IN14 Fachvortrag für Fortgeschrittene

Intelligente Klassifizierung von technischen Inhalten – Automatisierung und Anwendungspotenziale

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

Klassifikationen, z.B. nach PI-Class®, spielen bei Bereitstellung und Austausch von Content eine Schlüsselrolle. Eine nachträgliche manuelle Klassifizierung außerhalb von Content-Management-Systemen ist bei großen Datenmengen jedoch sehr aufwändig und fehleranfällig. Mit den von Industrie 4.0 eingeläuteten Veränderungen hat eine Automatisierung solcher Prozesse deshalb großes Potenzial und kann Basis für viele Anwendungen sein.

Der Beitrag erläutert den Stand der Forschung und beleuchtet drei konkrete Beispiele für die Anwendung: die Migration von Bestandsdaten, die Unterstützung bei der Erstellung von Inhalten und deren Bereitstellung durch Content-Delivery-Portale.

Intelligenter Content mit Klassifikationen

Metadaten sind für ein effizientes Content Management unerlässlich und machen aus einfachen Inhalte oft erst die gewünschten „intelligenten Informationen“. Neben den beschreibenden Metadaten spielt hier besonders die Klassifikation eine entscheidende Rolle.

Im Bereich der Technischen Dokumentation ist die Methode PI-Class® hervorzuheben, die für die Klassifikation von Modulen verwendet werden kann (vgl. Drewer/Ziegler 2011). PI-Klassifikationen werden als Taxonomien definiert und können grundsätzlich 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 geplante (auch mehrfache) Verwendung des Moduls für Produktmodelle und Dokumenttypen.

Klassifizierter Content lässt sich automatisiert zu Dokumenten für ein bestimmtes Endprodukt zusammenführen oder wird durch ein Content-Delivery-Portal semantisch filterbar und über eine Facettensuche zugänglich gemacht. Aber auch beim Erstellungsprozess kann eine gute Klassifikation dem Redakteur helfen, Module im Redaktionssystem zu finden und dadurch die Wiederverwendung zu steigern.

Maschinelles Lernen

Als Maschinelles Lernen (engl.: „Machine Learning“) bezeichnet man Verfahren, die auf Basis von Erfahrung neues Wissen generieren. Dabei werden Lerndaten verwendet, um Muster und Gesetzmäßigkeiten zu erkennen, welche dann auf dem System unbekannte Daten angewendet werden können (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.

Statistische Sprachverarbeitung

Um beim überwachten Lernen aus textuellen Daten Muster zu erkennen, muss definiert werden, welche Eigenschaften den Text charakterisieren (vgl. Manning/Schütze 1999). Bei der statistischen Sprachverarbeitung werden dafür häufig einzelne Wörter oder Wortgruppen (N-Gramme) verwendet, deren Häufigkeit und Verteilung in den Klassen gezählt wird (Bag-of-Words-Modell). Aus diesen Informationen wird dann eine Gewichtung pro ausgewählter Eigenschaft berechnet (z.B. Vorkommenshäufigkeit mal inverse Dokumenthäufigkeit). Dieser Vorgang entspricht dem eigentlichen Lernprozess und erzeugt die abstrakten Muster, die aus den Daten gewonnen werden.

Vektorraum-Modell

Für jede Klasse entsteht so eine individuelle Verteilung von Gewichtungen. Diese wird in einem Vektor repräsentiert, dessen Komponenten den gewählten Eigenschaften bzw. deren Gewichtung entsprechen. Jede Klasse hat somit eine bestimmte Position in einem mehrdimensionalen Raum. Soll nun ein bisher unbekannter Text klassifiziert werden, kommen die gleichen Verfahren wie in der Lernphase zum Einsatz, um daraus einen eigenschaftsbasierten Vektor zu formen. Die eigentliche Klassifizierung besteht dann darin, zu berechnen, welchem der Klassen-Vektoren der Eingabevektor am nächsten ist.

Anwendungspotenziale

Module aus Content-Management-Systemen eignen sich prinzipiell sehr gut für Maschinelles Lernen, da viele Herausforderungen der klassischen Textklassifikation vernachlässigt werden können: Synonymie und Ambiguität, uneinheitliche Terminologie, Emotionen (Sentiment Detection) sowie unstrukturierte Lerndaten. Dem entgegen steht die im Gegensatz zu Dokumenten geringere Anzahl an auszuwertenden Eigenschaften bei Modulen (durchschnittlich ca. 1/75).

