DIPLOMARBEIT

Analyse, Auswertung und Visualisierung umfangreicher textueller Daten mit dem Schwerpunkt der elektronischen Bürgerzufriedenheitserhebung als Aspekt des e-Government

zur Erlangung des akademischen Grades
Diplomingenieur
(Dipl.-Ing.)

ausgeführt am
Institut für Rechnergestützte Automation
Forschungsgruppe Industrial Software

der Technischen Universität Wien

unter der Anleitung von
Univ.-Prof. Dipl.-Ing. Dr. Thomas Grechenig

durch
Frau Elham Hedayati-Rad
Rosensteingasse 60-62, A-1170 Wien

Wien, 28. Juni 2008

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen nicht benützt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Alle Internet-Quellen und Internet-Verweise sind in Form von URLs angegeben und wurden zuletzt im Anfang Juni 2008 überprüft.

Wien, 28. Juni 2008

Danksagung

Meinen besonderen Dank möchte ich an dieser Stelle Herrn Univ. Prof. Dr. Thomas Grechenig, meinem Betreuer Herrn Dipl.-Ing. Gerald Fischer sowie Herrn Roman Trabitsch in seiner Funktion als Zweitbetreuer meiner Diplomarbeit aussprechen, die mir mit Ihren zahlreichen Ratschlägen stets hilfreich bei der Entstehung dieser Arbeit zur Seite gestanden sind.

Wesentlich zur Vollendung dieser Diplomarbeit haben auch die Antworten von Herrn Univ. Prof. Dr. Dieter Merkl beigetragen, dem ich hiermit für seine Bereitschaft danken möchte, meine Fragen zu beantworten.

Ganz besonders möchte ich hiermit meinen Eltern danken, die mir ein Studium in Österreich ermöglicht haben und die mich stets ermuntert haben, den Mut nicht sinken zu lassen und auch gegen die Sprachbarriere anzukämpfen. Meine Brüder Hadi und Mehdi Hedayati Rad haben in diesen vergangenen Jahren des Studiums auch einen wichtigen Teil zum Gelingen meines Projektes „Studium“ beigetragen, indem sie immer da waren, wenn Not am Mann war. Ich möchte allen Beteiligten danken, die mir direkt und indirekt beim Fortschreiten meines Studiums und bei der Entstehung meiner und Diplomarbeit beigestanden sind.

Danken möchte ich auch Herrn Dipl.-Ing. Andreas Pongratz für wertvolle Hinweise während meines Studiums und auch zuletzt bei meiner Diplomarbeit.

Inhaltsübersicht

Eidesstattliche Erklärung I
Danksagung II
Inhaltsübersicht III
Inhaltsverzeichnis IV
Tabellenverzeichnis VII
Abbildungsverzeichnis VIII
Abkürzungsverzeichnis IX
Kurzfassung X
Abstract XII
1 Einleitung 1
2 Empirische Sozialforschung 5
3 Meinungsforschung 28
4 Datenerhebung, -analyse und Interpretation 36
5 Qualitative Datenanalysesoftware 42
6 Analyse unstrukturierter Texte 49
7 Textkategorisierung 92
8 Praktische Anwendung bei einem Projekt 95
9 Zusammenfassung und Ausblick 105
10 Literaturverzeichnis 107

Inhaltsverzeichnis

Eidesstattliche Erklärung I
Danksagung II
Inhaltsübersicht III
Inhaltsverzeichnis IV
Tabellenverzeichnis VII
Abbildungsverzeichnis VIII
Abkürzungsverzeichnis IX
Kurzfassung X
Abstract XII
1 Einleitung 1
1.1 Motivation für diese Arbeit 1
1.2 Forschungsumfeld 2
1.3 Zielsetzung 2
1.4 Aufbau der Arbeit 3
2 Empirische Sozialforschung 5
2.1 Definition 5
2.2 Ziele der empirischen Sozialforschung 6
2.3 Anwendungsbereich 6
2.4 Methoden 7
2.4.1 Quantitative Methoden der Datenerhebung. 8
2.4.2 Qualitative Methoden der Datenerhebung 11
2.5 Ablauf eines empirischen Forschungsprojektes 18
2.6 Abgrenzung 20
2.6.1 Data-Mining 21
2.6.2 Der Prozess des Data-Mining 23
2.6.3 Methoden und Techniken des Data-Mining 26
3 Meinungsforschung 28
3.1 Was ist Meinungsforschung 28
3.2 Grundlagen und praktische Anwendung 28
3.3 Hauptziele der Meinungsforschung 29
3.4 Arten von Interviews 29
3.5 Arten von Fragen und Antworten 33
4 Datenerhebung, -analyse und Interpretation 36
4.1 Datenerhebung 36
4.2 Datenauswertung 38
4.2.1 Quantitative Daten 38

4.2.1.1 Datenaufbereitung ... 38
4.2.1.2 Datenanalyse ... 39
4.2.2 Qualitative Daten ... 40
4.3 Interpretation ... 41
5 Qualitative Datenanalysesoftware ... 42
5.1 Ziele von QDA Software ... 42
5.2 Vor- und Nachteile der Produkte ... 43
5.2.1 Vorteile ... 43
5.2.2 Nachteile ... 43
5.3 Softwarelösungen ... 44
5.3.1 STATISTICA Text-Miner ... 44
5.3.2 SPSS ... 46
5.4 Zusammenfassung ... 48
6 Analyse unstrukturierter Texte ... 49
6.1 Information Retrieval (IR) ... 51
6.2 Indexierung ... 51
6.2.1 Manuelle Indexierung ... 52
6.2.2 Computergestützte Indexierung, vollautomatische Indexierung 53
6.3 Verfahren der automatischen Indexierung ... 54
6.3.1 Volltextinvertierung ... 54
6.3.2 Statistische Verfahren ... 57
6.3.3 Termgewichtungsmethode ... 60
6.3.4 Informationslinguistische Verfahren ... 66
6.3.5 Regelbasierte Verfahren ... 68
6.3.6 Wörterbuchbasierte Verfahren ... 69
6.3.7 Lexikonbasierte Morphologieanalyse ... 70
6.3.8 Graphemisch-phonologische Verfahren ... 72
6.3.9 Art der Reduktionform ... 73
6.3.10 Begriffsorientierte Verfahren ... 76
6.3.11 Mustererkennungsverfahren ... 77
6.4 Grundlegenden Verfahren der automatischen Schlagwortextraktion ... 79
6.4.1 Eliminieren der Stoppwörter ... 79
6.4.2 Stemming Algorithmus zur Grundformreduktion ... 80
6.4.3 Porter Stemmer Algorithmus zur Grundform-reduktion ... 81
6.4.4 Over Stemming und Under Stemming ... 85
6.4.5 LOVINS Algorithmus zur Grundformreduktion ... 86
6.4.6 Wortstammbildung nach dem N-Gram Verfahren ... 89
6.4.7 Korpusbasierte Verfahren ... 90
6.5 Zusammenfassung ... 91

7 Textkategorisierung 92
7.1 Definition 92
7.2 Wozu Textkategorisierung? 92
7.3 Art der Kategorisierung 93
7.4 Schritte der Textkategorisierung 93
7.5 Anwendung der Textkategorisierung 94
8 Praktische Anwendung bei einem Projekt 95
8.1 Das Projekt 95
8.2 Basisdaten 95
8.3 Beispielhafte Analyse 96
8.3.1 Basisdaten 96
8.3.2 Analyse 98
8.3.2.1 Tokenisierung (Word-Splitting) 98
8.3.2.2 Vereinheitlichung 98
8.3.2.3 Eliminierung der Stoppwörter 100
8.3.2.4 Stemming 101
8.3.2.5 Analyse Worthäufigkeiten 102
8.4 Ergebnis 104
9 Zusammenfassung und Ausblick 105
9.1 Erfolgsfaktoren 105
9.2 Offene Fragen und Ausblick 105
10 Literaturverzeichnis 107
10.1 Literatur 107
10.2 Internetquellen 109

Tabellenverzeichnis

Tabelle 1: Formen der Datenerhebung....37
Tabelle 2: QDA-Produkte....47
Tabelle 3: Ergebnisse der Index....56
Tabelle 4: Alphabetisch Sortierung....56
Tabelle 5: Invertierte Datei....56
Tabelle 6: Term- und Dokumenthäufigkeit....64
Tabelle 7: Termgewichtung....65
Tabelle 8: Invertierter Index....65
Tabelle 9: Flexionsanalyse (nach Lezius,1995)....72
Tabelle 10: Reduktionsalgorithmenund ihre Wirkungensweisen nach KUHLEN....75
Tabelle 11: Beispiel für Patternklassen aus FIPRAN....78
Tabelle 12: Regeln der Porter Stemmer Algorithmus....82
Tabelle 13: Porter Stemmer Algorithmus für die deutsche Sprache....84
Tabelle 14: Bedingungen und Wortendungen nach Lovins....87
Tabelle 15: Ähnlichkeitsmatrix....90
Tabelle 16: Kookkurrenzdaten....90

Abbildungsverzeichnis

Abbildung 1: Textanalytischer Ansätze 17
Abbildung 2: Phasen des Forschungsablaufs 18
Abbildung 3: Schritte im Data Mininag-Prozess (nach Fayyad, U. et al. 1996) 25
Abbildung 4: Zuordnung von Data-Mining Methode zu Aufgaben 27
Abbildung 5: Forschungsphasen nach der Definition der Forschungsfrage 36
Abbildung 6: Entscheidungsstärke bedeutsamer Begriffe 58
Abbildung 7: Termhäufigkeitsverteilung 63
Abbildung 8: Arten von Stoppwörtern 80
Abbildung 9: Auszug Basisdaten 96
Abbildung 10: gesplittete Daten 98
Abbildung 11: alphabetisch sortierter Volltextindex 99
Abbildung 12: Normalisierung der Umlaute und des ß 100
Abbildung 13: Stoppwörter markieren 100
Abbildung 14: Stopwörter endgültig entfernt 101
Abbildung 15: Grundformreduktion und Synonymbildung 102
Abbildung 16: Index mit Worthäufigkeit (absteigeend nach Häufigkeit sortiert). .. 103
Abbildung 17: Häufigste Worte in den Antworten 104

Abkürzungsverzeichnis

AI	Artificial Intelligence / Künstliche Intelligenz
CONDOR	Communikation in Natürlicher Sprache mit Dialogorientiertem Ret
COPSY	Context Operator Syntax
CTX	Computergestützte Texterschließung
DM	Data Mining
DMM	Deutsch Malaga Morphologie
DOKFREQ	Dokumentenfrequenz
FIPRAN	FIrmen und PRodukt Analyse
FREQ	Termfrequenz
HP	Homepage
IE	Information Extraction / Informationsextraktion
IR	Information Retrieval
KNN	künstlichen Neuronalen Netzen
o.J.	ohne Jahr
o.S.	ohne Seiten
o.V.	ohne Verfasser
QDA	Qualitative Daten Analyse
TM	Text Mining

Kurzfassung

In der vorliegenden Arbeit wird aufbauend auf der theoretischen Basis der empirischen Sozialforschung ein Einblick in die Problematik der Analyse unstrukturierter Texte im Umfeld von Bürgerzufriedenheitserhebungen erarbeitet.

Gerade im Bereich des e-Government ist die Partizipation der Bürger wünschenswert und wichtig, weil nur bei Vorliegen von Feedback der Bürger eine positive verbesserung Veränderung verwaltungstechnischer und organisatorischer Abläufe möglich wird. Ein wichtiges Mittel für die Erhebung der Bürgermeinung sind Umfragen und Interviews.

Die Zielsetzung dieser Arbeit ist es, einen Überblick über die relevanten Gebiete der empirischen Sozialforschung und über die technischen Aspekte der Auswertung von Umfragen zu schaffen, wobei neben der Erhebung und Analyse quantitativer Daten ganz besonders die Erhebung und Analyse qualitativer Daten eine wichtige Rolle spielt. Gerade bei Bürgerbefragungen sind häufig Antworten in Form unstrukturierter Texte zu finden, deren Auswertung mit hohem Zeitaufwand und somit hohen Kosten verbunden ist.

Im technischen Teil dieser Arbeit werden die grundlegenden Algorithmen vorgestellt, die notwendig sind, um qualitative Daten in Form von unstrukturierten Texten analysieren zu können, wobei ein Überblick über aktuelle Softwarelösungen gegeben wird.

Abschließend wird anhand einer, im Rahmen der Diplomarbeit, ausgewerteten Bürgerzufriedenheitsumfrage die Anwendung der vorgestellten Algorithmen demonstriert.

Schlagwörter: *Empirische Sozialforschung, Qualitative und Quantitative Methoden, Interviews, QDA Software, Bürgerzufriedenheitumfrage (Fragebogen), Datenerhebung, Datenauswertung, Schlagwortextraktion, Analyse unstruktierter Texte, Textkategorisierung*

Abstract

This thesis presents the problems of the analysis of unstructured text in the context of surveys regarding citizen satisfaction. We base this on the theoretical basis of empirical social research.

The participation of citizens in the domain of e-Government is especially desirable and important as only the availability of feedback from citizens makes a positive change in administrative and organizational procedures feasible. Polls and interviews are two important means to conduct citizen opinion surveys.

The objective of this thesis is to provide an overview of the relevant fields of empirical social research and the technical aspects of the evaluation of polls, whereas the surveying and evaluation of qualitative data plays a decisive role next to the quantitative of the same. A common occurrence in citizen interviews are replies in the form of unstructured text whose evaluation necessitates a high expenditure of time and therefore cost.

The technical section of this thesis presents the necessary fundamental algorithms to allow the analysis of qualitative data in the form of unstructured text together with an overview on recent software solutions.

Finally, the application of the presented algorithms is demonstrated within the scope of this thesis through an actual large-scale citizen satisfaction survey.

1 Einleitung

1.1 Motivation für diese Arbeit

Seit Beginn des 20. Jahrhunderts wird in den verschiedensten Bereichen wie Politik, Medien, Wirtschaft und Sozialforschung die Analyse, Auswertung und Visualisierung von Daten eingesetzt, um Trends festzustellen und vorherzusagen. Einer der ersten nennenswerten Einsätze dieser Verfahren ist in den USA um ca. 1920 im Bereich der Wahlprognosen zu verzeichnen.

So gut wie alle Bereiche der empirischen Sozialforschung haben dringenden Bedarf nach raschen und möglichst exakten Auswertungsverfahren. Gerade im Bereich der qualitativen Analyse ist die Balance zwischen Aufwand und exakten Ergebnissen sehr schwer herzustellen, weil bei Verringerung des Aufwandes eine starke Verschlechterung der Exaktheit festgestellt werden kann.

Es zeigt sich anhand durchgeführter Recherchen, dass die Analyse im Rahmen der empirischen Sozialforschung und insbesondere im Bereich der Meinungs- und Marktforschung nur mit entsprechender Software effizient möglich ist und damit ist die Brücke zur Informatik geschlagen, die erst Datenanalyse im großen Umfang möglich macht.

Aus den gewonnenen, analysierten Daten kann mit Hilfe von statistischen Methoden einfach und rasch eine Aussage gewonnen werden, deren Relevanz aber erst durch geeignete Interpretationsmethoden erreicht wird.

Mit zunehmender Datenmenge ist der vermehrte Einsatz effizienter Hard- und Softwarelösungen unverzichtbar, weil jede empirische Sozialforschung auch eine möglichst gute Kosten-Nutzen-Rechnung als Ziel verfolgt bzw. nur unter Einhaltung vorhandener Budgets durchführbar ist.

1.2 Forschungsumfeld

Seit dem 20. Jahrhundert hat sich die empirische Sozialforschung im Bereich quantitativer und qualitativer Daten entwickelt in deren Mittelpunkt hauptsächlich Markt- und Meinungsforschung stehen.

Empirische Sozialforschung findet man in Universitäten, unabhängigen Forschungsinstituten, in der akademischen Sozialforschung und besonders in Markt-, Meinungs- und Sozialforschung, Unternehmen, Regierungen und politischen Parteien.

Die Forschung sucht nach effizienten Methoden zur qualitativen Analyse unter Einbeziehung statistischer, syntaktischer und semantischer Verfahren, wodurch eine wesentlich exaktere qualitative Analyse ermöglicht werden soll. Das Hauptproblem bei qualitativen Analysen stellt das Verstehen von Bedeutungen von Wörtern dar, weil eine rein statistische Analyse ohne Berücksichtigung des Kontextes (Syntax, Semantik) nur minderwertige Ergebnisse liefern kann. Es stellt sich die Frage, wie gut es überhaupt möglich ist, mit Hilfe von IT semantische Analysen in einem hohen Maß an Korrektheit durchzuführen.

Verschiedene Softwareanbieter versuchen laufend Softwarelösungen mit besserer Qualität zu geringeren Kosten zu erstellen, um schnell möglichst korrekte Ergebnisse liefern zu können.

1.3 Zielsetzung

Das primäre Ziel dieser Diplomarbeit ist es, einen zusammenfassenden Überblick über den aktuellen Stand der Entwicklungen im Bereich der Datenanalyse und der Auswertung qualitativer Daten im Umfeld der empirischen Sozialforschung zu geben. Weiteres versucht diese Arbeit auf Basis der aktuellen Softwarelösungen und einer Vorstellung der verwendeten Verfahren und Methoden, einen Einblick in die Analyse unstrukturierten Textes zu bieten und diese Verfahren anhand eines Beispieles zu erläutern. An dieser Stelle sollte festgerhalten werden, daß es hauptsächlich um die qualitative Auswertung der Daten geht und nicht um quantitative Verfahren.

Die Analyse quantitativer Daten wird in dieser Arbeit so gut wie völlig außer Acht gelassen, weil es sich dabei primär um die Anwendung statistischer Verfahren handelt, die nicht zentraler Gegenstand dieser Arbeit sind

1.4 Aufbau der Arbeit

In Kapitel 2 dieser Arbeit wird der Begriff der empirischen Sozialforschung erläutert und es wird geklärt, was die theoretischen Grundlagen und die Anwendungsbereiche der empirischen Sozialforschung sind. Weiters werden die Hauptziele der Meinungsforschung erörtert: Auswertung und Interpretation, um eine Aussage zu produzieren, wobei auch die Datenpräsentation in Form grafischer Auswertungen zu berücksichtigen ist . Danach wird der Versuch einer Abgrenzung zwischen empirischer Sozialforschung und den Bereichen Data-Warehousing bzw. Data-Mining unternommen.

Kapitel 3 geht ein auf die Meinungsforschung als Teilgebiet der empirischen Sozialforschung, wobei die Bedeutung der Art der Interviews ebenso wie verschiedene Arten der Fragen und Antworten, die bei qualitativen Methoden verwendet werden, näher beleuchtet werden.

In Kapitel 4 wird der Prozess von der Datenerhebung bis zur Interpretation der analysierten Daten vorgestellt, der so gut wie jedem Forschungsprojekt grundlegend gemeinsam ist, wobei an dieser Stelle eine exakte Differenzierung zwischen quantitativen und qualitativen Daten und deren Analyse erarbeitet wird.

In Kapitel 5 werden bestehende Softwarelösungen und existierende Produkte im Anwendungsbereich der Analyse qualitativer Daten analysiert, wobei die Produkte und deren Vor- und Nachteile überblicksweise vergleichend vorgestellt werden.

Kapitel 6 stellt die Analyse unstrukturierter Daten und die dafür verwendeten Methoden und Algorithmen im Detail dar.

In Kapitel 7 wird das Problem der Textkategorisierung dargestellt, das inhaltlich zwar andere Ziele verfolgt als die Analyse qualitativer Daten der Markt- und Meinungsforschung, das aber technisch betrachtet viele Gemeinsamkeiten aufweist.

In Kapitel 8 wird anhand eines im Rahmen der Diplomarbeit durchgeführten Projektes die Analyse unstrukturierter Texte beispielhaft anhand ausgewählter Fragestellungen demonstriert.

In Kapitel 9 erfolgen eine Zusammenfassung der Arbeit und ein Ausblick auf die möglichen künftigen Entwicklungen sowie eine Abschätzung der möglichen Ergebnisse vollautomatischer Analyseverfahren für unstrukturierte Texte aus subjektiver Sicht der Autorin.

2 Empirische Sozialforschung

2.1 Definition

In unserem Alltagsleben ist im Lauf der letzten Jahrzehnte die Verwendung von Ergebnissen der empirischen Sozialforschung viel stärker geworden als wir denken: z.B. wird beim Einkauf (sehr oft unbewusst) auf Medien und Werbung geachtet, die man als Ergebnisse von Marktforschung betrachten kann. In der Politik und in den politischen Parteien werden kaum Schritte ohne Meinungsforschung gesetzt und es wird schon gar nicht in einen Wahlkampf gezogen ohne vorher die modernen Auguren der Meinungsforschungsunternehmen befragt zu haben. In Parlamenten und Regierungen demokratischer Staaten fließt die öffentliche Meinung teilweise in Form von Umfragen und deren Ergebnissen ein und resultiert demzufolge teilweise auch in der Gesetzgebung und in der öffentlichen Verwaltung. Allgemein sieht man in unserem heutigen Leben besonders im wirtschaftlichen und politischen Umfeld, dass empirische Sozialforschung ein sehr wichtiger und bedeutender Aspekt geworden ist.[1]

Grundlegend kann man sagen: Empirische Sozialforschung ist die systematische Erfassung und Deutung sozialer Tatbestände und Sachverhalte.

Empirisch bedeutet erfahrungsgemäß. Grundlage der empirischen Sozialforschung ist also, was wir durch unsere Sinne als Erfahrung sammeln können.

Systematisch bedeutet, dass die Erfahrung der Umwelt nach Regeln zu geschehen hat. Soziale Tatbestände: beobachtbares menschliches Verhalten, von Menschen geschaffene Gegenstände sowie durch Sprache vermittelte Meinungen, Informationen über Erfahrungen, Einstellungen, Werturteile, Absichten.[2]

[1] vgl. [Atte00], S. 3
[2] vgl. [Atte00], S. 3

Das menschliche Verhalten und gesellschaftliche Aspekte bedürfen aufgrund unterschiedlichster Anforderungen häufig einer wissenschaftlichen Untersuchung, die sich aus verschiedenen Methoden und Techniken zusammensetzt. Die Auseinandersetzung mit sozialen Fragestellungen mit den dazugehörigen wissenschaftlichen Methoden ist der Gegenstand der empirischen Sozialforschung. Sie analysiert die Hintergründe und die sozialen Zusammenhänge[3].

2.2 Ziele der empirischen Sozialforschung

Empirische Sozialforschung ergründet die Umstände der Natur oder der Gesellschaft und versucht, Aussagen über das Zusammenwirken von Menschen zu ermitteln und zu überprüfen. Somit stellt empirische Sozialforschung das wichtigste Werkzeug zur Aufstellung und zur Überprüfung von Annahmen im soziologischen Umfeld dar. Wann immer im Bereich menschlichen Zusammenlebens bzw. Verhaltens Ursachen und deren Wirkung untersucht werden, bedient man sich der der Methoden der empirischen Sozialforschung.[4]

Die Empirische Sozialforschung verfolgt als wichtigstes Ziel[5], die Ereignisse der realen Welt objektiv zu beschreiben, möglichst im Großen und Ganzen Regeln zu finden, die gültig sind, die durch die Erlebnisse in der realen Welt erklärt werden können und die Klassen von Ereignissen voraussagen können.

2.3 Anwendungsbereich

Empirische Sozialforschung wird unter anderem in folgenden Bereichen angewendet[6]:

- Soziologie: alle Bereiche der Soziologie (z.B. Motivationsforschung, Erforschung der Bevölkerungsentwicklung, …)

3 vgl. [Atte00], S. 5
4 vgl. [Enge04], S. 1-2
5 vgl. [ScHiEs05], S. 6-7
6 vgl. [Diek06], S. 19-21

- Politik: Zentralbereich ist die Wahlforschung (Wahlprognosen, Wählerpotentiale, …) im großen Umfang in der Marktforschung, wird auch bei Politischen Meinungsumfragen verwendet.
- Wirtschaft: Kundenzufriedenheitsanalysen, Trendforschung, …
- Sozialpsychologie
- Pädagogik und Erziehungswissenschaft: Erziehungsmethoden, lernpsychologische Forschung, …
- Geographie: Raumplanung, Stadtentwicklung, …
- Geschichtswissenschaft
- Ethnologie
- Rechtswissenschaften

Aus dieser umfangreichen Liste der Anwendungsgebiete der empirischen Sozialforschung scheint der Schluss zulässig, dass empirische Sozialforschung unser tägliches Leben stark durchdringt und dass die empirische Sozialforschung aus dem Alltag nicht mehr wegzudenken ist.

2.4 Methoden

Allen Anwendungsbereichen der empirischen Sozialforschung ist die Verwendung ähnlicher Methoden gemein, die angewendet werden, um Daten zu erheben. Die Datenerhebung stellt eine grundlegende Voraussetzung für jeden weiteren Schritt im Rahmen jedes Forschungsprojektes im Bereich der empirischen Sozialforschung dar. Erst durch die Datenerhebung, bei der von einer möglichst zufällig ausgewählten Anzahl von Stichproben Daten erhoben worden sind, wird ein Schluss auf die Gesamtheit möglich, der in Form von Analyse und anschließender Interpretation durchgeführt wird und unter anderem auf statistischen Methoden basiert.

Die Methoden der empirischen Sozialforschung gewährleisten verlässliche und gültige Daten. Durch die Auswertung dieser Daten kann man Aussagen über das untersuchte Objekt treffen. Die Zuverlässigkeit und Gültigkeit sind zwei wichtige Faktoren für die Bewertung der Methoden der empirischen Sozialforschung.

Je nach der konkreten Anforderung bieten die verschiedenen Methoden der Datenerhebung Vorteile oder Nachteile. Es liegt in der Verantwortung der Durchführenden, die jeweils beste und möglichst optimale Methode der Datenerhebung zu suchen und zu finden, wobei sehr häufig verschiedene Methoden kombiniert werden, um bestmögliche Ergebnisse zu erreichen. Um einen kurzen Überblick über die Vor- und Nachteile der verschiedenen Methoden zu schaffen, ist es sinnvoll, die Methoden gegenüberzustellen bzw. zu gliedern.

Ähnlich wie in den Naturwissenschaften existiert auch in der Sozialwissenschaft eine Reihe von Instrumenten mit denen die Sozialforscher das Aufgabengebiet bearbeiten und explorieren. Eine der wichtigsten Unterscheidungen der verwendeten Methoden ist die Unterscheidung in qualitative und in quantitative Methoden. Die verwendeten Instrumente lassen sich jeweils in diesen zwei Methoden zuordnen, wobei es hier zur Überschneidungen kommen kann (ein Instrument wird in beiden Methoden verwendet).

Aufgrund der Komplexität der zu erforschenden Zusammenhänge und auf Basis der jeweiligen Aufgabenstellung lässt sich schlüssig ableiten, dass die Wahl der Methoden erst aus der Problemstellung vorgenommen werden kann und darf und es ergibt sich, dass nicht jede Methode für jede Aufgabenstellung geeignet ist.

2.4.1 Quantitative Methoden der Datenerhebung.

Quantitative Methoden der Datenerhebung versuchen mit Zahlen und Statistiken die soziale Wirklichkeit zu erklären. Diese Gruppe der Methoden hat die Funktionen und Einzelheiten der Realität, die bei einer Untersuchung interessant sind, möglichst exakt und genau zu beschreiben bzw. darzustellen. Quantitative Methoden kontrollieren und überprüfen Hypothesen und Theorien. Quantitative Methoden in der Sozialwissenschaft kann man zusammenfassen unter den Stichwörtern „Zählen“, „Urteilen“, „Testen“, „Befragen“, „Beobachten“ und Physiologischen Messungen. Im Allgemeinen wird bei der Datenerhebung eine Kombination vorgenommen, wie z.B. gleichzeitiges Beobachten und Zählen oder Befragung und Schätzen oder Testen und Messen. Bei jeder Art der Datenerhebung wird die Art des Untersuchungsgegenstandes und der Untersuchungsteilnehmer betrachtet, weil neben den erhobenen Daten

auch andere Informationen möglicherweise von Bedeutung sind, wie z.B. finanzielle oder zeitliche Aspekte.[7]

Zählen: Mann kann zählen indem nur wenige beschreibende Merkmale herausgegriffen werden, um ihre Gleichheit in Bezug auf die Ausprägung von Merkmalen zu definieren. Die Aufgabe der Forschung ist es dabei, die Merkmale nach deren thematischer Relevanz zu ordnen. Die Voraussetzung für sinnvolles Zählen ist es zu wissen, was man vergleichen und zählen soll.

Urteilen: Für die Humanwissenschaften ist der Mensch nicht nur das zentrale Thema sondern parallel dazu ist der Mensch auch ein sehr wichtiges Erhebungsinstrument. Das Urteilen eignet sich dazu, schwer zu messende Eigenschaften mit Hilfe der menschlichen Urteilsfähigkeit zu ergründen. Obwohl menschliche Urteile subjektiv sind, nützt dieses Messverfahren die menschliche Urteilsfähigkeit auf verschiedene Arten: Rangordnung, Dominanz-Paarvergleiche oder Ähnlichkeits-Paarvergleiche. Eine sehr häufige Art der Anwendung basiert auf der direkten, quantitativen Einstufung von Urteilsobjekten oder einzelnen Objektmerkmalen. Sie verwenden diese Erhebungsart als Schätzverfahren. Für menschliches Urteilen sind zwei Punkte sehr wichtig: Es muss gefragt werden, welche Urteilsleistung für die Fragenstellungen verlangt werden soll und weiters muss geklärt werden, wie die weitere Verarbeitung der erhobenen Daten erfolgen soll.

Testen: Man versteht unter Testen statistische Prüfverfahren, bei denen zur Untersuchung eines (Persönlichkeits-)Merkmales anhand einer Checkliste eine Hypothese überprüft wird. Zur Formulierung von Testfragen und Testaufgaben muss man einzelne Testaufgaben zu Testskalen zusammenstellen und überprüft bei der Datenerhebung die Entsprechung der untersuchten Merkmale anhand der aufgestellten Skala. Durch den Einsatz von Skalen kann man mit Verzerrungen und Verfälschungen umgehen bzw. diese teilweise verhindern.

[7] vgl. [Döri02], S. 137-138

Befragen ist die Methode, die in der empirischen Datenerhebung am meisten verwendet wird. Sozialwissenschaftler schätzen, dass ungefähr 90% aller Daten mittels Befragungen erreichbar sind. Die Hauptvertreter der Befragungsmethoden sind die mündliche Befragung in Form von Interviews und die schriftliche Befragung mit Fragebögen. Eine der wichtigsten Unterschiede zwischen diesen beiden Methode liegt in der Erhebungssituation, also darin, ob der Befragte selbst die Fragen liest und die Antworten vermerkt, oder ob dies durch einen Interviewer durchgeführt wird. Bei persönlich durchgeführten schriftlichen Befragungen weiß man, wer den Fragenbogen ausgefüllt hat, bei postalisch durchgeführten Befragungen weiß man nicht, wer im Endeffekt den Fragebogen ausgefüllt hat. Auch bei persönlich durchgeführten Interviews gibt es einige Probleme, so kann es z.B. sein, dass sich der Interviewer für das Thema interessiert oder nicht. Bei jedem Verfahren gibt es Vor- und Nachteile und die Entscheidung, ob eine Befragung mündlich oder schriftlich durchzuführen ist, hängt von den Anforderungen ab. Es spielt eine Rolle, welche Personen befragt werden sollen, wie der finanzielle Rahmen der Befragung ist und auf welche späteren Auswertungsmethoden Wert gelegt wird.

Beobachten: *In einem sehr allgemeinen Begriffsverständnis beruht somit jede Datenerhebung auf Beobachtung , wenn dezidiert von Beobachtungsmethoden die Rede ist, ist damit eine engere Begriffsfassung gemeint. Laatz (1993, S.169) definiert: Beobachtung im engeren Sinne nennen wir das Sammeln von Erfahrungen in einem nichtkommunikativen Prozess mit Hilfe sämtlicher Wahrnehmungsmöglichkeiten*[8]. Keine Datenerhebungsmethode kann auf Beobachtungen verzichten. Alltagsbeobachtungen und wissenschaftliche Beobachtungen sind zwei Möglichkeiten, die von den Sozialwissenschaften verwendet werden. Wissenschaftliche Beobachtungen haben festgelegte Regeln (Standardisierung und Intersubjektivität) und die Beobachtung und ihre Ergebnisse müssen jederzeit überprüft werden können. Sie können qualitative Daten produzieren, die bei der statistischen Hypothesenprüfung spezifisch sind.

