Informatik
Hauptcampus

H O C H
S C H U L E
T R I E R

Konzeption und Realisierung eines Systems zur Informationssuche in einem Dokumentenarchiv basierend auf Textinhalt und Metadaten.

Conception and Realization of an Information Retrival System for a Document Archive based on Text Content and Metadata

Annika Kremer

Bachelor-Abschlussarbeit

Betreuer: Prof. Dr. Karl Hans Bläsius

Trier, Abgabedatum

Vorwort

Ein Vorwort ist nicht unbedingt nötig. Falls Sie ein Vorwort schreiben, so ist dies der Platz, um z.B. die Firma vorzustellen, in der diese Arbeit entstanden ist, oder einigen Leuten zu danken, die in irgendeiner Form positiv zur Entstehung dieser Arbeit beigetragen haben. Auf keinen Fall sollten Sie im Vorwort die Aufgabenstellung näher erläutern oder vertieft auf technische Sachverhalte eingehen.

Kurzfassung

In der Kurzfassung soll in kurzer und prägnanter Weise der wesentliche Inhalt der Arbeit beschrieben werden. Dazu zählen vor allem eine kurze Aufgabenbeschreibung, der Lösungsansatz sowie die wesentlichen Ergebnisse der Arbeit. Ein häufiger Fehler für die Kurzfassung ist, dass lediglich die Aufgabenbeschreibung (d.h. das Problem) in Kurzform vorgelegt wird. Die Kurzfassung soll aber die gesamte Arbeit widerspiegeln. Deshalb sind vor allem die erzielten Ergebnisse darzustellen. Die Kurzfassung soll etwa eine halbe bis ganze DIN-A4-Seite umfassen.

Hinweis: Schreiben Sie die Kurzfassung am Ende der Arbeit, denn eventuell ist Ihnen beim Schreiben erst vollends klar geworden, was das Wesentliche der Arbeit ist bzw. welche Schwerpunkte Sie bei der Arbeit gesetzt haben. Andernfalls laufen Sie Gefahr, dass die Kurzfassung nicht zum Rest der Arbeit passt.

The same in english.

Inhaltsverzeichnis

1 Einleitung und Problemstellung 1
- 1.1 Einleitung 1
- 1.2 Problemstellung 1
- 1.3 Teilprobleme 2
  - 1.3.1 Dynamisches einlesen der Metadaten 2
  - 1.3.2 Unterscheidung Metadaten und Freitext 2
  - 1.3.3 Metadatensuche 2
  - 1.3.4 Freitextsuche 2
  - 1.3.5 Verknüpfung mit UND/ODER 3
  - 1.3.6 Benutzeroberfläche 3

2 Information Retrieval 4
- 2.1 Bedeutung 4
  - 2.1.1 Information 4
  - 2.1.2 Information Retrieval 5
  - 2.1.3 Unterschied zur Datenbankensuche 5
- 2.2 Beispiel Websuche 5
- 2.3 Bezug zur Problemstellung 6
  - 2.3.1 Teilweise strukturierte Daten 6

3 Boolesches Retrieval 7
- 3.1 Eigenschaften des Verfahrens 7
- 3.2 Funktionsprinzip 7
  - 3.2.1 Attribut 7
  - 3.2.2 Anfragen 8
- 3.3 Notwendige Begriffe 9
  - 3.3.1 Term 9
  - 3.3.2 Index 9
- 3.4 Implementierungsansätze 9
  - 3.4.1 Inzidenz-Matrix 9
  - 3.4.2 Invertierte Liste 11

4 Das Vektorraummodell 14
4.1 Funktionsprinzip 14
4.2 Vektor und Vektorraum 14
4.2.1 Definition Vektor 14
4.2.2 Definiton Vektorraum 14
4.3 Definition Vektorraummodell 14
4.4 Gewichte 15
4.4.1 Termhäufigkeit 15
4.4.2 Dokumenthäufigkeit 16
4.4.3 Invertierte Dokumenthäufigkeit 16
4.4.4 TF-IDF-Gewichtung 16
4.5 Anfragen 17
4.6 Ähnlichkeitsfunktion 17
4.6.1 Euklidischer Abstand 17
4.6.2 Cosinus-Maß 18
4.6.3 Alternativen 19

5 Bewertung eines Information Retrieval Systems 20
5.1 Precisison 20
5.2 Recall 20

6 Implementierung 21

7 Die Benutzeroberfläche 22

8 Zusammenfassung und Ausblick 23

Literaturverzeichnis 24

Glossar 25

Erklärung der Kandidatin / des Kandidaten 26

Abbildungsverzeichnis

3.1 Term-Dokument-Inzidenz-Matrix. Die Zeile enthalten die Terme, die Spalten die Dokumente (hier durch deren *docID* repräsentiert). Alle Einträge mit einer 0 sind leere Einträge (Eigene Abbildung). . . 10
3.2 Invertierte Listen zu Beispieltermen. Die Zahlen sind die Dokumentindizes, in denen der jeweilige Term vorkommt (Eigene Abbildung). 11

4.1 Vektorraum mit den Termen $T1$ und $T2$ als Achsen, drei Dokumentvektoren und dem Anfragevektor. Das Cosinus-Maß liefert als ähnlichsten Dokumentvektor d_2 (Eigene Abbildung). 19

Tabellenverzeichnis

1

Einleitung und Problemstellung

1.1 Einleitung

Nahezu jeder nutzt heutzutage E-Mail-Dienste und kennt die Problematik, dass der Posteingang sich unter der Flut täglich eintreffender Nachrichten stetig füllt, bis der Ordner unübersichtlich voll ist.
Dies wird zu einem ernsthaften Problem, sobald etwas bestimmtes darin wiedergefunden werden soll.
Die benötigte Mail war wichtig, weil sie eine bestimmte Kontaktadresse enthielt, aber wie war nochmal der Absender? Längst vergessen. Das genaue Datum? Leider ist nur noch der Monat bekannt. Wer die Suche manuell bewältigen muss, ist an dieser Stelle verloren.
Abhilfe schafft ein Information Retrieval System, mit dem man bestimmte Suchkriterien eingeben und beliebig miteinander kombinieren kann.
Das System liefert idealerweise eine Reihe passender Resultate und der Anwender braucht sich nicht selbst durch hunderte von Mails durchzuarbeiten.
In diesem Fall könnte die Person beispielsweise im Freitext nach dem Wort *„Kontaktadresse“* suchen und weitere Kriterien wie *„Datum = Juni “* hinzufügen.
Besonderheit ist das beliebige Verknüpfen: Der Anwender kann sich entscheiden, ob er nur Resultate akzeptiert, auf die Beides zutrifft, oder ob es bereits reicht, wenn eines der Kriterien erfüllt ist. Dies erlaubt eine sehr individuelle, auf die Informationsbedürfnisse des Nutzers zugeschnittene Suche.
Ein solches System ist nicht nur für das Alltagsbeispiel E-Mail-Ordner, d.h. Posteingang, Postausgang etc. wertvoll, sondern lässt sich auch auf jede andere Art von Dokumentenarchiv, dessen Dokumente sowohl Freitext als auch Metadaten beinhalten, anwenden.

