Universität Leipzig

Fakultät für Mathematik und Informatik

Institut für Informatik

Diplomarbeit

Thema: Klassifikation von Dokumenten mit statistischen und regelbasierten Verfahren

vorgelegt von

Leipzig, Juli 98

Andre‘ Seharsch

„Worte, nichts als Worte.“

Johann Wolfgang von Goethe

Inhaltsverzeichnis

1 Einführung

Der Mensch hat gegenüber allem bisher bekanntem Leben den Vorteil, sein Wissen nicht nur an seine nächsten Nachkommen durch Vererbung und Gesten weiter zu geben, sondern es für nachfolgende Generationen in Form von Schriften und Büchern zu bewahren. Die Gesamtheit so konservierter Informationen bildet das Weltwissen, welches aufgrund verbesserter Strukturen zur Kommunikation, wie der elektronischen Post oder dem Internet, in diesem Jahrhundert einem exponentiellen Wachstum unterlag.

Der enorme Zuwachs an Publikationen macht den Zugang zu den darin enthaltenen Informationen immer schwieriger. Das ein einzelner Mensch über das gesamte Wissen seiner Zeit verfügt, ist in Anbetracht der Größe des heute bestehenden Weltwissens nicht mehr denkbar. Dies blieb nach Ansicht der Philosophen zuletzt Dante Alighieri (1265-1321) vorbehalten. Auch die Verwaltung des Wissensbestandes durch den Menschen in Bibliotheken ist keine befriedigende Lösung mehr, die Informationsflut schnell zu erfassen und geordnet zugänglich zu machen.

Auf der Suche nach Alternativen zur Erschließung des Wissens führten in den 50iger Jahren Wissenschaftler erste automatische Analysen speziell auf dem Gebiet geschriebener Texte durch. Ihre Arbeit umfaßte die Entwicklung von Methoden zur Archivierung und zur Auswertung maschinenlesbarer Texte. Ein Teil ihrer Erkenntnisse wird in dieser Arbeit Verwendung finden, um ein ganz ähnliches Problem im Umgang mit Dokumenten zu lösen.

Viele Unternehmen besonders der Dienstleistungsbranche, wie Banken und Versicherungen, stehen vor der Aufgabe, die Kommunikation mit dem Kunden über den Postweg abzuwickeln. Für sie ist ein großer Aufwand damit verbunden, die eingehenden Kundenbriefe zu sichten, dem entsprechenden Sachbearbeiter zuzuleiten und die erforderlichen Folgeschritte, etwa die Erstellung eines Antwortschreibens, einzuleiten. Im ersten Schritt der Vereinfachung dieses Ablaufs soll die zugesandte und in maschinenlesbare Form konvertierte Geschäftspost dem zuständigen Sachbearbeiter automatisch zugeordnet werden. Hierfür soll diese Arbeit mögliche Wege aufzeigen und eine prototypische Lösung erstellen.

Mit Blick auf die Zukunft ist speziell für häufige und einfache Kundenbriefe eine vollautomatische Beantwortung denkbar, wie die Auskunft über den Kontostand bei einer Bank. Dafür sind jedoch genauere Analysen des Briefeinhalts erforderlich etwa die Extraktion von Kundendaten, für die diese Arbeit wichtige Grundlagen aufzeigt.

2 Allgemeine Indexierungssysteme

Im Gegensatz zur Speicherung von Informationen in einer Datenbank, in der die Daten in Felder von Tabellen oder Objekten eingegeben und von dort über Abfragen ausgewertet werden können, erfordert die Textanalyse spezielle Verfahren zur Gewinnung von Informationen aus Dokumenten und ihrer thematischen Erschließung. Solche Verfahren beruhen auf der Annahme, daß die zur Beschreibung von Sachverhalten verwendeten Wörter, den Inhalt eines Textes hinreichend genau kennzeichnen, um aus diesen sogenannten Indextermen einen Rückschluß auf behandelte Themengebiete zu ziehen. Der allgemeine Ablauf bei der Analyse und der Auswertung einer Menge von Dokumenten soll in den folgenden Abschnitten im Mittelpunkt stehen.

2.1 Allgemeine Textanalyse

Um ein Dokument automatisch auszuwerten, wird von ihm eine interne Repräsentation benötigt. Entsprechend der obigen Annahme wird diese Repräsentation des Inhalts durch die Menge der Indexterme t gebildet und zum Termvektor oder auch Dokumentvektor d zusammengefaßt (siehe Formel 1). In ihm werden die Wörter eines Dokuments zu einer Liste zusammengefaßt, die während der weiteren Verarbeitung für das Dokument steht.

$$d = (t_1, t_2, ..., t_n)$$

Formel 1 Termvektor eines Dokuments

Die Voraussetzung für die inhaltliche Analyse einer Menge von Dokumenten ist ihre Zerlegung in Terme. Die gebildeten Termvektoren werden in einer Matrix zusammengefaßt (siehe Formel 2 (a)), in der für alle Dokumente vermerkt wird, welchen Term sie enthalten. Für die Auswertung wird der Index aus den Dokumenten und ihren zugehörigen Termen invertiert. Mit dieser Methode des invertierten Indexes erhält man eine Zuordnung von Termen zu Dokumenten (siehe Formel 2 (b)).

(a)
$$\begin{array}{c|ccc} & t_1 & \cdots & t_n \\ d_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ d_m & 1 & \cdots & 1 \end{array}$$

(b)
$$\begin{array}{c|ccc} & d_1 & \cdots & d_m \\ t_1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ t_n & 0 & \cdots & 1 \end{array}$$

Formel 2 Beziehung zwischen Matrix der Termvektoren (a) und invertiertem Index (b)

Die Invertierung der Beziehung von Dokumenten zu ihren enthaltenen Termen entspricht der Sicht, die bei der Dokumentanalyse benötigt wird. Bei einer Abfrage werden Terme, die mit logischen Operatoren verknüpft sein können, im invertierten Index gesucht. Aus den gefunden Termen können die korrespondierenden Dokumente bestimmt werden, die als Ergebnis zurückgegeben werden. Im konkreten Beispiel aus Formel 3 liefert die Anfrage (a) auf dem Index (b) als Ergebnis das Dokument zu Termvektor d_3.

(a) $t_1 \wedge \neg t_2$

(b)
$$\begin{array}{c|ccc} & d_1 & d_2 & d_3 \\ t_1 & 1 & 0 & 1 \\ t_2 & 1 & 1 & 0 \\ t_3 & 0 & 1 & 1 \end{array}$$

Formel 3 Beispiel für eine Anfrage auf einem invertierten Index

Im Allgemeinen bildet ein Indexierungssystem I eine Menge von Termen T in eine Menge von Dokumenten D ab. Die Menge der Terme kann beispielsweise in einer boolschen Formel an das System übergeben werden, die mittels der invertierten Matrix ausgewertet wird. Als Rückgabewert der Abbildung wird eine Gruppe von Dokumenten erwartet, die inhaltlich mit den Termen der Anfrage korrelieren.

$$I : T \rightarrow D$$

Formel 4 Abbildung eines allgemeinen Indexierungssystems

2.2 Bewertung eines Indexierungssystems

Die vorgestellten Verfahren ermöglichen aus einer Menge von Dokumenten ein Indexierungssystem zu erstellen und mittels einer Abfrage daraus Dokumente zu extrahieren. Zur Einschätzung der Qualität der erzielten Ergebnismenge stehen die Maße Recall und Precision aus der Informationstheorie zur Verfügung [Pao89],

[Gil83]. Sie geben nicht nur Aufschluß über das Ergebnis, sondern lassen auch Rückschlüsse auf die Effektivität des gesamten Indexierungssystems zu.

Betrachtet man die Menge aller Dokumente, so wird diese durch eine Anfrage in gefundene und nicht gefundene Dokumente zerlegt. Für den Benutzer setzt sich die Gesamtheit der Dokumente aus einem relevanten und nicht relevanten Teil (siehe Abbildung 1) zusammen, wenn er die Dokumente nach einem bestimmten Thema durchsucht. Die aus den beiden Zerlegungen entstandenen Teile werden im folgenden zur Bewertung eines Indexierungssystems herangezogen.

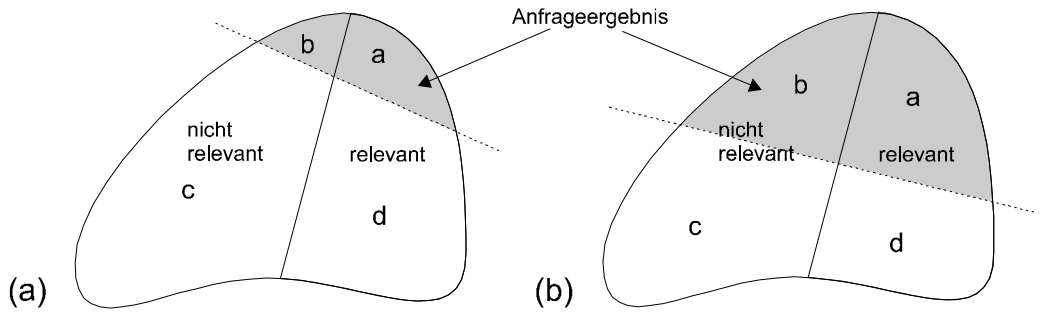


Abbildung 1 Recall und Precision für speziell (a) und allgemein (b) formulierte Abfragen [Sal89]

Der Parameter Recall R gibt an wie umfassend ein Anfrageergebnis ist. Er beschreibt das Verhältnis von relevanten gefundenen zu existierenden relevanten Dokumenten. Ein hoher Recall ($R \approx 1$) für das Ergebnis einer Anfrage weist auf ihren großen Anteil relevanter Dokumente bezüglich der Gesamtheit aller relevanten Dokumente im System hin, und spricht somit für ein Indexierungssystem.

$$R = \frac{a}{a + d}$$

Formel 5 Berechnung des Recalls

Die Precision P zeigt die Genauigkeit eines Anfrageergebnisses. Dazu werden die relevanten gefunden zu allen gefundenen Dokumenten ins Verhältnis gesetzt. Liefert eine Anfrage ein Ergebnis mit hoher Precision ($P \approx 1$), so ist ein großer Teil der gefunden Dokumente auch relevant, das Indexierungssystem arbeitet gut.

$$P = \frac{a}{a+b}$$

Formel 6 Berechnung der Precision

In der Praxis verhalten sich die Parameter gegensätzlich zueinander. Der Wunsch nach hohem Recall und hoher Precision ist selten erfüllbar. Vielmehr bleibt es Aufgabe des Benutzers zu entscheiden, welchem Parameter er den Vorzug gibt. Liegt der Schwerpunkt auf einer möglichst großen Ergebnismenge, ist ein hoher Recall ausschlaggebend bei gleichzeitig sinkender Precision. Wird der Anspruch auf die Genauigkeit der gefundenen Dokumente gelegt, sollte die Precision hoch sein, wobei der Recall abfällt.

2.3 Optimierung der Textanalyse

Der bisher beschriebene Prozeß der Bewertung von Dokumenten bietet viele Ansatzpunkte für eine Verbesserung. Es können Indexterme, Termvektoren und Anfragen spezialisiert und generalisiert werden, um ihre Konzentration auf den Inhalt der Dokumente zu erreichen. Die folgenden Abschnitte erläutern mögliche Verfahren mit einer abschließenden Zusammenfassung ihrer Wirkung.

2.3.1 Verdichtung der Indexterme

Zunächst sollen die Indexterme betrachtet werden, die man aus einem Text durch Identifizierung der Wörter gewinnt. Der Nachteil der direkten Übernahme ganzer Wörter liegt im separaten Behandeln verschiedener grammatikalischer Formen eines Wortstamms. Obwohl das zugrundeliegende Wort nur eine Bedeutung besitzt, wird es mehrfach im Termvektor aufgeführt. Eine Lösung für diese Verzerrung des Textinhalts bietet das Stemming. Es beseitigt vorhandene Vorsilben und Endungen von einem Wortstamm und gleicht gegebenenfalls Umlaute an (siehe Abbildung 2). Diese Generalisierung der Wortformen verdichtet den Inhalt eines Texts und reduziert den Termvektor.

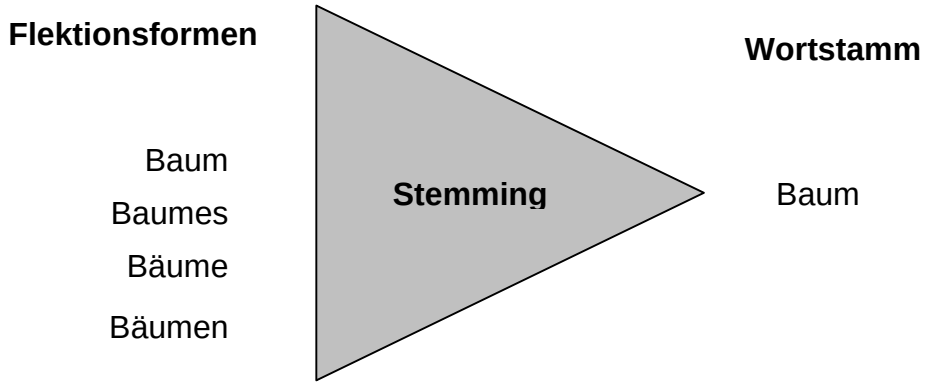


Abbildung 2 Beispiel für das Stemming

Eine weitere Einschränkung auf wichtige Terme für die Indexierung, wird durch die Entfernung häufiger Funktionswörter erreicht [Fra92]. Sie werden auch als Stopwörter bezeichnet. Zu ihnen zählen vor allem Wörter aus abgeschlossenen Wortklassen wie die Artikel `der` oder `ein`, welche in Texten zwar oft vorkommen aber keine Informationen tragen. Für spezielle Themengebiete sollten auch spezielle Stopwörter ausgeschlossen werden. So wird in Texten zum Thema Medizin das Wort `Arzt` kaum zur Differenzierung der Inhalte beitragen. Mit der Eliminierung von Stopwörtern wird einerseits der Termvektor aussagekräftiger und andererseits der Aufwand für die Klassifizierung gesenkt, da die Menge der Indexterm reduziert wird.

2.3.2 Erweiterungen des Termvektors

Neben den Indextermen selbst, bietet auch der Termvektor Möglichkeiten der Optimierung. So können die Indexterme nicht nur ungewichtet, also nur nach Identifizierung im Text, sondern auch gewichtet, beispielsweise nach Häufigkeit des Auftretens in einem Dokument, in den Termvektor aufgenommen werden (siehe Formel 7). Einem Indexterm kann über ein Gewicht x unterschiedlich hohe Bedeutung zugewiesen werden. Dies ermöglicht ein besseres Differenzieren der Dokumente. Ein weiterer Vorteil der Gewichtung ist die leichte Auswertung einer Anfrage, etwa über die Bildung einer Summe der Gewichte für Terme einer Konjunktion. Die Ergebnismenge kann dann geordnet nach den Summen ausgegeben werden, die der Relevanz der Dokumente im System entsprechen.

$$\begin{array}{cccc} & d_1 & \cdots & d_m \\ t_1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ t_n & x_{n1} & \cdots & x_{nm} \end{array}$$

Formel 7 Invertierter Index mit Termgewichten

Eine andere häufige Erweiterung des Termvektors, ist die Angabe einer Entfernung zu einem anderen Term als numerischer Wert, welcher den Bereich der nach links und rechts zu untersuchenden Wörter festlegt. Hierfür wird neben dem Indexterm zusätzlich seine Position im Satz zusammen mit einer eindeutigen Nummer für den Satz in den Termvektor aufgenommen. Aus diesen Angaben läßt sich ermitteln, in welchen Dokumenten das Wort `treten` in der Nähe von `ab` und welchen in es in der Nähe von `an` zu finden ist. Der Vorteil dieser Entfernungsangabe in einer Anfrage resultiert aus einer besseren Spezialisierung eines Suchbegriffs, wodurch eine höhere Genauigkeit des Ergebnisses erzielt wird. Der größere Aufwand bei der Ermittlung der relevanten Dokumente wirkt sich aber negativ auf die Geschwindigkeit der Verarbeitung einer Anfrage aus. Die Verarbeitungszeit steigt stark mit wachsender Distanz, die deshalb möglichst klein gehalten werden sollte.

2.3.3 Unterstützung der Anfrage

Stand bisher die Optimierung der internen Repräsentation eines Dokuments im Mittelpunkt, so sollen jetzt Verbesserungen der Abfrage betrachtet werden. Bei der Nutzung eines Indexierungssystems besteht für den Anwender das Problem, für sinnverwandte Wörter verschiedene Anfragen formulieren zu müssen. Abhilfe schafft hier der Einsatz eines Thesaurus, der generalisierte Termbeziehungen enthält [Fra92]. Er bildet automatisch weitere Anfragen, indem er Terme der Nutzeranfrage durch Synonyme ersetzt. Die Ergebnisse der verschiedenen Anfragen werden wieder zu Einem zusammengefaßt und dem Nutzer präsentiert.

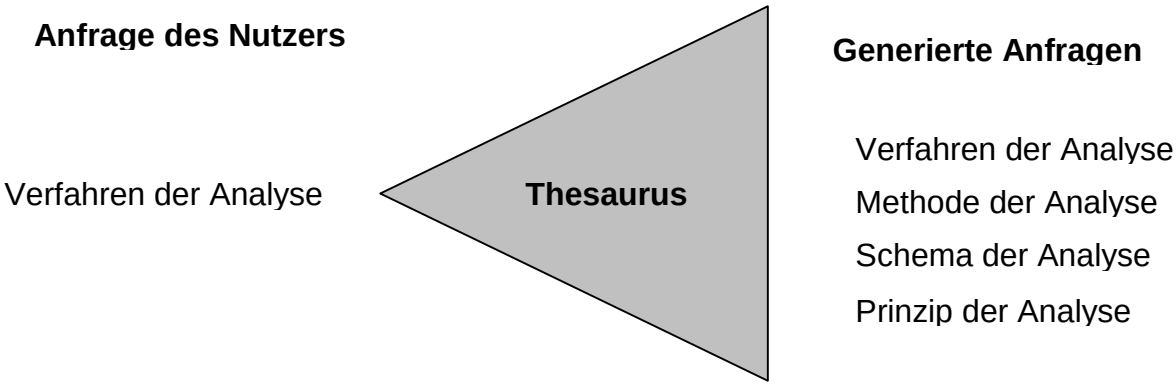


Abbildung 3 Beispiel für den Einsatz eines Thesaurus

Der Nachteil bei der Anwendung eines Thesaurus ist die rein syntaktische Analyse und die Ersetzung von Termen in der Anfrage. Eine weiter Steigerung der Qualität des Ergebnisses verspricht die semantische Analyse einer Anfrage [Gre89]. Diese basiert auf linguistischen Methoden, welche die Anfrage des Nutzers zunächst auf ihre Bedeutung untersuchen. Unter Verwendung eines Lexikons, das Wörter gemeinsam mit ihren grammatischen Eigenschaften beinhaltet, und einer Grammatik, die zulässige Sequenzen aus Wortklassen des Lexikons definiert, werden aus den gewonnen Informationen sinnverwandte Phrasen generiert. Die Phrasen werden analog zum Gebrauch eines Thesaurus als Anfragen verwendet. Aufgrund der genauen Spezialisierung der Anfrage, hebt diese Methode die Precision eines Ergebnisses stark an.

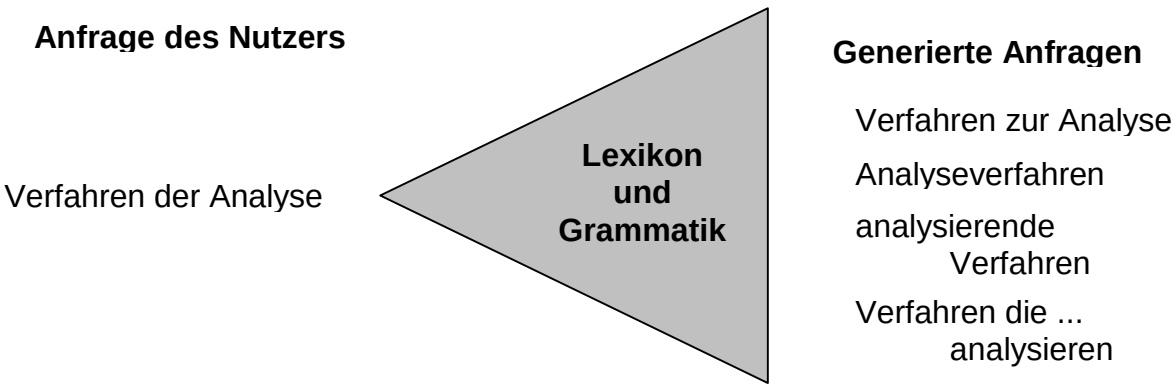


Abbildung 4 Beispiel für den Einsatz von Lexikon und Grammatik

Eine Entlastung des Benutzers von der Konstruktion einer Anfrage aus logischen Operatoren gelingt durch die Quorum-Level-Suche [Sal89]. Hierbei sind lediglich die Terme für die Anfrage zu bestimmen, aus denen der Algorithmus logische Formeln erstellt (siehe Tabelle 1). Je nach Rang der Anfrage, liefern hoch eingestufte Formeln (0) eine kleine aber genaue Ergebnismenge. Anfragen auf

niedrigem Niveau (2) haben ein großes aber ungenaues Resultat. Die Hierarchie der Anfragen ermöglicht es, die gefunden Dokumente ihrer Relevanz nach geordnet auszugeben, um dem Benutzer die Auswahl wichtiger Dokumente zu erleichtern.

