

# VORSCHAU

Dieses Dokument enthält lediglich einen Auszug aus der gesamten Studie.  
Besuchen Sie bitte [www.W3L.de](http://www.W3L.de), um eine vollständige Fassung dieser Studie zu erwerben.



## Studie

### Online-Suchfunktionen deutscher Hochschulen

#### Analyse und Optimierungsmöglichkeiten

Lennart Lorenz

Dr. Kai Schmitz-Hofbauer

Jakob Pyttlik

# Online-Suchfunktionen deutscher Hochschulen Analyse und Optimierungsmöglichkeiten

Veröffentlicht im Februar 2010

Herausgeber:  
W3L GmbH Herdecke Witten

W3L GmbH  
Wittener Industrie und Technologie Park  
Stockumer Str. 28

58453 Witten

Geschäftsführung:  
Dr. Olaf Zwintscher  
Prof. Dr. Helmut Balzert

Tel.: +49 (0)2302 – 988 65 0  
Fax.: +49 (0)2302 – 988 65 10

E-Mail: [suche@W3L.de](mailto:suche@W3L.de)  
<http://www.W3L.de/suche>

## Inhaltsverzeichnis

1	Einleitung.....	5
1.1	Motivation.....	5
1.2	Einsatz von Suchtechniken.....	5
1.3	Zielsetzung und Aufbau der Studie.....	6
2	Suchmaschinen aus theoretischer Sicht.....	7
2.1	Funktionsweise der klassischen Volltextsuche.....	7
2.1.1	Aufbau eines Suchsystems.....	7
2.1.2	Zerlegung eines Textes in Wörter.....	9
2.1.3	Entfernung von Stoppwörtern.....	10
2.1.4	Erstellung eines Indexes.....	10
2.1.5	Suche unter Berücksichtigung des Indexes.....	13
2.1.6	Ergebnispräsentation.....	13
2.2	Schwächen der Volltextsuche.....	14
2.2.1	Aufbereitung des Suchraums.....	14
2.2.2	Suchanfrage und Ergebnispräsentation.....	15
2.3	Semantische Suchmaschinen.....	16
2.3.1	Definition.....	16
2.3.2	Abgrenzung gegenüber der klassischen Volltextsuche.....	16
2.4	Basistechniken der semantischen Suche.....	17
2.4.1	Kompositazerlegung.....	17
2.4.2	Grundformbildung.....	18
2.4.3	Zuordnung von Wortarten.....	20
2.4.4	Ermittlung von verwandten Begriffen.....	20
2.4.5	Automatische Verschlagwortung.....	23
2.4.6	Dokumentenähnlichkeit.....	24
2.4.7	Clusterbildung.....	27
2.4.8	Zusammenfassung.....	28
3	Schema zur Bewertung von Suchmaschinen.....	29
3.1	Aufbau des Schemas.....	29
3.2	Kriterien und Faktoren des Bewertungsschemas.....	31
3.2.1	Manuelle Suchmöglichkeiten und Operatoren.....	31
3.2.2	Suchraum- und Anfragenaufbereitung.....	33
3.2.3	Ergebnispräsentation und Benutzerführung.....	36
3.2.4	Sonstiges.....	38

3.3	Spezifizierung der Suchbegriffe.....	38
4	Ergebnisse der Bewertung.....	40
4.1	Ablauf des Bewertungsvorgangs.....	40
4.2	Tabellarischer Überblick über die Ergebnisse.....	40
4.3	Bewertungen der einzelnen Suchmaschinen im Detail.....	49
4.4	Diskussion der Ergebnisse.....	81
5	Fazit.....	85
	Literaturverzeichnis.....	86

# 1 Einleitung

## 1.1 Motivation

»*Wer sucht, der findet*«. Dieses häufig gebrauchte Sprichwort ist leider vielfach nicht zutreffend. Während die Verfügbarkeit von digitalen Inhalten in der heutigen Zeit kein Problem mehr darstellt, ist das Auffinden von Informationen ein langwieriges, mühsames und zeitintensives Unterfangen. Die größte Schwierigkeit besteht darin, unter den zahlreichen angebotenen Informationen die relevanten Informationen herauszufiltern. Lange und übersichtliche Trefferlisten von Suchmaschinen mit teilweise unpassenden Ergebnissen erschweren den effizienten Zugang.

Dieser Sachverhalt ist gerade für Hochschulen von entscheidender Bedeutung. Eine schlechte Auffindbarkeit von Inhalten ist hierbei mit folgenden Konsequenzen verbunden:

- Studieninteressenten finden passende Angebote nicht und entscheiden sich ggf. für eine andere Hochschule.
- Die Qualität von Lehre und Forschung wird in Mitleidenschaft gezogen, da vorhandene Inhalte nicht genutzt werden.
- Studierende nutzen die vielfältigen Angebote und Einrichtungen der Hochschule nicht, da sie diese nicht finden.

Das Hauptproblem bildet somit nicht die Verfügbarkeit, sondern die Informationsüberflutung. Deren Überwindung und somit die Verbesserung der Auffindbarkeit von Inhalten stellt eine große Herausforderung dar.

## 1.2 Einsatz von Suchtechniken

Ein sehr großer Teil der Informationen liegt in Form von semi-strukturierten oder unstrukturierten Daten, d. h. Text, vor. Dabei ist es entscheidend, in den Texten implizit enthaltenes Wissen auffindbar zu machen [HQW06, S. 1 ff.]. Zum Auffinden von Informationen, und somit zur Überwindung der Informationsflut, kommen Such-Techniken zum Einsatz. Ein sehr häufig genutztes Verfahren im Bereich des Inter- und Intranets ist die Volltextsuche. Es wird versucht, die Informationsgewinnung zu ermöglichen, indem Dokumente beinahe komplett auf einen durchsuchbaren Index abgebildet werden. Ein Problem der Volltextsuche ist jedoch, dass die Suche auf eine rein syntaktische Ebene reduziert wird, die Semantik der Wörter und Texte hingegen vernachlässigt. Die Suche beschränkt sich auf den exakten Abgleich von Schlüsselwörtern. Da Sprache jedoch selten komplett ohne Mehrdeutigkeiten auskommt und der Kontext mitunter von entscheidender Bedeutung für die Interpretation eines Wortes ist, führt dies zu einer Reihe von Schwierigkeiten. So liefern einfache Suchanfragen oft eine unüberschaubare Anzahl von Dokumenten, deren inhaltliche Relevanz in den Augen des Benutzers oft eher niedrig ist. Es besteht dadurch die Gefahr, dass potenziell wichtige Informationen nicht gefunden werden, oder die Suche unverhältnismäßig aufwendig und zeitraubend ist.

Einen Lösungsansatz versucht das Konzept der semantischen Suche zu bieten. Zielsetzung ist hierbei, von einer starren Suche nach Schlüsselwörtern zu einer Suche nach der Bedeutung von Wörtern zu gelangen. Es wird versucht, den Benutzer zu unterstützen und die Suchfunktion intuitiver zu gestalten, um den Prozess der Wissensauffindung zu erleichtern.

### 1.3 Zielsetzung und Aufbau der Studie

Diese Studie widmet sich der Analyse und Bewertung von Suchmaschinen auf Websites von Hochschulen. Es ist Intention dieser Studie, Optimierungspotenziale und Verbesserungsmöglichkeiten aufzuzeigen. Sie bezieht neben Elementen der Volltextsuche auch die Möglichkeiten einer semantischen Suche – also die Berücksichtigung der Bedeutung – sowie ergonomische Aspekte in Form einer optimierten Benutzerführung mit ein.

Neben Hochschulen ist diese Studie für alle von Interesse, die das vorhandene Wissen in ihrer Organisation – verteilt in heterogenen Datenquellen – besser nutzen wollen und externen Kunden eine bessere Suche auf ihren Webseiten zur Verfügung stellen wollen.

Um den Anforderungen zu genügen, wird in den Kapiteln 2.1 und 2.2 zuerst ein kurzer Einblick in die Funktionsweise von klassischen Volltextsuchmaschinen inklusive der im Kontext ihrer Benutzung auftretenden Probleme gegeben. Danach erfolgt eine Analyse der Grundkonzepte der semantischen Suche. Die Analyse beschränkt sich auf die Basistechnologien, eine tiefere Analyse einzelner Verfahren wird nicht angestrebt. Die Definition des Begriffes der semantischen Suche sowie die Vorstellung der Grundkonzepte ist Inhalt der Kapitel 2.3 und 2.4.

Einen weiteren Hauptaspekt der Studie bildet die Entwicklung eines Klassifikationsschemas zur Beurteilung der Leistungsfähigkeit von Suchen auf Web-Auftritten. Das in Kapitel 3 entwickelte Schema wird anschließend auf ausgewählte Suchen der Web-Seiten von Hochschulen angewendet. Abschließend findet in Kapitel 4 eine Bewertung und Diskussion der Ergebnisse statt. Dies alles geschieht mit der Absicht, einen Überblick über die aktuelle Leistungsfähigkeit von Suchen auf Web-Seiten zu geben und zu zeigen, in welche Richtung eine Entwicklung möglich ist. Abschließend enthält Kapitel 5 ein Fazit bezüglich der Studie und ihrer Ergebnisse.

## 2 Suchmaschinen aus theoretischer Sicht

In den folgenden Unterkapiteln werden Suchmaschinen aus theoretischer Sicht näher betrachtet. Der Fokus liegt dabei auf Volltextsuchmaschinen und semantischen Suchmaschinen. Auf eine exakte Beschreibung möglicher Implementierungsdetails wird verzichtet. Das Kapitel 2.1 beschäftigt sich mit der generellen Funktionsweise der Volltextsuche, indem alle notwendigen Schritte bis zur Suche durch einen Benutzer kurz skizziert werden. Die im Kontext der Benutzung auftretenden Probleme werden im Kapitel 2.2 thematisiert. Auf Basis der Volltextsuche wird im Folgenden das Konzept der semantischen Suche beschrieben. Diese stellt eine Verbesserung der Volltextsuche dar, indem versucht wird, die Bedeutung von Wörtern stärker zu berücksichtigen, anstatt sich streng auf einzelne Schlüsselwörter zu beschränken.

### 2.1 Funktionsweise der klassischen Volltextsuche

Die folgenden Unterkapitel 2.1.1 bis 2.1.6 zeigen die Funktionsweise eines typischen Volltextsuchsystems. Zuerst wird kurz der Aufbau skizziert sowie auf die Heterogenität der Quellen und Inhalte hingewiesen. Die Abschnitte 2.1.2 bis 2.1.4 schildern die Schritte zur Aufnahme eines Dokuments in den Such-Index sowie den prinzipiellen Aufbau eines, für Suchmaschinen typischen, invertierten Index. Die Kapitel 2.1.5 bis 2.1.6 beschreiben, wie eine Suche über einen Index im Allgemeinen funktioniert und wie die Ergebnisse präsentiert und sortiert werden. Die Ausführungen beschränken sich hauptsächlich auf Aspekte der Verarbeitung von Text und der Repräsentation des Inhalts. Die technischen Implementierungsdetails werden zu Gunsten dieser Betrachtungsweise vernachlässigt.

#### 2.1.1 Aufbau eines Suchsystems

Grundsätzlich besteht ein Volltextsuchsystem aus verschiedenen Komponenten oder Modulen, die jeweils eine bestimmte Aufgabe erfüllen. Einen beispielhaften Aufbau zeigt die Abbildung 1. Die Pfeile zwischen den einzelnen Komponenten symbolisieren die jeweiligen Daten- und Kontrollflüsse.