1.2 Problemstellung

Ziel der Arbeit ist die Konzeption und Realisierung eines Systems zur Informationssuche (engl. Information Retrieval System) in einem Dokumentenarchiv, wobei die Dokumente von teilweise strukturierter Natur sind.
Dies bedeutet, dass sie sowohl gewöhnlichen Freitext als auch Metadaten enthalten. Der Nutzer soll spezifizieren können, in welchen Metadaten er suchen möchte,

zudem soll die Freitextsuche auswählbar sein. Alle Suchanfragen sollen hierbei beliebig mit den booleschen Operatoren *AND* (engl. und) sowie *OR* (engl. oder) verknüpfbar sein.
Hauptanwendungszweck des Systems sind E-Mail-Archive wie Posteingang und Postausgang, allerdings soll das System so flexibel sein, dass es auch auf andere teilweise strukturierte Dokumentenarchive anwendbar ist.

1.3 Teilprobleme

Aus der Aufgabenstellung ergeben sich die im folgenden beschriebenen Teilprobleme.

1.3.1 Dynamisches einlesen der Metadaten

Die genauen Metadaten sind, da das System flexibel sein soll, vor dem Ausführen des Systems noch nicht bekannt. Demnach muss das IR-System die Namen der Metadaten beim Starten des Programms dynamisch einlesen und diese dem Nutzer auf der grafischen Oberfläche anzeigen.

1.3.2 Unterscheidung Metadaten und Freitext

Damit das System die Metadaten dynamisch einlesen kann, müssen die folgenden Punkte erfüllt sein:

1. Das System muss zwischen Metadaten und Freitext unterscheiden können.
2. Metadaten setzen sich aus Name und Inhalt zusammen, weshalb beides erkannt und voneinander abgegrenzt werden muss. Im folgenden werden die Namen als „Keywords “ bezeichnet.
3. Der Inhalt kann unterschiedlichen Datentyps sein, z.B. String oder Liste, weshalb dieser bestimmt werden muss.

1.3.3 Metadatensuche

Es muss erkannt werden, welche Metadaten der Nutzer ausgewählt hat und in genau diesen muss, unter Berücksichtigung des jeweiligen Datentyps der Inhalte, gesucht werden.
Im Gegensatz zur Freitextsuche muss zu jedem Dokument vor der Suche zunächst geprüft werden, ob das entsprechende Schlüsselwort überhaupt darin auftritt.

1.3.4 Freitextsuche

Bei der Freitextsuche ist die Wortzahl weitaus größer als bei der Metadatensuche. Daraus resultieren zwei Probleme:

1. Wie kann effizient in großen Wortmengen gesucht werden?
2. Wie kann die Suche bei begrenztem Speicher bewältigt werden?

Zudem stellt sich die Frage nach einem geeigneten Verfahren, das bei längeren Anfragen auch teilweise passende Ergebnisse liefert und messen kann, wie gut die erzielten Treffer zur Nutzeranfrage passen.

1.3.5 Verknüpfung mit UND/ODER

Alle Anfragen sollen beliebig mit den logischen Operatoren *AND*, *OR* sowie *NOT* verknüpfbar sein. Dies beinhaltet die folgenden Problemstellungen:

1. Keywordsuche und Freitextsuche müssen miteinander verknüpft werden.
2. Sind meherere Keywords ausgewählt, müssen die Teilergebnisse verknüpft werden.
3. Stellt der Nutzer mehrere Anfragen, müssen die Ergebnisse der einzelnen Anfragen verknüpft werden.

1.3.6 Benutzeroberfläche

Der Nutzer benötigt eine verständliche Benutzeroberfläche, die es ihm ermöglicht, seine Suchanfragen beliebig zusammenzustellen. Hierzu muss die Oberfläche folgende grundlegenden Funktionalitäten aufweisen:

1. Das Suchverzeichnis, d.h. das Dokumentenarchiv in welchem die Suche stattfindet, muss auswählbar sein.
2. Alle im Archiv auftretenden Keywords sowie die Freitextsuche müssen auswählbar sein.
3. Logische Operatoren (AND,OR,NOT) zur Verknüpfung müssen auswählbar sein.
4. Die Suchanfrage muss für den Nutzer verständlich angezeigt werden.

2

Information Retrieval

Dieses Kapitel soll dem Leser einen Überblick über die Bedeutung des Begriffs „Information Retrieval “ vermitteln.

2.1 Bedeutung

Der englische Begriff „Information Retrieval “ lässt sich mit „Informationsrückgewinnung“ ins Deutsche übersetzen. [Aca] Hierbei wird explizit von *Rück*gewinnung gesprochen, da keine neuen Informationen erzeugt werden.
Bevor erklärt wird, wofür Informationsrückgewinnung tatsächlich steht, wird zunächst auf den Teilbegriff „Information“ eingegangen.

2.1.1 Information

Die Bedeutung des Begriffs Information ist sehr weit gefasst, was eine einheitliche Definition unmöglich macht.
Er stammt aus dem Lateinischen (informare = Gestalt geben) und heißt im Übertragenen Sinne so viel wie jemanden durch Unterweisung bilden.
Dies betont den Aspekt, dass eine Information stets einen Empfänger besitzt, welcher „gebildet“ wird. Dies kann eine Person, aber auch ein geeignetes, nach außen wirksames System sein.
Daten als Träger von Informationen müssen vom Empfänger aufgenommen und korrekt interpretiert werden, damit aus den Daten tatsächlich Informationen entstehen.
Die Informationen müssen deshalb auf irgendeine Weise dargestellt werden, z.B. durch alphabetische Zeichen, außerdem muss es hierfür einen geeigneten Träger geben. Dies kann beispielsweise ein Textdokument sein.
Information lassen sich in die folgenden drei Bestandteile zerlegen:

- Syntaktischer Teil: Ist die Struktur der Information syntaktisch zulässig? Beispiel hierfür ist die Einhaltung von Rechtschreibung und Grammatik bei Texten.
- Semantischer Teil: Welchen inhaltliche Bedeutung besitzt die Information?
- Pragmatischer Teil: Welchem Zweck dient sie?

