

d-Class

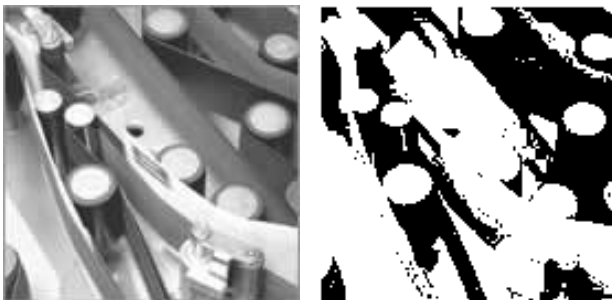
AUTOMATISCHE TEXTBASIERTE DOKUMENTENKLASSIFIKATION

Die Informationsflut rollt.

Unternehmen werden mit Informationen geradezu überschwemmt - aus allen Richtungen und schneller als je zuvor.

Anstatt Abmachungen per Handschlag gelten in unserer komplexen Welt immer vielfältigere und umfassendere Geschäftsdokumente: Anschreiben, Verträge, Anhänge, Geschäftsbedingungen, Sondervereinbarungen, Abrechnungen, Gutschriften und vieles mehr. Und dies in Form verschiedenster Medien: Briefpost, eMails, WEB-Informationen, Sprachnachrichten, Faxe, Bilder und Zeichnungen.

Die Folge: Posteingänge quellen über, Suchvorgänge im WWW, in Intranets und Archiven erbringen zu viele oder nicht relevante Treffer. Wer sucht, der findet nicht.



Informationen klassifizieren & strukturieren.

Daten sind nur so hilfreich wie der Zusammenhang, in dem sie stehen. Wer ihren Kontext vernachlässigt, der hat keine Grundlage für die digitalisierte Verarbeitung geschäftlicher Prozesse. Kontext ist der bestimmende Faktor.

Ein bedeutender Anteil der entscheidungsrelevanten Informationen liegt **in unstrukturierten Dokumenten** vor, hauptsächlich in Form von Texten. Schätzungen gehen von einem Anteil von rund 80% unstrukturierter Information in Unternehmen aus. Oder in Form von **Dokumentenbildern**, oftmals nur mit einem Minimum an Index-Information versehen.

Die mangelnde Informationsorganisation ist hier das Hauptproblem und das hat seinen Grund:

Die Masse der anfallenden Dokumente macht eine manuelle Strukturierung und Auswertung beider Informationsarten oft unzureichend und teuer, wenn nicht gar unmöglich. Das Management dieser Form der Information stellt alle Organisationen vor zunehmend größere Probleme. Deshalb geht heute kein Weg vorbei an **automatischer** Unterstützung durch Informationstechnik (IT).

Probleme der Informationsflut jetzt lösbar.

Schlüssel zur Problemlösung ist das Vorhandensein von Kategorien für hierarchische Informationsstrukturen, welche das Suchfeld reduzieren und relevante Informationen schneller zugänglich machen.

Wir bieten diese grundlegende Voraussetzung der automatischen Kategorisierung: Ob in **Archiven** oder beim **Posteingang**, ob als codierte Information (ASCII-Zeichen) oder als nicht codierte Information (Dokumenten-Images):

d-Class sortiert Dokumente vollautomatisch und sprachunabhängig nach vorgegebenen Kategorien. Das „Aschenputtel“ der Neuzeit.

DOCSTREAM-PRODUKTE

d-Class Textbasierte Klassifikation auf Basis codierter Zeichen zielt auf die Kategorisierung von unstrukturierten Textdokumenten, auf eMails oder ähnliche Informationsinhalte.

d-Class setzt dem PC quasi eine „Brille“ auf, mit der dieser erkennt, um welche „Sorte“ Dokument es sich handelt.