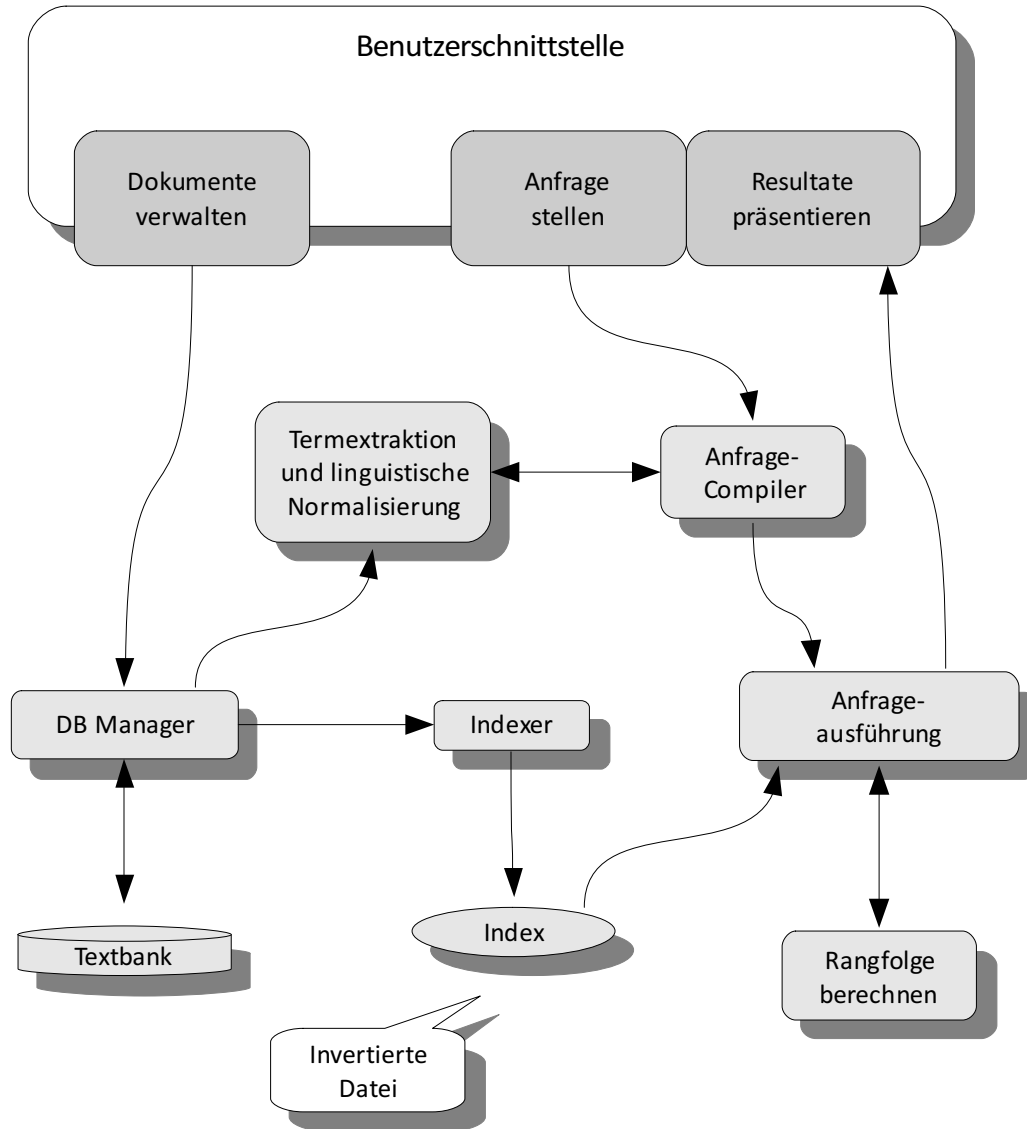


Abbildung 1: Komponenten eines Volltextsuchsystems (in Anlehnung an: [DGS04, S. 483])

Um eine Suchfunktion zu ermöglichen, muss zuerst ein **Index** auf Grundlage der zu durchsuchenden Dokumente aufgebaut werden. Weiterhin wird vorausgesetzt, dass bereits eine Umwandlung der Dokumente in ein von den entsprechenden Modulen des Suchsystems lesbares **Dateiformat** stattgefunden hat. Hier kommen formatspezifische **Konverter** zum Einsatz, die den Text aus den einzelnen Dokumenten extrahieren und für die Indizierung bereitstellen. Dies ist insofern von Bedeutung, da Inhalte oft in verschiedenen Formaten (z. B. PDF, Word oder HTML) vorliegen. Dabei muss es das Ziel sein, sämtliche heterogenen Quellen weitestgehend in das Suchsystem zu integrieren und durchsuchbar zu machen. Dies

ist von enormer Wichtigkeit für die Mächtigkeit des Suchsystems und somit auch für die Auffindung von Informationen. Den weiteren Ablauf zur Indexierung eines Dokuments zeigen die folgenden Abschnitte.

Mit Hilfe der Benutzungsschnittstelle ist es möglich, die Dokumente zu verwalten. Der Zugriff auf die Textbank, die im wesentlichen die zu durchsuchenden Dokumente enthält, erfolgt dabei über den DB Manager. Dieser ist ebenfalls mit dem Modul zur Termextraktion und linguistischen Normalisierung sowie dem Indexer verbunden. Nachdem die einzelnen Terme aus den jeweiligen Dokumenten extrahiert wurden (siehe Kapitel 2.1.2 und Kapitel 2.1.3), werden diese durch das Indexer-Modul dem Index hinzugefügt. Dieser wird im Allgemeinen in Form einer **Invertierten Datei** realisiert (siehe Kapitel 2.1.4). Im Falle einer Suchanfrage über die Benutzungsschnittstelle (siehe Kapitel 2.1.5) wird zuerst auf den Anfrage-Compiler zugegriffen. Dieser unterzieht die Suchanfrage gegebenenfalls einer linguistischen Vorverarbeitung, bevor die für das Suchsystem interpretierbaren Daten an das Modul zur Anfrageausführung weitergeleitet werden. Anschließend werden die der Suchanfrage entsprechenden Daten aus dem Index an das Anfrageausführungsmodul geleitet. Abschließend findet die Berechnung der Rangfolge statt, bevor die Ergebnisse zwecks Präsentation an die Benutzungsschnittstelle zurückgegeben werden. Bei der Suche wird nur auf den Index und nicht auf die Originaldokumente zugegriffen.

### 2.1.2 Zerlegung eines Textes in Wörter

Um die **Deskriptoren**, d. h. das Dokument beschreibende und repräsentierende Wörter, eines Dokuments und damit die notwendigen Indexterme zu gewinnen, müssen die Texte in kleinere Einheiten zerlegt werden. Bei der klassischen Volltextsuche handelt es sich dabei um einzelne Wörter, da jedes Dokument durch seine einzelnen Wörter repräsentiert wird. Der Vorgang der Zerlegung und des Erfassens jedes einzelnen Wortes eines Textes heißt **Tokenisierung** [Hala04, S. 218].

Auf den ersten Blick erscheint eine solche Zerlegung eines Textes in einzelne Wörter trivial, da »ein Wort [...] eine Einheit aus alphanumerischen Zeichen [ist], die zu ihrer Rechten und Linken durch Leerraumzeichen (engl. white space) oder Interpunktion begrenzt wird« [Hala04, S. 218]. Würde man jedoch einen simplen, musterbasierten Algorithmus anwenden, der eine Trennung entsprechend der obigen Wortdefinition vornimmt, so könnten in einigen Fällen Probleme auftreten. Ein einfaches Beispiel ergibt sich bei folgendem Satzbeginn: »Der PRO 7-Moderator Aiman Abdallah...«. Hier würde eine Trennung zu dem Begriff »7-Moderator« führen, der keine sinnvolle Wortform darstellt [Hqw06, S. 67]. Abgesehen von Leerzeichen enthaltenden, zusammengehörigen Begriffen stellen Zahlen eine weitere Herausforderung dar. [BaRi99, S. 166 f.] empfehlen Zahlen nicht als Indexterme zu gebrauchen, da sie ohne ihren eigentlichen Kontext zu unklar bzw. vage sind. Ausnahmen lassen sich über reguläre Ausdrücke, spezielle Ausnahmelisten oder andere, aufwendigere Verfahren regeln. Allerdings sind Zahlen in vielen Fällen zur Unterscheidung notwendig und sollten deshalb als Indexterme aufgeführt werden, um z. B. Suchanfragen zu vergangenen Ereignissen zu ermöglichen. Des Weiteren werden in den meisten Fällen alle Wörter entweder komplett in Groß- oder Kleinschreibung überführt, da dies in der Regel unwichtig für die Identifikation der Indexterme ist [BaRi99, S. 167].

Es lässt sich festhalten, dass die Zerlegung der Dokumente in einzelne Wörter grundsätzlich relativ leicht möglich ist. Dabei können wie bei jedem automatischen Verfahren Fehler auftreten, die sich aber in einem engen Rahmen befinden, da durch die Verwendung von

regulären Ausdrücken o. ä. ein Großteil der gebräuchlichen Satz- und Wortkonstruktionen abgedeckt werden kann. Sinnvoll ist jedoch »zu prüfen, ob die nach der Wortseparierung erhaltenen Objekte aus einer Menge von erlaubten Zeichen bestehen, d. h. auf den ersten Blick wie Wortformen aussehen« [HQW06, S. 67 f.]. Hierzu gehören zumindest die Buchstaben des Alphabets und der Bindestrich, für weitere Symbole sollten entsprechend des Einsatzgebiets der Suche individuelle Entscheidungen getroffen werden.

### 2.1.3 Entfernung von Stoppwörtern

Nachdem ein Dokument in einzelne Wörter zerlegt wurde, findet üblicherweise eine Entfernung der Stoppwörter statt, da sie keine potenziellen Indexterme darstellen [BaRi99, S. 167 f.]. **Stoppwörter** sind vor allen Dingen Artikel, Konjunktionen, Hilfsverben, Präpositionen etc., d. h. »inhaltsleere Worte«, die keine Bedeutung für den eigentlichen Inhalt haben [DGS04, S. 482]. Diese Worte machen meist einen großen Teil eines Textes aus. Die Eliminierung der Wörter geschieht mit Hilfe einer Stoppwortliste. Befindet sich ein Wort des Dokuments in dieser Liste, so wird es nicht indexiert, sondern stattdessen normalerweise mit einem Platzhalter ersetzt [Chak03, S. 48 f.]. Dies hat gegenüber einer einfachen Entfernung den Vorteil, dass die Suche nach Stoppwörter enthaltenden Phrasen weiterhin ermöglicht wird. Zwei Beispiele aus dem Englischen verdeutlichen die Auswirkungen [Chak03, S. 48 f.]:

- 1. Phrasensuche nach »*gone with the wind*«  
Mit »*with*« und »*the*« enthält die gesuchte Phrase zwei typische Stoppwörter. Es kann jedoch trotzdem ein passendes Dokument gefunden werden, das die Phrase enthält, da »*with*« und »*the*« durch Platzhalter ersetzt wurden.
- 2. Phrasensuche nach »*to be or not to be*«  
Diese Suchanfrage kann nicht mehr beantwortet werden, da je nach Umfang der Stoppwortliste alle enthaltenen Wörter als Stoppwörter behandelt werden.

Um dieses Problem zu umgehen, nehmen einige Suchmaschinen tatsächlich sämtliche Wörter eines Dokuments in den Index auf [DGS04, S. 482]. Nachteilig ist hier vor allen Dingen der Effizienzverlust und erhöhte Speicherbedarf. So gehen [BaRi99, S. 167] davon aus, dass durch den Verzicht auf Stoppwörter im Index die Größe der Indexstruktur um 40 % komprimiert werden kann. Es ist außerdem logisch, dass Stoppwortlisten separat für jede Sprache erstellt werden müssen und in ihrer Größe vom jeweiligen Zweck der Suchmaschine abhängig sind. Neben Stoppwortlisten können außerdem sogenannte *Black Lists* benutzt werden [Glög03, S. 56]. Der Unterschied zur Stoppwortliste besteht darin, dass ein in der *Black List* gefundenes Wort zum Ausschluss des gesamten Dokuments führt. Auf die Hintergründe bzgl. der *Black Lists* wird hier jedoch nicht weiter eingegangen.