([PDVC06], S.314-315)

2.1.2 Information Retrieval

Nachdem bekannt ist, worum es sich bei Informationen handelt, wird im Folgenden beschrieben, worauf sich Information Retrieval bezieht.
Auch hier ist es problematisch, eine einheitliche Definition zu finden. Eine mögliche Erklärung lautet so:

Definition 2.1. *(Information Retrieval)*

Mit Information Retrieval, kurz IR, wird das Auffinden von in unstrukturierter Form vorliegender und ein Informationsbedürfnis befriedigender Materialien innerhalb großer Sammlungen bezeichnet.

Mit unstrukturierten Materialien sind hierbei meist Dokumente in Textform gemeint. Üblicherweise liegen die Sammlungen auf dem Computer gespeichert vor ([CDM08], S.1).

2.1.3 Unterschied zur Datenbankensuche

Information Retrieval unterscheidet sich stark von der Suche in Datenbanken.
In einer Datenbank liegen die Daten strukturiert in Form von Werttupeln bekannten Datentyps vor, was Definition 2.1 widerspricht.
Im Gegensatz zum Information Retrieval kann bei der Datenbankensuche nicht mit vagen Anfragen umgegangen werden. In der Datenbank kann zwar nach $(Miete < 300)$ gesucht werden, aber mit *„günstige Miete“* wäre das System überfordert: Wie ist günstig zu interpretieren? ([Fer03], S.10)
Ein Information Retrieval System muss solche Anfragen mit unklarer Bedeutung verarbeiten können.

2.2 Beispiel Websuche

Zum besseren Verständnis soll an dieser Stelle ein Beispiel zur Veranschaulichung gegeben werden.
Nahezu jeder benutzt im Alltag Web-Suchmaschinen. Websuche stellt eine Form Information Retrieval dar, da hier Freitext beinhaltende Dokumente (z.B. im HTML- oder pdf-Format) aufgefunden werden sollen, um das Informationsbedürfnis des Internetnutzers zu befriedigen. ([Fer03], S.6)
Möchte dieser zum Beispiel seinen nächsten Urlaub planen, könnte seine Suchanfrage *„Hotel günstig Kreta“* lauten.
Die gesuchten Informationen dienen also dem Zweck, den Urlaub zu planen.
Problematisch ist hierbei der semantische Teil der Informationen: Die inhaltliche Bedeutung der Resultate muss mit der ursprünglichen Intention des Nutzers

übereinstimmen.
Ein von der Suchmaschine geliefertes Resultat kann zwar zum syntaktischen Teil passen, da die korrekten Wörter darin auftauchen, allerdings in einem ganz anderen Kontext, sodass der Nutzer mit dem Dokument nichts anfangen kann.

2.3 Bezug zur Problemstellung

Die Aufgabe besteht darin, nach vom Nutzer auswählbaren, logisch verknüpften Kriterien innerhalb eines Dokumentenarchivs zu suchen (1.2).
Demnach ist die Definition 2.1 erfüllt, da hier Materialien innerhalb einer Sammlung, dem Dokumentenarchiv, aufgefunden werden sollen, um ein Informationsbedürfnis zu befriedigen.
Dieses Bedürfnis unterscheidet sich natürlich von Anfrage zu Anfrage, besteht aber allgemein gefasst darin, Dokumente wiederzufinden, z.B. eine bestimmte E-Mail.

2.3.1 Teilweise strukturierte Daten

Besonderheit der Problemstellung ist hierbei, dass die Dokumente teilweise strukturiert sind, d.h. es liegt zwar Freitext vor, aber zusätzlich sind Metadaten vorhanden.
Im Falle der Freitextsuche lässt sich aufgrund der unstrukturierten Textform eindeutig von Information Retrieval sprechen.
Anders sieht es bei den Metadaten aus, welche alle die folgende Syntax und damit Struktur besitzen:

(Name Inhalt)

Es liegt jedoch immer noch ein Information Retrieval Problem vor, da der Begriff auch die Suche in teilweise strukturierten (engl. semistructured) Dokumenten einschließt ([CDM08], S.1-2).
Wobei hierbei angemerkt sei, dass selbst die Metadaten nicht vollkommen strukturiert sind: Der Datentyp des Inhalts ist offen gelassen und es gibt keinerlei Vorgaben, welche Keywords in den Dokumenten auftreten müssen.
In den folgenden Kapiteln wird beschrieben, auf welche Weise ein Information Retrieval System konzipiert und realisiert werden kann, welches in der Lage ist, die Problemstellung 1.2 zu lösen.
Hierzu werden zunächst die hierfür benötigten Modelle boolesches Retrieval (Kapitel 3) sowie das Vector Space Model (Kapitel 4) erläutert.

3

Boolesches Retrieval

Dieses Kapitel stellt das klassische Information-Retrieval-Verfahren boolesches Retrieval (engl.Boolean Retrieval) vor.

3.1 Eigenschaften des Verfahrens

Boolesches Retrieval überprüft Dokumente darauf, ob eine bestimmte Bedingung zutrifft.
Somit gibt es nur die Unterteilung in passende Dokumente und solche, welche die Bedingung nicht erfüllen. Eine weitere Bewertung der Ergebnisse findet nicht statt ([Fer03], S.33). Das fehlende Ranking der Ergebnisse ist ein häufiger Kritikpunkt des Verfahrens.

3.2 Funktionsprinzip

Boolesches Retrieval basiert auf Mengenoperationen. Deshalb werden Dokumente Mengen zugeordnet, die jeweils durch bestimmte Attribute charakterisiert sind.
Dokument bezeichnet die Einheit, auf welcher das Retrieval stattfindet. Ein Dokument kann deshalb eine kleine Textmemo, aber auch ein ganzes Buchkapitel sein ([CDM08], S.4).

3.2.1 Attribut

Ein solches Attribut ist eine Abbildung, welche jedem Dokument einen Wert für dieses Attribut zuordnet. Die Abbildung erzeugt somit Attribut-Wert-Paare, was in Formel 3.1 gezeigt wird.

$$t : D \to T, t(d) = t_i \tag{3.1}$$

Hierbei bezeichnet t die Abbildung (d.h. das Attribut), D die Menge aller Dokumente und T den Wertebereich des Attributs t.