8 [Döri02], S. 262

Physiologische Messung: Als Datenerhebungsmethode für Sozial- und Humanwissenschaften werden besonders in der Psychologie, der Biologie und der Medizin physiologische Messungen verwendet. Im Rahmen dieser quantitativen Datenerhebungsmethode werden methodische Messungen von Indikatoren unter Berücksichtigung der jeweiligen somatischen Prozesse durchgeführt. Die meisten physiologischen Messverfahren werden heutzutage mit hoch entwickelten Messgeräten durchgeführt wodurch es – bei entsprechender Eichung bzw. Justierung - zu höchst objektiven Messergebnissen kommt.

2.4.2 Qualitative Methoden der Datenerhebung

Bei qualitativen Methoden der Datenerhebung wird versucht, die subjektive Realität mit verbalen Mitteln zu beschreiben. Daten, die nicht einfach auf numerische Daten abgebildet werden können und nur in Form von Beschreibungen erfasst werden können, stellen den Arbeitsbereich der qualitativen Methoden dar. Bei den qualitativen Methoden der Datenerhebung ist nach der reinen Erhebung eine Interpretation unerlässlich und es wird versucht, die verbalen Beschreibungen durch Anwendung verschiedener Verfahren im Endeffekt doch auf numerische – und damit wesentlich leichter analysierbare - Werte abzubilden. Es wird versucht, die Meinung von Menschen über gewisse Situationen oder Umstände in Erfahrung zu bringen und anhand der Antworten auch über die Motive und die subjektive Bedeutung der Dinge für die Befragten mehr in Erfahrung zu bringen. Es geht also auch ganz speziell um den persönlichen Kontext der Befragten zum Zeitpunkt der Befragung.

In der empirischen Sozialforschung wird bei qualitativen und quantitativen Methoden nicht die Art der verwendeten Methoden beurteilt, sondern es werden vor allem aufgrund der Ergebnisse der Stellenwert und der Informationsgehalt beurteilt. Bei den qualitativen Datenerhebungsmethoden wird über den Weg von einer Hypothese zu einer Theorie versucht, anhand von Erhebungen Rückschlüsse zu ziehen und eventuell auch Verbesserungspotential ermittelt.

Die wichtigsten Techniken qualitativer Methoden zur Datenerhebung sind:[9]

Qualitative Befragung:

Befragung ist wie bei den quantitativen Methoden als Standardinstrument der empirischen Sozialforschung bei der Ermittlung von Wissen, Fakten, Realität, Meinungen und Einstellungen im Einsatz. Befragung bedeutet eine Kommunikation zwischen zwei oder mehr Personen und durch verbale Stimuli in Form von Fragen werden Antworten als Reaktionen induziert. Das Ziel einer Befragung ist nicht eine Zusammenfassung sozialen Verhaltens, sondern ausschließlich die Erhebung verbaler Aussagen. z.B. möchte eine politische Partei wissen, mit wie vielen Stimmen sie bei einer bevorstehenden Wahl rechnen kann, wenn gewisse politische Ziele in Angriff genommen werden, oder ein Unternehmer möchte über seine Produkte bzw. die Zufriedenheit seiner Kunden mit seinen Produkten mehr in Erfahrung bringen. Seit den 1930er Jahren hat die Markt und Meinungsforschung in verschiedenen Bereichen der Gesellschaft eine rasante Entwicklung eingeschlagen. Meinungsumfragen werden in der Öffentlichkeit oft mit empirischer Sozialforschung gleichgesetzt. Im heutigen Leben gibt es sehr selten Gespräche im Alltag, die nicht irgendwie durch Fragen und Antworten bzw. Gegenfragen geprägt sind. Alltägliche Befragungen und wissenschaftliche Befragungen sind als Informationsaustausch kategorisierbar. Ein wesentlicher Unterschied zwischen alltäglichen Befragungen und wissenschaftlichen Befragungen ist die Tatsache, dass alltägliche Befragung ein sozialer Prozess ist. Mindestens zwei Personen stehen einander bei einer Befragung gegenüber und bei jeder Befragung ist ein Interesse erkennbar. Alle Befragungen sind gezielt, verfolgen also ein oder mehrere Ziele. Das Ziel des Fragestellers ist die Information, die in Verbindung mit seiner Frage steht. Sprache, Fragesituation und bei der Befragung anwesende (andere) Personen, Zeitdruck und die Situation der Befragung sind Faktoren die berücksichtigt werden müssen. Bei wissenschaftlichen Befragungen geht es primär um die systematische Zielgerichtetheit und die zugrundeliegende Theorie. Den Hauptunterschied zwischen wissenschaftlicher und alltäglicher Befragung stellt die theoriegeleitete Kontrolle der gesamten Befragung dar, bei der man folgende Arten einer Befragung unterscheidet[10]:

[9] vgl. [Döri02], S. 295-296
[10] vgl. [Diek06], S. 373-374

- Das persönliche Interview (Face-to-Face)
- Das telefonische Interview
- Die schriftliche Befragung (Fragebogen)

Häufig sind aufgrund der Gegebenheiten mündliche Befragungen das geeignetste Mittel der Datenerhebung. Aufgrund der damit verbundenen hohen Kosten wird versucht, mittels schriftlicher Befragungen Kosten einzusparen. Bei mündlichen Befragungen übt der Interviewer bewusst oder unbewusst Einfluss auf den Gesprächsverlauf aus und erzeugt dadurch unvermeidbar eine Verzerrung der Antworten. Andererseits ist bei mündlicher Befragung eine Regel- und Kontrollfunktion durch den Interviewer möglich. Bei schriftlichen Befragungen kann oft nicht einmal genau festgestellt werden, wer eigentlich den Fragebogen unter welchen Umständen ausgefüllt hat und ob die Person alleine oder unter Anleitung eines Dritten ihre Antworten gegeben hat. Die heutzutage am meisten eingesetzte Methode stellt das Telefoninterview dar, die das früher üblichere persönliche Interview in weiten Bereichen abgelöst hat. Ein Interview wird auf Basis eines standardisierten Fragebogens durchgeführt (für alle Befragten die gleiche Fragen in gleicher Formulierung und Reihenfolge). Objektivität, Reliabilität und Validität sind drei wichtige Anforderungen. Fragebögen bei einer Befragung werden in Strukturierung und Standardisierung unterschieden. In standardisierten Interviews werden allen befragten Person die gleichen Fragen in der gleichen Reihenfolge gestellt und bei geschlossenen Fragen auch die gleichen Antwortkategorien angeboten. Bei strukturierten Interviews werden häufig auch offene Fragen ohne Antwortvorgaben durchgeführt.

Qualitative Beobachtung[11]:

Unter Beobachtung versteht man *das systematische Erfassen, Festhalten und Deuten sinnlich wahrnehmbaren Verhaltens zum Zeitpunkt seines Geschehens*[12]. Qualitative Beobachtungen beschäftigen sich mit offenen Kategorien bzw. Fragestellungen, dokumentieren größere Gesamtheiten des Verhaltens und Erlebens und werden im natürlichen Lebensumfeld bei aktiver Unterstützung des Beobachters durchgeführt. Im Vergleich mit Befragungen läuft die Beobachtung viel stärker als aktiver Prozess ab,

11 vgl. [Atte00], S. 73-74
12 [Atte00], S. 73

der hohe soziale Anforderung an die Forscher stellt. Auf der einen Seite ist die Beobachtung, die Erfassung und die Deutung sozialen Handels durchzuführen, auf der anderen Seite steht das eigene soziale Handeln. Die Beobachtung ist häufig als anfängliche Datenerhebungstechnik zu finden, weil ein erster Einblick in die Zusammenhänge möglich wird. Wissenschaftliche Beobachtung wird systematische Beobachtung genannt.

Die Beobachtung wird als empirische Datenerhebungstechnik dann brauchbare Ergebnisse liefern können, wenn sie einen genauen Forschungsplan hat, systematisch geplant und durchgeführt wird und möglichst wenig dem Zufall überlassen wird. Es ist auch wichtig, dass systematisch aufgezeichnet wird und auf eine generelle Bewertung der Ergebnisse Wert gelegt wird und nicht eine Sammlung von Besonderheiten erstellt wird. Gültigkeit, Zuverlässigkeit und Genauigkeit müssen überprüft werden. Das Ziel der wissenschaftlichen Beobachtung ist eine Beschreibung bzw. Erklärung und Ermittlung der sozialen Wirklichkeit auf Basis einer Forschungsfrage. Die wissenschaftliche Beobachtung unterscheidet sich von der alltäglichen Beobachtung durch die soziale Realität, durch systematische Wahrnehmungsprozesse und die Resultate dienen dazu, wissenschaftliche Fragen und Diskussionen zu überprüfen.[13]

Nonreaktive Verfahren: (Unobtrusive Measure, Nonreactive Research, nonintruding Measures) sind eine Datenerhebungsmethode, welche keinen Einfluss durch die untersuchende Person, Ereignisse und Prozesse erlaubt bzw. erlauben sollte. Bei dieser Methode wird entweder die verdeckte Beobachtung oder die indirekte Beobachtung verwendet. Der Beobachter und die Untersuchungsobjekte stehen nicht miteinander in Kontakt. Reaktionen wie bei Interview oder Befragung und die Verhaltensmuster werden nicht registriert. Beispiele für Nonreaktive Verfahren sind: Physische Spuren (z.B. verschmutze Seite eines Buches als Indikator für häufig verwendete Buchteile oder markierte Seite in einem Buch für wichtige oder unklare Texte), Hinweistafeln, Schilder (z.B. fremdsprachige Hinweise als Indikator für den Grad der Integration von Ausländern), Bücher, Zeitschriften, Filme oder andere Medien als Indikator z.B. in der Politik: Ermittlung, wie eine Partei oder ein Politiker bewertet wird, kann tlw. auf Basis von Nachrichtensendungen oder anderen Fernsehberichter-

13 vgl. [ScHiEs99], S. 358

stattungen erfolgen. Symbole wie Autoaufkleber oder Buttons können als Indikatoren für soziale Gruppenzugehörigkeit verwendet werden. Archive, Verzeichnisse und Verkaufsstatistiken sowie Einzeldokumente wie Tagebüchern sind auch Beispiele für nonreaktive Verfahren[14].

Probleme nicht-reaktiver Verfahren: Der wichtigste Nachteil der nonreaktiven Verfahren ist, dass kaum klare Gütekriterien (Reliabilität, Validität) existieren. Es gibt auch noch andere Probleme z.B. bei den Untersuchungen von Spuren ist es schwer möglich zu klären, welche ethnische Gruppe untersucht wird. Auch beim Beispiel der Symbole wie Autoaufklebern als Indikator für soziale Gruppenzugehörigkeit ist es schwer, die untersuchte Gruppe zu beurteilen, weil eine Gruppenzugehörigkeit nicht feststellbar ist. Nonreaktive Verfahren sind inhaltlich relativ stark limitiert. Man kann diese Methoden nur in wenigen Forschungsgebieten einsetzen. Zusätzlich zu den Nachteilen und Problemen dieser Methode darf nicht vergessen werden, dass wegen großteils fehlender Gütekriterien diese Methode im Vergleich zu anderen Instrumenten der qualitativen Methoden, eher selten verwendet werden.[15]

Gütekriterien[16]: Es ist ein wichtiger Grundsatz empirischer Sozialforschung, dass nach der Beendigung einer Studie die Ergebnisse mit Hilfe von Gütekriterien überprüft und kontrolliert werden müssen, um die Forderungen nach Objektivität, Reliabilität und Validität erfüllen zu können. Diese drei Faktoren bezeichnet man als zentrale Gütekriterien bei quantitativen und qualitativen Messungen. Die Objektivität eines Tests stellt sicher, dass die Ergebnisse einer Untersuchung unabhängig von der Person oder Gruppe des Untersuchers sind. Es ist sehr wichtig, dass bei Durchführung, Auswertung und Interpretation eine möglichst geringe Beeinflussung durch die durchführenden Personen erzielt wird, also dass die Beeinflussung minimiert wird. Die Reliabilität (Zuverlässigkeit) beschäftigt sich mit der Genauigkeit, wie exakt ein Test oder eine Untersuchung gemessen wird und was genau gemessen werden soll, ohne die Validität zu beachten. Validität gilt wie bei quantitativer Forschung auch bei qualitativer Forschung als das wichtigste Gütekriterium einer Datenerhebung.

14 vgl. [Döri02], S. 325-326
15 vgl. [ScHiEs99], S. 384-386
16 vgl. [Döri02], S. 326-328

Die Validität beschäftigt sich mit der Frage, wie weit die Untersuchungsergebnisse einen nachweisbaren Zusammenhang mit der Fragestellung haben und ob die verwendeten Indikatoren tatsächlich gültig sind.

Inhaltanalyse[17]:

„Content Analysis" ist eine Methode, die Kommunikationsinhalte, insbesondere Texte aller Art wie Bilder, Filme, Fernsehen, usw. untersucht. Inhaltanalyse stellt eine Form von Analyse von Texten und Datenerhebungsverfahren dar. In verschiedenen Bereichen wird Inhaltanalyse häufig verwendet: In der Soziologie zur Analyse des Lebensstils, der Lebensqualität und den Änderungen von Einstellungen von Menschen in der Gesellschaft. In der Publizistik, Pädagogik, Medien und Kommunikationswissenschaft, Ethnologie, Geschichte, Psychologie, Theologie und in der Literaturwissenschaft wird sie zur Analyse von Texten eingesetzt. Einer der Hauptanwendungsbereiche der Inhaltanalyse ist in der politischen Kommunikation, bei der Analyse von Massenmedien. Die Inhaltanalyse ist ebenfalls eine empirische Datenerhebungsmethode. Sie beschäftigt sich stark mit hermeneutischen Verfahren (Hermeneutik ist z.B. in der Literaturwissenschaft und der Psychologie üblich) und der Sprache. Datenbanken und Linguistik als zwei Ansätze werden ebenfalls unterschieden. Bei Datenbanken geht es primär darum, Text zu suchen und Informationen zu finden wie z.B. beim Recherchieren in Datenbanken oder Suchmaschinen im Internet. Linguistische Ansätze werden z.B. als Lexikographie oder Lemmatisierung von Worten verwendet (z.B. „gegangen" kann auf „gehen" zurückgeführt werden). Zwischen Inhaltsanalyse und hermeneutischen Verfahren gibt es große Unterschiede. Die Unterscheidung zwischen empirisch und hermeneutisch, wie in Abbildung 1 ersichtlich, legen die Bereiche quantitativ und qualitativ fest.

17 vgl. [Atte00], S. 201-204

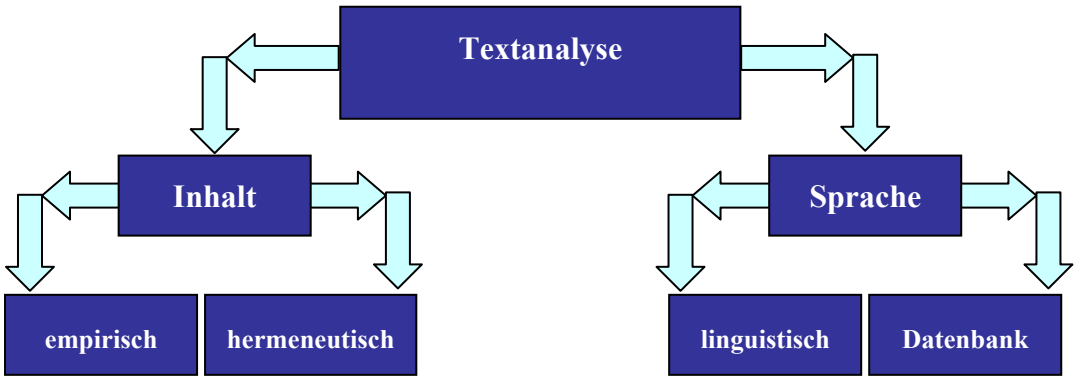

Abbildung 1: Textanalytischer Ansätze
Quelle: vgl. [Atte06], S. 181, eigene Darstellung

Experiment[18]:

Das Experiment ist eine bestimmte Untersuchungsanordnung. Es ist keine besondere Methode der Datenerhebung oder der Messung sozialer Daten. Es macht Sinn, jede Untersuchung als Experiment zu bezeichnen und ein Experiment dient primär der Überprüfung einer Hypothese. Ein Experiment ist somit die Überprüfung von bestimmten Aussagen. Das Experiment in der Sozialforschung ist der Versuch eines Beweises einer zu untersuchenden Hypothese, die zwei oder mehr Faktoren in einer begründeten Beziehung zueinander bringen will und diesen Zusammenhang in unterschiedlichen Situationen untersucht. Das Experiment hat gegenüber der Beobachtung und der Befragung folgende Vorteile:

Bei einem Experiment können Extremsituationen ausgearbeitet werden und Hypothesen unter strengen Prüfforderungen wiederholt getestet werden. Ein Experiment bietet also eine Möglichkeit, die Versuchsperson und -gegenstände in einen künstlich modellierten Prozess zu versetzen und unter ständiger Kontrolle können die sozialen Zusammenhänge reproduziert werden.

Folgende Grundbedingungen für ein Experiment sind wesentlich:

1. die notwendigen Faktoren und Variablen der Forschungsfrage zur Hypothesenbildung müssen bekannt sein

18 vgl. [Atte00], S. 183-200

2. die Kausalitätsbedingung ist für die Hypothesenbildung zwingend (abhängige und unabhängige Variablen)
3. die untersuchten Variablen müssen voneinander trennbar sein, Zweck: Kontrolle
4. die Variierbarkeit der unabhängigen Variable
5. Wiederholbarkeit der Studie muss gegeben sein.

2.5 Ablauf eines empirischen Forschungsprojektes

Man teilt den Ablauf eines typischen Forschungsprojektes in 5 verschiedene Phasen ein:[19]

- Formulierung und Präzisierung des Forschungsproblems
- Planung und Vorbereitung der Erhebung
- Datenerhebung
- Datenauswertung
- Berichterstattung

Abbildung 2: Phasen des Forschungsablaufs
Quelle: vgl. [Attes06], S. 17, eigene Darstellung

Der Ablauf eines empirischen Forschungsprojektes ist immer unabhängig von den verwendeten Methoden (Befragung, Narrative Interviews, Inhaltanalysen, …). Das Ziel ist es, überprüfbare Ergebnisse zu erarbeiten und dazu ist eine Standardisierung notwendig, die den Ablauf eines Forschungsprojektes normiert.

Am Beginn einer Untersuchung steht ein Problem bzw. eine Fragestellung, die geklärt werden soll. Als Problembenennung wird das soziale Problem als wissenschaft-

[19] vgl. [Diek06], S. 162-165

liche Fragestellung formuliert. Forscher können mit Hilfe von Hypothesen die theoretischen Zusammenhänge der sozialen Wirklichkeit feststellen.

Im allgemeinen Sinn ist eine Hypothese eine Vermutung über einen bestehenden Sachverhalt[20].

Eine Hypothese ist eine deutliche und klare Idee, die als Aussage mündlich oder schriftlich ausgedrückt werden muss und ist somit eigentlich die Voraussetzung für jede wissenschaftliche Forschung. Die Formulierung von Hypothesen ist ein induktiver Prozess d.h. im Endeffekt kann vom Einzelnen auf das Allgemeine geschlossen werden. Durch den Einsatz von Hypothesen und ihre empirischen Überprüfungen wird versucht, „wahres" Wissen über die soziale Realität[21] zu erlangen und sich von er Methode des „Trial and Error" zu entfernen. Die Widerlegung einer Hypothese nennt man Falsifizierung. Zur Bestätigung der Korrektheit einer Hypothese ist in Einzelfällen auch eine positive Überprüfung in Form von Stichproben sinnvoll und notwendig.

Nachdem die Problembenennung durchgeführt ist und man formuliert hat, wie eine Erscheinung auf ihre Ursachen zurückzuführen ist, wird das Forschungsdesign aufgebaut: (Konzeptspezifikation und Operationalisierung, Bestimmung der Untersuchungsformen). Mit Hilfe der Instrumente der Datenerhebung als Methoden der empirischen Sozialforschung werden die für die Hypothesenprüfung relevanten Daten gesammelt. Im nächsten Schritt erfolgt die Aufbereitung und Auswertung der erhobenen Daten und zum Schluss werden die erhobenen Daten mit verschiedenen statistischen Verfahren ausgewertet. Unter ständigem Rückbezug auf die vorangegangenen Arbeitsschritte werden die Daten interpretiert und besonders bei der Auswertung wird bei den statistischen Verfahren das Messniveau der Daten der empirischen Realität angepasst. Die Auswertung der in der Problemstellung beschriebenen Anforderungen muss im Hinblick auf die Problemstellung gewährleistet werden. Zum Schluss nach der Auswertung und Interpretation des erhobenen Materials und den

20 [Diek06], S. 107
21 vgl. [Krom95], S. 133

Rückschlüssen, die sich ergeben, werden die Ergebnisse in einem schriftlichen Bericht in Form einer Publikation zusammengestellt und verwendet.

Schließlich gilt es, darauf zu achten, wie die Beziehung zwischen den erhobenen Daten und ihrer Bedeutung im realen Leben ist. Die Hilfstheorien, welche von der ersten Phase bis zu letzten Phase im Feld zur Anwendung kommen können, sind ein notwendiger Bestandteil der Datenerhebung, -gewinnung und –analyse und die Erhebungsmethoden und statistischen Methoden sollten überlegt und expliziert bzw. argumentiert werden.

2.6 Abgrenzung

Häufig wird nur wenig zwischen der tatsächlichen empirischen Sozialforschung und den Bereichen Data-Warehousing bzw. Data-Mining unterschieden. Diese Unterscheidung ist aber von verhältnismäßig großer Bedeutung, weil es sich bei Data-Warehousing und Data-Mining nur um Vorgangsweisen handelt, die auf bereits erhobenen Daten basieren. Beide letztgenannten Techniken sind auf die Existenz von erhobenen und bereits digitalisierten Daten angewiesen und stellen im Grunde genommen keine Methoden der qualitativen Analyse dar. Data-Mining umfasst Datenanalyse, den Neugewinn von Zusammenhängen, Klassifizierung und Relevanzanalyse von Informationen aus großen Datenbeständen (Datenbanken), um danach die neuen Informationen für Entscheidungen zu nutzen. Es geht also um die Untersuchung von Mustern, Profilen und Trends, um aus Informationen mehr Nutzen zu erhalten. Unter einem Data-Warehouse versteht man ein Datenbanksystem oder ein „Daten-Lagerhaus“, in dem viele bzw. alle Geschäftsprozesse eines Unternehmens zur Unterstützung von Managemententscheidungen in Form einer (meist) zentralen Sammlung von Unternehmensdaten dargestellt und zur Verfügung gestellt werden. Durch den Einsatz von Datenbanken werden alle enthaltenen Informationen und Unternehmensdaten in einen sinnvollen Zusammenhang gebracht und ein Data-Warehouse erlaubt es, aus diesen Daten Wissen und Informationen zu extrahieren. Data-Warehousing ermöglicht es jedem berechtigten Benutzer einer Organisation, jederzeit zu individuellen Informationen zu kommen und damit seine Informationsbedürfnisse zu befriedigen. Mit Hilfe von vorbereiteten Abfragen und individuell

erstellten Abfragen können Informationen abgefragt, analysiert und visualisiert werden. Im Normalfall werden die Daten beim Data-Warehousing in einer oder mehreren Datenbanken vorgehalten, wobei einer der wesentlichen Unterschiede zu konventionellen Datenbanken die Tatsache darstellt, dass beim Data-Warehousing meist historisierte und konsolidierte Daten gespeichert werden, deren Veränderung oder Löschung nicht mehr möglich ist und diese Daten somit eine permanent wachsende Informationsquelle darstellen.[22]

2.6.1 Data-Mining

Der Begriff Data-Mining wird im Bereich der Statistik und im Bereich der Forschung bei Datenbankmanagement verwendet. Data-Mining beschäftigt sich mit der Behandlung und der Analyse großer Datenbestände und verwendet verschiedene Techniken und Methoden, um versteckte Informationen, Zusammenhänge und Wissen in Datenbeständen zu finden bzw. offenzulegen, die in Form von Mustern innerhalb der Datenbestände auffindbar sind. Unter Wissen sind in diesem Zusammenhang Informationen gemeint, die von Organisationen oder Unternehmen genutzt werden können, um den Geschäftsverlauf beeinflussen zu können. Manchmal wird Data-Mining als Datenmustererkennung übersetzt.

Durch Data Mining können also Muster und Zusammenhänge gefunden werden, die vorher nicht erkennbar gewesen sind.

Data-Mining kann bei (großen) Unternehmen manchmal helfen, wichtige Fragestellungen zu beantworten oder Probleme zu lösen, die ohne den Einsatz derartiger Technologien schwer oder nicht zu beantworten wären. Zeitraubende und aufwändige Datenanalysen großer Datenbestände werden durch Methoden des Data-Mining teilweise wesentlich vereinfacht bzw. beschleunigt, indem rasch die relevanten Informationen lokalisiert werden und aufgrund der Zusammenhänge Entscheidungen ermöglicht werden. Typische Anwendung von Data Mining umfassen Absatzsegmentierung, Kundenprofilierung, Auswertung von Kleinförderungen und Kreditrisikoanalyse. Die Hauptbenutzer von Data-Mining sind Kreditkartenunternehmen, Ein-

[22] vgl. [Nent00], S. 2-3

zelhändler, Finanzdienstleister, Telemarketing- und Direktvertriebsuntenehmen, Fluglinien, Produktionsbetriebe und Telekommunikationsunternehmen.

Data-Mining verwendet intensiv Methoden der Statistik zur Datenexploration und Datentransformation und Methoden der Mustererkennung, um Validierungen, Beschreibungen und Visualisierungen der Resultate zu erstellen. In der Datenbankforschung wird Data-Mining verwendet, um Datenverfügbarkeit, Datenintegrität und Plausibilität von Datenbanken zu analysieren und zu kontrollieren.

Data-Mining verwendet darüber hinaus Verfahren und Methoden der künstlichen Intelligenz. Dazu zählen künstliche Neuronale Netze (KNN), genetische Algorithmen, Entscheidungsbäume, Fallbasiertes Schließen und Assoziationsanalyse, die neben den statistischen Methoden zur Anwendung kommen.

Eine populäre Variante des Data Mining ist in den letzten Jahren Text-Mining geworden. Text-Mining beschäftigt sich mit dem Herausfinden von Mustern in unformatierten bzw. unstrukturierten Daten (Zeitungstexten, Patenten, elektronischen Nachtrichten, usw.) und kann als Erweiterung der klassischen Informationsgewinnung bzw. Informationsermittlung angesehen werden. In den Letzten Jahren spricht man von Web-Mining, Web-Log-Mining und Web-Content-Mining usw. Die Methoden, die beim Web- Mining verwendet werden, sind prinzipiell dieselben Methoden, wie sie auch beim klassischen Data-Mining verwendet werden. Teilweise sind auch Ansätze entwickelt worden, die spezifisch auf die Untersuchungen des WWW angewendet werden und somit Erweiterungen des klassischen Data-Mining darstellen.[23] Die Werkzeuge des Data-Mining sind Softwareprogramme, die Daten und Informationen so aufbereiten können, dass mit Hilfe von bestimmten wählbaren Methoden Analyseergebnisse ermittelt werden können und diese Ergebnisse dann nützlich und verständlich präsentiert werden können.

Der Markt bietet heute umfangreiche und ausgereifte Softwarepakete mit bestimmten Besonderheiten wie der Fähigkeit zur Aufbereitung von Datenbeständen, der

[23] vgl. [AlNi00], S. 3-5

Auswahl der passenden Data-Mining Methode (Algorithmen), Produktskalierbarkeit und verschiedenen Möglichkeiten zur Visualisierung der Ergebnisse. Die bekanntesten und am häufigsten anzutreffenden Werkzeuge sind Clementine von Integral Solution Ltd., Data Crusher von Data Mind Corp, Intelligent Miner von IBM, Mineset von Silicon Graphics Inc., Data Mining Suite von Information Discovery Inc. und SAS System von SAS Institute Inc.[24]

2.6.2 Der Prozess des Data-Mining

Der Prozess des Data-Mining besteht aus folgenden Phasen[25]:

1. Auswahl der Daten und Feststellung der Aufgaben: Der erste Schritt ist die Auswahl der zu untersuchenden Daten aus den existierenden Daten bzw. der Objekte (Datensätze) die abgebildet sind und ihrer Merkmale (Datenfelder). In dieser Phase ist es wichtig, dass Data- Mining sich oft mit einer Stichprobe begnügt, die repräsentativ für den gesamten Datenbestand ist und daher bei der Auswahl auf Verwendbarkeit als repräsentative Stichprobe überprüft werden muss.

2. Vorbereitung der Daten (Preprocessing): In dieser Phase werden die Daten bereinigt, was häufig notwendig ist (z.B. falls eine Datenfeld ungültige Werte enthält) und es werden fehlende Werte behandelt (z.B. durch Weglassen der entsprechenden Datensätze oder Ersetzen der fehlenden Werte durch Standardwerte) und weitere ähnliche Vorarbeiten durchgeführt. Es erfolgt eine Vorbereitung der Daten und Informationen für die folgende Verarbeitung. Aus Produktionssystemen oder aus Data-Warehouse Datenbeständen entsteht als Ergebnis eine Data-Mining Basis die die Basis für weitere Transformationen und Verarbeitungsschritte darstellt. Der Aufwand für die Datenvorbereitung ist nicht zu unterschätzen, weil häufig mehr Fehler in den Daten vorhanden sind, als anfangs angenommen.

[24] vgl. [DaSe04], S. 9
[25] vgl. [AlNi00], S. 6-9

3. Transformation der Daten: Bei der Transformation der Daten werden im Bedarfsfall Datenbereiche verändert, normiert und quantitative Daten werden in verschiedene Kategorien eingeteilt. Neue Datenfelder werden durch Aggregation oder andere Berechnungen generiert.

4. Auswahl von Data-Mining Methoden: In dieser Phase wird entschieden, welche Methoden des Data-Mining verwendet werden sollen. Die Aufgabenstellung spielt eine große Rolle für die Auswahl der Methoden.

5. Anwendung der Data-Mining Methoden: In dieser Phase werden die ausgewählten Methoden des Data-Mining nach verschiedene Kriterien für die Analyse der einzelnen Fragestellungen benützt. Es werden auch verschieden viele Klassifikationsebenen genutzt. Bei Chamoni wird eine Ebene mit ausgewählten „Verfahren“ wie Clusteranalyse, Bayes-Klassifikation, Induktivem Lernen und künstlichen Neuronalen Netzen (KNN) identifiziert, bei Schinzer et al. werden zwei Ebenen identifiziert: Verfahren wie Segmentierung, Klassifizierung, Assoziierung und Techniken wie Entscheidungsbäume und KNN. Bei Fayya et.al. werden drei Ebenen unterschieden. Zwei Ziele wie Vorhersage oder Beschreibung, die Methoden wie Klassifikation und Regression. und die Algorithmen wie Entscheidungsbäume, nichtlineare Regression bilden die Elemente der drei Ebenen.

6. Interpretation und Evaluation der Data Mining Ergebnisse: In dieser Phase werden die gefundenen Muster interpretiert und evaluiert und es wird überprüft, ob das Data-Mining brauchbare Ergebnisse geliefert hat oder ob mehr Daten für eine gültige und sinnvolle Aussage benötigen werden oder ob andere Methoden für die Analyse verwendet werden sollten. Die Interpretation der Daten wird mit Hilfe von Filtern und der Aufbereitung der Ergebnisse meist mittels Grafiken visualisiert. Das Ziel dieser Phase ist es, die Ergebnisse des Data-Mining auch für Nichtexperten deutlich zu machen. Die gewonnenen Ergebnisse werden zuerst interpretiert und die gefundenen Aussagen bzw. Informationen sollten interessant und neu sein und möglichst weitgehend die Fragen der Aufgabestellung beantworten. Falls die erwarteten Muster nicht gefunden werden können, so muss der Grund dafür ermittelt klar werden und zu einem der vorherigen Schritte im

Data-Mining Prozess zurückgekehrt werden, um die Aufgabenstellung durch bessere Auswahl der Algorithmen bzw. bessere Datenvorbereitung und -auswahl zu ermöglichen.

