Automatische transformatie van ingescande tabellen naar gestructureerde digitale data

Een proof-of-concept

Nazari Milad, Bram Vandewalle, Martijn Saelens
Hogeschool Gent, Valentin Vaerwyckweg 1, 9000 Gent
contact@miladnazari.be

Abstract

Met dit onderzoek wordt het ontwerp en ontwikkeling van een open source prototype-software bestudeerd die het mogelijk zou maken om tabellen in ingescande documenten te digitaliseren. Dit zou niet enkel digitaliseringsprocessen versnellen maar zou het eveneens gebruikt kunnen worden voor verschillende digitalisatietaken, zoals de digitalisatie van medicatieschema's. Er wordt eerst een literatuurstudie uitgevoerd om de stand van zaken rond tabeltransformatie te verduidelijken. Vervolgens worden systeemvereisten en non-goals voor de proof-of-concept gespecifieerd. Op basis hiervan worden de algoritmes voor tabeltransformatie en de technologieën geselectioneerd. Hierna wordt de proof-of-concept in detail toegelicht, bovendien werd een nieuwe algoritme voorgesteld voor structuuranalyse. Uiteindelijk wordt de software op een dertig tal afbeeldingen getest. Hoewel bij de test alle tabellen juist gedetecteerd worden, wordt de nauwkeurigheid van de tabeltransformatie verlaagd door de minder performante tabelstructuuranalyse. De voorgestelde nieuwe algoritme echter verhoogt de nauwkeurigheid van de structuuranalyse, al kan hierdoor de software nog niet als perfect beschouwd worden. Men kan bij dit onderzoek concluderen dat tabeltransformatie een complex domein is. Kant en klaar software-pakketten bestaan, maar zijn betalend en niet open source. Verder kan besloten worden dat tabelstructuuranalyse uit meerdere, van elkaar afhankelijke, subprocessen bestaat. Tenslotte kan men concluderen dat tabeldetectie enerzijds zeer nauwkeurig is, terwijl anderzijds structuuranalyse minder optimale resultaten kan leveren. De voorgestelde algoritme die een niet onbelangrijke verbetering van de tabeltransformaties teweegbrengt, toont aan dat optimalisatiemogelijkheden zeker nog mogelijk zijn.

Introductie

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 en bedrijven hun factures nog steeds per brief. 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. 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 besloten om een proof-of-concept van een tabeltransformatiesoftware te creëren die bij een foto automatisch tabellen detecteert en deze tabellen digitaliseert.

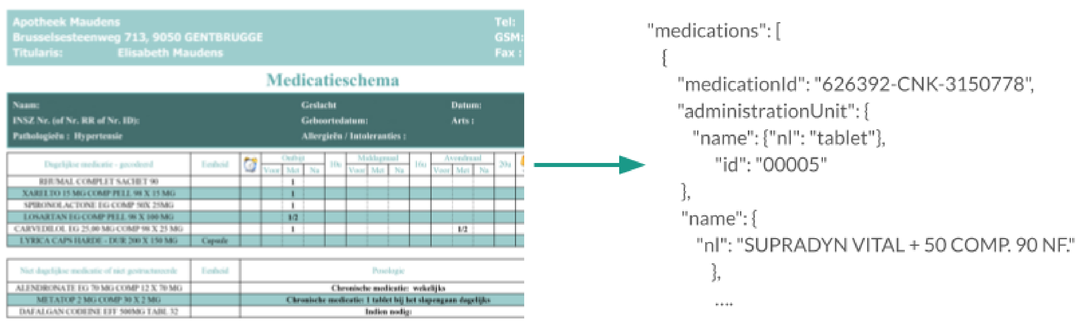


Figure 1: Tabeltransformatie

Een open source tabeltransformatiesoftware zal automatisch bijv. medicatieschema's kunnen omzetten in een uniform digitale datastructuur, zoals weergegeven in bovenstaande figuur. 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 niet te maken hebben met softwarelicentiekosten of privacyschendingrisico's. 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.

