DocuWare

White Paper Intelligent Indexing

September 2022

This document was created using AuthorIT™.

Disclaimer

The content of this guide is furnished for informational use only, is subject to change without notice, and should not be construed as a commitment by DocuWare GmbH. DocuWare GmbH assumes no responsibility or liability for any errors or inaccuracies that may appear in the informational content contained in this guide.

DocuWare GmbH
Planegger Straße 1
D-82110 Germering
www.docuware.com

DocuWare

Contents

1 Introduction

1.1 Objectives of this White Paper

Intelligent Indexing is a system based on self-learning algorithms that recognizes common document types independently and suggests the relevant document contents as index words. Indexing is performed automatically behind the scenes.

For the sake of transparency, this white paper explains the following aspects of the Intelligent Indexing system:

- Its architecture
- Its technique for identifying index words and self-learning algorithms, and
- Its security

This gives the reader a thorough picture of the complete workings of Intelligent Indexing.

This white paper addresses clients (users), consulting companies, IT magazines, and distribution partners. It assumes a certain level of technical knowledge about the structure of modern software applications, ideally of document management systems. Detailed knowledge of current or previous DocuWare versions is not necessary.

1.2 The concept of Intelligent Indexing

With Intelligent Indexing, DocuWare classifies documents into different types and automatically searches for the relevant index words in or for the documents, and suggests them to the user. The user only has to confirm the suggestions or improve them. Guided by the feedback, the system constantly continues to "learn".

Intelligent Indexing learns not only from the documents and feedback from each individual DocuWare user, but collectively from all users in a DocuWare organization (which usually encompasses the company of a customer). In this way, many documents can be automatically assigned the right index words, without having to be learned separately by each user.

After a quick learning curve, Intelligent Indexing largely replaces manual indexing for the user. Consequently, electronic document management is quicker than traditional paper filing when it comes to archiving documents as well.

1.3 Architecture

The Intelligent Indexing system runs in a data center. This consists of a number of computers running the Intelligent Indexing Service and a database (SQL Azure). The database stores full-text extractions, index data, user feedback, and general information like the document language, date format, etc., for the documents analyzed by Intelligent Indexing.

The entire Intelligent Indexing System is currently hosted on Windows Azure, a Cloud platform by Microsoft. This ensures high scalability and fail-safety. The architecture of Windows Azure Cloud Services avoids downtime even when software updates to the Intelligent Indexing System are being installed. Furthermore, a user and roles structure ensures that only authorized users receive access to the stored document information.

The following data centers are used:

- Amsterdam (Netherlands) for customers from the EMEA region.
- Virginia (USA) for customers from North and South America
- Tokyo (Japan) for customers from Japan
- New South Wales (Australia) for customers from Australia and some other Asia-Pacific countries

For DocuWare Cloud customers, the data center used for Intelligent Indexing is always in the same region as the DocuWare Cloud data center.

2 Intelligent Indexing in the DocuWare System

2.1 Linking Intelligent Indexing and DocuWare

DocuWare customers with on-premises installations must register for the service separately. Upon doing so, the customer will receive a configuration file in XML format to import into DocuWare Configuration for that customer's DocuWare system. Then the DocuWare system can use the data contained in that file to link to the Intelligent Indexing Service. For DocuWare Cloud customers, the system is pre-configured.

Index words are suggested for documents within DocuWare document trays, which must be configured accordingly. In addition to enabling Intelligent Indexing Services, this involves selecting a store dialog prepared for Intelligent Indexing. Within the store dialog the assignment of the categories for which Intelligent Indexing should make suggestions, such as document type, date, contact, amount, etc., to particular DocuWare index fields is defined. When storing documents to which Intelligent Indexing has assigned index words, the index words are entered in the index fields of the store dialog.

More about the configuration of Intelligent Indexing.

2.2 Workflow with Intelligent Indexing

Whenever a document is sent to a document tray that is set up for Intelligent Indexing, full-text extractions are generated, then automatically transferred to the Intelligent Indexing Service. The service analyzes the full text extractions, looks for similar documents that are already known, and offers suggested index words. Depending on Intelligent Indexing's confidence of having correctly identified the suggested index words, the documents are highlighted in the document tray with one of three colors following the "traffic light system". For customers using the Intelligent Indexing Cloud Service in combination with an on-premises DocuWare system, the document is added to the customer's quota at this point.

When the user would like to store a document in the file cabinet via the assigned store dialog, the index words suggested by Intelligent Indexing are shown in the dialog's respective index fields. Once again, the three-level color coding helps recognize the probability of each index word's correctness. The document is also displayed in the DocuWare Viewer.

The user provides the Intelligent Indexing system with feedback by accepting or changing the index words. The system employs self-learning algorithms to analyze the feedback, enabling Intelligent Indexing to correctly index similar documents in the future. To achieve the strongest learning effect possible, the user should not type index words directly into the store dialog when modifying or adding them, but should apply them using One-Click Indexing instead. This is a feature in the DocuWare Viewer that carries over words/numbers/data from the document displayed in the store dialog. That way Intelligent Indexing receives feedback on the word's position in the document as well as the word itself, improving how effectively it learns.

