Formularerkennung, Freitexterkennung und Dokumenten-Management-Systeme

Im folgenden wollen wir auf drei wesentliche Elemente der elektronischen Dokumentenanalyse und -archivierung eingehen. Dabei beleuchten wir folgende Bereiche:

1. die Formularerkennung
2. die Freitexterkennung
3. das klassische Dokumenten-Management-System

Vorbemerkung

Warum unterscheiden wir?

In diesem Dokument unterscheiden wir bewusst zwischen Formularerkennung, Freitexterkennung und dem klassischen Dokumenten-Management-System (DMS), obwohl man auch alle drei Elemente unter dem Oberbegriff DMS zusammenfassen könnte. In der Praxis jedoch treffen diese drei Elemente bislang selten aufeinander. So benutzt z.B. ein Umfrage-Institut möglicherweise eine Formularerkennung, hält jedoch die eigentlichen Dokumente denen die Informationen entstammen nicht in Dateiform oder Volltext vor. Umgekehrt indiziert vielleicht eine Rechtsabteilung alle eingehenden Dokumente, ohne jemals Daten aus diesen in eine Datenbank übernommen zu haben.

Wesentliche Voraussetzungen, Definitionen

Was ist ein Dokument im Sinne dieser Abhandlung? Welche Hard- und Software wird benötigt?

Als Dokument wollen wir eine schriftliche, in unserem Betrachtungsfalle maschinenlesbare Information werten, die im wesentlichen den Merkmalen eines Briefes, einer Rechnung, Bestellung o.ä. entspricht. Diese liegt entweder in Papierform oder elektronisch vor.

Die Analyse von Dokumenten bedingt, dass ebendiese in einem vordefinierten Dateiformat vorliegen. D.h. Dokumente in Papierform müssen zunächst einen Umwandlungsprozess durchlaufen. Dafür ist ein geeigneter Stapelscanner notwendig, welcher die Dokumente in einem vordefinierten Verzeichnis im Netzwerk ablegt.

Eine OCR-Engine (OCR=Optical Character Recognition) kümmert sich anschließend um die Umwandlung in analysierbaren Text.

Auch elektronische Dokumente in nicht unterstützten Formaten werden zunächst umgewandelt, also etwa ein Bild in ein Textformat.

Hardwarevoraussetzungen

Welche Hardware muss bereitgestellt werden?

Benötigt wird ein Stapelscanner der an einem Scannarbeitsplatz (Computer mit WIN OS) betrieben wird. Computer und Scanner sind in das interne Netzwerk eingebunden. Die eigentliche Textanalyse sollte auf einem separaten Server durchgeführt werden.

Softwarevoraussetzungen

Welche Software kommt zum Einsatz?

Jeder Computer, welcher auf das DMS zugreifen soll benötigt einen entsprechenden Software-Client. Auf dem Server läuft eine Texterkennungs- (OCR-) Engine. Alternativ können die Client-Rechner über einen Terminalserver betrieben werden. Die Anzahl der gleichzeitigen Zugriffe auf das DMS richtet sich dann nach der Anzahl der auf dem Terminalserver vergebenen Lizenzen.

Texterkennung

Was verstehen wir unter Texterkennung? Welche Unterschiede gibt es?

Prinzipiell unterscheiden wir zwischen der klassischen *Formularerkennung*, also uns in seinem Layout bekannten, auf Vorlagen basierenden Dokumenten und der *Freitexterkennung* semistrukturierter Dokumente. Bei letzteren erwarten wir eine gewisse Struktur, so wie z.B. in einer Rechnung oder Bestellung eine gewisse Anzahl an Argumenten (Rechnungsnummer, Empfänger, Produktbezeichnung etc.) allgemein üblich ist. Hierbei ist das Layout des Dokuments und die genaue Wahl der zu identifizierenden Schlagwörter meist unbekannt. Beispielweise erscheint auf einer Rechnung der Bezeichner für "Rechnungsnummer" häufig anders: "ReNr", "Rechnung Nr", "Invoice #"...
Ebenso ist uns die Anordnung dieser Elemente unbekannt. Auch kann sowohl eine Rechnung als eine Reklamation z.B. eine Rechnungsnummer enthalten.

Formularerkennung

Erkennung vordefinierter Dokumente.

Die klassische Formularerkennung basiert auf Dokumentvorlagen, welche uns in Layout und Inhalt bekannt sind. Als bestes Beispiel mögen hier Fragebögen gelten, welche von Umfrage-Instituten in Umlauf gebracht werden. Die Texterkennung ist hier darauf ausgerichtet, an definierten Stellen eines Dokuments einen erwarten Ausdruck (z.B. Ja, Nein) auszuwerten. Die Schwierigkeit liegt hier eher in der Deutung der unterschiedlichen Handschriften und Abweichungen von definierten Dokumentbereichen. Die Fehlerquote kann jedoch immer dann erheblich minimiert werden, wenn es möglich ist, diese Dokumente weitestgehend zu personalisieren.

Freitexterkennung

Die Freitextanylyse semistrukturierter Dokumente.

Semistrukturierte Dokumente zeichnet im allgemeinen aus, dass diese nicht von der empfangenden, verarbeitenden Person designed wurden, es diese in großer Anzahl verschiedener Layouts gibt, welche von Absender zu Absender erheblich differieren. Sie alle eint, dass sie durch das Vorkommen bestimmter Inhalte definiert sind. Die traditionelle Formularerkennung greift hier nicht.

