Großputz im CMS! Semantische Ähnlichkeitsanalyse für XML-Module

Jan Oevermann, ICMS GmbH, Karlsruhe
Dr. Timo Fleschutz-Balarezo, Siemens AG, Berlin

Dubletten und sehr ähnliche Inhalte führen in einem Redaktionssystem zu sinkender Wiederverwendung, höheren Übersetzungskosten und weniger Standardisierung. Automatisierte Ähnlichkeitsanalysen helfen beim Durchforsten großer Datenbestände, scheitern jedoch an den Modulvarianten, die in der Technischen Dokumentation häufig vorkommen.

In einem gemeinsamen Projekt haben ICMS und Siemens eine webbasierte Anwendung für die semantische Ähnlichkeitsanalyse von XML-Daten entwickelt, die bestimmte Textteile besonders gewichtet und damit die Trefferquote von korrekt identifizierten Löschkandidaten erfolgreich steigert. Mit Hilfe dieses Redaktionswerkzeugs konnte Siemens die Datenqualität der Bestandsdokumentation effizient steigern.

Duplikate, ähnliche Module und unkontrollierte Varianten

Auslöser und Ursachen für Duplikate

Für die Umstellung der Verwaltung der Dokumentation auf ein XML-basiertes Redaktionssystem mussten die bisherigen Inhalte migriert werden. Die bestehenden Word-Dokumente wurden automatisiert in XML-Bausteine im Informationsmodell PI-Mod zerlegt, jedoch gleichzeitig identische Texte mehrfach angelegt. Über die Jahre entstanden durch die Weiterentwicklung der Produkte zahlreiche ähnliche Texte, die inhaltlich sehr oft austauschbar sind.

Weitere Auslöser, die zu doppelten oder ähnlichen Inhalten führen sind die Arbeit in Multiautorensystemen (derzeit schreiben ca. 10 Redakteure für den betroffenen Geschäftsbereich), eine unkontrollierte Wiederverwendung (in der Regel ein Copy&Paste des wiederverwendeten Textes), sowie eine Neuanlage aufgrund von „Nicht-Wiederfinden“ des existierenden Contents.

In Summe sind 300.000 Textbausteine im System vorhanden, für die eine Analyse bzw. Bereinigung durchgeführt werden musste. Für einzelne Themenbereiche, wie etwa einen ausgewählten Strukturknoten, werden aktuelle 40-100 Bausteine verglichen.

Folgen duplikater Inhalte

Wird bereits vorhandener Content nicht wiedergefunden, muss er erneut geschrieben, geprüft und freigegeben werden. Davon ist auch die Übersetzung des Inhalts betroffen. Auswirkungen haben die Dubletten auch auf die Suchgenauigkeit im Content Delivery, wo zwar viele Treffer zurückgegeben werden (hoher Recall), der Nutzer jedoch nicht entscheiden kann, welche davon tatsächlich relevant sind (niedrige Precision).

Ähnlichkeitsanalyse

Grundlagen

Die Ähnlichkeitsanalyse basiert auf Textmerkmalen, die mit Methoden der statistischen Sprachverarbeitung ermittelt werden. Verwendet werden in der Regel die Häufigkeiten und Verteilungen von Wortgruppen oder einzelne Wörtern. Diese werden in Vektoren überführt, die deren Ähnlichkeit anschließend über den dazwischenliegenden Kosinus berechnet werden kann (sog. Kosinusähnlichkeit). Die Verfahren zur Textauswertung und Ähnlichkeitsberechnung basieren auf vorherigen Arbeiten im Bereich des Maschinellen Lernens (Oevermann/Ziegler 2018). Dieses Verfahren bietet sich an, da es zu großen Teilen der menschlichen Einschätzung zur Ähnlichkeit entspricht, indem es z.B. Satzumstellungen mit gleichem Inhalt weiterhin hohe Ähnlichkeiten zuweist aber gleichzeitig Änderungen durch Weglassen/Hinzufügen erkennt.

Semantische Gewichtung

Die in der Technischen Dokumentation häufigen gewollten Varianten (die in der Regel auf Produktvariante zurückzuführen sind) stellen jedoch ein Problem für konventionelle Ähnlichkeitsanalysen dar. Ist z.B. nur eine Leistungsangabe im Text geändert (2000 PS vs. 5000 PS), so wird eine hohe Ähnlichkeit berechnet obwohl ein erheblicher semantischer Unterschied besteht.

Um diese gewollten Varianten herausfiltern zu können, konnte auf die Intelligenz von semantischen Informationsmodelle zurückgegriffen werden. Bestimmte XML-Elemente eines Dokumentschemas können semantisch gewichtet werden, so dass ein darin liegender Unterschied höher bewertet wird. Innerhalb von PI-Mod kann dies z.B. das Element `<si-value>` sein, welches eine Einheitsangabe auszeichnet.

