Docuware

From Documents to Value

White Paper

Intelligent Indexing

2013

Dieses Dokument wurde erstellt mit ***AuthorIT™, Total Document Creation*** (***http://www.author-it.com***).

DocuWare GmbH
Therese-Giehse-Platz 2
82110 Germering
www.docuware.com (***http://www.docuware.com***)

Inhalt

1 Zielsetzung des White Papers

Intelligent Indexing ist ein auf selbstlernenden Algorithmen basiertes System, das gängige Dokumenttypen selbstständig erkennt und die relevanten Dokumentinhalte als Indexbegriffe vorschlägt. Die Indexierung erfolgt automatisch und für den Nutzer unsichtbar.

Um Transparenz zu schaffen, erläutert dieses White Paper zum Intelligent-Indexing-System

- die Architektur
- die Arbeitsweise, also die Erkennung der Indexbegriffe und die selbstlernenden Algorithmen
- sowie die Sicherheit

Der Leser erhält dadurch einen fundierten Einblick in die gesamte Funktionsweise von Intelligent Indexing.

Die technischen Mitarbeiter bei Kunden, Beratungsunternehmen, Fachzeitschriften und Vertriebspartnern werden gleichermaßen angesprochen. Vorausgesetzt wird lediglich ein technisches Grundlagenwissen über den Aufbau moderner Software-Anwendungen, idealerweise von Dokumentenmanagement-Systemen. Detaillierte Kenntnisse aktueller oder früherer Versionen von DocuWare sind nicht nötig.

2 Einführung zu Intelligent Indexing

Mit Intelligent Indexing klassifiziert DocuWare Dokumente in verschiedene Typen und sucht automatisch die relevanten Indexbegriffe in beziehungsweise zu den Dokumenten und schlägt sie dem Benutzer vor. Dieser bestätigt nur noch die Vorschläge oder verbessert sie. Anhand des Feedbacks lernt das System ständig hinzu.

Da Intelligent Indexing auch Crowd-Learning-Funktionen unterstützt, lernt das System nicht nur von den Dokumenten und Feedbacks der eigenen DocuWare-Organisation (entspricht meist der eigenen Firma), sondern auch von denen anderer Benutzer. So können viele Dokumente automatisch mit den passenden Indexbegriffen versehen werden, ohne dass sie in der eigenen Organisation angelernt werden mussten.

Spätestens nach einer kurzen Einlernphase entfällt somit durch Intelligent Indexing für den Anwender größtenteils die manuelle Indexierung. Damit ist das elektronische Dokumentenmanagement nun auch bei der Archivierung der Dokumente schneller als die klassische Papierablage.

3 Architektur

Das Intelligent Indexing System läuft in einem Rechenzentrum. Es besteht aus mehreren Rechnern, auf denen der Intelligent Indexing Service läuft, und einer Datenbank (SQL Azure). In dieser werden der Volltextauszug, die Indexdaten, das Benutzerfeedback und allgemeine Informationen wie Dokumentsprache, Datumsformat etc. der von Intelligent Indexing ausgewerteten Dokumente gespeichert.

Das gesamte Intelligent Indexing System wird zurzeit auf Windows Azure gehostet, einer Cloud-Plattform von Microsoft. Dadurch ist eine hohe Skalierbarkeit und Ausfallsicherheit gewährleistet. Sogar bei Software-Updates des Intelligent Indexing Systems werden Ausfallzeiten durch die Architektur der Windows Azure Cloud Services vermieden. Ein Nutzer- und Rollenkonzept stellt zudem sicher, dass ausschließlich autorisierte Benutzer Zugriff auf die abgelegten Dokumentinformationen erhalten. Das verwendete Data-Center befindet sich in Amsterdam (Niederlande).

4 Intelligent Indexing im DocuWare-System

4.1 Verbindung von Intelligent Indexing mit DocuWare

Hat ein DocuWare-Kunde eine On-Premise-Installation, muss er sich für den Service registrieren. Er erhält dann eine Konfigurationsdatei im XML-Format, die er innerhalb der DocuWare Administration in sein DocuWare-System einspielt. Anhand der enthaltenen Daten kann sich dann das DocuWare-System mit dem Intelligent Indexing Service verbinden.
Für Kunden von DocuWare-Online ist das System schon vorbereitet.

