarzetten van de medicatie.



Apotheek Maudens Tel:
Brusselsesteenweg 713, 9050 GENTBRUGGE GSM:
Titularis: Elisabeth Maudens Fax :

Medicatieschema

Naam: **Geslacht** **Datum:**
INSZ Nr. (of Nr. RR of Nr. ID): **Geboortedatum:** **Arts :**
Pathologieën : Hypertensie **Allergieën / Intoleranties :**

Dagelijkse medicatie - gecodeerd	Eenheid	Ontbijt Voor	Ontbijt Met	Ontbijt Na	10u	Middagmaal Voor	Middagmaal Met	Middagmaal Na	16u	Avondmaal Voor	Avondmaal Met	Avondmaal Na	20u
RHUMAL COMPLET SACHET 90			1										
XARELTO 15 MG COMP PELL 98 X 15 MG			1										
SPIRONOLACTONE EG COMP 50X 25MG			1										
LOSARTAN EG COMP PELL 98 X 100 MG			1/2										
CARVEDILOL EG 25,00 MG COMP 98 X 25 MG			1								1/2		
LYRICA CAPS HARDE - DUR 200 X 150 MG	Capsule												

Niet dagelijkse medicatie of niet gestructureerde	Eenheid	Posologie
ALENDRONATE EG 70 MG COMP 12 X 70 MG		**Chronische medicatie: wekelijks**
METATOP 2 MG COMP 30 X 2 MG		**Chronische medicatie: 1 tablet bij het slapengaan dagelijks**
DAFALGAN CODEINE EFF 500MG TABL 32		**Indien nodig:**

Figuur 1.2: Voorbeeld medicatieschema. Bron: apotheektsjoen.be

Zoals men in figuur 1.2 kan zien, wordt dit schema grafisch in tabulaire vorm gepresenteerd. Echter is de lay-out hiervan niet gestandaardiseerd; afhankelijk van de apotheker of andere zorgverstrekker worden andere kolomnamen, kolomverdeling, rand- en verdelingstijl, celgrootte en andere tabelelementen aangewend. Dit bemoeilijkt ernstig het ontwikkelen van een transformatiesysteem die ingescande medicatieschema's omzet in instanties van een uniform digitale datastructuur in bv. XML- of JSON-formaat voor digitale verwerking van de medicatiedata in gezondheidszorgplatformen.

Een open source tabeltransformatiesoftware zal automatisch medicatieschema's kunnen omzetten in een uniform digitale datastructuur. Hierdoor zal er geen manuele werk uitgevoerd moeten worden, wat tijd- en kostenreductie als positieve gevolgd heeft. Daarbovenop, omdat het open source zal zijn, zal men verzekerd zijn dat Into.care niet zal te maken hebben met softwarelicentiekosten of privacyschending.

Hoewel het digitaliseren van medicatieschema's een belangrijke toepassing is, zijn er tal van andere potentiële toepassingen, aangezien tabellen zo vaak gebruikt worden. Zo zou men tabeltransformatie eveneens kunnen gebruiken voor het inscannen van kassatickets, het analyseren van een sudokuspel, het digitaal weergeven van een - op een whiteboard gemarkeerde - matrix voor online leerplatformen, het verwerken van een foto van een

voedingswaardetabel op de verpakking van voedsel, en meer. Het is duidelijk dat een open source tabeltransformatiesoftware een beduidende universeel meerwaarde zal aanbieden.

1.2 Onderzoeksvraag

Men kan zich bij tabeltransformatie, en dus bij dit onderzoek, enkele vragen stellen.

- Uit welke processen bestaat tabeltransformatie? In welke volgorde deze plaats?
- Hoe kan men de performantie van tabeltransformatiesoftware best evalueren?
- Is preprocessing van de afbeelding nodig om de nauwkeurigheid van de resultaten te bewaren? Indien ja, uit welke stappen bestaat deze preprocessing?
- Analoog, is postprocessing van de verkregen tabel noodzakelijk? Indien ja, uit welke stappen bestaat deze postprocessing?
- Op welke manieren kan men de resultaten verbeteren, indien men in bezit is van domeinkennis? Zo zou men bijvoorbeeld kennis van de gezondheidszorg kunnnen gebruiken om medicatieschema's nauwkeuriger te digtaliseren.

