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 [6]. 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 [9]. 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 [2].
Dieses Modell wir im Rahmen dieser Arbeit genauer beleuchtet und als Grundlage für die Beispielimplementierung genutzt.

1.3.2 Probabilistische Ansätze

Probabilistische Ansätze basieren auf Wahrscheinlichkeiten. Hierbei wird eine Abschätzung der Wahrscheinlichkeit berechnet, mit der ein Dokument d bezüglich einer Anfrage q relevant ist [2]. Zur Abschätzung der Wahrscheinlichkeit gibt es verschiedene Ansätze, die hier jedoch nicht weiter thematisiert werden sollen. Bei allen praktisch nutzbaren Ansätzen sind jedoch eine - je nach Ansatz - große oder kleine Menge von Zusatzinformationen über die Dokumentenkollektion nötig [2].

2 Information Retrieval - Theoretische Grundlagen

2.1 Problemstellung

Wie in der Einleitung bereits angemerkt, beschreibt „Information Retrieval“das Bereitstellen spezieller Informationen aus einer meist großen, unsortierten Datenmenge. Dabei bekommt das System eine vom Nutzer gestellte Query (Abfrage) und versucht auf dessen Basis, Daten, die meist als Dokumente vorliegen, zurückzuliefern. Im Gegensatz zu Abfragen im Datenbankumfeld beinhaltet die Query jedoch keinerlei Informationen, um ein spezielles Element eindeutig identifizieren zu können. Dies soll ein IR-System auch nicht leisten. Vielmehr sollen Ergebnisse zurückgeliefert werden, die mit hoher Wahrscheinlichkeit Relevanz bzgl. der gestellten Query besitzen. Der Nutzer selektiert dann die für diesen nötigen Dokumente.
Mathematisch lässt sich dies folgendermaßen formulieren: Aus einer Dokumentenmenge D soll mithilfe einer Funktion eine Teilmenge D_1 von D ermittelt werden, die relevant für eine Abfrage q ist.
Um diese Funktion sinnvoll definieren zu können, muss jedoch zuvor die Menge aller Queries, sowie die Menge aller Tokens definiert werden:

Definition 2.1 *Sei $d \in D$ ein Dokument. Die Menge T_d ist nun die Menge aller Wörter, die in dem Dokument d enthalten sind: $T_d = \{t_1, .., t_n\}$.*
Die Menge T ist die Menge aller Terme, die in den Dokumenten aus D vorkommen, also: $T = T_{d1} \cup .. \cup T_{dn}$ mit $d_i \in D$.

Definition 2.2 *Mithilfe der letzten Definition kann nun die Menge aller möglichen Queries definiert werden: $Q \subseteq 2^T$*

Definition 2.3 *Eine Funktion $f: Q \to D_1$ heißt Retrievalfunktion, wobei $D_1 \subseteq D$ gilt und Q die Menge aller Queries ist.*

Nachdem die Problemstellung formuliert ist, muss eine Strategie entwickelt werden, wie die Funktion f dargestellt bzw. umgesetzt werden kann.