Das Vorschlagen von Indexbegriffen für Dokumente erfolgt in den DocuWare Briefkörben, die dafür entsprechend konfiguriert werden müssen. Neben dem Aktivieren des Intelligent Indexing Services wählt man einen Ablagedialog aus. Anhand dieses Ablagedialogs weist man anschließend den Kategorien, für die Intelligent Indexing Vorschläge machen soll, wie Dokumenttyp, Datum, Kontakt, Betrag etc., DocuWare Indexfelder zu. Bei der Ablage von Dokumenten, die per Intelligent Indexing mit Indexbegriffen versehen worden sind, sind die Indexbegriffe dann in den entsprechenden Indexfeldern des Ablagedialogs eingetragen.

Weiterführende Informationen zur Konfiguration von Intelligent Indexing finden Sie im ***Handbuch*** (***http://help.docuware.com/de/#t59014***).

4.2 Arbeitsablauf mit Intelligent Indexing

Für alle Dokumente, die in einen für Intelligent Indexing eingerichteten Briefkorb gelangen, werden zunächst automatisch Volltextauszüge erstellt und anschließend an den Intelligent Indexing Service transferiert. Dieser wertet die Volltextauszüge aus, sucht nach ähnlichen schon bekannten Dokumenten und macht Vorschläge für die Indexbegriffe. Je nachdem, für wie sicher Intelligent Indexing die Erkennung der vorgeschlagenen Indexbegriffe erachtet, werden die Dokumente im DocuWare Briefkorb mit drei verschiedenen Farben im Ampelsystem markiert.

Sobald der Benutzer ein Dokument über den zugewiesenen Ablagedialog im Archiv speichern möchte, werden die von Intelligent Indexing vorgeschlagenen Indexbegriffe in den einzelnen Indexfeldern des Dialogs angezeigt. Erneut lässt die drei-stufigen Farbmarkierung die Probabilität der einzelnen Indexbegriffe erkennen. Zudem wird das Dokument im DocuWare Viewer angezeigt.

Indem der Benutzer die Indexbegriffe akzeptiert oder ändert, gibt er Feedback an das Intelligent Indexing System. Dieses wertet das Feedback durch die selbstlernenden Algorithmen aus, sodass ähnliche Dokumente in Zukunft korrekt von Intelligent Indexing indexiert werden können. Um einen möglichst hohen Lerneffekt zu erzielen, sollte der Anwender, wenn er Indexbegriffe ändert oder ergänzt, diese nicht direkt in den Ablagedialog tippen, sondern per One-Click-Indexing übernehmen. Dies ist eine Funktion im DocuWare Viewer zum Übertragen von Wörtern/Zahlen/Daten aus dem angezeigten Dokument in den Ablagedialog. Intelligent Indexing erhält als Feedback dann nicht nur den Begriff als solchen, sondern auch dessen Position im Dokument, was den Lernerfolg erhöht.

Weiterführende Informationen zur Verwendung von Intelligent Indexing finden Sie im ***Handbuch*** (***http://help.docuware.com/de/#t58569***).

5 Automatische Indexerkennung

Die automatische Indexerkennung ist das Herzstück von Intelligent Indexing. Es basiert hauptsächlich auf drei Bereichen: den verschiedene Methoden zum Auslesen und Analysieren der einzelnen Dokumente, den Modellräumen, in denen nach ähnlichen bereits von Intelligent Indexing verarbeiteten Dokumenten gesucht wird, und den selbstlernenden Algorithmen.

5.1 Methoden der Indexerkennung

Intelligent Indexing verwendet eine Vielzahl von Methoden, um die richtigen Indexbegriffe zu Dokumenten herauszufinden. Einige davon sind aktuell beim Deutschen Patentamt zum Patent angemeldet. Das System ist trotz der vielen verschiedenen Algorithmen, die pro Dokument ausgeführt werden, hoch performant. Außerdem ist es flexibel für unterschiedliche Sprach- und Kulturräume, arbeitet auch mit leicht schief eingescannten Dokumenten problemlos und wertet Dokumentenelemente unabhängig davon aus, auf welcher Dokumentseite und wo sie sich innerhalb von einer Dokumentseite befinden.