Stufe	Formel	Umfang	Genauigkeit
0	$t_1 \wedge t_2 \wedge t_3$	klein	groß
1	$(t_1 \wedge t_2) \vee (t_1 \wedge t_3) \vee (t_2 \wedge t_3)$	mittel	mittel
2	$t_1 \vee t_2 \vee t_3$	groß	klein

Tabelle 1 Anfragen einer Quorum-Level-Suche mit drei Termen

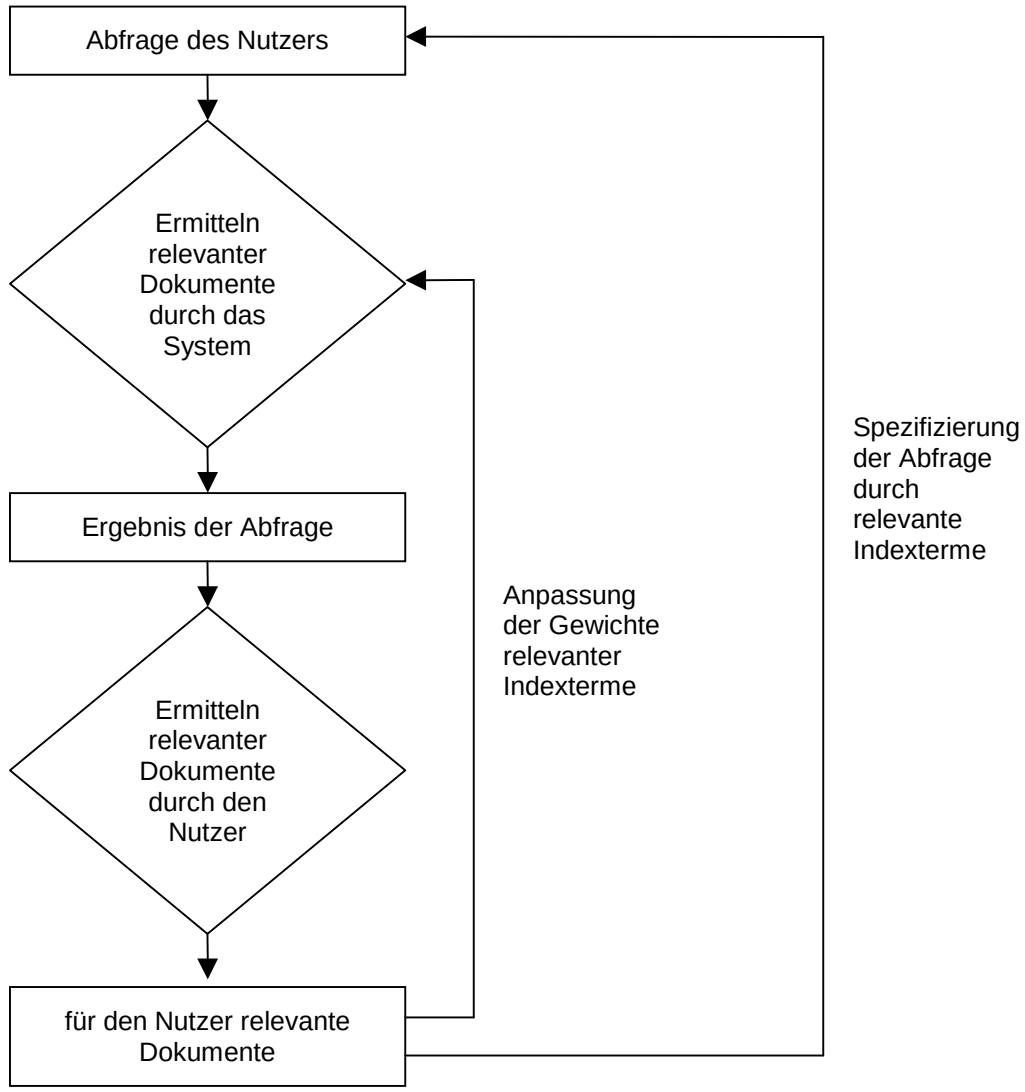


Abbildung 5 Ablauf des relevance Feedbacks

Eine letzte sehr effektive Methode der Optimierung eines Systems arbeitet mit dem von einer Anfrage erzielten Ergebnis, das dem Nutzer zur Auswahl relevanter

Dokumente präsentiert wird (siehe Abbildung 5) [Boc94]. Aus diesen ausgezeichneten Dokumenten können Terme gewonnen werden, um die ursprüngliche Anfrage zu präzisieren. Für die zutreffenden Terme der Anfrage lassen sich außerdem die Gewichte in den relevanten Dokumentvekoren verstärken, so daß sie an Wirksamkeit gewinnen. Dieses Verfahren des relevance Feedbacks steigert Recall und Precision, durch eine kombinierte Auswertung der Inhalte der Dokumente und der Interessen des Nutzers.

2.3.4 Übersicht über die Optimierungsmöglichkeiten

Abschließend sollen die in den letzten Abschnitten vorgestellten Methoden der Textanalyse noch einmal zusammengefaßt und in ihrer Wirkung verglichen werden.

Bereich	Methode	Generalisierung	Spezialisierung
Indexterm	Stemming	X	
	Entfernung von Stopwörtern		X
Termvektor	Gewichtung der Terme	X	X
	Entfernungsangaben		X
Anfrage	Thesaurus	X	
	Grammatik und Wörterbuch		X
	Ouorum-Level-Suche	X	X
	Relevance Feedback	X	X

Tabelle 2 Methoden zur Textanalyse und ihre Wirkung

2.4 Gewichtung der Indexterme

Auf die Optimierung eines Indexierungssystems durch gewichtete Terme wurde bereits in Abschnitt 2.3.2 hingewiesen. Das Ziel der Gewichtung ist die Hervorhebung der Indexterme, die besonders zur Differenzierung von Dokumenten beitragen. Bei einer Anfrage aus konjugierten Termen wird z.B. durch Addition ihrer Gewichte die Relevanz der Dokumente bestimmt, wobei höher bewertete Terme stärker in die Bewertung einfließen. Die Ausgabe erfolgt daraufhin geordnet nach der Summe, die den Rang des Dokuments in Bezug auf die Anfrage beschreibt. So lautet das Ergebnis der Anfrage (a) auf einem Index

von Gewichten (b) aus Formel 8 absteigend nach der Bewertung (c) sortiert d_2, d_3, d_1. Verschiedene Arten der Berechnung von Gewichten und ihre Wirkung auf die Indexierung werden im folgenden diskutiert [Rus94].

(a) $t_1 \wedge t_3$

(b)

	d_1	d_2	d_3
t_1	0.3	0.5	0.2
t_2	0.7	0.2	0.4
t_3	0	0.3	0.4

(c)

	d_1	d_2	d_3
$t_1 + t_3$	0.3	0.8	0.6

Formel 8 Beispiel einer Anfrage auf einem invertierten Index von Gewichten

Zum Aufbau eines Indexierungssystems verwendete Dokumente, können in ihrer Größe stark variieren. Setzt man allein die Termhäufigkeit in den Dokumenten als Gewicht ein, ist eine Verzerrung der Auswertung die Folge. In größeren Dokumenten treten Indexterme häufiger auf, die Dokumente werden überbewertet. Darum ist es nötig die Termhäufigkeiten oder auch Termfrequenzen tf_{ij} des Terms i zunächst auf die Termanzahl k_j des Dokuments j zu normieren. Die so erhaltenen normierten Termfrequenzen ntf_{ij} sind für die selben Terme in verschiedenen Dokumenten vergleichbar. Ihr Einsatz als Gewicht verspricht einen hohen Recall bei geringer Precision. Häufige Terme korrelieren zwar oft thematisch mit dem Dokument, sind aber zu unspezifisch für eine genauere Differenzierung des Inhalts.

$$ntf_{ij} = \frac{tf_{ij}}{k_j}$$

Formel 9 Normierung der Termfrequenz

Eine Verbesserung der Precision ist zu erwarten, wenn mehr Augenmerk auf jene Terme gelegt wird, die nur in wenigen Dokumenten anzutreffen sind. Solche niederfrequenten Terme zeichnen das Besondere am Dokumentinhalt aus und differenzieren ihn somit stärker als hochfrequente Terme. Diese Steigerung der Genauigkeit kann über eine Aufwertung der normierten Termfrequenz durch die inverse Dokumenthäufigkeit eines Terms erreicht werden (siehe Formel 10). Diese Häufigkeit berechnet sich aus dem Logarithmus des Quotienten der Dokumentanzahl N des Systems und der Dokumentfrequenz df_i des Terms i. Als Dokumentfrequenz geht dabei die Anzahl von Dokumenten ein, in denen ein Term auftritt.

$$x_{ij} = ntf_{ij} \cdot \log \frac{N}{df_i}$$

Formel 10 Gewichtung unter Beachtung seltener Terme

Neben allgemeinen Eigenschaften von Termhäufigkeiten kann auch die spezielle Wirkung eines Terms für die Unterscheidung von Dokumenten betrachtet werden, um ein Termgewicht zu ermitteln. Einerseits kann die Aufnahme eines Terms in ein Indexierungssystem negative Folgen für die Differenzierung der Dokumente haben, da Hochfrequente Terme in vielen Dokumenten auftreten und darum eine Angleichung der Dokumentvektoren nach sich ziehen. Anderseits tragen mittelfrequente Terme zur stärkeren Differenzierung bei, weil sie einen Sachverhalt nicht zu spezifisch oder zu allgemein beschreiben. Zu spezielle niederfrequente Terme hingegen ändern aufgrund ihrer Seltenheit die Unterscheidbarkeit kaum.

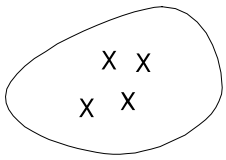
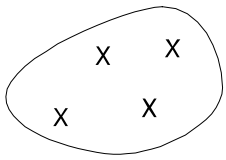
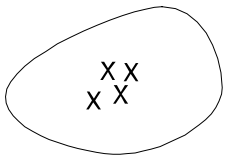
Termfrequenz	niedrig	mittel	hoch
Differenzierung	unverändert	besser	schlechter
Beispiel			
Termunterschied	$dv_i=0$	$dv_i>0$	$dv_i<0$

Tabelle 3 Wirkungsweisen eines Indexterms

Wie nützlich ein Term zur Differenzierung ist, kann durch die folgende Methode bestimmt werden. Zunächst wird für alle Dokumentvektoren eines Indexierungssystems ein Ähnlichkeitskoeffizient Q ermittelt (siehe Formel 11 (a)). Dazu wird eine Funktion b verwendet (siehe Abschnitt 3.2), welche die Unterschiede von Vektoren bewertet. Nach Aufnahme des Terms i in die Dokumentvektoren, ist ein weiterer Ähnlichkeitskoeffizient Q_i zu berechnen. Der Differenzierungskoeffizent dv_i, der sich aus der Differenz der beiden Ähnlichkeitskoeffizienten ergibt (siehe Formel 11 (b)), beschreibt dann die Veränderung die durch den Indexterm i erzielt wurde. Unveränderte Dokumentvektoren ergeben für dv_i gleich Null. Sind sich die Vektoren ähnlicher geworden, so ist dv_i kleiner Null, der Term hat nicht zur Differenzierung beigetragen. Hat sich ihr Unterschied aber vergrößert, wird dv_i größer Null sein.

(a) $$Q = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{k=1 \\ k \neq i}}^{N} b(d_i, d_k)$$ (b) $$dv_i = Q - Q_i$$

Formel 11 Differenzierung von Dokumentvektoren

Fließt das Maß dv_i für die Differenzierung von Dokumentvektoren in die Gewichtung des Terms i ein, so kommt seine spezielle Wirkung auf das Indexierungssytem zum Tragen. Die Terme gehen ihrer Funktion für die Unterscheidung der Vektoren nach gewichtet in die Bewertung ein und der Recall bzw. die Precision der Ergebnisse des Systems optimieren sich.

$$x_{ij} = ntf_{ij} \cdot dv_i$$

Formel 12 Gewichtung unter Beachtung der Differenzierung

2.5 Ablauf einer Indexierung von Dokumenten

Die Arbeit mit einem Indexsystem läßt sich grundsätzlich in zwei Teilaufgaben zerlegen. Der erste Teil betrifft den Aufbau des invertierten Indexes und ist in folgende Schritte zerlegt:

1. Identifizieren jedes Wortes in einer Menge von Dokumenten.
2. Entfernen aller Stopwörter, die keine Bedeutung für die Unterscheidung der Dokumente besitzen.
3. Bilden des Wortstamms für jeden Indexterm (Stemming).
4. Errechnen des Gewichts für jeden Wortstamm.
5. Repräsentieren jedes Dokuments durch die Menge seiner Wortstämme und der zugehörigen Gewichte.

Der zweite Aufgabenbereich eines Indexierungssystems bezieht sich auf die Verarbeitung von Anfragen. Voraussetzung hierfür ist ein existierender invertierter Index, der in die nachfolgenden Arbeitsschritte einbezogen wird:

1. Manuelle oder automatische Konstruktion einer Anfrage.
2. Bilden der Wortstämme aus den Termen der Anfrage (Stemming).
3. Ermitteln der Menge zutreffender Dokumente.
4. Ausgeben der Ergebnismenge geordnet nach der Relevanz.

Abschließend sollen die Aufgaben, die bei der Arbeit mit einem Indexsierungssytem zu bewältigen sind, anhand einer zusammenfassenden Grafik veranschaulicht werden. Es wird noch einmal die Trennung in die Vorverarbeitung der Dokumente bei ihrer Indexierung und die darauf aufbauende Auswertung deutlich.

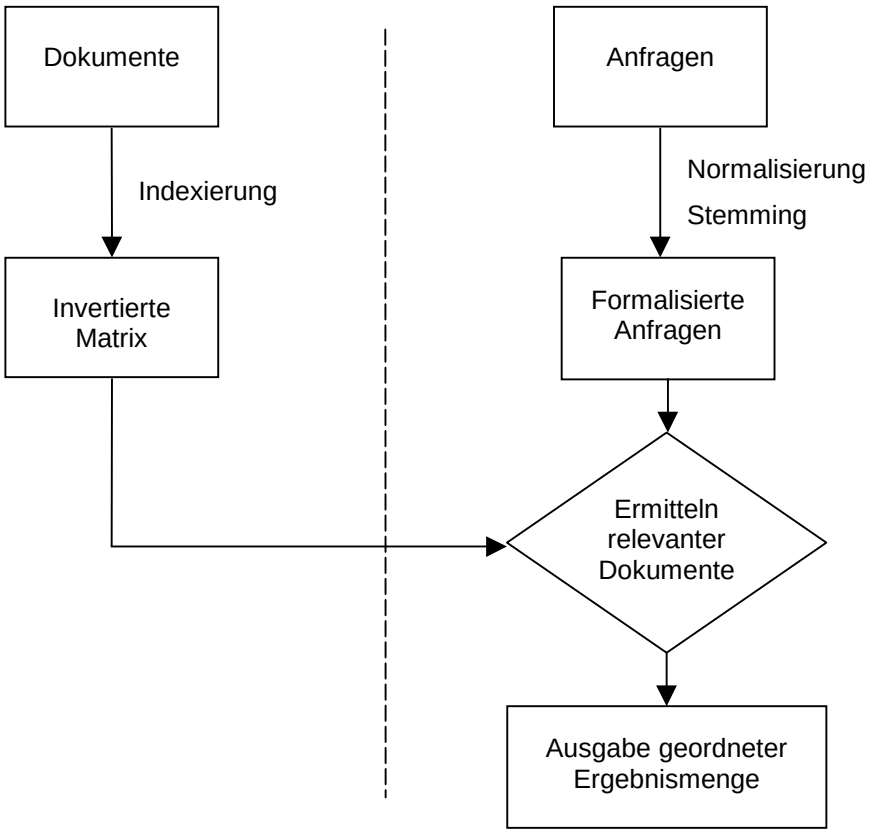


Abbildung 6 Aufgaben eines Indexierungssystems

3 Klassifikation von Dokumenten

Nach der Betrachtung eines allgemeinen Indexierungssystems, sollen in den folgenden Abschnitten die speziellen Belange bei der Klassifikation von Dokumenten im Mittelpunkt stehen. Es wird eine Einordnung des klassifizierenden Systems nach Kriterien der allgemeinen Indexierung vorgenommen. Anschließend wird ein Überblick über den Ablauf der Klassifikation gegeben, deren Aspekte später vertieft werden.

3.1 Klassifizierende Indexierung

Wie bei allen Indexierungssystemen ist auch hier der Text bzw. der Inhalt eines Dokuments der Ausgangspunkt für eine Klassifikation. Das System bezieht keine objektiven Daten über einen Text wie z.B. Angaben über seinen Autor in die Bewertung ein, da diese nicht immer bereit gestellt werden können. Stehen solche Angaben später zur Verfügung, sollten sie ergänzend eingesetzt werden, da sie eine Steigerung der Menge und Genauigkeit von klassifizierten Dokumenten versprechen.

Der Unterschied gegenüber dem allgemeinen Ansatz (siehe Formel 4) liegt bei dem Indexierungssystem I' zur Klassifikation in der geänderten Abbildungsfunktion. Der Definitionsbereich wird wiederum durch eine Menge von Termen T gebildet. Der Wertebereich ist jetzt auf eine feste Anzahl Klassen K beschränkt. Die Terme weisen in ihrer Bedeutung auf eine Menge von Klassen hin und nicht mehr wie bisher auf Dokumente.

$$I': T \rightarrow K$$

Formel 13 Abbildung eines Indexierungssystems zur Klassifikation

Eine weitere Besonderheit des Systems liegt in der Gewinnung von Termen zur Klassifikation. Sie werden nicht mehr vom Benutzer angegeben, sondern sind Bestandteil des Termvektors d eines zu klassifizierenden Dokuments. Soll ein Dokument einer Klasse zugeordnet werden, wird sein Termvektor an das System I'' übergeben, welches daraus die entsprechende Klassen ermittelt.

$$I'': d \rightarrow K$$

Formel 14 Abbildung eines Indexierungssystems zur Klassifikation von Dokumenten

Der Unterschied in der Funktion zieht einen veränderten Aufbau der invertierten Matrix nach sich. Dabei wird von einer Dokumentenmenge ausgegangen, die bereits klassifiziert ist. Die daraus erzeugten Termvektoren werden aber nicht mehr für jedes Dokument separat gespeichert, sondern zu Klassenvektoren zusammengefaßt, indem die Indexterme aller Dokumentvektoren einer Klasse in einen Termvektor übertragen werden. Bei einer Abfrage wird über den Vergleich von Dokument- und Klassenvektor die Klasse des Dokuments abgeleitet.

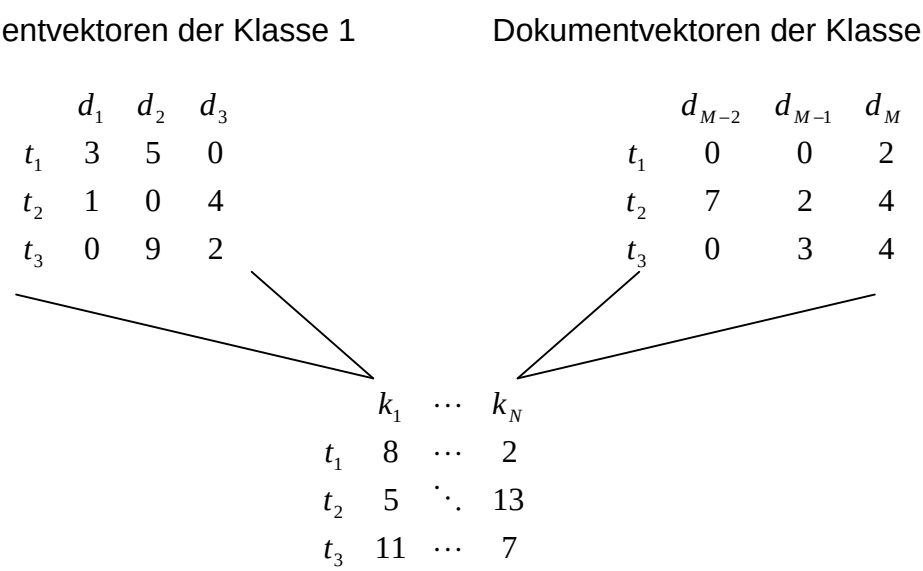


Abbildung 7 Zusammenfassen von Dokument- zu Klassenvektoren

Im Gegensatz zu allgemeinen Indexierungssystemen, in denen nur bereits indexierte Dokumente ausgewertet werden können, ist jetzt auch die Untersuchung systemfremder Dokumente möglich. Die Gruppierung der charakteristischen Wörter einer Klasse erlaubt, im Vergleich mit einem unbekannten Dokumentvektor, immer noch einen Schluß auf die Klasse des Dokuments.

3.2 Bewertung eines Dokuments

Die Zuordnung eines Dokuments zu einer Klasse wird von der Ähnlichkeit seines Dokumentvektors d zum jeweiligen Klassenvektor k bestimmt. Je ähnlicher sich

diese Vektoren sind, desto wahrscheinlicher ist die Zugehörigkeit des Dokuments zu dieser Klasse. Zur Bestimmung der Ähnlichkeit zweier Vektoren stehen verschiedene Bewertungsfunktionen b zur Verfügung, die auf der Ermittlung von Abständen zwischen Vektoren beruhen [Koh95].