### 2.1.4 Erstellung eines Indexes

Der nächste Schritt, der nach der Tokenisierung der Dokumente sowie der Entfernung von Stoppwörtern stattfindet, ist das Schreiben eines Indexes bzw. das Einfügen der gewonnenen Indexterme in eine bereits vorhandene Indexstruktur. Die bevorzugte Indexstruktur ist dabei die der **invertierten Datei** bzw. des **invertierten Indexes** [BaRi99, S. 191 ff.], [Glög03, S. 57 ff.]. Der invertierte Index erlaubt es schnell und effizient jedem Schlüsselwort eine Liste aller Dokumente zuzuordnen, in denen das Wort vorkommt. Für eine Betrachtung der Größe des

Indexes in Abhängigkeit zur Dokumentensammlung und mögliche Kompressionstechniken (vgl. [WMB94, S. 72 ff.]). Nach [Glög03, S. 58 ff.] ergibt sich dabei ein System, das auf drei verschiedenen Strukturen basiert (siehe auch Abbildung 2):

■ Direkte Dateien:

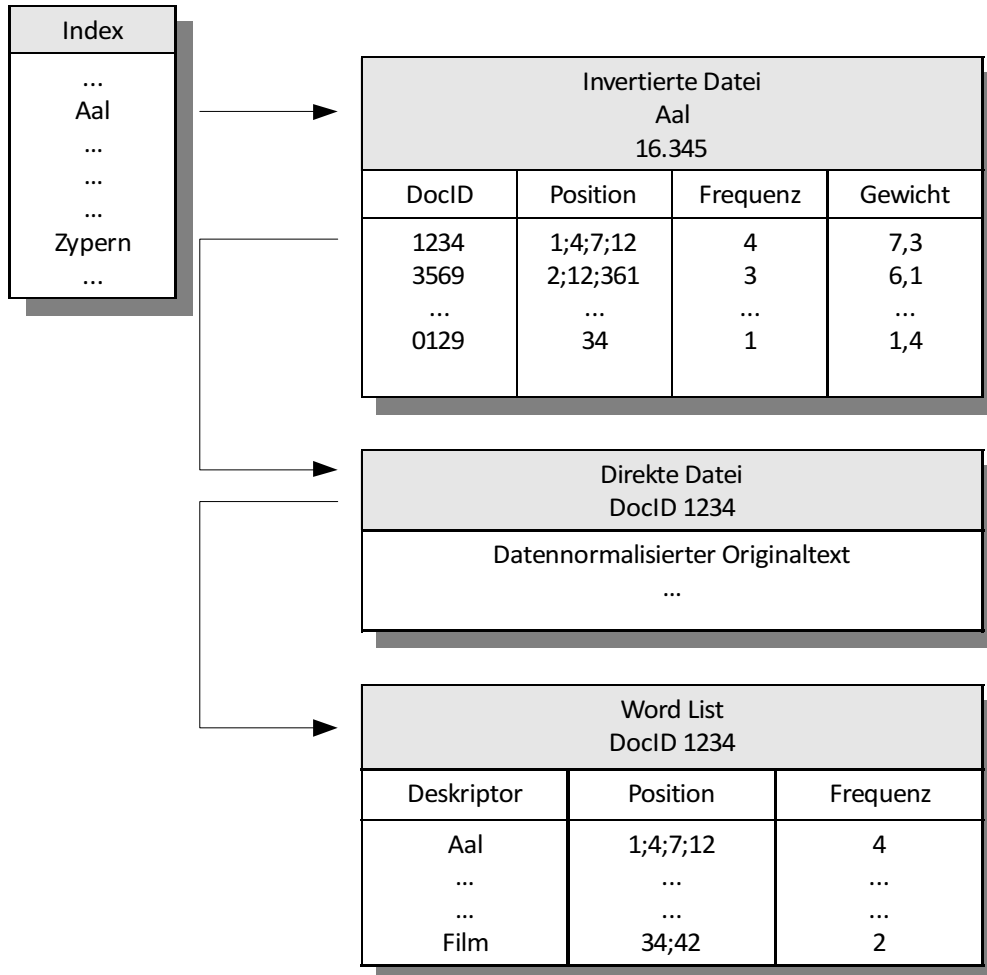
Jedes Dokument wird in ein vom System bevorzugtes Dateiformat umgewandelt und sein Textkörper komplett oder teilweise gespeichert. Zusätzlich kann die Datei um *Head*-Informationen, wie den Dokumententitel, oder Meta-Angaben erweitert werden. Diese Informationen können jedoch auch separat gespeichert werden, um einen effizienteren Zugriff zu ermöglichen. Außerdem enthält die Datei eine Wort-Liste (*word list*), in der sich alle zuvor gewonnenen Indexterme des Dokuments befinden sowie eine eindeutige Identifikationsnummer (*DocID*) zur effizienten Verwaltung.

■ Invertierte Dateien:

Eine invertierte Datei zu einem bestimmten Schlüsselwort enthält Verweise (*DocID*) auf alle direkten Dateien, die in ihrer jeweiligen Wort-Liste das Schlüsselwort der invertierten Datei beinhalten. Außerdem werden zu jedem Verweis weitere Informationen bzgl. der Anzahl des Auftretens des Schlüsselwortes im Dokument sowie der Positionen gespeichert. Weitere Details, die später für die Gewichtung der Ergebnisse hilfreich sein können, wie z. B. das Auftreten des Schlüsselwortes im Titel oder einer Überschrift, werden möglicherweise ebenfalls gespeichert.

■ Indexdatei:

Die Indexdatei enthält grundsätzlich eine Liste aller, aus den Dokumenten gewonnenen, Indexterme. Jedem Indexterm ist über eine ID eine eindeutige invertierte Datei zugeordnet. Bezüglich der Datenstruktur der Indexdatei gibt es verschiedene Ansätze, die hier nicht weiter behandelt werden. Für eine genauere Betrachtung sei auf [BaRi99, S. 196 ff.] verwiesen.



**Abbildung 2: Invertiertes Dateisystem - vereinfachte Darstellung (in Anlehnung an: [Glög03, S. 59])**

Wurde nun aus einem Dokument mit Hilfe der in Kapitel 2.1.2 und Kapitel 2.1.3 beschriebenen Verfahren eine Menge von Indextermen gewonnen, so wird diese Menge mit den in der Indexdatei enthaltenen Schlüsselwörtern verglichen. Außerdem wird für das Dokument eine direkte Datei angelegt. Für alle in der Indexdatei bereits enthaltenen Schlüsselwörter wird in der entsprechenden invertierten Datei ein Verweis auf die erzeugte direkte Datei hinzugefügt sowie die Informationen bzgl. Auftrittshäufigkeit und -position aktualisiert. Sollte ein Indexterm noch nicht in der Indexdatei enthalten sein, so wird eine neue invertierte Datei erzeugt, die einen Verweis auf die zuvor generierte direkte Datei sowie Informationen bzgl. Auftrittshäufigkeit und -position des Schlüsselwortes in der direkten Datei erhält. Der noch nicht enthaltene Indexterm wird in die Indexdatei aufgenommen und bekommt die erzeugte invertierte Datei zugeordnet.

### 2.1.5 Suche unter Berücksichtigung des Indexes

Nachdem in den vorigen Kapiteln das Grundprinzip der Aufbereitung der Dokumente durch eine Volltextsuchmaschine beschrieben wurde, wird nachfolgend die eigentliche Suche durch einen Benutzer thematisiert. Dies bedeutet, dass der Nutzer meist über ein einfaches Formular eine Suchanfrage formuliert. Die Suchanfrage kann sowohl aus nur einem Wort bestehen und somit sehr einfach, als auch komplex sein. Hier wäre z. B. die **Phrasensuche** oder die Benutzung **Boolescher Operatoren** zu nennen. Dabei lassen sich nach [BaRi99, S. 195 f.] drei generelle Schritte zur Bearbeitung einer gestellten Suchanfrage durch das System unterscheiden:

- *Vocabulary search:*  
Die in der Suchanfrage enthaltenen Worte werden isoliert, d. h. der Suchtext wird in einzelne Wörter zerlegt. Dies geschieht unabhängig davon, ob Suchbegriffe in Phrasen vorkommen. Danach wird jedes Wort einzeln im Index gesucht.
- *Retrieval of occurrences:*  
Es werden die Listen aller gefundenen Wörter abgerufen.
- *Manipulation of occurrences:*  
Weitere Bearbeitungen auf Basis der ermittelten Listen finden statt, um die eigentliche Suchanfrage (z. B. Phrasensuche, Suche mit Hilfe von Booleschen Operatoren) zu erfüllen.

Im Falle einer Suchanfrage bestehend aus einem Wort erübrigt sich der dritte Schritt, da durch die zurückgegebene Liste die Suchanfrage erfüllt ist. Interessanter ist z. B. der Fall einer Phrasensuche. Hier besteht der dritte Schritt in der Kombination der zu den Einzelwörtern ermittelten Listen, d. h. der gemeinsame Abgleich, um nur die Dokumente zu finden, in denen die Einzelwörter in der passenden Sequenz vorhanden sind. Dieser Abgleich der verschiedenen Listen gehört zu den zeitaufwendigsten Operationen auf invertierten Indexen [BaRi99, S. 195 f.]. Die Suche findet dabei ausschließlich über den Index statt, d. h. der Text der Originaldokumente wird im Gegensatz zur sequentiellen bzw. Online-Suche nicht mehr durchsucht. Dadurch ist die Suche vor allen Dingen in großen Datenbeständen deutlich schneller. **Erweiterte Suchmöglichkeiten** wie die unscharfe Suche (fuzzy search), die unscharfe Phrasensuche (proximity search), die Joker- oder Wildcardsuche etc. werden an dieser Stelle nur erwähnt und nicht weiter verfolgt (vgl. hierzu [BaRi99, S. 100 ff.], [DGS04, S. 487 f.]).

### 2.1.6 Ergebnispräsentation

Nachdem eine zur Suchanfrage passende Menge von Dokumenten gefunden wurde, müssen diese noch in einer sinnvollen Reihenfolge sortiert werden. Die Präsentation findet meist in Form einer Liste der ermittelten Dokumente statt. Dabei wird oft eine **Sortierung nach Relevanz** bevorzugt. Dies ist notwendig, da besonders bei einfachen Suchanfragen und einer größeren Datenmenge eine Vielzahl von Dokumenten gefunden werden, die den gesuchten Begriff oder die gesuchten Begriffe enthalten. Der Benutzer wird in der Regel jedoch nicht sämtliche zurückgegebenen Dokumente sichten können, sondern erwarten, dass sich die für ihn relevanten Dokumente in der Ergebnisliste ganz oben befinden [Lewa05, S.36 ff.]. Die dazu eingesetzten *Ranking*-Faktoren sowie die zugrunde liegenden Modelle sollen an dieser Stelle nur kurz skizziert werden. Sie werden in der Literatur ausführlich beschrieben und diskutiert (vgl. u. a. [BaRi99, S. 24 ff.], [Chak03, S. 53 ff.], [Glög03, S. 67 ff.], [Lewa05, S. 80 ff.], [DGS04, S. 484 ff.]).