Für alle von Intelligent Indexing verarbeiteten Dokumente wird zunächst eine Spracherkennung anhand von Wortbausteinen durchgeführt. Die erkannte Sprache bzw. der Kulturkreis eines Dokuments ist auch relevant für die korrekte Auswertung von Datumsangaben. Zum Beispiel lernt Intelligent Indexing, ob in einem englischen Dokument ein Datum im Format MM.TT.JJJJ oder als TT.MM.JJJJ vorliegt. Auch um Zahlenangaben korrekt auszuwerten, stützt sich Intelligent Indexing auf die erkannte Sprache bzw. den Kulturkreis des Dokuments. Als zusätzliches Kriterium dient hier die Position der Trennzeichen (Punkt und Komma als Dezimal- und Tausendertrennzeichen) innerhalb einer Zahl.

Des Weiteren macht sich Intelligent Indexing zu Nutze, dass viele wichtige Angaben in einem Dokument häufig in der Nähe von zugehörigen Schlüsselwörtern stehen, zum Beispiel das Datum neben/unter dem Begriff "Datum" oder der Betrag einer Rechnung neben/unter dem Begriff "Summe". Zudem stehen bei ähnlichen Dokumenten, zum Beispiel Rechnungen von einer bestimmten Firma, die einzelnen Elemente der Dokumente immer an ähnlicher Stelle, beispielsweise, Datum und Rechnungsnummer.

Weitere Methoden dienen zur Bestimmung des Dokumenttyps, wie etwa Rechnung, Lieferschein etc. und der Entscheidung, ob es sich um ein ein- oder ausgehendes Dokument handelt. Insbesondere bei Dokumenten, zu denen keine ähnlichen, bereits gelernten Dokumente im System existieren, werden Algorithmen mit festen Regeln angewendet, um die Indexbegriffe zu einem Dokument zu bestimmen. Diese Regeln basieren auf typischen Dokumentstrukturen und -inhalten von häufig verwendeten Dokumentarten. So wird zum Beispiel bei einer Rechnung vermutet, dass die größte Zahl mit einem Währungszeichen die Rechnungssumme ist.

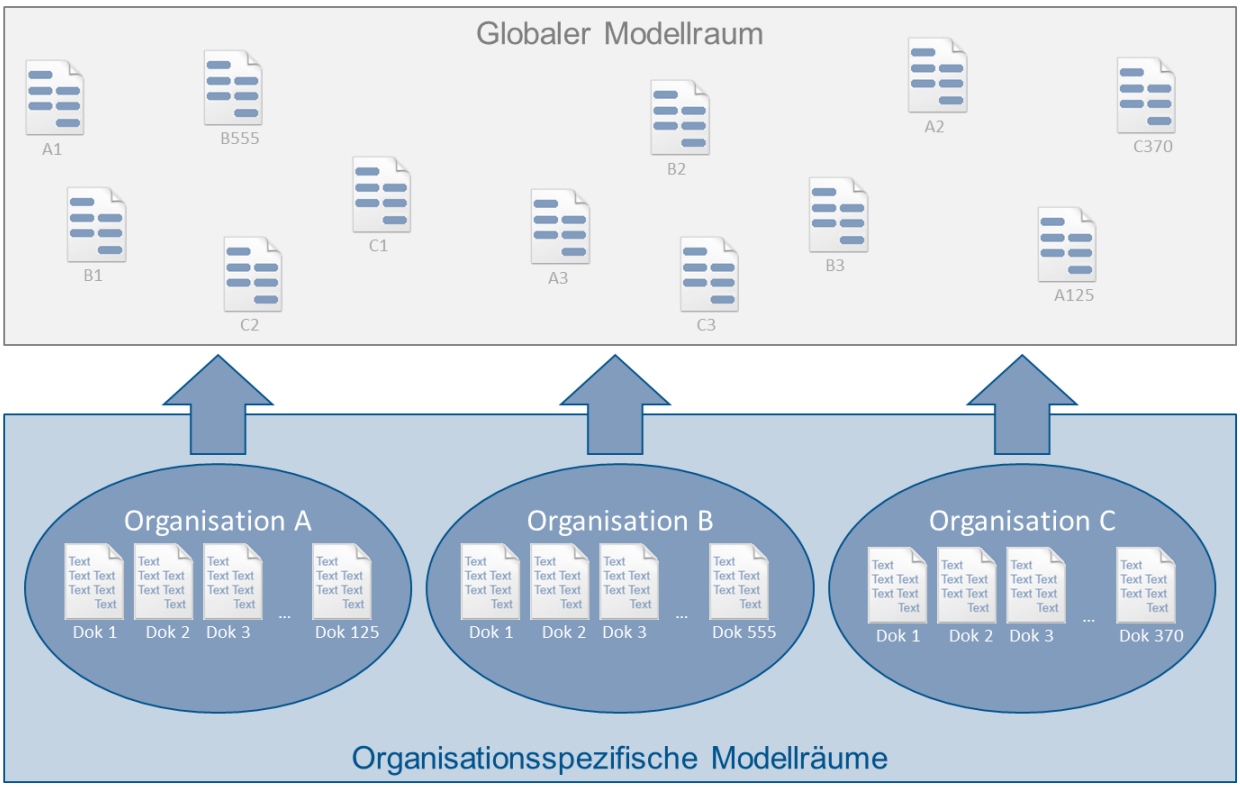
Trotz der Menge an Automatismen beachtet Intelligent Indexing auch die manuellen Eingaben des Nutzers. Ähnelt ein neues Dokument vielen, bereits gelernten Dokumenten, bei denen manuell ein bestimmter Indexbegriff eingegeben wurde, so wird dieser auch für das aktuelle Dokument verwendet. Dabei muss der Begriff nicht unbedingt in dem Dokument selbst enthalten sein.