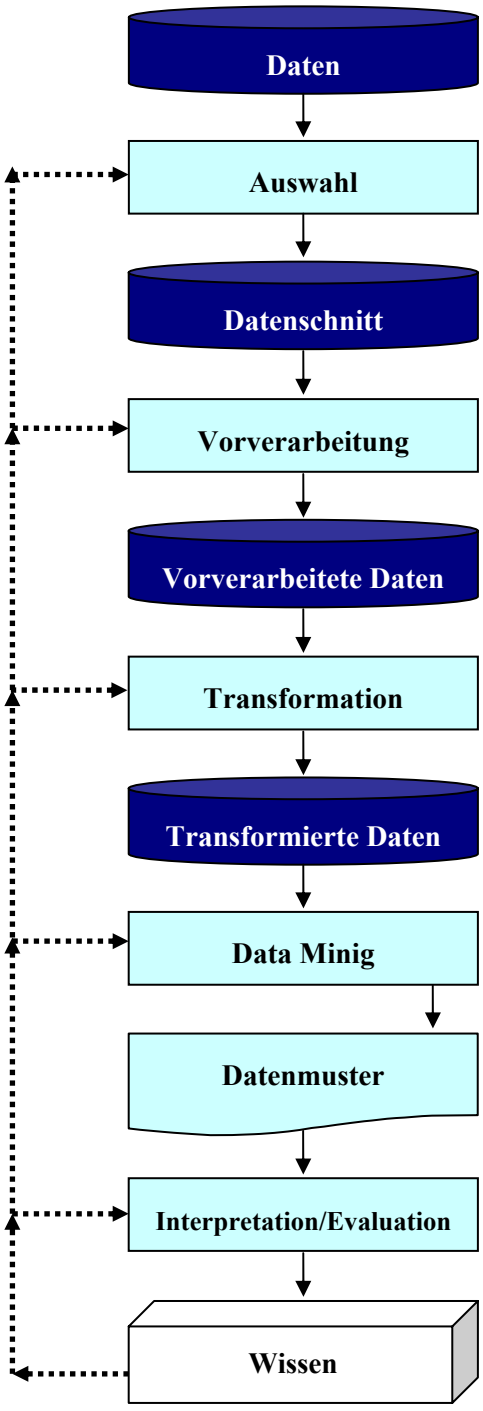

Abbildung 3: Schritte im Data Mininag-Prozess (nach Fayyad, U. et al. 1996)
Quelle: [AlNi00], S. 7, eigene Darstellung

2.6.3 Methoden und Techniken des Data-Mining[26]

Kern jedes Data Mining Systems sind die Analysealgorithmen[27].

Die Methoden und Techniken des Data Mining werden in der Literatur nach verschiedenen Eigenschaften klassifiziert. Die einzelnen Methoden und Algorithmen sind für eine bestimmte Problemstellung anwendbar. Meistens wird beim Data-Mining die Aufgabenstellung als erste Ebene bezeichnet und als zweite Ebene die Methoden und Algorithmen.

- Klassifikation (Classification): Die Hauptaufgabe der Klassifikation ist es, die betrachteten Objekte in bestimmte Klassen einzuteilen bzw. die Objekte bestimmten Klassen zuzuordnen. Die Einordnung wird aufgrund der Objektmerkmale und Klasseneigenschaften vorgenommen. Die bekanntesten Methoden für die Klassifikation sind die Entscheidungsbaummethode, Fallbasiertes Schließen und Neuronale Netze.

- Segmentierung (Sequential Patterns): Unter Segmentierung versteht man die Zerlegung der Datenbasis in einzelne Segmente die immer aus gleichartigen Datensätzen besteht. Bei der Segmentierung werden die Objekte in verschiedene Gruppen und Kategorien unterteilt. Die Bedeutung der Gruppe wird von Benutzern auf Basis gemeinsamer Eigenschaften der Mitglieder der neuen Gruppe festgelegt z.B. DINKs (double income, no kids) wäre eine mögliche Gruppe bei der Analyse von Bewohnerdaten einer Stadt. Bei der Segmentierung wird am häufigsten das Verfahren der Clusteranalyse verwendet, um Gruppen von Datensätzen , die Ähnlichkeiten aufweisen, auffinden zu können.

- Prognose (Prediction): Der Begriff Prognose wird als eine Vorhersage für die Zukunft oder für die Gesamtheit verwendet. Die Prognose nützt eine Vorhersage unbekannter Merkmalswerte entweder auf der Basis von anderen Merkmalen oder von Werten des gleichen Merkmales eines früheren Zeitraumes oder einer Stich-

26 vgl. [AlNi00], S. 9-13 und [Nent00], S. 9-14
27 [Nent00], S. 9

probe. Für die Prognose werden beim Data-Mining meist Entscheidungsbäume und KNN verwendet.

- Abhängigkeitsanalyse: Es wird die Abhängigkeit, also eine Beziehung zwischen Merkmalen eines Objekt oder verschiedener Objekte, untersucht. Die Abhängigkeitsbeziehung kann entweder in einer bestimmten Zeitperiode oder in verschiedenen Zeitperioden bestehen. Am Beispiel einer Warenkorbanalyse kann sich herausstellen, dass zwei Produkte oft gleichzeitig gekauft werden. Bei der Untersuchung von Kreditkartentransaktionen könnte analysiert werden, dass oft vier bis sechs Monate nach dem Kauf eines Videorecorders eine Videokamera gekauft wurde. Analysieren der Zeitverläufe und der Korrelation verschiedener Objekte ist die Aufgabe der Abhängigkeitsanalyse. Bei der Abhängigkeitsanalyse wird häufig das Informationsflussgraph - Verfahren verwendet.

- Abweichungsanalyse: Die Abweichungsanalyse kann als eine Weiterführung der vorigen Methoden betrachtet werden. Es geht darum, die Objekte und deren Regelmäßigkeiten festzustellen und untypische Merkmalsausprägungen zu identifizieren. Mit Hilfe der Abweichungsanalyse werden Objekte gesucht, deren Merkmale sogenannte „Ausreißer“ darstellen, die also mit hoher Wahrscheinlichkeit falsch erhobene Daten beinhalten und diese Objekte werden dann meist aus der weiteren Untersuchung ausgeschieden. Die Abweichungsanalyse wird oft in der Vorbereitungsphase eingesetzt.

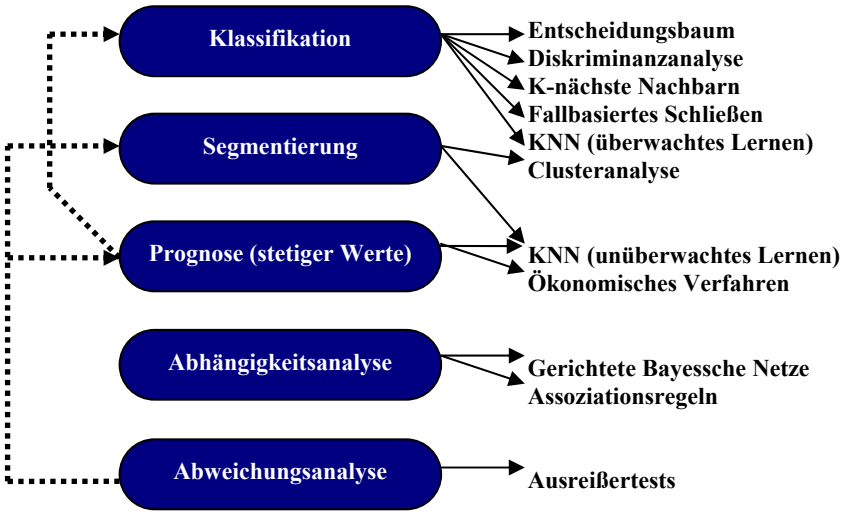

Abbildung 4: Zuordnung von Data-Mining Methode zu Aufgaben
Quelle: [AlNi00], S. 13, eigene Darstellung

3 Meinungsforschung

3.1 Was ist Meinungsforschung

Meinungsforschung (opinion research) ist eine Methode, die durch Befragung oder statistische Untersuchungsmethoden die Meinungen und Einstellungen von Menschen beobachtet und Daten und Informationen erfasst und analysiert. Die Meinungsforschung versucht durch Befragung die Einstellungen von Menschen, besonders zu politischen, wirtschaftlichen und sozialen Fragen herauszufinden. Eine Befragung kann, wie bereits oben angeführt, per Telefon, durch persönliche Befragung oder durch einen Online-Fragenbogen durchgeführt werden. Meinungsforschung basiert auf Datenanalyse, welche eine repräsentative Stichprobe der Bevölkerung als Basis verwendet, um die gesammelten Daten ohne große Verzerrungen auszuwerten und zu interpretieren.[28]

3.2 Grundlagen und praktische Anwendung

Die theoretische Grundlage der Meinungsforschung basiert auf Erkenntnissen und Grundlagen der empirischen Sozialforschung. Sie ist ein Teilbereich der empirischen Sozialforschung und die empirische Sozialforschung findet eine ihrer hauptsächlichen Anwendungen in der Markt- und Meinungsforschung. Die Ursprünge der Markt- und Meinungsforschung gehen bis ins 18. Jahrhundert zurück und zwar in den U.S.A., wo die ersten Probestimmungen zu Vorhersage der Wahlergebnisse vorgenommen wurden. In Europa, vor allem in Deutschland, wurde 1945 das Institut Allensbach gegründet, welches bis heute eine der bekanntesten Institute in diesem Bereich ist. Markt und Meinungsforschung wird häufig in der Verbraucherforschung, Werbeforschung und Wählerforschung angewendet. Da durch die Meinungsforschung viele Informationen für die Politik und die Vorhersage der Wahlen gewonnen werden können, wird sehr oft auf diese Möglichkeit zurückgegriffen. Praktische Anwendung der Meinungsforschung ist weiters im Bereich der Betriebswirtschaft und Volkswirtschaft, Politik, Mediendokumentation (Clippingservice und

[28] vgl. [o.V.08a], o.S.

ähnliches) und Psychologie zu finden und fundiert Entscheidungen mit oft weitreichenden Konsequenzen. Durch die Methoden wie Befragung, Beobachtung oder Interviews, die in der Meinungsforschung verwendet werden, wird ein Verfahren zur Analyse von Zahlen und Daten eingeleitet, um Aussagen über die Meinungen von Bevölkerungsteilen zu treffen. In der Politik und in der Wahlforschung wird die Meinungsforschung und in der Wirtschaft wird vor allem die Marktforschung verwendet.[29]

3.3 Hauptziele der Meinungsforschung

Hauptziel der Meinungsforschung ist die Erhebung von repräsentativen Meinungsstrukturen in der Bevölkerung zu wesentlichen „Issues". Hierbei spielt die Repräsentativität der erhobenen Daten eine wesentliche Rolle. Die Meinungsforschung hat im Rahmen der gesellschaftlichen Tätigkeit verschiedene Aufgaben:

- Die Problemanalyse und Beratung
- Das Erstellen von Marktforschungskonzepten
- Die Planung des Untersuchungsdesigns
- Die Entwicklung und Modifikation von Erhebungsmodellen
- Die Projektkonzeption und -koordination und die Durchführung empirischer Untersuchungen und Datenerhebung
- Die Auswertung und Datenanalyse und Aufbereitung
- Die Interpretation und Präsentation der Untersuchungsergebnisse
- Die Beratung der Klienten in allen relevanten Bereichen
- Die Wahrung der ihr anvertrauten Geschäftsgeheimnisse[30]

3.4 Arten von Interviews

In der Realität der empirischen Sozialforschung spielt das qualitative Interview in verschiedenen Varianten eine wichtige Rolle. Das Interview ist eine Gesprächsitua-

[29] vgl. [Meye07], o.S.
[30] vgl. [o.V.o.J.c], o.S.

tion, bei der von Interviewer gezielt und bewusst Fragen gestellt werden, die die befragte Person beantwortet. Ein Interview ist eigentlich eine Befragung durch Interviewer mit dem Ziel, bestimmte Informationen zu bekommen. Wie bei der Befragung ist beim Interview auch die soziale Situation zu berücksichtigen. Nicht nur die Personen die miteinander sprechen, sondern auch die Umgebung und die soziale Situation und Reaktion beeinflussen das gesamte Interview. z.B. wenn eine Person allein auf einen Fragebogen antwortet oder telefonisch befragt wird.[31]

Bei den qualitativen Forschungsmethoden gibt es verschiedene Arten des Interviews:[32]

- wenig strukturiertes Interview: Beim wenig strukturierten Interview verwendet der Interviewer keinen Fragebogen. Die Gesprächführung ist flexibel. Der Interviewer hat Freiheit und kann die Form und die Reihenfolge der Fragen selbst wählen. Wenn der Befragte zu tief bzw. zu viel auf eine Frage antwortet oder vom Ziel des Interviews zu weit abkommt, kann der Interviewer die Gesprächsrichtung ändern. Die Gesprächrichtung ist flexibel. Der Interviewer hört vor allem zu und nimmt die Hinweise und Informationen, die der Befragte gegeben hat, auf. Das Gespräch bei dieser Art des Interviews hängt nicht nur von den Fragen des Interviews ab, sondern die nächsten Fragen ergeben sich aus den Aussagen des Befragten. Das Ziel des wenig strukturierten Interviews ist die Meinungsstruktur der befragten Person herausbekommen.

- teilstrukturiertes Interview: Es besteht im Wesentlichen aus einem Gespräch mit vorformulierten und vorbereiteten Fragen. In der Regel wird dazu ein Gesprächplan benützt. Auf Grund der Antworten und der Reaktionen der befragte Person kann im Lauf des Gespräches die Richtung des Interviews verloren gehen und einzelne Perspektiven oder Themenbereiche können weiter vertieft werden. Ein wichtiges Ziel ist es, die Grundlinie des Interviews zu

31 vgl. [Atte00], S. 117-118
32 vgl. [Atte00], S. 140-150

erhalten, sodass mit der befragten Person der Themenschwerpunkt strukturiert bearbeitet wird.

- stark strukturiertes Interview: Beim stark strukturierten Interview muss zuerst ein Fragebogen konstruiert werden. Wichtig ist, dass durch den Fragebogen die Freiheitsspielräume des Interviewers stark beschränkt sind. Fehler im Fragebogen, die aus der Erhebungsphase resultieren, lassen sich nicht korrigieren und werden nicht berücksichtigt. Der Fragebogen legt den Inhalt, die Anzahl der Fragen und die Reihenfolge der Fragen fest. Darüber hinaus werden bei der Erstellung des Fragebogens die sprachliche Formulierung der Fragen, die Methode der Anwendung und die Antwortkategorien festgelegt. Durch die theoretische Problemstellung werden Inhalt, Anordnung und Anzahl der Fragen festgelegt, sodass über das Ziel der Untersuchung eine vollständige Informations- und Datenerhebung durchgeführt werden kann. Die Gesamtdauer eines Interviews soll zwischen 30 und 60 Minuten liegen.

- mündliches Interview[33]: Mündliche Befragungen werden meist als Interview bezeichnet. Es ist eine Befragungsmethode, um bestimmtes Wissen von Personen zu erfragen. Mündliche Interviews können als quantitative oder qualitative Befragung oder als Mischform durchgeführt werden. In der Sozialforschung werden häufig qualitative Interviews verwendet, weil in vielen Fällen nur ein mündliches Interview möglich ist. Beim mündlichen Interview ist der Interviewer in der Lage, Regel und Kontrollfunktion zu übernehmen und oft nimmt beim mündlichen Interview der Interviewer direkten Einfluss auf den Gesprächverlauf. Mündliche Interviews können mit Hilfe vorbereiteter Fragebögen so wie strukturierte Interviews durchgeführt werden oder ganz offen wie narrative Interviews geführt werden. Ein mündliches Interview verläuft in drei Phasen: Vorbereitungsphase, Durchführungsphase und Nachbereitungsphase. In der Vorbereitungsphase wird die Zielsetzung festgelegt und auch die Zielgruppe wird ausgewählt. In der Durchführungsphase werden die Fragebogen verteilt und die mündlichen Interviews durchgeführt und um ein hohe Beteiligung zu bekommen kann den befragten Personen eine Belohnung

[33] vgl. [Mitt05], o.S.

angeboten werden. Die Nachbereitungsphase ist die Endphase und es werden die Ergebnisse der Fragebögen beurteilt und für die Abschlusspräsentation aufbereitet. Telefoninterviews und Face-to-Face Interviews sind zwei bekannte Arten von Interviews, die bei Markt und Meinungsforschungsinstituten häufig verwendet werden. Es gibt viele Gründe für die Durchführung von Telefoninterviews z.B. eine verbesserte Telefontechnologie und auch dass fast die gesamte Bevölkerung heute per Telefon erreichbar ist. Nachteile sind, dass der Interviewer nicht genau weiß, wer eigentlich am Telefon ist und ob die befragte Person die Fragen ehrlich beantwortet, weil der visuelle Eindruck fehlt. Erinnerungsstützen und Hilfen durch Vorlage von Tabellen entfallen und bei unterschiedlichen Antwortmöglichkeiten wird entweder die erste oder die letzte Antwort überdurchschnittlich oft gewählt. Das Face-to-Face Interview ist eine Befragungssituation bei der Interviewer und befragte Person physisch am selben Ort anwesen sind. Auch längere Interviews sind möglich und der Interviewer kann Einfluss auf den Verlauf des Interviews nehmen, falls dies notwendig sein sollte.

- Schriftliches Interview: Das schriftliche Interview ist eine Sonderform der Befragung und ist nur mit voll strukturierten Fragebögen sinnvoll. Schriftliche Befragungen können in einer Gruppe von gleichzeitig anwesenden Personen durchgeführt werden, wobei die Fragebögen in Anwesenheit eines Interviewers ausfüllt werden. Ein Vorteil der schriftlichen Befragung liegt in den geringen Kosten. Es kann in kürzerer Zeit mit weniger Personalaufwand eine größere Zahl von Befragungen mit Hilfe von Fragebögen durchgeführt werden als dies bei mündlichen Interviews möglich ist. Bei schriftlichen Interviews ist zu bemerken, dass die Antworten ehrlicher sind, wenn kein Interviewer anwesend ist und es sind besser überlegte Antworten möglich, wenn ausreichend Bedenkzeit verfügbar ist, weil der Befragte mehr Zeit und höhere Konzentration für die Beantwortung der Fragen hat. Die Nachteile des schriftlichen Interviews liegen in der schlecht kontrollierbaren Befragungssituation weil z.B. andere Personen die Antworten der befragten Person beeinflussen können. Da kein Interviewer zusätzliche Erläuterungen zu den Fragen geben kann, muss bei der Gestaltung der Fragen stark darauf geachtet werden, dass die Fragen gut verständlich sind und es besteht eine relativ hohe

Wahrscheinlichkeit, dass Fragen unvollständig oder gar nicht beantwortet werden Weitere Nachteile des schriftlichen Interviews bestehen in der hohen Ausfallquote und in der Tatsache, dass die Antworten weniger spontan erfolgen.

3.5 Arten von Fragen und Antworten

Im folgende Teil werden die Arten von Fragen und Antworten erklärt[34]. Man unterscheidet grundsätzlich zwei verschiedene Fragetypen: „offene“ und „geschlossene“ Fragen.

- Offene Fragen haben keine fixen Antwortkategorien und können auch nicht mit Ja oder Nein beantwortet werden. Die befragte Person verfasst und formuliert die Antwort in eigenen Worten und es sollte bei mündlichen Interviews eine möglichst exakte Aufzeichnung der Antworten erfolgen. Im Rahmen der Auswertung müssen die Antworten dann bestimmten Kategorien zugeordnet werden. Ein Nachteil offener Fragen ist die Notwendigkeit der späteren Kategorisierung und der Aufwand für die Anpassung an die Ausdrucksweise des jeweiligen Befragten bei der Auswertung. Offene Fragen sind unverzichtbar, wenn es um die Erhebung von Meinungen geht.

- Bei der geschlossenen Frage wird dem Befragten vorgegeben, welche Antwortmöglichkeiten erlaubt sind und der Befragte muss die bestzutreffende Antwort aus den vorhandenen Antwortmöglichkeiten auswählen. Ein Nachteil geschlossener Fragen ist die Richtungsgebung der Befragung durch die Vorgabe der Antwortmöglichkeiten. Der wesentliche Vorteil geschlossener Fragen ist die wesentlich einfachere Auswertbarkeit geschlossener Fragen.

Gegenüberstellung offener und geschlossener Fragen:
Bei offenen Fragen wird erwartet, dass der Befragte sich an etwas erinnert bzw. mit eigenem Aufwand eine Antwort überlegt und formuliert. Bei geschlossenen Fragen wird einfach eine Antwort aus einer der Antwortmöglichkeiten ausgewählt. In der

[34] vgl. [Atte00], S. 158-170

Regel werden offene Fragen in geringerem Ausmaß beantwortet, enthalten aber aufschlussreiche Tatsachen, weil die Beschäftigung des Befragten mit der Fragestellung stärker ist. Geschlossene Fragen hingegen sind leichter auswertbar und einfacher vergleichbar, bieten aber ein gewisses Fehlerpotenzial, wenn sich der Befragte vor der Befragung noch nie über den Gegenstand der Frage eine Meinung gebildet hat, weil geschlossene Fragen dazu verleiten können, einfach eine der Antwortmöglichkeiten auszuwählen. Offene Fragen helfen dabei, Sachverhalte zu erfahren, die schwer kategorisierbar sind. Allgemein sollten Fragen und Antworten unter Berücksichtigung von Regeln formuliert sein: Fragen sollen einfache Worte enthalten und sie sollten kurz, konkret und nicht hypothetisch formuliert sein. Doppelte Negationen sollten nicht verwendet werden, um die Befragten nicht zu überfordern. Geschlossene Fragen sollten zumindest formal balanciert werden, sodass zumindest eine negative und eine positive Antwortmöglichkeit angeboten wird.
Geschlossene Fragen werden in verschiedene Typen unterteilt:

- Identifikationstyp: (W-Frage) Eine Frage, welche nach einer Person, Gruppe, Zeit, Zahl oder einem Ort gefragt wird: Wer, wo, wann, wie viel oder welche?

- Selektions-Typ oder Multiple-Choice-Typ: Eine Frage mit verschiedenen alternativen Antwortmöglichkeiten, wobei der Befragte eine von zwei oder mehreren Antworten auswählen muss. Der Selektionstyp wird auch als Alternative Frage bezeichnet, wenn es um eine Frage mit nur zwei Antwortmöglichkeiten geht. Wenn es mehr als zwei Antwortkategorien gibt, wird der Fragetyp mehrfachauswahl-fragend genannt.

- Eine besondere Form der Mehrfachauswahl-Frage ist „Skala-Frage“, mit der Werte, Meinungen, Gefühle oder Inhalt bezüglich ihrer Intensität oder Häufigkeit abgefragt werden

- Die „Dialogfrage“ ist auch eine Sonderform des Selektionstyps, bei der eine alternative Frage „eingekleidet“ wird.

- Ja-Nein Typ: Eine Frage, die mit Ja oder Nein als antwort beantwortet werden kann.

Direkte und Indirekte Fragen:

Bei indirekten Fragen wird versucht, eine Gesprächsituation zu formen, sodass der Befragte seine Gedanken über Gefühle und subjektiv wichtige Probleme frei formulieren kann, die er vorher aus sozialen bzw. persönlichen Gründen zurückgestellt hat oder nicht formulieren wollte oder konnte. Indirekte Fragen helfen bei der Beschaffung von Informationen über Zusammenhänge, die dem Befragten selbst tlw. nicht bewusst sind. Die Erwartung dass über den Weg der indirekten Fragen im Gegensatz zu direkten Fragen ein höherer Wahrheitsgehalt der Antworten erzielbar ist, konnte nicht bestätigt werden. Die Bewertung indirekter Fragen und der korrespondierenden Antworten ist äußerst schwierig, weil dabei Schlüsse gezogen werden müssen, die sich nur schwer wissenschaftlich definieren lassen. Meist wird eine Kombination aus direkten und indirekten Fragen verwendet. Auf Grund der Menge der verschiedenen möglichen Gebiete für Befragungen ist logischerweise auch die Anzahl der Sachgebiete und der Themenbereiche so groß, dass die Auswertung von Antworten auf indirekte Fragen in starkem Zusammenhang mit dem jeweiligen Themenbereich der Befragung zusammenhängt und nicht vereinheitlicht werden kann. Ein Vorteil indirekter Fragen ist die Tatsache, dass mit diesem Fragetyp viel Hintergrundinformation beschafft werden kann, allerdings ist eine Steuerung der Antwort durch den Interviewer nur schwer möglich, wenn die Antwort in die „falsche“ Richtung geht. Die anderen wichtigsten Fragetypen sind[35]:

Rhetorische Fragen , Suggestivfragen, Motivationsfragen, Informationsfrage (Interessenfrage), Alternativfrage (Entscheidungsfrage), Kontroll- oder Bestätigungsfrage, Verfolgerfrage, begründete Frage, hypothetische Frage, Fangfrage, Gegen- oder Rückfrage, Provokationsfrage, Unterstellungsfrage etc.

35 vgl. [Väth00], o.S.

4 Datenerhebung, -analyse und Interpretation

In diesem Kapitel befasst sich die Arbeit mit der Erhebung, Auswertung und Präsentation von Daten. Vor allem kommt es hier zu einer Definition der unterschiedlichen Terminologien, um eine Systematik zu entwickeln.

Es darf nicht vergessen werden, dass es für jede Forschung zuerst eine Forschungsfrage gibt. Die Phasen der Datenerhebung, der Analyse und der Visualisierung sind alle nach der Definition der Forschungsfrage als Folgephasen anzusehen. In diesem Kapitel wird auf die Formulierung und Präzisierung des Forschungsproblems kein Bezug genommen, sondern die Folgephasen beschrieben. Die Reihung dieser Phasen sieht wie in Abbildung 5 aus:

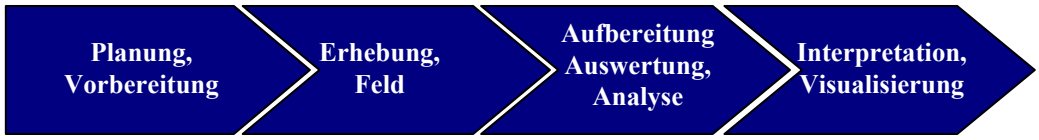

Abbildung 5: Forschungsphasen nach der Definition der Forschungsfrage
Quelle: eigene Darstellung

4.1 Datenerhebung

Es wurde schon im vorherigen Kapitel über die unterschiedlichen Arten und Möglichkeiten der Datenerhebung berichtet. Hier sollte es zu einem ganzheitlichen bzw. zusammenfassenden Überblick kommen, vor allem mit dem Fokus, eine bestimmte Definition dieses großen Fachgebiets vorzunehmen.

Als Datenerhebungsinstrument bzw. -methode wird heutzutage am häufigsten die Befragung bzw. das Interview verwendet. Verglichen mit anderen Methoden wie Beobachtung oder Inhaltsanalyse wird die Befragung viel häufiger in der empirischen Forschung eingesetzt.[36]

[36] vgl. [Diek06], S. 371-372

Die unterschiedlichen Formen der Befragung werden wie bereits erwähnt in diese Gruppen unterteilt: persönliche Interviews, telefonische Interviews, schriftliche Interviews. Eine andere sehr sinnvolle und notwendige Unterscheidung der Formen der Befragung ergibt die Unterteilung in strukturierte Befragung und unstrukturierte Befragung. Zusammenfassend werden die strukturierten Interviews zu den quantitativen Interviews bzw. quantitativen Forschung gezählt. Umgekehrt werden die unstrukturierten bzw. offenen Interviews zu den qualitativen Forschungsmethoden gezählt[37].

Die Unterscheidung zwischen diesen zwei Hauptgruppen ist manchmal nicht einfach. Eine Befragung kann als Grundlage einen Fragebogen haben, welcher von Anfang bis zum Schluss strukturiert ist, sowohl bei jeder einzelnen Frage als auch bei allen Antwortmöglichkeiten (vollständige Strukturiertheit). Andererseits kann man doch bei einer unstrukturierten Befragung manchmal mehr Einfluss auf das Gespräch nehmen, um falls notwendig die befragte Person mehr zum Forschungsthema zu bewegen. Es kann für eine Forschungsfrage die Möglichkeit gewählt werden, beide Methoden einzusetzen. Man lässt am Anfang eine komplett offene Befragung laufen (qualitativ), woraus man später einen strukturierten Fragebogen erstellt, um der Forschungsfrage quantitativ nachzugehen. Um einen Gesamtüberblick zu bekommen, sind in der Tabelle 1 die Hauptformen der Datenerhebung dargestellt.

QUANTITATIVE FORSCHUNG	QUALITATIVE FORSCHUNG
Strukturierte Befragung	Unstrukturierte Befragung (offen)
Face-to-Face	Face-to-Face
Telefonisch	-
Schriftlich	-

Tabelle 1: Formen der Datenerhebung
Quelle: vgl. [Diek06], S. 273-274

[37] vgl. [Diek06], S. 373-375

4.2 Datenauswertung

Die Auswertung der Daten setzt sich aus mehreren Arbeiten zusammen. Aufgaben wie Datenaufbereitung, Analyse und Interpretation werden zur Auswertung dazu gezählt.[38]

Im Zuge der Beschreibung der Datenauswertung ist es wichtig eine Trennung zwischen der Auswertung der quantitativen und der qualitativen Daten zu machen. Aus diesem Grund werden diese zwei unterschiedlichen Auswertungen getrennt voneinander beschrieben.

4.2.1 Quantitative Daten

Für die Auswertung müssen die Daten ständig auf Vollständigkeit und Plausibilität geprüft werden (während der Feldarbeit). Somit kann man bei unvollständigen Antworten die befragten Personen – Zweck Vervollständigung – noch mal kontaktieren.[39] Die ständige Kontrolle hat zusätzlich den Vorteil, dass man bei einer Falschprogrammierung sofort agieren kann und die Mängel im Fragebogen aufheben kann.

