Dokumentenablage und Informationssuche - Anforderungen und Lösungen

Armin Hust, A-Z Technology AG

Dr. Armin Hust ist Mitbegründer und Aufsichtsratsvorsitzender der A-Z Technology AG, die in den Bereichen Wissensmanagement und Information Retrieval arbeitet.

Der Begriff „Informationssuche" wird heute größtenteils einer Suche im Internet assoziiert. Dabei wird häufig vergessen, dass die Suche nach Informationen, die innerhalb des Unternehmens in der Form von unstrukturierten Dokumenten vorliegen, weitaus häufiger stattfindet. Bisher hatte man nur geringe automatische Unterstützung, die lokalen Dokumentenarchive zu organisieren und zu durchsuchen. Neue Produkte, Methoden und Technologien bieten nun vielfältige Möglichkeiten, schnell und gezielt die gewünschten Informationen wieder zu finden.

Trotz Automatisierung der Geschäftsprozesse und Integration der Anwendungen fallen immer mehr Daten und Informationen in unstrukturierten Dokumenten an. Dabei handelt es sich um verschiedenste Dokumentarten, wie z.B. Briefe, Kalkulationen, Präsentationen, E-Mails, wobei E-Mails auch beliebige Dokumente als Dateianhänge enthalten können. E-Mails und Dokumente werden für lange Zeit aufbewahrt und archiviert; sie können nicht gelöscht werden, sei es aus rechtlichen Gründen, aus revisorischen Gründen oder aus betrieblichen Gründen, die die Nachvollziehbarkeit von bestimmten Vorgängen erforderlich machen.

Bei der Ablage und dem Wiederauffinden von Information stellen sich verschiedene Herausforderungen, die mit den aktuellen Techniken des Wissensmanagements und des Information Retrievals gut zu meistern sind.

Anforderungen

Horizontale und vertikale Ablage. Der einzelne Benutzer, der ein Dokument erstellt oder ein Dokument per E-Mail erhält, ist zunächst der „Besitzer" des Dokuments. Kollegen können von den Informationen des Dokuments nicht profitieren, wenn das Dokument bei der Ablage im „Privatbesitz" bleibt. Damit Dokumente in den „Allgemeinbesitz" übergehen können, müssen organisatorische und technische Voraussetzungen geschaffen werden, damit alle Berechtigten den Zugriff zu diesen Dokumenten erlangen können, wobei in den Zugriffsrechten nach „Lesen" und „Schreiben" zu unterscheiden ist.

Eine Verfeinerung der Ablage ist dabei auch nach thematischen Untergliederungen möglich. Ein Angebot ist dann nicht nur als Angebot zu betrachten, sondern als „Angebot" eines „Lieferanten" über ein „Produkt". Diese in Datenbanken bewährte Ablageform führt bei unstrukturierten Dokumenten direkt zum Problem der Mehrfachablage.

Mehrfachablage. Dokumentkopien können in verschiedenen Ordnern abgelegt werden, E-Mails jedoch können in den meisten Systemen nur in einem einzigen Ordner abgelegt werden, beispielsweise bei „Angeboten". Einige

In diesem Beitrag lesen Sie:

- welche Schwachstellen gegenwärtig bei der Dokumentenablage und der Informationssuche existieren.
- wie die Anforderungen an die Dokumentenablage und Informationssuche durch kommerzielle Systeme abgedeckt werden können.
- welche Funktionalitäten das Produkt A-Z Finder der Firma A-Z Technology AG abdeckt.

Bild 1: Indices.

Verfügbare Indizes im aktuellen Indexordner (Eintrag 1 von 18)

Indexname	Status	Worte gesamt	Worte versch.	Analysierte Dateien	Indexgröße (KB)	Zuletzt geändert am
Test1 - Index vom 01.12.2004 ...	vollständig	416207	29835	3346	3284	14.12.2004 18:55:23
Test2 - Index vom 01.12.2004 ...	vollständig	416207	29835	3346	3284	14.12.2004 18:58:13
Artikel	vollständig	1199682	59678	3164	8219	14.12.2004 11:28:16
Mail	vollständig	2804509	192904	16239	19477	14.12.2004 19:02:18
AZ	vollständig	330874	32637	2129	2527	21.12.2004 15:20:21
Link auf Netzwerk-Laufwerk	vollständig	38263	8150	318	413	14.12.2004 11:16:03
Alle Dokumente im Ordner 'Eige...	neu angelegt	0	0	0	5	30.11.2004 14:24:15
Alle lokalen Bilder	neu angelegt	0	0	0	9	30.11.2004 14:24:15
Alle lokalen Filme und Musik	neu angelegt	0	0	0	9	30.11.2004 14:24:15
Alle lokalen Texte und Dokume...	neu angelegt	0	0	0	9	30.11.2004 14:24:15
Testindex	neu angelegt	0	0	0	5	30.11.2004 14:24:15