2.2 Strategiefindung

Dieser Abschnitt soll eine Übersicht bieten, wie das im Folgenden vorgestellte IR-System arbeiten soll.
Als Vorarbeit müssen alle Dokumente, die im Index aufgenommen werden sollen, in eine Codierung wie ASCII oder Unicode umgewandelt werden. Dazu wird ein Tool genutzt, das hier nicht weiter von Relevanz sein wird. Es sollen mindestens all diejenigen Dokumente in den Index aufgenommen werden, die im PDF-Format vorliegen.
Der erste Schritt, der das IR-System an sich leisten muss, ist das Erstellen von Tokens. Dazu wird jedes Dokument in Tokens aufgespalten. Ein Token ist in den meisten Fällen ein Wort, Satzzeichen wie Leerzeichen, Kommata usw. sollen nicht als Tokens behandelt werden und werden ignoriert.
Für jeden Token wird es später im Index einen Eintrag geben, der eine Liste mit weiteren Informationen hält. Diese Liste muss mindestens die Dokument-ID speichern, in dem das Token steht. In diesen Listen werden häufig noch weitere Informationen hinterlegt, beispielsweise die Häufigkeit eines Tokens.
Der zweite große Schritt besteht darin, einen Algorithmus zu entwerfen, der eine Query entgegennimmt und auf Basis der Query und des Index eine Liste von relevanten Dokumenten ausgibt. Dieser wird das in der Einleitung kurz vorgestellte Vektorraummodell verwenden. Weiter wird dieser für die Ermittlung der Relevanz die sogenannte TF-IDF-Gewichtung nutzen. Diese wird später noch ausführlich vorgestellt.
Neben diesen beiden Punkten wird der Index komprimiert, um Speicherplatz zu sparen und die Performance zu erhöhen.

2.3 Tokenization

2.3.1 Vorarbeiten

Bevor aus Dokumenten Tokens erzeugt werden können, müssen einige Fragen beantwortet werden. Eine Frage ist, welche Dokumente betrachtet werden sollen und wie man ein Dokument definiert. Zur Veranschaulichung dieses Problems, soll ein Beispiel dienen:
Angenommen das IR-System soll dazu dienen Dokumente auf der Festplatte eines Computers zu finden. In diesem Szenario kann jede in einem Ordner gelistet Datei als Dokument angesehen werden. Dies wäre der einfachste Fall. Jedoch ist dies meist nicht erwünscht. So sollen beispielsweise bestimmte Dateitypen von der Suche ausgeschlossen werden. In UNIX existiert ein Dateityp, welcher Mails speichert. Hier soll jede Mail als einzelnes Dokument angesehen werden, die Datei muss also in mehrere Dokumente aufgespalten werden [4]. Umgekehrt gibt es Szenarien, in denen mehrere Dokumente zu einem Dokument zusammengefasst werden müssen, um bei der Suche nutzbare Ergebnisse zu erzielen [4].
Ein weiteres Problem, das gelöst werden muss, um die Dokumente verarbeiten zu können, ist die Codierung der Inhalte der Dokumente. Hierbei müssen Dokumente, die meist in vielen unterschiedlichen Codierungen vorliegen, zu einer definierten Codierung überführt werden [4].
Diese Probleme der „Vorarbeit“werden in der später gezeigten Beispielimplementierung nicht behandelt, sondern von anderen Tools behandelt.

2.3.2 Tokenerzeugung

Sobald definiert ist, was als Dokument verstanden wird (engl. document unit) und ist eine einheitliche Codierung der Dokumente vereinbart, können die Dokumente in Tokens aufgeteilt werden.
Zunächst jedoch drei wichtige Definitionen:

Definition 2.4 *Unter einem Token wird eine zusammenhängende Zeichenkette verstanden, die innerhalb eines Dokuments vorkommt [4].*

Definition 2.5 *Ein Typ bezeichnet eine Klasse von Tokens, die dieselben Zeichen enthalten [4].*

Definition 2.6 *Ein Term ist ein Typ, welcher im Dictionary eines IR-System vorkommt [4].*

Eine wichtige Frage, die im Rahmen der Tokenerzeugung geklärt werden muss, ist, welche Zeichenketten als Token behandelt werden. Kommata, Punkte und sonstige Satzzeichen haben keine sinnvolle Bedeutung im Zusammenhang mit Information Retrieval, diese Zeichen können somit aus Tokens entfernt bzw. während der Tokenerzeugung überlesen werden [4].

