IN17 Fachvortrag für Fortgeschrittene

Smart Content Delivery - Von Dokumentation zu intelligenter Information mit Machine Learning

Jan Oevermann, Steinbeis-Transferzentrum Institut für Informations- und Content-Management

Content Delivery beschreibt das Konzept, Informationen aus verschiedenen Quellen zielgruppengerecht, situativ und kontextabhängig auszuliefern. Grundlage für diese passgerechte Informationsverteilung sind Klassifikationen, z.B. nach PI-Class®. Doch was tun, wenn die Daten (noch) nicht dafür aufbereitet sind? Machine-Learning-Verfahren können Inhalte automatisiert klassifizieren und anreichern. So kann in einem nahtlosen Prozess aus bestehender Dokumentation intelligenter Content werden: durch Smart Content Delivery.

Methoden

Was ist Content Delivery?

Unter dem Schlagwort „Content Delivery“ werden mittlerweile alle Anwendungen und Methoden zusammengefasst, die sich mit der situativen, kontext- und zielgruppenabhängigen Bereitstellung von Content beschäftigen. In der Regel sind modulare und klassifizierte Inhalte die Grundlage für diese Funktionalitäten, doch die Integration unstrukturierter Datenquellen (z.B. PDFs) rückt durch die Anforderungen aus der Industrie immer weiter in den Fokus (vgl. Ziegler 2013).

Was ist PI-Class®?

Bei PI-Class® handelt es sich um eine Methode, die für die Klassifikation von Modulen verwendet werden kann (vgl. Drewer/Ziegler 2011). PI-Klassifikationen werden als Taxonomien definiert und können systemunabhängig eingesetzt werden. Intrinsische Klassifikationen kategorisieren eindeutig die Informationsart des Inhalts (Informationsklasse) und verknüpfen ihn mit den beschriebenen Produktkomponenten (Produktklasse). Extrinsische Klassifikationen ergänzen die Methode um die vorgesehene (auch mehrfache) Verwendung des Contents für Produktmodelle und Dokumenttypen. Mit Hilfe vergebener Klassifikationen lassen sich Module automatisiert zu Dokumenten zusammenführen oder in einem Content-Delivery-Portal filtern (vgl. Ziegler 2015). Beim Erstellen von Inhalten kann eine gute Klassifikation dem Redakteur helfen, Module im CMS zu finden oder voneinander abzugrenzen und dadurch deren Wiederverwendung zu steigern.

Anwendungen und Potentiale

Durch den Einsatz von Klassifikationen und Content Delivery ergeben sich zahlreiche Anwendungsszenarien, die von kodierten Filterinformationen in RFID-Chips oder QR-Codes über Augmented-Reality-Anwendungen bis hin zu vollständig integrierten Service-Prozessen mit auftragsbezogener Informationsaufbereitung reichen. Davon ausgeschlossen sind derzeit jedoch noch nicht-modulare Informationen wie PDF-Dokumente, auf die ein Zugriff nur über das Inhaltsverzeichnis oder die Volltextsuche möglich ist.

Machine Learning

Grundlagen

Als „Machine Learning“ (dt.: maschinelles Lernen) bezeichnet man im Allgemeinen Verfahren, die auf Basis von Erfahrung neues Wissen generieren. Dabei werden Lerndaten verwendet, um Muster und Gesetzmäßigkeiten zu erkennen, welche dann auf Daten angewendet werden können, die dem System nicht bekannt sind (sog. Lerntransfer). Wird dem System während der Lernphase mitgeteilt, welche Ergebnisse für die jeweiligen Daten erwartet werden, spricht man von „Überwachtem Lernen“, zu dem auch die automatisierte Klassifizierung gezählt wird. Im Allgemeinen sind Machine-Learning-Verfahren eine Unterform der Künstlichen Intelligenz (KI).

Automatisierte Klassifizierung

Da die meisten Inhalte einer Technischen Dokumentation weiterhin textbasiert sind, ist besonders die maschinelle Textklassifizierung interessant. Maßgeschneiderte Verfahren zur automatisierten Vergabe von intrinsischen PI-Klassifikationen für Module aus dem Bereich der Technischen Kommunikation sind Gegenstand aktueller Forschung (Oevermann/Ziegler 2016).

Um die Modelle, die aus dem Training mit klassifizierten Modulen gewonnen wurden, auch auf unstrukturierte Inhalte anwenden zu können, werden Dokumente wie PDFs in zahlreiche Textfragmente zerteilt und im Anschluss automatisiert klassifiziert (Oevermann 2016a). Anhand der Resultate können nun Rückschlüsse auf die Struktur des Dokuments gezogen werden (siehe Abb. 1).