System erlauben, die E-Mail als Kopie abzulegen, z.B. in den Ordnern „Lieferanten“ und „Produkte“. Dies führt jedoch in kurzer Zeit zu redundanten Strukturen.

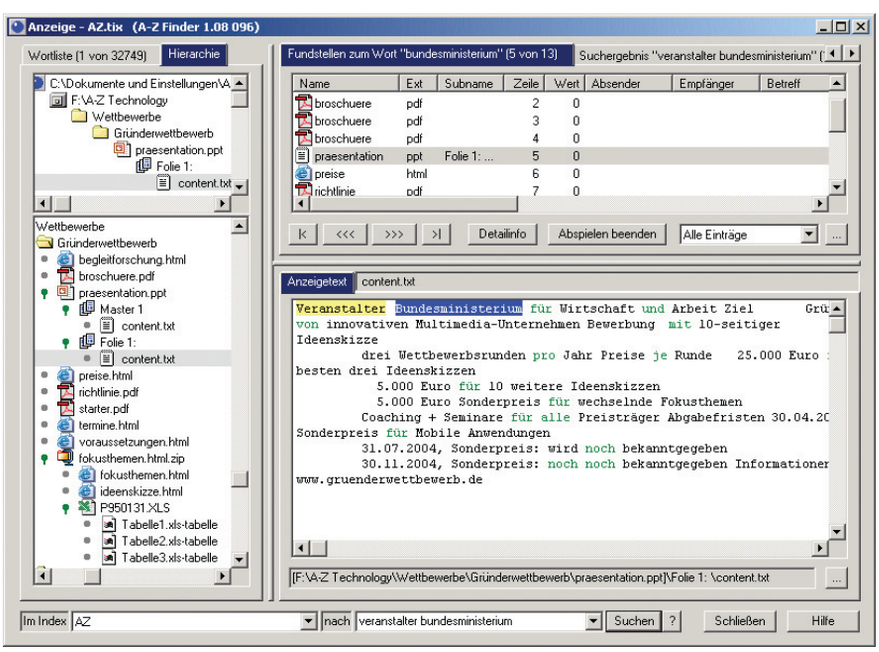
Darüber hinaus entsteht eine Mehrfachablage fast zwangsläufig bei E-Mails mit Dateianhängen in folgendem Szenario: Ein Benutzer erstellt ein Dokument (Original, 1. Kopie), versendet es als Dateianhang an einen Kollegen (Original, 2. Kopie), der Kollege empfängt die Datei in der E-Mail (Original, 3. Kopie) und speichert sie ab (Original, 4. Kopie). Wenn der Kollege Änderungen an dem Dokument vornimmt und es zurückschickt, entstehen von der geänderten Version wieder 4 Kopien. Ohne aggressive Löschstrategie bleibt das Dokument in zwei verschiedenen Versionen in jeweils 4 Kopien erhalten, da die meisten E-Mail-Systeme es standardmäßig nicht ermöglichen, die Dateianhänge zu löschen, die E-Mail selbst aber beizubehalten. Nach kurzer Zeit lässt sich kaum mehr ermitteln, welche Duplikate eines Dokuments existieren, welches die neueste Version eines Dokuments ist, und wie weit eine neuere Version eines Dokuments mit einer älteren Version übereinstimmt.

