Transformative Datensicherung mit Hilfe eines Webcrawlers – Bachelorarbeit

Marc Grunwald

Technische Hochschule Mittelhessen

Fachbereich MND
Wilhelm-Leuschner-Straße 13
61169 Friedberg
E-Mail:
marc.grunwald@mnd.thm.de

Prof. Dr. Harald Ritz

Technische Hochschule Mittelhessen

Fachbereich MND
Wilhelm-Leuschner-Straße 13
61169 Friedberg
E-Mail:
harald.ritz@mni.thm.de

Denis Malolepszy

DenktMit eG.

Fachbereich MND
Wilhelm-Leuschner-Straße 13
61169 Friedberg
E-Mail:
denis.malolepszy@dmalo.de

Kategorie

Abschlussarbeit

Schlüsselwörter

Webcrawler, transformative Datensicherung, Wiki, CRM, Tiefensuche, Algorithmen, Datenimport - export

Zusammenfassung

Viele Unternehmen nutzen für ihr Wissensmanagement unterschiedliche Wissensmanagementsysteme in Form von Content-Management-Systemen oder Wikis und stellen diese oftmals im Intranet bereit. Da Daten in solchen Systemen langlebiger als die Software selbst sind, stellt sich die Frage, wie Daten unabhängig von externen Dienstleistern zur Verfügung gestellt werden können. Die Repräsentation der Daten im HTML-Format funktioniert bei einigen Anbietern ausschließlich, während die entsprechende Software gebucht ist. Dadurch entsteht nicht nur eine Abhängigkeit von zeitlich begrenzt lizenzierter Software, Unternehmen sind meist gegen Preisaufschläge von Lizenzen machtlos, da sonst ein Verlust der Daten droht.

Im Rahmen der folgenden Arbeit werden zunächst mehrere Wikis und CMS beleuchtet, um aufzuzeigen, welche Kosten für Unternehmen entstehen und inwieweit eine Abhängigkeit zu den vorgestellten Systemen besteht, wenn diese genutzt werden.

Anschließend wird die transformative Datensicherung mit Hilfe eines Webcrawlers als mögliche Lösung des aufgezeigten Problems präsentiert. Der Webcrawler sammelt und speichert die gerenderten Daten der jeweiligen Plattform, ein Transformationsprozess passt diese im Nachgang an. Ein Archivserver stellt diese daraufhin unabhängig vom bisherigen Anbietersystem in einem lokalen Archiv zur Verfügung. Die Daten werden hierbei im ursprünglich dargestellten Format (HTML, PDF etc.) bereitgestellt, wobei ein zusätzliches Such-Interface die Suche und Darstellung der archivierten Daten ermöglicht.

Dieses Vorgehen hat das Ziel, die Abhängigkeit von externen Anbietern zu minimieren und einem Unternehmen zusätzlich eine Perspektive zu bieten, Kosten zu sparen, ohne einen Verlust der eigenen Wissensdatenbank befürchten zu müssen.

Hierfür wird zunächst definiert, was unter dem Begriff eines Webcrawlers zu verstehen ist und welche Eigenschaften dieser mit sich bringen soll. Unter Betrachtung verschiedener Quellen wird anschließend ein solcher Webcrawler konzeptioniert und in der Programmiersprache Kotlin umgesetzt.
Abschließend werden Möglichkeiten betrachtet, inwieweit die transformative Datensicherung erweitert werden kann. Hierfür bietet sich beispielsweise die Option an, das Archiv durchsuchbar zu machen, indem es indexiert wird.

Literatur

Angelika Steger (2018): Algorithmen & Komplexität. Lektüre. Institut für Theoretische Informatik.

Antonio-Jose Aledo-Hernandez; Jose-Manuel Martinez-Caro (2018): A Comparative Study of Web Content Management Systems. Universidad Politecnia de Cartagena. Spanien. Online verfügbar unter https://www.mdpi.com/2078-2489/9/2/27.

B. Leuf; W. Cunningham (2001): The Wiki Way: Quick Collaboration on the Web: Addison-Wesley Professional.

Ferber, Reginald (2003): Information Retrieval. Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web.
1. Aufl. Heidelberg: dpunkt-Verl.

J. Cho; H. Garcia-Molina; L. Page (2008): Efficent Crawling Through URL Ordering. Hg. v. Diglib Stanford. Online verfügbar unter http://ilpubs.stanford.edu:8090/347/1/1998-51.pdf.

L. Page, S. Brin, R. Motwani; T. Winograd (1998): The Pagerank citation algorithm: bringing order to the web.: Stanford Digital Library Technologies Project.

Norbert Fuhr (1997): Information Retrieval: Springer Berlin Heidelberg (2013).

W.F Cody; J.T. Kreulen; V. Krishna (2002): The Integration of Business Intelligence and Knowledge Management.