- Hochqualitative Brillen besorgt man sich beim Optiker und schleift sie nicht selbst. So wird auch der jeweilige Klassifikator beim Hersteller DocStream produziert – unter unserer Qualitätsverantwortung. Eine Sache von wenigen Tagen, nicht von Wochen.
- Doch einfachere „Kaufhausbrillen“ gibt es sofort: in Form eines Trainingstools D-TRAIN für den Anwender.

d-Class läuft als Batchverarbeitung auf jedem PC mit Windows XP. Angabe des Quellverzeichnisses und des Ausgabezeichnisses genügt. Pro Dokument wird die Kennung der ermittelten Dokumentenklasse (plus weniger wahrscheinlicher Alternativen) ausgegeben.

d-Class plus d-ImageSearch: Auch Texte aus Dokumentenimages, die als Ergebnis von Schriftenleseverfahren (OCR) gewonnen werden, sind erfolgreich klassifizierbar. Werden die OCR-Ergebnisse auch gleich noch dem hochkomprimierten Luratech-Farbbild unterlegt, so sind farbige Dokumentenimages nicht nur sehr kompakt, sondern auch durchsuchbar. Viewer ist der Acrobat Reader (V7.0).

Kommt es bei Dokumentenimages auf hohen Durchsatz an, so setzen d-Class und d-PDF-ImageSearch einen modernen, leistungsfähigen PC mit Windows 2000/ XP voraus.

Die Lizenzierung kann d-Class und d-PDF-ImageSearch nach Verbrauch erfolgen. Es gibt Cartridges verschiedener Größe, welche Seitenlizenzen enthalten.

DOKUMENT-KATEGORISIERUNG

Erster Schritt zum automatischen inhaltlichen Dokumentenverstehen ist die automatische Sortierung von Dokumenten in Gruppen mit bestimmten gemeinsamen Eigenschaften.

Aufgabenstellung der Dokumentenklassifikation ist also das Selektieren von Dokumenten aus einer Gesamtmenge und deren Verteilung in vorgegebene Untermengen („Töpfe“, Klassen oder Kategorien genannt).

Zielsetzung ist, weder alleine die zu 100% „deckungsgleichen“ Dokumente heran zu ziehen noch gar alleine diejenigen mit nur einem einzigen voll übereinstimmenden Begriff („Wortsuche“), sondern es soll idealerweise die Bandbreite relevanter Dokumente auf Grund der Gesamtheit ihrer Merkmale bewertet und mit bestmöglicher Wahrscheinlichkeit in Kategorien sortiert werden.

Der Natur des Verfahrens nach ist die Hauptzielrichtung **textbasierter** Klassifikation das Finden und Vergleichen von relevanten Schlüsselbegriffen / Textsequenzen aus ASCII-Zeichenstrings, deren Bewertung im Hinblick auf das Maß der Relevanz, einer bekannten Dokumentenklasse zugehören und die Zuordnung eines Dokuments zu bekannten Dokumentenklassen über den Vergleich mit bekannten Informationen aus der Belehrung.

Die **textbasierte Klassifikation** setzt dort an, wo die **Dokumenten-Layout-basierten** Analyseverfahren für strukturierte Dokumente scheitern.

Dokumentenklassen, die bei layout-basierten Verfahren auf Grund von Lesefehlern nicht einwandfrei kategorisiert werden, können nachfolgend zusätzlich der textbasierten Klassifikation unterzogen werden, um die Treffsicherheit zu steigern. Aber auch das umgekehrte Verfahren ist sinnvoll: Vorsortieren mit d-Class und geometriebasierte Detailsortierung. Eben je nach Anwendungsfall.

MERKMALE

Bei d-Class bedienen wir uns hauptsächlich hochwertiger statistischer Verfahren, die kontrollierbar bleiben und auch nachbelehrt werden können (also keine Verfahren auf Basis neuronaler Netze). Unser Anspruch ist es, höchst mögliche Klassifikationsqualität zu liefern.

Der Kern von d-Class basiert auf einem Verfahren, welches die relevanten Zeichengruppen, Worte oder Textsequenzen in ASCII-Strings identifiziert und in ihrer Gesamtheit mit dem gelernten Wissen vergleicht, Ähnlichkeiten feststellt und dem entsprechend klassifiziert. Je höher die Zahl vorgestellter und belehrter, repräsentativer Dokumente desto besser (Einstiegsschwelle sind 100 Belehrdokumente pro Klasse).