Die einfachste Art die Übereinstimmung zweier Vektoren festzustellen, ist das Zählen gemeinsamer Komponenten. Für die Termvektoren heißt das, nach gemeinsam enthaltenen Termen zu suchen. Der Nachteil dieser Methode liegt in einer relativ ungenauen Ermittlung der Ähnlichkeit, da die verschiedenen Bedeutungen der Terme für jede Klasse unbeachtet bleiben.

$$b(d,k) = \sum_{t \in d} \sum_{s \in k} a(t,s) \;\; mit \;\; a(t,s) = \begin{cases} 0 \, wenn \, t \neq s \\ 1 \, wenn \, t = s \end{cases}$$

Formel 15 Anzahl übereinstimmender Terme als Ähnlichkeit

Verbesserung verspricht neben der Beachtung der Häufigkeit $h(t,d)$ eines Terms t in einem Dokument d, die Einbeziehung seines Gewichtes $x(t,k)$ für eine Klasse k in die Bewertungsfunktion. Durch die Gewichte fließen die Vorteile der Bewertung von Indextermen aus Abschnitt 2.4, etwa bei der Summierung aller Termgewichte für ein Dokument, in die Berechnung ein. Dabei wird der Unterschied in der Größe der Vektoren vernachlässigt. Verschiedene Bewertungen sind nicht vergleichbar, wenn sich ihre Dokument- oder Klassenvektoren in der Länge stark unterscheiden.

$$b(d,k) = \sum_{t \in d} \sum_{s \in k} h(t,d) \cdot x(t,k) \cdot a(t,s)$$

Formel 16 Inneres Produkt von Vektoren als Ähnlichkeit

Erst eine Normierung der Bewertung nach Größe der Vektoren, wie es bei der Ermittlung des Winkels zwischen zwei Vektoren im Raum geschieht, sichert die Vergleichbarkeit zweier Ergebnisse. Dafür werden die Vektoren als Angehörige eines hochdimensionalen Vektorraums aufgefaßt, der durch die Menge aller Terme im System gegeben wird. Der Sinus des Winkels zwischen Dokument- und Klassenvektor in diesem Raum bestimmt ihre Ähnlichkeit. Der Nachteil dieser Funktion ist die sehr komplexe Berechnung.

$$b(d,k) = \sqrt{1 - \left[\sum_{t \in d} \sum_{s \in k} \left(\frac{h(t,d) \cdot x(t,k)}{\sqrt{\sum_{t \in d} h(t,d)^2 \cdot \sum_{s \in k} x(s,k)^2}} \cdot a(t,s) \right) \right]^2}$$

Formel 17 Sinus eines Winkels zwischen Vektoren als Ähnlichkeit

3.3 Ermittlung relevanter Bewertungen

Nach Größe der erzielten Bewertungen lassen sich die Klassen in zutreffend und abwegig für ein Dokument unterscheiden. Niedrige Bewertungen sind die Ursache zufälliger Übereinstimmungen zwischen Dokument- und Klassenvektor, und können nicht zur Klassifikation herangezogen werden. Um diese schwachen von den bedeutenden Bewertungen mit Sicherheit abzugrenzen, ist die Einschränkung eines Bewertungsergebnisses unter Annahme einer bestimmten Signifikanz erforderlich.

Grundlage für die Beschränkung der Berwertungen eines Dokuments bildet eine ansteigende Bewertungsfunktion b mit wachsender Ähnlichkeit von Dokument- und Klassenvektor. Nach der Forderung an die Bewertung, ist das Dokument d am aussagekräftigesten für die Klasse k, wenn die Bewertung $b(d,k)$ ihr Maximum annimmt. Allein das Maximum stellt aber noch kein befriedigendes Maß dar, da hierfür eine Beschreibung der Signifikanz fehlt.

$$I'': d \rightarrow K' \text{ und für } k \in K' \text{ gilt } b(d,k) = \max_{k \in K}(b(d,k))$$

Formel 18 Klassifikation nach Maximum der Bewertung

Am einfachsten läßt sich eine Signifikanz über die prozentuale Höhe der erreichten Bewertung gegenüber der Summe aller erteilten Bewertungen für ein Dokument angeben. Das Ausmaß der Bewertung wird von einem Parameter p zwischen 0 und 1 beschrieben. Liegt eine Bewertung des Dokuments für eine Klasse im Bereich von p, so wird ihm diese Klasse zugeschrieben. Der Abstand zu den erzielten Bewertungen der anderen Klassen wird dabei vernachlässigt. Der Vorteil der prozentualen Betrachtung gegenüber einem absoluten Grenzwert ist die Unabhängigkeit von den Dimension der konkreten Ergebnisse, die sich für verschiedene Gewichte und Bewertungen um Zehnerpotenzen unterscheiden können.

$$I'': d \rightarrow K' \; und \;\; für \;\; k \in K' \;\; gilt \;\; b(d,k) \geq p \cdot \sum_{k \in K} b(d,k)$$

Formel 19 Grenzwert als Signifikanzmaß

Der prozentuale Abstand der Klassen untereinander kann durch einen weiteren Parameter q zwischen 0 und 1 beschrieben werden. Hierzu wird das Ergebnis einer Bewertung für eine Klasse herausgegriffen und mit denen für die restlichen Klassen verglichen. Unterschreitet für eine gewählte Bewertung keine andere den Abstand q, wird das Dokument der entsprechenden Klasse zugeordnet.

$$I'': d \rightarrow K' \; und \;\; für \;\; k \in K' \;\; gilt \;\; b(d,k) > (1+q) \cdot b(d,k') \;\; für \; alle \; k' \neq k$$

Formel 20 Abstand zwischen den Bewertungen als Signifikanzmaß

Eine genauere Anpassung des Signifikanzmaßes an die erzielten Bewertungen für alle Klassen ermöglicht ihre Eingrenzung durch ein Konfidenzintervall [Bro91]. Die bestehenden Bewertungen werden dazu als Stichprobe einer normalverteilten Zufallsgröße angesehen, welche der Höhe einer Bewertung für ein Dokument in einer Klasse entspricht. Um eine Normalverteilung der Bewertungen zu erreichen, wird die Normierung der Termfrequenzen in den Klassenvektoren vorausgesetzt. Unter Verwendung des Erwartungswertes a und der Streuung σ der Stichprobe wird jede Bewertung auf ihren Wert für eine Standardnormalverteilung $N(0,1)$ zurückgeführt. Aus der Standardnormalverteilung läßt sich für eine Sicherheitswahrscheinlichkeit a durch Lösung des entsprechenden Wahrscheinlichkeitsintegrals ein Grenzwert r errechnen (siehe Formel 21), der die signifikanten Bewertungen abtrennt. Überschreitet eine konvertierte Bewertung den Grenzwert r, so wird dem Dokument die zugehörige Klasse zugewiesen. Die Wahrscheinlichkeit a beschreibt somit die Signifikanz der Bewertung.

$$I'': d \rightarrow K' \; und \;\; für \;\; k \in K' \;\; gilt \;\; \frac{b(d,k) - e}{\sigma} \geq r \;\; wobei \;\; r = \int_0^a N(0,1)$$

Formel 21 Konfidenzintervall als Signifikanzmaß

Für die Signifikanz zur Einschränkung erzielter Bewertungsergebnisse für ein Dokument gelten die Parameter Recall und Precision wie bei einer Anfrage aus Abschnitt 2.2. Wird die Signifikanz hoch angesetzt, steigt die Genauigkeit der Klassifikation, wobei ihr Umfang sinkt. Ist sie zu niedrig gewählt, nehmen die

Klassifikationen zu, das Ergebnis wird insgesamt aber ungenauer. In der Praxis muß ein Kompromiß bei der Wahl des Parameters eingegangen werden, wobei bei der Klassifikation von Dokumenten die Genauigkeit im Vordergrund steht.

3.4 Spezielle Eigenschaften einer Klassifikation

Bereits in Abschnitt 2.3 wurde auf allgemeine Möglichkeiten zur Optimierung der Textanalyse eingegangen. In den folgenden Abschnitten sollen diese Methoden für ein klassifizierendes Indexierungsystem spezieller vorgestellt werden.

3.4.1 Ein- und Mehrwortterme

Im Abschnitt 2.3.2 wurde bereits der Einsatz einer Entfernungsangabe als mögliche Optimierung von Indexierungssystemen besprochen. Sie soll auch hier zum Einsatz kommen, da die einzelnen Indexterme häufig eine zu geringe Aussagekraft für die Klassifikation besitzen. Ein Nachteil der allgemeinen Methode ist ein hoher Aufwand bei der Auswertung von Satznummer und Satzposition für einen Indexterm, darum soll eine modifizierte Darstellung für die Entfernung verwendet werden. Während der Generierung des Termvektors, werden hintereinanderfolgende Terme zu Mehrworttermen zusammengefaßt. Zudem werden auch Mehrwortterme mit Auslassungen gebildet, um eine Verallgemeinerung ihrer Bedeutung zu erreichen (siehe Tabelle 4). Für diese Termarten läßt sich direkt ein Vergleich von Dokument- und Klassenvektor durchführen, ohne das eine Zwischenauswertung weiterer Angaben nötig wird. Dabei wächst aber die Anzahl der Einträge im Termvektor und somit der benötigte Speicherplatz.

Termart	Spezifik	Anzahl	Beispiel
Einwortterm	niedrig	klein	Besitzer
Mehrwortterm	hoch	groß	Besitzer eines Grundstücks
Mehrwortterm mit Auslassungen	mittel	mittel	Besitzer ... Grundstücks

Tabelle 4 Eigenschaften der Termarten

3.4.2 Statistische und regelbasierte Verfahren

Bei der Art der Indexierung stehen statistische und regelbasierte Verfahren zur Auswahl. Ein statistisches oder auch automatisches Verfahren beruht auf der

Auswertung einer Menge klassifizierter Dokumente. Aus ihr wird eine invertierte Matrix von Gewichten gewonnen, die sich zur Klassifikation systemfremder Dokumente verwenden läßt. Diese Verfahrensweise setzt keine Vorkenntnisse über den Themenbereich der zu klassifizierenden Dokumente voraus. Die Methode liefert aber nur unzureichende Ergebnisse, da die Zuschreibung von Gewichten zu den Termen bei der Bewertung fehlerbehaftet ist. Die Entscheidung, welche Terme in einem unbekannten Dokument besonders auf seine Klasse hinweisen, kann nicht automatisch getroffen werden. Zudem sind die Ergebnisse schlecht nachvollziehbar, weil sie sich aus einer komplexen Matrix ableiten.

Zur Verbesserung der Klassifikation wird deshalb zusätzlich ein regelbasiertes oder auch manuelles Verfahren verwendet. Es basiert auf Termen, die direkt vom Benutzer den Klassen zugeordnet werden. Dieses Vorgehen bietet eine hohe Kontrolle über das Ergebnis einer Klassifikation, da für die Zuweisung einer Klasse verantwortliche Terme leicht identifizierbar sind. Es setzt aber gute Kenntnisse über das Fachvokabular in den Dokumenten voraus.

Kriterien	statistische Verfahren	regelbasierte Verfahren
Grundlage	klassifizierte Dokumente	klassifizierte Terme
Fachkenntnisse	nicht nötig	erforderlich
Kontrolle	schwach	stark
Sicherheit	gering	hoch

Tabelle 5 Vergleich statistischer und regelbasierter Verfahren

Bei der manuellen Zuordnung eines Terms zu einer Klasse, kann die Beziehung durch eine starke oder schwache Korrelation beschrieben werden. Spricht ein Term mit hoher Sicherheit für genau eine Klasse, korreliert er stark mit ihr. So soll etwa die Klasse Beschwerde sofort einem Dokument zugewiesen werden, in dem das Wort `beschweren` auftritt. Dagegen ist ein Term mit schwacher Korrelation lediglich ein Indiz für die Zugehörigkeit zu einer Klasse. Beispielsweise können die Begriffe `Vertragsende` und `beendigen` auf eine Klasse Kündigung hinweisen. Die Intensität der Beziehung zwischen Term und Klasse läßt sich durch einen numerischen Wert ausdrücken. Dies ermöglicht die Gewichtung der benutzerdefinierten Terme.

Die Kombination dieser Verfahren verspricht einen Ausgleich ihrer Vor- und Nachteile. Die Ungenauigkeit des automatischen Verfahrens wird durch die

Spezifik der regelbasierten Methode kompensiert und umgekehrt. Die Kompensation der Nachteile in einem System läßt eine Steigerung von Recall und Precision seiner Ergebnisse erwarten.

3.4.3 Eindeutige Klassifikation

Häufig ist die Klassifikation eines Dokuments eindeutig vorzunehmen. Damit reduziert sich der Wertebereich des Indexierungssystems I''' von einer Menge von Klassen auf höchstens Eine. Für die Reduktion ist ein Verfahren nötig, das aus der Menge relevanter Klassen die für das Dokument zutreffende auswählt. Ist die Abgrenzung einer solchen Klasse nicht möglich, gilt das Dokument als nicht klassifizierbar.

$$I''': d \rightarrow k$$

Formel 22 Abbildung eines eindeutig klassifizierenden Indexierungssystems

Wie in Abschnitt 3.4.2 erläutert, kann ein Dokument aufgrund starker bzw. schwacher Korrelation nach regelbasierten Verfahren oder unter Verwendung von Gewichten nach statistischen Verfahren bewertet werden. Für jede dieser Kategorien der Bewertung lassen sich relevante Klassen für ein Dokument ermitteln. Auf der Suche nach der eindeutigen Klasse eines Dokuments werden aus den relevanten Klassen verschiedener Kategorien Schnittmengen gebildet. Diese Zusammenfassung unterschiedlicher Kategorien erfolgt nach der Reihenfolge, die von den Prioritäten in Tabelle 6 vorgegeben wird. Für die einelementige Schnittmenge mit der höchsten Priorität wird die enthaltene Klasse dem Dokument zugeordnet. Führt der Durchschnitt der Kategorien in allen Fällen zu Mengen mit keinen oder mehreren Elementen, so gilt das Dokument als nicht klassifizierbar, da ihm keine Klasse eindeutig zuzuschreiben ist.

Priorität	Kategorien in der Schnittmenge		
	starke Korrelation	**schwache Korrelation**	**Gewicht**
1	x	x	x
2	x	x	
3	x		x
4	x		
5		x	x
6		x	
7			x

Tabelle 6 Priorität der Schnittmengen aus den Kategorien

Die Priorität eines Durchschnitts entspricht seiner Genauigkeit für die Klassifikation, ähnlich der Rangordnung von Abfragen bei der in Abschnitt 2.3.3 beschriebenen Quorum-Level-Suche. Die Rangverteilung orientiert sich dabei an der Spezifik der Lexika und der Mehrheitsfähigkeit eines Ergebnisses. So führen regelbasierte Verfahren zu spezifischeren Bewertungen und somit zu genaueren Ergebnissen. Für die Priorisierung folgt eine Bevorzugung der Kategorien aus dem Benutzerlexikon. Schnittmengen aus mehreren Kategorien werden höher eingestuft, weil die Genauigkeit eines Ergebnisses mit der Zahl der einbezogenen Lexika steigt.

Abschließend soll die Methode zur eindeutigen Klassifikation noch einmal an einem Beispiel verdeutlicht werden. Aus gegebenen relevanten Klassen für die verschiedenen Kategorien werden Schnittmengen in der Reihenfolge ihrer Priorität gebildet (siehe Tabelle 7). Erst mit Priorität 3 läßt sich so eine eindeutige Klasse ableiten.

Priorität		**1**	**2**	**3**
Einbezogene Kategorien	**starke Korrelation**	{1,3,7}	{1,3,7}	{1,3,7}
	schwache Korrelation	{1,3,4}	{1,3,4}	
	Gewicht	{5,7}		{5,7}
Schnittmenge		∅	{1,3}	{7}
Ergebnis		kein Klasse	mehrere Klassen	eindeutige Klasse

Tabelle 7 Beispiel für die Bestimmung einer eindeutigen Klasse

3.5 Ablauf einer Klassifikation

Auch beim Umgang mit Indexierungssystemen zur Klassifikation bleibt die Teilung der Aufgabenbereiche in Aufbau des Systems und Verarbeitung einer Anfrage bestehen. Die einzelnen Arbeitsschritte variieren aber gegenüber der allgemeinen Indexierung. Deshalb soll ein entsprechender Ablauf hier noch einmal vollständig dargestellt werden.

Zur Erzeugung des Systems wird wie folgt verfahren:

1. Identifizieren jedes Wortes in einer Menge von Dokumenten.
2. Entfernen aller Stopwörter, die keine Bedeutung für die Unterscheidung der Dokumente besitzen.
3. Bilden des Wortstamms für jeden Indexterm (Stemming).
4. Bilden von Mehrworttermen aus den Wortstämmen.
5. Zusammenfassen aller Terme von Dokumenten einer Klasse zu einem Klassenvektor.
6. Errechnen der Gewichte für jeden Term in allen Klassen.
7. Repräsentieren jeder Klasse durch die Menge ihrer Terme und der zugehörigen Gewichte.

Für die Bearbeitung einer Abfrage, also die Klassifikation eines Dokuments, sind folgende Schritte nötig:

1. Erzeugen des Termvektors für das zu klassifizierende Dokument.
2. Bewerten des Dokuments durch den Vergleich von Dokument- und Klassenvektor.
3. Bestimmen der relevanten Klassen für die Kategorien der Bewertung.
4. Ermitteln einer eindeutigen Klasse aus den relevanten Klassen.
5. Ausgeben des Ergebnisse der Klassifikation.

4 Aufbau eines Klassifikationssystems

Nachdem sich die vorherigen Abschnitte mit den theoretischen Voraussetzungen der Indexierung und Klassifikation von Dokumenten auseinandersetzten, soll jetzt ein konkretes Klassifikationssystem beschrieben werden. Es stellt einen Prototyp zur Klassifizierung von Dokumenten am Beispiel von Geschäftspost vor.

4.1 Gewinnung des Datenmaterials

Die Geschäftspost eines Unternehmens besteht im Regelfall aus Briefen, die vom Kunden zugesandt werden. Diese schriftlich vorliegenden Dokumente müssen zunächst konvertiert werden, um sie der Analyse zugänglich zu machen. Eine OCR-Software erzeugt aus ihnen maschinenlesbare Dokumente im ASCII-Format. In dieser Form können die Kundenbriefe dann klassifiziert und dem zuständigen Sachbearbeiter zugeleitet werden. Da eine Zuordnung aller Dokumente nicht erwartet werden kann, ist weiterhin eine manuelle Bearbeitung vorzusehen.

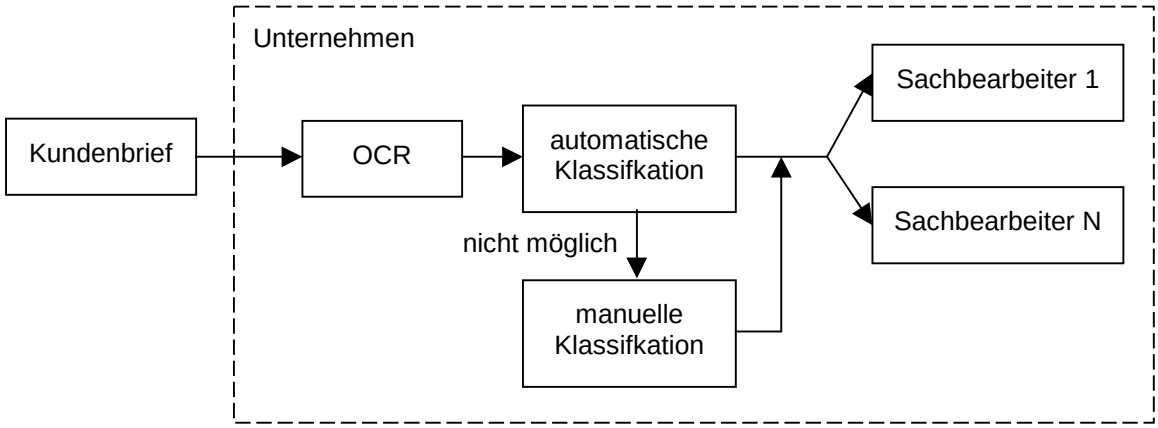


Abbildung 8 Weg eines Kundenbriefes

Im Idealfall entsprechen die zu klassifizierenden Dokumente der deutschen Rechtschreibung und weisen keine Abnormitäten im Schriftbild wie z.B. Sperrschrift auf. Es ist aber zu erwarten, daß die gewonnenen Daten mit Fehlern behaftet sind. Die Ursachen liegen beispielsweise in Tippfehlern, Konvertierungsverlusten oder Problemen bei der Texterkennung. Diese Einflüsse senken die Aussicht auf eine erfolgreiche Klassifikation und sollten deshalb möglichst gering gehalten werden.

4.2 Vom Dokument zum Dokumentvektor

Das Ziel des im folgenden beschriebenen Prozesses ist die Verdichtung der in einem Dokument enthaltenen Informationen, um seine schnelle aber zuverlässige Klassifikation zu ermöglichen. Im Mittelpunkt stehen hierbei die aus den Wörtern gewonnenen Indexterme, die als Träger der Informationen dem Inhalt eines Dokuments eine Klasse zuordnen. Sie werden aus dem Text extrahiert, analysiert und zum Aufbau einer Repräsentation des Dokuments verwendet. Ihr Aufbau erfolgt durch die sequentielle Verarbeitung eines Dokuments durch den Parser, den Typisierer, die morphologische Analyse und den Extrahierer zu einem Dokumentvektor [Len86].