Ein Ansatz für eine Sortierung der Ergebnisse bildet das **Vektorraummodell**. Dabei werden die einzelnen Dokumente als n-dimensionale Vektoren dargestellt, wobei n die Anzahl der Indexterme des Dokuments ist. Somit wird jedes Dokument durch genau einen Vektor repräsentiert, der aus den n Schlüsselwörtern des Dokuments besteht (siehe auch Kapitel 2.4.6). Um das Modell weiter zu verbessern, kann man eine Gewichtung der einzelnen Indexterme vornehmen. Je häufiger ein Wort in einem Text vorkommt, desto aussagekräftiger ist dieses für den Inhalt des Textes. Die Anzahl der Vorkommnisse wird dabei über die Anzahl aller Wörter des Textes normiert. Im Umkehrschluss lässt sich auch feststellen, dass Wörter eines bestimmten Dokuments zu Schlagwörtern werden, je seltener sie im gesamten Index außerhalb des Dokuments vorkommen. Über diese **relative Termhäufigkeit** (TF, *term frequency*) und **inverse Dokumentenhäufigkeit** (IDF, *inverse document frequency*) lassen sich die einzelnen Indexterme gewichten, so dass der Vektor eine bessere Repräsentation des eigentlichen Dokuments darstellt (siehe Kapitel 2.4.5 und 2.4.6). Im folgenden Schritt ist es möglich, zwei Dokumente oder ein Dokument und eine Suchanfrage (repräsentiert durch einen m-dimensionalen Vektor bestehend aus den m einzelnen Suchbegriffen) zu vergleichen und so ein Maß für deren Ähnlichkeit zu bestimmen. Dies geschieht z. B. über den Kosinus des Winkels zwischen beiden Vektoren. Je kleiner dieser ist, desto ähnlicher sind die Dokumente [HQW06, S. 203 ff.].

Weitere mögliche Faktoren für ein *Ranking* sind z. B. der Wortabstand zwischen zwei Suchbegriffen oder die Position der Suchbegriffe im Text bzw. in den *Meta-Tag*-Angaben.

Dies dient letztendlich alles dem Zweck, dem Benutzer eine Ergebnisliste mit möglichst relevanten Dokumenten zu präsentieren. Außerdem sind je nach Einsatzgebiet der Suche auch Sortierungen nach dem Datum oder anderen Kriterien möglich.

## 2.2 Schwächen der Volltextsuche

Die folgenden Abschnitte 2.2.1 bis 2.2.2 stellen kurz die Probleme dar, die bei der Benutzung von Volltextsuchsystemen auftreten können. Dabei wird eine Unterscheidung zwischen der Aufbereitung des Suchraums, d. h. der Indexierung der Dokumentenbasis durch das System und den entstehenden Problemen bei der Interaktion zwischen System und Benutzer getroffen. Der Fokus liegt zum einen auf dem möglichen Verlust der Semantik bzw. Bedeutung bei der Indexierung und der Suche, zum anderen auf den Schwierigkeiten des Nutzers, komplexe Suchanfragen zu formulieren und relevante Inhalte schnell zu finden.

### 2.2.1 Aufbereitung des Suchraums

Das in den vorherigen Kapiteln geschilderte System der Volltextsuche besitzt einige Schwächen. Ein Problem ergibt sich bei der Aufbereitung des Suchraums bzw. der Repräsentation der Dokumente. Jedem Dokument werden eine Vielzahl von Indextermen zugeordnet. In manchen Fällen, in denen keine Entfernung von Stoppwörtern stattfindet, wird sogar der komplette Text indiziert, d. h. jedes im Text enthaltene Wort wird als Repräsentant des Dokuments verwendet. Durch die Verwendung eines invertierten Dateisystems bzw. eines invertierten Indexes werden jedem so gewonnenen Schlüsselwort die entsprechenden Dokumente zugeordnet. Dies hat zwei entscheidende Folgen:

- Einerseits wird bei einer großen Datenbasis jedes im Index enthaltene Schlüsselwort mit einer Vielzahl von Dokumenten verbunden. Die Menge der Dokumente pro Schlüsselwort wird dabei sehr schnell so groß, dass es für den Benutzer unmöglich ist, alle Dokumente zu sichten.

- Andererseits kann nur nach im Index enthaltenen Wörtern gesucht werden. Wird in den Texten der Dokumentenbasis anstatt des gesuchten Wortes ein Synonym, d. h. ein anderes Wort, das jedoch dieselbe Sache bezeichnet, benutzt, so werden die entsprechenden Dokumente nicht gefunden. Dies kann dazu führen, dass keine Ergebnisse geliefert werden, obwohl Dokumente mit thematisch passendem Inhalt vorhanden wären.

Als Ergebnis einer Suche wird dem Benutzer oft eine **unüberschaubare Liste an Dokumenten** übermittelt, von denen trotz *Ranking*-Mechanismen in den Augen des Benutzers viele irrelevant sind. Dadurch kann eine Suche viel Zeit kosten, weil Dokumente geöffnet werden müssen, um ihre Relevanz zu beurteilen. Des Weiteren kann die **Suchdauer** verlängert werden, da neue Suchanfragen formuliert werden müssen, um eine neue Ergebnisliste zu erhalten. Dies kann dazu führen, dass der Nutzer trotz Vorhandenseins geeigneter Dokumente aufgibt, ohne dass sein Informationsbedürfnis befriedigt wurde oder er sich mit einem eher schlechten Ergebnis abfindet.

Eine Ursache dieses Problems ist, dass bei der Abbildung von Dokumenten auf einen Index mit Schlüsselwörtern die Semantik verloren geht, d. h. die Bedeutung einzelner Wörter ohne ihren Kontext abnimmt bzw. falsch interpretiert werden kann. Ein einfaches Beispiel sind **Homonyme**, also Wörter die verschiedene Bedeutungen haben. Das Wort »Bank« kann sowohl das Geldinstitut als auch eine Sitzgelegenheit bezeichnen. Im Index werden dem Wort »Bank« also sowohl Texte über Geldinstitute als auch Dokumente zugeordnet, die die Sitzgelegenheit zum Inhalt haben. Bereits erwähnt wurde die Problematik bzgl. der **Synonyme**. Kommen in einem Text beispielsweise nur die Wörter »Auto« bzw. »Kfz« vor, so wird ein Benutzer, der mit Hilfe des Begriffs »PKW« nach Texten über Autos sucht dieses Dokument nicht als Ergebnis erhalten, obwohl es sich hierbei wahrscheinlich um einen für ihn relevanten Treffer handeln würde. Einen möglichen Lösungsansatz stellen die in den folgenden Kapiteln behandelten semantischen Verfahren dar.

### 2.2.2 Suchanfrage und Ergebnispräsentation

Dem im vorigen Abschnitt beschriebenen Problem könnte man auch durch eine verbesserte Suchanfrage entgegenwirken. Um das Beispiel aus Kapitel 2.2.1 wieder aufzugreifen: Würde, um einen Text über Autos zu finden, die Suchanfrage von »PKW« auf »PKW OR Auto OR Kfz« erweitert, so würden deutlich mehr relevante Treffer gefunden werden. Gleichzeitig steigt jedoch auch die Anzahl der irrelevanten Treffer mit an. Folglich könnte man nun anfangen, einzelne Suchbegriffe aus der Anfrage explizit auszuschließen. Das Problem hierbei ist jedoch, dass »Suchmaschinen [...] eine heterogene Nutzerschaft [bedienen]«, wobei Laien-Nutzer einen Großteil der Suchmaschinennutzer ausmachen [Lewa05, S. 36 ff.]. Dies führt dazu, dass Boolesche Operatoren selten, oder sogar falsch benutzt werden und erweiterte Suchmöglichkeiten ungenutzt bleiben [BaRi99, S.278 f. bzw. S. 389 f.]. Der Nutzer erwartet meist auf einfache Anfragen mit einem Klick die benötigten Informationen zu bekommen. Dazu kommt, dass sich die durchschnittlichen Suchanfragen im Bereich der Internetsuchmaschinen meist auf zwei bis drei Wörter beschränken [Chak03, S. 79]. Auch hierbei können die mit Hilfe von Verfahren aus dem Bereich des *Text Mining* extrahierten Informationen genutzt werden, z. B. um dem Benutzer weitere **verwandte Begriffe zur Verfeinerung** seiner Suchanfrage vorzuschlagen [Lewa05, S. 154 ff.]. Unter dem Begriff *Text Mining* werden dabei computergestützte Verfahren für die semantische Analyse von Texten verstanden [HQW06, S. 3].

Ebenfalls problematisch kann die Form der Ergebnispräsentation über eine einfache Liste sein. Der Nutzer erhält so bei mehrdeutigen Suchanfragen Dokumente, die verschiedene Themenbereiche abdecken, in einer Liste. Das Verfahren der *Clusterbildung* (siehe Kapitel 2.4.7) scheint hier ein Ansatz zu sein, dieses zu visualisieren und dem Nutzer so die Betrachtung seiner Ergebnisse bzw. Spezialisierung seiner Anfrage zu erleichtern.

### 2.3 Semantische Suchmaschinen

Eine Lösung der in Kapitel 2.2 genannten Probleme bietet die semantische Suche. Es wird versucht, das Wissen der Suchmaschinen zu vergrößern, um dadurch die Suchergebnisse zu verbessern und ihre Benutzung intuitiver zu gestalten. Im Unterkapitel 2.3.1 erfolgt eine kurze Definition mit Hinweis auf die beiden gängigen Interpretationen des Begriffs der semantischen Suche. Der Abschnitt 2.3.2 beschäftigt sich mit der Unterscheidung vom Begriff der Volltextsuche und erklärt, warum eine Abgrenzung der beiden Begriffe schwierig erscheint.

#### 2.3.1 Definition

Eine genaue Definition der semantischen Suche erscheint sehr schwierig. Dies liegt zu einem großen Teil daran, dass das Konzept der Semantik, d. h. der Bedeutung von Wörtern selbst von vielen wissenschaftlichen Richtungen mit abgewandelten Bedeutungen verwendet wird und einem gewissen Interpretationsbedarf unterliegt. Eine Definition geht davon aus, dass die semantische Suche »[...] die Bedeutung eines Wortes beim *Ranking* der Dokumente einbezieht und dem Benutzer daraus resultierende Handlungsempfehlungen gibt.« [Zieg08, S. 118 ff.]. Eine zweite, weitergehende Definition sieht in der »[...] semantischen Suche weit mehr als nur die **Disambiguierung**, das Auflösen der begrifflichen Mehrdeutigkeit. Dort geht es darum Zusammenhänge und Relationen zwischen den Suchbegriffen formulieren zu können und dadurch der Suchmaschine Fragen zu stellen, wenn möglich in natürlicher Sprache.« [ebd., S.118 ff.]. Daneben gibt es noch weitere Sichtweisen und Interpretationen des Konzepts der semantischen Suche, die beiden hier genannten Definitionsansätze sind jedoch die gängigsten [ebd., S. 118 ff.]. Bereits hier wird erkennbar, dass der Begriff der semantischen Suche sehr unterschiedlich interpretiert werden kann. Gemeinsam ist allen Sichtweisen, dass es um die Bedeutung von Wörtern oder Texten geht, d. h. sich die Betrachtung des Textes von einer rein syntaktischen Ebene und einzelnen Wörtern teilweise auf eine abstraktere Ebene der Bedeutung und des Textverständnisses durch den Kontext verlagert.

#### 2.3.2 Abgrenzung gegenüber der klassischen Volltextsuche

Aufgrund der in 2.3.1 genannten Schwierigkeiten einer genauen Definition der semantischen Suche ist eine trennscharfe Abgrenzung gegenüber der klassischen Volltextsuche nur schwer möglich. Vielmehr ist es so, dass die Grenzen zwischen beiden Begriffen fließend sind. Die semantische Suche erweitert die Volltextsuche um das **Konzept der Bedeutung**. Damit wird versucht, den in Kapitel 2.2 beschriebenen Problemen, die mit der Indexierung jedes einzelnen Wortes eines Textes verbunden sind, zu begegnen.

Um also eine Suchmaschine als semantisch bezeichnen zu können, reicht es aus, wenn sie (zumindest in geringem Umfang) in der Lage ist, zwischen den einzelnen Bedeutungen homonymer bzw. polysemer Begriffe zu unterscheiden und die Ergebnisse dementsprechend

präsentiert (z. B. über *Clustering*) oder den Nutzer auf diese Mehrdeutigkeiten hinweist. Erfüllt eine Suchmaschine diese Kriterien, d. h. ist sie der Disambiguierung einzelner Wörter mächtig und passt sie ihre Interaktion mit dem Nutzer darauf hin an, so lässt sie sich als einfache Form einer semantischen Suchmaschine betrachten [Zieg08, S. 118 ff.].