4.2.1.1 Datenaufbereitung

Unter der Aufbereitung der gewonnen Daten versteht man die Codierung und Kategorisierung der Daten. Sollten die Daten nicht elektronisch erfasst geworden sein, dann muss man auch entscheiden mit welchem EDV-Programm man die Daten bearbeiten und analysieren will. Diesbezüglich muss man zwei Aspekte berücksichtigen. Einerseits geht es hierbei um Dateneingabe und andererseits um die statistische Auswertung der eingegebene Daten. Es ist zu empfehlen, diese beide Ausgaben mit unterschiedlichen, jeweils besser geeigneten Programmen zu machen, da nicht jedes Statistikprogramm[40] für die Verwaltung der Daten geeignet ist und nicht jedes Datenerfassungsprogramm[41] alle statistischen Auswertungstechniken besitzen.[42]

[38] vgl. [Atte06], S. 273
[39] vgl. [Atte06], S. 281
[40] SPASS, E-Views, R
[41] dBase, Access, Excel
[42] vgl. [Atte06], S. 282

Die Festlegung der Codes und Kategorien passiert vor der Dateneingabe. Für die Codierung der Daten muss darauf geachtet werden, ob die Daten ordinal-, intervall- oder ratioskaliert sind. Bei geschlossenen Fragen und vorgegebenen Antworten sind die Codes für die Auswertung normalerweise schon vorhanden[43]. Die Definition von Codes ist erst für die Auswertung der offenen Fragen notwendig. Hierfür sollten zuerst alle möglichen Antworten definiert und falls notwendig gruppiert werden. Für die jeweilige Gruppe bzw. Kategorie kommt es dann zur Vergabe von Codes. Die Verwendung von Kategorien hat zusätzlich den Nutzen die Fragen mit ungenauen Angaben bzw. mit Aussageverweigerungen zielgerechter auszuwerten. [44]

Nach der Festlegung der Codes und Kategorien und der erfolgten Eingabe der Daten, müssen die eingegebenen Daten auf Vollständigkeit und Richtigkeit geprüft werden. Es gibt unterschiedliche Möglichkeiten derartige Kontrollen durchzuführen. An dieser Stelle wird nicht mehr genauer auf diese Möglichkeiten eingegangen. Der Fokus liegt im Allgemeinen auf den Kontrollen und dies darf nicht vergessen werden.[45]

4.2.1.2 Datenanalyse

Im Allgemeinen wird es in der Datenanalyse zwischen der deskriptiven und analytischen Statistik unterschieden. Die deskriptive Statistik befasst sich mit der statistischen Beschreibung der Daten. Die Überprüfung – Verifizierung und Falsifizierung – der Hypothesen ist der Gegenstand von der analytischen Statistik. Folgende Kriterien sind für die Analyse der Daten ebenfalls von großer Bedeutung:[46]

- Das Skalenniveau der Daten (unterschiedliche Formen der Darstellung)
- Die Zahl der Variablen (eine Variable vs. Mehrdimensionale Daten)
- Die Zahl der möglichen Nennungen (Einzelnennung vs. Mehrfachnennung)
- Die Zahl und Verbundenheit von Stichproben (Vergleich verschiedener Stichproben miteinander)

[43] Beispiel: Welche Stellung haben Sie im Berufsleben? “1=berufstätig”, “2=nicht berufstätig”
[44] vgl. [Atte06], S. 284
[45] vgl. [Atte06], S. 285
[46] vgl. [Atte06], S. 285-286

Im Endeffekt müssen Hypothesen, welche am Anfang einer Studie aufgestellt und formuliert worden sind, geprüft werden. Hierfür gibt es dementsprechend festgelegte Merkmale[47], welche durch die Daten erhoben werden. Die definierten Merkmale werden dann sowohl einzeln als auch im Zusammenhang mit anderen Merkmalen ausgewertet. In der Einzelauswertung kommt es einerseits zu einer beschreibenden Auswertung[48] und andererseits zu einem analytischen Verfahren[49]. Für die Auswertung von mehreren Merkmalen gibt es eine großen Anzahl an Möglichkeiten für statistische Tests und Verfahren[50]. Die Zusammenführung von nominalen und ordinalen Daten wird sehr oft in Form von Kreuztabellen dargestellt. In einer solchen Tabelle werden sowohl Absolutbeträge als auch Anteilsberechnungen (Prozente) für jedes Merkmal und gesamt aufgestellt. Vorteil solcher Tabellen liegt in Übersichtlichkeit und leichtern Lesbarkeit der gewonnen Daten. Als statistische Analyse-Tools für Auswertungen mehrerer Merkmale können folgende Instrumente aufgezählt werden: Chi-Quadrat-Test, Varianzanalyse, Korrelations- und Regressionsrechnungen und Signifikanztest.[51]

4.2.2 Qualitative Daten

Für die Auswertung von Daten, die in einer qualitativen Studie (das Leitfadeninterview, das fokussierte Interview, das narrative Interview) erhoben wurden, kann man unterschiedliche Methoden wählen. Die Methode des zirkulären Dekonstruierens und die Grounded Theory sind gängige Methoden. Die Auswertung von qualitativen Studien beinhaltet Aufgaben wie Nacherzählung, das offene Kodieren, Erarbeitung eines Themenkataloges und Kategorienbildung. Die Nacherzählung betrifft einzelne Interviews, vor allem wenn Äußerungen sprunghaft sein sollten, um eine thematische Zusammenführung der verschiedenen Aspekte, Bedeutungsgehalte und Erstellung einer Übersicht zu ermöglichen. Durch das offene Kodieren werden Sätze oder Abschnitte auseinandergenommen und mit Codes versehen.

47 Beispiele: Geschlecht, Studienfach, Einkommen, Wünsche
48 Häufigkeitstabelle
49 Test auf Normalverteilung der Daten, Chi-Quadrat-Test
50 Uni-, bi- und multivariate descriptive und analytische Methoden
51 vgl. [Atte06], S. 287-297

Mit der Erarbeitung eines Themenkataloges (bestehend aus Konzepten und Codes) werden die ersten Kategorien und Subkategorien gebildet. Dadurch wird die Auswertung der Ergebnisse einfacher. Die Kategorienbildung ergibt sich durch das erfolgte Interview und das ständige und langsame Hinterfragen[52] der gewonnenen Daten.[53]

4.3 Interpretation

Es muss an dieser Stelle festgehalten werden, dass ein Teil der interpretativen Arbeit schon in der Auswertungsphase passiert. Allein die Überlegung und Entscheidung über die unterschiedlichen Analyseverfahren hat bereits interpretativen Charakter und Einfluss auf die Endergebnisse. Die Interpretation der ausgewerteten Ergebnisse erfolgt in zwei Stufen. Einerseits werden die Ergebnisse der durchgeführten Studie interpretiert. Andererseits kommt es zu einem Vergleich mit anderen Ergebnissen aus anderen Studien und bestehenden Theorien. Ein Forschungsbericht muss folgende vier Elemente beinhalten:

- Problemstellung (Fragestellung)
- Vorgehensweise (Forschungsmethode)
- Ergebnisse (Auswahl der relevanten Ergebnisse)
- Schlussfolgerungen

Die Schlussfolgerungen formulieren meistens Empfehlungen bzw. Handlungsempfehlungen. Es soll hierbei geachtet werden, dass es bei Schlussfolgerungen nicht um Vermutungen gehen darf. Vor allem bei Handlungsempfehlungen muss diese Empfehlung auch wissenschaftlich statistisch fundiert sein.[54]

---

52 Äußerungen, Ergänzungen, impliziten Sinngehalten

53 vgl. [DaHe98], o.S.

54 vgl .[Atte06], S. 298-300

5 Qualitative Datenanalysesoftware

5.1 Ziele von QDA Software

Die computergestützte Analyse qualitativer Daten ist ein Oberbegriff für viele voneinander abweichende Arbeitsweisen und Techniken zur Auswertung qualitativer Daten, die in Form von besonders dafür entwickelter Software (QDA-Software) angewendet werden.

Die erhebliche Geschwindigkeit zeitgemäßer Computer und die Fähigkeit, nahezu unermessliche Mengen an Daten speichern und rasch sortieren zu können, bedeuten einen deutlichen Zuwachs an analytischen Möglichkeiten gegenüber herkämmlichen manuellen Verfahren. Für die Marktforschung sind die neuen computerunterstützen und vollautomatischen Verfahren von sehr großem Nutzen, denn sie gestatten es, qualitative Methoden im Gebiet der Marktforschung anzuwenden, ohne dass dies mit einem kaum akzeptablen Zeit- und Ressourcenaufwand verbunden wäre. Schneller, preiswerter und intersubjektiv leichter überprüfbar. So lassen sich die drei charakteristischen Vorteile der computergestützten Methoden gegenüber der herkömmlichen manuellen Methode qualitativer Sozialforschung zusammenfassen. QDA Computerprogramme gestatten auch eine Fülle von neuen Auswertungstechniken und "Abkürzungsstrategien", die eine durchaus schnellere und gleichwohl methodisch kontrollierbarere Methode gestatten, als sie mit den gängigen althergebrachten sozialwissenschaftlichen Forschungsinstrumenten möglich ist.

Weltweit wurden Computerprogramme entwickelt, wie z.B. Aquad, Atlas-ti, Hyper Research, The Ethnograph, MAXQDA/winmax, Nvivo/Nudist, SPSS, STATISTICA Text-Miner und andere Programme, die immer wieder im Umfeld qualitativer Forschung zur Anwendung gelangen. Derzeit gehören diese analytischen Werkzeuge, also "QDA Software", zum gebräuchlichen Werkzeug der empirischen Forschung.

Moderne QDA Software enthält eine große Menge von Arbeitsmitteln, die die Durchführung verschiedenster Auswertungsschritte ermöglichen, wie Datenexploration, die lexikalische Suche nach Zeichenketten, Wörtern oder Wortkombinationen oder Teilgruppen von Texten und erlauben teilweise auch skalierte Suchprozesse im Sinn von Text Mining, d.h. die sukzessive Suche in Texten und in den Ergebnissen vorangegangener Suchläufe[55].

5.2 Vor- und Nachteile der Produkte

QDA-Software unterscheidet sich in der Herangehensweise grundlegend von der althergebrachten manuellen Methode und weist – je nach eingesetzten Verfahren und deren Parametrisierung – verschiedene Merkmale auf, die man in Vor- und Nachteile gliedern kann[56]:

5.2.1 Vorteile

- Geschwindigkeitsvorteile
- Komplexe Analysevorgänge werden tlw. erst durch Einsatz von QDA Software möglich bzw. sinnvoll
- Flexibilitätssteigerung möglich
- Neue Formen der Datenexploration werden möglich
- Aussagen der qualitativen Sozialforschung können verbessert werden

5.2.2 Nachteile

- Gefahr von Datenverlusten durch „Ungenauigkeit“
- Versuchung, eine Analyse „quick and dirty“ d.h. schnell und ungenau durchzuführen
- Gewonnene Transparenz kann durch die Menge und die hohe Komplexität des Datenmaterials wieder verloren gehen

55 vgl. [BuHo07], S. 715-717
56 vgl. [Frie08], o.S.

- Aufgrund von Fehlinterpretationen und nicht transparentem Einsatz von Algorithmen besteht die Gefahr von sinnlosen Analysen, deren Aussagen sehr geringen oder keinen Wert haben

5.3 Softwarelösungen

In Anwendungsbereich der Datenanalyse existiert eine Vielzahl – größtenteils kommerzieller – Produkte, die fast alle ihren Ursprung in der quantitativen Datenanalyse haben und daher im Bereich der qualitativen Analyse mehr oder minder große Schwierigkeiten haben. Auf Basis einer durchgeführten Evaluierung bei mehreren Unternehmen im Bereich der Meinungs- und Marktforschung ergibt sich eine Liste von häufig eingesetzten Produkten. In den vergangenen Jahren haben die meisten Hersteller bekannter Produkte (SPSS, SAS, …) zusätzlich zu den bereits existierenden Softwarelösungen auch noch optionale Zusatzprodukte oder -module herausgebracht, die sich hauptsächlich mit der Analyse qualitativer Daten und teilweise auch intensiv mit der Analyse unstrukturierter Texte beschäftigen.

Im folgenden Abschnitt wird eine Reihe bestehender QDA Softwarelösungen kurz vorgestellt, die im Zuge der Recherchen als häufig verwendet einzustufen sind. Diese Aufzählung von Softwareprodukten und ihren charakteristischen Eigenschaften erhebt keinen Anspruch auf Vollständigkeit.

5.3.1 STATISTICA Text-Miner

URL: http://www.statsoft.de/

STATISTICA Text-Miner ist eine optionale Erweiterung des STATISTICA Data-Miner. Die Software bringt unstrukturierte Textdaten in eine verständliche und verwendbare Form, damit die Prozesse der Entscheidungsfindung unterstützt werden können. Dabei bedient sich STATISTICA Text-Miner einer Vielzahl von Algorithmen zur Verarbeitung unstrukturierter Texte. In der Realität sind die erhobenen Daten oft nicht direkt auswertbar und müssen erst für die Analyse aufbe-

reitet werden. Diese Software bietet eine Möglichkeit, verborgene Informationen aufzudecken und vor allem stehen ausgefeilte Meachanismen für die Verarbeitung unstrukturierter Texte zur Verfügung. Das Produkt lässt sich sehr gut auf Textdokumente oder Webseiten anwenden, bietet aber zusätzlich auch die Möglichkeit, Bilder, Sounddateien und Videos zu verarbeiten. Unstrukturierte Informationen können kategorisiert, gruppiert oder in andere Analysen einbezogen werden. Der Hauptanwendungsbereich von STATISTICA Text-Miner ist die Analyse offener Interviewfragen oder allgemeiner unstruktierter Texte und es können damit große Mengen an Dokumenten rasch analysiert werden (wie auch z.B. Emails von Kunden zu klassifizieren, um Reklamationen identifizieren zu können und diese einer effizienten Bearbeitung zuführen zu können). Die Daten werden mit Text-Miner aufbereitet, es wird ein Wortindex erstellt und es werden folgende Schritte durchgeführt[57]:

- *Es lassen sich Ausschlussregeln anwenden, um gebräuchliche, aber analytisch nicht relevante Wörter wie "ein/e", "der", "die", "das" oder "ist" auszuschließen. Ein Wortstamm-Algorithmus stellt sicher, dass Wörter wie "gereist" und "reisen" als Fälle des Worts "Reise" zählen.*
- *Der STATISTICA Text Miner enthält Ausschlusslisten und Wortstamm-Algorithmen für Deutsch, Dänisch, Holländisch, Englisch, Französisch, Italienisch, Portugiesisch, Spanisch, Schwedisch und weitere Sprachen. Die Ausschlusslisten können vom Anwender bearbeitet und erweitert werden. Das Design der Software erlaubt die Unterstützung zusätzlicher Sprachen mit geringem Aufwand.*
- *Als nächstes erzeugt die Software aus den bereinigten Dokumenten einen Index, um die Häufigkeiten aller Wörter für alle Dokumente zu zählen. Diese Information ist die Basis für alle folgenden numerischen Analysen.*
- *Vor Erstellung einer STATISTICA-Datei mit den Häufigkeiten, die eine Verdichtung der Informationen aus den Dokumenten darstellt, können verschiedene zusätzliche Filter angewandt werden, die sich auf die Länge und Buchstabenstruktur der Worte beziehen. Die Häufigkeiten der Wörter lassen sich skalieren (unter Berücksichtigung der Häufigkeit Ihres Auftretens), transformieren (z. B. log-transformieren) und "komprimieren" (über einen Algorithmus zur Singulärwertzerlegung, der aus den Worthäufigkeiten zugrunde liegende Dimensionen extrahiert).*
- *Die resultierende Datei mit den numerischen Informationen (Dimensionen, Häufigkeiten, relativen Häufigkeiten usw.) steht dann für weitere Analysen zur Verfügung.*

[57] vgl. [o.V.o.J.e], o.S.

- *Es gibt verschiedene Optionen, um die aus dem Text extrahierten Informationen in das STATISTICA-Datenblatt oder in externe Datenbanken zu übertragen*

Im Rahmen der Analyse werden auf Basis der Eingabetexte numerische Zusammenfassungen gebildet, die dann innerhalb des Programmes (STATISTICA Data-Miner Basisversion) mit den umfangreichen, in der Software verfügbaren statistischen Modellen analysiert werden können.

Einfachste Analysen der betrachteten Dokumente erzeugen nur eine Worthäufigkeitsanalyse der Wörter, die am häufigsten verwendet werden. Durch Abbildung der Dokumente auf Dimensionen können Streuungsdiagramme erstellt werden, mit deren Hilfe die Ähnlichkeit von Dokumenten beurteilt werden. Ebenfalls durch Abbildung der Dokumente auf Dimensionen, die auch auf Worthäufigkeiten basieren, kann gleichzeitig das Mapping von Dokumenten und Wörtern erzeugt werden, das Bedeutungen der Dokumente reflektiert. Zur Analyse können prädiktive Data-Mining Techniken und Cluster Techniken wie EM oder K-Means eingesetzt werden, um Gruppen ähnlicher Dokumente zu identifizieren.

5.3.2 SPSS

URL: http://www.spss.com/de/spss/

SPSS steht für „Statistical Package for the Social Sciences“. SPSS wird seit 1968 in den USA entwickelt und ist die meistverbreitete Software, die im sozialwissenschaftlichen Bereich verwendet wird. Heute wird dieses Statistikprogramm als Standardsoftware in den verschiedensten Unternehmen, Schulen, Universitäten und Institutionen verwendet. Derzeit ist die Version 16.0 die aktuelle Version am Markt. SPSS ist eigentlich ein statistisches Rechenprogramm und ein Anwendungssystem mit graphischen Tools und ideal verwendbar für einfache und gehobene Analyse quantitativer Daten. SPSS enthält viele statistische Prozeduren für die Analyse von Häufigkeiten, Kreuztabellen, deskriptive Statistiken, Faktorenanalyse, Regression und Clusteranalyse. Zusatzmodule gibt es auch für Regressionsanalyse, Advances Models, Tables, Trends und Categories und für weitere

spezifische Probleme. Mit SPSS können in in wenigen Schritten wertvolle Informationen ermittelt werden und aussagekräftige Grafiken und Tabellen erstellt werden. Die Ergebnisse lassen sich als 2D-, 3D und pivotierbare Grafiken visualisieren. Die folgende Tabelle enthält einige weitere QDA Produkte:

	Atlas.ti www.atlasti.com	Hyper Research2.6 www.researchware.com	MAXqda www.maxqda.com	QSR Nvivo www.qsr.com.au
DATA ENTRY	Media types: Text (txt, rtf, doc), graphic (jpeg, bmp, tiff and others), audio (wav, au, snd, mp3), video (avi, mpeg, mov, qt)	Media types: Text (txt), graphic (jpeg, bmp, gif, png, pict), audio (wav, aif, mov, mp3), video (avi, mpeg, mov, swf, gif)	Media types: rich text Editing of coded documents supported Option of five character sets: West European, Arabic, Cyrillic, Greek, Hebrew.	Media types: Rich text not including embedded objects like pictures or tables Editing of coded documents supported
CODING				
Smallest unit that can be coded	1 character	1 character	1 character	1 character
On-screen coding	By marking the text with the mouse; quick-coding and in-vivo coding possible.	By marking the text with the mouse, quick-coding and in-vivo coding possible.	By marking the text with the mouse; quick-coding and in-vivo coding possible.	By marking the text with the mouse; quick-coding and in-vivo coding possible.
Display of code words and boundaries	Code words and boundaries are displayed in the right margin (applies to text and graphic documents).	Code words are displayed in the left margin.	Code words and boundaries can be displayed in the right or left margin. Different colour attributes for systematic differentiation of codes.	Code words and coding stripes can be displayed in the margin.
TEXT SEARCH AND AUTO CODING	String, category and GREP searches. Results of text searches can automatically be coded, the process can be controlled	Supports text searches for phrases or words. All search results can be auto-coded.	Supports text searches for unlimited number of words or phrases in documents, coded text passages and memos. Searches can be restricted to selected texts. Search strings can be combined using two logic operators. Search lists may be saved for further use. All search results can be auto-coded.	String, category and GREP searches in all or selected documents and nodes. Results of text searches can automatically be coded, the process can be controlled
Further analysis of search results	Search queries can be saved to super codes and used in further searches. Super codes can be turned into regular codes by creating snapshot codes and subjected to further analysis. Reports of search results can be copied into a memo and assigned as primary document. This way, they can be coded and analysed further.	HyperResearch offers the unique feature of testing hypothesis. This means that you can define rules. If a rule applies, action and consequences can be specified, like adding or removing codes to/from a case.	As the window with the search results and the originally data can be displayed side by side, further coding or modifications to the coded data segments are possible. The entire search result can also be coded to a new or existing code.	As search results are saved to a node, they can be handled like any other coded text, i.e. browsed and coded further.
QUANTITATIVE OUTPUT	List of code words including frequencies (plain text) / cross-tabulation of code words and primary documents (plain text) / list of word frequencies (tab-delimited or comma separated file) / document variables as tab-delimited or comma separated file / entire coding system as SPSS syntax file	Code frequencies can be exported (plain text)	List of codes including frequencies as tab delimited file / list of coded Nodes in all or a set of documents including character counts, numbers of paragraphs, documents and passages coded / character counts or number of passages coded at each node / results of matrix searches are displayed in table format. All tables can be exported as text or dat (SPSS) file. segments (all or only selected code) as tab delimited or html file / variable matrix as tabdelimited file.	

Tabelle 2: QDA-Produkte
Quelle: vgl. [Frie04], o.S.

5.4 Zusammenfassung

Abschließend sei hier festgehalten, dass es viele weitere Produkte gibt, die sich mit der Analyse quantitativer und qualitativer Daten beschäftigen. Im Laufe der Recherchen zu dieser Diplomarbeit hat sich herausgestellt, dass viele Organisationen neben den erwähnten Softwareprodukten noch zusätzlich selbst implementierte Speziallösungen im Einsatz haben, die einerseits im Bereich der Datenvorbereitung und andererseits ergänzend zu den vorgestellten Softwarelösungen bestehen, wenn die eingesetzten Softwarelösungen die Anforderungen der Benutzer nicht bzw. nur teilweise erfüllen.

Diese Auflistung und die Reihenfolge ihrer Erwähnung ist keine Wertung der Produkte und je nach gestellten Anforderungen der Benutzer ist eine Entscheidung für eine Softwarelösung im Einzelfall durchzuführen. Die genannten Eigenschaften und Merkmale sind diejenigen Merkmale, deren Erwähnung aus Sicht des Autors sinnvoll sind.

6 Analyse unstrukturierter Texte

In der heutigen Zeit ist der Bedarf an aussagekräftiger Datenanalyse infolge der häufig eingesetzten modernen Technologien sehr groß. Eine der Hauptursachen dafür, dass häufig in bestehenden Daten- und Informationssammlungen nur schwer und aufwändig bestimmte Informationen gefunden werden können, ist die Tatsache, dass die meisten Informationen in Form unstrukturierter Texte digital gespeichert sind. Speziell im Bereich der Wirtschaft geht man davon aus, dass zwischen 80 und 85 Prozent der relevanten Informationen in Form von Freitexten gespeichert sind. Es ist zwar schon ein Schritt in die richtige Richtung, dass Informationen digital abgelegt werden, allerdings fehlt den meisten Dokumenten die Struktur, was sie erst zu digital auswertbaren Texten machen würde.[58]

Die Komplexität und die grundlegenden Probleme des Verstehens und Interpretierens von unstrukturierten Texten bewirkt, dass vorhandenes Wissen nur schwer zugänglich ist. Daran ändern auch die Rechenleistung moderner Computer und der Einsatz ausgefeilter Algorithmen nicht sehr viel. Auf Grund des Informationsvolumens sind die meisten „Datensammlungen“ zu groß, um gelesen, durchsucht oder verwendet werden zu können. Die Folge ist, dass Daten und Informationen, die nicht zu gut durchsuchbarem Wissen umgewandelt werden können, als physische Dokumente oder in Form von digitalen Dokumenten in sogenannten Data-Repositories verbleiben und so gut wie unzugänglich bleiben, weil sie zwar vorhanden, aber nicht suchbar bzw. identifizierbar sind. Aus diesem Grund haben sich im Forschungsbereich des Wissensmanagements unterstützende Methoden und Verfahren entwickelt, die den Versuch unternehmen, aus Daten Wissen zu generieren, also computerunterstützt Daten zu analysieren und die Ergebnisse der Analysen verfügbar zu machen[59]. In den letzten Jahren haben sich verschiedene Modelle zur Repräsentation von Texten und Verfahren zur Ähnlichkeitsbestimmung von Texten und zur automatischen Klassifikation herauskristallisiert und es besteht eine Hoffnung, dass weitere Verfahren entwickelt werden können.

---

[58] vgl. [o.V.05], o.S.

[59] vgl. [Moen00], S. 61

Alltägliche Systeme zur Informationsaufbereitung, Informationsanalyse und zum Erschließen von Informationen sind fast nur dazu geeignet, auf Basis einer Volltextsuche die gefundenen Ergebnisse in Listenform anzuzeigen, wobei außer booleschen Operatoren so gut wie keine Möglichkeit der Ergebnisverfeinerung verfügbar ist. Komplexe Fragestellungen können oft maschinell nicht beantwortet werden und müssen daher manuell bearbeitet werden, was hohen Zeitaufwand mit sich bringt. Wertvolle Informationen bleiben ungenutzt, was besonders bei Zusammenhängen zwischen verschiedenen Texten zu deutlich schlechteren Ergebnissen führt, als dies durch manuelle Bearbeitung möglich ist.[60]

Bei komplexen Fragestellungen ist eine Beantwortung der Fragestellung nicht in Form einer einfachen Liste gefundener Begriffe durchführbar bzw. sinnvoll. Um qualitativ hochwertige Ergebnisse zu erhalten, muss man eine Kombination von Verfahren anwenden, die sich je nach Themen- und Anwendungsgebiet unterscheidet. Es ist also kein allgemeines Verfahren absehbar, das eine universelle hochqualitative Analyse unstrukturierter Texte in verschiedenen Anwendungs- und Sachgebieten verspricht.

Die bekannten Verfahren zur automatischen Analyse unstrukturierter Texte lassen sich unterteilen in Indexierungsverfahren, Stemmingverfahren, statistische Verfahren, linguistische Verfahren und Reduktionsverfahren. Das Hauptanwendungsgebiet der Reduktionsverfahren ist der Bereich des Information Retrieval (IR).

Die grundlegenden Verfahren können funktional eingeteilt und qualitativ bewertet werden und im Einzelfall ist aufgrund der Funktionalität und der Qualität der einzelnen Verfahren eine Entscheidung zu treffen, welche Verfahren und in welcher Kombination die ausgewählten Verfahren eingesetzt werden. Jedes Verfahren und jeder Algorithmus hat Vor- und Nachteile.[61]

60 vgl. [Webe03], S. 7
61 vgl. [Hali05], S. 3-4

6.1 Information Retrieval (IR)

Information Retrieval ist ein Fachgebiet des Wissensmanagements. Es ist auch ein Bereich der Informationswissenschaft, der Computerlinguistik und auch der Informatik. Information Retrieval konzentriert sich auf die Entwicklung intelligenter Suchverfahren mit dem Ziel, dem Benutzer möglichst zutreffende Antworten auf seine Fragen zu liefern. Das Hauptziel des IR ist neben der Optimierung des Auffindungsprozesses die Rückgewinnung von Informationen und Wissen aus Dokumenten und anderen unstrukturierten Texten unter Zugrundelegung eines konkreten und klaren Informationsbedürfnisses. Es handelt sich bei IR also um eine Kombination von Verfahren, die mit der Auffindung, Aufbereitung und Speicherung von Dokumenten zu tun haben[62]. Unstrukturierte Texte bzw. Daten bestehen aus Textdaten, die kein deutlich erkennbares Schema haben. **Das Grundproblem besteht darin, dass Computer die Bedeutung unstrukturierter Texte nicht verstehen können**. Eine mögliche Lösung für die Verbesserung der Auffindbarkeit unstrukturierter Texte ist es, den Texten eine Struktur zu geben und sie z.B. in einer Datenbank zu speichern[63]. Beim IR werden manuelle Verfahren mit automatischen Verfahren kombiniert, wobei die automatischen Verfahren grundsätzlich in die Extraktion von Schlüsselwörtern und die Extraktion von Schlüsselsätzen unterschieden werden. Die Basis für die Extraktion von Schlüsselwörtern und –sätzen stellt die automatische Indexierung von Texten dar.[64]

6.2 Indexierung

Unter Indexierung versteht man Inhaltserschließung. Die Erkennung einer inhaltlichen Repräsentation von Dokumenten unter Schlagwörtern (Deskriptoren) macht es erst möglich, nach Schlagwörtern zu suchen. Die Untersuchung der Schlagwörter muss in zwei grundlegend verschiedene Methoden unterschieden

62 vgl. [Kauf01], S. 2
63 vgl. [Hald02], S. 6
64 vgl. [Hali05], S. 4-5

werden: In die Extraktionsmethode, bei der einem Dokument Schlagwörter zugeordnet werden, die im Dokument vorkommen, und in die Additionsmethode, bei der dem Dokument künstliche Schlagwörter zugeordnet werden, die im Dokument nicht enthalten sind. Dokumente ohne Text (z.B. Audio-, Video- und Bilddokumente) können ausschließlich mittels additiver Methoden indexiert werden[65].

6.2.1 Manuelle Indexierung

Die manuelle Indexierung ist eine nach wie vor wichtige Methode für die Auswahl der Schlagwörter aus unstrukturierten Texten. Bei der manuellen Indexierung handelt es sich um intellektuelle Arbeit, bei der eine qualifizierte Person die Indexierung durchführt. Aufgrund des hohen Zeitaufwandes für die manuelle Indexierung erscheint diese Vorgangsweise zwar qualitativ hochwertige Ergebnisse zu liefern, eine Kosten-Nutzen-Rechnung muss aber in weiten Bereichen unzweifelhaft zu Gunsten computergestützter oder vollautomatischer Indexierungsmethoden ausfallen, weil diese Methoden zwar meist geringere Qualität aufweisen, aber wesentlich effizienter sind.. Ein weiteres Problem der manuellen Indexierung ist die Tatsache, dass verschiedene Personen aufgrund unterschiedlicher individueller Bewertung der Texte zu unterschiedlichen Ergebnissen gelangen können. Experimente haben ergeben, dass computergestützte oder vollautomatische Verfahren der manuellen Indexierung in Bezug auf die Qualität der Indexierung meist nur geringfügig unterlegen sind. Aufgrund dieser Zusammenhänge (Kosten-Nutzen-Rechnung, Qualitätsanforderungen) werden heutzutage meistens computergestützte Methoden der Indexierung angewendet[66].

65 vgl. [Lucko.J.], o.S.
66 vgl. [Bünz01], S. 23

6.2.2 Computergestützte Indexierung, vollautomatische Indexierung

Von computergestützter Indexierung wird gesprochen, wenn sowohl ein Mensch als auch ein IT-System an der Indexierung beteiligt sind. Je nach verwendetem System ist die Interaktion der Person mit dem System in unterschiedlichem Umfang und in unterschiedlichen Bereichen zu finden. Meist handelt es sich dabei um Kontroll- und Korrekturfunktion, die von einem Menschen ausgeübt werden, um die Ergebnisse der maschinellen Verarbeitung einer Plausibilitätsprüfung zu unterziehen[67]. Mit Hilfe von Software wird entweder nur der Index erstellt oder auch die Relevanz der Worte für das betrachtete Dokument ermittelt und vom Menschen dann aus den indexierten Worten die Menge der relevantesten Worte ermittelt, die man in der Folge als Schlagwörter bezeichnet. Der Mensch hat bei computergestützten Verfahren zumindest Einfluss auf die Indexierung[68].

Automatische und computergestützte Indexierung müssen immer im Kontext der geforderten Qualität und Geschwindigkeit betrachtet werden. Manuelle Indexierung führt, wie bereits oben gesagt, zu höherem Aufwand und teilweise zu besseren Resultaten, wohingegen die Vorteile der computergestützten und der vollautomatischen Indexierung vor allem im Zeitaufwand zu finden sind[69].

Bei der **vollautomatischen Indexierung** erfolgt die Indexierung ohne Benutzerinteraktion ausschließlich durch ein IT-System. Daraus ergibt sich, dass diese Form der Indexierung auf jeden Fall als Ausgangsdaten Dokumente in digitaler Form benötigen. Im Vergleich zur manuellen Vorgangsweise können bei der vollautomatischen Indexierung multiple Methoden automatisch kombiniert werden (statistische, linguistische und andere Methoden), mit denen die Eingabedokumente analysiert und verarbeitet werden können, solange es sich um maschinell lesbare Dokumente handelt aus denen die notwendigen Informationen extrahierbar sind. Eine Abwägung der Vor- und Nachteile der Arten (manuell, computergestützt, vollautomatisch) fällt aufgrund der Komplexität der Bewertungskriterien

67 vgl. [Lucko.J.], o.S.
68 vgl. [Tres07], o.S.
69 vgl. [Lucko.J.], o.S.

schwer: Die automatischen Indexierungsmethoden sind eigentlich Extraktionsverfahren die hauptsächlich mit den im Dokument enthalten Ausdrucksweisen arbeiten. Folgt man Halip[70], dann gelingt eine Abschätzung der Vor- und Nachteile: Die Vorteile der automatischen Verfahren sind eher auf der Kostenseite und im Zeitmanagement zu finden, die Anzahl der Fehler bleibt niedrig, da nur die Deskriptoren extrahiert werden und weiters werden Suchanfragen natürlichsprachlich bearbeitet. Als Nachteil der automatischen Indexierung stellt sich die Situation dar, dass bei den automatischen Verfahren nach wie vor einige Probleme nicht gelöst sind, wie z.B. *Die nichtlexikalische Ausdrucksweise verstehen und lexikalisieren*, *die Mehrdeutigkeit der Ausdrucksweise zu überwinden, die Mannigfaltigkeit der Ausdrucksweise zu erkennen, die unvoraussehbaren Ausdrücke durch ein uniformisiertes Vokabular ersetzen, die Lückenhaftigkeit der natürliche Sprache*[71].

6.3 Verfahren der automatischen Indexierung

In der Literatur wird die automatische Indexierung von Texten üblicherweise in fünf Kategorien unterteilt: Freitextinvertierung ist eigentlich die einfachste Methode der Stichwortextraktion ohne linguistische Bearbeitung, statistische Verfahren, Mustererkennungsverfahren, computerlinguistische Verfahren und begriffsorientierte Verfahren.

