Technische Hochschule
Brandenburg
University of
Applied Sciences
Fachbereich
Informatik und Medien

Evaluation eines Word-Embedding-basierten Information-Retrieval-Systems

Hüseyin Çelik

Masterarbeit • Studiengang Informatik • Fachbereich Informatik und Medien • 08.05.2023

Aufgabenstellung

Im Rahmen dieser Arbeit wird eine Ausführungsvariante des von der Fraunhofer-Gesellschaft zum Patent angemeldeten Word-Embedding-basierten Information-Retrieval-Systems (IR-Systems) *Verfahren und Vorrichtung zur Ermittlung ähnlicher Dokumente* (FhG, 2019) im Vergleich zu einer Volltextsuche und einer Volltextsuche, bei der die Suchanfragen durch einen domänenspezifischen Thesaurus erweitert werden, evaluiert. Dafür werden geeignete wissenschaftliche Evaluationsmethoden identifiziert, ihre Vor- und Nachteile herausgearbeitet, ihre Evaluationsmetriken beschrieben, Möglichkeiten zur Erhebung der Relevanzbewertungen erläutert und die Auswirkungen der in der Patentschrift beschriebenen Schwellwerte untersucht.

Evaluation von Information-Retrieval-Systemen

Im Zuge dieser Arbeit werden drei verschiedene Evaluationsmethoden und ihre Evaluationsmetriken zur Beurteilung der Effektivität vorgestellt: Die Online-Evaluation, die Interactive-Evaluation und die Testdatensatz-basierte Evaluation (s. Abb. 1).

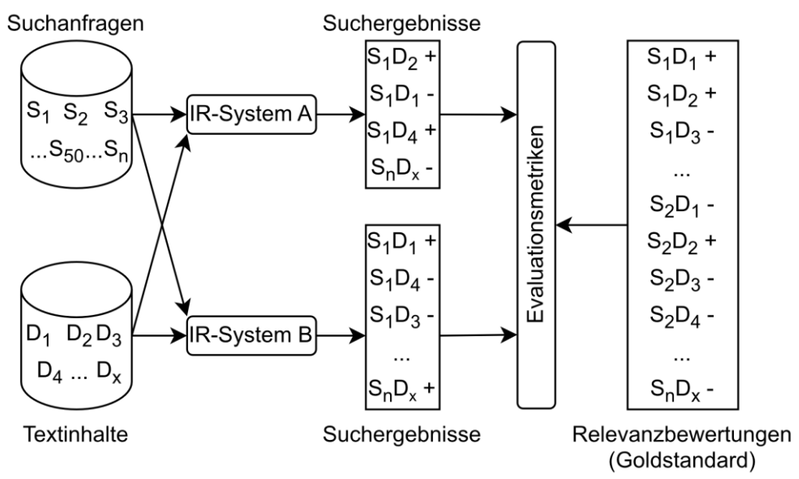

Abb. 1: Schematische Darstellung der Testdatensatz-basierten Evaluation. Eigene Darstellung basierend auf Zhai und Massung (2016, S. 169). Die Relevanzbewertungen werden durch die Symbole + (relevant) und – (nicht relevant) dargestellt.

Verfahren und Vorrichtung zur Ermittlung ähnlicher Dokumente

Das zu evaluierende Word-Embedding-basierte IR-System besteht aus zwei Hauptphasen: Der Indexierungsphase (s. Abb. 2) und der Anfragephase (s. Abb. 3). Die Indexierungsphase ist nutzerunabhängig und dient dem IR-System zur Vorverarbeitung und Vorbereitung der benötigten Textinhalte und Datenstrukturen. Zentrale Datenstruktur dieses Ansatzes sind „SimSets“, mit denen Cluster ähnlicher Word-Embeddings bezeichnet werden. Durch die Indexierungsphase soll die benötigte Zeit zur Ermittlung der Suchergebnisse bei der Nutzung des IR-Systems minimiert werden. Auf die Indexierungsphase folgt die Anfragephase, in welcher das IR-System einem Nutzer zur Verfügung steht und eine Suchanfrage übermittelt werden kann. In dieser Phase wird die Suchanfrage vorverarbeitet, die Textinhalte selektiert und anschließend nach Relevanz geordnet.