Der ungleich komplexere Fall einer semantischen Suchmaschine, die keine Probleme bei der Behandlung der natürlichen Sprache und dem Textverständnis hat, soll an dieser Stelle keine Berücksichtigung finden, da eine Implementierung aktuell noch nicht möglich erscheint. Aber auch der Einsatz anderer Techniken, wie die Grundformreduktion oder Kompositazerlegung, die dafür sorgen, dass eine Suche nicht mehr streng auf Schlüsselwörter begrenzt wird, ist Kennzeichen einer semantischen Suche. Zielsetzung ist hierbei, eine Suche nach Bedeutung zu ermöglichen.

## 2.4 Basistechniken der semantischen Suche

In den folgenden Abschnitten werden einige Grundkonzepte und Verfahren, die im Bereich der semantischen Suche eingesetzt werden, vorgestellt. Dies umfasst sowohl Ansätze aus dem Feld der **Computerlinguistik** als auch aus der Domäne des **Text Mining** und anderen Wissensbereichen. Es erfolgt zunächst in Kapitel 2.4.1 eine kurze Beschreibung der Kompositazerlegung, die z. B. im Hinblick auf eine automatische Erstellung von Begriffshierarchien und -netzen verwendet wird. Abschnitt 2.4.2 beschäftigt sich mit der Grundformbildung, d. h. die Reduktion verschiedener Wortformen auf einen gemeinsamen Stamm. Eine Beschreibung der Verfahren hätte durchaus auch im Bereich der Volltextsuche erfolgen können, dies korrespondiert mit dem in Kapitel 2.3.2 beschriebenen Abgrenzungsproblem gegenüber der Volltextsuche. Da die Grundformbildung jedoch wichtige Voraussetzung für weitere Analysen ist und eine Anwendung in klassischen Volltextsuchsystemen nur teilweise Anwendung findet, wird sie den semantischen Basistechnologien zugeordnet. Im folgenden Kapitel 2.4.3 wird die Zuordnung von Wortarten zu den Wortformen eines Satzes, das sogenannte *Part-of-Speech-Tagging*, erläutert. Die anschließenden Abschnitte 2.4.4 bis 2.4.7 beschäftigen sich mit der Ermittlung von verwandten Begriffen, der automatischen Verschlagwortung von Dokumenten, der Dokumentenähnlichkeit sowie den Verfahren des *Clusterings*. Diese Verfahren bauen partiell auf den vorigen Methoden auf und überschneiden sich teilweise. Sie lassen sich außerdem alle dem Bereich des *Text Mining* zuordnen. Das Kapitel 2.4.8 liefert eine kurze Zusammenfassung über das Potenzial der semantischen Techniken.

### 2.4.1 Kompositazerlegung

Die Kompositazerlegung, auch Dekomposition genannt, trennt Wortformen, die aus mehreren freien Morphemen bestehen, in ihre Bestandteile auf. Ein **Morphem** bezeichnet »das kleinste bedeutungstragende Element der Sprache« [HQW06, S. 327]. Es wird zusätzlich als freies Morphem bezeichnet, wenn es affixfrei als Wort in einem Satz vorkommen kann [HQW06, S. 327]. Das Gegenteil zum freien Morphem bildet das gebundene Morphem. So setzt sich z. B. das Wort »Schornstein« aus dem gebundenen Morphem »Schorn« und dem freien Morphem »Stein« zusammen während »Autohaus« aus den zwei freien Morphemen »Auto« und »Haus« besteht. Es wird direkt ersichtlich, dass eine Auftrennung des Wortes »Schornstein« wenig Sinn ergibt. Im Kontext der Suche kann die Kompositazerlegung eingesetzt werden, um eine **Suche in Wortbestandteilen** zu ermöglichen.

Ein grundsätzliches Problem bei der Zerlegung von Komposita in ihre Bestandteile ist, dass es im Deutschen möglich ist, jederzeit neue Komposita zu bilden. Aufgrund dessen können lexikalische Lösungen nicht in jedem Fall zum Erfolg führen [Nohr05, S. 67]. Die regelbasierte Dekomposition von zusammengesetzten Wörtern ist allerdings kaum lösbar [Nohr05, S. 66]. Aufgrund dessen scheint hier der Ansatz über Wörterbücher zu operieren, um dann mit Hilfe von morphologischen Analysen und Ausnahmeregeln die wahrscheinlichsten Bestandteile von Komposita zu ermitteln, am erfolgversprechendsten.

Dies kann z. B. so implementiert werden, dass ein Kompositum zerlegt wird, falls sich die Einzelbestandteile alle im Wörterbuch oder Index befinden. Dieser einfache Ansatz führt jedoch in vielen Fällen zu Problemen. So könnten bei der Zerlegung des Wortes »Staatsexamen« statt der korrekten Einzelwörter »Staat« und »Examen« (das »s« als Fugenelement fällt weg) fälschlicherweise die Wörter »Staat«, »Sex« und »Amen« ermittelt werden [Nohr05, S. 67]. Dieses Problem lässt sich zum einen über Ausnahmelisten, die das Kompositum sowie die korrekte Zerlegung enthalten, regeln, zum anderen kann versucht werden, über Analysen die Wahrscheinlichkeit einer solchen Trennung zu berechnen. Einen weiteren Ansatz stellt die Zerlegung des zusammengesetzten Wortes in N-Gramme, d. h. Buchstabenfolgen der Länge n, dar. Allerdings können auch hier Probleme auftreten, falls z. B. in einem Kompositum ein Wort der Länge n enthalten ist, das keine Beziehung zu dem zusammengesetzten Begriff besitzt (vergleichbar mit dem Beispiel »Staatsexamen«) [Lewa05, S. 107].

Die Kompositazerlegung hat einige Anwendungsszenarien. Neben der Suche in Wortbestandteilen kann sie z. B. bei der automatischen Verschlagwortung eingesetzt werden (siehe Kapitel 2.4.5). Sie führt aber auch bei Fachtexten zu interessanten Ergebnissen. Bei in der Fachterminologie benutzten Komposita gibt es meist einen »Kopf, der die Wortart des Ganzen bestimmt und dessen Bedeutung (fast immer) durch die Bedeutung der anderen Teile – den so genannten Modifikatoren – eingeschränkt wird [...]. Auf diese Weise entstehen Begriffshierarchien, da der Kopf eines Kompositums meist ein Oberbegriff (Hyperonym) des Ganzen ist« [HQW06, S. 241].

### 2.4.2 Grundformbildung

Die Grundformreduktion (im Englischen *stemming* genannt) reduziert Wörter auf eine gemeinsame Grund- oder Stammform. Notwendig wird dies durch die, durch **Flexion** und **Derivation** entstehenden, unterschiedlichen Wortformen eines Wortes. »Flexion dient der Ableitung grammatischer Vollformen aus einer Grundform und schafft somit syntaktische Oberflächenvarianten ein und desselben Wortstammes.« (z. B. »Kind« wird zu »Kind-er«) während »Derivation [...] neue Wortformen durch Anfügung (*Affigierung*) von Derivativen an Wortstämme [schafft]« (z. B. wird aus dem Verbstamm »lös« das Adjektiv »lös-bar«) [HQW06, S. 328 f.]. Im Kontext der Suche kann durch eine Grundformbildung erreicht werden, dass auch **grammatikalische Veränderungen der ursprünglichen Suchbegriffs** gefunden werden.

Vertritt man den Standpunkt, dass sich dabei die Semantik kaum ändert, d. h. die Bedeutung verschiedener Wortformen eines Wortes ähnlich ist, so ist eine Grundformreduktion wünschenswert. »Für den Benutzer einer Suchmaschine ist es meist nicht von Bedeutung, ob in einem Text von »Haus« oder »Häusern« die Rede ist. Beides hat fast denselben semantischen Gehalt bzw. der Benutzer der Suchmaschine möchte auch »Häuser« finden, wenn er nach »Haus« sucht.« [HQW06, S. 242]. Es kann ebenfalls notwendig sein, eine Grundreduktion der einzelnen Wörter eines Dokuments vorzunehmen, falls danach weitere Analyseverfahren benutzt werden, die nicht zwischen den verschiedenen Wortformen unterscheiden sollen, z. B. zur Bestimmung der Frequenz eines Wortes zur Bestimmung von möglichen Schlagwörtern für den Text. [Nohr05, S. 68 f.] unterscheidet dabei zwischen drei verschiedenen **Grundformen**:

- Die formale Grundform
- Die lexikalische Grundform
- Die Stammform

Bezogen auf das englische Wort *absorbancy* (lexikalische Grundform) stellt *absorbanc* die formale Grundform und *absorb* die Stammform dar [Nohr05, S. 70]. Dieses Beispiel wird von der Abbildung 3 verdeutlicht. Führt man eine Grundformreduktion im Zuge der Vorverarbeitung vor der Indexierung eines Dokumentes durch, so hat dies mehrere Auswirkungen. Geht man von einem kurzen Text aus, in dem z. B. viermal das Wort »Haus«, zweimal das Wort »Häuser« und einmal das Wort »Häusern« vorkommt, so würde eine fehlerfreie Grundformreduktion alle Wortformen auf die Form »Haus« zurückführen, welche dann insgesamt siebenmal auftreten würde. Das führt zu einem kleineren Index, da nur der Begriff »Haus« und nicht zusätzlich noch die Begriffe »Häuser« und »Häusern« in den Index aufgenommen würden, zum anderen erhöht sich die Trefferanzahl bei der Suche eines Benutzers [BaRi99, S. 168 f.].

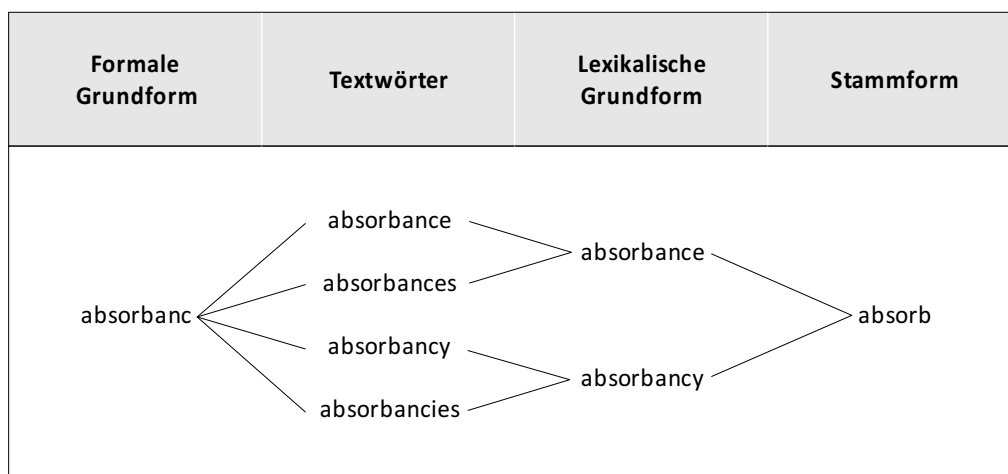


Abbildung 3: Beispiel verschiedener Wortformen (in Anlehnung an: [Nohr05, S. 70])

Für die Grundformreduktion existieren verschiedene Ansätze. Die gebräuchlichsten sind jedoch die Entfernung von Prä- und Suffixen (*affix removal*), die Erkennung und Zurückführung mit Hilfe von digitalen Wörterbüchern (*table lookup*) und die N-Gramm-Methode [Lewa05, S. 107]. Zu beachten ist jedoch, dass sich die Verfahren unterschiedlich gut für verschiedene Sprachen eignen. So führt der Ansatz einer Grundformreduktion mit Hilfe eines regel- und musterbasierten Verfahrens bei Texten in englischer Sprache in den meisten Fällen zu guten und sehr brauchbaren Ergebnissen, während er bei auf deutsch verfassten Dokumenten deutlich schlechtere Ergebnisse liefert. Erklären lässt sich dieser Unterschied durch die Tatsache, dass die englische Sprache deutlich weniger Flexive als die deutsche Sprache besitzt [HQW06, S. 242 f.], [Nohr05, S. 68].