Proof-of-concept en resultaten

Zoals de volgende figuur het weergeeft, worden bij de proof-of-concept meerdere processen doorgelopen. Als eerst uploadt de gebruiker een afbeelding naar de transformatieserver d.m.v. de GUI. Vervolgens wordt de afbeelding verwerkt in de preprocessing-stap om de nauwkeurigheid van de tabeltransformatie te verhogen. Hierna worden de tabellen in de afbeelding gedetecteerd m.b.v. een diep neurale netwerk en geïsoleerd. Dan wordt elke geïsoleerd tabel getransformeerd in de structuuranalyse-stap waarin tekstherkenning, lijndetectie, cellsegmentatie en hiërarchische clustering plaatsvinden (voorgestelde algoritme). Een postprocessing-stap wordt eveneens uitgevoerd om de nauwkeurigheid verder te verhogen. Uiteindelijk worden de getransformeerde tabellen in JSON-formaat teruggestuurd naar de gebruiker.

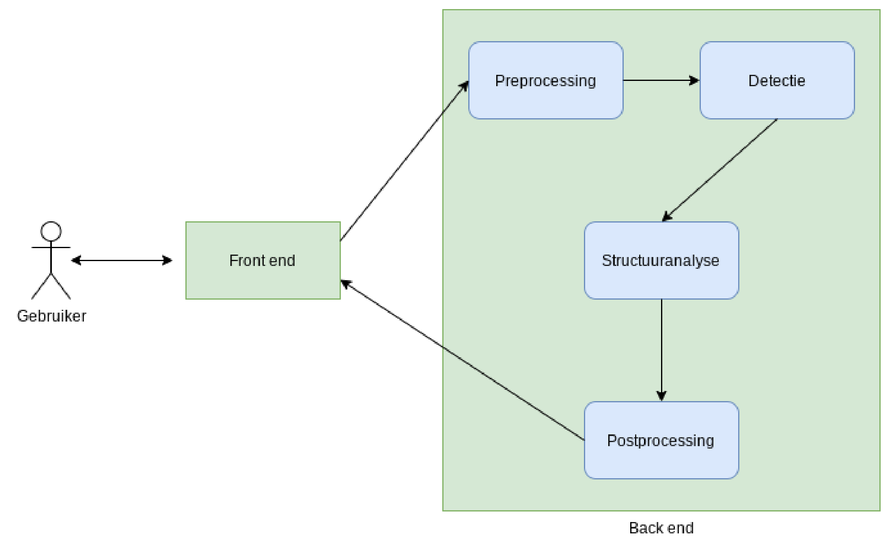
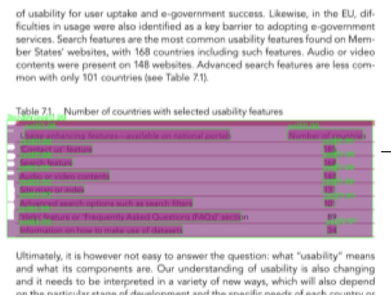


Figure 2: Architectuur van de proof-of-concept

Hoewel bij de performantietest alle tabellen juist gedetecteerd werden, werd de nauwkeurigheid van de tabeltransformatie verlaagt door de minder performante tabelstructuuranalyse. De voorgestelde algoritme echter verhoogde de nauwkeurigheid van de structuuranalyse, al kan hierdoor de software nog niet als perfect beschouwd worden.



	Usage-enhancing features—available on national portals	Number of countries
0	'Contact us' feature	185
1	Search feature	168
2	Audio or video contents	148
3	Site map or index	131
4	Advanced search options such as search filters	101
5	'Help' feature or 'Frequently Asked Questions ...	89
6	Information on how to make use of datasets	34

Figure 3: Voorbeeld tabeltransformatie d.m.v. proof-of-concept

Conclusies

Men kan bij dit onderzoek concluderen dat tabeltransformatie een complex domein is. Kant en klaar software-pakketten bestaan, maar zijn betalend en niet open source. Een open source versie bestaat momenteel niet. Verder kan besloten worden dat tabelstructuuranalyse niet een simpel eenvoudig proces is. Eerder is het een complex procedure die uit meerdere, van elkaar afhankelijke, subprocessen bestaat. Tenslotte kan men concluderen dat tabeldetectie enerzijds zeer nauwkeurig is, terwijl anderzijds structuuranalyse minder optimale resultaten kan leveren. De voorgestelde algoritme die een niet onbelangrijke verbetering van de tabeltransformaties teweegbrengt, toont aan dat optimalisatiemogelijkheden zeker nog mogelijk zijn.