Zum Beispiel gibt ein Benutzer bei Dokumenten, die er von einer bestimmten Person erhält, immer den Namen dieser Person als Indexbegriff ein (auch, wenn der Name nicht auf dem Dokument steht). Für ein weiteres Dokument dieser Person würde Intelligent Indexing dann auch den Namen der Person als Indexbegriff vorschlagen, weil das System es so von den bereits indexierten Dokumenten gelernt hat.

Für jeden Indexwert eines Dokuments wertet Intelligent Indexing die Ergebnisse der einzelnen Methoden aus und berechnet über kombinatorische Algorithmen den plausibelsten Indexbegriff. Dieser wird dem Anwender im Ablagedialog direkt angezeigt, weitere etwas weniger plausible Indexbegriffe werden in einer Auswahlliste angeboten.

5.2 Modellräume

Mit Modellraum wird bei Intelligent Indexing die Komponente bezeichnet, die Informationen von einem bereits gelernten Dokument für die Indexierung eines neu zu indexierenden Dokuments nutzt und auch Trainingsergebnisse speichert. Unterschieden wird dabei zwischen dem *organisationsspezifischen Modellraum* und dem *globalen Modellraum*. Der organisationsspezifische Modellraum beinhaltet die Volltextauszüge der Dokumente mit den gelernten Indexierungsmethoden einer bestimmten DocuWare Organisation (dies ist meist eine Firma, die mit DocuWare arbeitet). Der globale Modellraum fasst mehrere dieser organisationsspezifischen Modellräume zusammen. Dabei werden von den einzelnen organisationsspezifischen Modellräumen nur die Strukturinformationen derjenigen Dokumente in den globalen Modellraum übertragen, die einen Dokumenttyp haben, der von Intelligent Indexing als Standard-Dokumenttyp betrachtet wird (Eingang-/Ausgangsrechnung, Angebot, Bestellung, Kontoauszug, Brief und viele mehr). Alle anderen Dokumente werden als nicht-relevant für globale Indexierungsstrategien erachtet.



Wenn Intelligent Indexing ein neues Dokument erhält, wird zunächst geprüft, ob in dem organisationsspezifischen Modellraum, also bei den Dokumenten der gleichen Firma, ähnliche Dokumente vorliegen, deren Indexierungsmethoden auf das aktuelle Dokument übertragen werden können. Dabei kann bereits ein einziges sehr ähnliches Dokument ausreichend sein. Mit einer größeren Anzahl an Referenzdokumenten steigt jedoch die Wahrscheinlichkeit für eine erfolgreiche automatische Indexdatenextraktion. Bis zu fünf Dokumente werden für jedes neu zu indexierende Dokument berücksichtigt. Wenn also ähnliche Dokumente im organisationsspezifischen Modellraum gefunden werden, wird anhand dieser für das neue Dokument die Indexierung vorgenommen.

