Knipsel Info Service maakt stevige digitaliseringsslag door

DERTIG JAAR OUD PRIMAIR PROCES

Al meer dan dertig jaar verzorgt de in Almere gevestigde Knipsel Info Service de bekende knipselkranten voor abonnees. Miljoenen kranten en tijdschriften zijn in de loop der tijd gelezen, artikelen werden geselecteerd, uitgeknipt, opgeplakt, gefotokopieerd en gedistribueerd. Het bedrijf uit 1979 verschilt in vele opzichten flink van het bedrijf anno 2010. Het is vooral IT die dat mogelijk maakt. Zo wordt er software gebruikt die begrijpend kan lezen en artikelen op inhoud kan beoordelen.

Door Hans Lamboo

Op 1 oktober 2009 bestond de Knipsel Info Service 30 jaar. Een heugelijk feit dat werd gevierd samen met klanten en relaties en werd opgeluisterd door de aanwezigheid van mevrouw Jorritsma, de burgemeester van Almere. Een knappe prestatie om een bedrijf zo lang te laten bestaan en bovendien almaar te laten groeien. Algemeen directeur Mark Reisz heeft altijd goed grip gehouden op het proces en is tijdig begonnen met een transitie naar digitaal. “Alles begon handmatig,” vertelt hij. “We hadden een groep lezers die alle beschikbare kranten, huis-aan-huisbladen en tijdschriften lazen en de artikelen die voor abonnees interessant konden zijn selecteerden. Vervolgens werden die artikelen met de hand uitgeknipt en van een label met daarop de herkomst voorzien. We hadden een wand van ongeveer 20 meter lang vol met plastic bakjes, waar per klant alle knipsels in werden gelegd. Dat moest aan het eind van de dag allemaal vergaard en geteld worden – want er moest gefactureerd kunnen worden – en vervolgens in een envelop gedaan en naar het postkantoor gebracht. Dat was in een notendop het gehele proces. Heel arbeidsintensief, allemaal.”

Fifty-fifty

De hoeveelheid gedrukte media is de afgelopen jaren enorm gegroeid, wat een steeds zwaardere belasting voor het proces werd. “We verwerken nu ruim 2000 titels. Per dag betekent dat zo’n 15.000 (kranten)pagina’s; dat resulteert in een kleine 60.000 artikelen. Van internet worden 60.000 nieuwsitems geplukt, en van radio en TV wordt zo’n kleine 100 uur per dag opgenomen. We hebben landelijk 100 procent dekking van de gangbare kranten, tijdschriften en huis-aan-huisbladen – ook technische en vaktijdschriften,” zegt Reisz. “Dat kunnen we nu doen met 170 mensen en een hoge mate van gedigitaliseerd werk. In het begin konden we dat nog wel af met 60 mensen, maar de groei van het aanbod handmatig blijven volgen was op den duur niet meer mogelijk. Inzet van technische hulpmiddelen was onze enige kans.”

> Het knippen doen we in de PDF met speciale software

Vooral het lezen van elk artikel in de 2000 titels en dat beoordelen op relevantie voor één van de abonnees van de Knipsel Info Service is extreem arbeidsintensief. Dat beaamt Reisz. “Een groep lezers zat aan tafel alle publicaties te lezen. Elke lezer moest werkelijk alles uit het hoofd weten: welke abonnees hebben we, wat vindt elke individuele abonnee relevant.

Foto: Arjen van den Berg

Mark Reisz: “We onderzoeken nu de mogelijkheden van de Autonomy Sentiment Analysis software”.