Der Attributwert t_i mit $t_i \in T$ und $i \in N$ wird durch die Abbildung t dem Dokument $d \in D$ zugeordnet.

3.2.2 Anfragen

Elementare boolesche Anfrage

Ein Attribut-Wert-Paar wird auch als elementare boolesche Anfrage bezeichnet. Bei der elementaren booleschen Anfrage (t, t_1) werden zum Beispiel alle Dokumente gesucht, deren Attribut t den Wert t_1 annimmt.
Mathematisch kann die Ergebnismenge D_{t,t_i} für eine Anfrage (t, t_i) wie in Formel 3.2 beschrieben werden.

$$D_{t,t_i} = \{d \in D | t(d) = t_i\} \tag{3.2}$$

Verknüpfung

Mehrere Attribut-Wert-Paare lassen sich mit booleschen Operatoren AND, OR und NOT verknüpfen.
(t, t_1) AND (s, s_1) bedeutet, dass alle Dokumente gesucht sind, bei denen sowohl $t(d) = t_1$ als auch $s(d) = s_1$ gilt, d.h. hier muss der Durchschnitt dieser beiden Ergebnismengen gebildet werden, wie in Formel 3.3 gezeigt.

$$D_{t,t_1} \cap D_{s,s_1} \tag{3.3}$$

Wird hingegen der Operator OR verwendet, werden die Ergebnismengen vereinigt, wie in Formel 3.4 gezeigt.

$$D_{t,t_1} \cup D_{s,s_1} \tag{3.4}$$

Außerdem kann der unäre Operator NOT verwendet werden, der das Komplement der Ergebnismenge erzeugt. Demnach wird für die Anfrage NOT (t, t_1) erst die Menge aller Dokumente bestimmt, bei denen $t(d) = t_1$ zutrifft, und diese anschließend von der Gesamtmenge aller Dokumente abgezogen. Dies wird in Formel 3.5 dargestellt.

$$D \setminus D_{t,t_1} \tag{3.5}$$

Da bei jeder Operation neue Ergebnismengen entstehen, lassen sich hierauf erneut die oben beschriebenen Operatoren anwenden. Auf diese Weise können Anfragen beliebig tief geschachtelt werden ([Fer03], S.34).

3.3 Notwendige Begriffe

3.3.1 Term

Im Laufe der Arbeit wird der Begriff Term häufig auftauchen, darum sei er an dieser Stelle erklärt.
Wenn im Folgenden von Termen gesprochen wird, ist dies nicht äquivalent zu Wörtern: Bei einem Term kann es sich zwar um ein Wort handeln, dies muss jedoch nicht zwangsläufig der Fall sein.
Manche Implementierungen verwenden beispielsweise Stammformen, um ähnliche Wörter zu einem Term zusammenzufassen, was den Speicherbedarf reduziert.
Außerdem wird nicht unbedingt jedes Wort zu einem Term: Handelt es sich beispielsweise um sehr häufig auftretende und zum Sinn des Textes wenig beitragende Wörter, wie z.B. „und“ oder „dann“, können diese wegfallen, um Speicherplatz zu sparen ([Fer03], S.37).
Dokumente werden zu Beginn in solche Terme zerlegt, um später verarbeitet werden zu können (siehe Abschnitt 3.4.1 und 3.4.2). Die Terme bilden somit die indexierten (oder indizierten) Einheiten der Dokumente ([CDM08], S.3).

3.3.2 Index

Terme als indexierte Einheiten zu bezeichnen bedeutet, dass jedem unterschiedlichen Term in der Dokumentensammlung ein in der Sammlung einmaliger Index zugeordnet wird. Die Menge unterschiedlicher Terme wird als „Vokabular“ bezeichnet.
Auch die Dokumente selbst werden mit einem Index, der *docID* (kurz für *document identification*), versehen. Dabei handelt es sich meist um einen ganzzahligen Wert ([CDM08], S.7).

3.4 Implementierungsansätze

In diesem Abschnitt werden typische Implementierungen für das boolesche Retrieval vorgestellt. Diese dienen als Grundlage für die tatsächliche Realisierung des Projektes, welche jedoch erst im zweiten Teil der Arbeit vorgestellt wird.

3.4.1 Inzidenz-Matrix

Eine mögliche Implementierung des booleschen Retrieval stellt die Umsetzung mittels einer Term-Dokument-Inzidenz-Matrix dar.
Dies bedeutet, dass die Zeilen der Matrix die Terme enthalten und die Spalten die Dokumente, was auch umgekehrt realisierbar ist. Genau betrachtet handelt es sich hier nicht um die Terme und Dokumente selbst, sondern deren Indizes.

Tritt Term t in Dokument d auf, so lautet der Eintrag für (t, d) der Matrix 1. Alle Einträge für nicht vorkommende Terme sind hingegen mit einer 0 versehen.
Abbildung 3.1 zeigt eine Beispielmatrix, wobei die tatsächliche Anzahl an Termen und Dokumenten in einer Sammlung weitaus größer ausfällt.
Eine Term-Inzidenz-Matrix verbraucht unnötig Speicherplatz, da sehr viele Einträge der Matrix eine 0 enthalten. Gerade bei sehr großen Sammlungen bzw. Dokumenten ist dies praktisch nicht realisierbar.

	1	2	3	4	5	6	7
Kontaktadresse	0	1	1	0	0	0	1
Seminar	1	0	1	0	1	0	0
Termin	1	1	1	0	0	0	0

Abb. 3.1. Term-Dokument-Inzidenz-Matrix. Die Zeile enthalten die Terme, die Spalten die Dokumente (hier durch deren *docID* repräsentiert). Alle Einträge mit einer 0 sind leere Einträge (Eigene Abbildung).

Verarbeitung einer Anfrage

Um eine Anfrage wie $Kontaktadresse\ AND\ Seminar\ AND\ Termin$ mithilfe einer Matrix zu verarbeiten, werden einfach die entsprechenden Zeilen der Matrix genommen und bitweise logisch verknüpft, was für die obige Anfrage wie folgt aussieht:

0110001 AND
1010100 AND
1110000
———
0010000

Demnach wird das Dokument mit der *docID* 3 zurückgegeben. Analog funktioniert die *OR*-Verknüpfung:

0110001 OR
1010100 OR
1110000
———
1110101

Dieses Beispiel führt demnach zur Ergebnismenge $\{1, 2, 3, 5, 7\}$ ([CDM08], S.4).