6.3.1 Volltextinvertierung

Volltextinvertierung bzw. Freitextinvertierung ist eines der schnellsten und wirtschaftlichsten Verfahren und bei diesem Verfahren werden alle Wörter des Textes für die Stichwortextraktion verwendet. Ausschließlich die sogenannten Stoppwörter (wie Pronomen, Hilfsverben, Konjunktionen) werden vor der Volltextinvertierung entfernt und somit nicht verwendet. Alle nicht als Stoppwörter in Stopplisten stehenden Wörter eines Textes werden beim Freitextinvertieren in den Index aufgenommen. Dieser Vorgang wird bei der Schilderung der Wortnormalisierung der

[70] vgl. [Hali05], S. 6
[71] [Hali05], S. 6

Informationslinguistischen Verfahren näher beschrieben. Alle restlichen Wörter werden zur Stichwortextraktion verwendet. Es wird zuerst eine invertierte Datei angelegt, in der alle Wörter mit einem Verweis auf ihre jeweilige Quelle im Dokument eingetragen werden. Freitextinvertierung kann als das einfachste Verfahren eingestuft werden, weil nach Dokumenten nicht über den Text des Dokumentes gesucht wird, sondern es wird in der invertierten Indexdatei gesucht[72]. Die Freitextinvertierung besteht aus zwei Schritten[73]:

Erstens wird eine invertierte Datei gebildet und und zweitens wird durch eine Extraktion aller Wörter ohne Stoppwörter mit dem Verweis auf das Dokument, gesucht. Das Verfahren ist wirtschaftlich interessant, ist aber von der Indexierungssprache abhängig. Für Benutzer ist dies das einfachste Verfahren. Nachteilig wirkt sich aus, dass oft unerwünschte Ergebnisse resultieren und Informationsverlust droht, weil aufgrund fehlender Synonymbehandlung die Ergebnisse verfälscht sein können[74].

Obwohl Volltextinvertierung automatisch durchgeführt werden kann und alle Wörter in einem Text in einer invertierten Liste gespeichert und verarbeitet werden, kann Volltextinvertierung nicht zur automatischen Indexierung gezählt werden[75].

Beispiel für Textinvertierung[76]

Dokument 1: Eine Schwalbe macht noch keinen Frühling

Dokument 2: Frühling lässt sein blaues Band wieder flattern durch die Lüfte

Schritt 1: Ein Index in der Reihenfolge des Vorkommens im Text wird erstellt und jedes Wort bekommt bei seinem ersten Vorkommen eine Adresse in Form einer fixen Nummer

72 vgl. [Weis00], o.S.
73 vgl. [Kais97], o.S.
74 vgl. [Hali05], S. 7
75 vgl. [Nohr05], S. 217
76 vgl. [Hali05], S. 21

Adr.	Stichwort
1	eine
2	Schwalbe
3	macht
4	noch
5	keinen
6	Frühling
7	lässt

Adr.	Stichwort
8	Sein
9	blaues
10	Band
11	Wieder
12	Flattern
13	Durch
14	die
15	Lüfte

Tabelle 3: Ergebnisse der Index
Quelle: [Hali05], S. 21

Danach wird der Index alphabetisch sortiert.
Schritt 2: Stoppwörter werden aus dem Index eliminiert und das Ergebnis wird als invertierte Datei mit Verweisen auf die Quelldateien der Indexwörter fertig gestellt.

Adr	Stichwort
10	band
9	blaues
14	die
13	durch
1	eine

Adr	Stichwort
12	flattern
6	frühling
5	keinen
7	lässt
15	lüfte

Adr	Stichwort
3	macht
4	noch
2	schwalbe
8	sein
11	wieder

Tabelle 4: Alphabetisch Sortierung
Quelle: [Hali05], S. 21

Adr	DocID
10	2
9	2
12	2
6	1,2

Adr	DocID
5	1
7	2
15	2
3	1

Adr	DocID
2	1
11	2

Tabelle 5: Invertierte Datei
Quelle: [Hali05], S. 21

Das nun vorliegende Ergebnis kann sehr einfach in einer Datenbank gespeichert werden, indem drei Tabellen erzeugt werden, die man z.B. Document, Word, WordInDocument nennen könnte.

6.3.2 Statistische Verfahren

In Vergleich zur Volltextinvertierung gehen die statistischen Verfahren davon aus, dass nicht alle Terme eines Textes gleich wichtig sind und auch nicht alle Terme eines Textes geeignete Kandidaten für Schlagworte sind. In der Praxis bilden statistische Verfahren die bekannteste Klasse von Verfahren. Die statistischen Verfahren arbeiten meist auf Basis von Termen, die mittels sprachorientierter Ansätze ermittelt werden. Alle statistischen Verfahren funktionieren prinzipiell so, dass die Bedeutung eines indexierten Wortes anhand der Vorkommenshäufigkeit des Wortes im Text und anhand anderer Faktoren bewertet wird. Häufig wird darüber hinaus noch die Anzahl der Vorkommen des Wortes in gleichartigen oder übergeordneten Gesamttexten in die Berechnung einbezogen, was zu einer Relativierung der Gewichtung führt und diese im Gesamtkontext korrekter und gegenüber Ausreißern unempfindlicher macht[77]. Bei diesen Verfahren geht es eigentlich um die Zählung der Wörter und um die Ermittlung der Relevanz von Wörtern für die Gewichtung. Ziel der statistischen Verfahren ist es, in einem Dokument die Bedeutung von Begriffen mit der Häufigkeit ihres Auftretens in einem Dokument und in anderen gleichartigen Dokumenten in Zusammenhang zu bringen. Damit wird eine Identifizierung der inhaltsrelevanten Stichwörter in einem Dokument erreicht. Mit Hilfe dieser Verfahren werden Wörter und Datensätze gezählt und die gezählten Wörter werden in Relation gesetzt, woraus eine Relevanz von Termen für die Beschreibung des Inhaltes eines Textes abgeleitet werden kann[78].

Die statistischen Indexierungsansätze basieren auf zwei Grundannahmen[79]:

- Nicht alle Terme eines Dokuments sind als Indexterme geeignet. Es muss eine Auswahl getroffen werden.
- Nicht alle ausgewählten Indexterme haben bzgl. der inhaltlichen Bedeutung die gleiche Wertigkeit und es muss eine Gewichtung der Indexterme vorgenommen werden.

77 vgl. [Henr02], S. 9
78 vgl. [Henr02], S. 9
79 vgl. [Nohr05], S. 217

In den folgenden Punkten werden die verschiedene Methoden und Faktoren der Stichwortextraktion zur Termgewichtung und die entsprechende Modelle dafür vorgestellt. Die Elemente für Termgewichtung sind[80]:

- **Einfache Zählung von Worthäufigkeiten**

  Die einfache Zählung von Worthäufigkeiten baut auf dem Modell von H. P. Luhn (1958) auf und ist als eine der ersten Methoden für die automatische Indexierung bekannt geworden. Dieses Modell basiert auf allen Wörtern eines Dokumentes und auch in der gesamten Datenbasis werden Wörter gezählt, woraus dann ermittelt wird, wie oft ein Wort vorkommt. Eine Voraussetzung muss bei diesem Verfahren berücksichtigt werden: Nicht alle Begriffe, die in der Datenbasis oft vorkommen, sind als Schlagwörter oder Deskriptoren nützlich. Der Vorteil des Luhn-Modells liegt in der einfachen Durchführung des Zählens von Termen und der einfachen Sortierung nach Häufigkeit. Ein Nachteil dieses Modells ist die Schwierigkeit der Analyse der unteren und oberen Schwellwerte.

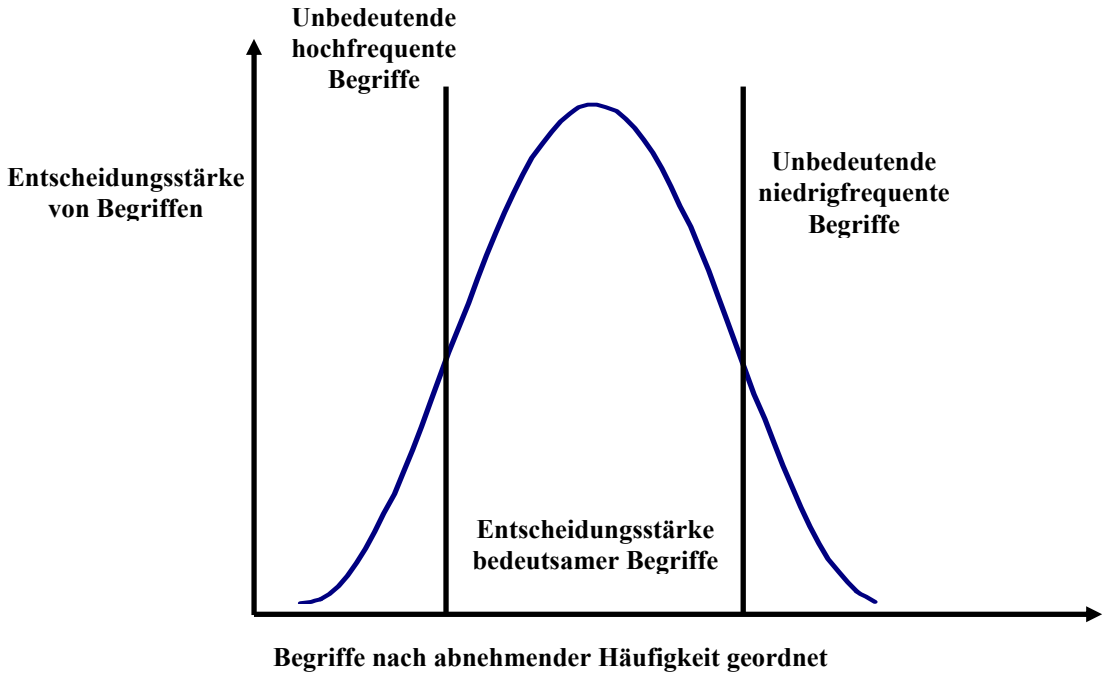

Abbildung 6: Entscheidungsstärke bedeutsamer Begriffe
Quelle: vgl. [Kais93], S. 27, eigene Darstellung

80 vgl. [Weis00], o.S.

Das Modell von Luhn kann durch die Elimination von Stoppwörtern und das Reduzieren der unterschiedlichen Wortformen auf ihren Wortstamm deutlich verbessert werden.

- **Dokumentspezifische Wortgewichtung:**
  Aufgrund der unterschiedlichen Textlänge verschiedener Dokumente kann die einfache Zählung der Worthäufigkeit in einem Dokument für deren Inhalt fehlerhafte Ergebnisse liefern. Aus diesem Grund wird häufig zur Ermittlung der relativen Häufigkeit des Vorkommens von Wörtern die Methode zur Ermittlung der relativen Häufigkeit von Wörtern verwendet:

$$TF_{td} = (ld\ [FREQ_{td} + 1]) / ld\ GESAMT_{td}$$

- **Gewichtung nach der Position im Text:**
  Die dokumentspezifische Wortgewichtung wird durch die Position der Wörter im Dokument verbessert. Bei diesem Verfahren wird ein Text in mehrere Blöcke unterteilt. Unter der Annahme, dass Terme am Anfang des Textes wichtiger sind als Terme in der Mitte des Textes, muss berücksichtigt werden, dass ein Vorkommen eines Schlagwortes am Anfang des Textes verglichen mit in der Mitte bzw. dem Ende des Textes mit einem höheren Gewichtungsfaktor bewertet wird.[81].

- **Inverse Dokumenthäufigkeit:** Die inverse Dokumenthäufigkeit basiert auf dem Modell von Sparck Jones und setzt voraus, dass die Vorkommenshäufigkeit eines Wortes in einem Text auch die Bedeutung des Wortes für den Text repräsentiert. Es wird zuerst nach einem bestimmten Wort in den Texten gesucht. Danach wird die Anzahl der gesamten Texte gezählt. Aus dem Verhältnis zwischen diesen beiden Messungen setzt sich die inverse Dokumenthäufigkeit zusammen. Ein Wort wird wichtiger gewertet wenn weniger Dokumente dazu in der Datenbank bzw. den zu betrachtenden Daten zur Verfügung stehen. Bei der inversen Dokumentenhäufigkeit wird logarithmisch gerechnet, damit der Wertbereich nicht allzu groß wird.

$$IDF(t) = (\log_2 N / n) + 1$$

81 vgl. [Stoc00], S. 162

- **Wortabstand:**

  Wenn sich der Suchbegriff aus mehreren Wörtern zusammensetzt, dann ist der Wortabstand relevant. Falls die einzelnen Wörter knapp nebeneinander stehen (kleiner Wortabstand), wird die Gewichtung für die Suche höher ausfallen, als im umgekehrten Fall (also bei weitem Abstand zwischen den betreffenden Wörtern). Das Auftreten von Mehrwortkombinationen mit geringem Wortabstand kann als Indiz dafür gesehen werden, dass eine Mehrwortkombination relevant für das Dokument ist.

6.3.3 Termgewichtungsmethode

Termgewichtung basiert auf der passenden Gewichtung der Indexterme bzw. auf der passenden Auswahl der Schlagwörter (Deskriptoren). Dieses Verfahren ist dazu geeignet, eine Differenzierung von statistischen Häufigkeiten des Auftretens der einzelnen Terme zu ermitteln. Das statistische Maß wird als semantischer Indikator eingesetzt. Die Häufigkeit eines Worts in einem Dokument wird als Indikator verwendet, wie repräsentativ das Wort für den Inhalt des Dokumentes ist. Die Termfrequenz (TF) in einem Dokument wird als Maßeinheit für das Gewicht eines Worts in einem Dokument und auch als Maß für die Bedeutung des Inhalts verwendet.

Der Termfrequenzansatz hat zwei Voraussetzungen:

- Terme die häufig auftreten, haben im Vergleich mit selten auftretenden Termen für die Bedeutung eines Dokuments eine höhere Signifikanz.
- Selten auftretende Terme haben im Vergleich mit häufig auftretenden Termen innerhalb einer Menge von Dokumenten einen höheren Diskriminanzeffekt.

Diskriminanzeffekt[82]:

Wenn die Qualität von intellektueller Indexierung bewertet wird, liegen zwei Kriterien zu Grunde

1. Gute Deskriptoren sind **signifikant für den Inhalt** des erschlossenen Dokuments.

[82] vgl. [o.V.08b], o.S.

2. Gute Deskriptoren sind dazu geeignet, die unterschiedlichen Dokumente einer Dokumentensammlung inhaltlich voneinander zu unterscheiden. (**Diskriminanzeffekt**)

Die Termfrequenz (TF) wird mit folgender Formel errechnet und es wird die sich daraus resultierende Signifikanz berechnet:

(1) $TF_{td} = FREQ_{td} / GESAMT_{td}$

Abkürzung: t = Term
d = Dokument
TF = Termfrequenz
$FREQ_{td}$ = Frequenz eines Terms im Dokument
$GESAMT_{td}$ =

Termfrequenz ist die Häufigkeit eines Wortes im Dokument durch die Anzahl aller Wörter des Dokuments.

(2) $TF_{td} = (ld [FREQ_{td} + 1]) / ld\ GESAMT_{td}$

Meist wird ebenfalls mit logarithmischen Verfahren gerechnet, um die Spannbereiche der Gewichte klein zu halten, weil bei langen Texten sonst Ergebnisse mit vielen Nachkommastellen als Ergebnis herauskommen, was durch logarithmische Verfahren vermieden werden kann.

Wenn der Term t in einem Dokument mit 200 Wörtern 4 Mal auftritt, so ist TF_{td} = 4/200 = 0,02.

(3) $TF_{tk} = FREQ_{tk} / GESAMT_{tk}$

Es wird in der gesamten Dokumentmenge die Termfrequenz ermittelt

Abkürzung: $GESAMT_{tk}$ = Gesamtzahl der Terme im Dokument
$FREQ_{tk}$ = Frequenz eines Terms in der Kollektion

Wenn der Term t in einer Menge von Dokumenten, die insgesamt 200.000 Wörter enthält, 600 mal auftritt, so ist TF_{tk} = 600 / 200.000 = 0,003.

(4) $S = TF_{td} - TF_{tk}$

Zum Schluss wird die Signifikanz errechnet. Es ergibt sich für das obige Beispiel eine Signifikanz von S = 0,02 - 0,003 = 0,017.

Die inverse Dokumenthäufigkeit wird verwendet, wenn TF_{td} und TF_{tk} in einer direkten Beziehung zueinander stehen. Die Frequenz eines Term t wird in einem

Dokument berechnet und dann in Beziehung gesetzt zur Anzahl aller Dokumente in denen t auftritt. Klassische Logarithmen werden verwendet. Die Formel lautet:

(5) $IDF(t) = FREQ_{td} / DOKFREQ_t$
(6) $IDF(t) = (\log_2 N / n) + 1$

Abkürzung: IDF(t) = inverse Dokument Frequenz des Wortes t
N = Gesamtzahl der Dokumente in der Datenbank
N = Anzahl der Dokument, in denen t vorkommt

Term t tritt in einer Menge von 2000 Dokumenten in 40 Dokumenten auf, und im betrachteten Dokument 4 mal, so ergibt sich die IDF = 4 / 40 = 0,1.

„Gute Indexterme“ weisen eine hohe Frequenz bei gleichzeitig niedriger Dokumentfrequenz auf. Sie finden sich relativ häufig in einzelnen Dokumenten und zusätzlich in wenigen Dokumenten der Kollektion. Wenn der Wert der IDF höher wird, wird die Entscheidungsstärke eines Indexterms höher. Die Entscheidungsstärke gibt Auskunft über die Möglichkeit, mittels eines Indexterms im Retrievalprozess relevante Dokumente aus einer Menge von Dokumenten zu selektieren und zusätzlich natürlich dazu, um irrelevante Dokumente auszuschließen.

Die Entscheidungsstärke eines Terms liegt in folgendem Diagramm:

- B im mittleren Frequenzbereich
- A im hohen Frequenzbereich
- C im niedrigen Frequenzbereich

D und E sind Schwellwerte.

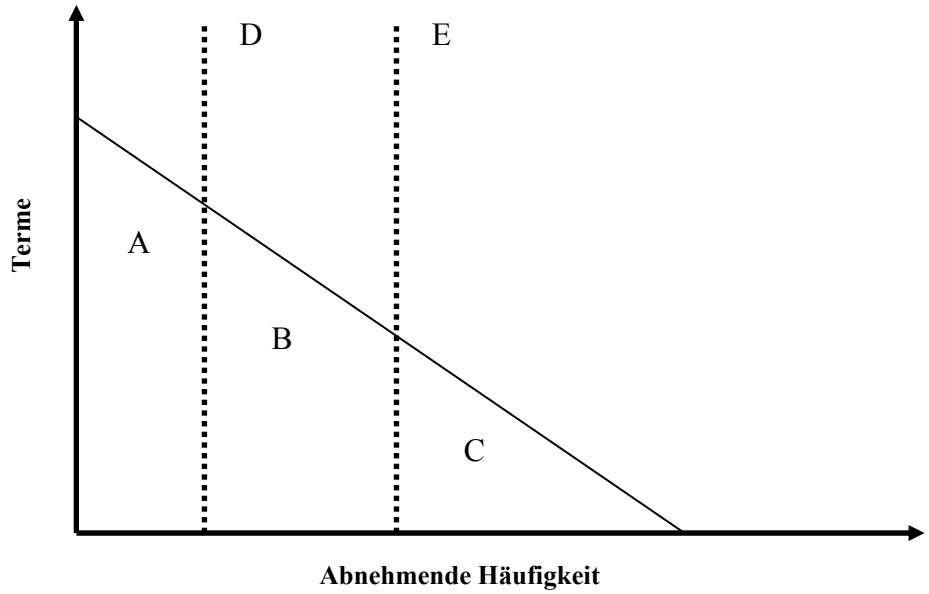

Abbildung 7: Termhäufigkeitsverteilung
Quelle:vgl. [Nohr05], S. 218, eigene Darstellung

C: Niedrigfrequente Terme sind nicht signifikant.
A: Hochfrequente Terme wie Artikel, Pronomen, Adverba sind relevant für die Syntax eines Textes.

Die Feststellung und Untersuchung geeigneter Schwellenwerte ist eine der wichtigsten Aufgabe bei der Anwendung der statistischen Verfahren[83].

Bei Auswahl guter Indexterme wird ein gutes Ergebnis zu erwarten sein durch eine hohe Frequenz bei niedriger Dokumentfrequenz, wobei berücksichtigt werden muss, dass zu hohe und zu niedrige Werte eliminiert werden sollten[84].

Eine ablauflogische Methode für die Termgewichtung ergibt folgenden Algorithmus[85]:

- *Eine Stoppwortliste erstellen. Restliche Substantive werden extrahiert und auf Nominativ, Singular, Grundform reduziert.*
- *Die Term- und Dokumentfrequenz wird errechnet*
- *Die IDF wird ermittelt*

[83] vgl. [Nohr05], S. 217,218
[84] vgl. [Nohr03], S. 33-37
[85] [Hali05], S. 9

- *Ein unterer Schwellenwert für die Auswahl guter Indexterme wird festgelegt und somit wird ein invertierter Index erstellt. Hierbei werden die Indexterme den Texten gegenübergestellt, und das Auftreten in den entsprechenden Texten wird festgehalten. Dieses Ergebnis kann bereits als automatische Indexierung benutzt werden.*

Beispiel für ein statistisches Verfahren[86]:

Text 1: Computer werden im Information Retrieval eingesetzt. Es existieren Verfahren auf Computern für ein automatisches Retrieval. Moderne Computer ermöglichen ein effizientes Retrieval nach spezifischer Information.
Text 2: Nutzer von Systemen zum Information Retrieval wurden befragt. Viele Nutzer waren mit der Funktionalität des Retrieval zufrieden. Die vorhandenen Systeme zum Information Retrieval genügen den Anforderungen der Nutzer. Es existiert eine Reihe von Softwaresystemen.
Text 3: Die Entwicklung neuer Systeme für das Information Retrieval wird von vielen Nutzern begrüßt. Die Entwicklung zielt auf neue Methoden des Retrievals mit Computern ab. Systeme zum effizienten Retrieval nach Information befinden sich derzeit in der Entwicklung.

Schritt 1: Mit Hilfe eine Stoppwortliste werden die Substantive extrahiert und die extrahierten Wörter werden bearbeitet und auf Nominativ, Singular reduziert.

Schritt 2: In dieser Bearbeitungsphase wird die Termfrequenz (FREQ) und die Dokumentfrequenz (DOKFREQ) errechnet. Die Addition der Frequenz eines Termes in einem Dokument wird in FREQn berechnet. Die Summe der Anzahl der Dokumente aus der gesamten Kollektion wird in DOKFREQ gerechnet.

Indexterm	FREQ1	FREQ2	FREQ3	DOKFREQ
Computer	3	1	1	3
Information	2	2	2	3
Retrieval	3	3	3	3
Nutzer	-	3	1	2
System	-	3	2	2
Entwicklung	-	-	3	1

Tabelle 6: Term- und Dokumenthäufigkeit
Quelle: [Nohr03], S. 42

86 vgl. [Nohr03], S. 40-44

Schritt 3: Nachdem die Termfrequenz und Dokumentfrequenz berechnet wurde, wird der Termgewichtung für jeden Indexterm berechnet.

$$IDF(t) = FREQ_{dt}/DOKFREQ_t$$

Indexterm	Text1	Text2	Text3
Computer	1	0,34	0,34
Information	0,23	0,23	0,23
Retrieval	1	1	1
Nutzer	0	1,5	0,5
System	0	1,5	1
Entwicklung	0	0	3

Tabelle 7: Termgewichtung
Quelle: [Nohr03], S. 43

Schritt 4: Zur Selektion guter Indexterme kann ein unterer Schwellenwert behilflich sein, der ein Steuerungsinstrument für die Auswahl von geeigneten Termen darstellt und es kann der niedrigfrequenteste Term bei der weiteren Bearbeitung außer Acht gelassen werden**.**

Zum unten folgenden invertierten Index führt ein unterer Schwellenwert von 0,5, sodass ein Schwellenwert IDF > 0,5 und für das Auftreten in Texten ein Schwellwert von 0,5 fixiert wird. Ein oberer Schwellenwert ist nicht notwendigerweise zu definieren, weil hochfrequente Wörter über eine Stoppwortliste bereits im ersten Schritt ausgeschlossen worden sind. Diese Ergebnisse können als automatische Indexierung als erstes Resultat verwendet werden. Der erzeugte invertierte Index kann als Zuganginstrument zu den Dokumenten dienen.

Indexterm	Texte		
Computer	Text1		
Retrieval	Text1	Text2	Text3
Nutzer		Text2	
System		Text2	Text3
Entwicklung			Text3

Tabelle 8: Invertierter Index
Quelle: [Nohr03], S. 43

6.3.4 Informationslinguistische Verfahren

Bei informationslinguistischen Verfahren geht es um bestimmte Aufgaben und Problembereiche, die sich mit regelbasierten Verfahren bzw. wörterbuchbasierten Verfahren beschäftigen. Bei diesen Verfahren werden die Terme eines Textes vor ihrer Indexierung bearbeitet, damit bei der Indexierung selbst qualitativ bessere Ergebnisse erzielt werden können. Diese Verfahren kommen bei der Vorbereitung für statistische Indexierung und bei der Vorbereitung der Produktion automatischer Indexe zum Einsatz.

Informationslinguistische Verfahren sind grundsätzlich Extraktionsverfahren. Es werden sprachliche Problemstellungen der Textanalyse untersucht. Diese Verfahren stellen eine Schnittstelle zwischen Computerlinguistik und Informationswissenschaft dar. Zusatzlich zu den lingusitischen Verfahren wird bei den statistischen Verfahren eine Zusammenarbeit und eine Verarbeitung der sprachlichen Einheiten wie Textterme angewendet. Teile der informationslinguistischen Verfahren sollten unbedingt bei statistischen Ansätzen berücksichtigt werden, damit bessere Ergebnisse bei der Termfrequenzermittlung gewährleistet werden können[87].

Die Anwendung der linguistischen Verfahren hat folgende grundlegende Ziele[88]:

- Nicht sinntragende Wörter zu eliminieren, damit diese aus der Indexierung gestrichen werden können (Pronomen, Präpositionen, …)
- Grammatikalische Flexionsformen in eine Stammform oder Grundform zu bringen wie „Autos → Auto“, „Häuser → Haus“
- Komposita sinnvoll zu zerteilen wie „Glücksautomaten → Glück ‚S’Automaten“; Autoversicherung → Auto und Versicherung“; „Wasserhahndichtung“ z.b. wird zerlegt in „Wasserhahn“ und „Dichtung“. Es wird nicht zerteilt in „Wasser“ und „Hahn“ und “Dichtung“. Es erfolgt keine morphologische Zerlegung, sondern eine semantische Zerlegung.

87 vgl. [Nohr03], S. 47-48
88 vgl. [Nohr03], S. 49

- Mehrwortbegriffe zu erkennen wie z.B. „elektronischer Marktplatz“, um Mehrwortlexeme als zusammengehörige Elemente identifizieren zu können.
- Pronomina korrekt einzusortieren und zuzuordnen (d.h. ein Wort, das für ein Nomen steht wie er, dieser, …)

Bei den statistischen Verfahren werden Wörter als eigenständiges Wort erkannt und gewertet und die linguistischen Eigenschaften bleiben unberücksichtigt. Für eine korrekte Ermittlung der Wörthäufigkeit müssen alle Wörter auf ihre Grundformen reduziert werden. Bei informationslinguistischen Verfahren wird die Zusammengehörigkeit von Wörtern ermittelt und bei der Indexierung mitberücksichtigt. Die linguistischen Verfahren beschäftigen sich mit der Lemmatisierung (Stemming) von Worten, wobei die Erkennung von Mehrworttermen und die Erkennung von Beziehungen zwischen Termen mit Hilfe syntaktischer, semantischer und morphologischer Analysen durchgeführt wird. Diese drei Ebenen der Textanalyse sind unabhängig von der verwendeten sprachlichen Ausdrucksform. Die **morphologische Analyse** (alle Wortformen werden einem Wortstamm zugeordnet) arbeitet auf der Wortebene und bearbeitet die vier oben genannten Aufgabenbereiche. Sie beschäftigt sich mit der inneren Struktur von Wörtern, der Bildung von Wortklassen und der Erkennung struktureller Gesichtspunkte. Es werden dabei verschiedene Verfahren zur Bildung von Stammformen und zur Erkennung der inneren Struktur von Worten verwendet. Die morphologische Analyse beschäftigt sich mit den Bereichen Flexion, Derivation und Komposition.[89]

Die Flexionsmorphologie befasst sich mit den Änderungen von Wörtern im grammatikalischen Kontext innerhalb der Satzstruktur wie z.b. geht—ging , kann—könnte. Die Derivationsmorphologie beschäftigt sich mit Wortzusammensetzungen, bei denen sich durch Hinzufügung eines Morphems ein Wort mit neuer Bedeutung entsteht. Bei der Kompositionsmorphologie werden Komposita als Zusammensetzung von mehreren Wörtern analysiert und untersucht[90].

[89] vgl. [Nohr03], S. 53
[90] vgl. [Uszk01], o.S.

Die **syntaktische Analyse** (Analysieren der Satzstruktur) führt ihre Arbeit auf der Satzebene durch. Es wird die grammatikalische Struktur einzelner Sätze untersucht, wobei die Zusammengehörigkeit von Wörtern in Wortgruppen sowie Haupt- und Nebensätzen analysiert wird. Weiters wird im Rahmen der syntaktischen Analyse versucht, mittels semantischer Analyse der (korrekt) reduzierten Grundformen Eindeutigkeit für mehrdeutige Wörter zu erreichen (Homographe und Homonyme). Das Ziel der syntaktischen Analyse ist es, auf Basis der Syntax eine möglichst korrekte Grundformreduktion zu erzielen. Es gibt Systeme, die dabei immer Strukturen gesamter Sätze verarbeiten (z.B. CONDOR, COPSY) und Systeme, die nur partielle Analysen durchführen (z.B. CTX). Syntaktische Analyse alleine liefert im Regelfall keine brauchbaren Ergebnisse und wird erst in Kombination mit der semantischen Analyse sinnvoll einsetzbar. Die **semantische Analyse** (Bedeutung von Sätzen und Satzteilen wird erschlossen) arbeitet auf Dokumentebene und versucht, die Bedeutung eines Teiltextes bzw. Satzes im Kontext des Dokumentes zu erfassen. Mit Hilfe von semantischen Analyseverfahren gelingt es teilweise, bis zu diesem Zeitpunkt bestehende Mehrdeutigkeiten zu eliminieren und Probleme wie Begriffsauswahl, Synonyme und Homonymie (gleich klingend oder geschrieben, aber mit unterschiedlicher Bedeutung in Abhängigkeit vom Kontext) in den Griff zu bekommen. Die Semantische Analyse von Freitexten befindet sich noch im Forschungsstadium.[91]
Informationslinguistische Verfahren können in regelbasierte Verfahren und wörterbuchbasierte Verfahren unterteilt werden:

6.3.5 Regelbasierte Verfahren

Regelbasierte Verfahren versuchen die Regelstruktur einer Sprache für die Indexierung in Form von Algorithmen zu verarbeiten und anhand von Zwischenergebnissen zu Resultaten zu gelangen. Im Vergleich zu wörterbuchbasierten Verfahren sind regelbasierte Verfahren mit geringerem Aufwand verbunden. Manche Probleme sind mit regelbasierten Verfahren nicht bzw. nur schlecht lösbar, wie man am Beispiel der Zerlegung von Komposita erkennen kann, wenn z.B. das

[91] vgl. [Nohr03], S. 51-53

Wort „Glücksspielautomat“ in seine Teile zerlegt werden soll. Eine Abhilfe bieten syntaxanalytische Verfahren.

Die Bearbeitung der unregelmäßigen Pluralbildungen wie („Haus“ - „Häuser“), die eine Stammformveränderung enthalten, ist ebenfalls ein Problem bei den regelbasierten Verfahren[92]. Regelbasierte Verfahren haben den Vorteil, dass nach einmaliger Definition der Regeln eine Anwendung auf alle Texte der jeweiligen Sprache erfolgen kann., wobei dabei vorsichtig vorgegangen werden muß, weil es sonst zu Unter- bzw. Übergeneralisierung kommen kann, weil sich dies aus dem Charakter des Modells der Wortformenreduktion ergibt.

