Information Retrieval

Seminararbeit

des Studienganges Angewandte Informatik
an der Dualen Hochschule Baden-Württemberg Mannheim

von

Jesse-Jermaine Richter, Jonas Seng

27.08.2018

Matrikelnummer, Kurs:	8787549/1980179, TINF16AIBI
Ausbildungsfirma:	DZ BANK AG, Frankfurt
Betreuer der Ausarbeitung:	Herr Prof. Dr. Karl Stroetmann

Erklärung

Wir versichern hiermit, dass wir unsere Seminararbeit mit dem Thema: „Information Retrieval“ selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt haben.

Wir versichern zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Ort, Datum	Unterschrift
Ort, Datum	Unterschrift

In dieser Seminararbeit wird das Thema „Information Retrieval“ anhand einer lokalen Suchmaschine näher erläutert...

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungstabelle

Abkürzung:	**Bedeutung:**
Abkürzung	Erklärung

1 Einleitung

1.1 Was ist Information-Retrieval?

Information-Retrieval (IR) beschreibt das Bereitstellen spezieller Informationen aus einer großen und unsortierten Datenmengen. Dieses Themengebiet fällt unter Informatik, Informationswissenschaften sowie Computerlinguistik und ist ein wesentlicher Bestandteil von Suchmaschinen wie Google.
Das Thema besitzt bereits seit einigen Jahren eine hohe, aber dennoch steigende Relevanz. Die Gründe der hohen Relevanz von IR liegen vor allem beim Einsatz von Suchmaschinen. Diese sind in Zeiten des Internets die wohl wichtigste Form der Informationsbeschaffung - und das in Bruchteilen von Sekunden. Aufgrund der immer schneller steigenden Informationsmengen wird das Thema künftig weiter an Relevanz gewinnen. Unternehmen, ebenso wie Privatanwender, wird eine immer weiter wachsende Menge von Informationen zugänglich, die organisiert werden muss, damit relevante bzw. spezifisch gesuchte Informationen jederzeit und ohne Verzögerung gefunden werden kann.
Um das Ziel der Bereitstellung von Informationen gewährleisten zu können, werden sämtliche Informationen bzw. Dokumente, welche später gefunden werden können sollen, durchsucht und gewichtet. Das zentrale Objekt der Informationsrückgewinnung stellt der invertierte Index dar, dessen Aufbau und Funktionsweise in den nächsten Kapiteln ausführlich erläutert wird. Weiter wird im Verlauf dieser Arbeit die Komprimierung des Indexes sowie das Tf-idf-Maß, welches zur Beurteilung der Relevanz eines Dokumentes genutzt wird, im Fokus stehen.
Die theoretischen Hintergründe des invertierten Index, der Komprimierung und des Tf-idf-Maß werden durch eine Beispiel-Implementierung einer lokalen Suchmaschine in Programmiersprache Python veranschaulicht.

1.2 Ziel der Arbeit

Ziel der Arbeit soll es sein, ein grundlegendes Verständnis des Themenkomplexes Information-Retrieval zu vermitteln. Das umfasst einerseits die theoretischen Hintergründe, die für die später vorgestellte Beispielimplementierung notwendig sind, sowie die Vorstellung der Beispielimplementierung an sich.
Die Beispielimplementierung soll hauptsächlich die folgenden Themengebiete umfassen:

- Aufbau eines invertierten Indexes
- Approximierende Beurteilung der Relevanz eines gefundenen Dokuments mittels tf-idf
- Komprimierung des invertierten Indexes

Die in dieser Arbeit vorgestellte Implementierung hat nicht den Anspruch auf hohe Performance, vielmehr dient diese dem Zwecke der praxisnahen Veranschaulichung der Funktionsweise von IR-Systemen.

1.3 Stand der Forschung

Dieser Abschnitt soll den aktuellen Stand der Forschung kurz umreißen. Es sollen dazu zwei Modelle von Information Retrieval knapp beschrieben werden, die für die Entwicklung einer lokalen Suchmaschine, von Bedeutung sind. Es wird jedoch nur ein Modell im Verlauf dieser Arbeit gezeigt.

1.3.1 Vector Space Model

Das Vector Space Model, zu deutsch Vektorraummodell, repräsentiert Dokumente und Anfragen als hochdimensionale, metrische Vektoren [2]. Der Anfrage-Vektor wird beim Retrieval-Prozess mit den Dokumenten-Vektoren verglichen. Dabei werden jedoch nur Dokumente betrachtet, welche mit der Anfrage in Verbindung stehen könnten [3]. Welche Dokumente mit der Anfrage in Verbindung stehen könnten, wird mithilfe des invertierten Index ermittelt.
Es gibt verschiedene Maße, mit denen die Vektoren miteinander verglichen werden können. Der einfachste Ansatz besteht darin, den Abstand zu berechnen, jedoch ist dies kein sehr gutes Maß. Besser und weit verbreitet ist deshalb das Cosinus-Maß (heißt das echt so?!), welches den Winkel zwischen Anfrage-Vektor und Dokumenten-Vektor angibt. Je kleiner der Winkel, desto höher ist die Relevanz des Dokuments [1].

1.3.2 Probabilistische Ansätze

Literatur

[1] *Information Retrieval*. 2007. URL: http://www.is.informatik.uni-duisburg.de/courses/ie_ss07/folien/folien-ir.pdf (siehe S. 8).

[2] *Vektorraum-Retrieval*. 2017. URL: https://de.wikipedia.org/wiki/Vektorraum-Retrieval (siehe S. 8).

[3] *Klassische Information Retrieval Modelle Einführung* (siehe S. 8).