Der Text

Beispielsatz, der ein Komma hat.

erzeugt diese Tokenmenge:

$$Tokens = \{\textit{Beispielsatz}, \textit{der}, \textit{ein}, \textit{Komma}, \textit{hat}\}$$

Einige Information Retrieval-System nutzen darüber hinaus sogenannte „stop words“. Das sind Wörter, die in sehr vielen Dokumenten in großer Anzahl vorkommen und damit wenig Bedeutung für die Suche besitzen [4]. Beispiele für solche Wörter sind „ist“, „sein“ und „und“. Jedoch funktioniert diese Technik später beim Suchen schlechter als zunächst angenommen. Das Wort „sein“ kann beispielsweise als Verb oder als Pronomen in einem Dokument vorkommen. Als Pronomen kann dieses Wort durchaus wichtig sein für eine Suche (beispielsweise innerhalb eines Buchtitels), wird jedoch als stop word aussortiert. Daher werden in neuen IR-Systemen entweder gar keine stop words oder nur eine geringe Anzahl stop words genutzt.
Eine weitere Möglichkeit solche Wörter zu filtern ist Stemming. Diese Methode führt Wörter auf ihren Wortstamm zurück [4] [3]. Dadurch wird die Anzahl der Terme, die im Index gespeichert werden müssen, stark gesenkt. Allerdings bringt diese Methode eine Unschärfe mit sich. Damit ist gemeint, dass zwei nicht verwandte Wörter auf denselben Wortstamm zurückgeführt werden, wodurch bei der späteren Suche nach einem der beiden Urpsrungswörter auch Ergebnisse zurückgeliefert werden, die irrelevant für die Query sind [4].

Weiter existieren sprachspezifische Probleme. Beispielsweise wird im Englischen häufig mit Kurzformen von Wörtern gearbeitet, so wird „are not“zu „aren't“. Es muss geklärt werden wie mit solchen Formen umgegangen werden soll. Ein Ansatz ist Query Preprocessing. Hierbei werden Wörter dieser Art, die in der Query stehen, wie während der Tokenerzeugung in eine einheitliche Form gebracht [4]. Darüber hinaus gilt es zu beachten, dass Eigennamen wie „Hewlett-Packard“nicht oder nur nach bestimmten Regeln prozessiert werden dürfen. Welche Wörter als Eigenname behandelt werden und welche nicht, kann mittels Machine Learning-Verfahren oder auf Basis eines großen Vokabulars gelöst werden.

2.4 Invertierter Index

Dieser Abschnitt wird die Idee des invertierten Index vorstellen sowie die Erzeugung des invertierten Index erläutern. Darüber hinaus wird aufgezeigt, wie die Verarbeitung einer Query mittels des invertierten Index funktioniert. Ab jetzt werden die Ausdrücke „invertierter Index“und „Index“synonym verwendet.

2.4.1 Grundlegender Aufbau

Der invertierte Index kann als eine Liste betrachtet werden, welche für jeden Term, der in der Dokumentenmenge vorkommt, einen Eintrag hält [4] [3]. Dieser Eintrag wiederum ist eine weitere Liste. Diese Liste hält mindestens eine eindeutige Dokumenten-ID, meist jedoch darüber hinaus weitere Informationen. Beispiele für solche weiteren Informationen sind Häufigkeit, in der ein Term in einem Dokument vorkommt, die Position und die umliegenden Wörter in der Nähe zum Term t [4].
Den invertierten Index kann man folgendermaßen visualisieren:

Nummer	Term	Dokumente
1	als	5
2	auch	3
3	das	1, 2, 6
4	der	1, 2, 5
5	dritter	3
6	ein	3
7	ende	6
8	er	4
9	erste	1, 5
10	folgt	3
11	hier	2, 6
12	ist	1, 2, 6
13	länger	4
14	satz	1, 3
15	und	2, 5, 6
16	wobei	4
17	zweite	2, 5

Abbildung 2.1: Beispiel für einen invertierten Index [1]