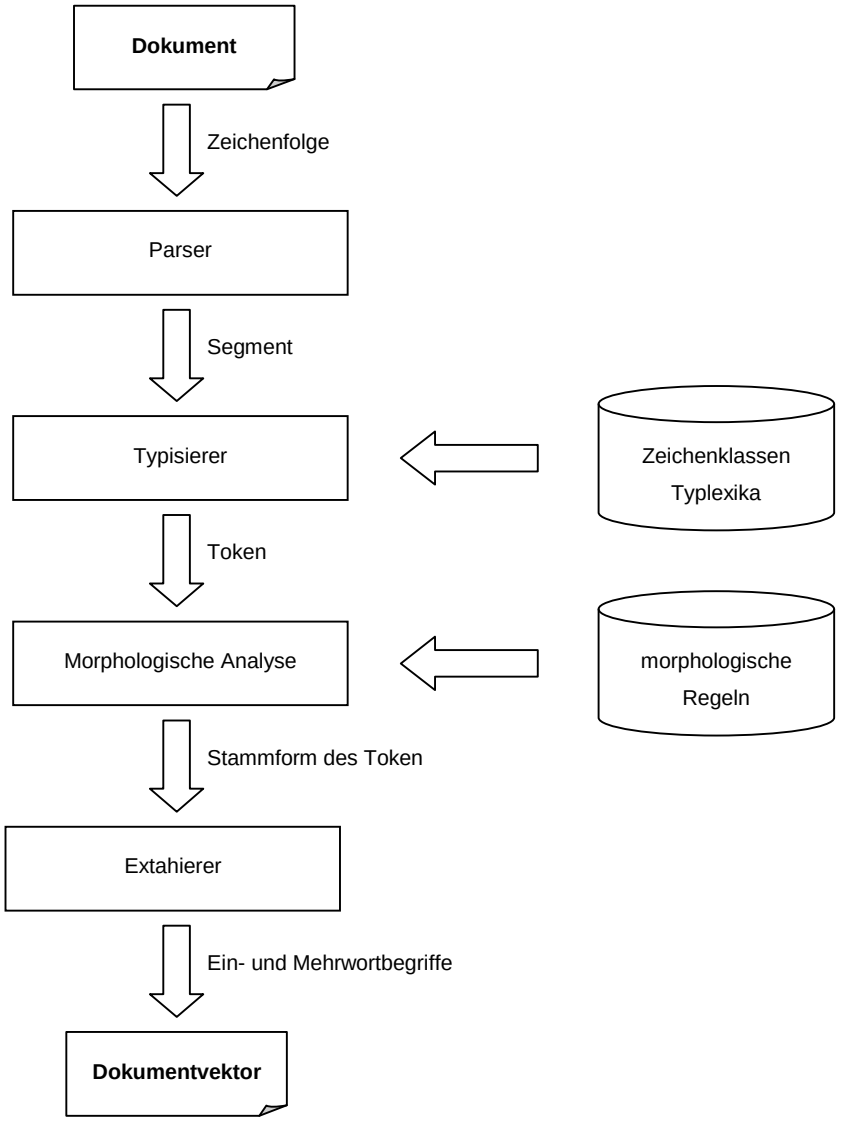


Abbildung 9 Analyse eines Dokuments

Das aufgearbeitete Datenmaterial wird dafür zunächst durch den Parser (siehe 4.2.1) in Segmente zerlegt. Diese werden durch den Typisierer (siehe 4.2.2) in typisierte Wörter gegliedert, die auch als Token bezeichnet werden. Das Token wird einer morphologischen Analyse (siehe 4.2.3) unterzogen, um die Flektionsform des enthaltenen Wortes auf seinen Wortstamm abzubilden. Anschließend werden durch den Extrahierer (siehe 4.2.3) unter Verwendung der Typinformationen aus den Stammformen Ein- und Mehrwortterme gebildet. Die Gesamtheit aller erzeugten Terme eines Dokuments wird zu seinem Dokumentvektor zusammengefaßt. Er stellt die interne Repräsentation eines Dokuments dar und bildet die Grundlage für die weiteren Verarbeitungsschritte in Lexikonaufbau und Bewertungsprozeß.

Der Stand der Verarbeitung eines Textes soll anhand eines Beispielsatzes veranschaulicht werden. Er wird nach Bearbeitung der verschiedenen Module wiedergegeben, um den momentanen Zustand der Daten aufzuzeigen.

```
Ein Beispielsatz für die Schritte, die bei der Dokument-
analyse durchlaufen werden.
```

Abbildung 10 Beispielsatz für die Verarbeitung eines Dokuments

4.2.1 Zerlegung in Segmente

Ein Dokument wird als endliche Folge von Zeichen aufgefaßt. Der **Parser** hat die Aufgabe, diese Folge nach dem Auftreten eines Leerzeichens, Tabulatorzeichens oder Zeilenumbruchs zu segmentieren. Das Ergebnis des Parsens sind Zeichenfolgen oder auch Segmente, die ein oder mehrere Token enthalten. So besteht beispielsweise das Segment `war.` aus einem KLEINWORT und einem SATZENDE. Die Zerlegung durch den Parser birgt keine Gefahr des Informationsverlustes, da im Deutschen keines der Trennzeichen Bestandteil eines Wortes ist.

Bei der Erkennung eines Tabulators oder Zeilenumbruchs als Trennzeichen besteht die Möglichkeit, daß das folgende Segment zu Beginn eines Satzes steht, obwohl ihm kein Token vom Typ SATZENDE vorausging. Ein solcher Zustand wird als Pseudosatzanfang bezeichnet und kann vom Parser abgefragt werden, um die Groß- und Kleinschreibung von Token sicherzustellen. Signalisiert der Parser einen Pseudosatzanfang, gilt die Schreibweise des folgenden Wortes als

unbekannt. Dies ist von Bedeutung, wenn nach einer Überschrift, die für gewöhnlich ohne Satzzeichen endet, ein neuer Satz mit einem Artikel oder einem Verb beginnt.

Als zusätzliche Information führt der Parser die aktuelle Zeilennummer mit, die beim Auftreten eines Zeilenumbruchs aktualisiert wird. Sie kann unterstützend bei der Eingrenzung eines Textabschnittes eingesetzt werden, findet bisher aber noch keine Verwendung.

```
Ein                         bei
Beispielsatz                der
für                         Dokument-
die                         analyse
Schritte,                   durchlaufen
die                         werden.
```

Abbildung 11 Beispielsatz nach Verarbeitung durch den Parser

Nachdem die benötigten Informationen aus den Seperatorzeichen gezogen wurden, gehen diese nicht weiter in die Verarbeitung ein. Die Segmente werden zur weiteren Analyse an den Typisierer übergeben. Enthält ein Segment mehrere Token, wird es von ihm wiederum zerlegt. Tritt beispielsweise ein Komma nach einem Wort auf, ist es durch den Typisierer abzutrennen. Das Wort wird behandelt und das Komma muß zwischengespeichert werden. Zu diesem Zweck bietet der Parser die Möglichkeit, eine beliebige Anzahl von Zeichen zurückzulegen. Diese werden ähnlich einem Kellerautomaten bei der nächsten Abfrage eines Segments zuerst berücksichtigt [App92].

4.2.2 Tokenbildung und Typzuweisung

Die Aufgabe des **Typisierers** ist die Gewinnung von Token aus den Segmenten und ihre Zuordnung zu einem Typ. Hierfür wird das Segment gegebenenfalls zerlegt und den enthaltenen Token ein Standard- oder Lexikontyp zugewiesen. Findet sich keine Entsprechung für einen Typ, gilt das Token als undefiniert.

Die vom Parser zur Verfügung gestellten Segmente lassen sich nicht sofort in einen Typ einordnen. Sie müssen zunächst vom Typisierer nach der Grammatik in Abbildung 12 zerlegt werden, um die enthaltenen Token zu separieren. Daraus folgt zunächst die Entfernung ignorierter Zeichen vom vorderen und hinteren Teil des Segments. Anschließend werden Komma- oder Satzendezeichen und gegebenenfalls nochmals ignorierte Zeichen vom Ende des Segments abgetrennt.

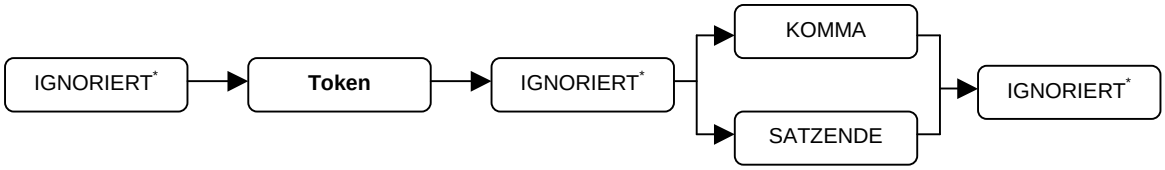


Abbildung 12 Allgemeiner Aufbau eines Segments

Im letzten Schritt der Bereinigung sollen Bindestriche aus den Token eliminiert werden. Dies dient nicht dem Zusammenführen getrennter Wörter, sondern der Verallgemeinerung verschiedener Schreibweisen eines Wortes, beispielsweise dem Abbilden von `Schadens-Klasse` auf `Schadensklasse`. Ein Wort wird am Bindestrich zerlegt und die Typen der Teilworte bestimmt. Sind beide Teile vom Typ KLEIN- oder GROSSWORT, werden sie zu einem Wort verbunden. Die Einschränkung auf diese zulässigen Typen verhindert unerwünschte Zusammenfassungen, so soll z.B. das Wort `Versicherungs-AG` nicht in `Versicherungsag` umgewandelt werden.

Schritt	Beschreibung	Beispiel
0	Entfernung der ignorierter Zeichen	„Hallo!“
1	Abtrennen des Satzendezeichens	Hallo!
2	typisierbares Token	Hallo

Abbildung 13 Beispiel für die Zerlegung eines Segments

Die Grundlage der Typisierung nach Standardtypen (siehe Tabelle 8) bilden Zeichenguppen, die aus einer Initialisierungsdatei geladen werden (siehe Abschnitt VI.i). Diese Datei ermöglicht eine einfache Anpassung der Typzuschreibung, über die Manipulation der dem Typ zugehörigen Zeichengruppe. So kann je nach Art der zu verarbeitenden Texte etwa das Auftreten eines Paragraphen-Zeichens ignoriert werden und somit für die weitere Analyse ohne Bedeutung sein, oder als Beginn einer Abkürzung gewertet werden und so in die Terme einfließen.

Für Token aus einem und mehreren Zeichen werden verschiedene Standardtypen vergeben. Danach unterscheiden sich auch die Einträge in der Initialisierungsdatei. Zu jedem Einzeichentyp ist jeweils genau eine Zeichengruppe definiert. Gehört ein einzeln im Text auftretendes Zeichen der Gruppe an, bekommt es den entsprechenden Typ zugeordnet. Für einen Mehrzeichentyp sind je zwei Zeichengruppen vorgesehen, eine für das Startzeichen und eine weitere

für die Folgezeichen. Einer Zeichenkette wird ein Mehrzeichentyp zugewiesen, wenn ihr erstes Zeichen in den Startzeichen und jedes weitere Zeichen in den Folgezeichen diese Typs enthalten ist.

Typ	Beispiel	Bemerkung
KLEINBUCHSTABE	a...zäöüß	
GROSSBUCHSTABE	A...ZÄÖÜ	
ZIFFER	0...9	
SATZENDE	.!?	Indiz für Satzende
KOMMA	,;:	separiert Satzteile
IGNORIERT	()<>'"=\/%~*	verwerfen bei der Verarbeitung
KLEINWORT	sein	Folge von Kleinbuchstaben
GROSSWORT	Verlag	erstes Zeichen Groß- sonst Kleinbuchstaben
WORT	Das	Wort mit unklarer Groß- oder Kleinschreibung
NUMMER	1,34	Ziffernfolge mit Trennzeichen
ABKUERZUNG	StGB	Groß- und Kleinbuchstaben und Sonderzeichen

Tabelle 8 Standardtypen des Typisierers

Die Vorlage für das Zuordnen von Lexikontypen bilden Lexikondateien, die ebenfalls während der Initialisierung des Typisierers eingeladen werden. Eine solche Datei mit der Extension .lex beinhaltet Wörter eines Typs, die durch Zeilenumbrüche getrennt sind. Ein Wort wird einem Lexikontyp zugewiesen, wenn es seiner Lexikondatei angehört. Durch das Ändern der Lexika ist die Zuweisung der Typen beeinflußbar. So können der existierenden Lexikondatei Stopwort.lex, die sehr häufige Wörter der Deutschen Sprache enthält, neue Wörter hinzugefügt werden. Zudem ist die Anzahl der Lexika nicht festgelegt, wodurch neu erstellte Dateien berücksichtigt werden können.

Die Typisierung der Token ist anhand der bisherigen Definitionen der Typen nicht eindeutig möglich. So kann z.B. das Wort `ist` sowohl als STOPWORT oder KLEINWORT betrachtet werden. Um die Mehrdeutigkeiten aufzulösen, erhalten die Typen eine Priorität. Die Einordnung erfolgt nach dieser Reihenfolge, bis ein Typ erkannt ist. Höchste Priorität besitzen die Lexikontypen, d.h. es werden zuerst die Lexikondateien nach dem Token durchsucht. Bleibt die Suche ohne Ergebnis,

sind die Standardtypen nach der Reihenfolge in Abbildung 14 zu prüfen. Kann auch hier kein Typ zugewiesen werden, gilt das Token als undefiniert.

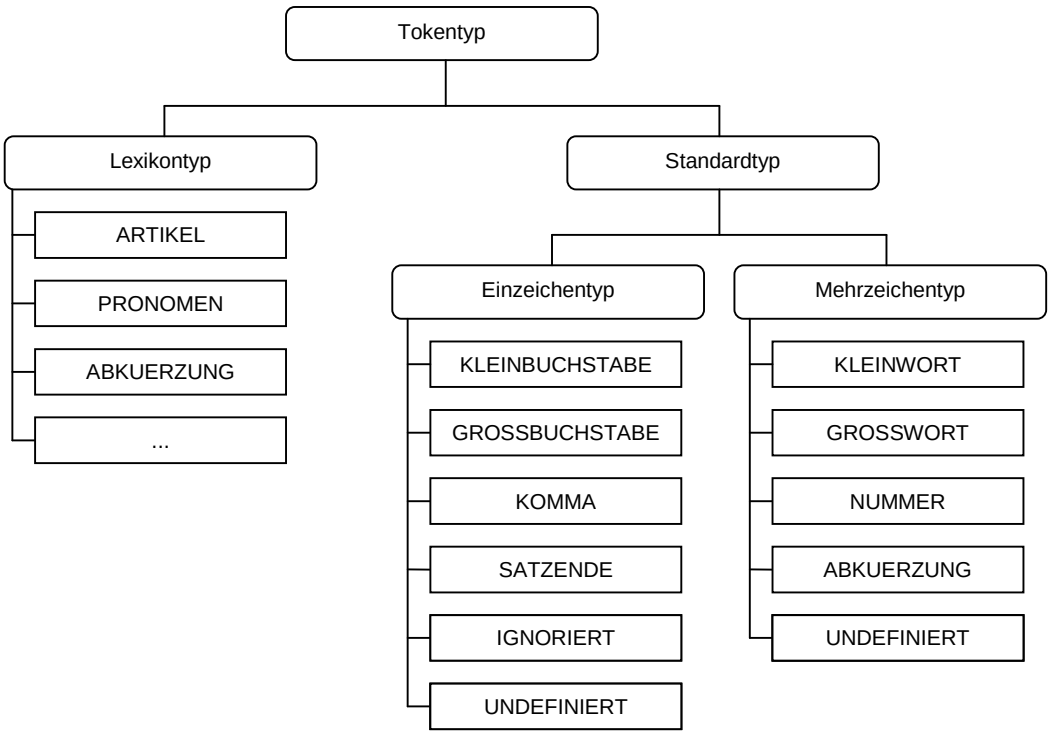


Abbildung 14 Prioritäten der Typen

Die Standardtypen haben die Aufgabe die Indexterme nach brauchbar wie GROSSWORT oder KLEINWORT und unbrauchbar wie IGNORIERT oder UNDEFINIERT für die Klassifikation zu unterscheiden. Durch die Lexikontypen ist eine weitere Einschränkung der wichtigen Terme für die Indexierung möglich, da sie Stopwörter eliminieren. Die erzeugten Angaben bieten aber auch Ansatzpunkte zur tieferen Analyse des Dokumentinhalts. So kann die nähere Betrachtung der Token vom Typ NUMMER zur Gewinnung von Kundennummern führen.

```
Ein             [ARTIKEL]      bei             [PRAEPOSI]
Beispielsatz    [ABKUERZUNG]   der             [ARTIKEL]
für             [PRAEPOSI]     Dokument-       [ABKUERZUNG]
die             [ARTIKEL]      analyse         [KLEINWORT]
Schritte        [GROSSWORT]    durchlaufen     [KLEINWORT]
,               [KOMMA]        werden          [STOPWORT]
die             [ARTIKEL]      .               [SATZENDE]
```

Abbildung 15 Beispielsatz nach Verarbeitung durch den Typisierer

4.2.3 Bildung von Ein- und Mehrworttermen

Bereits in Abschnitt 3.4 wurde auf die Vorteile des Einsatzes von Mehrworttermen eingegangen. Der **Extrahierer** bildet Ein- und Mehrwortterme in den Satzgrenzen, indem er die Folge typisierter Token in Sätze segmentiert. Eine Generierung dieser Terme über einen Satz hinaus ist nicht sinnvoll, da er als abgeschlossenes grammatikalisches Konstrukt alle in Beziehung stehenden Wörter enthält. Die Informationen über den Typ werden verwendet, um die Aussagekraft der gebildeten Terme zu erhöhen und die entstehende Datenmenge einzuschränken, indem undefinierte Token nicht einbezogen werden.

Vor der Erzeugung der Terme sollen aus der Tokenfolge getrennte Wörter eliminiert werden. Tritt ein Token mit abschließendem Bindestrich auf und ist das folgende Token vom Typ KLEINWORT, so wird die Trennung aufgehoben. Die Einschränkung des Typs verhindert falsche Zusammenführungen, wie der Typ STOPWORT des Tokens `und` in der Folge `Vor- und Nachbearbeitung.` Über die Lexikontypen wird so die häufigste Fehlerquelle bei der Aufhebung von Trennungen eliminiert.

Im letzten Schritt der Vorbereitung werden die Token einer morphologischen Analyse unterzogen, um die verschiedenen Flektionsformen eines Wortes zu vereinheitlichen [Her98]. Wie im Abschnitt 2.3.1 beschrieben, dient das Stemming der sinnerhaltenden Verdichtung der Indexterme.

Als Indiz für die Segmentierung eines Satzes wird ein Satzendezeichen angesehen. Um die Sicherheit bei der Erkennung der Satzgrenze zu erhöhen, wird das vorangehenden Token näher untersucht. Ist es vom Typ GROSSBUCHSTABE, KLEINBUCHSTABE, ZIFFER, NUMMER oder LEXABKUERZUNG[1], so wird das Satzendezeichen entkräftet. Dadurch sinkt die Gefahr falsch angenommener Satzgrenzen, wie sie z.B. bei den Wortgruppen `H. Müller` oder `50. Geburtstag` besteht.

Im Prozeß der Termbildung werden ignorierte und undefinierte Token verworfen, alle anderen nach bedeutungstragend, wenn sie vom Typ KLEINWORT,

[1] Dieser spezielle Lexikontyp, der den Worten der Lexikondatei Abkuerzung.lex zugewiesen wird, wird von einem Punkt abgeschlossen, wie beispielsweise ggf. .

GROSSWORT, WORT oder ABKUERZUNG sind, und ansonsten in bedeutungslos unterschieden. Die Einschränkung auf diese Typen filtert Token mit schwacher Aussagekraft für die Indexierung heraus und verringert somit den Aufwand für die Klassifikation. Insbesondere werden hier die Wörter der Lexika vernachlässigt. In den Dokumentvekor gehen bedeutungstragende Token, auf ihre Stammform reduziert, als Einwortterme ein.

Für den Aufbau der Mehrwortterme werden von jedem Token des Satzes aus schrittweise alle folgenden Token angefügt. Hier fließen auch bedeutungslose Token ein, da sie zur Spezifizierung der Terme beitragen. Ein Mehrwortbegriff wird aber nur in den Dokumentvektor aufgenommen, wenn er wenigstens ein bedeutungstragendes Token enthält. Der Aufbau wird unterbrochen, wenn das Ende des Satzes, eines Teilsatzes oder die definierte maximale Länge für Mehrwortbegriffe erreicht ist. Das Ende eines Teilsatzes wird anhand eines Tokens vom Typ KOMMA festgestellt. Für die maximale Länge eines Mehrwortbegriffes hat sich ein Wert von 5 als sinnvoll erwiesen. Längere Begriffe gehen kaum noch in die Bewertung ein, da sie zu spezifisch sind.

```
ein Beispielsatz
ein Beispielsatz für
ein Beispielsatz für die
ein Beispielsatz für die
  Schrit
ein ~ Schrit
Beispielsatz
Beispielsatz für
Beispielsatz für die
Beispielsatz ~ die
Beispielsatz für die Schrit
Beispielsatz ~ Schrit
für die Schrit
für ~ Schrit
die Schrit
Schrit
die bei der Dokumentanalys
die ~ Dokumentanalys
die bei der Dokumentanalys
  durchlauf
die ~ durchlauf
bei der Dokumentanalys
bei ~ Dokumentanalys
bei der Dokumentanalys
  durchlauf
bei ~ durchlauf
bei der Dokumentanalys
  durchlauf werden
der Dokumentanalys
der Dokumentanalys durchlauf
der ~ durchlauf
der Dokumentanalys durchlauf
  werden
Dokumentanalys
Dokumentanalys durchlauf
Dokumentanalys durchlauf
  werden
Dokumentanalys ~ werden
durchlauf
durchlauf werden
```

Abbildung 16 Beispielsatz nach Verarbeitung durch den Extrahierer

Um eine allgemeinere Darstellung der Mehrwortterme zu erhalten, werden während ihrer Erstellung Terme mit variabler Länge gebildet. Hierfür wird das

erste und letzte Token des aktuellen Mehrwortterms durch einen Platzhalter verbunden, der als Tilde dargestellt wird. Ist eines der beiden Token bedeutungstragend wird der Term in den Dokumentvekor übernommen. Auf diesem Weg werden die Tokenfolgen `beantworte die Frage` und `beantworte eine Frage` zusammengefaßt zu `beantworte ~ Frage`, ihre gleiche Bedeutung wird durch den selben Term repräsentiert.

4.3 Das Lexikon

In den folgenden Abschnitten soll das Problem der Speicherung von Beziehungen zwischen Indextermen und Dokumenten bzw. Klassen im Mittelpunkt stehen. Traditionell wurde der invertierte Index in Dateisystemen untergebracht [Fra92]. Die Verwaltung der Dateien war Aufgabe des Indexierungsystems, das allein für ihre Pflege verantwortlich war. Das bedeutete neben einem hohen Programmieraufwand für die Erstellung des Systems einen schlechten Zugang zu den Daten für andere Anwendungen.

Um diesen Nachteilen aus dem Weg zu gehen, werden die gewonnen Informationen zur Klassifikation in einer relationalen Datenbank untergebracht. Sie ist im Format von Mircrosoft Access 97 implementiert und kann aus den Anwendungen der Programmiersprache Mircrosoft Visual C++ über die Schnittstelle DAO[2] 3.5 angesprochen werden. Die Struktur der Datenbank und die Aufgaben bei ihrer Verwendung werden in den folgenden Abschnitten erläutert.