6.3.6 Wörterbuchbasierte Verfahren

Wörterbuchbasierte Verfahren bauen auf der Systematik auf, jeden zu analysierende Term mit allen Möglichkeiten und möglichen Lösungen der Behandlung in einem Wörterbuch zu speichern. Es wird versucht, die auftretenden Begriffe dann auch in einem Wörterbuch wiederzufinden z.B. wird die mögliche Bearbeitung von Wörtern wie Wortformenreduktion, Dekomposition, Komposition und Derivation in einem Wörterbuch gespeichert. Es wird auch festgehalten dass z.B. „Glücksspielautomat“ in die Teile „Glück“, “S“, „Spiel“ und „Automaten“ zu zerteilen ist[93]. Der Aufwand bei den wörterbuchbasierenden Verfahren ist verhältnismäßig hoch, bringt aber im Vergleich zu regelbasierten Verfahren zuverlässigere Ergebnisse und diese Verfahren können oft auch unregelmässige sprachliche Phänomene zu einer brauchbaren Lösung verarbeiten[94]. Ein vergleichbares Wörterbuch müsste nicht nur Flexionsformen sondern auch Derivationsformen auf ihre Grundform Rückführen können[95]:

- Flexionsform → Grundform: lief → laufen, Häuser → Haus
- Derivationsform → Grundform: Lieblosigkeit → lieblos, Berechnung → rechnen

92 vgl. [Nohr03], S. 55
93 vgl. [Kump06], S. 51
94 vgl. [Strao.J.], S. 18
95 vgl. [Henr07], S. 109

Wörterbuchbasierte Verfahren bzw. lexikonbasierte Grundformreduktionen erzeugen also eine linguistisch korrekte Grundformreduktion mit Hilfe eines oder mehrerer elektronischer Wörterbücher. Regelbasierte Vorgänge sind für eine Sprache wie Englisch zweckmässig, sind aber für eine Sprache wie Deutsch, die stark konjugiert und dekliniert, häufig nicht zweckmässig, um korrekte Grundformreduktionen automatisch durchführen zu können.

Lexikalisch semantische Wortnetze wie WordNEt, GermaNet und EuroNET sind mit ihren lexikalisch semantischen Informationen für linguistische Verfahren benutzbar[96].

6.3.7 Lexikonbasierte Morphologieanalyse

In der deutschen Sprache ist es nicht wie in der englischen Sprache so, dass die Wörter mit Hilfe einzelner Regeln auf Zeichenketten (bzw. Mengen von Wörtern) abgebildet werden, um ihre Grundformen zu ändern. Es wird z.B. nicht nur am Wortende ein Suffix angehängt wie bei „das Haus – des Hauses“ sondern es kann z.B. zur Einfügung von Umlauten kommen, wie bei „alt - älter, Apfel - Äpfel“ und es wird manchmal ß und ss in Wörtern ausgetauscht. Bei Präfixen muß berücksichtigt werden ob das Präfix getrennt werden kann bzw. muß oder nicht, wie z.B. beim Wort „mitbringen“, wo im Imperfekt statt „mitbringen“ (Präfix „mit“) das Präfix abgetrennt wird („er brachte den Brief mit“), wohingegen beim Wort „überbringen“ im Imperfekt keine Abtrennung des Präfix erfolgt („sie überbrachte den Brief“). Die Trennung des Präfix kann die Bedeutung eines Wortes auch stark verfälschen. Weiters kommen in der deutschen Sprache häufig Komposita vor, wobei Worte zu einem neuen Wort zusammengefügt werden. Daneben werden unterschiedliche „Verfugungen“ benutzt die so gut wie keinen Regelmässigkeiten folgen (Schwein-s-stelze, Rind-s-braten). Derartige Probleme lassen sich bzgl. der deutschen Sprache nicht in allgemein gültige Regeln fassen, was eine Problemlösung mit rein regelbasierten Verfahren so gut wie unmöglich macht. Erst durch Verwendung eines Lexikons wird mit Hilfe der morphologischen Analyse die Ermittlung von Wortstämmen möglich. Die Pflege eines derartigen Lexi-

96 vgl. [Kim07], S. 15

kons muß manuell erfolgen. Lexikonbasierte Morphologieanalyse verwendet ein Vollformenlexikon oder ein Wortstammformenlexikon. Beim Vollformlexikon weist jede Wortform einem Wortstamm im Lexikon auf. Beim Stammformlexikon werden sowohl der Wortstamm als auch die verschiedenen Flexionsklassen gespeichert (wie e, s/e in Nominativ Plural der Berg**e** oder Genitiv Singular in Berg**(e)s**). An der Universität Erlangen wurde eine interessante Software namens "Deutsche Malaga Morphologie" (DMM) entwickelt. Diese Software ist in der Lage, eine automatische Erkennung von Wortformen durchzuführen und weiters erfolgt eine Kategorisierung, eine Lemmatisierung und eine Segmentierung[97]. Ein weiteres Programm, das auf dem Prinzip der lexikonbasierten Morphologieanalyse basiert, ist die Software „Morphy" von Wolfgang Lezius[98]. Die Reduktion auf die jeweilige lexikalische Grundform wird durch „Lemmatisierung durch Generierung" erreicht. Es werden dabei folgende Schritte durchgeführt:

- Es wird nach der Wortform in einer kleinen Liste gesucht, die häufig vorkommende Worte mit ihrer Grundform enthält. Wird das Wort gefunden, dann wird die Lemmatisierung beendet.
- Flexionsanalyse*:* Schrittweise wird nach Abtrennung der letzten Buchstaben eine Suche im Stammformlexikon durchgeführt. Wird das „Restwort", also das Wort bei dem der oder die letzten Buchstaben abgetrennt wurden, im Stammformlexikon gefunden, dann wird überprüft, ob der Wortstamm generiert werden kann, indem die zulässigen Umwandlungen überprüft werden. Fallweise werden mehrere mögliche Wortstämme gefunden, die dann lexikographisch überprüft werden müssen. Wenn Grundformen gefunden werden, werden die Gruundformen mit der Wortklassenzugehörigkeit ausgegeben und die Lemmatiserung wird beendet.

97 vgl. [Lore96], o.S.
98 vgl. [Lezi06], o.S.

Fall/Endung	-	n	en	Sen	…
Normal	Flüssen	Flüsse-n	Flüss-en	Flüs-sen	…
Umlaut	Flussen	Flusse-n	**Fluss**-en	Flus-sen	…
ß/ss	Flüßen	Flüsse-n	Flüß-en	Flü-ßen	…
beides	Flußen	Fluße-n	**Fluß**-en	Flu-ßen	…

Tabelle 9: Flexionsanalyse (nach Lezius,1995)
Quelle: [Ferb03], o.S.

- Kompositionsanalyse: Die längsten Wortformen werden von rechts zurückgehend abgeschnitten und im Lexikon gesucht. Dies führt zu einer Zergliederung in Teilworte. Die Lemmatisierung wird erst nach der Zerteilung in Teilworte beendet.
- Falls das Wort nicht lemmatisiert werden kann, dann wird für das Wort auf Basis einer empirischen Häufigkeitstabelle über die Wortklassenzugehörigkeit eine Vermutung angestellt.

Lexikonbasierte Morphologieanalyseprogramme sind deutlich aufwändiger und teurer als andere Programme zur Grundformenreduktion. Diese Programme produzieren jedoch deutlich mehr und genauere Informationen über die gesuchten Wortformen. Besonders bei Verfahren, die isolierte Wörter und Wörter im syntaktischen Zusammenhang auswerten, sind derartige Informationen hilfreich und notwendig[99].

6.3.8 Graphemisch-phonologische Verfahren

Graphemisch-Phonologische Verfahren berücksichtigen Fehlschreibungen von Wörtern, wenden also Fehlertoleranz praktisch an. Sie werden vor der Indexierung angewendet, um falsch geschriebene Worte zu korrigieren, die andernfalls bei der Indexierung zu falschen Deskriptoren führen würden. Die Rechtschreibprüfung ist ein wichtiger Teil von Textverarbeitungsprogrammen, wobei in deutschsprachigen Texten häufig Abweichungen gegenüber der Normschreibweise des Duden vorkommen. Wenn Benutzer die „richtige“ Schreibweise verwen-

[99] vgl. [Ferb03], o.S.

den, führt das möglicherweise zu Informationsverlusten. Folgende Unterschiede müssen korrigiert werden:

1. Am Satzanfang groß- und kleingeschriebene Wörter
2. ss und ß ,
3. Umlaute, Ablaute abgelegt als Vokal + „e“: Köln, Koeln

SOUNDEX- und N-Gram-Verfahren werden für die Erkennung und Zusammenführung von Schreibvarianten und auch für die Korrektur von Schreibfehlern verwendet[100]. Der Soundex-Algorithmus versucht Worte zu finden, die gleich ausgesprochen werden, aber unterschiedlich geschrieben werden. Dazu wird für jedes Wort mittels eines Algorithmus ein Soundex-Code ermittelt. Soundex-Codes für Worte, die ähnlich ausgesprochen werden, sind gleich bzw. annähernd gleich. Damit lassen sich Worte finden, die von der Phonetik her sehr ähnlich sind, wie das betrachtete Wort. Wird der Soundex-Code eines betrachteten Wortes nicht gefunden, dann wird als Vorschlag das Wort mit dem nächstähnlichen Soundex-Code verwendet. Der Algorithmus wurde primär als Fehlerbehandlungsverfahren entwickelt, um bei Suchanfragen Fehlschreibungen abzufangen und kann auch zur fehlertoleranten Behandlung natürlichsprachlicher Dokumente verwendet werden. Zur Indexierung von Dokumenten ist der Soundex-Algorithmus so gut wie nicht brauchbar[101].

6.3.9 Art der Reduktionform

Für die englische Sprache wurden erfolgreich verschiedene regelbasierte Verfahren entwickelt, die aber leider – wie z.B. der Porter Stemmer Algorithmus – für deutsche Texte nur sehr eingeschränkt verwendbar sind. Der Grund für die schlechte Verwendbarkeit dieser Algorithmen liegt in der geringen Komplexität und den seltenen Wortstammänderungen der englischen Sprache gegenüber der deutschen Sprache.

100 vgl. [Nohr03], S. 50-51
101 vgl. [Kump06], S. 52

KUHLEN (1974) folgend, sind folgende Reduktionen durchzuführen[102]:

- auf die formale Grundform
- auf die Stammform
- auf die lexikalische Grundform

Es werden drei Reduktionsformen differenziert[103]:

Grund- und Stammformenreduktion: Das Verfahren der Grundformreduktion führt verschiedene Wortformen auf ihre jeweilige formale Grundform zurück. Dies geschieht dadurch, dass die Flexionsendung abgetrennt wird. Zusätzlich kann die lexikalische Grundform dadurch erreicht werden, dass man bei Substantiven den Nominativ Singular und bei Verben den Infinitiv bildet. Wird anschließend die Derivationsendung entfernt, wird das Wort auf seinen Wortstamm zurückgeführt, und somit die Stammform gebildet. Die Reduktion auf die Stammform besteht also primär aus der Entfernung der Derivationsendungen z.B. ist der Wortstamm der englischen Wörter computer, compute, computation und computerization das Wort comput[104].

Dass verschiedene Reduktionsalgorithmen unterschiedlich bis auf die Formel oder die lexikalische Grundform vereinfachen oder sogar Reduktionen bis auf die Stammform vornehmen, wird hierbei besonders deutlich. Die Zurückführung auf Grundform oder Stammform wird als Lemmatisierung oder Stemming bezeichnet. Es hängt davon ab ob man mit der deutschen oder englischen Sprache zu tun hat. Bei automatischen Verfahren wird durch die Reduktion zur Grundform die Anzahl der Wörter stark reduziert. Es wird nur ein Wort einer Wortfamilie gespeichert und es wird eine Generalisierung durchgeführt, bei der die einzelnen Wörter einer Wortfamilie auf eine grammatikalische Form zurückgeführt werden. Es gibt einige erfolgreiche Verfahren die aber nur für die englische Sprache optimal funktionieren: Porter Stemmer Algorithmus oder lexikografische Grundformenreduktion nach Kuhlen. Durch einen Vergleich zwischen der deutschen und englischen Sprache erkennt man, dass in der englischen Sprache die Wörter weniger Wortformen haben und weniger zusammengesetzte Wörter existieren als in der

102 vgl. [Nohr03], S. 57
103 vgl. [Kump06], S. 50-51
104 vgl. [Nohr03], S. 57

deutschen Sprache. Deshalb haben die lexikonbasierten Verfahren bei Anwendungen in der deutschen Sprache große Bedeutung. Von Lemmatisierung spricht man wenn eine Wortform auf ein Lemma zurückgeführt wird[105].

Formale Grundform	Textwörter	Lexikalische Grundform	Stammform
Absorb ….. ….. ….. ….. …..	Absorb Absorbed Absorbing Absorbs	Absorb ….. ….. …..	Absorb ….. ….. ….. ….. …..
	Absorber absorbers	Absorber …..	
absorbab	Absorbable absorbably	Absorbable …..	….. …..
Absorbanc ….. ….. …..	Absorbance Absorbances	Absorbance …..	….. ….. ….. …..
	Absorbancy absorbancies	Absorbancy …..	
Absorbant ….. …..	Absorbant Absorbants absorbantly	Absorbant ….. …..	….. ….. …..
Absorbtion ….. absorbtiv …..	Absorbtion Absorbtion Absorbtively absorbtive	Absorbtion ….. absorbtive …..	….. ….. ….. …..

Tabelle 10: Reduktionsalgorithmenund ihre Wirkungensweisen nach KUHLEN
Quelle: [Nohr03], S. 58

- **Bindestrichergänzung:** Das Weglassen gleicher Worteile unter Zuhilfenahme eines Platzhalters für die weggelassenen gleichen Wortteile ist eine Besonderheit der deutschen Sprache. Durch Weglassen des Bindestriches und Ersetzen des Bindestriches durch den jeweils gleichen, mehrfach vorkommenden Wortteil kann man eine Vervollständigung durchführen. Aus den Worten „Film- und Fernsehwirtschaft“ wird durch Bindestrichergänzung „Filmwirtschaft und Fernsehwirtschaft“[106].

105 vgl. [Kump06], S. 50-51
106 vgl. [Kump06], S. 51

- **Phrasenerkennung / Mehrwortgruppen**: Die Mehrwortgruppenerkennung (Phrasenerkennung) ist eine wichtige Aufgabe für lingusitische Verfahren. Um die Mehrwortgruppen zu erkennen wird ein Text in Text-„klumpen“ zerlegen. Mit Hilfe einer Stoppwortliste (Hilfsverben, Adverba und andere Verben) und Satzzeichen erfolgt eine Zerteilung von Sätzen, wobei die Stoppwortliste und die Satzzeichen als Delimiter verwendet werden[107]. Indikatorbegriffe helfen bei der Erkennung von Phrasen. Eine Phrase ist ein Begriff, der aus mehreren einzelnen Wörtern besteht. Neben einzelnen Wörtern und ihren Wortstämmen können Phrasen als Einzelworte begriffen werden und somit auch als Schlagworte dienen. Begriffe, die in Phrasen häufig auftreten, nennt man Indikatorbegriffe. Bei Unternehmensnamen findet sich häufig die Rechtsform im Unternehmensnamen und das Wort „GmbH“ ist z.B. ein Indiz dafür, dass in der Wortfolge „der Geschäftserfolg der XYZ GmbH“ „XYZ GmbH“ eine Phrase ist. Bedeutsam ist es, dass man durch die Suche nach Indikatorbegriffen die Phrasenerkennung wesentlich verbessern kann. Namen von Organisationen finden sich häufig in Verbindung mit Indikatorbegriffen wie College, Center, Church, Comitte oder Universität usw..[108]

6.3.10 Begriffsorientierte Verfahren

Begriffsorientierte Verfahren sind eine Methode, die menschliche Indexierung zu simulieren, bei der versucht wird, die Bedeutung eines Textes herauszufinden. Es wird versucht die Wörter eines Textes zu extrahieren und sie anschließend normiert und gewichtet einer (gemeinsamen) Bedeutung zuzuordnen. Das Ziel derartiger Verfahren ist das Verstehen von Texten indem die Algorithmen versuchen, die Bedeutungen eines Textes, also Wortbedeutungen und Bedeutungen von Worttermen herauszufinden. Es existieren verschiedenste Ansätze mit teilweise großen Unterschieden in der Herangehensweise und in den verwendeten Algorithmen[109]. Auch derartige Verfahren sind nicht in der Lage, Texte wirklich zu verstehen. Die begriffsorientierten Vefahren versuchen häufig mit Hilfe einer kontrollierten Dokumentationssprache (z.B. Thesaurus) die Bedeutung von Dokumenten zu erfassen. Begriffsorientierte Verfahren werden im Forschungsbe-

107 vgl. [Nohr03], S. 60
108 vgl. [Kump06], S. 51
109 vgl. [Kump06], S. 52

reich der Künstlichen Intelligenz (AI) erforscht und bilden einen Teil des Forschungsgebietes der wissensbasierten Systeme[110].

6.3.11 Mustererkennungsverfahren

Mustererkennungsverfahren sind Methoden zur Auffindung von Informationen in großen Dokumentenbeständen ohne Indexierung von Dokumenten. Bei den Mustererkennungsverfahren wird prinzipiell durch den Vergleich von Zeichen oder Bildern mit Mustern in einer Wissensbasis versucht, übereinstimmende Lösungen anhand ihres (Wort-)Musters zu identifizieren. Gewisse Eigenschaften werden als Erkennungsparameter herangezogen: bündig am Wortanfang (Präfix), bündig am Wortende (Suffix), Teil eines Wortes.

Die Algorithmen der Mustererkennung werden in der Praxis z.B. vom Indexierungssystem FIPRAN implementiert. Mit Hilfe einer Kombination von Mustererkennungsverfahren und linguistischen Verfahren wird analysiert und die Ergebnisse werden unter Berücksichtigung des Bezuges zum Originaltext gespeichert. Es werden drei Schritte unterschieden:

Zuerst wird eine Schlüsselworterkennung durchgeführt, deren Ziel die Zuordnung zu einer Patternklasse ist. Elemente einer Patternklasse weisen einen Schlüssel und eine Liste prüfbarer Parameter auf, anhand derer dann eine weitere Analyse möglich wird. z.b. werden Länderschlüssel („den“) und Adjektive („dänisch“) und nominale Formen („Dänemark“) identifiziert. Bündige Wortteile am Wortanfang (Präfixe) und bündige Wortteile am Wortende (Postfixe) kommen als Worterkennungsparameter in Frage. Um ein Muster in einem Dokument zu erkennen müssen sowohl der Schlüssel als auch die hinterlegten prüfbaren Parameter übereinstimmen.

- Bündig am Wortanfang heißt, das Wort (in diesem Fall „Dän“) darf keine vorhergehenden Buchstaben haben. Das Muster für „Dän“ ist akzeptabel „Dänemark“ aber das Wort „Dänisch“ wird nicht akzeptiert.

110 vgl. [Nohr03], S. 79-81

- Bündig am Wortende: Am Wortende steht ein sinnvolles Wort, das akzeptiert wird.
- Über Wortgrenzen hinaus wird teilweise versucht, mehrere Worte als zusammengesetzte Einheit zu betrachten und auf diese Weise wird es z.B. möglich gemeinsame Vorkommen von Vor- und Nachnamen zu erkennen.

Parameter / Schlüssel	Wortanfang bündig	Wortende bündig	über Wortgrenzen hinweg
Patternklasse Länder (z.B.“dän“)	Ja	Nein	Nein
Patternklasse Produkt (z.B.“Flugzeug“)	Nein	Nein	Nein

Tabelle 11: Beispiel für Patternklassen aus FIPRAN
Quelle: [Hali05], S. 13

Textblöcke werden auf Regelbasis festgestellt. FIPRAN z.B. definiert die Blockgrenzen über ein Begrenzungsverfahren und basiert auf Regeln. z.B. vor und nach Hilfsverben oder vor und nach einem Punkt, oder bei einem Doppelpunkt. *Die Regeln, um Blockgrenzen innerhalb von Texten zu ermitteln sind: Bei Satzende, bei Semikolon oder Doppelpunkt, vor und nach Verben, vor und nach Hilfsverben, vor Konjunktionen, vor einer Präposition, vor Artikeln, wenn davor keine Präposition steht*[111].

Danach werden Textblöcke ausgewertet und es wird versucht, auf den Inhalt zu schließen. Die Auswertung der Blockstruktur wird in Form von Regeln durchgeführt und es bietet sich die Möglichkeit der automatischen Inhaltserschließung, wobei Kombinationen von neuronalen Netzen und Mustererkennungsverfahren eingesetzt werden. Die wichtigsten Algorithmen bei Mustererkennungsverfahren sind folgende Algorithmen: Brute-Force Algorithmen, Knuth-Morris-Pratt Algorithmen, Boyer-Moore Suche und die Suche mit endlichen Automaten.

111 [Strao.J.], S. 21

Obwohl die Mustererkennungsverfahren sehr wichtige Formen der automatischen Indexierung sind, werden sie sowohl in der Praxis als auch in der Forschung nur in geringem Umfang bei der Analyse von Freitexten eingesetzt. Bei der Analyse und Erschließung von Bildern, Filmen und anderen Multimediainhalten könnten die Mustererkennungsverfahren in Zukunft aufgrund des rasch zunehmenden Bedarfes stark an Bedeutung gewinnen[112].

6.4 Grundlegenden Verfahren der automatischen Schlagwortextraktion

Data-Mining befasst sich mit unstruktierten Datenbeständen in Datenbanken. Die Menge von Informationen, die aber in Form unstrukturierter Texte vorliegt, erhöht sich ständig. Nicht zuletzt sei hier das WWW explizit erwähnt. Text-Mining ist eine Umschreibung für alle Methoden und Verfahren, die den Versuch unternehmen, Informationen und Wissen aus unstrukturierten Texten zu erschließen. Beim Text-Mining geht es also um automatisierbare Bearbeitungstechniken zur Inhaltserschliessung von unstrukturierten Texten unter Zuhilfenahme verschiedenster Verfahren (statistische, linguistische und musterbasierte Methoden).

In der Folge wird ein Überblick über die verschiedenen Algorithmen gebracht, die bei der Extraktion von Schlagwörtern verwendet werden[113]:

6.4.1 Eliminieren der Stoppwörter

Wörter die häufig in Texten vorkommen, für den Retrievalzweck keinen Sinn haben und deren Bedeutung in Bezug auf den Dokumentinhalt nicht relevant ist nennt man Stoppwörter. Stoppwörter werden von der Verarbeitung eines Textes ausgenommen, damit das Ergebnis der Textuntersuchung nicht negativ beeinflusst wird. Je nach verwendetem Verfahren können Stoppwörter allerdings für die Trennung eines Textes als trennende Markierungen Verwendung finden.

112 vgl. [Nohr03], S. 75-79
113 vgl. [Kim07], S. 7

Stoppwörter der deutschen Sprache können neben Artikeln auch Konjunktionen, Präpositionen, Hilfsverben, usw. sein.

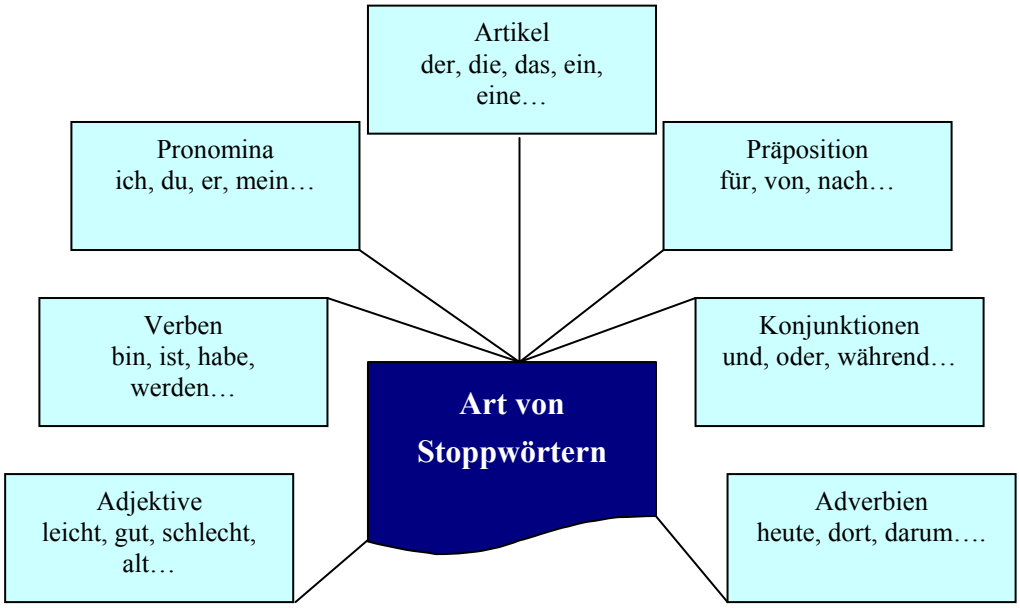

Abbildung 8: Arten von Stoppwörtern
Quelle: [Batt05], S. 32, eigene Darstellung

6.4.2 Stemming Algorithmus zur Grundformreduktion

Unter Stemming versteht man die Reduktion von Wörtern auf ihre Grundform und Stammform. Stemming ist eigentlich eine Form der Normalisierung. Durch Stemming wird der Term zu einer Art "Wortkern" reduziert, der nicht unbedingt Element der Sprache sein muß. Es ist möglich, dass ein Stemming-Algorithmus sich auf linguistisches Wissen über die sprachspezifischen Suffixe und Endungen stützt. Es wird versucht, die Flexionsendungen und Derivationssuffixe zu entfernen. Wenn in einem Dokument ein Wort in verschiedenen Formen vorkommt, so werden mit Hilfe von Stemming alle vorkommenden Wortformen auf ihre jeweilige gemeinsame Grundform reduziert. Die Methoden die bei Stemming verwendet werden, sind:

- Affixabtrennung (einfaches Entfernen von Prä- und Suffixen)
- tabellengestützte Abtrennungen (es wird ein Lookup in einer Tabelle durchgeführt, um abtrennbare Teile zu finden und zu einer Grundform zu gelangen)
- N-Gram-Verfahren
- Porter Stemmer Verfahren

Stemmingverfahren sind für manche Sprachen gut verwendbar und für manche Sprachen weniger gut geeignet. Für die englische Sprache ist Stemming verhältnismässig einfach implementierbar, wohingegen die komplexere Wortbildung der deutschen Sprache eine Verwendung von regelbasierten Verfahren einschränkt. Als Alternative für „problematische“ Sprachen wie z.B. Deutsch bieten sich lexikonbasierende Verfahren an. Bei lexikonbasierenden Verfahren müssen neu auftretende Wörter, die bisher noch nicht im verwendeten Lexikon enthalten sind, eingepflegt werden, was meist nur manuell erfolgen kann und somit wieder verhältnismässig aufwändig ist. Es folgt ein Überblick über den Ablauf der verschiedenen Stemming Algorithmen[114].

6.4.3 Porter Stemmer Algorithmus zur Grundformreduktion

Um englische Wörter zu lemmatisieren wird häufig der Porter-Stemmer Algorithmus oder der Lovins-Stemmer Algorithmus (Lovins1968) verwendet. Der Porter-Stemming Algorithmus befasst sich mit einer 100% richtigen linguistischen Grundformreduktion und mit einer effizienten Berechenbarkeit der Grundformreduktion. Im Gegensatz zum Lovins-Stemmer Algorithmus werden zwei Operationen wie Suffixstripping und recording beim Porter-Stemmer Algorithmus durchgeführt[115]. Zur Vorstellung des Algorithmus werden einige Definitionen benötigt [116]:

114 vgl. [Lewa05], S. 107-108
115 vgl. [Kim07], S. 14
116 vgl. [Hein99], S. 162-163

- Ein Konsonant in einem Wort ist ein Buchstabe ungleich A, E, I, O ,U und Y, dem ein Konsonant vorausgeht. Z.B. sind in TOY die Konsonanten T und Y enthalten. In SYZYGY sind S,Z,G enthalten.
- wenn ein Buchstabe kein konsonant ist, ist der Buchstabe ein **Vokal.**
- Eine Konsonant wird durch ein **C** bezeichnet und ein Vokal durch ein **V**.
- Falls eine folge „**ccc**“ mit mehr als 0 Konsonanten besteht, wird sie mit **C** bezeichnet.und eine Folge **vvv**… mit mehr als 0 Vokalen wird als **V** bezeichnet. Das heisst dass jedes Wort kann in vier Formen dargestellt werden:
  - ●CVCV…C
  - ●CVCV…V
  - ●VCVC…C
  - ●VCVC…V

Es kann $[C](VC)^m[V]$ verwendet werden (die eckigen Klammern stehen für optionale Teile)

- Ein Wort kann also grundlegend in der Form $[C](VC)_m[V]$ dargestellt werden.

$(VC)_m$: m-malige Wiederholung von VC, wobei für m=0 ein Nullwort resultiert.

Es gibt beim Porter-Stemming Algorithmus acht Bearbeitungsregeln[117]:
Die Regeln (a) bis (e) beziehen sich auf Substantive, die Regeln (f) bis (h) sind die Bearbeitungsregeln für Verbformen. z.B. das Wort „Making“ wird zuerst durch den zweiten Fall der Regel (f) bearbeitet, wodurch die die Zeichenfolge „ing“ durch das Zeichen „e“ ersetzt wird und damit wird die Grundform „make“ generiert.

Notation		Regel des Porter Stemmming.	
%	alle Vokale,einschl.Y	a)	IES→*Y*
*	alle Konsonanten	b)	ES→*§ nach * O/CH/SH/SS/ZZ/X*
!	Länge des wortes	c)	S→*§ nach */E%Y%O/OA/EA*
/	„oder“	d)	S→*§ , IES→Y, ES→§*
§	Leerzeichen	e)	´S→*§*, ´→*§*
→	„zu“	f)	ING→*§ nach **/%/X* *ING→E nach %**
←	„aus“		
\	„nicht“	g)	IED→*Y*
		h)	*ED→§ nach **/%/X* ED→*E nach %**

Tabelle 12: Regeln der Porter Stemmer Algorithmus
Quelle: [Nohr03], S. 59

117 vgl. [Nohr03], S. 58-59

Bei manche Regeln gibt es auch leere Beendigungen, wie folgende Beispiele zeigen[118]:

SSES→*SS*	*caresses→caress*
IES→I	*Ponies→poni , ties→ti*
S→§	*cats→cat*
EED→EE	*feed→feed , agreed→agree*
ED→§	*plastered→plaster, bled→bled*
ING→§	*motoring→motor , sing→sing*

Aufgrund die relativ unregelmäßigen Morphologie der deutschen Sprache und den damit einhergehenden Problemen beim Einsatz von Stemming muss bei deutschen Stemmern unter Umständen eine Anpassung durchgeführt werden, um brauchbare Ergebnisse zu erhalten (in Abbildung 15 wird eine deutsche Version des Porter Stemmer Algorithmus dargestellt). Der Vorbereitungsschritt hat die Aufgabe, den gesamten Text zu bereinigen und kümmert sich um eine Umwandlung von Umlauten und dem scharfen s. Es folgen die drei eigentlichen Schritte des Stemming und es werden nach bestimmten Regeln die Suffixe verarbeitet und abgetrennt. Jedes Wort wird in unterschiedliche Bereiche (Regionen) eingeteilt und diese Bereiche werden für die Verarbeitung R1 und R2 gennant. Die Reduktion eines Wortes wird immer abhängig von R1 und R2 durchgeführt. Der Sinn und die Schwierigkeit ist es, Suffixe zu reduzieren, aber Ausnahmen von den Regelfällen zu erkennen und zu berücksichtigen. z.B. wird im Schritt drei die Endung „isch“ reduziert, was beim Wort „italienisch“ zum korrekten Wortstamm „italien“ führt, wohingegen das Wort „Tisch“ zu „T“ reduziert würde, was eine fehlerhafte Verarbeitung darstellt. Stemming ist effektiver als lexikonbasierende Methoden und es entsteht kein zusätzlicher Aufwand bei der Erstellung bzw. Pflege von Wörterbüchern. Stemming ist ein regelbasierendes Verfahren, daher ist eine Erkennung von Ausnahmen nur schwer realisierbar. Die regelbasierten Verfahren haben – wie alle anderen Verfahren auch - eine bestimmte Fehlerquote, liefern aber in Abhängigkeit von der exakten Problemstellung meist recht brauchbare Ergebnisse, wie aus der Tabelle 13 ersichtlich.