3.4.2 Invertierte Liste

In der Regel werden zur Implementierung des booleschen Retrieval invertierte Listen verwendet ([Fer03], S.36).
Der Name basiert auf den darin gespeicherten invertierten Indizes. Diese werden deshalb als invertiert bezeichnet, weil sie vom Term zurück auf die Postion, in welcher der Term aufgetreten ist, schließen lassen.
In einer geeigneten Speicherstruktur, zum Beispiel einem Dictionary, werden zu jedem Term alle Dokumente gespeichert, in denen der Term auftritt.
Diese Dokumentlisten werden als invertierte Listen (engl. inverted lists) bezeichnet (siehe Abbildung Abbildung 3.2). Manche Implementierungen beinhalten neben dem Dokumentindex zusätzliche Informationen wie die genaue Wortpostion im Dokument.
Dieses Vorgehen setzt voraus, dass zuvor eine Indizierung (siehe **??**) stattgefunden hat ([CDM08], S.5-6).
Das Verfahren ermöglicht sehr schnelle Zugriffe, ist allerdings speicherintensiv ([Fer03], S.36). Im Vergleich zur Inzidenz-Matrix wird jedoch deutlich weniger Speicher benötigt, da die vielen leeren Einträge entfallen.

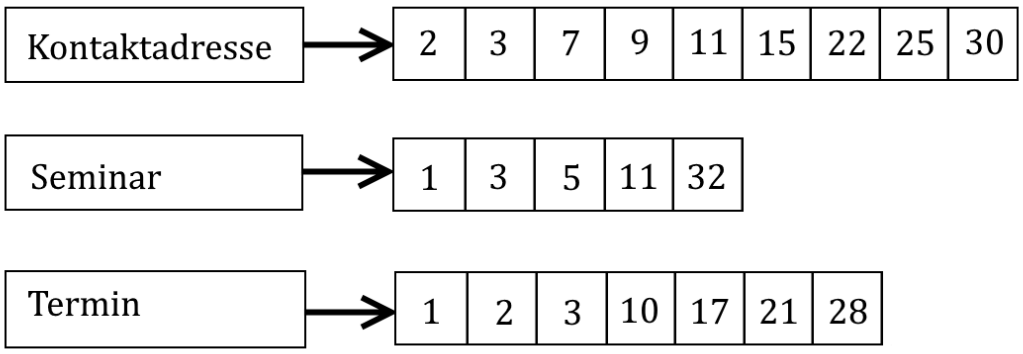

Abb. 3.2. Invertierte Listen zu Beispieltermen. Die Zahlen sind die Dokumentindizes, in denen der jeweilige Term vorkommt (Eigene Abbildung).

Verarbeitung einer Anfrage

Hier stellt sich die Frage, wie denn nun eine boolesche Anfrage, wie in Abschnitt 3.2.2 beschrieben, mithilfe invertierter Listen umgesetzt werden kann.

Elementare Anfrage

Angenommen, es liegt eine elementare boolesche Anfrage in der Form (t, t_1) vor. In der Praxis ist meist nach dem Vorkommen eines bestimmtes Wortes gefragt. Demnach entspricht das Attribut t dem Term des gesuchten Wortes.

Da man auf dessen Vorkommen prüft, gelten für den Wertebereich $T = \{true, false\}$ und für den Attributwert $t_1 = true$.
Die Verarbeitung einer solchen elementaren Anfrage geht relativ einfach: Über den Index kann auf den Term schnell zugegriffen werden, vorausgesetzt dieser ist in der Sammlung enthalten. Trifft dies zu, kann einfach die gesamte zugehörige invertierte Liste als Resultat ausgegeben werden, da für alle enthaltenen Dokumente $t_1 = true$ gilt.

AND-Verknüpfungen

Wie sieht nun die Verarbeitung aus, wenn mehrere Anfragen miteinander verknüpft werden? Hierzu wird zunächst der AND - Operator betrachtet. Eine Anfrage liegt dann in der Form $(t, t_1)\ AND\ (s, s_1)$ vor, wie etwa bei dem Beispiel *Kontaktadresse AND Seminar*, wobei gilt $t = Kontaktadress, t_1 = true$ sowie $s = Seminar, s_1 = true$.
Demnach werden alle Dokumente gesucht, in denen beide Terme auftauchen. Dazu wird der Durchschnitt aus den zugehörigen invertierten Listen gebildet. Betrachtet man die Abbildung 3.2, so ist der Durchschnitt für $D_{t,t_1} \cap D_{s,s_1}$ bzw. für $Kontaktadress \cap Seminar$ gleich der Ergebnisliste $\{3, 11\}$.

OR-Verknüpfungen

Lautet die Anfrage hingegen $(t, t_1)\ OR\ (s, s_1)$ bzw. *Kontaktadresse OR Seminar*, so sind alle Dokumente gesucht, in denen entweder t_1 oder s_1 oder auch beide Terme vorkommen.
Gesucht ist also die Vereinigung $D_{t,t_1} \cup D_{s,s_1}$ bzw. $Kontaktadress \cup Seminar$. Dies ist die Vereinigung der invertierten Listen beider Terme. Im Falle des Beispiels 3.2 lautet die Ergebnisliste für die Anfrage *Kontaktadresse OR Seminar* $\{1, 2, 3, 5, 7, 9, 11, 15, 22, 25, 30, 32\}$.
Für beide Listenoperationen gilt, das die Indizes in den Ergebnislisten sortiert und Duplikate entfernt werden ([CDM08], S.11).

AND NOT-Verknüpfung

Da es sich bei NOT um einen unären Operator handelt, könnte dieser theoretisch alleine auftreten.
Eine Anfrage der Form *NOT Seminar* kann sehr viel Laufzeit kosten, wenn die Sammlung aus vielen Dokumenten besteht: Es muss über die gesamte Sammlung iteriert werden und für jedes Dokument geprüft werden, ob es in der invertierten Liste für den Term Seminar auftaucht. Der alleinstehende NOT-Operator ist deshalb so ineffizient, dass er bei den meisten booleschen Retrieval Systemen nur im Zusammenhang mit einem binären Operator zugelassen ist.
Da die Kombination $OR\ NOT$ keinen Sinn macht, wenn man einen Term ausschließen möchte, ist dies in der Regel AND.

Hierbei wird aus den beiden Listen die Differenz gebildet. Lautet die Anfrage beispielsweise $Kontaktadresse\ AND\ NOT\ Seminar$, so werden aus der Ergebnisliste für Kontaktadresse alle Elemente entfernt, die in der Liste für Seminar enthalten sind ([Hen08], S.174).