More about using Intelligent Indexing.

3 Automatic Index Recognition

Automatic index recognition is the core of Intelligent Indexing. It draws primarily on three areas: (1) the various methods for reading out and analyzing individual documents, (2) the model spaces that are searched for similar documents previously processed by Intelligent Indexing, and (3) the self-learning algorithms.

3.1 Index Recognition Methods

Intelligent Indexing uses numerous methods to determine the correct index words for documents. For some of these, DocuWare has patents in Germany and the USA. The system has high performance even though it runs many different algorithms for each document. Moreover, it can work flexibly in different languages and cultural areas, process documents scanned on an angle without a hitch, and analyze document elements regardless of their page within the document or their placement on that page.

Language recognition based on word fragments is performed right away for all documents processed by Intelligent Indexing. The recognized language or cultural area of a document is also relevant for correctly interpreting dates. For example, Intelligent Indexing can learn whether a document in English is using a date in the format mm/dd/yyyy or dd/mm/yyyy. Intelligent Indexing also relies on a document's recognized language or cultural area to interpret numbers correctly. An additional criterion for determining this is the position of delimiter characters within a number, that is, whether the thousand separators and decimal points are commas, periods, or something else.

Intelligent Indexing also makes use of the fact that many important details in a document are often in close proximity to associated keywords. For instance, the date is often adjacent to or beneath the word "Date" and an invoice's total is often adjacent to or beneath the word "Sum." Likewise, similar documents such as invoices from a particular company tend to always put the same elements in the same places, e.g. the date and invoice number.

Additional methods are used to classify the document's type, such as invoice, delivery note, etc. Particularly for documents without any previously learned similar documents in the system, algorithms with fixed rules are applied to determine the document's index words. These rules rest upon typical document structures and the content of frequently used document types. Thus for an invoice, for example, it is assumed that the largest number with a currency sign is the invoice sum.

Even with so many forms of automation, Intelligent Indexing also considers manual user input. If a new document resembles many documents previously learned by the system which were manually assigned a particular index word, the index word will also be used for the current document. In such cases, the word does not necessarily have to be contained in the document itself.
One example of this might be if a user always lists the name of the person the document was received from as an index word (even if the person's name is not specified in the document). Subsequently, Intelligent Indexing system would know to suggest the person's name as an index word, having learned it from the previously indexed documents.

For each index field in a document, Intelligent Indexing evaluates the results of each method and uses combinatorial algorithms to determine the most plausible index word. The word determined is displayed to the user directly in the store dialog, while somewhat less plausible index words are presented in a select list.

3.2 Model space

In Intelligent Indexing, model space refers to the component that uses information from an already learned document for then indexing a new document and also stores all training results. A model space is always organization-specific, i.e. the full-text excerpts and training results are summarized per organization and are strictly separated from the data of other DocuWare organizations.

When Intelligent Indexing receives a new document, it checks whether there are similar documents in the model space whose indexing methods can be applied to the current document. Even a single, very similar document can be sufficient. However, with a larger number of reference documents, the probability of successful automatic index data extraction increases. Currently, three documents are considered for each new document to be indexed. Since the methods are tested and adapted again and again, the number is not fixed; up to five documents can also serve as references.

Regardless of the model space, hard-coded rules are also used for indexing suggestions. These rules are based on typical documents in many languages and are adapted from time to time.

3.3 User Feedback

Whenever a user confirms or modifies index words, Intelligent Indexing analyzes this feedback, manages it in the model space, and uses the information collected for future similar documents.
One example of this is the way Intelligent Indexing extracts information about corrections a user has made to suggested index words. For instance, if the optical character recognition reads out *Docuware GmbH* (lowercase "w") instead of *DocuWare GmbH* and the user corrects it accordingly, the *DocuWare GmbH* will be suggested with the proper capitalization the next time a similar document is processed.
But the words themselves are not all the system can learn. It also picks up corresponding metadata such as their position in the document. The next time there is another document of the same kind, a word from the same position in the new document will be suggested as an index word.

4 Security Concept

4.1 Transferring Document Content and Index Words

The Web Client and the Intelligent Indexing Service communicate with one another to upload full-text extractions of documents, send indexing suggestions, and send feedback. All such communication is HTTPS encrypted, protecting the documents' contents and index words from third-party access.

4.2 Saving Document Content

The Intelligent Indexing System stores full-text extractions, index data, user feedback, and general information like the document language, date format, etc., for the documents it analyzes. The database used for this is hosted by Microsoft Azure, ensuring high scalability and fail-safety. Furthermore, a user and roles structure ensures that only authorized users receive access to the stored document information. Hence, with a customer's permission, DocuWare support can access full text extractions in order to analyze and resolve potential problems.
Upon request, the data can also be removed again from the Intelligent Indexing system.

4.3 Deleting of Data when Leaving the System

If a DocuWare customer decides to stop using the Intelligent Indexing system, the associated organization-specific model space as well as the full-text extractions of the documents are deleted from the Intelligent Indexing System.