[118] vgl. [Hein99], S. 166, (fast unverändert übernommen)

Porter Stemmer Algorithmus für die deutsche Sprache

Vorbereitung:

- ersetze u und y in Großbuchstaben, wenn sie zwischen Vokalen stehen
- e ersetze ß durch ss
- ersetze ae durch ä
- ersetze oe durch Ö
- ersetze ue durch ü, wenn kein q vorrausgeht

Schritt 1:

- Suche das längste der folgenden Suffixe
  e em en ern er es
  s, wenn b,d,f,g,h,k,I,m,n,r ofrt t vorausgeht und lösche, wenn in R1

Schritt 2:

- Suche das längste der folgenden Suffixe
  en er est
  st, wenn b,d,f,g,h,k,l,m,n oder t und mindestens drei Buchstaben vorausgehen
- und lösche , wenn in R1

Schritt 3:

- Suche das längste der folgenden Suffixe und führe die entsprechende Anweisung aus
  end ung → lösche, wenn in R2 und ig oder e vor
  ig ik isch → lösche, wenn in R2 und kein e vorausgeht
  lich heit → lösche, wenn in R2 und er oder en vorausgeht oder w. in R1
  keit → lösche, wenn in R2 und lich oder ig vorausgeht oder wenn in R1

Nachbereitung:

- ersetze U und Y wieder durch Kleinbuchstaben

Definition von R1 und R2

R1 ist die Region nach dem ersten Konsonanten, der auf einen Vokal folgt oder das Nullwort, wenn diese Region nicht existiert. (Beispiel: Wirtschaftsinformatik**: | < R1 > |),** R2 ist die Region nach dem ersten Konsonanten, der auf einen Vokal folgt in R1 (Beispiel: Schaufenster: **| < R1 > | < R2 > |)**

Tabelle 13: Porter Stemmer Algorithmus für die deutsche Sprache
Quelle: [Batt05], S. 34

6.4.4 Over Stemming und Under Stemming

Die Porter Stemmer Regeln, die oben erklärt sind, gelten für englischsprachige Texte, aber für andere Sprachen existieren angepasste Regeln und Praxiserfahrungen. Ein Nachteil dieses Verfahren ist die Fehleranfälligkeit und damit die Erzeugung falscher Wortstammreduktionen, weil diese Reduktionsalgorithmen nur die implementierten Regeln abarbeiten und keinerlei „Verstehen“ der Wörter vorliegt. Men unterteilt beim Stemming die Fehler in zwei Fehlerklassen: Overstemming und Understemming[119].

- Overstemming:
  Eine zu lange Zeichenkette wird gekürzt, unterschiedliche Wortformen mit gleichen Grund- bzw. Stammformen werden falsch vereinheitlicht (z.B. werden die Worte Kommunismus, Kommunikation und kommunizieren zum gemeinsamen Wortstamm „kommun“ reduziert)
- Understemming:
  Eine zu kurze Zeichenkette wird gekürzt, es wird also zu wenig abgeschntten und unterschiedliche Wortformen mit gleichen Grund- bzw. Stammformen werden nicht zur gleichen Stammform reduziert, we im Folgenden zu erkennen ist:

  Kommunismus → Kommun
  Kommunikation → Kommunika
  kommunizieren → kommuniz
  oder
  die Themen→ them
  des Themas→ thema

Weitere Bespiele: http://www.bui.haw-hamburg.de/pers/ursula.schulz/astep/le4_step_3.html

119 vgl. [Nohr03], S. 58-59

6.4.5 LOVINS Algorithmus zur Grundformreduktion

Der Lovins-Algorithmus zur Grundformreduktion, der besonders für Texte im Englischen gut geeignet ist, wird hier exemplarisch für eine Reihe von ähnlichen regelbasierten Algorithmen vorgestellt. Die Funktionsweise vom Lovins-Algorithmus ist zweistufig. Im ersten Schritt trennt er Endungen ab und im zweiten Schritt wird eine Umgestaltung der verbliebenen Endung des Wortes erledigt. Das Ergebnis des Lovins-Algorithmus ist nicht immer die bestmögliche Endung eines Wortes, aber wichtig ist, dass alle Wörter auf die gleiche Weise umgeformt werden.

Abbildung 19. zeigt die Art und Weise der Abtrennung und Sortierung der Endungen. Also, die Endungen sind absteigend nach ihrer Länge und innerhalb der gleichen Länge alphabetisch aufsteigend sortiert. Der Algorithmus beginnt seine Arbeit von der ersten Zeile der Tabelle an und vergleicht die Endungen mit der Endung eines zu reduzierenden Wortes. Neben der Konkordanz der Endung muss zudem die in Tabelle 14 ebenfalls für eine Endung angegebene Bedingung erfüllt sein.

Erläuterung der weiteren Teilschritte:
1. Endet ein Wort mit einem Konsonanten ungleich „s“, der von einem s gefolgt ist, dann wird „s“ gelöscht..
Beispiele:
• *stems wird zu stem*
• *Aber stress bleibt stress*

2. Endet ein Wort mit *es*, dann wird das abschließende *s* entfernt

Beispiele:
• *places wird zu place*
• *likes wird zu like*
• *theses wird zu these (»Fehler!«)*
• *indices wird zu indice (»Fehler!«)*
• syntheses wird zu synthese (»Fehler!«)

Länge	Endung	Bedingung
11	Alistically	B
11	Arizability	A
11	Izationally	B
10	Antialness	A
10	Arisations	A
10	Arizations	A
10	Entialness	A
9	Allically	C
9	Antaneous	A
4	Able	A
4	Ably	A
4	Ages	B
4	Ally	B
3	Ism	B
1	e	A

Tabelle 14: Bedingungen und Wortendungen nach Lovins
Quelle: [Henr07], S. 106

Bedingungen:
A: keine Begrenzungen
B: verbleibender Wortstamm min. 3 Zeichen lang
C: verbleibender Wortstamm min. 4 Zeichen lang

Die weitere Schritte sind:

3. Ändern *iev* zu *ief* und *metr* zu *meter*.

Beispiel:

• *believable* wird zu *believ* und nun zu *belief*

Schritt = Abtrennung der Endung)

4. Endet ein Wort mit *ing* dann wird *ing* gelöscht, es sei denn, das Wort besteht nach der Löschung nur aus einem Buchstaben oder aus *th*.

Beispiele:

• *thinking wird zu think*

• *singing wird zu sing*

• *sing wird zu sing (keine Änderung)*

• *thing wird zu thing (keine Änderung)*

• *preceding wird zu preced (Fehler; kein Wort!)*

Den Fehler im letzten Beispiel könnte man z.B. durch eine nachgeschaltete Regel vermeiden, die lautet: Endet ein Wort nach der Reduktion mit *et*, *ed* oder *es*, dann wird ein *e* hinzugefügt.

5. Endet ein Wort mit *ed* und ein Konsonant vorausgeht, dann wird *ed* gelöscht. Es sei denn, das Wort besteht nach der Löschung nur aus einem Buchstaben.
Beispiele:
• *ended wird zu end*
• *red wird zu red*
• *proceed wird zu proceed (es geht kein Konsonant voraus)*
• *proceeded wird zu proceed*

6. Endet nach dem Entfernen der Endung ein Wort mit *bb, dd, ..., tt*, dann wird einen der doppelten Buchstaben entfernt.
Beispiel:
• *embedded wird zu embedd und nun zu embed*

7. Endet ein Wort mit *ion,* dann wird *ion* entfernt. Falls das verbleibende Wort mehr als 2 Buchstaben hat und der letzte Buchstabe des Stammes ein Konsonant und der vorhergehende ein Vokal ist, dann wird zusätzlich ein *e* zugefügt.
Beispiele:
• *direction wird zu direct*
• *polution wird zu polute*
• *plantation wird zu plantate (Fehler!?)*
• *zion wird zu zion*
• *scion wird zu scion*
• *anion wird zu anion*
• *cation wird zu cate (Fehler!)*

Ein leistungsfähiges System zur Grundformreduktion für die englische Sprache braucht zwischen 10 bis 20 Regeln. Um die Qualität des Systems zu verbessern sind zahlreiche Ausnahmen z.B. für irreguläre Verben erforderlich. Das System sollte die iterative Anwendung der Regeln auch unterstützen. So wird z.B. zuerst

aus *directions direction* und durch wiederholte Anwendung der Regel die Grundform *direct* erstellt[120].

6.4.6 Wortstammbildung nach dem N-Gram Verfahren

Das N-Gram Verfahren gehört zu den Stemmingverfahren. Es wird dabei versucht mittels Zerteilung von Worten in Kombination mit statistischen Verfahren eine Grundformreduktion zu erreichen. Beim N-Gram Verfahren werden Worte (auch Komposita) in Wortteile der Länge N zerteilt. Diese Zerteilung funktioniert schnell und effizient. (die Verarbeitungsgeschwindigkeit ist von der Länge der N-Gramme abhängig und der Wert von N darf nicht beliebig gewählt werden). Aufgrund der verwendeten statistischen Verfahren liefert das N-Gram Verfahren ebenfalls Ergebnisse mit einer gewissen Ungenauigkeit. Die N-Gram-Häufigkeiten erlauben keine genaue Aussage über die Stelle, an der die gesuchten N-Grame auftreten. Bei diesem Verfahren werden auch die N-Grame untersucht, die auf einen Begriff verweisen, der im Ausgangswort nur durch seine Reihenfolge von Buchstaben, nicht jedoch von der Bedeutung her enthalten ist [121]. Das Verfahren modelliert Paare von Wörtern mit einer gemeinsamen Anzahl von n-Grammen. Für n=2 (Bigramme) wird als Ergebnis im folgenden Beispiel für die Wörter „Statistics“ und „Statistical“ folgende Abarbeitung resultieren und es ergeben sich 6 gemeinsame Bigramme[122]:

statistics →st ta at ti is st ti ic cs
s1 = {at, cs, ic, is, st, ta, ti}
statistical →st ta at ti is st ti ic ca al
s2 = {al, at, ca, ic, is, st, ta, ti}
S1 ∩ S2 = {at, ic, is, st, ta, ti }

Die Gleichheit S zweier Wörter wird mit folgender Formel berechnet:

$$S = (2C)/(A+B)$$

A: Zahl der einzelnen Diagramme aus dem ersten Wort;
B: Zahl der einzelnen Diagramme aus dem zweiten Wort;
C: Gesamtzahl der Diagramm.
S: Ähnlichkeitsmaß

120 vgl. [Henr07], S. 105-108
121 vgl. [Lewa05], S. 107-108
122 vgl. [Nits04], S. 28

	Word1	Word2	Word3	…	$Word_{n-1}$
$Word_1$		S_{12}	S_{13}	$S_{1}…$	$S_{1(n-1)}$
$Word_2$	S_{21}		S_{23}	$S_{2}…$	$S_{2(n-1)}$
$Word_3$	S_{31}	S_{32}		$S_{3}…$	$S_{3(n-1)}$
….	$S_{…1}$	$S_{….2}$	$S_{…3}$		$S_{..(n-1)}$
$word_n$	S_{n1}	S_{n2}	S_{n3}	$S_{n…}$	$S_{n(n-1)}$

Tabelle 15: Ähnlichkeitsmatrix
Quelle: [Nits04], S. 28

6.4.7 Korpusbasierte Verfahren

Die korpusbasierten Verfahren versuchen einen wichtigen Nachteil anderer wichtiger Verfahren zu vermeiden. Während bei den meisten Verfahren davon ausgegangen wird, dass die Häufigkeit eines Wortes in einem Dokument von seinem allgemeinen Vorkommnis in einer Sprache abhängt, tritt diese Fehlannahme bei korpusbasierten Verfahren gar nicht erst auf. Stattdessen wird berücksichtigt, dass das Auftreten von Wörtern auch von ihren Bedeutungen abhängt. Veranschaulicht wird dies durch die Tatsache, dass verschiedene Wörter aufgrund ihrer Relation zueinander gemeinsam in einem Dokument auftauchen und nicht aufgrund ihres Vorkommens in der Sprache. Deswegen unterscheiden sich Verfahren, je nachdem ob sie Gemeinsamkeiten über das Auftreten von Termen dazu nutzen Daten auszuwerten oder aber die Termhäufigkeit. Terme, die gemeinsam auftreten, werden als sogenannte Kookkurrenzdaten in einer Häufigkeitstabelle gespeichert wie z.B. im folgenden Beispiel, in dem für Wörter die am häufigsten in Verbindung mit einem der drei Wörter Tax, Fruit und Sin die Auftrittshäufigkeit aufgelistet wird.

Tax	Fruit	Sin
Income: 71.81	eggs: 56.69	crime: 107.11
Fiscal: 66.96	meat: 56.69	doctrine: 98.31
Taxes: 61.99	foods: 55.09	morality: 92.00
Profits: 56.67	fresh: 54.99	adam: 87.57

Tabelle 16: Kookkurrenzdaten
Quelle: [Kump06], S. 79-80

In dieser Tabelle wird zudem der Quotient aus der relativen Häufigkeit des gemeinsamen Auftretens und dem Produkt der relativen Häufigkeiten der einzelnen

Terme ermittelt, also das gemeinsame Auftreten in Beziehung zur Dokumenthäufigkeit der Wörter gesetzt. Dadurch lässt sich deutlich erkennen, ob es einen inhaltlichen Zusammenhang zwischen den betrachteten Wörtern gibt. Ist dies nicht der Fall müssen alle Häufigkeiten ungefähr bei Eins liegen. Möchte man nun vergleichen wie ähnlich sich verschiedene Dokumente sind, werden verschiedene Ähnlichkeitsmaße verwendet. Darunter unter anderem das Cosinus-Maß, das Dice-Maß und das Jaccard-Maß. Dabei ist es in einigen komplizierteren Kookkurrenzverfahren möglich, nur Terme einzubeziehen, die bestimmten syntaktischen Regeln genügen oder beispielsweise drei aufeinanderfolgende Hauptwörter beinhalten. Auf dem Gebiet des Information Retrieval unterstützt die Ähnlichkeitsanalyse verschiedener Terme folgende Schritte: Die Indexierung bietet die Möglichkeit, direkt weitere Terme abzuleiten, um die ein verfügbarer Term ergänzt werden kann. Ein weiteres Anwendungsgebiet ist die Auswahl von entsprechenden Wörtern aus einem überprüften Vokabular. Wenn es die Vorgabe verlangt, dann werden die Stichwörter nicht aus einem Index des Dokuments verwendet, sondern es werden ähnliche Wörter aus einem Thesaurus verwendet.[123].

6.5 Zusammenfassung

Neben den hier angeführten Algorithmen existieren noch eine Vielzahl anderer Algorithmen und Subformen der bereits angeführten Algorithmen, deren Erwähnung aber den Umfang dieser Arbeit sprengen würde. Die oben zusammengefaßten Algorithmen. spiegeln die Mannigfaltigkeit unterschiedlicher Ansätze wider, wobei hier auf die Komplexittät dieser Ansätze und die Sinnhaftigkeit von Einsatz dieser Algorithmen geachtet werden soll. Die vorliegende Arbeit erhebt keinen Anspruch auf Vollständigkeit, sondern hat exemplarisch einige besonders weit verbreitete Methoden herausgegriffen und näher erläutert.

[123] vgl. [Kump06], S. 79-80

7 Textkategorisierung

7.1 Definition

Textkategorisierung (Klassifikation) ist die Aufgabe, jedem Paar $(d_j,c_i) \in D \times C$ *einen booleschen Wert zuzuordnen, wobei D eine Menge von Dokumenten und* $C = [c_1, \ldots, c_{|C|}]$ *eine Menge gegebener Kategorien ist. Der Wert T für das Paar* (d_j, c_i) *heißt, dass das Dokument* d_j *zu* c_i *zuzuordnen ist, während der Wert F bedeutet, das das Dokument* d_j *nicht zu* c_i *gehört*[124]. (Bemerkung: D=Dokument , C=Kategorie)
Wenn es z.B. 10 Dokumente und zwei Kategorien gibt, dann existieren 20 Paare (d_j, c_i) für die jeweils entschieden werden muß, ob das Paar T oder F bekommen muß.

Textkategorisierung ist die Zuordnung eines Textes zu einer oder mehreren Kategorien, also zu einem oder mehreren Themenbereichen.

7.2 Wozu Textkategorisierung?

Ziel der Textkategorisierung ist es, die schnelle und gezielte Auffindbarkeit von Dokumenten zu erreichen, die einer Textkategorie zugeordnet worden sind. Nur dann, wenn die Dokumente, die einer Kategorie zugeordnet wurden, auch wirklich zu dieser Kategorie „passen“, handelt es sich um eine gut verwendbare Kategorisierung von Dokumenten. Gerade die Genauigkeit der Kategorisierung stellt in Wirklichkeit das Problem dar, weil es nicht einfach ist, „gute“ Textkategorisierungsmethoden zu entwickeln[125]. Ein weiteres Ziel ist es selbstverständlich auch, Dokumente, die mehreren Kategorien zugeordnet sind, suchbar und auffindbar zu machen. In der Praxis wird meist auf Basis einer Volltextsuche in Kobination mit einer Suche in einer (oder mehreren) Kategorien gearbeitet, so dass man über den Volltext Dokumente finden kann, aber durch Auswahl einer oder mehrerer Katgorien die Suchergebnisse auf die gewählten Kategorien beschränken kann.

124 [Serb06], S. 15
125 vgl. [Gild02], S. 3

7.3 Art der Kategorisierung

Es gibt verschiedene Methoden und Algorithmen die bei Textkategorisierung verwendet werden[126]:

- Bayesian (Naive)
- Relevanzfeedback (Rocchio)
- Regelbasiertes Lernen (Ripper)
- Nächster Nachbar (fallbasierend)
- Support Vektor Maschinen (SVM)
- Neuronale Netze
- …

Häufig werden Kombinationen verschiedener Methoden und Algorithmen eingesetzt, um möglichst optimale Ergebnisse zu erreichen.

7.4 Schritte der Textkategorisierung

Die Textkategorisierung besteht aus folgenden grundlegenden Phasen:

- Dokumentvorbereitung:
  Das Dokument wird bearbeitet, eine Wortliste wird erstellt und es wird eine Indexierung durchgeführt. Dabei werden Terme extrahiert, Stoppwörter werden eliminiert und meist wird auch eine Grundformenreduktion mittels Stemmingverfahren durchgeführt.
- Klassifikation:
  Unter Zuhilfenahme von Textkategorisierungsalgorithmen wird das Dokument einer oder mehreren Kategorien zugeordnet und diese Zuordnungsergebnisse werden dauerhaft abgespeichert, um sie für die (spätere) Dokumentensuche verfügbar zu machen.
- Optional erfolgt eine Evaluierung der Klassifikationsergebnisse, bei der durch einen Experten eine Überprüfung der maschinell durchgeführten Kategorisierung durchgeführt wird. Dabei können dann wieder statistische Verfahren unterstützend verwendet werden.[127].

126 vgl. [HoSt07], S. 8
127 vgl. [Serb06], S. 15

7.5 Anwendung der Textkategorisierung

Aufgrund der großen Menge an Information mit der man es aktuell zu tun hat, wäre ein Leben ohne Kategorisierung von Texten und sonstigen „Dokumenten“ (auch Grafiken, Videos, Musik, …) nur noch schwer vorstellbar. Daraus folgt, dass nicht nur in Wissenschaft, Forschung und Industrie, sondern auch im Alltagsleben mittlerweile sehr häufig Anwendungen der Textkategorisierung anzutreffen sind. Einige Beispiele für derartige Kategorisierungen sind Webseiten, Newsgroups, Bibliotheksverzeichnisse, usw.

In Abhängigkeit davon, welche Algorithmen bei der Kategorisierung verwendet werden und welche „Schärfe“ bei der Kategorisierung angewandt wird, sind die Ergebnisse der Kategorisierung sehr unterschiedlich.

Eine typische Anwendung ist die Filterung von „Top News“ für eine bestimmte Interessentengruppe, die z.B. Neuigkeiten über Sport haben möchte, die aber nicht interessiert ist an Wirtschafts- oder Börsennachrichten[128]. Im Internet wird bei derartigen Anwendungen meist eine sehr einfache Form der Volltextkategorisierung verwendet, die oft entsprechend schlechte Ergebnisse liefert oder es wird eine manuelle Textkategorisierung angewendet, bei der Redakteure einzelne Dokumente manuell bestimmten Kategorien zuordnen.

128 vgl. [HoSt07], S. 7

8 Praktische Anwendung bei einem Projekt

8.1 Das Projekt

Anhand der Bürgerzufriedenheitsbefragung einer österreichischen Bezirkshauptstadt lässt sich die praktische Anwendung der vorgestellten Algorithmen mit besonderem Augenmerk auf der Analyse unstrukturierter Textdaten anschaulich demonstrieren. Die gegenständliche Umfrage ist in der Bezirkshauptstadt selbst und in den zugehörigen Katastralgemeinden durchgeführt worden, wobei im Rahmen der Umfrage eine Klassifizierung der teilnehmenden Personen nach Geschlecht, Alter und Katastralgemeinde vorgenommen worden ist.

Es sind insgesamt 3000 Fragebögen nach dem Zufallsprinzip an über 16 jährige Bewohner verschickt worden und 947 ausgefüllte Fragebögen haben an der Auswertung teilgenommen. Die Bezirkshauptstadt führt alle zwei Jahre eine derartige Bürgerzufriedenheitsumfrage durch, deren Hauptziel die Beurteilung der Qualität der Leistungen der Stadtverwaltung durch die Bürger ist.

8.2 Basisdaten

Der Fragebogen besteht aus 18 Fragen, die zum Teil noch in einzelne Sub-Fragen untergliedert sind. Ein Großteil der Fragen ist quantitativer Art, wobei eine ungefähre Gleichverteilung zwischen Multiple-Choice Fragen und anderen quantitativen Fragen gegeben ist.

Die erhobenen Daten liegen in Form einer Textdatei und als Abschrift der erhobenen Daten in Form einer PDF-Datei vor. Zur Demonstration wird eine offene Frage bzw. die Antworten auf diese Frage herangezogen. Bei der Auswahl der Frage für die Analyse ist der ausschlaggebende Grund für die Auswahl dieser Frage der Umfang der Bentwortung dieser Frage und der Charakter der Fragestellung gewesen. Die Antworten sind umfangreich, wobei die ausgewählte Frage lautet, worüber besser bzw. mehr informiert werden sollte.

8.3 Beispielhafte Analyse

Aus dem Fragebogen wurde Frage 3 ausgewählt und diese Frage wurde von 221 Teilnehmern der Befragung beantwortet. Die nun folgende beispielhafte Datenanalyse wurde absichtlich mit sehr einfachen Mitteln durchgeführt, um etwas Transparenz beim Ablauf zu ermöglichen und beispielhaft den Ablauf der Datenanalyse zu demonstrieren.

8.3.1 Basisdaten

Auszug der zu analysierenden Antworten auf Frage 3:

Frage 3:			**Über welche Inhalte wird Ihrer Meinung nach zu wenig informiert? Was müsste noch mehr bekanntgemacht werden?**
G	A	K	Text
1	1	3	B 301, einzelne Bauabschnitte, Flughafen dritte Piste
1	1	4	Arbeiten, Taxi
1	1	4	Sportliche Ereignisse
1	2	1	Umweltbelastung durch den Flughafen, die Industrien und durch die geplante S1
1	2	3	Jugendberatung und Jugendförderung
1	2	3	S1 (noch immer zu wenig Info über Baufortschritt)
1	2	3	Infrastrukturveränderungen, derzeitiger Wohnungsbestand der Gemeinde
1	2	4	Gesundheitsinformationen--> Fachärzte, Veranstaltungen, Werbeveranstaltungen, Infrastruktur der Rettung
1	2	4	Veranstaltungen in der Körnerhalle
1	2	4	Veranstaltungen im Forum und im Fehlmayergarten
1	3	4	Busfahrpläne
1	3	4	Haustierhaltung (Anmeldung, Leinenpflicht, Freilaufzone,...)
1	3	4	Neubau Sonderschule/Musikschule, Planung Frauenfeld, Inhalt und Ergebnisse von Gemeinderatssitzungen, Ziele (konkret) des Bürgermeisters für je ein Jahr
1	3	4	Umwelt, Energie, Wasserwirtschaft, Kultur
1	3	4	Deutsch- bzw. Englischkurse
1	3	4	Genaue Kennzeichnung der Kurzparkzonen
1	4	1	Veranstaltungen, Interviews
1	4	2	In Mannswörth werden Veranstaltungen zu wenig und zu spät plakatiert.

Abbildung 9: Auszug Basisdaten
Quelle: eigene Auswertung, eigene Darstellung

Spalteninhalte:

- G Geschlecht (1: männlich, 2: weiblich, 3: keine Angabe)
- A Alter

	Alter (A):
1	Bis 19 Jahr
2	19-24
3	25-29
4	30-39
5	40-49
6	50-59
7	60-69
8	70 Jahre oder älter
9	Keine Angabe

- K Katastralgemeinde (durchnummeriert)
- Text Antwort des Befragten

8.3.2 Analyse

8.3.2.1 Tokenisierung (Word-Splitting)

Um vom der Satzebene zur Wortebene gelangen zu können, muss jede Antwort in ihre einzelnen Worte zerteilt werden. Die Fragebogennummer (Spalte D) und die anderen vorhandenen Eckdaten bleiben erhalten und die vorher noch als Sätze vorliegenden Antworten werden aufgeteilt. Diesse Aufteilung von Sätzen bzw. Wortgruppen erfolgt auf Basis der Satzzeichen und von Leerzeichen.

	A	B	C	D	E	F	G	H	I	J	K
1	1	1	3	1	B	301	einzelne	Bauabschnitt	Flughafen	dritte	Piste
2	1	1	4	2	Arbeiten	Taxi					
3	1	1	4	3	Sportliche	Ereignisse					
4	1	2	1	4	Umweltbelastung						
5	1	2	3	5	Jugendberatung	und	Jugendförderung				
6	1	2	3	6	S1	(noch	immer	zu	wenig	Info	über
7	1	2	3	7	Infrastrukturveränderungen	derzeitiger	Wohnungsbe	der	Gemeinde		
8	1	2	4	8	Gesundheitsinformationen-	Fachärzte	Veranstaltung	Werbeveranst	Infrastruktur	der	Rettung
9	1	2	4	9	Veranstaltungen	in	der	Körnerhalle			
10	1	2	4	10	Veranstaltungen	im	Forum	und	im	Fehlmayergarten	
11	1	3	4	11	Busfahrpläne						
12	1	3	4	12	Haustierhaltung	(Anmeldung	Leinenpflicht	Freilaufzone	)		
13	1	3	4	13	Neubau	Sonderschule	Planung	Frauenfeld	Inhalt	und	Ergebnisse
14	1	3	4	14	Umwelt	Energie	Wasserwirtsc	Kultur			
15	1	3	4	15	Deutsch-	bzw	Englischkurse				
16	1	3	4	16	Genaue	Kennzeichnu	der	Kurzparkzonen			
17	1	4	1	17	Veranstaltungen	Interviews					
18	1	4	2	18	In	Mannswörth	werden	Veranstaltung	zu	wenig	und
19	1	4	2	19	Pläne	über	öffentliche	Verkehrsmitte	Infos	für	Kinder
20	1	4	2	20	Kunstausstellungen						
21	1	4	2	21	Mannswörth						
22	1	4	3	22	Bauvorhaben	im	Bezirk	Neuschaffung	von	Wohnungen	
23	1	4	3	23	B	301					
24	1	4	3	24	Kultur-	und	Freizeitangebote				
25	1	4	3	25	Luftgüte	diverse	Veranstaltungen				
26	1	4	3	26	B	301/S1					
27	1	4	4	27	Zukünftige	Planung	z	B	Verbauung	Frauenfeld	
28	1	4	4	28	Straßenumbau	Thurnmühlstr	Sportveransta	anfällige	Stadtmeisterschaften		
29	1	4	4	29	Brauerei	Lärmschutz-F	Ökologie	Hundeexperimente			
30	1	4	4	30	Einzelne	Bereiche	Siedlungen	Stadtkern			
31	1	4	4	31	Abfallwirtschaft	Sperrmüllsam	Sicherheit				
32	1	4	4	32	Umwelt	Luftgüte	"Made	in	XYZ"		

Abbildung 10: gesplittete Daten
Quelle: eigene Auswertung, eigene Darstellung

8.3.2.2 Vereinheitlichung

Aus dieser tokenisierten Darstellung der Antworten lässt sich nun leicht ein erster Volltextindex (mit bis zu diesem Zeitpunkt noch mehrfach vorkommenden Worten) erstellen, indem alle Einzelworte in eine Liste mit nur mehr einem Wort pro Datensatz transformiert werden. Für spätere Analyseschritte muß natürlich die Fragebogennummer mitgeführt werden, um diese Information noch verwenden zu können.

Nun erfolgt eine Umwandlung aller Worte des bisherigen Volltextindexes in Kleinbuchstaben, um den Einfluss von unterschiedlichen Behandlungen, die aus Groß- und Kleinschreibung resultieren, auszuschließen.

Eine abschliessende Sortierung des Volltextindexes führt zu einer alphabetisch sortierten Liste aller Worte (mit Mehrfachvorkommen einzelner Worte).

	A	B	C	D
1	1	B	b	
2	1	301	301	
3	23	301	301	
4	46	301	301	
5	48	301	301	
6	49	301	301	
7	132	301	301	
8	210	301	301	
9	124	39570	39570	
10	26	301/S1	301/s1	
11	141	30km/h-Zone	30km/h-zone	
12	102	a	a	
13	142	aber	aber	
14	90	Abfahrtszeiten	abfahrtszeiten	
15	31	Abfallwirtschaft	abfallwirtschaft	
16	159	Abfallzentren	abfallzentren	
17	192	Abgaben	abgaben	
18	150	Abläufe	abläufe	
19	44	Ahnung	ahnung	
20	120	Aktionen	aktionen	
21	122	Aktivitäten	aktivitäten	
22	135	Aktivitäten	aktivitäten	
23	171	Alanova-Radweg	alanova-radweg	
24	184	Alanova-Radweg	alanova-radweg	
25	131	alleinerziehende	alleinerziehende	
26	125	Alleinerzieher	alleinerzieher	
27	42	allem	allem	
28	87	Alles	alles	
29	159	Alles	alles	
30	168	alles	alles	
31	181	alles	alles	

Abbildung 11: alphabetisch sortierter Volltextindex
Quelle: eigene Auswertung, eigene Darstellung

Spalten:

- A Fragebogennummer
- B Wort aus den Basisdaten
- C Zu Kleinbuchstaben konvertiertes Wort

Auffällig ist an dieser Stelle, dass die Wort „B“ und „301“ getrennt wurden, weil sie in allen Antworten durch ein Leerzeichen getrennt vorliegen. Bei Anwendung eines lexikonbasierten Verfahrens zur Tokenisierung hätte bereits eine Erkennung des Wortes „B 301“ stattfinden können. („B 301“ ist die Bezeichnung für eine neue Bundesstraße).

Um unterschiedliche Schreibweisen einheitlich behandeln zu können werden nun alle Umlaute und das „ß“ in ihre jeweilige zweistellige Schreibweise transformiert:

	A	B	C	D	E	F	G	H	I
1			**klein**	**ä -> ae**	**ö -> oe**	**ü -> ue**	**ß -> ss**	**endwerte**	
2	142	aber	aber	aber	aber	aber	aber	aber	1
3	90	Abfahrtszeiten	abfahrtszeiten	abfahrtszeiten	abfahrtszeiten	abfahrtszeiten	abfahrtszeiten	abfahrtszeiten	1
4	31	Abfallwirtschaft	abfallwirtschaft	abfallwirtschaft	abfallwirtschaft	abfallwirtschaft	abfallwirtschaft	abfallwirtschaft	1
5	159	Abfallzentren	abfallzentren	abfallzentren	abfallzentren	abfallzentren	abfallzentren	abfallzentren	1
6	192	Abgaben	abgaben	abgaben	abgaben	abgaben	abgaben	abgaben	1
7	150	Abläufe	abläufe	ablaeufe	ablaeufe	ablaeufe	ablaeufe	ablaeufe	1
8	44	Ahnung	ahnung	ahnung	ahnung	ahnung	ahnung	ahnung	1
9	120	Aktionen	aktionen	aktionen	aktionen	aktionen	aktionen	aktionen	1
10	122	Aktivitäten	aktivitäten	aktivitaeten	aktivitaeten	aktivitaeten	aktivitaeten	aktivitaeten	1
11	135	Aktivitäten	aktivitäten	aktivitaeten	aktivitaeten	aktivitaeten	aktivitaeten	aktivitaeten	1
12	171	Alanova-Radweg	alanova-radweg	alanova-radweg	alanova-radweg	alanova-radweg	alanova-radweg	alanova-radweg	1
13	184	Alanova-Radweg	alanova-radweg	alanova-radweg	alanova-radweg	alanova-radweg	alanova-radweg	alanova-radweg	1
14	131	alleinerziehende	alleinerziehende	alleinerziehende	alleinerziehende	alleinerziehende	alleinerziehende	alleinerziehende	1
15	125	Alleinerzieher	alleinerzieher	alleinerzieher	alleinerzieher	alleinerzieher	alleinerzieher	alleinerzieher	1
16	42	allem	allem	allem	allem	allem	allem	allem	1
17	87	Alles	alles	alles	alles	alles	alles	alles	1
18	159	Alles	alles	alles	alles	alles	alles	alles	1
19	168	alles	alles	alles	alles	alles	alles	alles	1
20	181	alles	alles	alles	alles	alles	alles	alles	1

Abbildung 12: Normalisierung der Umlaute und des ß
Quelle: eigene Auswertung, eigene Darstellung

8.3.2.3 Eliminierung der Stoppwörter

Da in der Regel in Antworten auf offene Fragen eine beträchtliche Menge an Stoppwörtern vorhanden ist, müssen nun die Stoppwörter entfernt werden. Stoppwörter sind Worte mit einer hohen Auftrittshäufigkeit und so gut wie keinem Informationsgehalt haben (Details in Abschnitt 6.4.1).