Kann im organisationsspezifischen Modellraum aber kein Referenzdokument gefunden werden, sucht Intelligent Indexing im globalen Modellraum, ob es schon ein ähnliches Dokument verarbeitet und indexiert hat. Wenn das der Fall ist, übernimmt Intelligent Indexing von einem solchen die Indexierungsstrategie und nutzt das Crowd-Learning-Prinzip aus. Der Anwender profitiert davon, dass in einer anderen DocuWare-Organisation (einer anderen Firma) schon einmal für ein ähnliches Dokument die Indexierung gelernt worden ist.
Aus Sicherheitsgründen wird beim Verwenden eines Dokuments aus dem globalen Modellraum nie direkt ein Indexbegriff von diesem Dokument übernommen, sondern nur Strukturdaten, wie zum Beispiel die Position des verwendeten Begriffs auf dem Dokument. Eine Ausnahme hiervon ist das Feld, das bei eingehenden Dokumenten die Absenderfirma enthält. Diese Information wird als sicherheitstechnisch unbedenklich angesehen.

Aufgrund der kulturellen Unterschiede und den damit verbundenen verschiedenen Dokumentstrukturen arbeitet das Intelligent Indexing System aktuell mit drei globalen Modellräumen: einen für europäische DocuWare-Kunden, einen für amerikanische und einen für die DocuWare-Kunden der restlichen Regionen.

5.3 Alterungsstrategien für verarbeitete Dokumente

Um die Modellräume aktuell zu halten, werden Dokumente nicht dauerhaft als Vergleichsobjekte für die Indexierung verwendet.
Je seltener ein Dokument zur Indexierung anderer Dokumente herangezogen wird, desto unwahrscheinlicher ist es, dass es in der Zukunft als Referenzdokument dient. Ebenso werden im Laufe der Zeit falsch oder ungenügend indexierte Dokumente erkannt und aus dem System entfernt.
Neue Dokumente und Dokumente, die schon häufig als Referenzen gedient haben, werden hingegen bei der Übertragung der Indexierungsmethoden präferiert.
Durch diese Alterungsmethoden für Dokumente arbeitet Intelligent Indexing nur mit relevanten Dokumenten für die Indexierung neuer Dokumente.

5.4 Benutzer-Feedback

Sobald ein Benutzer Indexbegriffe bestätigt oder ändert, analysiert Intelligent Indexing dieses Feedback, verwaltet es im Modellraum und nutzt die gewonnenen Informationen für folgende ähnliche Dokumente.
Zum Beispiel extrahiert Intelligent Indexing Informationen über Korrekturen, die ein Nutzer an vorgeschlagenen Indexbegriffen durchgeführt hat. Wenn die optische Zeichenerkennung beispielsweise *Docuware GmbH* anstatt *DocuWare GmbH* ausliest, und der Anwender es entsprechend korrigiert, wird beim nächsten entsprechenden Dokument gleich *DocuWare GmbH* vorgeschlagen.
Aber es können nicht nur die Begriffe als solche gelernt werden, sondern auch zugehörige Metadaten wie zum Beispiel die Position der Begriffe auf dem Dokument. Für ein neues Dokument der gleichen Art würde dann an entsprechender Position ein Begriff aus dem Dokument als Indexbegriff vorgeschlagen werden.

6 Sicherheitskonzept

6.1 Übertragung der Dokumentinhalte und der Indexbegriffe

Zum Hochladen der Volltextauszüge der Dokumente, Senden der Indexvorschläge und Senden des Feedbacks kommunizieren der Web Client und der Intelligent Indexing Service miteinander. Die komplette Kommunikation ist HTTPS verschlüsselt, die Dokumenteninhalte und Indexbegriffe sind so vor fremden Zugriffen gesichert.