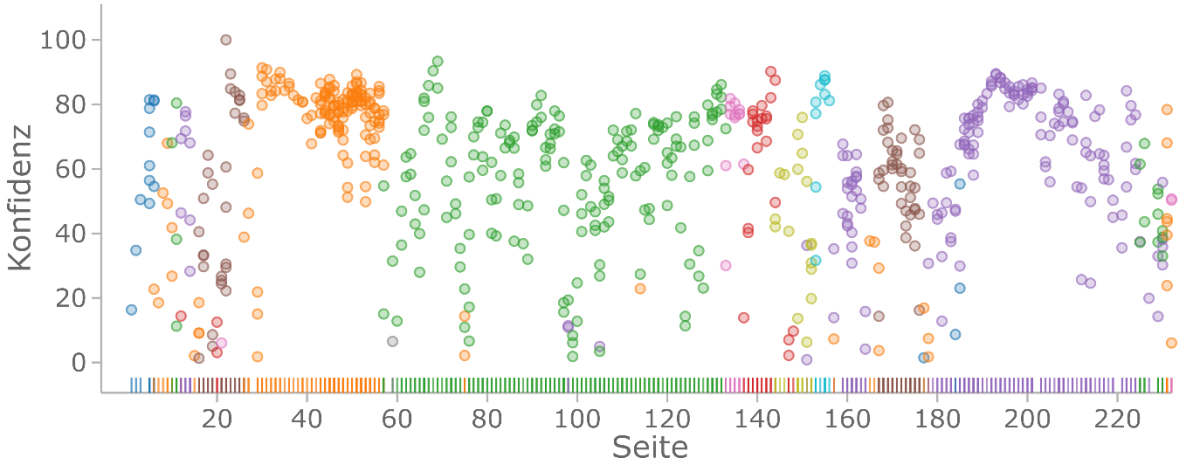

Abb. 1: Automatisierte Klassifizierung einer ca. 230 Seiten lange PDF-Betriebsanleitung nach Informationsarten (z.B.: orange: Sicherheit, grün: Betrieb, lila: Wartung, etc.). Von: fastclass.de

Diese Informationen über die Eigenschaften der Dokument-Abschnitte können bei einer facettierten Suche durch ein Content-Delivery-Portal ausgewertet und dem Nutzer die entsprechende Stelle des segmentierten PDFs präsentiert werden.

Smart Content Delivery

Prozess

Durch die Kombination von Content Delivery, PI-Klassifikation und Machine-Learning-Verfahren entsteht „Smart Content Delivery“: Unstrukturierte Dokumente, die in traditionellen Content-Delivery-Portalen nur begrenzt nach Klassifikationen filterbar waren, werden beim Import automatisch aufbereitet und durch das Portal zugänglich gemacht (Oevermann 2016b). Dabei wird der Import-Prozess des CDPs um eine automatische Anreicherung der Daten im Hintergrund erweitert. Alte oder zugelieferte PDF-Dokumente können so für einen großen Nutzerkreis komfortabel zugänglich gemacht werden (z.B. können auf die Inhalte dann die gleichen Filterfunktionen wie bei modularen Inhalten angewendet werden).

Qualitätssicherung

Die für die Technische Dokumentation wichtige Qualitätssicherung kann durch das Setzen eines Grenzwerts für die sog. Konfidenz des Klassifikators durchgeführt werden. Diese gibt die Sicherheit an, mit der eine korrekte Vorhersage getroffen werden kann. Bei der Qualitätssicherung können so Module oder Dokumente, bei deren automatisierter Klassifizierung nur eine geringe Konfidenz erreicht wurde, aussortiert und einem Technischen Redakteur zur Kontrolle vorgelegt werden.

Zusammenfassung

Der große Wunsch nach dem „magischen“ Importprozess, der aus unstrukturierten Dokumenten intelligente Informationen macht, könnte sich bald erfüllen. Die erforderlichen Technologien sind bereits heute verfügbar und müssen nur noch zu in einem durchgängigen Prozess kombiniert werden.

Weiterführende Literatur

- Drewer, Petra / Ziegler, Wolfgang (2011): Technische Dokumentation. Vogel, Würzburg.
- Oevermann, Jan / Ziegler, Wolfgang (2016): „Automated Intrinsic Text Classification for Component Content Management Applications in Technical Communication“. In: Proceedings of the 2016 ACM Symposium on Document Engineering. Seiten 95-98. ACM, NewYork, USA.
- Oevermann, Jan (2016a): „Reconstructing Semantic Structures in Technical Documentation with Vector Space Classification“. In: Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems. CEUR Workshop Proceedings.
- Oevermann, Jan (2016b): „Intelligente Klassifizierung von technischen Inhalten – Automatisierung und Anwendungspotenziale“. In: Tagungsband zur tekom Jahrestagung 2016, Seiten 185-187. tekom, Stuttgart.
- Ziegler, Wolfgang (2013), "Alles muss raus! Content Delivery für Informationsportale". In: Tagungsband zur tekom Jahrestagung 2013, Seiten 47-48. tekom, Stuttgart.
- Ziegler, Wolfgang (2015), " Content Management und Content Delivery. Powered by PI-Class". In: Tagungsband zur tekom Jahrestagung 2015. tekom, Stuttgart.

für Rückfragen:
jan.oevermann@stw.de