	A	B	C	D	E
1	Antwort	Wort	Anzahl	Stoppwort	Klartext (0: kein Stopwort; 1: ist ein Stoppwort)
2	1	b	1	#NV	0
3	1	301	1	#NV	0
4	23	301	1	#NV	0
5	46	301	1	#NV	0
6	48	301	1	#NV	0
7	49	301	1	#NV	0
8	132	301	1	#NV	0
9	210	301	1	#NV	0
10	124	39570	1	#NV	0
11	26	301/s1	1	#NV	0
12	141	30km/h-zone	1	#NV	0
13	102	a	1	#NV	0
14	**142**	**aber**	**1**	**1**	**1**
15	90	abfahrtszeiten	1	#NV	0
16	31	abfallwirtschaft	1	#NV	0
17	159	abfallzentren	1	#NV	0
18	192	abgaben	1	#NV	0
19	150	ablaeufe	1	#NV	0
20	44	ahnung	1	#NV	0
21	120	aktionen	1	#NV	0
22	122	aktivitaeten	1	#NV	0
23	135	aktivitaeten	1	#NV	0
24	171	alanova-radweg	1	#NV	0
25	184	alanova-radweg	1	#NV	0
26	131	alleinerziehende	1	#NV	0
27	125	alleinerzieher	1	#NV	0
28	42	allem	1	#NV	0
29	87	alles	1	1	1
30	159	alles	1	1	1
31	168	alles	1	1	1

Abbildung 13: Stoppwörter markieren
Quelle: eigene Auswertung, eigene Darstellung

Anhand einer sogenannten Stoppwortliste erfolgt eine Markierung der im Volltext vorhandenen Stoppworte in Spalte E (0: nicht in Stoppwortliste enthalten, 1: in Stoppwortliste enthalten). Es wurde für dieses Beispiel eine Stoppwortliste aus dem Internet (http://www.phpbar.de/w/Stoppwortliste_deutsch) verwendet. Das Wort „aber“ ist in der verwendeten Stoppwortliste nicht enthalten und wird daher nicht als Stoppwort erkannt (siehe obenstehende Abbildung).

Nach der Markierung der Stoppwörter können diese aus den zu analysierenden Daten entfernt werden,

	A	B	C	D	E
1	Antwort	Wort	Anzahl	Stoppwort	Klartext (0: kein Stopwort; 1: ist ein Stoppwort)
2	1	b	1	#NV	0
3	1	301	1	#NV	0
4	23	301	1	#NV	0
5	46	301	1	#NV	0
6	48	301	1	#NV	0
7	49	301	1	#NV	0
8	132	301	1	#NV	0
9	210	301	1	#NV	0
10	124	39570	1	#NV	0
11	26	301/s1	1	#NV	0
12	141	30km/h-zone	1	#NV	0
13	102	a	1	#NV	0
14	90	abfahrtszeiten	1	#NV	0
15	31	abfallwirtschaft	1	#NV	0
16	159	abfallzentren	1	#NV	0
17	192	abgaben	1	#NV	0
18	150	ablaeufe	1	#NV	0
19	44	ahnung	1	#NV	0
20	120	aktionen	1	#NV	0
21	122	aktivitaeten	1	#NV	0
22	135	aktivitaeten	1	#NV	0
23	171	alanova-radweg	1	#NV	0
24	184	alanova-radweg	1	#NV	0
25	131	alleinerziehende	1	#NV	0
26	125	alleinerzieher	1	#NV	0
27	42	allem	1	#NV	0
28	149	allgemein	1	#NV	0

Abbildung 14: Stopwörter endgültig entfernt
Quelle: eigene Auswertung, eigene Darstellung

Es resultiert ein stoppwortfreier Volltextindex.

8.3.2.4 Stemming

Da im vorliegenden Volltextindex Worte in unterschiedlichen Flexionsformen vorkommen muß nun eine Grund- bzw. Stammformreduktion durchgeführt werden. In diesem Beispiel wird eine Analyse der einzelnen Worte mit Hilfe einer Internetressource (http://www.canoo.net) durchgeführt, um erste brauchbare Ergebnisse zu erhalten. Details zur Grundformreduktion finden sich in Kapitel 6 dieser Arbeit.

	A	B	C	D	E
1	Antwor	Wort	Anzahl	Wort gestemmt (canoo.net)	Wort
2	132	301	1	301	301
3	210	301	1	301	301
4	124	39570	1	39570	39570
5	26	301/s1	1	301/s1	301/s1
6	141	30km/h-zone	1	30km/h-zone	30km/h-zone
7	102	a	1	a	a
8	90	abfahrtszeiten	1	abfahrtszeit	abfahrtszeit
9	31	abfallwirtschaft	1	müllwirtschaft	müllwirtschaft
10	159	abfallzentren	1	abfall	abfall
11	192	abgaben	1	abzug	abzug
12	150	ablaeufe	1	entwicklung	entwicklung
13	82	aeltere	1	alt	alt
14	92	aerztliche	1	aerztlich	aerztlich
15	44	ahnung	1	ahnung	ahnung
16	120	aktionen	1	aktionen	aktionen
17	122	aktivitaeten	1	handlung	handlung
18	135	aktivitaeten	1	handlung	handlung
19	171	alanova-radweg	1	alanova-radweg	alanova-radweg
20	184	alanova-radweg	1	alanova-radweg	alanova-radweg
21	131	alleinerziehende	1	alleinerziehend	alleinerziehend
22	125	alleinerzieher	1	alleinerzieher	alleinerzieher
23	42	allem	1	allem	allem
24	149	allgemein	1	allgemein	allgemein
25	150	allgemein	1	allgemein	allgemein
26	56	an	1	an	an
27	69	an	1	an	an
28	192	an	1	an	an
29	28	anfaellige	1	anfaellige	anfaellige
30	95	angebote	1	angebot	angebot
31	100	angebote	1	angebot	angebot
32	138	angestellten	1	angestellt	angestellt
33	173	anliegen	1	anliegen	anliegen

Abbildung 15: Grundformreduktion und Synonymbildung
Quelle: eigene Auswertung, eigene Darstellung

Durch diese durchgeführten Vereinheitlichungen lässt sich nun ein stoppwortfreier Index bilden, in dem Worte nicht mehr mehrfach vorkommen.

8.3.2.5 Analyse Worthäufigkeiten

Durch einfaches Zählen der Vorkommen der einzelnen Worte im Index ergibt sich nun eine Worthäufigkeit für jedes Wort. Wenn man dieses Zwischenergebnis nach der Anzahl des Auftretens absteigend sortiert, dann gelangt man zu einer Worthäufigkeitstabelle der häufigsten Wörter bzw. Wortstämme. Eine Anwendung komple-

xerer statistischer Verfahren, um die Termfrequenz innerhalb einer Frage mit der Termfrequenz über alle Fragen gegenüberzustellen, macht bei den vorliegenden zu analysierenden Daten wenig Sinn, weil aufgrund der Kürze der Antworten die Wahscheinlichkeit sehr gering ist, dass überhaupt ein Nicht-Stoppwort in einer Antwort öfter als ein Mal vorkommt.

	A	B
1	**Wort**	**Anzahl**
2	veranstaltungen	30
3	in	28
4	b	14
5	diverse	11
6	im	11
7	zu	11
8	XYZ	10
9	bauvorhaben	9
10	angebot	8
11	koernerhalle	8
12	mannswoerth	8
13	s1	8
14	301	7
15	es	7
16	weiß	7
17	wenig	6
18	z	6
19	beiraete	5
20	XYZer	5
21	stadt	5
22	gemeinde	4
23	information	4
24	kultur	4
25	sport	4
26	an	3
27	ausreichend	3
28	flughafen	3
29	info	3
30	inhalte	3
31	kinder	3
32	kulturelle	3
33	moeglichkeiten	3
34	omv	3
35	reicht	3
36	schulden	3
37	sicherheit	3
38	sportveranstaltungen	3
39	alanova-radweg	2

Abbildung 16: Index mit Worthäufigkeit (absteigeend nach Häufigkeit sortiert).
Quelle: eigene Auswertung, eigene Darstellung

Aus den nun vorliegenden Daten lässt sich sehr einfach eine Visualisierung der häufigsten Worte erstellen:

	A	B
2	veranstaltung	30
3	in	28
4	b	14
5	diverse	11
6	im	11
7	zu	11
8	XYZ	10
9	bauvorhaben	9
10	angebot	8
11	koernerhalle	8
12	None name	8
13	s1	8
14	301	7
15	es	7
16	weiß	7
17	wenig	6
18	z	6
19	beiraete	5
20	schwechater	5
21	stadt	5
22	gemeinde	4
23	information	4
24	kultur	4
25	sport	4
26	an	3
27	ausreichend	3
28	flughafen	3
29	info	3
30	inhalte	3
31	kinder	3
32	kulturelle	3
33	moeglichkeite	3
34	omv	3
35	reicht	3
36	schulden	3
37	sicherheit	3
38	sportveransta	3

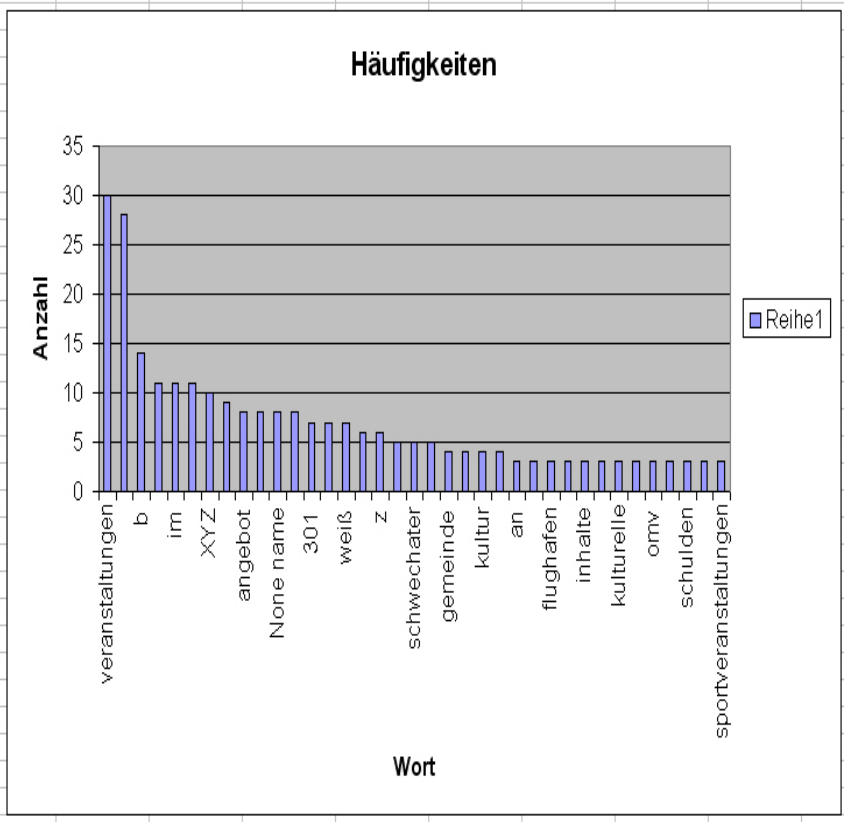

Abbildung 17: Häufigste Worte in den Antworten
Quelle: eigene Auswertung, eigene Darstellung

8.4 Ergebnis

In Abhängigkeit von der verwendeten Tokenisierungsmethodik, der eingesetzten Stoppwortliste und vor allem der Methode der Grundformreduktion lässt sich grundsätzlich das Ergebnis beeinflussen. Bei der Analyse der vorliegenden Beispieldaten ist aufgrund der Kürze der Antworten und aufgrund der Struktur der Fragestellung auch durch Einsatz komplizierterer Methoden für die Datenanalyse kein wesentlich aussagekräftigeres Ergebnis erreichbar, wenn nicht durch Erzeugung eines individuellen Lexikons und Verwendung lexikonbasierter Verfahren und/oder durch semantische Analyse der Aufwand für die Analyse deutlich erhöht wird.

9 Zusammenfassung und Ausblick

9.1 Erfolgsfaktoren

Der Einsatz von Software zur Analyse unstrukturierter Texte hat unter Beachtung hoher Qualitätskriterien zu erfolgen, weil infolge der vollautomatischen Verarbeitung und der verwendeten Algorithmen und ihren verwendeten Parametern gegenüber manuellen Analysemethoden große Gefahr besteht, infolge von automatisierten Abläufen in der verwendeten QDA-Software Irrtümern zu unterliegen, diese Irrtümer aber nicht zu bemerken.

Da infolge sprachspezifischer Eigenheiten verschiedene Behandlung unterschiedlicher Sprachen bei der Analyse unstrukturierter Texte notwendig ist, wurde im Rahmen der vorliegenden Arbeit deutlich aufgezeigt, daß neben den Grundanforderungen an QDA-Software die Unterstützung verschiedener Sprachen eine wichtige Voraussetzung für den Erfolg zukünftiger QDA-Software darstellt.

Das zu analysierende Datenmaterial spielt eine sehr große Rolle für die möglichen sinnvollen Analysen, weil selbst hochkomplexe Analysemethoden bei unzureichenden Basisdaten keine brauchbaren Ergebnisse liefern (können).

Speziell der Einsatz lexikonbasierter Verfahren scheint die Qualität der Analyse selbst bei rudimentären Basisdaten wesentlich verbessern zu können, ist aber mit entsprechendem Aufwand für die Erstellung und Pflege eine Lexikons verbunden.

9.2 Offene Fragen und Ausblick

Selbst bei intensiver Nutzung von Statistik und Informatik fehlt jeder vollautomatischen Analysemethode unstrukturierter Texte eine der wichtigsten Grundlagen für eine tatsächlich korrekte Analyse: das Verständnis des Inhaltes unter Beachtung sprachspezifischer Syntax und Semantik.

Da der Forschungsbereich der Künstlichen Intelligenz zwar Fortschritte macht, aber in näherer Zukunft keine Quantensprünge zu erwarten sind, ist aus Sicht der Autorin auch in näherer Zukunft nicht mit rasanten Fortschritten bei der Analyse qualitativer Daten in Form unstrukturierter Texte zu rechnen, was aber keinen Widerspruch zum sinnvollen und intensiven Einsatz der vorhandenen Methoden der computerunterstützten und vollautomatischen Analyse qualitativer Daten darstellt, wenn die Analyse unter Wahrung der dargestellten Qualitätskriterien durchgeführt werden.

10 Literaturverzeichnis

10.1 Literatur

[Alle72] K. Allerbeck: Datenverarbeitung in der empirischen Sozialforschung, B. G. Teubner Verlag, Stuttgart, 1972.

[AlNi00] P. Alpar, J. Niedereichholz: Data Mining im praktischen Einsatz; Vieweg Gabler Verlag, 1.Auflage, 2000.

[Atte00] P. Atteslander: Methoden der empirischen Sozialforschung; Walter de Gruyter-Verlag, Berlin, New York, 9. Auflage, 2000.

[Atte06] P. Atteslander: Methoden der empirischen Sozialforschung; Walter de Gruyter-Verlag, Berlin, New York, 10. Auflage, 2006.

[Bohn99] R. Bohnsack: Rekonstruktive Sozialfoschung; Leske+Budrich Verlag, 3. Auflage, 1999.

[Bros07] F. Brosius Felix: SPSS für Dummies; Wiley-VCH Verlag GmbH & Co.KG, 2007.

[BuHo07] R. Buber, H. H. Holzmüller: Qualitative Marktforschung – Konzepte-Methoden-Analyse; Gabler; 1. Auflage, 2007.

[BüZö02] A. Bühl, P. Zöfel: SPSS11 - Einführung in die moderne Datenanalyse unter Windows; Pearson Studium Verlage, 8. Auflage, 2002.

[Diek06] A. Diekmann: Empirische Sozialforschung - Grundlagen, Methoden, Anwendungen; Rowohlts Enzyklopädie Verlag, Reinbek bei Hamburg, 15. Auflage, 2006.

[Döri02] B. Döring: Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler; Springer Verlag, 3. Auflage, 2002.

[Frie90] J. Friedrichts: Methoden der empirischen Sozialforschung; Westdeutscher Verlag, Opladen, 1990.

[FrLu03] U. Froschauer, M. Lueger: Das qualitative Interview (zur Praxis Interpretative Analyse sozialer Systeme); WUV Universität-Verlag, UTB; 1. Auflage, 2003.

[Fugm99] R. Fugmann: Inhaltserschließung durch Indexieren: Prinzipien und Praxis, Informationswissenschaft der DGD; Frankfurt a.M., 3. Auflage, 1999.

[Kais93] A. Kaiser: Computerunterstütztes Indexieren in Intelligenten Information Retrieval Systemen - ein Relevanz- Feedbackorientierter Ansatz zur Informationserschließung in unformatierten Datenbanken; Dissertation Wirtschaftsuniversität Wien, Wien, 1993.

[Kauf01] E. Kaufmann: Das Indexieren von natürlichsprachigen Dokumenten und die inverse Seitenhäufigkeit; Lizentiatsarbeit, Universität Zürich, 2001.

[KrWiZi98] D. Krahl, U. Windheuser, K. F. Zick: Data Mining Einsatz in der Praxis; Addison-Wesley Verlag, 1. Auflage, 1998.

[Mayr93] P. Mayring: Einführung in die qualitative Sozialforschung; Beltz Psychologie Verlagsunion, Weinheim, 2. Auflage, 1993.

[Mayr97] P. Mayring: Qualitative Inhaltanalyse – Grundlagen und Techniken; Beltz Verlag, 6. Auflage, 1997.

[MaZi05] P. Mayring, M. G. Zikuda: Die Praxis der Qualitativen Inhaltanalyse; Beltz Velag, 2005.

[MeWi99] P. Mertens, H. W. Wieczorrek: Data X Strategien; Springer Verlag, Deutschland, 5. Auflage, 1999.

[Moen00] M. F. Moens: Automatic Indexing and Abstracting of Document Texts; Springer Verlag, USA, 1. Auflage, 2000.

[Nohr03] H. Nohr: Grundlagen der automatischen Indexierung; Logos Verlag Berlin, 2003.

[Nohr05] H. Nohr: Grundlagen der automatischen Indexierung; Logos Verlag Berlin, 2005.

[OtWi96] T. Ottmann, P. Widmayer: Algorithmen und Datenstrukturen; Spektrum akademischer Verlag, Heidelberg Berlin Oxford, 3. Auflage, 1996.

[RoHo99] E. Roth, H. Holling: Sozialwissenschafliche Methoden. Lehre und Handbuch für Forschung und Praxis; Oldenberg, 5 Auflage, 1999.

[ScHiEs05] R. Schnell, P. B. Hill, E. Esser: Methoden der empirischen Sozialforschung; R. Oldenbourg Verlag, 7. Auflage, München Wien, 2005.

[ScHiEs99] R. Schnell, P. B. Hill, E. Esser: Methoden der empirischen Sozialforschung; R. Oldenbourg Verlag, München Wien, 6.Auflage, 1999.

[Schr99] N. Schreiber: Wie mache ich Inhaltanalysen – vom Untersuchungsplan zum Ergebnisbericht?; R. G. Fischer Verlag, 1999.

[Sedg02] R. Sedgewick: Algorithmen; Addison Wesley Verlag, 2. Auflage, 2002.

[SeJaDeCo72] C. Selltiz, M. Jaboda, M. Deutsch, S. W. Cook: Untersuchungsmethoden der Sozialforschung Teil II; Hermann Luchterhand Verlag, 1972.

[Stoc00] W. G. Stock: Informationswirtschaft: Management externen Wissens; Oldenbourg Wissenschaftsverlag GmbH, München Wien, 2000.

[Stra94] A. L. Strauss: Grundlagen qualitativer Sozialforschung; Wilhelm Fink Verlag, 1994.

[Vier97] R. Vierler: Einführung in die Stochastik; Springer Verlag, Wien New York; 2. Auflage, 1997.

[Webe03] C. Weber: Realisierung einer automatischen Klassifizierungs- und Beschlagwortungskomponente zur Nutzung im Projekt META-AKAD; Universität Kaisesrslautern, 2003.

[Wiek99] J. H. Wieken: Der Weg zum Data Warehouse; Addison-Wesley München Verlag, 1999.

[Wind08] T. Winder: Skriptum – Auswertung von Marktforschungsdaten mit SPSS; Institut für Absatzwirtschaft, Wirtschaftsuniversität Wien, 2008.

10.2 Internetquellen

[Batt05] J. Battenfeld: Benutzer - Matching auf Basis automatischer Textanalyse - Ein Ansatz zur Ähnlichkeitsbestimmung von Benutzern durch Dokumentenanalyse für das Expert Finder Framework.

http://www.uni-siegen.de/fb5/wirtschaftsinformatik/publikationen/diplomarbeiten/pdf/da_battenfeld--benutzer-matching_auf_basis_automatischer_textanalyse--2005.pdf (Zugriff: 18.05.2008), 2005

[Biemo.J.] C. Biemann: Morphologische Grundformreduktion.

http://www.asv.informatik.uni-leipzig.de/opencms/opencms/asv/de/Ueber_die_ASV/ASV-Verfahren/Grundformreduktion.html (Zugriff: 02.04.2008), o.J.

[Bünz01] A. Bünzli: Information Retrieval – Eine Einführung in das Indexieren.

http://www.ifi.unizh.ch/cl/hess/classes/seminare/semrep/irarbeit.pdf (Zugriff: 15.03.2008), 2001

[Dast00] P. Dastani: Data Mining – eine Einführung.

http://www.database-marketing.de/miningmining.htm (Zugriff: 21.11.2007), 2000

[DaHe98] G. Dabiri, D. Helten: Psychologie und Internet - Psychologische Grundlagenstudie zum Phänomen Internet Relay Chat Qualitative Analyse der Bedeutungsschwerpunkte für die Anwender.

http://userpage.fu-berlin.de/~chlor/ (Zugriff: 24.05.2008), 1998

[DaSe04] A. Dachtler, J. Senske: Data Mining.

http://www.heindl.de/KI2004/datamining/Ausarbeitung_KI.pdf (Zugriff: 28.11.2007), 2004

[Enge04] U. Engel: Statistik und empirische Sozialforschung.

http://mlecture.uni-bremen.de/intern/ws2004_2005/fb08/vak-08-b709/20041029/ (Zugriff: 26.10.2007), 2004

[Ferb03] R. Ferber: Information Retrieval – Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web.

http://information-retrieval.de/irb/ir.part_1.chapter_3.section_2.subdiv1_3.html (Zugriff: 15.04.2008), 2003

[Frie04] S. Friese: Software-Vergleich / Software Overview.

http://www.quarc.de/software_overview_table.pdf (Zugriff: 24.05.2008), 2004

[Frie08] S. Friese: Software-Vergleich / Software Overview.

http://www.quarc.de/einfuehrung.html (Zugriff: 24.05.2008), 2008

[Gild02] S. Gildner: Textkategorisierung.

http://wwwcs.uni-paderborn.de/cs/ag-klbue/de/courses/ws01/wwwsearch01/gildner-seminar-categorization.pdf (Zugriff: 27.05.2008), 2002

[Hald02] L. Haldemann: Datenstrukturen im WWW.

http://www.inf.uni-konstanz.de/dbis/teaching/ss01/data-on-the-web/local/datenstrukturen.pdf (Zugriff: 15.03.2008), 2002

[Hali05] I. Halip: Automatische Extrahierung von Schlagworten aus unstrukturierten Texten.

http://www-wi.uni-muen-ster.de/pi/lehre/ss05/seminarSuchen/Ausarbeitungen/IoanaHalip.pdf (Zugriff: 15.03.2008), 2005

[Henr02] J. Henrich: Indexierung und Kategorisierung von Dokumenten – Verfahren, Grenzen und praktische Anwendung.

http://www.wifrankfurt.de/veranstaltung/Groffmann_SS02/5_henrich.pdf (Zugriff: 15.03.2008), 2002

[Henr07] A. Henrich: Information Retrieval Grundlagen, Modelle und Anwendungen.

http://www.uni-bam-berg.de/fileadmin/uni/fakultaeten/wiai_lehrstuehle/medieninformatik/Dateien/Publikationen/2007/henrich-ir1-1.1.pdf (Zugriff: 18.05.2008), 2007

[Heis08] A. Heistinger: Qualitative Interviews – Ein Leitfaden zu Vorbereitung und Durchführung inklusive einiger theoretischer Anmerkungen.

http://www.univie.ac.at/igl.geschichte/kaller-dietrich/WS%2006-07/MEXEX_06/061102Durchf%FChrung%20von%20Interviews.pdf (Zugriff: 12.01.2008), 2008

[Hein99] A. Heinrich: Information Retrieval Grundlagen, Modelle, Implementierung und Anwendungen.

http://131.246.18.10/courses/proseminar/WS0405/Vorlesungsunterlagen/Information_Retrieval.full.pdf (Zugriff: 02.04.2008), 1999

[HoSt07] A. Hotho, G. Stumme: Textkategorisierung (Text classification).

http://www.kde.cs.uni-kassel.de/lehre/ws2007-08/IR/folien/14_TextCategorization.pdf (Zugriff: 27.05.2008), 2007

[Kaas03] M. Kaase: Meinungsforschung.

http://www.bpb.de/wissen/02613286908526470897531845992175,0,0,Meinungsforschung.html (Zugriff: 26.10.2007), 2003

[Kais97] A.Kaiser: Volltextinvertierung als sonderfall der automatischen Indexierung.

http://wwwai.wu-wien.ac.at/~kaiser/diss/node22.html (Zugriff: 26.12.2007), 1997

[Kim07] D. Kim: Semantische Analyse und automatische Gewinnung von branchenspezifischem Vokabular für E-Commerce.

http://edoc.ub.uni-muenchen.de/8420/1/Kim_Daewoo.pdf (Zugriff: 15.05.2008), 2007

[Kump06] D. Kumpe: Methoden zur automatischen Indexierung von Dokumenten

http://cis.cs.tuberlin.de/Dokumente/Diplomarbeiten/2006/kumpe.pdf (Zugriff: 15.03.2008), 2006

[Lewa05] D. Lewandowski: Web Information Retrieval – Technologien zur Informationssuche im Internet.

http://www.durchdenken.de/lewandowski/web-ir/download/Web-IR-Buch.pdf (Zugriff: 18.05.2008), 2005

[Lezi06] W. Lezius: Morphy

http://www.wolfganglezius.de/doku.php?id=public:cl:morphy (Zugriff: 01.05.2008), 2006

[Lore96] O. Lorenz: Automatische Wortformenerkennung für das Deutsche im Rahmen von Malaga.

http://www.linguistik.uni-erlangen.de/~orlorenz/DMM/DMM.html (Zugriff: 01.05.2008), 1996

[Lucko.J.] H. D.Luckhardt: Virtuelles Handbuch Informationswissenschaft. Automatische und intellektuelle Indexierung.

http://is.uni-sb.de/studium/handbuch/exkurs.ind.html (Zugriff: 18.04.2008), o.J.

[Meye07] Meyers Lexikonverlag: Meinungsforschung.

http://lexikon.meyers.de/index.php?title=Meinungsforschung&oldid=185309 (Zugirff: 05.03.2008) 2007

[Mitt05] A. Mittelmann: Wissensmanagement Methoden/Werkzeuge.

http://www.artm-friends.at/am/km/WM-Methoden/WM-Methoden-98.htm (Zugriff: 15.01.2007), 2005

[Nent00] T. Nentwich.: Data Mining – Umfeld , Prozeß, Methoden.

http://wwwai.wu-wien.ac.at/~koch/lehre/inf-sem-ws-00/nentwich/mining.pdf (Zugriff: 21.11.2007), 2000

[Nits04] D. Nitsche: UNICO-WebRoboter: Konzept einer spezialisierten Suchmaschine.

http://deposit.ddb.de/cgi-bin/dokserv?idn=974429910&dok_var=d1&dok_ext=pdf&filename=974429910.pdf (Zugriff: 18.05.2008), 2004

[Nohro.J.] H. Nohr: Theorie des Information Retrieval II: Automatische Indexierung.

http://www.inf-wiss.uni-konstanz.de/People/RK/Zulassung/b08-nohr-END.pdf (Zugriff: 24.10.2007), o.J.

[o.V.05] o. V.: Text Mining Conference Brochure 2005.

http://www.textminingnews.com/ (Zugriff: 05.03.2008), 2005

[o.V.08a] o.V.: Was ist Marktforschung.

http://www.vmoe.at/show_content2.php?s2id=1 (Zugriff: 01.05.2008), 2008a

[o.V.08b] o. V.: a Step-Automatische inhaltserschliessung(Statistische Verfahren).

http://www.bui.haw-hamburg.de/pers/ursula.schulz/astep/le6_step_2.html (Zugriff: 02.04.2008), 2008b

[o.V.o.J.a] o. V.: Empirische Sozialforschung / Methodenlehre.

http://www.uni-trier.de/uni/fb4/soziologie/faecher/empirik/header_main.html (Zugriff: 28.08.2007), o.J.

[o.V.o.J.b] o. V.: Der Prozess des Data Mining.

http://wissensexploration.de/datamining-kdd-prozess.php (Zugriff: 28.11.2007), o.J.

[o.V.o.J.c] o. V.: Markt- und Meinungsforschung.

http://www.fachverbandwerbung.at/de-brancheninfos-marktforschung.shtml (Zugriff: 13.01.2008), o.J.

[o.V.o.J.d] o. V.: Der Soundex-Algorithmus.

http://www.sound-ex.de/index.html (Zugriff: 02.04.2008), o.J.

[o.V.o.J.e] o.V.: STATISTICA Data Miner – Software für den Erfolg.

http://www.statsoft.de/pro_text_miner.html (Zugriff: 01.05.2008), o.J.

[Serb06] W. Serber: Kann man automatisch klassifizieren? Probleme und Ansätze automatischer Klassifikation.

http://www.dmv2006.uni-bonn.de/minisymposien/29/vortraege/060919_Bonn_DMV_Classification.ppt (Zugriff: 27.05.2008), 2006

[SiGv02] E. Sinanovic, S. Gvozden: Data Mining.

http://www.dke.univie.ac.at/extern/bi_ws20012002/ss2002/DATA-MINING.pdf (Zugriff: 23.11.2007), 2002

[Strao.J.] C. Strang: Methoden zur Indexierung von News-Feeds und deren Verteilung anhand benutzerspezifischer Interessen.

http://www.gm.fh-koeln.de/~bbertels/bachelorarbeit-christian-strang.pdf (Zugriff: 15.03.2008), o.J.

[Tres07] M. Tressl: Tagging, die andere indexierung des Internet.

http://www.contentmanager.de/magazin/artikel_1617_tagging.html (Zugriff: 07.03.2008), 2007

[Uszk01] H. Uszkoreit: Repräsentationen und Prozesse in der Sprachverarbeitung.

http://www.coli.uni-saarland.de/~hansu/Verarbeitung.html (Zugriff: 12.05.2008), 2000-2001

[Väth00] C. Väth: Besprechungs- & Verhandlungsmanagement im Bauwesen.

http://www.ibl.uni-stuttgart.de/03studium/html/diplomarbeiten/aushaenge/vaeth/ (Zugriff: 22.01.2008), 2000

[Weis00] M.Weiss: Automatische Indexierung mit besondere Berücksichtigung deutschsprachiger Text.

http://wwwai.wu-wien.ac.at/~koch/lehre/inf-sem-ws-00/weiss/index.html (Zugriff: 22.01.2008), 2000