Ein bekannter Algorithmus für die englische Sprache ist der **Porter-Algorithmus** (vgl. [Port80, S. 130 ff.]). Er wendet eine Reihe von Regeln und fallbasierten Unterscheidungen an, um so Affixe zu entfernen, bleibt dabei jedoch relativ simpel und liefert in relativ kurzer

Laufzeit gute Ergebnisse. Die Funktionsweise wird in [BaRi99, S. 168 f., S. 433 ff.] ausführlicher beschrieben. Wie bereits erwähnt existieren ähnliche Verfahren auch für das Deutsche. Aufgrund der komplexeren Wortformenbildung im Vergleich zum Englischen, sind diese jedoch weitaus aufwendiger zu implementieren - da z. B. zahlreiche Ausnahmen bei der Flexion beachtet werden müssen - und liefern weniger genaue Ergebnisse [HQW06, S. 243].

Ein weiterer genannter Ansatz funktioniert mit Hilfe von Wörterbüchern. Dazu wird jedoch ein **Vollformenlexikon** benötigt, in dem für jede Wortform die entsprechende Grundform aufgeführt ist. Die Vor- und Nachteile eines solchen Ansatzes ergeben sich intuitiv. So führt dieser Ansatz zu sehr genauen Ergebnissen und funktioniert auch bei Wortformen, an denen ein regelbasierter Ansatz scheitert. Nachteilig ist jedoch, dass bei diesem Verfahren die Lexika mit Grund- und Vollformen in der Regel manuell erstellt und erweitert werden müssen. So führt dieser Ansatz zwar zu sehr genauen Ergebnissen, im Gegensatz zum musterbasierten Ansatz, bei dem das Regelwerk einmalig erstellt wird, ist jedoch eine kontinuierliche Pflege und Erweiterung des Wörterbuchs notwendig. Um die Vorteile beider Verfahren zu erhalten, werden deshalb in der Praxis oft kombinierte Methoden eingesetzt, die sowohl mit Hilfe von festen Mustern und Regeln arbeiten als auch ein Wörterbuch zur Hilfe nehmen [Chak03, S. 49]. Beiden Verfahren gemeinsam ist außerdem der Nachteil, dass sie sprachabhängig sind, d. h. sie lassen sich im Allgemeinen nur für eine bestimmte Sprache einsetzen [Lewa05, S. 107]. Das ebenfalls erwähnte N-Gramm-Verfahren soll an dieser Stelle nicht weiter vertieft werden. Es genügt lediglich zu wissen, dass ein N-Gramm eine Folge von  $n$  Symbolen bzw. Zeichen darstellt. So lässt sich ein Wort z. B. in mehrere Bigramme (jeweils zwei Zeichen) oder Trigramme (drei Zeichen) zerlegen. Dies wird unter anderem auch verwendet um die Wahrscheinlichkeit von Buchstaben- und Wortfolgen zu berechnen (vgl. [HQW06, S. 102 ff.]).

Letztendlich stellt sich jedoch die Frage nach der Nützlichkeit der Grundformreduktion. Nachdem bereits erwähnt wurde, dass sie für weitergehende Textanalysen von großer Wichtigkeit sein kann, sind die Meinungen über die Verwendung in Suchmaschinen, die auf einer großen und heterogenen Datenbasis operieren - wie Internet-Suchmaschinen - gespalten. [Chak03, S. 49] weist darauf hin, dass die Anwendung einer Grundformreduktion zwar zu einer Vergrößerung der Treffermenge führen kann, damit aber gleichzeitig auch mehr irrelevante Treffer erscheinen. Er führt das Beispiel einer Reduktion der semantisch unterschiedlichen Worte *university* und *universal* auf den Stamm *univers* an. Des Weiteren ist es in manchen Fällen erwünscht nur Dokumente mit bestimmten Formen eines Wortes zu finden. Aufgrund der genannten Gründe ist eine Verwendung von *Stemming*-Algorithmen im Umfeld der Indexierung und Suche nicht komplett positiv zu beurteilen, es kann jedoch zu einer besseren inhaltlichen Erfassung von Dokumenten führen und für semantische Analysen hilfreich sein. Optional könnte dem Anwender erlaubt werden, die Grundformreduktion bei Bedarf abzuschalten. So könnte in den Sonderfällen, in denen eine Grundformreduktion zu keinen guten Ergebnissen führt, auf sie verzichtet werden.

### 2.4.3 Zuordnung von Wortarten

Der Inhalt dieses Abschnittes ist nicht in dieser Vorschau enthalten.

### 2.4.4 Ermittlung von verwandten Begriffen

Ein weiterer wichtiger Bereich umfasst die Ermittlung von verwandten Begriffen bzw. das Erkennen von semantischen Zusammenhängen zwischen verschiedenen Wörtern und Wortformen. Ein Ansatz zur Bestimmung semantisch verbundener Begriffe unterstellt, »dass häufiges gemeinsames Auftreten zweier Wortformen in enger textueller Nachbarschaft ein

starkes Indiz für einen semantischen Zusammenhang ist.« [HQW06, S. 134]. Dieses gemeinsame Auftreten zweier Wortformen in einem festgelegten Textbereich, z. B. einem Satz oder einem Textfenster von beispielsweise fünf Wörtern, wird als **Kookkurrenz** bezeichnet. Geschieht dieses Auftreten so oft, dass es statistisch auffällig ist, so spricht man von einer **signifikanten Kookkurrenz** [HQW06, S. 135]. Um also zu einem gegebenen Begriff verwandte Begriffe zu finden, müssen die signifikanten Kookkurrenzen dieses Begriffes ermittelt werden. Signifikant heißt, dass zwei Begriffe auffällig oft zusammen in einem definierten Bereich vorhanden sind. Um zu beurteilen, ab wann aus einer einfachen Kookkurrenz eine signifikante Kookkurrenz wird, muss eine statistische Überprüfung durchgeführt werden. Man kann dazu die Annahme treffen, dass alle verwendeten Begriffe voneinander statistisch unabhängig sind, d. h. nichts miteinander zu tun haben und das Auftreten und die Reihenfolge der Begriffe rein zufällig ist. Üblicherweise wird das Textfenster außerdem ab so festgelegt, dass man entweder nur den unmittelbaren rechten und linken Nachbarn eines Wortes (Nachbarschaftskookkurrenz) oder das gemeinsame Auftreten in einem Satz (Satzkookkurrenz) betrachtet. Dessen ungeachtet sind je nach Anwendungszweck natürlich auch andere Fenster üblich. Aus der obigen Definition der Kookkurrenzen ergibt sich außerdem, dass es sich bei signifikanten Nachbarschaftskookkurrenzen auch immer um signifikante Satzkookkurrenzen handelt [HQW06, S. 137]. Möchte man nun zu zwei gegebenen Wortformen A und B einen Signifikanzwert berechnen, so benötigt man nach [HQW06, S. 138 f.] vier verschiedene Größen:

- a, b: Anzahl der Sätze, die Wortform A bzw. B enthalten
- k: Anzahl der Sätze, die A und B gemeinsam enthalten
- n: Gesamtzahl der Sätze

Mit Hilfe der Poisson-Verteilung, welche Aussagen über die Wahrscheinlichkeit des Auftretens zufälliger und voneinander unabhängiger Ereignisse macht, lässt sich nun ein **Signifikanzwert** berechnen. Liegt dieser Wert über einem festzulegenden Schwellwert, so liegt eine signifikante Kookkurrenz vor. Die Festlegung des Schwellwertes entscheidet darüber, wieviele signifikante Kookkurrenzen gefunden werden und wie gut oder schlecht diese aus menschlicher Sicht sind. Es ist offensichtlich, dass es sich hierbei um eine individuelle Entscheidung handelt. Aufgrund dessen gibt es auch keinen perfekten Schwellwert, sondern vielmehr einen Übergangsbereich [HQW06, S. 137 ff.]. Als Ergebnis einer Berechnung von signifikanten Nachbarschaftskookkurrenzen lassen sich nach [HQW06, S. 148 f.] verschiedene semantische Relationen aufdecken:

- Über- oder Unterordnung: »Medaillen« zu »Silber«
- Typische Eigenschaften: »rotes« bei »Ampellicht«
- Handlungen und Handlungsträger: »hält« zu »Torwart«, »Saft« zu »trinken«
- Maßangaben zu Stoffen: »Liter« zu »Trinkwasser«

Bei signifikanten Satzkookkurrenzen lassen sich außerdem unter anderem folgende Relationen beobachten:

- Teil-Ganzes-Beziehung: »Seiten« bei »Buch«
- Wirkung zur Ursache: »Feuer« bei »Kurzschluss«
- Hilfsmittel und Werkzeuge: »Lupe« zu »Vergrößerung«

- Gegenteil: »heiß« zu »kalt«
- Orts- oder Personenangaben: »Rathaus« zu »Bürgermeister«, »Heimwerker« zu »Hammer«
- Synonyme: »Schreibkraft« bei »Sekretärin«
- Titel oder Berufe zu Personen: »Präsident« zu »Obama«, »Fußballspieler« zu »Ballack«
- Produktnamen zu Firmennamen: »Astra« zu »Opel«

Die aufgeführten Relationen sind entweder syntagmatischer oder paradigmatischer Natur. **Syntagmatische Relationen** existieren zwischen solchen Wörtern, die oft zusammen in einem gemeinsamen Kontext auftreten. Auf die oben genannten Relationen bezogen bedeutet dies, dass es sich z. B. bei typischen Eigenschaften, Handlungen und Handlungsträgern oder Maßangaben zu Stoffen um syntagmatische Relationen handelt.

**Paradigmatische Relationen** hingegen bestehen zwischen Wörtern, die häufig in vergleichbaren Kontexten auftreten, dies aber nicht zusammen tun. Ein klassisches Beispiel dafür sind Synonyme, aber auch z. B. Wörter der Gegenteilbeziehung [HQW06, S. 162]. Für eine genauere Beschreibung der syntagmatischen und paradigmatischen Relationen (vgl. [Saus66]). Betrachtet man nun die signifikanten Nachbarschafts- oder Satzko-kkurrenzen eines Wortes, so fällt auf, dass es sich dabei oft um Wortformen handelt, die mit dem Referenzwort gemeinsam auftreten. Iteriert man nun das Verfahren der Kookkurrenzbildung, d. h. erzeugt man Kookkurrenzen höherer Ordnung, indem man die ermittelte Menge an Kookkurrenten einer Wortform als Ausgangsbasis für eine erneute Kookkurrenzermittlung nimmt, so erhält man interessante Ergebnisse. So sind z. B. die Kookkurrenzen zweiter Ordnung Wortformen, die häufig in den Kookkurrenzen erster Ordnung erscheinen. Daraus folgt, dass es sich hierbei um Wortformen handelt, die in ähnlichen Kontexten auftreten, dies aber nicht unbedingt zusammen tun [HQW06, S. 163].

Eine andere Anwendung ist die Auffindung fremdsprachlicher Wörter oder typischer Wörter eines Dialekts in einer großen Textsammlung. Hierbei reicht es, eines dieser Wörter anzugeben, da man über die signifikanten Kookkurrenzen schnell einen Großteil der weiteren enthaltenen Wörter dieser Sprache oder des Dialekts erhält. Des Weiteren ist eine Auflösung von Wörtern, die mehrere Bedeutungen haben, möglich, indem ihre signifikanten Kookkurrenten analysiert werden.