We zaten op zo’n kleine 25.000 onderwerpen; de inwerktijd voor een lezer was dan ook langer dan één jaar.” Bovendien zijn de meeste lezers vrouwelijke parttimers die hooguit vier uur per dag het intensieve leeswerk willen komen doen. Het lezen werd de bottleneck. Reisz: “We realiseerden ons dat het zo niet langer kon. We konden op de oude, handmatige manier nooit meegroeien met het aanbod. Dus zijn we het proces nauwkeurig gaan bekijken om te zien hoe we de taak van de lezers konden verlichten. Inmiddels was het digitale tijdperk aangebroken en werd het met de komst van de Pentium III PC mogelijk artikelen te scannen en zo dus om te zetten in een digitaal bestand. We besloten om alle publicaties in te scannen en vervolgens door net op de markt gekomen OCR-software te laten ‘lezen’ en omzetten in een digitaal tekstbestand. Daar lieten we dan een zoekprogramma op los en alleen de als relevant aangemerkte artikelen, de ‘hits’, gingen naar de leesafdeling.” De invoering van deze digitaliseringsslag verliep aanvankelijk moeizaam, omdat de OCR (Optical Character Reading) software nog niet de kwaliteit had die het vandaag de dag heeft. De omzettingen van de gescande artikelen naar tekstbestanden wemelden dan ook van de fouten; ook de zoeksoftware werkte verre van optimaal. Dat samengevoegd, leverde in het begin tegenvallende resultaten. “Maar we kregen het proces steeds verder onder de knie, de software werd steeds beter en uiteindelijk verwerkten we alle titels op deze manier, in zwart-wit. Tegenwoordig doen we alles ook nog eens in kleur.”
Wat opvallenderwijs niet is veranderd door het digitaliseringsproces is de verhouding tussen het aantal lezers en andere medewerkers. Reisz: “Van de 60 mensen die we vroeger hadden zat ongeveer de helft te lezen; we hebben nu 170 mensen in dienst, maar nog steeds is de helft daarvan lezer. Het selecteren van de artikelen, het bepalen van de relevantie voor onze klanten is nu eenmaal mensenwerk. Wel kunnen de lezers veel grotere hoeveelheden artikelen uit een veel groter aanbod selecteren, ze zijn veel productiever geworden.” Dat laatste geldt ook voor de rest van het traject, hoewel het scannen en knippen nog steeds veel tijd kost. De winst is vooral te vinden in het gemakkelijk kunnen selecteren, distribueren en factureren per abonnee van de knipsels.
Het scannen gebeurt met hele (kranten)pagina’s tegelijk; de pagina’s worden omgezet in een PDF-bestand. “Het knippen doen we in de PDF met speciale software, Mimotek geheten. We participeren ook in de ontwikkeling van die software,” vertelt Reisz. “Veel kranten kunnen we zo automatisch knippen. Daarna kost het nog zo’n vijf minuten per knipsel om te zien of het allemaal goed gegaan is. Het lukt ons nu om ongeveer de helft automatisch te knippen.”

Zoeken in context

Ook probeert Reisz de uitgevers ertoe te bewegen meer artikelen rechtstreeks in PDF-vorm bij de Knipsel Info Service aan te leveren. “We hoeven niet meer te scannen en bovendien is de kwaliteit van de direct aangeleverde pdf veel beter. Ongeveer 20 procent van de uitgevers werkt daar nu aan mee, maar het gaat moeizaam. Het maken van een PDF valt bij veel redacties buiten de workflow.”
Knipsel Info Service heeft elk artikel dat sinds oktober 2006

in Nederland verscheen in haar database, die nu ongeveer 40 Terabyte (40.000 Gigabyte) aan data bevat. Die asset wilde Reisz ook beter exploiteren en hij moest daarvoor de gesprekken met de uitgevers intensiveren. Tijdens de gesprekken ontmoette Reisz een andere partij waarmee het bedrijf Profactys werd opgezet, dat e-papers en archiefdiensten aanbiedt waardoor een win/win-situatie met de uitgevers ontstaat.
Vooral het zoeken in de grote hoeveelheid heterogene artikelen baarde Reisz zorgen. Hij raakte gecharmeerd van de software van Autonomy. "Ik zag de kracht ervan. Autonomy IDOL (Intelligent Data Operating Layer) zou ons kunnen helpen de bulk aan data efficiënt te beheren en het veel gemakkelijker maken om er informatie uit te trekken," vertelt hij. "En allemaal met de 'Jip-en-Janneke methode': je toetst een paar woorden in en de software presenteert de hits. Dat klinkt als Google, maar IDOL gaat veel verder dan alleen keyword-search. Ook geavanceerde vormen zoals Meaning Based Retrieval (het automatisch begrijpen van de context van content) en pattern-search. Bovendien gaat de keyword-search van Autonomy vele malen verder dan het zoeken op een woord met een aantal tikfouten erin. De hits kunnen ook artikelen zijn waar de zoekterm niet eens in voorkomt, maar waarvan de software vindt dat het eraan gerelateerd is."
IDOL levert de gerelateerde artikelen geclusterd aan op hoofdonderwerp: de zoekterm 'weer' levert bijvoorbeeld clusters van gerelateerde artikelen over zonneschijn, regen, wind en temperatuur.