6.2 Speicherung der Dokumentinhalte

Das Intelligent Indexing System speichert von den ausgewerteten Dokumenten den Volltextauszug, die Indexdaten, das Benutzerfeedback und allgemeine Informationen wie Dokumentsprache, Datumsformat etc. Die dafür verwendete Datenbank ist bei Microsoft Azure gehostet, was hohe Skalierbarkeit und Ausfallsicherheit gewährleistet. Ein Nutzer- und Rollenkonzept stellt zudem sicher, dass ausschließlich autorisierte Benutzer Zugriff auf die abgelegten Dokumentinformationen erhalten. So kann zum Beispiel der DocuWare Support mit Erlaubnis eines Kunden auf deren Volltextauszüge zugreifen, um eventuelle Probleme zu analysieren und zu beheben.
Auf Wunsch können die Daten aus dem Intelligent Indexing System auch wieder entfernt werden.

6.3 Abgrenzung der Informationen in globalen Modellräumen

Durch die globalen Modellräume, in denen Trainingsmodelle und die verwendeten Dokumentinformationen von vielen Anwendern gespeichert sind, profitieren alle Anwender voneinander. Insbesondere neue Anwender, deren organisationsspezifischer Modellraum noch leer ist, können sehr gute Ergebnisse durch die Nutzung des globalen Modellraums erzielen.
Der Datenschutz ist dabei gewährleistet: Abgesehen vom Sender eines Dokuments werden keine inhaltlichen Angaben aus den Dokumenten weitergegeben, sondern nur Strukturdaten geteilt. So wird sichergestellt, dass Intelligent Indexing niemals Dokumentinhalte von einem DocuWare-Kunden als Indexbegriffe an einen anderen DocuWare-Kunden weitergibt.

6.4 Löschen der Daten beim Verlassen des Systems

Wenn ein DocuWare-Kunde das Intelligent-Indexing-System wieder verlassen will, werden der zugehörige organisationsspezifische Modellraum, und damit die Volltextauszüge der Dokumente, aus dem Intelligent Indexing System gelöscht.

7 Glossar

DocuWare Administration	Die DocuWare Administration ist ein Modul zur Verwaltung einer gesamten DocuWare-Installation. Dazu gehören beispielsweise die DocuWare Server, DocuWare Archive und Benutzerrechte. Bei der Nutzung von DocuWare Online werden über die DocuWare Administration nur Elemente der DocuWare Organisation verwaltet wie Archive und Benutzer.
DocuWare Briefkorb	Die Briefkörbe in DocuWare sind die Orte, an denen sich die Dokumente befinden, die zur Bearbeitung anstehen. Das können Dokumente sein, die noch nicht archiviert sind, oder auch Kopien von Dokumenten, die sich bereits im Archiv befinden.
DocuWare Organisation	Ein DocuWare-System hat mindestens eine Organisation, die mit der Lizenzen, die beim Installieren genutzt wurden, korrespondiert. Mit zusätzlichen Lizenzen lassen sich weitere Organisationen einrichten, beispielsweise um das System mandantenfähig zu machen. Im Allgemeinen entspricht eine DocuWare-Organisation einer Firma, die DocuWare verwendet.
globaler Modellraum	Ein globaler Modellraum beinhaltet die Strukturdaten von Dokumenten mit den zugehörigen gelernten Indexierungsmethoden von verschiedenen DocuWare Organisationen. Für verschiedene Sprach-/und Kulturräume gibt es verschiedene globale Modellräume.
Indexbegriffe	Indexbegriffe sind Metadaten zu einem Dokument. Sie werden als Ordnungskriterien einem Dokument zugeordnet, um dieses strukturiert in DocuWare abzulegen und um das spätere Finden des Dokuments zu erleichtern.
Indexierung	Das Zuweisen von Indexbegriffen zu einem Dokument wird als Indexierung bezeichnet.
Modellraum	Mit Modellraum wird bei Intelligent Indexing der Bereich bezeichnet, aus dem Indexierungsmethoden und Informationen von einem bereits gelernten Dokument für die Indexierung eines neu zu indexierenden Dokuments übertragen werden.
organisations-spezifischer Modellraum	Der organisationsspezifische Modellraum beinhaltet die Volltextauszüge der Dokumente mit den gelernten Indexierungsmethoden einer bestimmten DocuWare Organisation.
Volltextauszug	Der Volltextauszug zu einem Dokument beinhalten den Text, der in einem Dokument enthalten ist zusammen mit den Positionen der einzelnen Textelemente innerhalb des Dokuments. Bei gescannten Dokumenten wird der Volltextauszug mithilfe einer optischen Zeichenerkennung (Optical Character Recognition - OCR) erstellt.