[SeBl08, S. 25 ff.] beschreiben vergleichbar mit der oben dargestellten Herangehensweise, dass die meisten Ansätze zur Ermittlung ähnlicher Wörter auf der Annahme beruhen, dass ähnliche Wörter im gleichen Kontext benutzt werden. Es werden zum einen Verfahren vorgestellt, die auf dem Vektor-Raum-Modell basieren, zum anderen Ansätze mit Hilfe syntaktischer Analysen des Kontextes [SeBl08, S. 29 ff.]. Dabei wird davon ausgegangen, dass z. B. zwei Nomen ähnlich sind, falls sie als Subjekt oder direktes Objekt der gleichen Verben auftauchen. Notwendig für dieses Verfahren zur Ermittlung ähnlicher Wörter ist eine vorausgegangene Zuordnung der einzelnen Wortformen zu einzelnen syntaktischen Kategorien (*POS-Tagging*, siehe Kapitel 2.4.3), also eine Kennzeichnung eines Wortes als Nomen, Adjektiv, Verb, o. ä.. Davon abgesehen liegt der Fokus des vorgestellten Verfahrens auf der Ähnlichkeit von Nomen. Nachdem eine syntaktische Zuordnung stattgefunden hat, ergibt sich für jedes Nomen eine Reihe von syntaktischen Relationen bzw. Kontexten (ADJ: Nomen wird vom Adjektiv modifiziert, NN: Nomen wird von anderem Nomen modifiziert usw.). Die Anzahl aller syntaktischen Relationen eines Nomens stellen dessen Attribute dar. [SeBl08, S. 30] nennen als Beispiel das Nomen *cause*, dass in einer medizinischen Textsammlung 83 mal auftritt und dabei 67 einzelne und verschiedene Attribute hat. Diese Attribute, die den Kontext des Nomens darstellen, erhalten nun verschiedene Gewichte. Als letzter Schritt wird die Ähnlichkeit zweier Wörter ermittelt, indem die Summe der Gewichte aller gemeinsamen Attribute durch die Summe der Gewichte der Attribute geteilt wird, die

nur einem von beiden Nomen zugeordnet werden können. Das Verfahren liefert dabei relativ gute Ergebnisse. Es ist außerdem stark von dem zu Grunde liegenden Korpus abhängig, was domänenspezifische Analysen erleichtert [SeBl08, S. 29 ff.].

#### 2.4.5 Automatische Verschlagwortung

Ein weiteres wichtiges Verfahren aus dem Bereich des *Text Mining* umfasst die **automatische Ermittlung von Schlagwörtern** für ein Dokument. Diese sollten in der Regel so gewählt werden, dass sie den Inhalt eines Textes in einigen wenigen Wörtern prägnant zusammenfassen, ohne zu allgemein oder unpassend zu sein. Verschlagwortung bezieht sich dabei auf die Extraktion von Termen aus dem Text, um diese als Schlagworte zu benutzen. Begriffsorientierte Verfahren, die versuchen, die Bedeutung eines Textes zu erfassen, um dann Schlagworte aus einem kontrollierten Vokabular zu verwenden, werden hier nicht näher betrachtet (vgl. hierzu [Nohr05, S. 93 ff.]). Dies hat zur Folge, dass z. B. »Airport« und »Flughafen« als zwei eigenständige Begriffe behandelt werden, so dass unter Umständen beide als Schlagwort eines Textes ermittelt würden. Ein einfacher Ansatz, um potenzielle Indexterme bzw. Schlagworte eines Textes zu ermitteln, ist statistischer Art. Dabei geht man nach [Nohr05, S. 44] von zwei Grundannahmen aus:

- Nicht alle Terme eines Dokuments sind als Schlagworte geeignet – es muss daher eine geeignete Auswahl getroffen werden.
- Nicht alle ausgewählten Schlagworte besitzen hinsichtlich der inhaltlichen Bedeutung die gleiche Wertigkeit – es muss daher eine Gewichtung der Schlagworte vorgenommen werden.

Nicht geeignet sind insbesondere Stoppwörter (siehe Kapitel 2.1.3), da sie keinerlei inhaltliche Aussagekraft haben. Mit Abstand am besten für die Verschlagwortung eignen sich Substantive. Hier sind unter anderem **Eigennamen** interessant, theoretisch ebenfalls denkbar ist eine Substantivierung bzw. Normalisierung inhaltlich bedeutender Verben oder Adjektive. Dabei ist zu beachten, dass nicht alle Verben und Adjektive gleichermaßen für eine Normalisierung geeignet sind, im Zweifelsfall sollte eine Beschränkung auf Substantive stattfinden. Nach der Entfernung der ungeeigneten Wortformen sollte im nächsten Schritt eine **Grundformreduktion** der verbliebenen Wörter vorgenommen werden. Man erreicht dadurch, dass unterschiedliche Wortformen eines Wortes auf die gemeinsame Grundform abgebildet werden (siehe Kapitel 2.4.2), was sich für die statistische Indexierung als nützlich erweist.

Ebenfalls nützlich kann eine (sinnvolle) **Kompositazerlegung** sein (siehe Kapitel 2.4.1). So kann z. B. das Kompositum »Fußballweltmeisterschaft« in die Begriffe »Fußball« und »Weltmeisterschaft« zerlegt werden. Handelt ein Text nun inhaltlich von Fußballweltmeisterschaften, ist es vorteilhaft, wenn sowohl das Kompositum als auch die Einzelwörter als Schlagwörter verwendet werden. Wurden diese Vorverarbeitungen abgeschlossen, müssen nun aus den übrig gebliebenen Termen die inhaltlich repräsentativen Schlagwörter gewonnen werden. Auch hier geht [Nohr05, S. 46] wieder von zwei populären Grundannahmen aus:

- Häufig auftretende Wörter haben hinsichtlich der Bedeutung *eines Dokuments* eine höhere Signifikanz als jene Wörter mit einem geringen Vorkommen und sind aus dieser Sicht bessere Schlagwörter.
- Seltener auftretende Wörter haben innerhalb einer *Dokumentensammlung* einen höheren Diskriminanzeffekt als häufig vorkommende Wörter und sind damit aus dieser Sichtweise bessere Schlagwörter.

Es ist also nicht ausreichend, jedes Dokument für sich isoliert zu betrachten, um dann die im Dokument am häufigsten vorkommenden Wörter als Schlagworte zu verwenden. Kommt beispielsweise in einem Text sehr häufig das Wort »Computer« vor, so kann man davon ausgehen, dass es in dem Text um Computer geht und der Begriff »Computer« ein gutes Schlagwort darstellt. Ist dieser Text aber Teil einer großen Dokumentensammlung, die domänenspezifisch hauptsächlich Texte aus dem Computerbereich enthält, relativiert sich dieser Eindruck. Das Schlagwort »Computer« ist ohne Aussagekraft, da sich fast alle Dokumente mit diesem Thema beschäftigen.

Eine Lösung dieses Problems bietet die **Differenzanalyse** [HQW06, S. 95 ff.]. Mit ihrer Hilfe lassen sich potenzielle Schlagwörter bestätigen, indem der zu analysierende Text mit einem Referenzkorpus verglichen wird. Wörter die im Analysekorpus, also dem Text, für den die Beschlagwortung stattfinden soll, deutlich häufiger vorkommen als im Referenzkorpus, d. h. allen anderen Texten, sind besonders geeignet, um den Einzeltext zu charakterisieren und ihn von den anderen Dokumenten abzugrenzen und zu unterscheiden. Die Differenzanalyse kann außerdem als Teilverfahren verwendet werden, um aus domänenspezifischen Texten Fachterminologie zu extrahieren, eine weitere Betrachtung der dazu notwendigen Schritte erfolgt hier jedoch nicht (vgl. hierzu [HQW06, S. 272 ff.]). Um also diejenigen Wörter eines Textes zu finden, die den Inhalt eines Textes gut charakterisieren, ohne zu unspezifisch und allgemein zu sein, werden diejenigen Wörter ermittelt, die möglichst häufig im Text und möglichst selten in der Dokumentensammlung auftreten. Man kann dabei die Einschränkung machen, dass eine gewisse Mindestfrequenz des Terms im Referenzkorpus gegeben sein muss, um zu spezifische oder allgemeine unbekannte Begriffe als Schlagwörter auszuschließen [HQW06, S. 99]. Hier wird implizit die Wichtigkeit des **Referenzkorpus** klar. Zum einen spielt die Größe eine Rolle. Je größer das Korpus, desto differenzierter gestalten sich im Allgemeinen die Schlagwörter. Zum anderen ist die Domäne entscheidend. Je spezifischer die Dokumentensammlung, desto spezifischer auch die Schlagwörter.

Ebenfalls gut als potenzielle Kandidaten eignen sich **Eigennamen**, d. h. Namen von Personen, Organisationen, Orten, Produkten etc.. Zur Erkennung von Eigennamen kann z. B. der Einsatz von Wörterbüchern dienen, hier besteht natürlich die mehrfach angesprochene Problematik, dass die Erstellung und Pflege eines solchen Wörterbuchs sehr aufwendig ist. Man erreicht hiermit zwar eine gewisse Genauigkeit, jedoch keine Vollständigkeit. Auch sind viele Begriffe mehrdeutig und treten sowohl als Nach- oder Ortsnamen oder mit weiteren Bedeutungen auf. Falls bereits ein *Part-of-Speech-Tagging* (siehe Kapitel 2.4.3) stattgefunden hat, sind die Eigennamen bereits mit einem *Tag* versehen, hier ist also gar kein zusätzlicher Aufwand notwendig. Als weiteres Verfahren bietet sich eine **Mustererkennung** an. So sind Namen oft von der Form *Titel Vorname Nachname*, Firmen können an Zusätzen wie GmbH oder AG identifiziert werden [Nohr05, S. 131 f.].

Abschließend lässt sich festhalten, dass die automatische Verschlagwortung enorm hilfreich für die Repräsentation von Dokumenten sein kann. Sie bietet auch die Möglichkeit, verwandte Dokumente zu ermitteln, indem die Schlagworte verglichen werden. Schlagworte ermöglichen zudem eine schnelle **inhaltliche Schwerpunktbestimmung der Dokumente**. Die Schlagworte können außerdem als **Vorschläge für neue Suchanfragen** verwendet werden. Dabei muss es das Ziel sein, eine für den Menschen möglichst verständliche und nachvollziehbare Schlagwortstruktur zu generieren.

#### 2.4.6 Dokumentenähnlichkeit

Die Suche nach ähnlichen Dokumenten bzw. die Bestimmung der Ähnlichkeit zweier Dokumente ist ebenfalls von Bedeutung. Hat bereits eine Verschlagwortung stattgefunden, so besteht eine Möglichkeit in dem Vergleich der Schlagworte zweier Dokumente. Die Übereinstimmung lässt sich als Prozentwert darstellen und kann dem Benutzer präsentiert

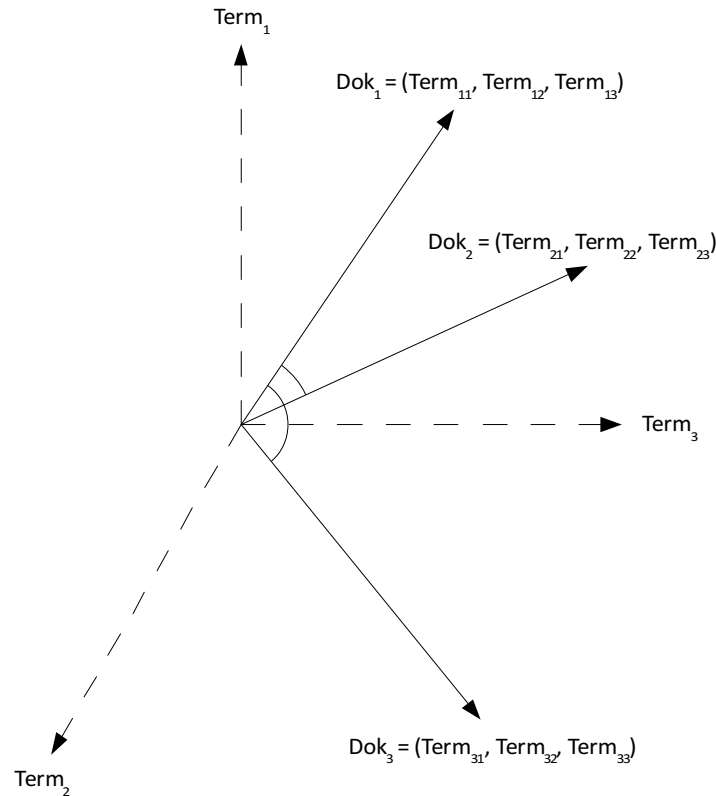
werden, um ihn so auf ähnliche Inhalte aufmerksam zu machen. Ein mögliches Anwendungsbeispiel ist ein **Online-Shop**, wo ein Benutzer auf Produkte aufmerksam gemacht wird, die ihn auch interessieren können.