Hinsichtlich der Anzahl der Dokumentenkategorien ist eine Obergrenze in technischer Hinsicht nicht erkennbar; insbesondere nicht angesichts der in der Praxis geforderten, eher kleineren Anzahl an Dokumentenklassen.

d-Class ist sprachunabhängig und sowohl für originäres, weitgehend ungestörtes ASCII-Textmaterial (wie z. B. eMails) als auch in Verbindung mit störanfälligen Verfahren (wie z. B. durch OCR-Falschlesungen) einsetzbar.

Zielsetzung des neuen, bei DocStream entwickelten, textbasierten Dokumentenklassifikationsverfahrens ist, in vertikalen Märkten mit der Zeit deren Dokumente-„Weltwissen“ von Haus aus mitzubringen.

D-CLASS IST PFEILSCHNELL UND PRÄZISE

Beispielsweise sortiert d-Class **3,5 Textseiten pro Sekunde** mit einer **Fehlerrate von 0,1%** (Testergebnis bei 10.000 Textseiten (ASCII) und 2 Dokumentklassen).

D-IMAGESEARCH MIT PC-POWER

Die Verarbeitungsgeschwindigkeit bei Dokumenten-Images wird hauptsächlich von der Bildverarbeitung und Zeichenerkennung (OCR) bestimmt, sie ist abhängig von der verwendeten CPU, vom Bus und von der Grafikkarte des PC, sowie vom Betriebssystem, von der Auflösung der Images, von der Textmenge der Seite und von deren inhaltlichen Mischung (Text, Grafik, Bild).

Je besser die Qualität der Bildverarbeitung und der Zeichenerkennung (OCR) desto präziser die nachfolgende Dokumentenklassifikation. Dennoch toleriert d-Class in bestimmtem Umfang auch gestörte Texte.

Die Eigenschaft von d-Class, das Layout eines Dokumentes völlig unberücksichtigt zu lassen, ist beispielsweise bei der automatischen Verarbeitung verkleinerter oder vergrößerter Dokumente nützlich, oder auch im Falle schlecht gescannter Images, bei denen Regionen des Dokuments bildlich ausgefallen sind. Auch Schief lagen von Dokumentenimages werden in weiten Grenzen toleriert, solange die OCR-Engine noch einigermaßen relevante Informationen generiert.

Durchsatzmessungen an repräsentativem Beleggut (Mischung von text- und grafikorientierten Geschäftsdokumenten aus der Praxis, Basis von 10.000 Dokumentenimages TIFF Gr. 4 am Beispiel von 45 Dokumentenklassen) ergab auf einem modernen PC-System (mit Pentium4, 3 GHz, 1 GB RAM, Windows XP Professional, SP4) eine mittlere Interpretationsdauer von 5,5 sec. pro Beleg oder 650 Dokumenten/Stunde. Dieser Durchsatz kann gesteigert werden, wenn dem Verfahren nur überwiegend textorientierte Dokumente zugeführt werden (Bild- und Grafikinhalte verlangsamen nur ohne irgendwelchen Informationsgewinn).

An diesem Beispiel wird deutlich, daß es sich bei der textbasierten Klassifikation von Images um ein sehr rechenintensives Verfahren handelt, für das PCs der höchsten Rechenleistung sinnvoll sind.

D-TRAIN IST SPONTAN

Für ergänzendes Spontantraining durch den Anwender liefern wir einen Trainingsmodul – ganz treffsicher, aber mit insgesamt weniger hohen Qualitätsansprüchen.

INDIVIDUELLE KLASSIFIKATOREN MIT „HQ“

Besonderer Vorteil von d-Class ist jedoch, für trainingsmüde Anwender mit höchsten Qualitätsansprüchen (**High Quality**) bereits fertige Klassifikatoren zu mitzuliefern, die alles im Praxiseinsatz erforderliche Wissen bereits ab Hersteller mitbringen und nicht vom Anwender belehrt bzw. trainiert werden müssen.

Beim Hersteller werden Dokumente gesammelt, gespeichert, gesichtet, kategorisiert und belehrt. DocStream als Hersteller des kundenspezifischen Dokumentenklassifikators übernimmt hier die Qualitätsverantwortung.

Nachbelehrungen sind nur dann erforderlich, wenn das zu kategorisierende Beleggut sich inhaltlich stark wandelt, was jedoch gewöhnlich auf absehbare Zeit nicht zu erwarten ist.