Informationssuche. Die bisherige Unterstützung zum Suchen von Inhalten in unstrukturierten Dokumenten auf dem eigenen PC oder im Netzwerk ist gering. In der Familie der Windows-Betriebssysteme steht dafür der Explorer zur Verfügung, der jedoch nur eingeschränkte Funktionalität bietet. Erst zukünftige Versionen dieses Betriebssystems werden über weitergehende Unterstützung verfügen. In letzter Zeit sind jedoch geeignete eigenständige Programmsysteme entstanden, die die Suche nach lokalen Dokumenten unterstützen. Diese Programme arbeiten überwiegend mit Methoden der Künstlichen Intelligenz, die in dem Forschungsbereich Information Retrieval (einem Teilgebiet des Wissensmanagements) entwickelt wurden.

Überblick über verwendete Datenelemente. Einzelne Datenelemente von Dokumenten sind nicht mehr sichtbar und über die Standardsuchtechniken nicht wieder auffindbar. So können mit den Standardmethoden z.B. Kommentare zu einer Zelle in einer Kalkulations-Tabelle nicht gefunden werden, Bilder und Videos innerhalb von Präsentationen können nicht mehr dargestellt werden und Dateiinhalte in Zip-Archiven sind nicht transparent. Jede Suche nach solchen Datenelementen erfordert das Öffnen eines Dokumentes mit der entsprechenden Anwendung und eine individuelle Suche oder manuelles Durchblättern.

Weiterbearbeitung. Wiederverwendung und Weiterbearbeitung von gefundenen Textstellen ist nur möglich, indem das Dokument zuerst mit der entsprechenden Anwendung geöffnet wird und danach der gewünschte Teil über die Zwischenablage anderen Anwendungen zur Verfügung gestellt wird. Extrahieren und Abspeichern von eingebetteten Bildern, Videos und Audios aus Präsentationen ist äußerst schwierig, teilweise sogar unmöglich.

Bild 2: Ansichten.



Lösungen

Am Beispiel des Produkts A-Z Finder der Firma A-Z Technology AG wird hier dargestellt, wie die Anforderungen durch kommerzielle Systeme abgedeckt sind.

Dokumentablage. Damit die Suche auch auf sehr großen Datenbeständen schnell durchgeführt werden kann, ist eine sogenannte Indizierung notwendig. Dabei werden Dokumente und E-Mails einmalig vor der Suche analysiert und Informationen in einem Index abgelegt. Ein Index bildet das globale Stichwortverzeichnis für alle analysierten Dokumente. Darüber können zu einem späteren Zeitpunkt Dokumente und E-Mails sehr schnell gefunden werden, ohne sie erneut lesen und aufwendig durchsuchen zu müssen.

Die in Bild 1 dargestellte Struktur von verschiedenen Indices unterstützt ein sehr flexibles Dokumentablagesystem. Einzelne Indices gehören dem Benutzer und sind von diesem erstellt, während der Index mit dem Namen ‚Link auf Netzwerk-Laufwerk’ ein gemeinsamer Index ist, der von einem Administrator mehreren Benutzern zur Verfügung gestellt wird. In dieser Form lassen sich Strukturen wie Gruppen-, Abteilungs- und Bereichsablagen mit den entsprechenden Zugriffsrechten abbilden.

Mehrfachablage. Beim Aufbau des Index können Dokumentduplikate und

der Grad der Übereinstimmung ermittelt werden. Zur Vermeidung von Duplikaten können übergeordnete Metastrukturen aufgebaut werden (z.B. elektronische Akte), die geeignete Automatismen zum Erlernen der Ablagestrukturen enthalten.

Informationssuche. Ein Suchvorgang benutzt das globale Stichwortverzeichnis, auch Wortliste genannt, um sehr schnell die relevanten Dokumente zu finden. Dabei sind neben den einfachen Suchvorgängen, wie z.B. Einwortsuche, auch sehr komplexe Suchen möglich, z.B. Mehrwortsuche, Ähnlichkeitssuche, Suche mit Jokerzeichen oder Numerische Suche sowie verschiedene Kombinationen davon.

Ein Beispiel: ‚*fuß&t!*fußball ~meyer' ist eine Mehrwortsuche, bei der im ersten Suchbegriff Jokerzeichen benutzt werden und im zweiten Begriff eine Ähnlichkeitssuche durchgeführt wird. Dabei werden alle Dokumente gefunden, in denen der Begriff ‚fuß' vorkommt , aber nicht der Begriff ‚fußball', und gleichzeitig der Name ‚meyer' sowie ähnliche Schreibweisen davon vorkommen (z.B. ‚maier', ‚mayer', ‚meier', ‚meiers' etc).