4.3.1 Struktur des Lexikons

Der relationale Aufbau des Lexikons verlangt die Speicherung von Informationen in Tabellen. Der invertierte Index muß aufgespalten werden, um die Relationen zwischen Indextermen und Klassen abzulegen. Die Klassen werden in einer Tabelle geführt, die in Beziehung zum automatischen und benutzerdefinierten Lexikon steht. Die Ursache für die Ausklammerung der Klassen liegt in ihren übergreifenden Gültigkeiten für beide Teillexika. Das automatische Lexikon trägt die Informationen, die zur statistischen Bewertung der Dokumente erforderlich

[2] Data Access Objects sind Objekte der Programmiersprache C++, die den Zugriff auf Datenbanken unterstützen.

sind, insbesondere die aus den Dokumenten extrahierten Indexterme und deren Gewichte. Das Benutzerlexikon hingegen beinhaltet die vom Benutzer erstellten Terme und ihre Korrelationen zu den Klassen, also Angaben die zur regelbasierten Einordnung der Dokumente eingesetzt werden.

Während des Zugriffs auf die Datenbank, werden ihre Integritätsbedingungen automatisch überwacht. Das betrifft einfache Forderungen, wie z.B. nach Eindeutigkeit der Menge von Indextermen. Es werden aber auch komplexere Operationen unterstützt, wie die Weitergabe einer Löschanweisung. So verursacht die 1:n Beziehung der Tabelle KLASSE zur Tabelle HAEUFIGKEIT beim Löschen einer Klasse, die Entfernung aller auf diese Klasse verweisenden Häufigkeiten aus der Datenbank.

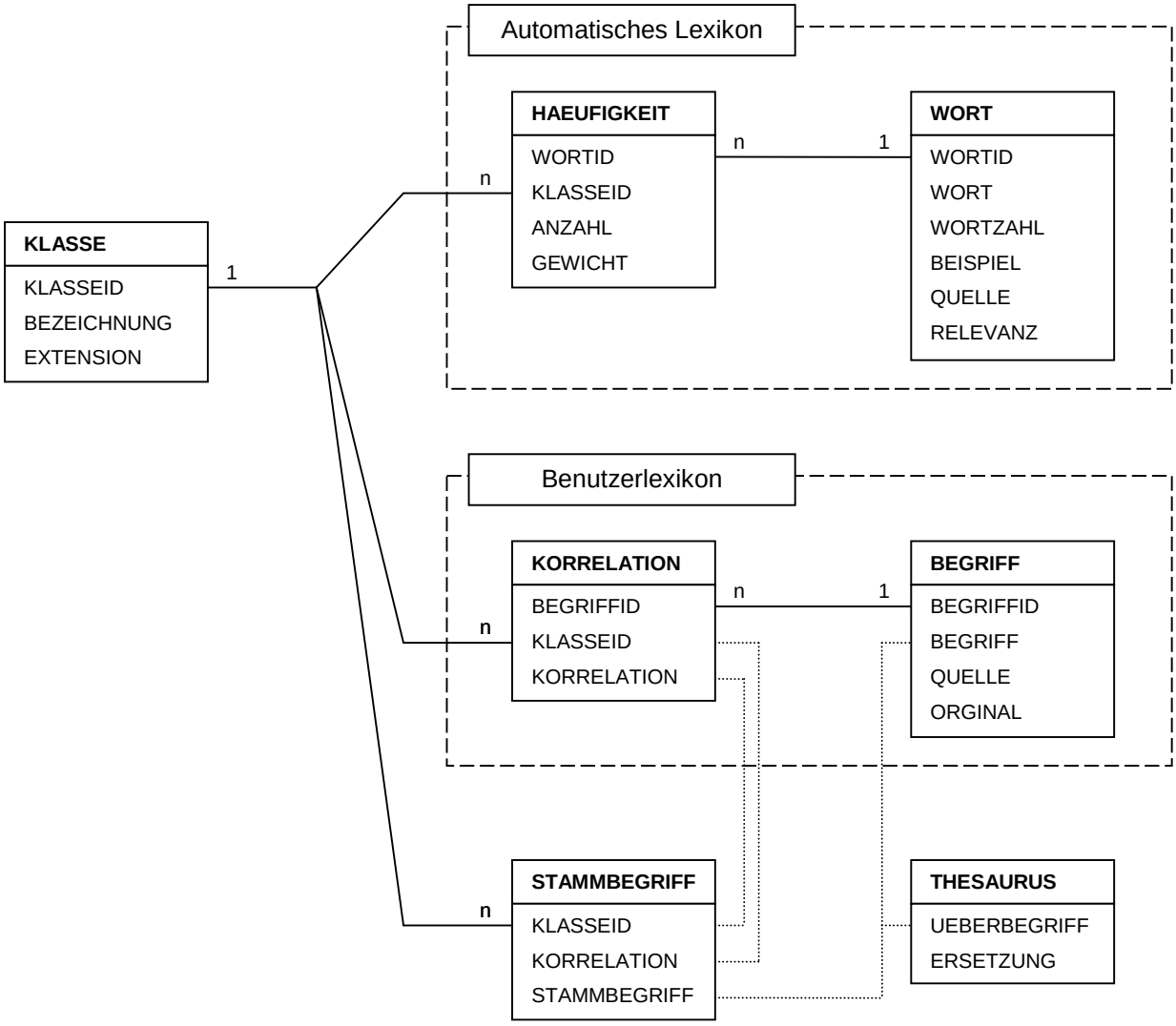


Abbildung 17 Aufbau der Datenbank für das Lexikon

Nachdem die allgemeine Struktur des Lexikons erläutert wurde, soll in den nächsten Abschnitten auf die einzelnen Tabellen näher eingegangen werden. Eine zentrale Stellung in der Struktur der Datenbank nimmt die Tabelle KLASSE ein. Sie enthält für jede Klasse neben einer eindeutigen Identifikationsnummer, über die sie in den Teillexika referenziert wird, eine Bezeichnung und eine Extension. Die Extension findet bei der Zuordnung klassifizierter Dokumente zu den Einträgen der Tabelle KLASSE Verwendung, wie in Abschnitt 4.4.1 beschrieben wird. Diese Eintrage können manuell hinzugefügt oder während des Lexikonaufbaus automatisch erzeugt werden.

Feld	Typ	Beschreibung
KLASSEID	LONGINT	eindeutige Nummer der Klasse
BEZEICHUNG	CHAR[255]	Bezeichnung der Klasse
EXTENSION	CHAR[3]	Dateiendung, die einem Dokument eine Klasse zugeordnet

Tabelle 9 Aufbau der Tabelle KLASSE

Das automatische Lexikon wird durch die in Relation stehenden Tabellen WORT und HAEUFIGKEIT repräsentiert. Die Indexterme, welche in die statistischen Verfahren eingehen, werden zusammen mit einigen zusätzlichen Informationen in der Tabelle WORT abgelegt. So ist ein Beispielsatz vorgesehen, um den Kontext eines indentifizierten Terms für den Benutzer deutlich zu machen. In der Tabelle HAEUFIGKEIT wird die Anzahl mitgeführt, mit der ein Term in einer Klasse auftrat. Nachdem das Einlesen aller Dokumente beendet ist, werden aus den Häufigkeiten der Terme ihre Gewichte für jede Klasse berechnet. Dafür kommen die Verfahren aus Abschnitt 2.4 zur Anwendung. Wenig aussagekräftige Indextterme können daraufhin über das Feld RELEVANZ der Tabelle WORT aus der Bewertung ausgeschlossen werden. So ist beispielsweise ein Term mit gleichen Gewichten in allen Klassen nicht von Bedeutung für ihre Differenzierung.

Feld	Typ	Beschreibung
WORTID	LONGINT	eindeutige Nummer des Terms
WORT	CHAR[255]	Zeichenfolge, die den Term repräsentiert
WORTZAHL	LONGINT	Anzahl der Teilterme (0 steht für variable Länge)
BEISPIEL	MEMO	Beispielsatz für den Term
QUELLE	MEMO	Dokument, aus dem der Term stammt
RELEVANZ	BOOLEAN	Ausschluß von Einträgen aus den Klassenvektoren

Tabelle 10 Aufbau der Tabelle WORT

Feld	Typ	Beschreibung
BEGRIFFID	LONGINT	eindeutige Nummer des Terms
KLASSEID	LONGINT	eindeutige Nummer der Klasse
ANZAHL	LONGINT	Anzahl des Terms in Dokumenten der Klasse
GEWICHT	FLOAT	Gewicht des Terms für die Klasse

Tabelle 11 Aufbau der Tabelle HAEUFIGKEIT

Neben der automatischen Erkennung von Beziehungen zwischen Termen und Klassen, kann ein Term auch direkt einer Klasse zugeordnet werden. Diese zur regelbasierten Klassifikation der Dokumente eingesetzten Terme werden in der Tabelle BEGRIFF des Benutzerlexikons abgelegt. Über die Tabelle KORRELATION ist ihnen eine in Abschnitt 3.4 erläuterte Korrelation zugewiesen. Für starke Korrelationen ist der Wert Null des Feldes KORRELATION reserviert. Schwache positive und negative Korrelationen werden durch Werte ungleich Null mit entsprechenden Vorzeichen beschrieben.

Feld	Typ	Beschreibung
BEGRIFFID	LONGINT	eindeutige Nummer des Terms
BEGRIFF	CHAR[255]	Zeichenfolge, die den Term repräsentiert
QUELLE	MEMO	Ersteller des Begriffs
ORGINAL	CHAR[255]	Begriff ohne Stammformreduktion

Tabelle 12 Aufbau der Tabelle BEGRIFF

Feld	Typ	Beschreibung
WORTID	LONGINT	eindeutige Nummer des Wortes
KLASSEID	LONGINT	eindeutige Nummer der Klasse
KORRELATION	LONGINT	Korrelation eines Terms zu einer Klasse =0 stark >0 schwach positiv <0 schwach negativ

Tabelle 13 Aufbau der Tabelle KORRELATION

Für die Wartung des Benutzerlexikons steht ein Werkzeug zur Verfügung [Her98], daß im Abschnitt 4.4.2 näher erläutert wird. Es basiert auf den Tabellen THESAURUS und STAMMBEGRIFF. In einer Vorstufe erlaubt es dem Nutzer die Definition von Gruppen synonymer Wörter, die in der Tabelle THESAURUS abgelegt werden. Ein Überbegriff erlaubt die Einbindung der Angehörigen einer Wortgruppe bei der Definition von Stammbegriffen. Sie werden gemeinsam mit ihrer Korrelation in der Tabelle STAMMBEGRIFF geführt. So lassen sich beispielsweise die Begriffe `am Morgen` und `am Abend` zusammenfassen unter dem Stammbegriff `am [Tageszeit]` mit dem Eintrag `[Tageszeit] =`

`Morgen, Abend` des Thesaurus. Bei der Generierung des Benutzerlexikons werden die Stammbegriffe unter Einsatz des Thesaurus aufgelöst, ähnlich wie es in Abschnitt 2.3.3 für Anfragen gezeigt wurde. Die erzeugten Indexterme werden morphologisch analysiert und gemeinsam mit ihren Korrelationen in die Tabellen BEGRIFF und KORRELATION übernommen.

Feld	Typ	Beschreibung
UEBERBEGRIFF	CHAR[255]	Beschreibung der Wortgruppe
ERSETZUNG	CHAR[255]	Angehörige der Wortgruppe

Tabelle 14 Aufbau der Tabelle THESAURUS

Feld	Typ	Beschreibung
KLASSEID	LONGINT	Eindeutige Nummer des Begriffs
KORRELATION	LONGINT	Analog Tabelle 13
STAMMBEGRIFF	CHAR[255]	Zeichenfolge, die Stammbegiff repräsentiert

Tabelle 15 Aufbau der Tabelle STAMMBEGRIFF

4.3.2 Funktionen zur Arbeit mit dem Lexikon

Zur Arbeit mit dem Lexikon werden von den Anwendungen Funktionen benötigt, welche die Datenbank entsprechend ihrer Struktur mit Werten füllen, die Auswertungen ihres Datenbestandes unterstützen und aus ihr die Klassenvektoren ermitteln. Diesen Funktionen liegen SQL[3]-Konstrukte zugrunde, die als eingebettetes SQL der Anwendung in der Datenbank schreiben, lesen, aktualisieren oder löschen können [Gei95]. Aus den Konstrukten erzeugt die Anwendung durch Modifikation von Parametern ein SQL-Statment, das an die Datenbank gesandt wird. So kann etwa das Konstukt SELECT * FROM WORT WHERE WORTID=%ul durch Einsetzen des Wertes 101 für den Parameter %ul in ein SQL-Statment überführt werden, um damit die Angaben über den entsprechenden Indexterm abzufragen. Das Ergebnis der Ausführung eines SQL-Statments wird an das Programm zurückgeliefert, welches die erhaltenen Daten gegebenenfalls weiterverarbeitet. Die vorgesehenen Funktionen für den Datenaustausch werden in den nächsten Abschnitten erläutert.

[3] Structured Query Language: Sprache zur Suche und Manipulation von Daten in Datenbanksystemen

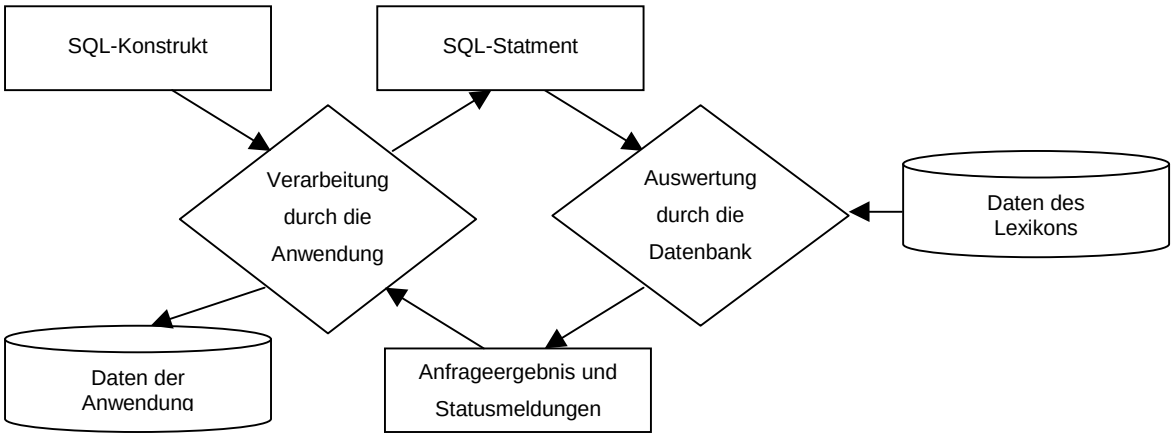


Abbildung 18 Prinzip des eingebetteten SQLs

4.3.2.1 Aufnehmen von Einträgen in das Lexikon

Beim Füllen des automatischen Lexikons sind klassifizierte Dokumentvektoren in die Datenbank zu übernehmen. Dazu steht eine Funktion zur Verfügung, welcher die Klasse und der erweiterte Termvektor eines einzulesenden Dokuments übergeben wird. Die Erweiterungen des Dokumentvektors betreffen neben der Anzahl eines Indexterms im Dokument die zusätzlichen Informationen, die zum Anlegen eines neuen Eintrags der Tabelle WORT benötigt werden. Um den Dokumentvektor entsprechend seiner Klassifikation in die Datenbank zu überführen, wird wie folgt vorgegangen:

1. Überprüfen, ob die Klasse des Dokuments in der Tabelle KLASSE eingetragen ist und eventuell Anlegen einer neue Klasse.
2. Entnehmen eines Indexterms aus dem Dokumentvektor.
3. Überprüfen, ob ein Eintrag für den Indexterm in der Tabelle WORT existiert und eventuell Anlegen eines neuen Eintrags aus den Zusatzinformationen.
4. Überprüfen, ob ein Eintag für den Indexterm in Bezug auf die Klasse des Dokuments in der Tabelle HAEUFIGKEIT existiert und eventuell Anlegen eines neuen Eintrags, dessen Feld ANZAHL auf Null zu setzen ist.
5. Addieren der Anzahl des Terms im Dokument zum Feld ANZAHL des Eintrags von Indexterm und Klasse in der Tabelle HAEUFIGKEIT.
6. Falls der Dokumentvektor weitere Indexterme enthält, gehe zurück zu Schritt 2.

Das Anlegen eines Eintrages im Benutzerlexikon wird durch eine Funktion mit den Argumenten Klasse, Term und Korrelation unterstützt. Diese Argumente werden wie folgt in die Datenbank eingefügt:

1. Überprüfen, ob die Klasse des Dokuments in der Tabelle KLASSE eingetragen ist und eventuell Anlegen einer neuen Klasse.
2. Überprüfen, ob der Term in der Tabelle BEGRIFF eingetragen ist und eventuell Anlegen eines Neuen.
3. Falls in der Tabelle KORRELATION ein Eintrag für den Term in Bezug auf die Klasse mit gleicher Korrelationsart existiert, Ausgeben einer Warnung und Verwerfen der Angaben, ansonsten Anlegen eines neuen Eintrags mit der übergebenen Korrelation.

4.3.2.2 Bestimmen der Gewichte und Relevanzen

Nachdem die klassifizierten Dokumente vollständig in das automatische Lexikon übernommen wurden, ist die Ermittlung der Gewichte in der Tabelle HAEUFIGKEIT erforderlich, um mit dem Teillexikon arbeiten zu können. Zu ihrer Berechnung stehen z.B. die in Abschnitt 2.4 vorgestellten Verfahren zur Verfügung. Um die normierte Termfrequenz für einen Indexterm mit WORTID 13 in einer Klasse mit KLASSEID 2 zu bestimmen, wird zunächst die Anzahl aller Wörter der Klasse (siehe Abbildung 19 (a)) und die Häufigkeit des Indexterms in der Klasse (siehe Abbildung 19 (b)) aus der Datenbank abgefragt. Anschließend führt die Anwendung die nötigen Rechnungen durch und aktualisiert das Gewicht in der Datenbank wie im Beispiel auf den Wert 3.4 (siehe Abbildung 19 (c)).

(a)
```
SELECT SUM(HAEUFIGKEIT.ANZAHL)
FROM WORT INNER JOIN HAEUFIGKEIT ON WORT.WORTID = HAEUFIGKEIT.WORTID
WHERE ((HAEUFIGKEIT.KLASSEID)=2)
```

(b)
```
SELECT HAEUFIGKEIT.ANZAHL
FROM HAEUFIGKEIT
WHERE ((HAEUFIGKEIT.WORTID)=13 AND ((HAEUFIGKEIT.KLASSEID)=2))
```

(c)
```
UPDATE HAEUFIGKEIT SET HAEUFIGKEIT.GEWICHT = 3.4
WHERE (((HAEUFIGKEIT.WORTID)=13) AND ((HAEUFIGKEIT.KLASSEID)=2))
```

Abbildung 19 SQL-Statments zur Ermittlung der normierten Termfrequenz

Ist die Auswertung des Datenbestandes abgeschlossen, werden für die statistische Bewertung irrelevante Einträge gesperrt. Das gilt für seltene Terme oder jene mit gleichen Gewichten in allen Klassen, die für eine Differenzierung der Dokumente ohne Bedeutung sind. Für diese ist das Feld RELEVANZ der Tabelle WORT nicht gesetzt, der Term geht nicht in die Klassenvektoren ein. Vorteil des eingeschränkten Klassenvektors ist die beschleunigte Klassifikation, da weniger Einträge zu verarbeiten sind.

4.3.2.3 Ermitteln der Klassenvektoren

Um die Angaben im Lexikon zur Klassifikation nutzen zu können, müssen sie durch eine Abfrage in Klassenvektoren überführt werden. Für die statistischen Bewertungsverfahren baut sich der Klassenvektor aus den Indextermen mit ihren zugehörigen Klassen und Gewichten auf. Diese Informationen werden durch eine Abfrage aus dem automatischen Lexikon ermittelt, wie z.B. in Abbildung 20 (a) für die Klasse mit KLASSEID 2. Für die regelbasierte Bewertung tritt an die Stelle des Gewichts die Korrelation. Der Klassenvektor wird in diesem Fall durch eine Abfrage des Benutzerlexikons bestimmt. So werden in Abbildung 20 (b) die manuell definierten Terme für die Klasse mit KLASSEID 2 aus dem Lexikon gewonnen.

(a)
```
SELECT WORT.WORT, HAEUFIGKEIT.KLASSEID, HAEUFIGKEIT.GEWICHT
FROM WORT INNER JOIN HAEUFIGKEIT ON WORT.WORTID = HAEUFIGKEIT.WORTID
WHERE WORT.RELEVANZ = 1 AND KLASSE.KLASSEID = 2
```

(b)
```
SELECT BEGRIFF.BEGRIFF, KORRELATION.KLASSEID, KORRELATION.KORRELATION
FROM BEGRIFF INNER JOIN KORRELATION
    ON BEGRIFF.BEGRIFFID = KORRELATION.BEGRIFFID
 WHERE KLASSE.KLASSEID = 2
```

Abbildung 20 SQL-Statments zur Abfrage eines Klassenvektors aus dem statistischen (a) und regelbasierten (b) Lexikon

4.4 Implementierung des Prozesses zum Aufbau des Lexikons

Der Aufbau des Lexikons gliedert sich in zwei unabhängige Abläufe für den automatischen bzw. regelbasierten Teil. Soll ein Teillexikon überarbeitet werden, ist der jeweilige Ablauf zu wiederholen. So kann der Benutzer zu einem

bestehenden Lexikon ein neues Dokument hinzuladen oder einen neuen Term definieren, um die Klassifikation zu präzisieren. Ähnlich dem unter Abschnitt 2.3.3 beschriebenen Verfahren des relevance Feedbacks können falsch oder nicht klassifizierte Dokumente in ein Lexikon aufgenommen werden, um die Gewichtung der Indexterme anzupassen. Welche Schritte für den Aufbau bzw. die Aktualisierung der Teillexika durchlaufen werden, wird in den nächsten Abschnitten erläutert.