Diese Dokumente können nach hinterlegten Algorithmen inzwischen jedoch mit einer hohen Treffergenauigkeit analysiert werden. Dabei werden Entscheidungsbäume abgearbeitet an deren Ende eine Analyse-Wahrscheinlichkeit steht. Mit anderen Worten, bei diesem Prozess arbeitet man gegen eine Fehlerquote. Erreicht die Dokumentenanalyse eine definierte Fehlerquote, wird der nächste Entscheidungsbaum abgearbeitet usw.

Diese Entscheidungsbäume werden je Dokumenttyp hinterlegt und müssen mit jedem nicht sicher erkannten Dokument weiter gepflegt werden.

Das bedeutet, die Freitexterkennung sucht nach dem Vorhandensein definierter Argumente und entscheidet, um welchen Dokumententyp es sich handelt. Anschließend wird der eigentliche Inhalt des identifizierten Dokumententyps analysiert.
Dokumente, die einem bekannten Muster entsprechen, werden automatisch zugeordnet, andere müssen manuell bearbeitet werden.
Die Praxis zeigt, dass Erkennungssysteme bei Ihrer Initialisierung im Unternehmen bereits ca 60% der bearbeiteten typischen Dokumente mit einer Fehlerquote von 0% abarbeiten.

Dokumenten-Management-Systeme

Elektronisches Dokumentenmanagement.

Unter DMS verstehen wir die strukturierte Archivierung von Dokumenten, inklusive Zugriffsrechten, Historien.
Die in einem DMS gepflegten Dokumente sind im Volltext recherchierbar, unter Einbeziehung eines Thesaurus sogar nach Synonymen.
Die in einem DMS gepflegten Dokumente durchlaufen zunächst ebenfalls einen Texterkennungsalgorithmus, d.h. herkömmliche Dokumente in Papierform werden gescannt und in verwertbaren Text umgewandelt. Ebenso wird mit Dokumenten in nicht unterstützten Dateiformaten verfahren.
Anschließend wird der Inhalt des Dokuments für die Suche indiziert und sowohl in einer Datenbank als reiner Text, als auch auf dem Server als Datei abgelegt. In beiden Fällen durchläuft der Text bzw. das abzulegende Dokument einen Verschlüsselungsprozess, d.h. auch das als Datei auf dem Server abgelegte Dokument ist für einen Unbefugten nicht lesbar.

Der Einsatz in der Praxis

Wann lohnt sich der Einsatz von Texterkennungsprogrammen und DMS?

In der heutigen Gesellschaft, in der Software zum Kulturgut gehört, ist ein Einsatz derselben immer reizvoll. Die Entscheidung ein Texterkennungsprogramm zum Einsatz zu bringen, sollte gut durchdacht werden, denn solche Systeme gehören in der Softwarelandschaft zu den derzeit komplexesten und sind somit zunächst kostenintensiv.

Eine *Formularerkennung* muss auf Anwender abgestimmt werden, d.h. es handelt sich hier jeweils um ein Unikat, eine Einzelanfertigung. Ist diese einmal eingerichtet, arbeitet sie softwareseitig zuverlässig, das Fehlerpotenzial tendiert gegen Null.

Wir können hier also von einer hohen Einstiegsinvestition ausgehen, während die fortlaufende Pflege unproblematisch ist.

In einigen Unternehmen beschäftigen sich meist gut qualifizierte Angestellte überwiegend mit dem Erfassen, also oft abschreiben, von Daten. Die dafür aufgewendete Arbeitszeit wird durch eine Formularerkennung frei und kann anderweitig genutzt werden. Nicht selten amortisiert sich die Einstiegsinvestition binnen einem Jahr. Bei einer durchschnittlichen Lebensdauer einer Branchensoftware von 7 Jahren lässt sich das Einsparpotential leicht errechnen.

Die *Freitextanylyse* ist ein komplexer Prozess, welcher fortwährende Pflege erfordert. Die notwendigen Anpassungen können nur teilweise vom Anwender vorgenommen werden. Ändert ein Lieferant das Design der Rechnung, wird diese unter Umständen nicht mehr erkannt.

Hier wird der Einsatz von Software sinnvoll, wenn man sich auf bestimmte Dokumententypen beschränkt, also zum Beispiel nur Bestellungen und/oder Eingangsrechnungen automatisch bearbeitet. Prinzipiell ist es immer dann empfehlenswert, wenn Dokumente Daten enthalten, die für eine Weiterverarbeitung aufbereitet werden müssen. Von der Anzahl solcher Dokumente hängt die Rentabilität für das Unternehmen ab.

Auf das klassische *Dokumenten-Management-System (DMS)* kann in rentabel arbeitenden Unternehmen nicht mehr verzichtet werden. Ein DMS stellt Informationen zu jeder Zeit und, bei entsprechenden Voraussetzungen, an jedem Ort zur Verfügung. De facto arbeitet ohnehin beinahe jedes Unternehmen mit einer Vorstufe des DMS, indem wichtige Dokumente in Dateiform auf einem Serverlaufwerk vorgehalten und mehreren Anwendern angeboten werden. Das macht ein DMS auch, nur strukturierter, volltextindiziert, mit Sicherheitsstufen und Versionierung. DMS ist abrufbares Wissen und wer mit Wissen arbeitet, kann mit dem Einsatz eines Dokumenten-Management-Systems nichts falsch machen.