Anforderungen aus Dokumentationssicht

Da die Bereinigung der Dokumentation eine neue und zusätzliche Aufgabe für die Redakteure ist, wurde das Tool in enger Abstimmung mit den Redakteuren entwickelte und kontinuierlich angepasst. So konnte im Projektverlauf sichergestellt werden, dass die Nutzung gut in den redaktionellen Alltag passt.

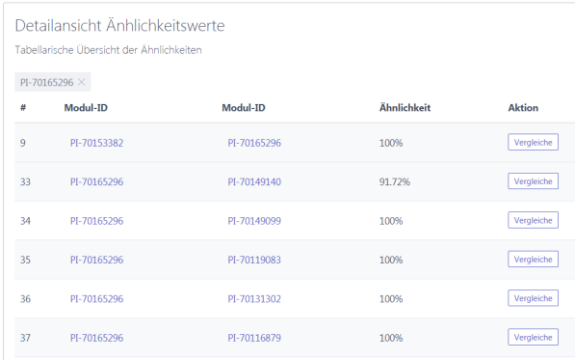
Anwendung

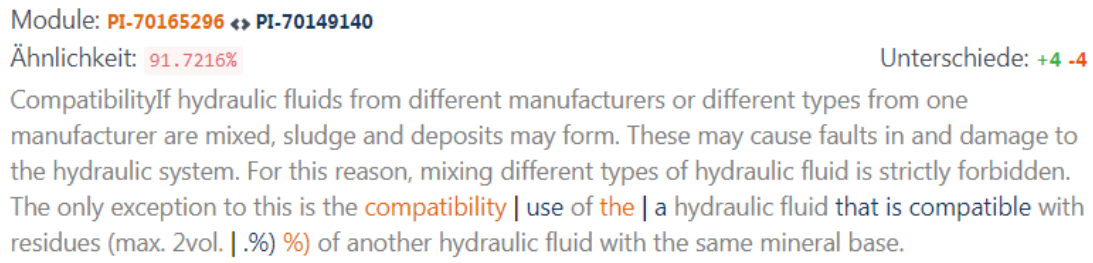
Die Lösung wurde als client-seitige Webanwendung in JavaScript implementiert und kann von den Redakteuren im Browser aufgerufen werden. Der Quellcode und ein Prototyp sind öffentlich verfügbar (https://github.com/j-oe/semsim). Alle Berechnungen werden auf dem Rechner des Anwenders ausgeführt. Ergebnisse können betrachtet, gespeichert und exportiert werden.

Erfahrungen aus dem Projekt

Unter Anderem wurden folgenden Anpassungen vorgenommen:

- Integration eines Filters zum Imports von XML-Daten
- Integration eines Textvergleichs / Diffings der Quelltexte in die Ähnlichkeitsanalyse
- Dynamische Filterung und Markierung der gewählten Bausteine
- Selektive Zusammenstellung der Quelltexte für den Vergleich auf Basis der inhaltlichen Struktur





Screenshots der Anwendung zur Auswertung durch den Redakteur

Zusammenfassung & Ausblick

Optimierungspotenziale

- Erweiterung der Oberfläche zum Abgleich neuer Texte mit bestehenden Bausteinen
- Integration von Metadaten / Klassifizierungen zur Unterstützung der Entscheidungsfindung

Anwendungsszenarien

- Überarbeitung von bestehenden Dokumentation zur Reduzierung der Textbausteine
- Versionsvergleich von Dokumentationen
- Prognose der Übersetzungskosten

Literatur

- Oevermann, Jan / Ziegler, Wolfgang (2018): In: Computational Intelligence. Volume 34, Issue 1. February 2018. P. 30-48. Wiley, New Jersey, USA.
- Oevermann, Jan / Lüth, Christoph (2018): Semantically Weighted Similarity Analysis for XML-based Content Components. Proceedings of the 18th ACM Symposium on Document Engineering. DocEng 2018, Halifax, Canada. ACM, New York, USA.
- Soto, Axel J. et al. (2015): Similarity-Based Support for Text Reuse in Technical Writing. In Proceedings of the 2015 ACM Symposium on Document Engineering. P. 97-106. DocEng 2015, Lausanne, Switzerland. ACM, New York, USA.
- Ziegler, Wolfgang (2011): PI-Mod: Ein Informationsmodell (nicht nur) für den Maschinen- und Anlagenbau. Website-Eintrag. http://www.i4icm.de/forschungstransfer/pi-mod/

Kontakt:
jan.oevermann@icms.de