4.4.1 Automatische Erstellung des Lexikons für statistische Verfahren

Es ist die Aufgabe des **Einlesers** die klassifizierten Dokumente in das automatische Lexikon aufzunehmen. Er bestimmt für jedes Dokument die zugeordnete Klasse, erzeugt seinen Dokumentvektor und liest diese Angaben in die Datenbank ein. Sind alle Dokumente in die Datenbank übernommen, berechnet er die Gewichte und bestimmt die Relevanz für jeden Indexterm. Erst dann ist das Lexikon vollständig erstellt und kann zur Klassifikation von Dokumenten verwendet werden.

Die Zuordnung eines Dokuments zu einer Klasse wird über die Extension seines Dateinamens vorgenommen. Diese durch einen Punkt abgetrennten letzten drei Zeichen des Dateinamens entsprechen dem Inhalt des Feldes EXTENSION der Tabelle KLASSE des Lexikons. Aus dem referenzierten Eintrag kann die Klasse eines Dokuments bestimmt werden. So bekommt im Beispiel aus Abbildung 21 die Datei Dokument.005 die Klasse Beschwerde zugewiesen.

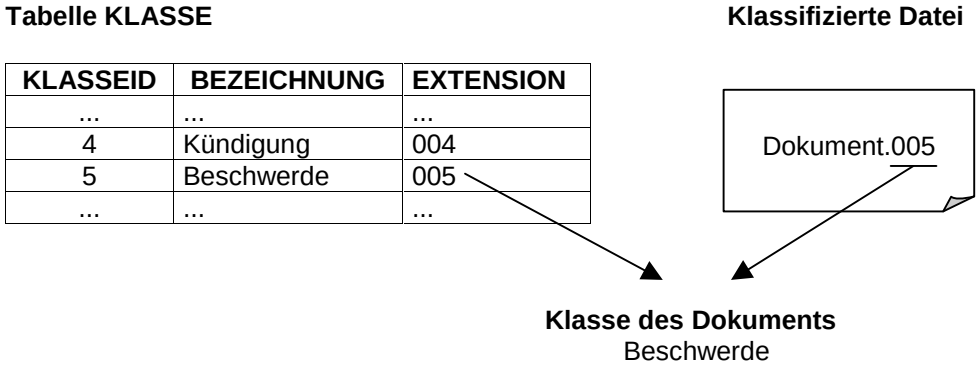


Abbildung 21 Beispiel für die Zuordnung einer Klasse zu einem Dokument

Für das Bilden der Dokumentvektoren wird das im Abschnitt 4.2 beschriebene Verfahren angewendet. Die Dokumentvektoren werden gemeinsam mit ihren Klasseninformationen an die Funktion aus Abschnitt 4.3.2.1 übergeben, die

daraufhin das Lexikon aktualisiert. Sind alle klassifizierten Dokumente in das Lexikon eingeladen, werden über die im Abschnitt 4.3.2.2 erläuterte Funktion die Gewichte und Relevanzen der Indexterme berechnet.

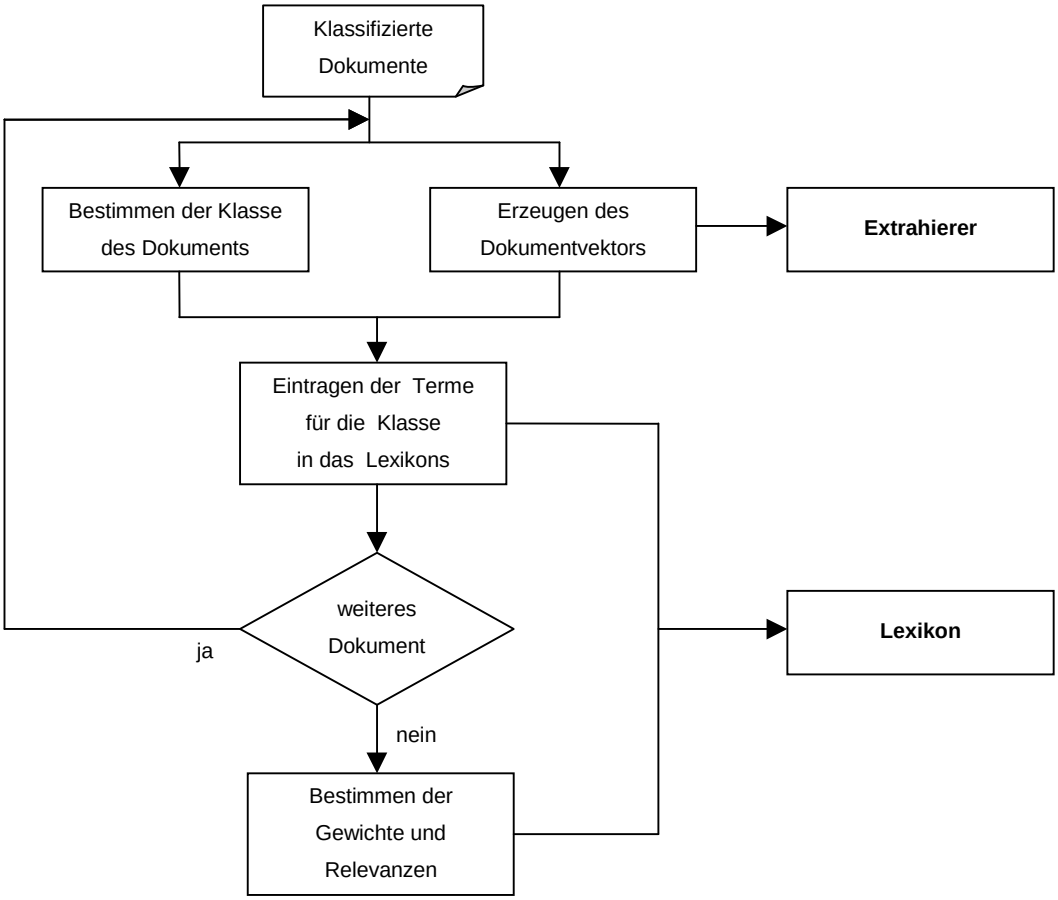


Abbildung 22 Aufbau des automatischen Lexikons

Der Einleser wird von der Kommandozeile mit folgenden Optionen aufgerufen:

c:\>Einleser [-n] Quellenpfad Lexikonpfad

Über den optionalen Schalter –n ist es möglich, den bestehenden Teil des Lexikons vor dem Laden neuer Dokumente zu löschen. Dies ist nötig wenn ein neues Lexikon erstellen werden soll.

Der Quellenpfad verweist auf ein oder mehrere Dokumente, die in das Lexikon aufgenommen werden sollen. Mehrere Dokumente können durch Wildcards im Quellenpfad übergeben werden. Der Lexikonpfad gibt eine Mircrosoft-Access-Datenbank an, die durch die Dateiextension .mdb gekennzeichnet ist. Diese Datenbank muß der unter Abschnitt 4.3.1 beschriebenen Struktur entsprechen.

Im folgenden Beispiel wird zunächst der Inhalt der Lexikondatei lexikon.mdb gelöscht und anschließend alle Dokumente, die mit der Zeichenkette „brief“ beginnen in die Datenbank lexikon.mdb geladen und ausgewertet.

c:\>Einleser –n brief*.* lexikon.mdb

4.4.2 Erstellen des Benutzerlexikons

Das Benutzerlexikon wird mit Hilfe eines Werkzeuges erstellt [Her98], das die Gruppierung synonymer Wörter zu Überbegriffen in einem Thesaurus erlaubt. Aus den Überbegriffen, Wörtern und dem Zeichen Tilde können verallgemeinerte Terme oder auch Stammbegriffe definiert werden. In einem Compile-Schritt werden die Wortgruppen des Thesaurus mit den Stammbegriffen zusammengeführt, ähnlich der Verallgemeinerung einer Anfrage aus Abschnitt 2.3.3. Die entstehenden Terme werden einer morphologischen Analyse durch den Extrahierer unterzogen und gemeinsam mit den Angaben zur Klasse und Korrelation an die im Abschnitt 4.3.2.1 aufgeführte Funktion übergeben, die daraufhin das Benutzerlexikon aktualisiert.

Angenommen in das Benutzerlexikon sollen für eine Klasse Beschwerde Begriffe eingetragen werden. So kann zunächst ein Thesauruseintrag `[Service] = Service, Beratung, Behandlung` angelegt werden. Der Überbegriff kann jetzt zur Definition von Stammbegriffen eingesetzt werden, wie z.B. in `schlechter [Service]` oder `[Service] ~ unzureichend`. Aus diesen Stammbegriffen werden die folgende Einträge gebildet:

Stammbegriff	`schlechter [Service]`	`[Service] ~ unzureichend`
Begriffe für das Benutzer-lexikon	`schlechter Service`	`Service ~ unzureichend`
	`schlechter Beratung`	`Beratung ~ unzureichend`
	`schlechter Behandlung`	`Behandlung ~ unzureichend`

Tabelle 16 Beispiel für das Anlegen von Begriffen für das Benutzerlexikon

Grammatikalische Ungenauigkeiten sind hier nicht von Bedeutung, da die Wörter durch die morphologische Analyse reduziert und wie im Abschnitt 2.3.1 verallgemeinert werden. Die im zweiten Stammbegriff verwendete Tilde steht für eine variable Anzahl eingeschlossener Wörter. So findet der Term `Beratung ~`

unzureichend eine Entsprechung in dem Satz Ihre Beratung war sehr unzureichend.

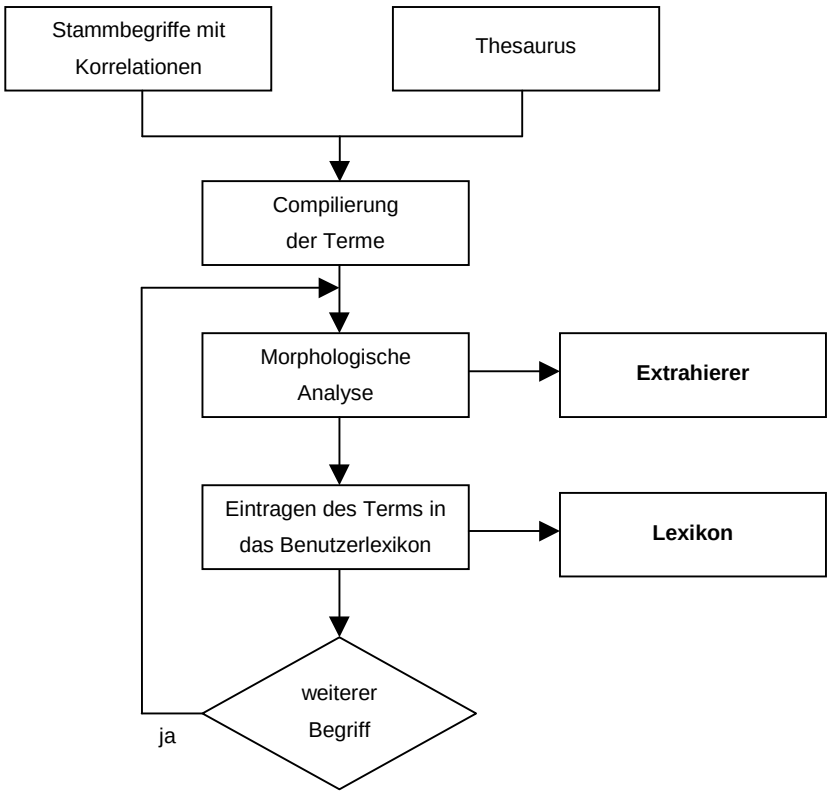


Abbildung 23 Aufbau des Benutzerlexikons

5 Klassifikation eines Dokuments

Nachdem der Aufbau des Lexkions abgeschlossen ist, kann es zur Klassifikation von Dokumenten herangezogen werden. Für ein zu klassifizierendes Dokument ist zunächst sein Termvektor zu ermitteln. Dieser Termvektor wird mit den Klassenvektoren aus dem Lexikon verglichen, um daraufhin aus den Bewertungen die Klasse des Dokuments zu ermitteln. Wie diese Schritte im Einzelnen ablaufen ist Gegenstand der folgenden Abschnitte.

5.1 Dokumentvektor als Anfrage

Eine Besonderheit bei der Klassifikation eines Dokuments durch ein Indexierungssytem ist die Verwendung seines Dokumentvektors als Anfrage. Die Erzeugung des Dokumentvektors erfolgt nach der Methode aus Abschnitt 4.2. Die Komponenten dieses Termvektors werden als konjugiert angesehen, d.h. jeder Term fließt als Hinweis auf die Klasse seines Dokuments in die Bewertung ein.

Da die Bewertung eines Dokuments auf dem Vergleich der Komponenten von Dokument- und Klassenvektor beruht, müssen die Dokumentvektoren für den Aufbau des Lexikons und für die Bewertung nach dem selben Verfahren erzeugt werden. Nur so kann die selbe Struktur der Terme in den Vektoren und somit ihre Vergleichbarkeit gesichert werden. Bei der Verwendung verschiedener Methoden zur morphologischen Analyse besteht beispielsweise die Gefahr, einen Begriff auf verschiedene Stammformen abzubilden. Für das Wort `berechnete` ist z.B. die Reduktion auf `berechn` oder `rechn` denkbar. Ist nun die eine Form im Lexikon gespeichert und fließt die andere in die Auswertung ein, bleibt beim Vergleich der Komponenten von Dokument- und Klassenvektoren die Übereinstimmung der originalen Wortformen unerkannt, die Bewertung wird verfälscht.

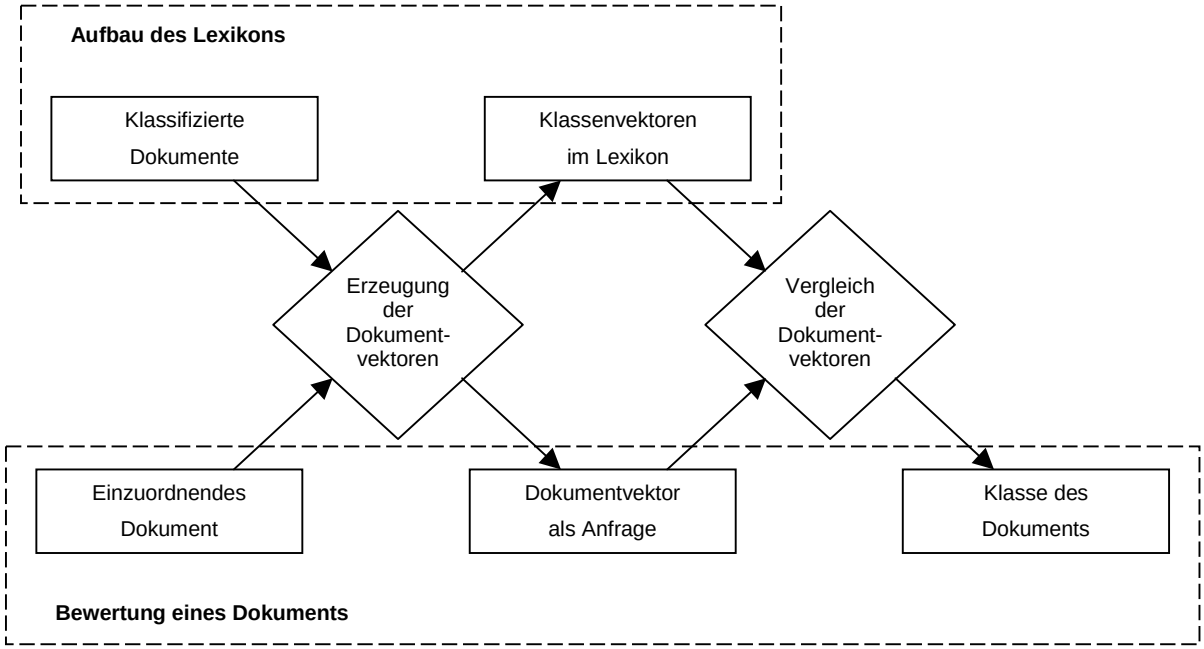


Abbildung 24 Aufgabe der Dokumentvektoren bei der Klassifikation

5.2 Bewertung eines Dokuments

Um die Zugehörigkeit eines Dokuments zu einer Klasse zu bestimmen, müssen zunächst die Klassenvektoren mittels der unter Abschnitt 4.3.2.3 beschriebenen Funktion aus dem Lexikon gewonnen werden. Eine Komponente dieser Vektoren enthält neben dem Indexterm, seine Gewichte bzw. Angaben über seine Korrelationen. Aus dem Vergleich der Terme einer Anfrage mit denen eines Klassenvektors, werden die für das Dokument zutreffenden Gewichte und Korrelationen ermittelt. Diese Werte bilden die Grundlage für die folgenden Berechnungen.

Die Angaben aus verschiedenen Teillexika gehen vorerst in getrennte Kategorien der Bewertung ein, d.h. ein Dokument wird gesondert nach statistischen und regelbasierten Verfahren beurteilt. Unter Verwendung einer Funktion aus Abschnitt 3.2 kann aus den ermittelten Werten für die Gewichte und die schwachen Korrelationen eine Bewertung für das Dokument in einer Klasse berechnet werden. Aus der Größe des erzielten Wertes für eine Klasse läßt sich ihre Relevanz für das Dokument ableiten. Die Auswertung der starken Korrelationen wird in einer gesonderten Kategorie behandelt. Für jeden Term der Anfrage mit dieser Art von Korrelation wird nur die referenzierte Klasse vermerkt,

da für eine daraus resultierende Zuordnung des Dokuments keine abgestufte Bewertung erforderlich ist.

Das Ergebnis der Bewertung eines Dokuments könnte wie in Tabelle 17 ausfallen. An diesem Beispiel sollen die weiteren Schritte bis zu eindeutigen Klassifikation eines Dokuments verdeutlicht werden.

Kategorie der Bewertung	**Klasse**				
	1	**2**	**3**	**4**	**5**
starke Korrelation	0	1	0	1	0
schwache Korrelation	0.3	10.4	3.5	0.4	2.7
Gewicht	85.3	87.8	9.6	5.4	7.2

Tabelle 17 Beispiel für das Bewertungsergbnis eines Dokuments

5.3 Kandidaten für die Klassifikation ermitteln

Um unbedeutende von aussagekräftigen Bewertungen zu trennen, wird die Auswahl von Kandidaten für die Klasse eines Dokuments nötig. Dazu wird für die Kategorie der schwachen Korrelationen bzw. der Gewichte ein Signifikanzwert nach einer Methode aus Abschnitt 3.3 berechnet. Daraufhin werden alle Bewertungen der jeweiligen Kategorie eliminiert, die unter diesem Wert liegen. So bleiben für die weiteren Untersuchungen die Bewertungen erhalten, deren erzielte Werte nicht Ergebnis zufälliger Übereinstimmung sondern von hoher Signifikanz sind. Für die starken Korrelationen gehen alle vermerkten Klassen als Kandidat in die Klassifikation ein, diese Kategorie wird nicht eingeschränkt.

Kategorie der Bewertung	**Klasse**				
	1	**2**	**3**	**4**	**5**
starke Korrelation	0	**1**	0	**1**	0
schwache Korrelation	0.3	**10.4**	3.5	0.4	2.7
Gewicht	**85.3**	**87.8**	9.6	5.4	7.2

Tabelle 18 Aussagekräftige Werte des Bewertungsergebnisses

Für das Beispiel einer Bewertung aus dem vorherigen Abschnitt erweisen sich die hervorgehobenen Werte in Tabelle 18 als aussagekräftig. Aus ihnen ist jetzt die Klasse des Dokuments abzuleiten.

5.4 Die Klasse eines Dokuments bestimmen

Um die Bewertung des Dokuments zu einem eindeutigen Ergebnis zu führen, wird die unter Abschnitt 3.4.3 beschriebenen Methode auf die relevanten Klassen der Kategorien angewandt. Die Zusammenfassung der relevanten Klassen der Kategorien im Beispiel aus Tabelle 19 ergibt für das bewertete Dokument die Klasse 2 mit der Priorität 1.

Kategorie der Bewertung	**Klasse**				
	1	**2**	**3**	**4**	**5**
starke Korrelation	0	**1**	0	**1**	0
schwache Korrelation	0.3	**10.4**	3.5	0.4	2.7
Gewicht	**85.3**	**87.8**	9.6	5.4	7.2

Tabelle 19 Eindeutige Klassifizierung aus den relevanten Klassen

Damit die Qualität des erzielten Ergebnisses bei erfolgreicher Klassifikation für den Nutzer einsichtiger wird, soll neben der Priorität noch eine Widerspruchsanalyse ausgegeben werden. Diese Analyse erstellt eine Beschreibung, die das Ergebnis der Klassifikation in die relevanten Klassen der Kategorien einordnet. Für jede Kategorie werden die Fälle aus Tabelle 20 geprüft und das entsprechende Kennzeichen ausgegeben. Die Untersuchung der starken Korrelationen aus dem Beispiel ergibt als Kennzeichen ein „I“, da die Klasse 2 in den Kandidaten dieser Kategorie liegt. Aus diesen Zeichen können Schlüsse zur Optimierung des Systems gezogen werden. So sollte ein richtig klassifiziertes Dokument, dessen Klasse einen Fehler oder Widerspruch zu den Kandidaten des automatischen Lexikons liefert, zur Aufarbeitung dieses Teillexikons herangezogen werden.