Komplexe Ausdrücke

Da sowohl Vereinigung als auch Durchschnitt eine neue Ergebnisliste liefern, kann auf dieser wiederum jeder Operator angewandt werden, was eine beliebig tiefe Schachtelung erlaubt. Dieser Abschnitt erklärt, wie komplex geschachtelte Ausdrücke verarbeitet werden.

Im Falle von mehreren AND-Operatoren, wie etwa in der Suchanfrage $Kontaktadresse\ AND\ Seminar\ AND\ Termin$, ist es effizient, zunächst die einzelnen invertierten Listen aufsteigend nach deren Länge zu sortieren und dann von links nach rechts zu verarbeiten, indem das nachfolgende AND auf die Ergebnisliste des vorherigen Durchschnitts angewandt wird:
$(Seminar\ AND\ Termin)\ AND\ Kontaktadresse$
Auf diese Weise werden die Listen, über die iteriert werden muss, möglichst klein gehalten. Besitzt die kleinste Liste beispielsweise die Länge eins, dann kann nach der ersten Iteration bereits abgebrochen werden, da zulässige Lösungen in allen drei Listen vorkommen müssen.

Bei mehreren OR-Operatoren werden die Ausdrücke analog von links nach rechts verarbeitet, wobei die Sortierung nach Länge hierbei keinen Vorteil bringt, da bei der Vereinigung zweier Listen ohnehin über alle Elemente iteriert werden muss. Die Verarbeitung würde demnach in der folgenden Reihenfolge erfolgen:
$(Kontaktadresse\ OR\ Seminar)\ OR\ Termin$

Ist die Anfrage hingegen gemischt, wie etwa in $(Kontaktadresse\ OR\ Seminar)\ AND\ (Termin\ OR\ Seminar)$, werden erst die inneren Ausdrücke ausgewertet und dann aus deren Ergebnislisten der Durchschnitt gebildet ([CDM08], S.11).

Die Verarbeitung mehrerer Wörter

Boolesches Retrieval kann auch mehrere zusammengehörende Wörter verarbeiten. Über die interne Verarbeitung besteht hierbei jedoch für den Nutzer kein Einblick: Das Information Retrieval System kann so realisiert sein, dass es die aus den Wörtern der Anfrage isolierten Terme mit OR veknüpft, es kann diese jedoch genauso gut mitAND verbinden ([Hen08], S.171).

4

Das Vektorraummodell

Dieses Kapitel stellt ein weiteres klassisches Information Retrieval Verfahren, das Vektorraummodell, vor.

4.1 Funktionsprinzip

Zunächst werden alle Dokumente in Terme zerlegt und indexiert, wie es auch beim booleschen Retrieval der Fall ist. Die Ermittlung des Vokabulars ist Grundvoraussetzung für alle weiteren Schritte.
Wie der Name bereits nahelegt, basiert das Verfahren auf Vektoren. Die Grundidee besteht darin, für jedes Dokument sowie für die Anfrage einen reellen Vektor zu erstellen und anschließend zu ermitteln, welche Dokumentvektoren am ähnlichsten zum Anfragevektor sind.
Jeder Vektor besitzt hierbei die Länge des Vokabulars, da er die Gewichte aller Terme enthält. Die Bedeutung des Gewichts wird in Abschnitt 4.4 erklärt.
Im Gegensatz zum booleschen Retrieval können die Resultate des Vektorraummodells basierend auf dem Grad der Ähnlichkeit in eine Rangfolge gebracht werden ([Fer03], S.62-63).

4.2 Vektor und Vektorraum

Um die Funktionsweise des Vektorraummodells zu verstehen, müssen zunächst die Begriffe Vektor und Vektorraum bekannt sein.

4.2.1 Definition Vektor

4.2.2 Definiton Vektorraum

4.3 Definition Vektorraummodell

Das soeben beschriebene Funktionsprinzip lässt sich mathematisch mithilfe von Attributen beschreiben.

Attribute stellen im Vektorraummodell eine Abbildung der Dokumentenmenge D auf die reellen Zahlen R dar. Damit ist der Wertebereich der Attribute im Gegensatz zum booleschen Retrieval eindeutig festgelegt.
Die Definition lautet somit wie folgt:

Definition 4.1. *(Vektorraummodell mit Attributen)*

Sei $D = d_1, ..., d_n$ eine Menge von Dokumenten oder Objekten und $A = A_1, ..., A_n$ eine Menge von Attributen $A_j : D \to R$ auf diesen Objekten. Die Attributwerte $A_j(d_i) =: w_{i,j}$ des Dokuments d_i lassen sich als Gewichte auffassen und zu einem Vektor $w_i = (w_{i,1}, ..., w_{i,n}) \in R^n$ zusammenfassen. Dieser Vektor beschreibt das Dokument im Vektorraummodell: Er ist seine Repräsentation und wird Dokumentvektor genannt.
Eine Anfrage wird durch einen Vektor $q \in R^n$ mit Attributwerten, den Anfragevektor oder Query-Vektor, dargestellt.
Eine Ähnlichkeitsfunktion $s : R^n \times R^n \to R$ definiere für je zwei Vektoren $x, y \in R^n$ einen reellen Ähnlichkeitswert $s(x, y)$.

Diese Definition ist aufgrund der Beschreibung durch Attribute sehr allgemein gehalten. Es ist deshalb nicht definiert, welche Einheiten des Dokuments gewichtet werden. Demnach sind theoretisch auch andere Dokumentformate wie etwa Bilder mit Pixeln bzw. Pixelgruppen als Attributen möglich ([Fer03], S.63).
Praktisch gesehen machen im Rahmen dieser Arbeit jedoch andere Formate als Texte keinen Sinn. Darum wird im folgenden davon ausgegangen, dass ausschließlich Terme gewichtet werden. Demnach lässt sich die Attributmenge A spezifisch auf die Problemstellung bezogen als Menge der Terme oder Vokabular T auffassen.

4.4 Gewichte

Bei Gewichten handelt es sich, wie bereits beschrieben, um reelle Zahlenwerte.
Ein Gewicht gibt die Wichtigkeit eines Terms basierend auf dessen statistischer Häufigkeit an ([CDM08], S.100).