Aber auch wenn noch keine Verschlagwortung stattgefunden hat, lässt sich eine Ähnlichkeit bestimmen. Der Ansatz dazu gleicht dem Verfahren der Schlagwortauswahl bei einzelnen Dokumenten. Man greift auch hier auf die Grundannahme zurück, dass sich der Inhalt eines Dokuments in seinen am häufigsten verwandten Begriffen widerspiegelt, Stoppwörter ausgeschlossen. Obwohl das Vorgehen dem der Verschlagwortung gleicht, soll es an dieser Stelle noch einmal kurz skizziert werden. Nach Entfernung der Stoppwörter ermittelt man für jeden übrig gebliebenen Term im Dokument seine Häufigkeit (**Termfrequenz**). Da die Häufigkeit jedes einzelnen Begriffes mit zunehmender Länge logischerweise steigt, wird eine **Normalisierung** notwendig. Hier gibt es zwei populäre Ansätze [HQW06, S. 203], [BaRi99, S. 28 f.]:

- Division der Termfrequenz durch die Anzahl aller im Text vorkommenden Wörter
- Verwendung der Frequenz des häufigsten Wortes des Textes als Divisor (die Häufigkeit des Terms, für den eine Normalisierung stattfinden soll, ist wieder der Dividend).

Wie im Bereich der Verschlagwortung wird auch hier die Frequenz der Wörter des Textes im Gesamtkorpus berücksichtigt, seltenere Wörter werden dabei bevorzugt (siehe Kapitel 2.4.5). Dies geschieht durch Berechnung der inversen Dokumentenfrequenz. Nun lässt sich für jedes Wort im Text ein Gewicht berechnen, dass sich aus der Multiplikation von normalisierter Termfrequenz und inverser Dokumentenfrequenz ergibt (oder es werden Variationen dieser einfachen Formel benutzt) [BaRi99, S. 29 f.]. Je höher das Gewicht, desto repräsentativer ist ein Term für den Inhalt des Dokuments. Bis hierhin gleicht das Verfahren ziemlich genau dem der Verschlagwortung, nun muss jedoch ein Vergleich der Ähnlichkeit der Dokumente auf Basis der gewichteten Terme stattfinden.

Dazu wird meist das **Vektorraum-Modell** benutzt, in welchem ein Dokument durch einen n-dimensionalen Vektor repräsentiert wird. n steht dabei für die Anzahl der Wörter des Dokuments, für die jeweils das Gewicht berechnet wurde. Um die Anzahl der Dimensionen und damit die Komplexität späterer Berechnungen zu reduzieren, ist es möglich, einen Mindestwert für das Gewicht der Terme festzulegen, so dass unwichtige Terme nicht in den Vektor aufgenommen werden [HQW06, S. 205]. Abbildung 4 zeigt ein einfaches Beispiel eines durch drei Termvektoren aufgespannten Vektorraums.



**Abbildung 4: Beispiel eines Vektorraums mit drei Dokumenten (in Anlehnung an: [Lewa05, S. 84])**

Möchte man nun Dokumente hinsichtlich ihrer Ähnlichkeit vergleichen, so vergleicht man ihre Dokumentenvektoren. Auch hier gibt es wieder unterschiedliche Möglichkeiten, um ein Maß der Ähnlichkeit zu bestimmen, z. B. indem man das Skalarprodukt der Dokumentenvektoren, die Euklidische Distanz der Dokumentenvektoren oder den Kosinus des Winkels zwischen den Dokumentenvektoren benutzt (vgl. [BaRi99, S.27 ff.], [Hqw06, S. 206 f.]). Verwendet man z. B. den Kosinus, um zwei Dokumentenvektoren zu vergleichen, so steht ein Zahlenwert von eins für eine vollständige Übereinstimmung (z. B. wenn man das Dokument mit einer exakten Kopie vergleicht) und Null für das absolute Gegenteil. [AIDo08] weisen auch auf die Nachteile des klassischen Vektorraum-Modells hin, die ihrer Meinung nach insbesondere durch die Behandlung von Synonymen als eigenständige Wörter und Homonymen und Polysemen als ein Begriff gekennzeichnet sind, auch wenn zugegeben wird, dass dies durch vorhergehende Analysen und Vorverarbeitungen teilweise aufgelöst werden kann. Ein weiterer Nachteil kann durch die bei großen Texten enorme Anzahl an Dimensionen im Vektorraum-Modell gegeben sein. Hier scheint die Beschränkung auf die wichtigsten Terme eine Lösung zu sein.

### 2.4.7 Clusterbildung

Ein elementares Verfahren aus dem Bereich des *Text Mining*, das sich ebenfalls zur Aufbereitung von Suchergebnissen einsetzen lässt, ist das sogenannte **Clustering**. »Bei einer Cluster-Analyse wird eine Menge von Elementen (Daten, Objekte) in *Cluster* (Teilmengen, Gruppen, Klassen, Kategorien) eingeteilt, die einen natürlichen Zusammenhang zwischen den Elementen widerspiegeln sollen« [HQW06, S. 196]. Die Cluster werden in der Regel aus den zu analysierenden Texten abgeleitet und nicht fest vorgegeben. Auch hier geht es wieder darum, die Ähnlichkeit von Dokumenten oder Wortformen zu bestimmen, um diese dann in semantisch möglichst ähnlichen Teilmengen anzuordnen. »Das Ziel eines jeden *Clustering*-Verfahrens ist es, das Ergebnis so zu optimieren, dass der »Abstand« von Texten innerhalb eines *Clusters* möglichst minimal und die Distanz zwischen verschiedenen *Clustern* möglichst maximal ist« [DGS04, S. 493]. Es ist aber nicht nur eine Einteilung der Gruppen in Bezug auf den Inhalt der Dokumente denkbar. Ebenso kann z. B. eine Zuordnung der Dokumente nach Quelle, Datum, Dokumenttyp, Länge oder Sprache wünschenswert sein [Lewa05, S. 165].

Die Nützlichkeit von *Clustering*-Verfahren wird schnell deutlich, wenn man ein Beispiel aus dem Webkontext betrachtet. So nennt [Chak03, S. 79] das englische Wort *star* als möglichen Suchbegriff. Problematisch hierbei ist jedoch, dass das Wort *star* im Englischen mit einer Vielzahl von Dingen in Verbindung gebracht werden kann, so z. B. mit populären Schauspielern und Sängern, der Astronomie, verschiedenen Orten, Sportmannschaften, diversen Zeitungen, patriotischen Liedern aus den USA. Um dem Benutzer die Mehrdeutigkeit seines Suchbegriffes sowie die verschiedenen Bedeutungen aufzuzeigen, sind *Clustering*-Verfahren hervorragend geeignet. Sie bilden die verschiedenen Bedeutungen auf einzelne *Cluster* ab, ordnen Dokumente diesen Clustern zu und ermöglichen dem Benutzer dadurch, sich schnell auf die von ihm gewünschte Interpretationen des Suchbegriffs zu beschränken [Chak03, S. 79 f.]. [Chak03, S. 80 f.] beschreibt weiterhin die sogenannte **cluster hypothesis**, die er als Grundlage für den Nutzwert von *Clustering* sieht. Sie sagt aus, dass ein Benutzer, der ein Dokument interessant findet, voraussichtlich weitere Dokumente aus demselben *Cluster* ebenfalls interessant findet. Die Verfahren zur *Cluster*-Bildung lassen sich dabei nach verschiedenen Gesichtspunkten unterscheiden [HQW06, S. 198 f.]. Eine Unterscheidungsmöglichkeit stellt die **Clusterzugehörigkeit** dar:

- **hart**: Jedes Element wird exakt einem *Cluster* zugeteilt, es ist also Teil genau eines *Clusters*.
- **soft**: Jedes Element kann mehreren *Clustern* zugeteilt werden. Zusätzlich wird ein Zugehörigkeitswert angegeben.

Ein weiteres Anwendungsszenario für *Clustering*-Verfahren besteht z. B. in der Verwendung zur (halb-)automatischen Erstellung von Taxonomien. »Unter einer **Taxonomie** versteht man ein hierarchisches Klassifikationsschema, das dazu geeignet ist, eine Wissensdomäne inhaltlich zu strukturieren« [DGS04, S. 494]. Eine Herausforderung stellen dabei vor allem die oberen Ebenen der Struktur dar, da hier vom Text abstrahiert werden muss und das Vokabular des Textes möglicherweise nicht ausreichend ist, um die Kategorien angemessen zu beschreiben [DGS04, S. 494].

Einige typische Probleme beim Einsatz von *Clustering*-Verfahren im Webkontext nennt [Lewa05, S. 164]. So kann es z. B. passieren, dass Akronyme als *Cluster*bezeichnungen verwendet werden oder Synonyme zu einer falschen Zuordnung führen. Ebenfalls nicht optimal ist die Verwendung unvollständiger Phrasen oder von zu allgemeinen Begriffen, wie z. B. von Stoppwörtern als *Cluster*bezeichnung. Diesen Fehlern kann jedoch mit einer

Analyse durch Verfahren, wie sie in den vorherigen Abschnitten skizziert wurden, begegnet werden. Insgesamt betrachtet scheinen *Clustering*-Verfahren gut geeignet, um Muster oder Strukturen aufzudecken und anschaulich zu visualisieren.

#### 2.4.8 Zusammenfassung

Die in den vorigen Kapiteln genannten semantischen Techniken haben ein großes **Potenzial** in Bezug auf die Verbesserung von Suchmaschinen. Sie ermöglichen es, durch die Kombination von linguistischen und statistischen Analysen, den eigentlichen Inhalt und **die Bedeutung von Texten stärker zu erfassen** und zu berücksichtigen, als dies durch das Konzept der klassischen Volltextsuche erreichbar ist. Verfahren, wie die Kompositazerlegung oder Grundformbildung, berücksichtigen semantische nahe Begriffe bei der Suche. Die Zuordnung von Wortarten versucht, den grammatikalischen Kontext zu erhalten. Weitere Ansätze, wie die automatische Verschlagwortung oder die *Clusterbildung*, helfen zudem auch bei der Einordnung der Dokumente und einer prägnanten Beschreibung und Darstellung ihres Inhalts. Durch die Weiterentwicklung der beschriebenen Verfahren und anderer Ansätze ist eine Steigerung der Intelligenz der Suchmaschinen zu erwarten. Auch in Bezug auf die automatisierte Textanalyse und die damit verbundene Wissensauffindung sind Verbesserungen anzunehmen. Die semantischen Techniken besitzen das Potenzial, auf diesen Gebieten zu entscheidenden Entwicklungen beizutragen.

Der Inhalt der folgenden Kapitel ist in dieser Vorschau nicht enthalten.

Besuchen Sie bitte [www.W3L.de](http://www.W3L.de), um eine vollständige Fassung dieser Studie zu erwerben.