Das Ergebnis der zusammenfassenden Auswertung und seine Prioritätsstufe wird gemeinsam mit der Widerspruchsanalyse ausgegeben, die Klassifikation ist abgeschlossen. Für das betrachtete Beispiel sieht die Ausgabe wie folgt aus:

Klasse: 2 Priorität: 1 Widerspruch: IEI

Zeichen	Anzahl der Kandidaten	Einordnung in die Klassifikation	Qualität
E	1	**E**xakte Übereinstimmung mit dem Kandidaten	sehr gut
I	>1	**I**n der Menge der Kandidaten	gut
K	0	**K**ein Kandidat in der Kategorie	befriedigend
F	1	**F**ehlende Übereinstimmung mit einem Kandidaten	schlecht
W	>1	**W**iderspruch zu einer Gruppe von Kandidaten	schlecht

Tabelle 20 Kennzeichen der Widerspruchsanalyse

5.5 Implementierung des Prozesses zur Klassifikation

Die Zuordnung einer Klasse zu einem Dokument ist die Aufgabe des **Bewerters**. Er liest vor Beginn der Klassifikation aus dem Lexikon die abgelegten Indexterme gemeinsam mit den Informationen zu den Klassen in die Klassenvektoren ein. Dazu wird die unter Abschnitt 4.3.2.3 beschriebene Funktion des Lexikons benutzt. Anschließend wird von einem zu bewertenden Dokument mit Hilfe der unter Abschnitt 4.2 erläuterten Methode des Extrahierers der Dokumentvektor erzeugt. Seine Komponenten werden mit denen der Klassenvektoren verglichen, um die Bewertung des Dokuments in jeder Klasse zu ermitteln. Aus den Bewertungen werden mit einem Signifikanzmaß die relevanten Kandidaten für jede Kategorie bestimmt. In einer abschließenden Auswertung wird aus den Kandidaten die Klasse des Dokuments und eine Widerspruchsanalyse gewonnen. Das Ergebnis der Klassifikation wird gegebenenfalls mit seiner Prioritätsstufe und seinen Kennzeichen für den Widerspruch ausgegeben.

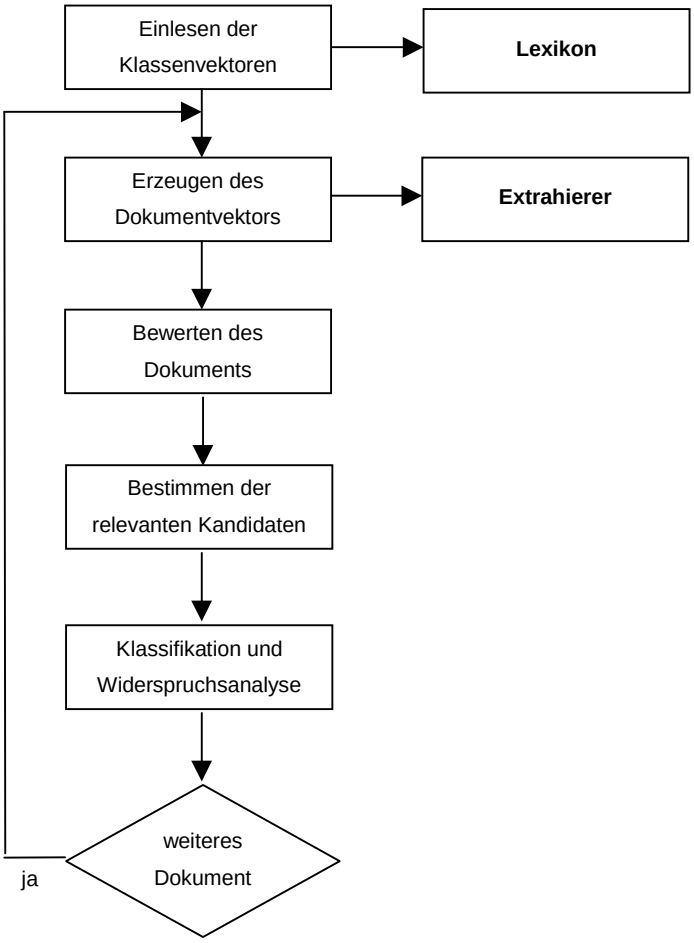


Abbildung 25 Ablauf der Bewertung von Dokumenten

Der Bewerter wird von der Kommandozeile mit den folgenden Optionen aufgerufen:

c:\>Bewerter [-b] Quellenpfad Lexikonpfad

Über den optionalen Schalter –b wird das Programm angewiesen eine detaillierte Beschreibung des Bewertungsprozesses für jedes Dokument auszugeben. Die Ausgabe erfolgt in eine Datei die den Namen des bewerteten Dokuments trägt und mit der Extension .bgd gekennzeichnet ist. Aus dieser Beschreibung sind neben den gefundenen Einträgen der Teillexika auch die Bewertungen und relevanten Kandidaten für das Dokument ersichtlich (siehe Abschnitt VI.ii). Diese Datei soll vor allem bei Fehlklassifikationen die Fehlersuche und Optimierung des Bewertungsprozesses unterstützen.

Der Quellenpfad verweist auf ein oder mehrere Dokumente, die bewertet werden sollen. Mehrere Dokumente können durch Wildcards im Quellenpfad übergeben werden. Der Lexikonpfad gibt eine Mircrosoft-Access-Datenbank an, die durch die

Dateiextension .mdb gekennzeichnet ist. Diese Datenbank muß der unter Abschnitt 4.3.1 beschriebenen Struktur entsprechen.

Im unten angegebenen Beispiel wird das Dokument brief10.txt unter zur Hilfenahme des Lexikons lexikon.mdb bewertet. Für das Ergebnis der Bewertung wir eine Begründung in der Datei brief10.bgd abgelegt.

c:\>Bewerter –b brief10.txt lexikon.mdb

6 Test eines Klassifikationssystems

Im Folgenden sollen die bisher vorgestellten Verfahren auf ein Beispiel in der Praxis angewendet werden. An die Stelle zu klassifizierender Dokumente tritt die Geschäftspost eines Versicherungsunternehmens. Diese wird nach dem Empfang in maschinenlesbare Form konvertiert und ist für die Verteilung im Unternehmen in bestehende Klassen einzuordnen. Nach den Klassen ist die Zustellung der Briefe an die entsprechenden Sachbearbeiter möglich.

Aufgrund der großen Menge von ca. 60000 eingehenden Briefen pro Tag, ist eine möglichst schnelle Zuordnung einer Klasse zu einem Brief gefordert, um das Indexierungssystem sinnvoll nutzen zu können. Neben der Geschwindigkeit des Systems ist der zu erwartende Anteil erfolgreicher Klassifikationen und seine Genauigkeit von Interesse. Durch einen Prototyp sollen Näherungen für diese Parameter ermittelt werden.

6.1 Voraussetzungen für den Test

Bevor auf die Durchführung der Tests eingegangen wird, sollen seine technischen Grundlagen erläutert werden, um die erzielten Ergebnisse besser einschätzen zu können. Zudem werden die zur Verfügung gestellten Daten beschrieben, die für die Bewertung genutzt werden.

6.1.1 Eingesetzte Hard- und Software

Die Testumgebung wird unter Windows 95 auf einem PC mit Intel Pentium II Prozessor und 64 MB Hauptspeicher simuliert. Die Software für das Indexierungssystem ist mit der Programmiersprache Mircrosoft Visual C++ 5.0 implementiert. Das Lexikon wurde mit der Datenbankverwaltung Mircrosoft Access 97 erstellt. Der Zugriff der Software auf die Datenbank wird durch die Schnittstelle Mircrosoft DAO 3.5 realisiert.

6.1.2 Datenmaterial zur Klassifikation

Für die Klassifikation liegen nach Klassen geordnete Textbausteine vor, aus denen der entsprechende Sachbearbeiter einen Antwortbrief erstellt. Sie werden auch als AKO-Texte bezeichnet und stellen mit 4MB den größeren Anteil des Datenmaterials für die Klassifikation.

Ergänzend stehen 130 Kundenbriefe im ASCII-Format zur Verfügung, denen ebenfalls eine Klasse zugeordnet ist. Sie wurden aus originalen Briefen mittels OCR-Software gewonnen und manuell von Fehlern der Konvertierung befreit, um Verfälschungen der Ergebnisse zu verhindern. Mit 80 KB liefern sie nur eine kleine Menge von Daten für die Klassifikation.

6.2 Ergebnisse des Tests

Nachdem die Daten für die Bewertung vorgestellt wurden, sollen in den nächsten Abschnitten die mit ihnen erzielten Ergebnisse erläutert werden. Die betrachteten Ergebnisse resultieren aus Versuchen mit dem automatischen Lexikon und seinem kombinierten Einsatz mit dem Benutzerlexikon.

6.2.1 Automatische Verfahren

Im ersten Testlauf wurden die AKO-Texte in das automatische Lexikon übernommen. Das Teillexikon enthielt daraufhin ca. 170000 Indexterme, die insgesamt etwa 1,3 Millionen mal in den Texten auftraten. Diese Verdichtung der Terme um den Faktor 10, obwohl Stopwörter entfernt wurden, weist auf viele Wiederholungen in den Texten hin. So tritt etwa der aus der Redewendung `Name und Anschrift des Kontoinhabers` gewonnene Term `Anschrift ~ Kontoinhaber` in allen Texten 212 mal auf.

Die Bewertung der Kundenbriefe unter Verwendung der AKO-Texte führt kaum zu erfolgreichen Klassifikationen. Das liegt nicht an der falschen Gewichtung der Indexterme, die zu falschen Ergebnissen bei der Klassifikation führen würde, sondern an ihren seltenen Übereinstimmungen mit den Termen in den Briefen. Da sich nur wenige Terme des Lexikons in den Kundenbriefen wiederfinden, ist eine aussagekräftige Bewertungen nicht möglich. Somit eignen sich die AKO-Texte nicht zur Klassifikation der Kundenbriefe.

Der Aufbau des automatischen Lexikons aus allen Kundenbriefen führt zu ca. 21000 Indextermen, die zusammen ca. 23000 mal in den Briefen vorkommen. Der größte Teil von ca. 19700 Termen tritt also genau einmal auf. Trotz der wenigen Wiederholungen läßt sich aus den Häufigkeiten pro Indexterm bereits der Einfluß der Länge eines Mehrwortterms für die Klassifikation absehen. Je mehr Teilterme ein Mehrwortterm enthält, desto seltener wird er wiederholt identifiziert, d.h. er

fließt immer weniger in die Klassifikation ein. Die Terme mit variabler Länge ordnen sich zwischen Einworttermen und sehr langen Mehrworttermen ein. Aufgrund ihrer Häufigkeit pro Indexterm eignen sie sich gut für eine Klassifikation, da sie weder zu allgemeine noch zu spezifische Beschreibungen liefern.

Die Verteilung der Gewichte in Bezug auf die Länge der Mehrwortterme macht eine verlagerte Bedeutung für die Indexterme nach der Berechnungsmethode deutlich. Für die Gewichtung nach normierten Termfrequenzen wächst der Stellenwert eines Terms mit seiner Häufigkeit, das Gewicht sinkt mit der Länge eines Mehrwortterms. Bei der Verstärkung des Gewichts für seltene Terme kehrt sich das Verhältnis um, das Gewicht steigt mit der Länge eines Mehrwortterms. Die Differenzierung der Dokumente wird für diesen Datenbestand am besten durch die Einwortterme bewältigt, da diese Terme in mehreren aber nicht allen Dokumenten auftreten. Für eine größere Menge von Daten wird sich das Gewicht auf die Mehrwortterme mit zwei bis drei Teiltermen verschieben, da sie dann eine mittlere und die Einworterme eine hohe Dokumenthäufigkeit aufweisen werden.

Teilterme		**1**	**2**	**3**	**4**	**5**	**variabel**
Indexterme		1975	3378	3728	3531	2780	7044
Häufigkeit		2631	3604	3829	3581	2804	7315
Häufigkeit pro Indexterm		1,33	1,07	1,03	1,01	1,01	1,04
Gewicht pro Indexterm	**normierte Termfrequenzen**	1,206	0,964	0,927	0,914	0,909	0,937
	Verstärkung seltener Terme	2,203	2,322	2,328	2,329	2,324	2,326
	Beachtung der Differenzierung	1,404 $x10^{-4}$	0,047 $x10^{-4}$	0,011 $x10^{-4}$	0,003 $x10^{-4}$	0,002 $x10^{-4}$	0,039 $x10^{-4}$

Tabelle 21 Verteilung der Mehrwortterme

Die Bewertung der Kundenbriefe anhand eines Lexikons, das aus ihnen selbst erstellt wurde, kann höchstes zur Prüfung der Stabilität eines Verfahrens genutzt werden. Daraus können aber keine allgemeinen Aussagen über eine erfolgreiche Klassifikation getroffen werde, da ihr in der Praxis systemfremde Briefe zugrunde liegen. Deshalb sollen aus den 130 Kundenbriefen verschiedene Szenarien zur Bewertung gebildet werden. Die 10 Briefe jeder Klasse werden jeweils in einen Teil für das Lexikon und die Bewertung zerlegt (siehe Tabelle 22). So werden einerseits Verbesserungen der Klassifikation für eine steigende Anzahl von

Dokumenten im Lexikon sichtbar gemacht, andererseits wird die Gefahr der Überbewertung von Fehlklassifikationen gesenkt, wie sie bei einer geringen Zahl klassifizierter Dokumente droht.

Szenario	**A**	**B**	**C**	**D**	**E**	**F**
Briefe für das Lexikon	65	78	91	104	117	130
Briefe für die Bewertung	65	52	39	26	13	130
Indexterme	12423	12106	14425	14425	18913	21160
Häufigkeit	12455	13203	15843	20011	21109	23764

Tabelle 22 Szenarien der Bewertung

Für die Szenarien sollen zunächst Bewertungen mit unterschiedlich gewichteten Indextermen im Lexikon durchgeführt werden. Als Bewertungsfunktion wird dabei das Innere Produkt verwendet und ein Brief nach dem Maximum seiner Bewertungen klassifiziert. Anhand der Vorklassifizierung läßt sich die Zuordnung der Klasse in richtig oder falsch unterschieden.

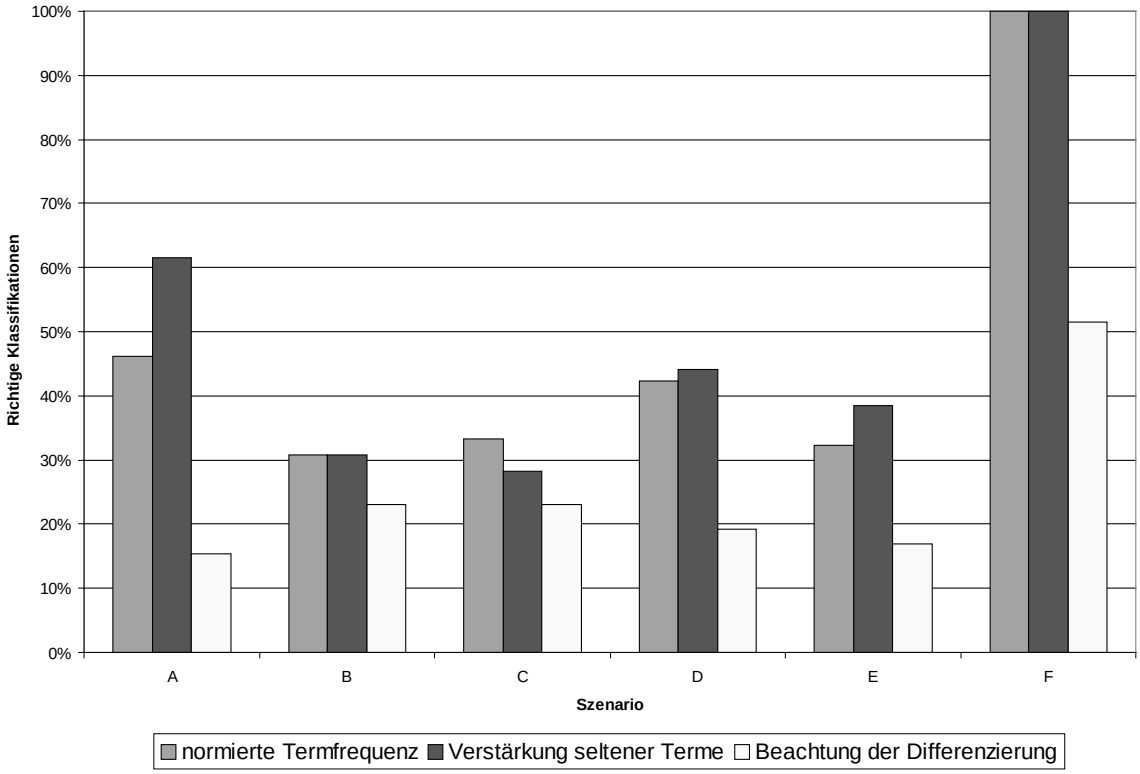


Abbildung 26 Bewertung mit verschiedenen Gewichten

In den Ergebnissen zeigt sich ein geringer Vorsprung für die Verstärkung seltener Terme gegenüber den releativen Häufigkeiten. Dieser Vorteil beruht auf der

stärkeren Beachtung spezifischer Terme, welche die Klassifikation verbessern. Das Verfahren der Gewichtung nach Differerzierung eines Terms arbeitet auf der kleinen Datenmenge nicht korrekt, da die von ihm benötigten mittelfrequenten Terme nicht im Lexikon vorhanden sind. So enthält ein nach diesem Verfahren aus allen Briefen aufgebautes Lexikon nur 734 relevante Terme, die für eine erfolgreiche Klassifikation aber nicht ausreichen.

Für die weiteren Betrachtungen soll die Gewichtung nach normierten Termfrequenzen herangezogen werden, da sie ähnlich gute Ergebnisse wie die Verstärkung seltener Terme liefert, ihre Gewichte aber schneller zu berechnen sind. Anhand dieser Gewichtung sollen im Folgenden verschiedene Bewertungsfunktionen getestet werden, wobei wiederum nur ihre Maxima zur Bestimmung des Klassifikationsergebnisses verwendet werden.

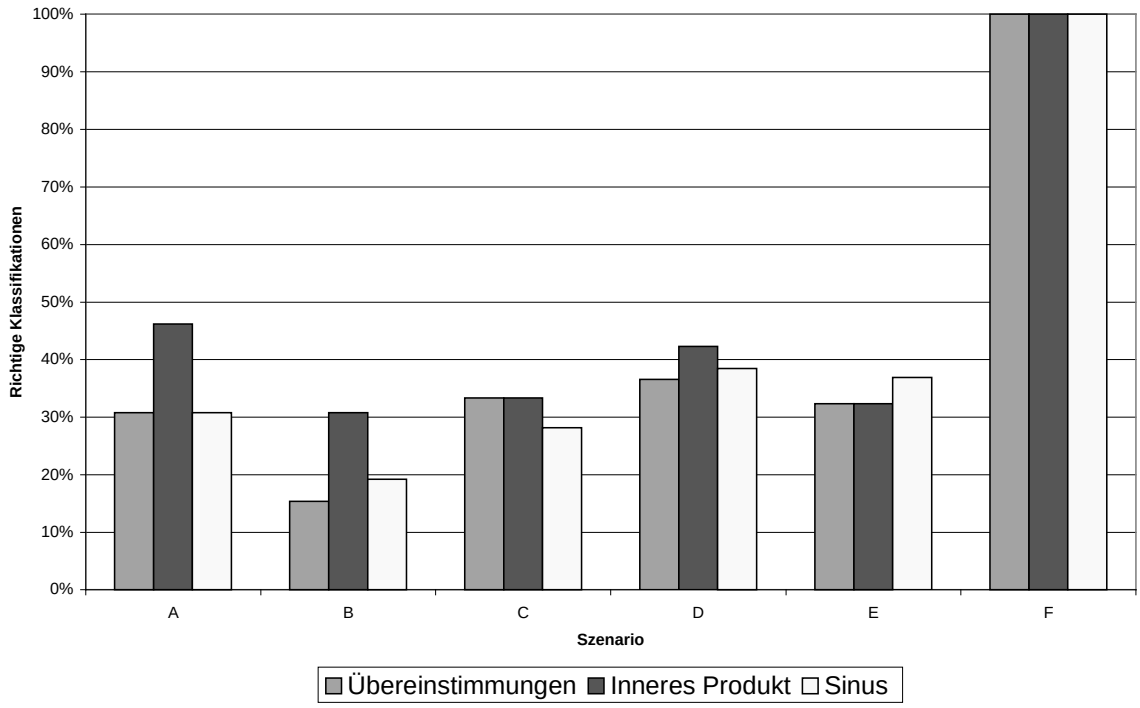


Abbildung 27 Bewertung mit verschiedenen Bewertungsfunktionen

Bei diesem Test zeichnet sich die Bewertung durch das Innere Produkt als beste Variante ab. Das schlechte Abschneiden der Bewertung nach Übereinstimmungen begründet sich aus der ungenauen Ermittlung der Ähnlichkeit, da für die relativ kurzen Briefe mit ca. 300 Termen so nur eine schlechte Differenzierung möglich ist. Auch die Bewertung nach dem Sinus zwischen den Dokumentvektoren ist

noch zu ungenau, da sich zu wenige Entsprechungen zwischen den Indextermen der Briefe und des Lexikons finden lassen, was zu einer verzerrten Bewertung führt.

Bevor ein Signifkanzmaß bei einer Bewertung der Szenarien eingesetzt wird, soll zunächst die Bestimmung eines Parameters für das Maß zur Abgrenzung der aussagekräftigen Bewertungen demonstriert werden. Dafür wird das Szenario C nach dem Inneren Produkt bewertet. Die Indexterme sind hierbei wiederum nach ihren normierten Termfrequenzen gewichtet. Die Signifikanz des Bewertungsergebnisses wird nach der Schätzung eines Konvidenzintervalls bestimmt, deren Sicherheitswahrscheinichkeit zwischen 97,72% und 99,86 % variiert wird. Mit steigender Signifikanz sinkt der Umfang des Ergebnisses, es erhöht sich aber seine Genauigkeit, da die falschen Klassifikationen abnehmen. In die weiteren Untersuchungen wird der Parameter 98,92% einfließen, da für ihn der Anteil richtiger Klassifikationen bei geringster Einschränkung des Ergebnisses überwiegt. Für einen anderen Datenbestand sollte der Parameter mit dieser Methode erneut bestimmt werden, um ihn optimal an die veränderte Bewertung anzupassen.