1.3 Onderzoeksdoelstelling

Aangezien het doel van deze studie het creëren van een end-to-end tabeltransformatie-tool is, zal er niet alleen gestreefd worden subprocessen zoals OCR of preprocessing geïsoleerd te bestuderen maar evenwel de subprocessen te implementeren in code. Eveneens is het de bedoeling dat de componenten met elkaar op een geïntegreerde manier zullen kunnen functioneren.

Dit betekent dat de prototype niet enkel zal bestaan uit tabelanalysesoftware, maar alsook uit een Graphical User Interface (GUI), een backend server, een preprocessing pipeline, en meer.

1.4 Opzet van deze bachelorproef

De rest van deze bachelorproef is als volgt opgebouwd:

In Hoofdstuk 2 wordt een overzicht gegeven van de stand van zaken binnen het onderzoeksdomein, op basis van een literatuurstudie.

Verder wordt in Hoofdstuk 3 de methodologie toegelicht en worden de gebruikte onderzoekstechnieken besproken om een antwoord te kunnen formuleren op de onderzoeksvragen.

In Hoofdstuk 4 wordt vervolgens de architectuur van de proof of concept uitgelegd. Eveneens worden de verschillende algoritmen in detail besproken.

Verder worden in Hoofdstuk 5 de met de proof of concept verkregen resultaten besproken en vergeleken.

In Hoofdstuk 6 worden enkele optimalisatiemogelijkheden om de nauwkeurigheid van het systeem te verhoden, besproken.

En tenslotten in Hoofdstuk 7, wordt de conclusie gegeven en een antwoord geformuleerd op de onderzoeksvragen. Daarbij wordt ook een aanzet gegeven voor toekomstig onderzoek binnen dit domein.

2. Stand van zaken

Dit hoofdstuk bevat je literatuurstudie. De inhoud gaat verder op de inleiding, maar zal het onderwerp van de bachelorproef *diepgaand* uitspitten. De bedoeling is dat de lezer na lezing van dit hoofdstuk helemaal op de hoogte is van de huidige stand van zaken (state-of-the-art) in het onderzoeksdomein. Iemand die niet vertrouwd is met het onderwerp, weet nu voldoende om de rest van het verhaal te kunnen volgen, zonder dat die er nog andere informatie moet over opzoeken (Pollefliet, 2011).

Je verwijst bij elke bewering die je doet, vakterm die je introduceert, enz. naar je bronnen. In LaTeX kan dat met het commando `\textcite{}` of `\autocite{}`. Als argument van het commando geef je de "sleutel" van een "record" in een bibliografische databank in het BibLaTeX-formaat (een tekstbestand). Als je expliciet naar de auteur verwijst in de zin, gebruik je `\textcite{}`. Soms wil je de auteur niet expliciet vernoemen, dan gebruik je `\autocite{}`. In de volgende paragraaf een voorbeeld van elk.

Knuth (1998) schreef een van de standaardwerken over sorteer- en zoekalgoritmen. Experten zijn het erover eens dat cloud computing een interessante opportuniteit vormen, zowel voor gebruikers als voor dienstverleners op vlak van informatietechnologie (Creeger, 2009).

2.1 Inleiding

2.2 Tabulair data

2.2.1 Layouts

2.2.2 Digitale representaties

2.3 Tabeldetectie

2.3.1 Regelgebaseerde technieken

2.3.2 Datagedreven technieken

2.3.3 Performantieberekening

2.4 Tabelanalyse

2.5 End-to-end-systemen

3. Methodologie

3.1 Systeemvereisten

3.1.1 Goals

3.1.2 Non-goals

3.1.3 Vereisten

3.2 Selectie technologieën

3.2.1 Programmeertaal

3.2.2 Interne tabelmodel

3.2.3 Tabeldetectie en Tabelstructuuranalyse

3.2.4 OCR

3.2.5 Back end server

3.2.6 Front end

3.3 Scoresysteem

4. Proof of concept

5. Resultaten

6. Optimalisatiemogelijkheden

6.1 Domeinkennis

6.2 Natural Language Processing

6.3 Anomaliedetectie

7. Conclusie

A. Onderzoeksvoorstel

Het onderwerp van deze bachelorproef is gebaseerd op een onderzoeksvoorstel dat vooraf werd beoordeeld door de promotor. Dat voorstel is opgenomen in deze bijlage.