2.4.2 Umsetzung eines invertierten Index

Nachdem das Prinzip des invertierten Index klar ist, steht die Frage im Raum wie dieser umgesetzt werden kann. Es liegt nahe als Datenstruktur einen Trie einzusetzen. Ein Trie ist ein spezieller Suchbaum, welcher besonders gut zum Suchen von Zeichenketten geeignet ist [4].
Alternativ kann über den Einsatz einer Hasmap nachgedacht werden, jedoch führt dies zu folgendem Problem: Angenommen eine vom Nutzer eingegebene Query ist nicht in der Hashmap vorhanden. Die Hash-Funktion wird einen Fehler werfen. Da in einer Hashmap Wörter, die ähnlich zueinander sind, nicht unbedingt benachbart gespeichert werden und keinerlei Information darüber bekannt ist, wo zur Query ähnliche Wörter gespeichert sind, kann im Falle, dass ein oder mehrere Wörter der Query nicht vorhanden sind, nicht mit geringem Aufwand nach ähnlichen Wörtern gesucht werden. Bei Tries besteht dieses Problem nicht [4].
Da in der Beispielimplementierung ein Trie als Datenstruktur zum Einsatz kommen wird, soll diese kurz vorgestellt werden.

Tries

Ein Trie wird auf der Basis einer Menge von Zeichenketten aufgebaut. Jede Zeichenkette, die gefunden werden muss, ist innerhalb des Tries repräsentiert.
Erreicht wird dies dadurch, dass ein Knoten jeweils ein Zeichen repräsentiert und eine Liste mit Verweisen auf die nächsten möglichen Knoten hält, basierend auf einem weiteren Zeichen. Die folgende Abbildung zeigt einen Trie.

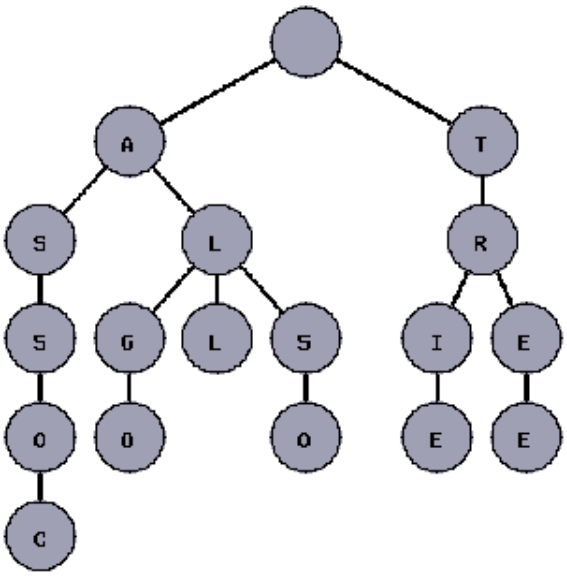

Abbildung 2.2: Beispiel eines Tries [8]

Im Folgenden soll eine formale Definition eines Tries gegeben werden:

Definition 2.7 *Sei $\sum$ eine endliche Menge von Zeichen (Alphabet) und $\sum^*$ die Menge aller Wörter, die über $\sum$ gebildet werden können. Sei $S \subseteq \sum^*$. Dann ist $T = (V, E)$ ein Trie, wobei V die Menge aller Knoten und E die Menge aller Kanten ist. Darüber hinaus muss gelten:*

- $\forall e \in E$: *e ist mit Zeichen aus $\sum$ beschriftet.*
- $\forall v \in V$: *alle ausgehenden Kanten von v sind unterschiedlich beschriftet mit einem Zeichen $z \in \sum$*
- $\forall S_i \in S : \exists v \in V$: *S_i ist ein Präfix der Konkatenation der Beschriftungen des Pfades vom Wurzelknoten bis v.*
- $\forall b \in V : \exists S_i \in S$: *Die Konkatenation der Beschriftungen von der Wurzel bis b ergibt S_i, sofern b ein Blatt des Tries ist.*

Definition 27 ist aus [7] entnommen.