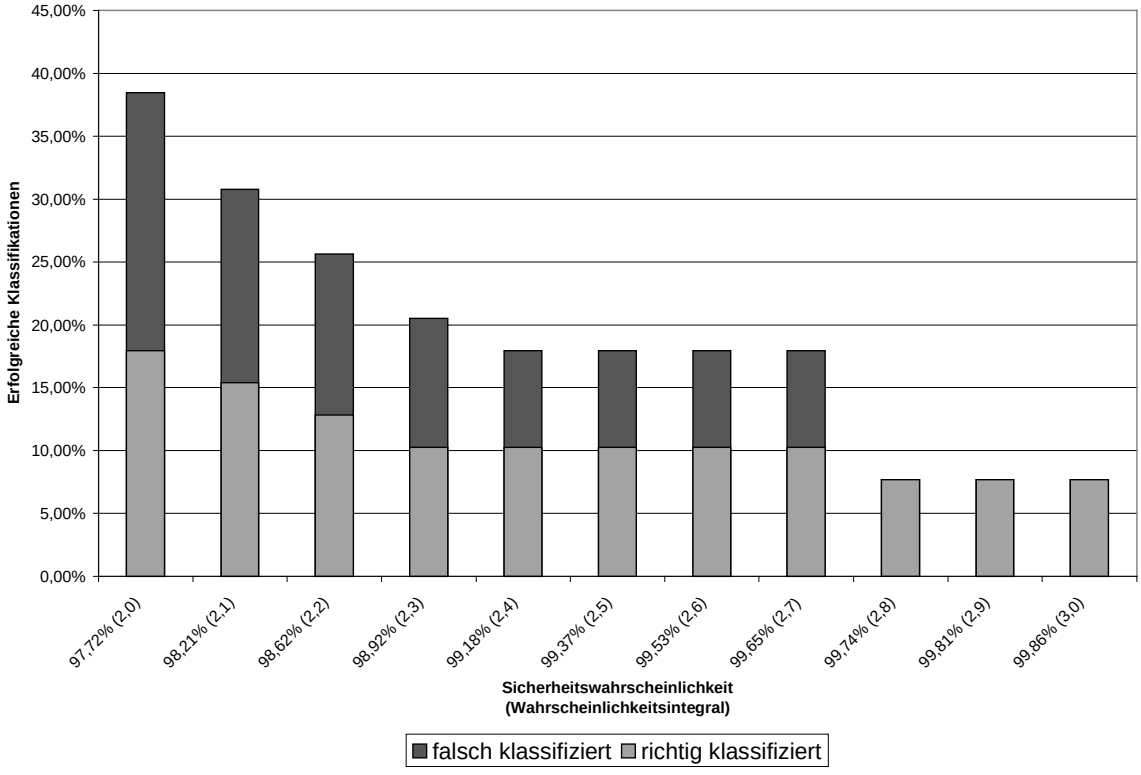


Abbildung 28 Beispiel für die Bestimmung eines Parameters für ein Signifikanzmaß

Wird die ermittelte Signifikanz auf alle Szenarien angewandt (siehe Abbildung 29) zeigt sich mit wachsender Zahl der Indexterme im Lexikon ein vergrößerter Umfang richtig klassifizierter Dokumente im Ergebnis. Aber auch das Verhältnis von richtig zu falsch klassifizierten Dokumenten verbessert sich mit größerem Lexikon, die Klassifikation läßt sich genauer durchführen. Dies zeigt wie wichtig ein möglichst umfangreiches Lexikon für eine erfolgreich Klassifikation ist.

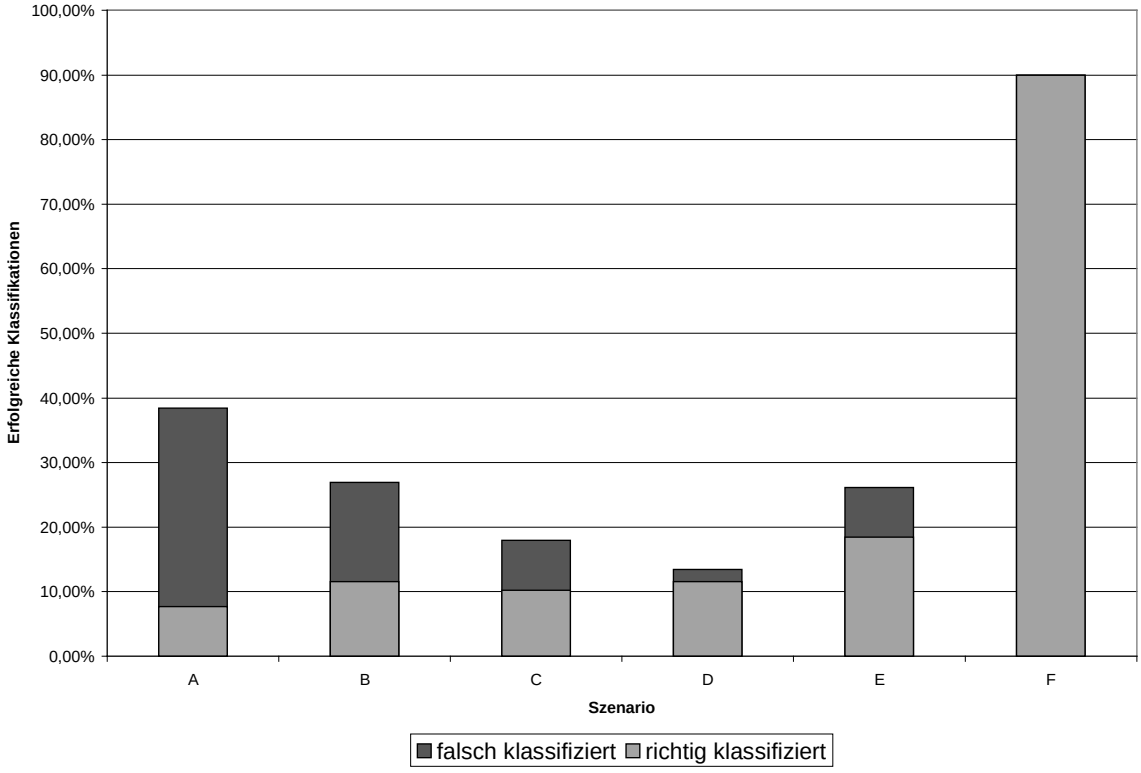


Abbildung 29 Bewertung mit einem Signifikanzmaß

Insgesamt verringert sich die Zahl richtig klassifizierter Briefe bei der Einschränkung durch ein Signifikanzmaß gegenüber den Ergebnissen, die durch die Auswertung des Maximums gewonnenen wurden. Die Ursache hierfür liegt im Verlust der nach dem Maximum richtig klassifizierten Dokumente, wenn sich ihre Bewertung von denen anderer Klassen nicht ausreichend abgrenzen läßt.

6.2.2 Kombination von statistischen und regelbasierten Verfahren

Um abschließend die Verbesserung der Klassifikation durch regelbasierte Verfahren zu betrachten, sind zunächst Terme für das Benutzerlexikon zu gewinnen. Diese Terme werden aus dem vollständig geladenen automatischen Lexikon durch die Abfragen aus Abbildung 30 ermittelt. Für die Beschreibung der

starken Korrelationen werden Terme bestimmt, die nur in einer Klasse häufig auftreten. Die schwachen Korrelationen werden aus den häufigen Termen ermittelt, die in mehreren Klassen auftreten. Als häufig gilt für diesen Datenbestand ein Term bereits, wenn er öfter als 2 mal in Dokumenten einer Klasse gefunden wurde. So ergeben sich 77 Einträge für die starken und 204 Einträge für die schwachen Korrelationen, die durch eine manuelle Bereinigung von nicht relevanten Termen auf 10 bzw. 21 eingeschränkt werden. Diese Terme werden gemeinsam mit ihren Korrelationen in das Lexikon geladen, und fließen in die folgende Bewertung ein. Dieses Verfahren zur Gewinnung von Termen für das Benutzerlexikon eignet sich auch für einen anderen Datenbestand, um für die Indexierung relevante Terme und ihre Korrelationen zu bestimmen.

(a)
```
SELECT KLASSE.EXTENSION, HAEUFIGKEIT.GEWICHT, WORT.WORT
FROM WORT INNER JOIN
    (KLASSE INNER JOIN HAEUFIGKEIT ON KLASSE.KLASSEID =
    HAEUFIGKEIT.KLASSEID) ON WORT.WORTID = HAEUFIGKEIT.WORTID
WHERE (((HAEUFIGKEIT.GEWICHT)>0) AND ((HAEUFIGKEIT.ANZAHL)>2));
```

(b)
```
SELECT KLASSE.EXTENSION, HAEUFIGKEIT.ANZAHL, WORT.WORT, WORT.BEISPIEL
FROM (WORT INNER JOIN [WortID und Anzahl der Klassen] ON
    WORT.WORTID = [WortID und Anzahl der Klassen].WORTID) INNER JOIN
    (KLASSE INNER JOIN HAEUFIGKEIT ON KLASSE.KLASSEID =
    HAEUFIGKEIT.KLASSEID) ON WORT.WORTID = HAEUFIGKEIT.WORTID
WHERE (((HAEUFIGKEIT.ANZAHL)>2) AND
    (([WortID und Anzahl der Klassen].[Anzahl von KLASSEID])=1));
```

Abbildung 30 SQL-Statments zur Gewinnung der Terme für das Benutzerlexikon für (a) starke und (b) schwache Korrelationen[4]

Die kombinierte Bewertung durch statistische und regelbasierte Verfahren führt neben der Steigerung des Anteils richtiger Klassifikationen auch zur Verbesserung des Verhältnisses von richtig zu falsch klassifizierten Briefen, da die Bewertungen aus dem Benutzerlexikon genauer sind.

[4] Die Abfrage [WortID und Anzahl der Klassen] bestimmt für jede WORTID eines Indexterms die Anzahl der referenzierten Klassen, also die Anzahl der Klassen, in denen der Indexterm vorkommt.

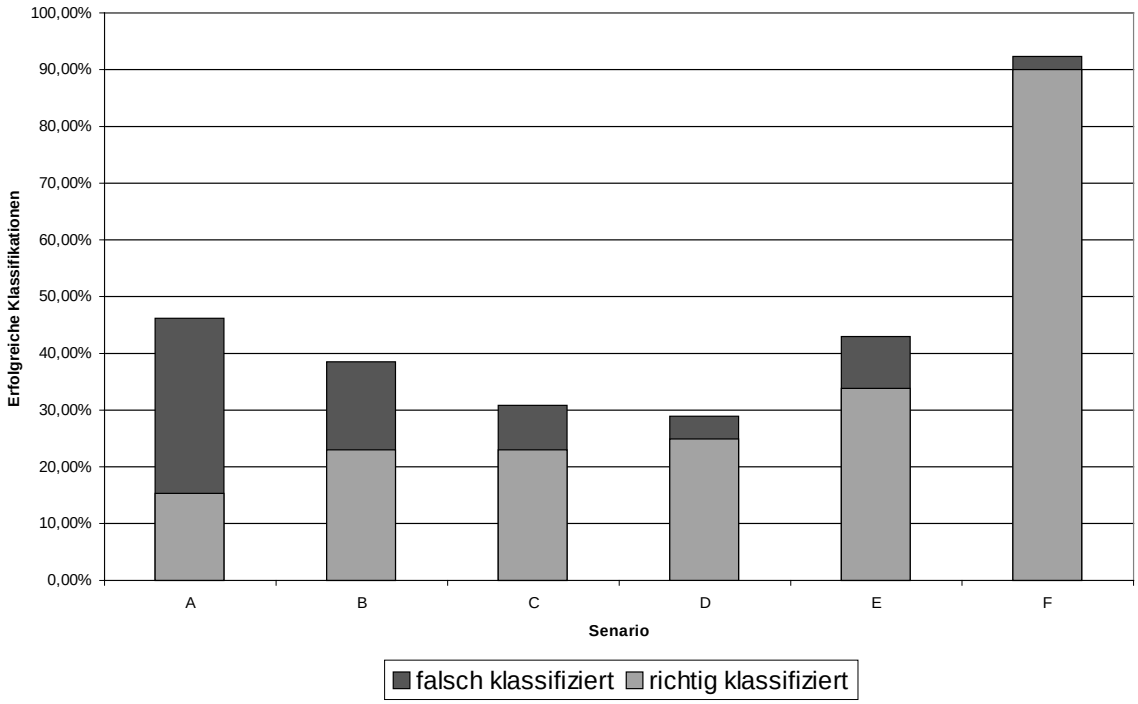


Abbildung 31 Kombinierte Bewertung durch alle Lexika

6.2.3 Zusammenfassung

Der Einsatz einer automatischen Klassifikation für Briefe ist aus zwei Gründen sinnvoll. Erstens liegt der Zeitaufwand für eine Klassifikation durch ein Indexierungssystem mit etwa 0,3 s viel niedriger als bei einer manuellen Einordnung, so daß auch bei einem nicht klassifizierbaren Dokument kein Nachteil durch die Klassifikation entsteht. Zweitens erzielt das System für ein Drittel der Briefe eine richtige Klassifikation mit steigender Tendenz für ein umfangreicheres Lexikon.

Eine manuelle Klassifikation der Briefe wird aber auch weiterhin nötig sein, da selbst bei optimaler Anpassung des Indexierungsystems der automatischen Klassifikation Grenzen gesetzt sind. So treten beispielsweise für sehr kurze und allgemein formulierte Briefe Probleme bei der Bewertung auf, da ihnen aufgrund ihres geringen Inhalts keine Klasse zugeordnet werden kann. Ähnlich verhält es sich mit Briefen, die mehrere Anliegen enthalten und darum nicht eindeutig einer Klasse zugeordnet werden können.

I Kurzzusammenfassung

Das Ziel dieser Arbeit ist die Entwicklung eines Indexierungssytems zur eindeutigen Klassifikation von Dokumenten anhand einer vorgegebenen Zahl von Klassen. Die Grundlage zum Aufbau des Klassifikationssystems bilden allgemeine Indexierungsysteme für Dokumente.

Diese allgemeinen Systeme werden für die Klassifikation modifiziert, um die Verarbeitung der Dokumente zu beschleunigen und die Klassifikationsergebnisse zu präzisieren. Diese optimierten Methoden zur Indexierung sowie der kombinierte Einsatz von statistischen und regelbasierten Verfahren zur Bewertung von Dokumenten bilden die Basis des Klassifikationssystems. Zudem werden spezielle Techniken zur eindeutigen Klassifikation von Dokumenten erarbeitet, wobei besonderes Augenmerk auf die Transparenz des Klassifikationsprozesses für den Benutzer des Systems gelegt wurde.

Die positiven Ergebnisse, die erste Tests des Klassifikationssystems am Beispiel von Kundenbriefen eines Versicherungsunternehmens lieferten, rechtfertigen seine Erprobung in der Praxis. Der Vorteil des Systems liegt in der großen Geschwindigkeit für die Verarbeitung von Dokumenten bei gleichzeitig hoher Genauigkeit der erzielten Ergebnisse. Die Möglichkeit zur Erweiterung der Datenbasis für das System verspricht ferner einen wachsenden Umfang von klassifizierbaren Dokumenten.

II Literaturverzeichnis

[App92] H.-J. Appelrath, Jochen Ludewig, *Skriptum Informatik – eine konventionelle Einführung*, B. G. Teubner Verlag, Stuttgart, **1992**, 2

[Boc94] H.-H. Bock, W. Lenski, M. M. Richter, *Information Systems and Data Analysis, Prospekts – Foundations – Applications*, Springer Verlag, Berlin, Heidelberg, **1994**

[Bro91] I. N. Bronstein, K. A. Semendjajew, *Taschenbuch der Mathematik*, B. G. Teubner Verlagsgesellschaft Stuttgart, Leipzig, **1991**, 25

[Fra92] W. B. Frakes, R. Baeza-Yates, *Information Retrieval, Data Structures and Algorithms*, P T R Prentice Hall, Englewood Cliffs, New Jersey, **1992**

[Gei95] K. Geiger, *Inside ODBC*, Microsoft Press, Unterschleißheim, **1995**

[Gil83] M. J. McGill, G. Salton, *Information Retrieval – Grundlegendes für Informationswissenschaftler,* McGraw-Hill Book Company GmbH, Hamburg, **1983**

[Gre89] G. Grewendorf, F. Hamm, W. Sternefeld, *Sprachliches Wissen, Eine Einführung in moderne Theorien der grammatischen Beschreibung*, Suhrkamp Taschenbuch Verlag, Frankfurt am Main, **1989**, 3

[Her98] K. Herrmann, *Automatische Klassifikation des Inhalts von Dokumenten nach Sachgebieten unter besonderer Berücksichtigung der Geschäftspost von Lebensversicherungsunternehmen*, Diplomarbeit, Universität Leipzig, **1998**

[Koh95] T. Kohonen, *Self-Organizing Maps*, Springer Series in Information Science, Springer-Verlag, Berlin, Heidelberg, **1995**, 30

[Len86] W. Lenders, G. Willée, *Linguistische Datenverarbeitung, Ein Lehrbuch*, Westdeutscher Verlag, Opladen, **1986**

[Pao89] M. L. Pao, *Concepts of Information Retrieval*, Libraries Unlimited, Inc., Engelwood, Colorado, **1989**

[Rus94] G. Ruske, *Automatische Spracherkennung. Methoden der Klassifikation und Merkmalsextraktion*, R.Oldenbourg Verlag, München, Wien, **1994**, 2

[Sal89] G. Salton, Automatic, *Text Processing, The Transformation, Analysis and Retrieval of Information by Computer*, Verlag Addison-Wesley Publishing Company, Bonn, **1989**

III Abbildungsverzeichnis

IV Formelverzeichnis

V Tabellenverzeichnis

VI Anlagen

VI.i Initialisierungsdatei für den Typisierer

```
// Zeichengruppen für Einzeichentypen
KLEINBUCHSTABE=abcdefghijklmnopqrstuvwxyzäöüáéíóúàèìòùß
GROSSBUCHSTABE=ABCDEFGHIJKLMNOPQRSTUVWXYZÄÖÜÁÉÍÓÚÀÈÌÒÙ
ZIFFER=0123456789
SATZENDE=.!?
KOMMA=,;:
IGNORIERT=()<>'"=/\%~+*

// Zeichengruppen für Mehrzeichentypen
// eZ_TYP = Zeichen mit denen der Typ beginnen darf
// wZ_TYP = Zeichen mit denen der Typ bis zum Ende
//   fortgesetzt werden darf

// Wort in Kleinschreibung
eZ_KLEINWORT= abcdefghijklmnopqrstuvwxyzäöüáéíóúàèìòù
wZ_KLEINWORT= abcdefghijklmnopqrstuvwxyzäöüáéíóúàèìòùß

// Wort in Großschreibung
eZ_GROSSWORT=ABCDEFGHIJKLMNOPQRSTUVWXYZÄÖÜÁÉÍÓÚÀÈÌÒÙ
wZ_GROSSWORT=abcdefghijklmnopqrstuvwxyzäöüáéíóúàèìòùß

// Nummer
eZ_NUMMER=0123456789
wZ_NUMMER=0123456789.,-/

// Abkürzung
eZ_ABKUERZUNG=ABCDEFGHIJKLMNOPQRSTUVWXYZÄÖÜÁÉÍÓÚÀÈÌÒÙabcdefgh
     ijklmnopqrstuvwxyzäöüáéíóúàèìòù
wZ_ABKUERZUNG=ABCDEFGHIJKLMNOPQRSTUVWXYZÄÖÜÁÉÍÓÚÀÈÌÒÙabcdefgh
     ijklmnopqrstuvwxyzäöüáéíóúàèìòùß-
```

VI.ii Beispiel für eine Begründungsdatei

```
Treffer im Benutzerlexikon für starke Korrelation
-------------------------------------------------
       Methoden: Übereinstimmungen in den Terme

starke Korrelation    Anzahl Syntax
0      1      0      0      0      0      0      0      0      0      0      0      0      1x     ist ~ verstorb


Treffer im Benutzerlexikon für schwache Korrelation
---------------------------------------------------

       Methoden: Inneres Produkt

schwache Korrelation  Anzahl Syntax


Treffer im Automatischen Lexikon
--------------------------------

       Methode: Inneres Produkt

Gewichte       Anzahl Syntax
0.000  2.751  3.751  1.887  0.000  0.000  0.945  0.000  0.940  0.000  0.000  0.950  0.000  0.000       1x     Anleg
0.000  0.917  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000       1x     Ruckfrag steh
0.000  0.917  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000       1x     Ruckfrag steh ich
0.000  0.000  0.000  0.944  0.000  0.000  0.000  0.000  0.000  0.000  0.000  1.900  0.799  0.000       1x     VN
0.000  1.834  0.938  0.000  0.000  0.000  0.945  0.000  0.000  0.000  0.000  0.000  0.000  0.000       1x     der Anleg
0.000  0.917  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000       1x     für Ruckfrag steh
0.000  0.917  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000       1x     für Ruckfrag steh ich
0.000  0.917  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000       1x     für ~ steh
0.000  2.751  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000       1x     ist ~ verstorb
0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  1.799       1x     rn
0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.799  0.000       1x     setz
0.000  1.834  0.938  0.000  0.000  0.000  0.945  0.000  0.000  0.000  0.000  0.000  0.000  0.000       1x     sie ~ Anleg
0.000  0.917  0.000  0.000  0.000  0.000  0.000  0.906  0.940  0.000  0.000  0.000  0.799  0.900       1x     steh
0.000  0.917  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000       1x     steh ich
0.000  3.668  0.000  0.000  0.000  0.000  0.000  0.906  0.000  0.000  0.000  0.000  0.000  0.000       1x     verstorb
```

Klassenvektoren mit Abständen
---

starke Korrelation	0	1	0	0	0	0	0	0	0	0	0	0	0	0
schwache Korrelation	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Abstände	0.000	19.257	5.627	2.831	2.814	0.000	2.835	1.811	1.881	0.000	0.000	2.851	2.397	2.699

Verfahren zur Auswahl der Kandidaten

Methode: Angehörige eines Konfidenzintervalls

Grenzwert für schwache Korrelation: 0
Grenzwert für Abstände: 19.257

Klassenvektoren mit Kandidaten
---

starke Korrelation	0	1	0	0	0	0	0	0	0	0	0	0	0	0
schwache Korrelation	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Abstände	0.000	19.257	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Ergebnis der Klassifikation

Zugeordnete Klasse: 001
Prioritätsstufe: 3
Widerspruchsanalyse: EKE

Erklärung

„Ich versichere, daß ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.“

Leipzig, 19. Januar 1998 Andre‘ Seharsch