A.1 Introductie

Het medicatieschema is een geheel van gestandaardiseerde informatie over de actieve medicatie van een patiënt, met inbegrip van de identiteit van de geneesmiddelen, hun dosering, indicatie, relevante gebruiksaanwijzingen en bijkomende informatie waar nodig. Het omvat zowel voorgeschreven als niet-voorgeschreven geneesmiddelen en voedingssupplementen (Apothekersnetwerk, 2013).

Deze oplijsting van de actieve medicatie van de patiënt is niet enkel een essentieel hulpmiddel voor de patiënt bij de correct inname van medicatie maar ook voor medische professionelen om bv. over- of onderdosering, dubbelmedicatie, en andere geneesmiddelgebonden problemen te voorkomen. Ook wordt het gebruikt bij de communicatie tussen zorgverstrekkers. Het medicatieschema wordt eveneens door verpleegsters geraadpleegd voor het klaarzetten van de medicatie.

Dit schema wordt grafisch steeds in tabulaire vorm gepresenteerd. Echter is de lay-out hiervan niet gestandaardiseerd; afhankelijk van de apotheker of andere zorgverstrekker worden andere kolomnamen, kolomverdeling, rand- en verdelingstijl, celgrootte en andere tabelelementen aangewend. Dit bemoeilijkt ernstig het ontwikkelen van een transformatiesysteem die ingescande medicatieschema's omzet in instanties van een uniform digitale

datastructuur in bv. XML- of JSON-formaat voor digitale verwerking van de medicatiedata in gezondheidszorgplatformen.

Hierdoor is er een nood aan een digitalisatiesysteem die medicatieschema's van verschillende vormen en met verschillende lay-outs nauwkeurig omzet in corresponderende instanties van een uniforme datastructuurschema. Voor deze bachelorproef wordt gebruik gemaakt van het datastructuurschema van Into Care by Pridictiv NV. De doelstelling van dit onderzoek is het bestuderen van de mogelijkheden om een dergelijk systeem tot stand te brengen en het implementeren van een proof-of-concept van een optimale oplossing. De volgende onderzoeksvragen kunnen gesteld worden bij dit onderzoek:

- Wat zijn de structuren en de relaties tussen de entiteiten in tabulaire data?
- Wat zijn de uitdagingen en complicaties bij tabelherkenning en -analyse? Kan er meer complexiteit ondervonden worden bij medicatieschematabellen?
- Hoe kan de correctheid en nauwkeurigheid van de transformatie van een tabel geëvalueerd worden?
- Welke oplossingen bestaan er reeds voor tabelherkenning en/of tabelanalyse?
- Wat is de optimale oplossing voor medicatieschema's? Hoe kan deze bepaald worden?
- Hoe kan domeinkennis gebruikt worden om de oplossing te optimaliseren?

A.2 State-of-the-art

Verschillende oplossingen voor tabeldetectie zijn reeds beschikbaar:

- Vervormbare convolutionele neurale netwerken (Siddiqui e.a., 2018)
- Verticale en horizontale lijnendetectie (Gatos e.a., 2005)
- Naïve Bayes en documentstructuur (Li e.a., 2006)

Ook voor tabelanalyse zijn enkele oplossingen voorgesteld:

- Cellsegmentatie (Nazemi e.a., 2016)
- Fast CNN (Oliveira & Viana, 2017)
- Faster R-CNN (Schreiber e.a., 2017)
- Graafgebaseerde neurale netwerken (GNN's) (Qasim e.a., 2019)

A.3 Methodologie

Het uitvoeren van het onderzoek zal beginnen met het ontwerpen van een scoresysteem, ook wel een benchmarksysteem genoemd, waarbij de nauwkeurigheid, precisie, performantie en andere factoren van de tabelherkenningsoplossingen in rekening gebracht zullen worden.

