

# Intelligenz: Captured & Exchanged

## Kofax-Forum bietet marktgerechte Orientierung

Ein wohl unbestrittener Vorteil des Internet-Zeitalters ist die gewaltige Qualitätsverbesserung von Informationen durch die gemeinschaftliche Diskussion. Eine der besten Quellen für Informationen und Orientierungen im Technologiebereich sind ohne Frage die Peer-to-Peer-Diskussionsgruppen, Listserver und Forengruppen. Wer die Seiten von FaceBook, SalesForce.com oder Wikipedia besucht, erkennt schnell den Wert dieser Foren als Plattform, in der Annahmen in Frage gestellt und Erfahrungen ausgetauscht werden.

Im Capture-Sektor bietet Kofax ein hervorragendes Forum an (<http://forums.kofax.com>), das sich durch aktive Teilnahme, lebhaft Diskussionen und eine gut verwaltete Site auszeichnet. Es ist öffentlich zugänglich und lässt sich einfach nach bestimmten Fragen durchsuchen.

Der nachfolgende Meinungs austausch über INDICIUS (auch als Kofax Transformation Module bezeichnet), ist nur ein Beispiel. In zukünftigen Ausgaben von Inside Capture werden wir einige der interessantesten und aufschlussreichsten Postings aus den Kofax-Foren vorstellen. Sie wurden im Interesse der Prägnanz, Platzersparnis und Anonymität geringfügig überarbeitet.

## ICR mit Word-Spotting?

Besitzt INDICIUS (auch als Kofax Transformation Module bekannt) Funktionen für Word-Spotting/ Worterkennung zur Bewältigung wirklich schwieriger ICR-Aufgaben mit handschriftlichen Vorlagen? Ich nehme an, dass eine höhere Erkennungsrate in diesen Fällen zweitrangig ist und es vor allem auf das ICR-Ergebnis ankommt. Außerdem gehe ich davon aus, dass in einem gegebenen Volltextdokument bzw. Volltextformular der handschriftliche Text von ein- und derselben Person stammt.



Was ich mit „Word-Spotting“ meine: Die Idee ist, auf Wortebene statt buchstabenweise zu arbeiten. Man identifiziert ein Wort im Dokumentenbild, erstellt ein kleines Teilbild, ermittelt den zugehörigen Text und sucht dann im Dokumentenbild nach ähnlichen Teilbildern. Wird ein neues Wort (Teilbild) gefunden, dann wird einfach das bekannte Wort zur Textausgabe hinzugefügt.

Die Mustererkennung auf Wortbasis verspricht eine größere Genauigkeit als die Segmentierung, Identifizierung und Erkennung von Einzelbuchstaben in einem Wort.

Wer mit ICR in Kofax Capture experimentiert, kennt wohl den Einfluss auf die ICR-/OCR-Qualität bei handschriftlichem Text mit zu engem Buchstabenabstand (Buchstaben berühren sich) oder in Schreibschrift. Beim Word-Spotting braucht man sich keine Gedanken um die Segmentierung von Zeichen der schreibschriftähnlicher Buchstabenneigung zu machen. Zwischenräume bei Wörtern sind offenbar wahrscheinlicher als bei Buchstaben, und eine Schriftneigung dürfte viel weniger ins Gewicht fallen, weil das Wort mehr Musterinformationen besitzt als ein einzelnes Zeichen.

Ich bin kein wirklicher Fachmann dafür, glaube aber, dass ältere Handschrift generell mit zeichenweiser ICR schwerer zu handhaben ist und sich deshalb besser für Word-Spotting eignen würde.

Ich habe kein aktuelles Beispiel, würde aber gern Näheres dazu erfahren.

## Antwort 1:

Der Ansatz könnte dann sinnvoll sein, wenn ein strukturiertes oder halbstrukturiertes Hauptdokument („Main“) und ein Anhang („Att“) einer oder mehrerer handschriftlicher, möglicherweise schräg geschriebener Texte mit zusammengefügteten Zeichen und vielen Störungen auf jeder Seite vorliegt.

Das erste Dokument „Main“ ließe sich wahrscheinlich ohne große Probleme in Kofax Capture mit oder ohne INDICIUS verarbeiten, aber das „Att“-Dokument wird wahrscheinlich ein sehr schlechtes OCR-Texterkennungsergebnis liefern.

Man würde m.E. zu Produktionslinien mit einer „Hauptsoftware“ tendieren, deren Bediener dieselbe Software auf verschiedene Implementierungen und Dokumente anwenden. Somit ließe sich wahrscheinlich sowohl „Mail“ als auch „Att“ in derselben Produktionslinie mit jeweils derselben Software verarbeiten, statt eine eigene Lösung für „Att“ aufzubauen oder zu kaufen und dann die Ausgabe in die Produktionslinie für „Main“ zu integrieren.

Vielleicht wäre es machbar, einen benutzerdefinierten Job in INDICIUS mit APIs auf Bibliotheken von Fremdherstellern zu erstellen. Rein gefühlsmäßig halte ich das aber für riskant. Außerdem bräuchte man dazu wohl ein großes Entwicklungsprojekt, um sicherzustellen, dass die komplette Methode von denen verwendet wird, die Word-Spotting implementieren.

## Antwort 2:

Das hört sich nach einem Projekt an, das wirklich Spaß macht. Falls Sie einmal ein Projekt mit diesen Anforderungen haben, lassen Sie es mich wissen. Es würde mich sehr interessieren, wie Sie das angehen.

Ich habe ein ähnliches Problem. Es ging darum festzustellen, ob ein handschriftlicher Name auf zwei Seiten identisch ist. Die von uns eingesetzten ICR-Module lieferten fürchterliche Ergebnisse. Trotzdem erreichten wir das Ziel, weil wir einen Algorithmus verwendeten, um die Zeichen-Wort-Ähnlichkeit zwischen zwei Seiten zu ermitteln (und weil das ICR-Modul bei seinen Fehlern anhand desselben handschriftlichen Namens ziemlich konsistent war). Hier ein Beispiel:

Der handschriftliche Name John Smith wurde zu

Seite 1 - Lolii Snihh

Seite 2 - Iohi Smiih

## Antwort 3:

Sie sollten sich A2iA Field Reader ansehen, der kursive Handschriften beherrscht. Er liest den Text nicht zeichenweise, sondern nach Phonemen. Er hat eine benutzerdefinierte Modulintegration mit Kofax Capture.

Diesen Diskussionsfaden (und viele ähnliche) finden Sie unter: <http://forums.kofax.com>.