> De dienstverlening van Knipsel Info Service omvat inmiddels ook internet en RTV-content

De dienstverlening van Knipsel Info Service omvat inmiddels ook internet en RTV-content. Voor dat laatste werd het bedrijf ReporterService overgenomen. Daar wordt 24 uur per dag naar alle radio- en TV-programma's geluisterd en gekeken; men noteerde handmatig de onderwerpen en verzond alerts naar abonnees als zijzelf of hun branche in het nieuws werden genoemd. "Dat is natuurlijk veel te tijdrovend," stelt Reisz. "Ook daar zijn we gaan automatiseren met een combinatie van de Autonomy Virage-software en IDOL. De eerste zet gesproken tekst om in een digitaal tekstbestand, met Autonomy IDOL zoeken we naar de juiste fragmenten, die direct op tijdcode worden gepresenteerd. Dat scheelt heel veel zoekwerk. Ook hier komt de kracht van IDOL naar voren: Virage zet de teksten nooit helemaal foutloos om in tekstbestanden. Mensen spreken immers soms onduidelijk of met een dialect, of er is storend achtergrondgeluid. Autonomy IDOL kijkt naar de context en 'begrijpt' als het ware wat er gezegd is, ook als het niet helemaal correct is weergegeven. Het woord 'bos' in de context 'financiën' en 'balkenende' wordt door Autonomy begrepen als minister Bos, niet als een verzameling bomen."
Knipsel Info Service doorzoekt op internet alleen de bekende nieuwssites – en kijkt momenteel naar de mogelijkheden om ook *social media* zoals Twitter mee te nemen.
"We kunnen natuurlijk niet *alle* sites monitoren," stelt Reisz. "We hebben een selectie gemaakt. Behalve de nieuwssites van de kranten en bijvoorbeeld nu.nl, bekijken we ook een aantal toonaangevende consumentensites, waar mensen zelf vertellen hoe blij – of niet – ze zijn met een bepaald product of dienst. Dat is voor onze abonnees interessante informatie. We hebben nog steeds primair de leesafdeling, die de hits uit Autonomy beoordeelt op relevantie. Voor die mensen is de winst vooral dat het aantal *false hits* gigantisch is teruggebracht," zegt Reisz. "Toerisme is bijvoorbeeld een heel lastig onderwerp. Iedereen weet precies wat het is, maar het is vrijwel onmogelijk in een keyword-profiel te vatten. Juist met Conceptuele Search worden de zoekresultaten vele malen beter, worden de false hits geminimaliseerd. Daarnaast hebben we een aantal profilisten en taxonomiebouwers in dienst die de zoekprofielen voor Autonomy ontwikkelen en bouwen."
Nuttige bijkomstigheid van het gebruik van Autonomy is het terugvinden van in eerste instantie niet-geselecteerde artikelen, die vroeger dus in de prullenbak verdwenen.

Sentiment analyse

Reisz blijft zoeken naar groeimogelijkheden. "In Nederland kunnen we nog zoveel meer doen. Er zijn bijvoorbeeld nog heel veel tijdschriften die we niet hebben. Waar ik ook veel van verwacht zijn onze extra diensten zoals analyse, e-papers, archiefdiensten en het bedienen van de grote bedrijven. Er valt nog veel winst te halen bij onze bestaande klanten. Die zien de hoeveelheid informatie alleen maar toenemen en willen dat steeds beknopter zien. Uiteindelijk zal het in de richting gaan van media-analyse, dat er een rapportage wordt gemaakt van hoe bedrijf X of onderwerp Y in het nieuws is geweest vandaag, deze week of deze maand. Dus steeds meer *to-the-point*, steeds meer samengevat. Daar kunnen we nog veel verbeteren."
De behoefte aan informatie blijft altijd en zal alleen maar groeien. Wel verwacht Reisz dat het papieren aanbod zal afnemen en de focus meer richting het digitale traject zal verschuiven. "Alles moet snel en de zoekvragen worden steeds complexer. Daar zullen we in mee moeten gaan. Die stap denken we gezet te hebben met de implementatie in 2007 van Autonomy IDOL. We onderzoeken nu de mogelijkheden van de Autonomy Sentiment Analysis software, dat automatisch aangeeft of het nieuwsbericht een positieve of negatieve lading heeft. Dat is voor de klant ook weer een extra beknoptheid, omdat positieve berichten minder aandacht vragen dan negatieve – en snellere actie wellicht."

Hans Lamboo is hoofdredacteur van Business Process Magazine.