Das Suchen nach gespeicherten Wörtern gestaltet sich nun verhältnismäßig einfach: Um das Wort „Tree“im, in der Abbildung gezeigten, Trie zu finden, wird wie folgt vorgegangen. Stellt der User die Anfrage „Tree“wird diese Query nun Zeichen für Zeichen durchgegangen. Beginnend bei „T“wird im Wurzelknoten geprüft, ob es einen Verweis auf einen Knoten gibt, der ein „T“repräsentiert [8]. Existiert ein solcher Knoten, wird in diesem geprüft, ob es einen Knoten gibt, der das nächste Zeichen in der Query (das „r“) repräsentiert. Ist dies der Fall wird das Verfahren solange wiederholt bis die Query komplett eingelesen ist oder in der Query ein Zeichen steht, das durch keinen Knoten im Trie repräsentiert ist [8] [5].

Ähnlich leicht funktioniert das Einfügen neuer Wörter in den Trie. Dazu wird - wie beim Suchen - das Wort, das eingefügt werden soll, so weit wie möglich nach dem oben beschriebenen Muster eingelesen und es wird zu den entsprechenden Knoten gesprungen [8] [5]. Wird nun ein Zeichen eingelesen, das nicht durch einen Knoten repräsentiert ist, wird ein neuer Knoten erzeugt, welcher dieses Zeichen repräsentiert [8]. Alle nun noch einzulesenden Zeichen erhalten einen neuen Knoten, da der neu erzeugte Knoten natürlich nicht auf bereits vorhandene Knoten zeigen kann. Innerhalb dieses Knotens wird eine Liste angelegt, die Verweise auf weitere Knoten hält [8]. In dieser Liste wird ein Verweis auf den Knoten angelegt, der das nächste Zeichen des einzufügenden Wortes repräsentiert. Dieses Verfahren setzt sich solange fort, bis das neue Wort vollständig eingelesen ist.

Das Löschen soll in diesem Rahmen nicht aufgezeigt werden, da dies weitaus komplexer sein kann als das Finden oder Einfügen von Einträgen.

Ein Knoten, zu dem man mit dem Wort S_i gelangt, muss außerdem eine Liste mit Verweisen auf alle Dokumente, in denen das Wort S_i vorkommt, speichern. Gibt es keine Dokumente, in denen S_i vorkommt, ist die Liste leer.

2.5 Komprimierung des Index

Bei großen Dokumentenmengen wächst die Größe des Index ebenfalls. Doch nicht nur die Größe des Index wächst, sondern auch die benötigte Zeit, um auf eine Query zu antworten [4]. Die Index-Komprimierung adressiert genau dieses Problem. Über die Jahre der Forschung haben sich einige Komprimierungs-Techniken bewährt und sind bei den allermeisten aktuellen Suchmaschinen in Benutzung.
Dieser Abschnitt soll das Thema Komprimierung erläutern, das Hauptaugenmerk liegt dabei auf der Komprimierungs-Technik, die in der Beispielimplementierung eingesetzt wird.

2.6 TF-IDF Gewichtung

2.7 Retrieval

Literatur

[1] *Indexierung mittels invertierter Dateien*. 2001. URL: http://www.cis.uni-muenchen.de/people/Schulz/SeminarSoSe2001IR/Nagy/node4.html (siehe S. 12).

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

[3] Andreas Henrich. *Information Retrieval 1*. 2008 (siehe S. 11, 12).

[4] Hinrich Schütze Cristopher D. Manning Prabhakar Raghavan. *Introduction to Information Retrieval*. 2009 (siehe S. 10–13, 15).

[5] *Trying to Understand Tries*. 2017. URL: https://medium.com/basecs/trying-to-understand-tries-3ec6bede0014 (siehe S. 14).

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

[7] *Trie*. 2018. URL: https://de.wikipedia.org/wiki/Trie (siehe S. 14).

[8] *Introduction to Trie*. URL: https://guide.freecodecamp.org/miscellaneous/data-structure-trie/ (siehe S. 13, 14).

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