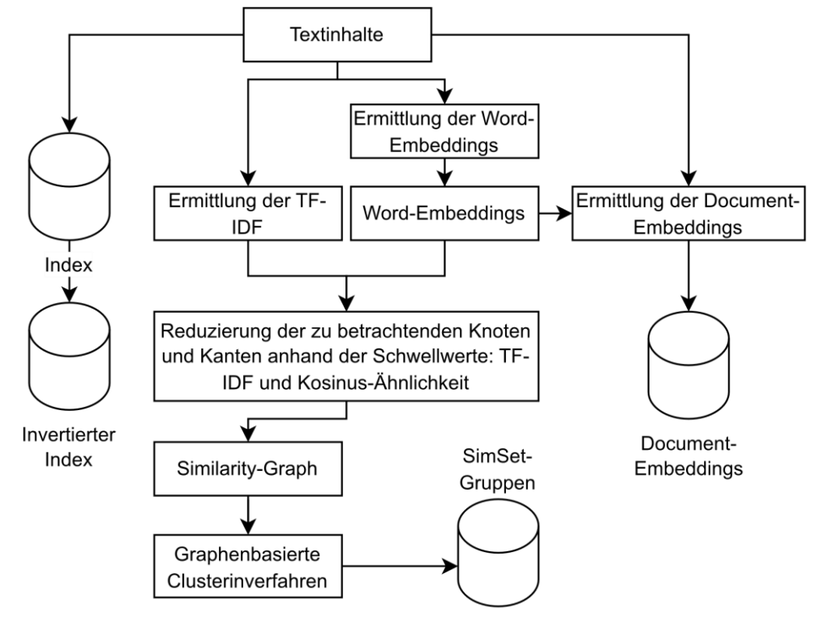

Abb. 2: Schematische Darstellung der Indexierungsphase. Basierend auf FhG (2019, S. 15-16, 18)

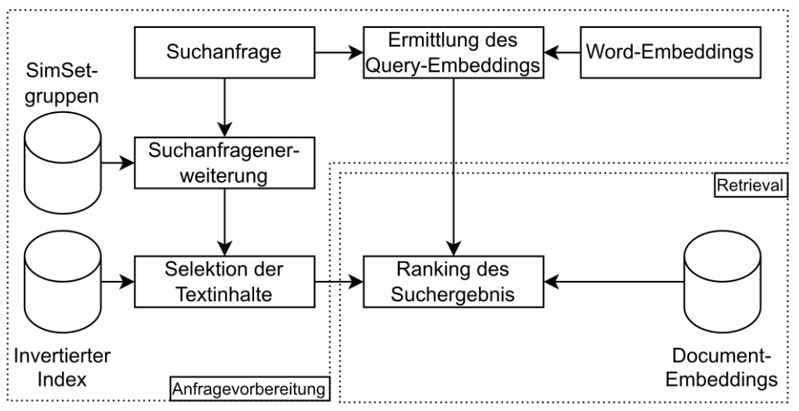

Abb. 3: Schematische Darstellung der Anfragephase

Ergebnisse der Evaluation

Die Ergebnisse zeigen erstens, dass das Word-Embedding-basierte IR-System aufgrund der Suchanfragenerweiterung durch die SimSets im Vergleich mit einer Volltextsuche zusätzliche Textinhalte ermittelt, und zweitens, dass die Document-Embeddings sowie die SimSets des Word-Embedding-basierten IR-Systems eine differenziertere Ermittlung der Reihenfolge der Textinhalte im Suchergebnis ermöglichen. Des Weiteren ist messbar, dass die Effektivität des Word-Embedding-basierten IR-Systems bei einer domänenspezifischen Textsammlung mit der Effektivität einer Volltextsuche, welche die Suchanfragen durch einen domänenspezifischen Thesaurus erweitert, vergleichbar ist. Mit den automatisch ermittelten SimSets können somit Ergebnisse erzielt werden, die mit denen einer Volltextsuche, die einen manuell modellierten domänenspezifisch Thesaurus nutzt, vergleichbar sind.

Quellen

Zhai, C. & Massung, S. (2016). Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining (Bd. 12). Association for Computing Machinery and Morgan & Claypool.

FhG (2019). Verfahren und Vorrichtung zur Ermittlung ähnlicher Dokumente (Nr. DE102019212421A1). Deutsches Patent- und Markenamt. Zugriff auf https://depatisnet.dpma.de/DepatisNet/depatisnet?action=pdf&docid=DE102019212421A1

Betreuung Prof. Dr.-Ing. Jochen Heinsohn • Technische Hochschule Brandenburg • Prof. Dr. rer. nat. Thomas Hoppe • Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V.