4.4.1 Termhäufigkeit

Die Häufigkeit, mit der ein Term t in einem Dokument d auftritt, wird als Termhäufigkeit (engl. term frequency) bezeichnet. Demnach wird die Termhäufigkeit pro Vorkommen von t in d um eins erhöht ([CDM08], S.71).
Es erscheint intuitiv logisch, dass ein Text, der das gesuchte Wort mehrmals beinhaltet wichtiger sein muss als ein Dokument, in welchem der Begriff nur ein einziges mal auftaucht.
Dies Gewichtungsschema erlaubt eine viel genauere Differenzierung als eine simple Unterscheidung zwischen true und false, wie es beim booleschen Retrieval der Fall ist.

4.4.2 Dokumenthäufigkeit

Die Termhäufigkeit stellt für sich genommen schon eine mögliche Gewichtungsmethode dar, allerdings keine besonders gute: Die Bewertung alleine aufgrund der Termhäufigkeit lässt außer Acht, dass nicht alle Terme gleich wichtig sind.
Taucht ein Term beispielsweise in jedem Dokument auf, kann es nicht besonders aussagekräftig sein. Demnach ist es sinnvoll, zusätzlich zur Termhäufigkeit $tf_{t,d}$ auch die Dokumenthäufigkeit (engl. document frequency) df_t zu bestimmen.
Diese entspricht der Anzahl Dokumente in D, welche t enthalten. Um den Einfluss nicht aussagekräftiger Terme zu reduzieren, wird das Gewicht umso stärker verringert, je größer die Dokumenthäufigkeit ausfällt ([CDM08], S.108).

4.4.3 Invertierte Dokumenthäufigkeit

Zur Reduktion des Gewichts basierend auf der Dokumenthäufigkeit wird als reduzierender Faktor die Invertierte Dokumenthäufigkeit (engl. inverse document frequency) verwendet. Die invertierte Dokumenthäufigkeit idf_t des Terms t berechnet sich wie in Formel 4.1 gezeigt.

$$idf_t = \frac{1}{df_t} \tag{4.1}$$

Oftmals werden modifizierte Formen verwendet, um die großen Werte seltener Terme durch den Logarithmus wieder zu dämpfen ([Fer03], S.68-69).
Formel 4.2 zeigt ein Beispiel für eine solche modifizierte invertierte Dokumenthäufigkeit. In der Regel beträgt die Basis des Logarithmus 10, dies spielt aber letztendlich für das korrekte Ranking der Resultate keine Rolle ([CDM08], S.108-109).

$$idf_t = log\frac{N}{df_t} \tag{4.2}$$

4.4.4 TF-IDF-Gewichtung

Die vollständige Methode zur Gewichtung einzelner Terme kombiniert Termhäufigkeit und invertierte Dokumenthäufigkeit, indem erstere mit letzterer multipliziert wird. Alle Formeln dieses Typs werden als TF-IDF Gewichtung (engl. tf-idf weighting) bezeichnet ([Fer03], S.71).
Das Gewicht für Term t in Dokument d berechnet sich somit wie in Formel 4.3 gezeigt ([CDM08], S.109).

$$tf - idf_{t,d} = tf_{t,d} \times idf_t, \tag{4.3}$$

Verwendet man für die invertierte Dokumenthäufigkeit den unmodifizierten Wert 4.1, so ergibt sich hieraus die Berechnung:

$$tf - idf_{t,d} = \frac{tf_{t,d}}{df_t} \tag{4.4}$$

Für die modifizierte Formel 4.2 lautet die TF-IDF-Gewichtung wie folgt:

$$tf - idf_{t,d} = tf_{t,d} \times (log \frac{N}{df_t}) \tag{4.5}$$

Sei T die Menge aller Terme der Sammlung bzw. das Vokabular, dann enthält der Gewichtsvektor w_i zu einem Dokument $d_i \in D$ für jeden Term $t_j \in T$ dessen Gewicht $w_{i,j} = tf - idf_{j,i}$, sodass $w_i = (tf - idf_{1,i}, ..., tf - idf_{n,i})$ gilt.

4.5 Anfragen

Beim Vektorraummodell gibt es keine booleschen Operatoren zur Verknüpfung von Termen, weshalb Anfragen in Freitextform gestellt werden. Diese Form wird auch in der Websuche verwendet und ist darum sehr bekannt.
Da die Reihenfolge von Wörtern weder bei Anfragen noch in den Dokumenten eine Rolle spielt, lassen sich Anfragen einfach als eine Menge von Wörtern bzw. als die daraus resultierende Menge von Termen betrachten.
Ein solches Modell, das lediglich die Anzahl, nicht aber die Reihenfolge von Wörtern berücksichtigt, wird auch als *bag of words model* bezeichnet.
Da für jeden Term ein anderer Ähnlichkeitswert erzielt wird, werden die Ähnlichkeitswerte aller in der Menge enthaltenen Terme addiert, sodass pro Dokument ein Gesamtwert berechnet wird ([CDM08], S.107).
Anfragetexte werden genau wie Dokumente behandelt und die Vektoren wie in Abschnitt 4.4.4 beschrieben bestimmt ([Fer03], S.82).

4.6 Ähnlichkeitsfunktion

Grundidee des Vektorraummodells ist das Ermitteln der Ähnlichkeit zwischen Vektoren. Deshalb muss hierfür eine geeignete Ähnlichkeitsfunktion gefunden werden.

4.6.1 Euklidischer Abstand

Eine typische Distanzfunktion für Vektoren ist der euklidische Abstand, bei dem die Differenz wie in Formel 4.6 gezeigt berechnet wird ([KMOB14],S.132).

$$d(\overrightarrow{x}, \overrightarrow{y}) = \sqrt{\sum_{i=1}^{M} (x_i - y_i)^2} \tag{4.6}$$

Allerdings besitzt der euklidische Abstand den gravierenden Nachteil, dass die Länge der Vektoren eine Rolle spielt.
Liegen zwei Dokumente $d1$ und $d2$ vor und $d2$ besitzt den Inhalt von $d1$ zweimal aneinandergereiht, so wird $d2$ als ähnlicher eingestuft, aus dem einzigen Grund weil die Termhäufigkeit doppelt so groß ist und der Vektor damit die doppelte Länge hat.
Das Problem, dass zwei unterschiedlich lange Dokumente, in denen die gesuchten Terme etwa gleich verteilt sind, dennoch vollkommen verschiedene Ähnlichkeitswerte erzielen macht den euklidischen Abstand zu einer ungeeigneten Ähnlichkeitsfunktion.