Hiervoor zullen reeds bestaande geannoteerde, geanonimiseerde medicatieschemadatasets gebruikt worden.

Hierna zullen de verschillende oplossingen geïmplementeerd en tevens geëvalueerd worden a.d.h.v. de benchmarksysteem. De optimale oplossing zal op deze manier bepaald worden.

Verder zullen potentiële optimalisatieopportuniteiten bestudeerd worden, zowel algemene optimalisaties als optimatlisatimogelijkheden binnen een medisch-farmaceutisch context zoals anomaliedetectie van tijdstippen van medicatieinnamen.

A.4 Verwachte resultaten

Enerzijds bestaan er in tabellen relaties tussen kolommen en cellen, en relaties tussen cellen onderling die voorgesteld kunnen worden door grafen en anderzijds vertonen de verschillende lay-outs van tabellen een patroon die door het menselijke brein maar dus ook door diepe neurale netwerken zeer snel herkend kan worden. Er wordt daarom verwacht dat een graafgebaseerde Deep Learning-oplossing de best resultaten zal opleveren.

A.5 Verwachte conclusies

Aangezien zowel state-of-the-art algoritmen als reeds bestaande softwareimplementatie-oplossingen beschikbaar zijn, wordt er verwacht dat een performante proof-of-concept van een digitalisatiesysteem voor medicatiesystemen successvol gecreëerd zal worden. Eveneens wordt er verwacht dat domeinkennis de nauwkeurigheid van het systeem zal verhogen.

Bibliografie

Apothekersnetwerk, V. (2013, juli 27). *Standpunt medicatieschema*. https://vlaamsapothekersnetwerk.be/index.php/informatie/nieuws/8-berichten-van/54-van-standpunt-medicatieschema

Creeger, M. (2009). CTO Roundtable: Cloud Computing. *Communications of the ACM*, *52*(8), 50–56.

Federale Overheidsdienst Economie, M. e. E., K.M.O. (2019). *Barometer van de informatiemaatschappij (2019)* (onderzoeksrap.). Federale Overheidsdienst Economie, K.M.O., Middenstand en Energie.

Gatos, B., Danatsas, D., Pratikakis, I. & Perantonis, S. (2005). Automatic Table Detection in Document Images. https://doi.org/10.1007/11551188_67

Knuth, D. E. (1998). *The art of computer programming, volume 3: (2nd ed.) sorting and searching*. Redwood City, CA, USA, Addison Wesley Longman Publishing Co., Inc.

Li, J., Tang, J., Song, Q. & Xu, P. (2006). Table Detection from Plain Text Using Machine Learning and Document Structure (X. Zhou, J. Li, H. T. Shen, M. Kitsuregawa & Y. Zhang, Red.). In X. Zhou, J. Li, H. T. Shen, M. Kitsuregawa & Y. Zhang (Red.), *Frontiers of WWW Research and Development - APWeb 2006*, Berlin, Heidelberg, Springer Berlin Heidelberg.

Nazemi, A., Murray, I., Fernaando, C. & McMeekin, D. A. (2016). Converting Optically Scanned Regular or Irregular Tables to a Standardised Markup Format to Be Accessible to Vision-Impaired. *World Journal of Education*, *6*(5), p9–19. Verkregen 2019, van https://eric.ed.gov/?id=EJ1158245

Oliveira, D. A. B. & Viana, M. P. (2017). Fast CNN-Based Document Layout Analysis, In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. https://doi.org/10.1109/ICCVW.2017.142

Pollefliet, L. (2011). *Schrijven van verslag tot eindwerk: do's en don'ts*. Gent, Academia Press.

Qasim, S. R., Mahmood, H. & Shafait, F. (2019). *Rethinking Table Recognition using Graph Neural Networks*.

Schreiber, S., Agne, S., Wolf, I., Dengel, A. & Ahmed, S. (2017). DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images, In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. https://doi.org/10.1109/ICDAR.2017.192

Siddiqui, S., Malik, M. I., Agne, S., Dengel, A. & Ahmed, S. (2018). DeCNT: Deep Deformable CNN for Table Detection. *IEEE Access*, *PP*, 1–1. https://doi.org/10.1109/ACCESS.2018.2880211