Technische Hochschule
Brandenburg
University of
Applied Sciences
Fachbereich
Informatik und Medien

Entwicklung einer prototypischen Web-Anwendung in Shiny zur Kuratierung eines Thesaurus mit teilautomatischen Themenvorschlägen

Marie-Christin Knoll

Bachelorarbeit • Studiengang Informatik • Fachbereich Informatik und Medien • 02.02.2018

Aufgabenstellung

Ziel der Arbeit war die Entwicklung einer prototypischen Web-Anwendung mithilfe des R-Pakets „Shiny" von RStudio. Mit dieser Anwendung können Mitarbeiter der mapegy GmbH auch ohne tiefe technische Kenntnisse Datenbankinhalte kuratieren, die in der Form eines Thesaurus vorliegen. Darüber hinaus werden dem Nutzer der Web-Anwendung Themenvorschläge präsentiert, die dem Thesaurus hinzugefügt werden können.

Konzept

Die Arbeit setzt sich aus mehreren Teilen zusammen: zum einen sollte das Bearbeiten der Metadaten der Themen eines Thesaurus ermöglicht werden, der in einer Datenbank gespeichert ist.
Ein weiterer Schwerpunkt der Arbeit war die Implementierung einer Anzeige von Themenvorschlägen, die zu einem vorgegebenen Thema eine Relation haben könnten. Dafür wurden Schlüsselphrasen mit Methoden des Text-Minings aus firmeninternen Dokumenten extrahiert.
Schließlich musste als eine Grundvoraussetzung für eine funktionierende Web-Anwendung auch ein dynamisches UI umgesetzt werden.

Terminologieextraktion

Für das Erstellen von Themenvorschlägen werden Bi- und Trigramme, Schlüsselphrasen mit einer Länge von zwei oder drei Wörtern, aus internen Dokumenten herausgezogen. (Siehe Abb. 1)
Hierbei handelt es sich um einen Anwendungsfall des Text-Mining-Verfahrens der Terminologieextraktion, welche der Erkennung von Fachtermen in Texten dient.
Das Text Mining ist eine Variante des Data Minings mit Textdaten. Es dient der Analyse von Textdaten nach Informationen, dafür werden Inhalte aus digitalen Texten strukturiert und extrahiert. [Meh14]

keyphrase	occurrence	N-gram
neural network	191	2
artificial neural	149	2
artificial intelligence	14	2
risk value	12	2
network artificial	11	2
frac partial	10	2
hidden layer	10	2
neural network artificial	10	3
security threat	9	2
communication service	8	2

Showing 1 to 10 of 100 entries Previous 1 2 3 4 5 … 10 Next

Abb. 1: Themenvorschläge aus 100 Dokumenten zu „Artificial neural network"

Thesaurus

Bei einem Thesaurus handelt es sich um eine Form einer Ontologie.
Ein Thesaurus ist definiert als ein strukturiertes Konzept bestehend aus Begriffen und drei Arten von semantischen Beziehungen zwischen Begriffen , die in dem Thesaurus verwendet werden. Dies ist auch der Grund, weshalb ein Thesaurus zur Konstruktion von Ontologien verwendet wird.
Die drei Arten der Beziehungen werden in hierarchisch, äquivalent und assoziativ unterschieden. [SR+17]

Testthesaurus

Der Thesaurus, welcher über 220.000 Begriffe (Themen) enthält, wurde zu Testzwecken auf sieben Themen begrenzt, die alle eine Beziehung (Verknüpfung) zum Thema „Artificial intelligence" („Künstliche Intelligenz") haben. Dieser Testthesaurus stellt damit nur einen Bruchteil des gesamten Thesaurus dar und wurde während der Entwicklung zur schnellen Erkennung und Behebung von Fehlern in der Web-Anwendung verwendet.

