

## Whitepaper



# NBCS: Selbstlernende Kategorisierung

Informationen in elektronischer Form liegen nur zu 20% in strukturierter und elektronisch zugriffsbereiter Form vor. Textuelle Information ist immer noch vorherrschend, denn sie kann vom Menschen wesentlich einfacher produziert und konsumiert werden.

Automatische Erschließung, Strukturierung und Ordnung solcher für die Maschine unstrukturierter Information verlangt daher spezielle Technologien und Verfahren. Ein Basisverfahren ist hierbei die Kategorisierung von Dokumenten in strukturierte Themenblöcke.

Solche Themenblöcke oder auch Kategoriensysteme erscheinen in den unterschiedlichsten Bereichen des Kommunikations- oder Informationsmanagement. Kategorisierungen sind überwiegend thematisch ausgerichtet und werden auf der Basis von Inhalten ermittelt. Sie spiegeln sich zum Beispiel wider in Dateiordnern, in E-Mail-Foldern oder in Verteilerschlüsseln. Sie ordnen aber auch Klassen von Anfragen, Informationszugriffen oder Suchaufgaben, auf die Standardantworten gegeben werden können.

Die Aufgabe eines selbstlernenden Kategorisierers besteht prinzipiell darin, den Inhalt eines Dokuments soweit zu verstehen, dass er eine oder mehrere thematische Zuordnungen, also Kategorisierungen vornehmen kann. Dazu muss der Kategorisierer in einer Lernphase die relevanten inhaltlichen Merkmale der Dokumente, Emails, etc einer Kategorie identifizieren und in Bezug auf die Dokumente in anderen Kategorien setzen. Die Lernbasis bilden eine Menge von Trainingsdokumenten, die beispielsweise derzeit schon in einem Kategoriensystem manuell gepflegt wurden.

Das Ergebnis dieser Lernphase ist ein Kategorisierungsmodell, das in der produktiven Kategorisierungsphase die Grundlage für die Entscheidungen des Kategorisierers bildet. Soll also nun ein neues Dokument kategorisiert werden, so wird es zunächst analysiert und dann mit Hilfe des Kategorisierungsmodells einer oder mehreren Kategorien zugeordnet.



NBCS ist ein neuartiges Verfahren aus der Gruppe der selbstlernenden Kategorisierer und zeichnet sich durch folgende Leistungsmerkmale aus:

- Erstellung eines Kategorisierungsmodells innerhalb von Sekunden:  
Zum Beispiel 15 Sekunden bei 19 Kategorien mit insgesamt 760 Trainingsdokumenten
- Robustheit gegenüber Anzahl und Güte von Trainingsdokumenten:  
In bestimmten Anwendungen reichen bereits 5 Trainingsdokumente je Kategorie.
- Sprachenunabhängigkeit und Skalierbarkeit auf Kategorienanzahlen bis zu 500 Kategorien.
- Erstellen von vollautomatischen Performanzanalysen bzgl. der prognostizierbaren Kategorisierungsgüte:  
Das System kann bereits während der Lernphase voraussagen, wie viele Dokumente zukünftig sicher kategorisierbar sind.
- Verwendung von Fuzzy Technologie: Der Kategorisierer ist robust gegenüber Orthografie- und Grammatikfehlern innerhalb der zu kategorisierenden Dokumente.
- Verfügbarkeit als Module:  
Als *dll* Bibliothek für Applikationen basierend auf Microsoft® .NET  
Als *jar* Bibliothek für Applikationen aus der Java Welt.