4.6.2 Cosinus-Maß

Um den Einfluss der Vektorlänge zu eliminieren, wird in der Regel das Cosinus-Maß verwendet. Das Cosinus-Maß ist das Skalarprodukt der normalisierten Vektoren ([CDM08], S.112), d.h. die Vektoren werden durch ihre Länge dividiert.
Da zur Berechnung des Cosinus-Maßes somit sowohl das Skalarprodukt als auch die Längenberechnung eines Vektors bekannt sein müssen, werden beide an diese Stelle vorgestellt.

Euklidische Länge

Sei $\overrightarrow{x}$ ein Vektor, dann wird seine euklidische Länge wie in Formel 4.7 angegeben berechnet.

$$|\overrightarrow{x}| = \sqrt{\sum_{i=1}^{M} x_i^2} \tag{4.7}$$

Skalarprodukt

Das Skalarprodukt zweier Vektoren $\overrightarrow{x}$ und $\overrightarrow{y}$ wird wie in Formel 4.8 gezeigt berechnet.

$$\overrightarrow{x} \cdot \overrightarrow{y} = \sum_{i=1}^{M} x_i y_i \tag{4.8}$$

Funktion

Das Cosinus-Maß multipliziert die Vektoren und dividiert sie durch deren Länge, sodass die Ähnlichkeitsfunktion von der Länge unbeeinflusst ist, was in Formel 4.9 gezeigt wird ([CDM08], S.111).

$$sim(\overrightarrow{x}, \overrightarrow{y}) = \frac{\overrightarrow{x} \cdot \overrightarrow{y}}{|\overrightarrow{x}| \cdot |\overrightarrow{y}|} \tag{4.9}$$

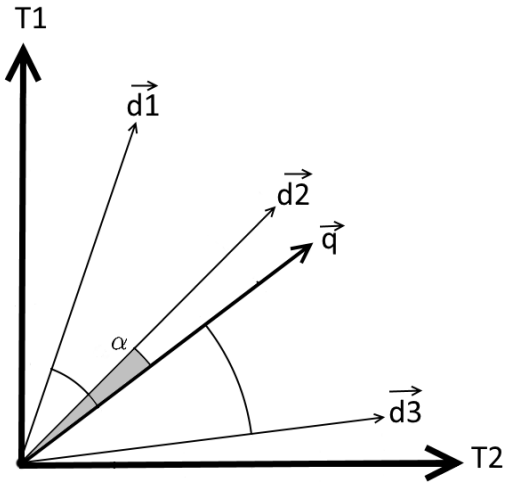

Abb. 4.1. Vektorraum mit den Termen $T1$ und $T2$ als Achsen, drei Dokumentvektoren und dem Anfragevektor. Das Cosinus-Maß liefert als ähnlichsten Dokumentvektor d_2 (Eigene Abbildung).

Beispiel

Abbildung 4.1 zeigt einen Vektorraum mit zwei Termen als Achsen, sodass er sich zweidimensional darstellen lässt. In der Realität gibt es meist weitaus mehr Achsen, da das Vokabular tausende Terme beinhalten kann. Die hier dargestellten Vektoren wurden bereits normalisiert.
Dieser beispielhafte Vektorraum beinhaltet insgesamt drei Dokumentvektoren $\overrightarrow{d_i}$ sowie den Anfragevektor $\overrightarrow{q}$.
Unter Verwendung des Cosinus-Maßes ergibt sich für $sim(\overrightarrow{d_2}, \overrightarrow{q}) = cos(\alpha)$ der höchsten Wert, da α der kleinste Winkel ist. Damit wird $d2$ als erstes Ergebnisdokument ausgegeben ([CDM08], S.112).

4.6.3 Alternativen

5

Bewertung eines Information Retrieval Systems

5.1 Precisison

5.2 Recall

6

Implementierung

7

Die Benutzeroberfläche

8

Zusammenfassung und Ausblick

In diesem Kapitel soll die Arbeit noch einmal kurz zusammengefasst werden. Insbesondere sollen die wesentlichen Ergebnisse Ihrer Arbeit herausgehoben werden. Erfahrungen, die z.B. Benutzer mit der Mensch-Maschine-Schnittstelle gemacht haben oder Ergebnisse von Leistungsmessungen sollen an dieser Stelle präsentiert werden. Sie können in diesem Kapitel auch die Ergebnisse oder das Arbeitsumfeld Ihrer Arbeit kritisch bewerten. Wünschenswerte Erweiterungen sollen als Hinweise auf weiterführende Arbeiten erwähnt werden.

Literaturverzeichnis

Aca. *Academic - Academic dictionaries and encyclopedias, Universal Lexikon.* http://universal_lexikon.deacademic.com/253489/Informationsr%C3%BCckgewinnung, letzter Zugriff am 05.06.2017.

CDM08. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze: *Introduction to Information Retrieval.* Cambridge University Press, 2008.

Fer03. Ferber, Reginald: *Information Retrieval, Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web.* dpunkt.verlag, 2003.

Hen08. Henrich, Andreas: *Information Retrieval 1 -Grundlagen, Modelle und Anwendungen*, 2008. https://www.uni-bamberg.de/fileadmin/uni/fakultaeten/wiai_lehrstuehle/medieninformatik/Dateien/Publikationen/2008/henrich-ir1-1.2.pdf, letzter Zugriff am 10.06.2017 .

KMOB14. Kweku-Muata Osei-Bryson, Ojelanki Ngwenyama: *Advances in Research Methods for Information Systems Research - Data Mining, Data Envelopment Analysis, Value Focused Thinking.* Nummer 34 in *Integrated Series in Information Systems.* Springer Verlag, 2014.

PDVC06. Prof. Dr. Volker Claus, Prof. Dr. Andreas Schwill: *Duden - Informatik A-Z, Fachlexikon für Studium, Ausbildung und Beruf.* Dudenverlag, 2006.

A

Glossar

DisASTer	DisASTer (Distributed Algorithms Simulation Terrain), A platform for the Implementation of Distributed Algorithms
DSM	Distributed Shared Memory
AC	Linearisierbarkeit (atomic consistency)
SC	Sequentielle Konsistenz (sequential consistency)
WC	Schwache Konsistenz (weak consistency)
RC	Freigabekonsistenz (release consistency)

B

Erklärung der Kandidatin / des Kandidaten

☐ Die Arbeit habe ich selbstständig verfasst und keine anderen als die angegebenen Quellen- und Hilfsmittel verwendet.

☐ Die Arbeit wurde als Gruppenarbeit angefertigt. Meine eigene Leistung ist ...

Diesen Teil habe ich selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet.

Namen der Mitverfasser: ...

Datum	Unterschrift der Kandidatin / des Kandidaten