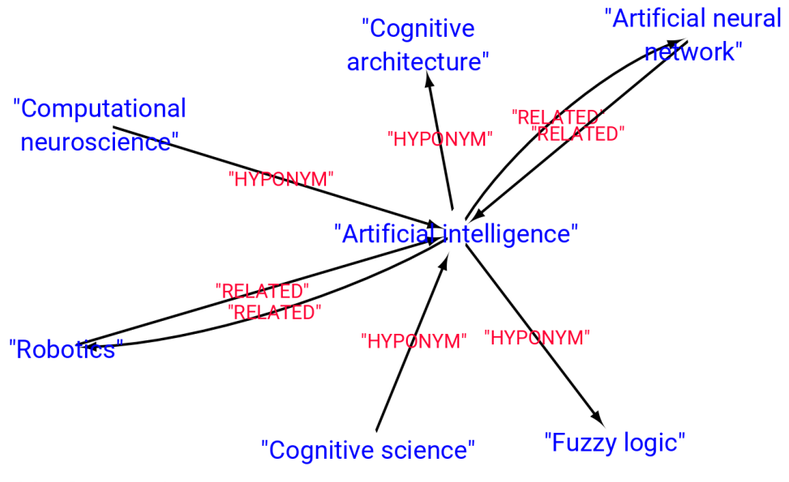

Abb. 2: Testthesaurus vor Kuratierung

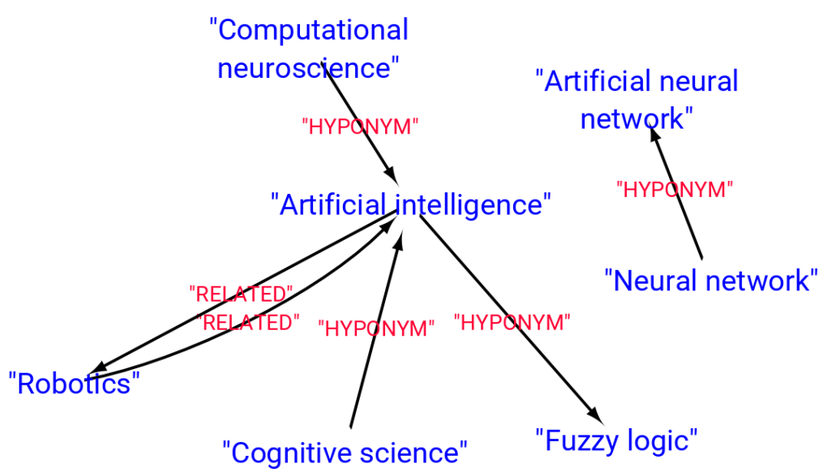

Abb. 3: Testthesaurus nach Kuratierung

Ergebnis

Die Abbildung 2 zeigt den unveränderten Testthesaurus. Der zweite Graph [Abb. 3] stellt das Resultat nach der Kuratierung dar: Mit Hilfe der Web-Anwendung wurde vom Nutzer ein vorhandenes Thema als gelöscht markiert, eine Verknüpfung zwischen zwei Themen gelöscht und ein neues Thema mit einer Verknüpfung hinzugefügt. Dabei wurde er durch automatisiert erstellte Themenvorschläge unterstützt.

Fazit

Mit dieser Arbeit wurde gezeigt, wie eine in Shiny entwickelte Web-Anwendung ein Unternehmen bei der Kuratierung ihres eigenen Thesaurus unterstützt. Das entwickelte Text-Mining-Verfahren für die teilautomatischen Themenvorschläge liefert aus den internen Dokumenten aus Sicht des Unternehmens interessante Resultate bzw. Vorschläge. Mit Hilfe der Themenvorschläge kann der Thesaurus manuell erweitert werden.

Quellen

[Meh14] Chris Biemann, Alexander Mehler.: *Text Mining: From Ontology Learning to Automated Text Processing Applications*. Springer, 2014.

[SR+17] C Sankaralingam, S Rajendran u.a.: *Onto-Thesaurus for Tamil language: Ontology based Intelligent System for Information Retrieval.* In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 13.-16. September 2017.

Betreuung: Dipl.-Inform. Ingo Boersch • Prof. Dr.-Ing. Jochen Heinsohn • Technische Hochschule Brandenburg • Dr. rer. nat. Matthias Plaue • mapegy GmbH