Numerische Suchen erlauben die Suche nach einzelnen Werten oder können Bereiche angeben, innerhalb derer ein numerischer Wert liegen muss (z.B. findet ‚NUM>=0.1&t<2.5' alle Werte zwischen 0.1 und 2.5, auch in der Schreibweise 0.1000 bzw. 0,1000).

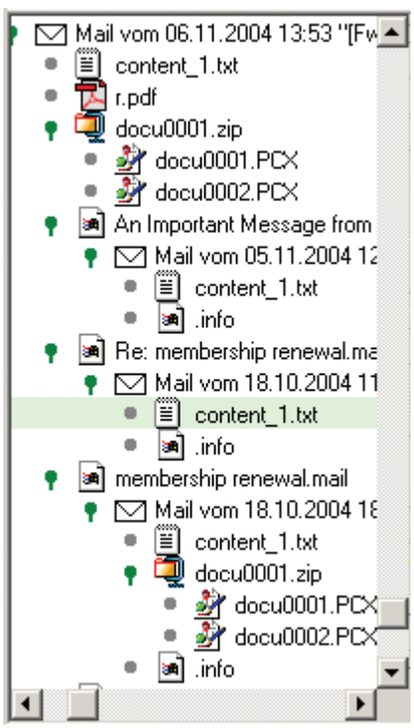
Durch die Tiefe der vorhergegangenen Analyse kann die Suche die Begriffe auch in Kommentaren zu einer Zelle in einer Excel-Tabelle oder in Notizen zu einer Folie in einer Powerpoint-Präsentation finden.

Überblick über die verwendeten Datenelemente. Nach einer Suche werden die gefunden Datenelemente in verschiedenen Sichten angezeigt. In Bild 2 ist links die Hierarchie-Ansicht ausgewählt. Darin werden die Datenelemente entsprechend der Ordnerstruktur dargestellt. Eine einfache Navigation ermöglicht so den schnellen Überblick über die vorhandenen Dokumente. Kleinste Einheiten sind hier die einzelnen Folien in einer Powerpoint-Präsentation bzw. die einzelnen Tabellenblätter in einer Excel-Tabelle (die mit anderen Dateien zusammen in einem zip-Archiv enthalten ist). Rechts oben ist die Fundstellen-Ansicht ausgewählt. Darin werden die Dokumente dargestellt, die die gesuchten Begriffe enthalten. Rechts unten werden bei Text-Dokumenten die gesuchten Begriffe im Kontext des Dokuments dargestellt, bei Grafik-Dateien wird das Bild angezeigt. Audio und Video-Dateien werden in einem separaten Media-Player-Fenster kurz angespielt.

Bild 3 zeigt die Struktur einer E-Mail in der Hierarchie-Ansicht. Die E-Mail enthält mehrere Anhänge: ein pdf-Dokument und ein zip-Archiv sowie weitere E-Mails (die selbst wieder Anhänge enthalten), die durch Beantworten und Weiterleiten entstanden sind.

Weiterbearbeitung. Aus allen Ansichten lassen sich die angezeigten Datenelemente sehr einfach extrahieren, entweder im Original-Dateiformat oder als Textformat abspeichern oder zur Weiterverwendung in die Zwischenablage kopieren.

Bild 3: E-Mail.



Ausblick

In zukünftigen Systemen wird die semantische Ebene einfließen. Dies beinhaltet das „Verstehen" eines Textes während der Vorverarbeitung (z.B. mit Techniken des Natural-Language-Processing) und den Einsatz von Ontologien während der Vorverarbeitung und zum Suchzeitpunkt.

Schlüsselwörter
Informationssuche, Wissensmanagement

Document Filing and Information Retrieval – Requirements and Solutions

Today, the term „information search" is normally associated with a search in the internet. Thereby, it is often disregarded that searches for information, which is avaliable in the company in structured or unstructured documents, occurs more frequently. Automatic assistance for organization of document archives as well as information searches on those locally stored documents was limited up to now. New products, methods and technologies now offer new and manifold potentials for quick and specific search support.

Keywords:
Information Retrieval, Knowledge Management

Kontakt

armin.hust@a-z-technology.de
info@a-z-technology.de
www.a-z-technology.de