Mit einem für Module angepassten Vektorraum-Modell können bei der Zuordnung intrinsischer Informationsklassen so Ergebnisse von bis zu 90% korrekter Klassifikationen erzielt werden (vgl. Oevermann/Ziegler 2016).

Migration von Bestandsdaten

Wird mit dem Wechsel des Redaktionssystems auch die Einführung einer PI-Klassifikation beschlossen, stellt sich zwingend die Frage, wie mit Bestandsdaten umgegangen wird. Bisher musste auf Grund der erheblichen Aufwände auf eine Klassifikation der bereits erstellten Inhalte verzichtet werden.

Mit einer automatisierten Klassifizierung können diese Aufwände reduziert werden. Nur ein Teil des alten Contents (z.B. 20%) wird manuell klassifiziert und dient als Trainingsmenge für das Maschinelle Lernen. Das entstandene Modell übernimmt dann die Klassifizierung der übrigen Module.

Autorenunterstützung

Technische Redakteure, die mit Hilfe eines Klassifikationssystems arbeiten, legen bereits beim Anlegen eines neuen Moduls die entsprechenden Informations- und Produktklassen für den Inhalt fest. Eine im Hintergrund aktive automatisierte Klassifizierung kann hier als eine zusätzliche Ebene der Qualitätssicherung dienen, indem sie die vom Redakteur vergebene Klassifikation mit der aus

dem Text gewonnenen abgleicht und auf etwaige Abweichungen hinweist, bevor das Modul eingecheckt wird. So kann falsch klassifizierter Content im System vermieden werden.

Die Ermittlung von Kennzahlen zur Messung der Klassifikationsqualität ist ein weiteres Einsatzgebiet des Maschinellen Lernens. Sind Lern- und Validierungsdaten identisch, kann das Modell Klassifikationen mit einer Korrektheit von nahezu 100% voraussagen. Starke Abweichungen nach sind deutlicher Indikator für falsch klassifizierte Module oder ein unscharfes Klassifikationsmodell. Problematische Module können dann den Redakteuren zur Überprüfung vorgelegt werden.

Content-Delivery-Portale

Viele Firmen möchte mit dem Einsatz eines Content-Delivery-Portals nicht nur ihre modularen Informationen online anbieten, sondern auch Zugriff auf Bestandsdaten geben, die nicht aus einem Redaktionssystem stammen (vgl. Ziegler 2013). Um den Anwendern auch auf diesen – in der Regel unklassifizierten – Content semantischen Zugriff zu geben, muss er klassifiziert werden. Hier existieren bereits Prototypen, die den Importprozess des Portals um eine automatisierte Klassifizierung des Contents erweitern.

Eine weitere Möglichkeit, die Vorteile des Maschinellen Lernens zu nutzen, ist die Zerlegung eines unstrukturierten PDF-Dokuments in semantische Sinneinheiten (z.B. nach Informationsklassen). Dabei wird der Text eines PDF-Dokuments in eine große Anzahl kleiner Abschnitte zerlegt, die anschließend automatisiert klassifiziert werden (vgl. Oevermann 2016). Werden Grenzen zwischen Informationsklassen erkannt, kann im Portal hinterlegt werden, an welcher Stelle im PDF die Sinneinheit steht (z.B. Seitenzahl) und bei einer Filterung entsprechend darauf verwiesen werden.

Weiterführende Literatur

- Drewer, Petra / Ziegler, Wolfgang (2011): Technische Dokumentation. Vogel, Würzburg.
- Manning, Christopher / Schütze, Hinrich (1999): Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, USA.
- 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. Pages 95-98. ACM, NewYork, USA.
- Oevermann, Jan (2016): „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.
- Ziegler, Wolfgang (2013), "Alles muss raus! Content Delivery für Informationsportale". In: Tagungsband zur tekom Jahrestagung, Seiten 47-48. tekom, Stuttgart.

für Rückfragen:
jan.oevermann@stw.de