METHODEN UND FORSCHUNGSLOGIK
DER KOMMUNIKATIONSWISSENSCHAFT

Andreas Niekler

Automatisierte Verfahren für die Themenanalyse nachrichtenorientierter Textquellen

HERBERT VON HALEM VERLAG

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

Andreas Niekler
Automatisierte Verfahren für die Themenanalyse nachrichtenorientierter Textquellen
Methoden und Forschungslogik der Kommunikationswissenschaft, 13
Köln: Halem, 2018

Andreas Niekler, geb. 21.7.1979, Dr. Ing., ist wissenschaftlicher Mitarbeiter an der Universität Leipzig. Derzeit arbeitet er an der Dynamik semantischer Kontexte in diachronen Korpora und der Extraktion von Aktivitäts- und Prozessinformationen aus großen Dokumentenquellen. Andreas Niekler studierte Medientechnik an der HTWK Leipzig und der University of West Scotland. Nach zwei Jahren als freier Programmierer und Dozent wechselte er an die Universität Leipzig. In seinem Dissertationsprojekt arbeitete er mit automatisierten Methoden zur Inhalts- und Themenanalyse.

Die Reihe *Methoden und Forschungslogik der Kommunikationswissenschaft* wird herausgegeben von Prof. Dr. Werner Wirth.

ISSN 1863-4966

ISBN (Print): 978-3-86962-261-3
ISBN (PDF): 978-3-86962-262-0

Den Herbert von Halem Verlag erreichen Sie auch im Internet unter http://www.halem-verlag.de
E-Mail: info@halem-verlag.de

SATZ: Herbert von Halem Verlag
LEKTORAT: Imke Hirschmann, Rabea Wolf
DRUCK: docupoint GmbH, Magdeburg
GESTALTUNG: Claudia Ott Grafischer Entwurf, Düsseldorf

Inhalt

1. EINLEITUNG

Die Benutzung digitaler Textarchive und -quellen erlaubt Zugriff auf darin manifestiertes Wissen und Erkenntnis. In zunehmendem Maße werden Archive, Bibliotheken und redaktionelle Inhalte digital verfügbar. Die maschinelle Verarbeitung von digitalem Text mithilfe der Informatik erleichtert und optimiert die Arbeit mit digitalen Textarchiven. Allerdings ist die analytische Arbeit mit digitalen Quellen ungleich komplizierter als die digitale Archivierung und Suche. Zum einen sind Volltexte nicht immer verfügbar und zum anderen ist das enthaltene Wissen meist nicht vorstrukturiert. Die Inhalte sind nicht effizient verfügbar. Die Analyse von Texten ist aber wichtiges und grundlegendes Arbeitsmittel vieler Tätigkeiten im Bereich des Journalismus und der Wissenschaft, insbesondere der Soziologie und der Kommunikationswissenschaft, weil viele Fachrichtungen bei der Arbeit auf Inhaltsanalysen setzen, um mit dem in Texten und anderen Medien enthaltenen Wissen, Antworten auf ihre Fragestellungen zu finden. Der Inhaltsanalyse geht es um die Erhebung empirischer Daten mittels einer strukturierten und offengelegten Suchstrategie, die aus materialisierter Kommunikation gewonnen werden können (vgl. FRÜH 2007: 147). In dieser Arbeit wird untersucht, welche Potenziale und Werkzeuge aus den Bereichen des Data- und Text-Minings für die inhaltsanalytische Arbeit in Textdatenbanken hilfreich und gewinnbringend eingesetzt werden können. Dabei konzentriert sich die Arbeit auf die Themenanalyse, einen Teilbereich der Inhaltsanalyse, welcher in Kapitel 2 konkretisiert und zur Definition grundlegender Anforderungen führt. Die Darstellung der Potenziale automatisierter Themenuntersuchungen in großen digitalen Textkollektionen in dieser Arbeit leistet dabei einen Beitrag zur Erforschung der automatisierten Inhaltsanalyse. Dabei setzt die Inhaltsanalyse einer-

seits in größeren repräsentativen Mengen von Inhalten auf die quantitative Analyse manifester Messgrößen, wie beispielsweise eines bestimmten Vokabulars. Der Gesamtzusammenhang der Inhalte steht im Hintergrund, da lediglich einzelne Messgrößen bewertet werden. Andererseits kann ein qualitativer Zugang in wenigen Dokumenten erfolgen, um vorliegende Inhalte durch einen detaillierten Zugriff auf alle inhaltlichen Zusammenhänge zu verstehen. Bei der analytischen Arbeit in Textarchiven gibt es verschiedene Dimensionen, die für die Analysten eine Rolle spielen. So gliedern sich Textsammlungen in verschiedene Dokumente auf, die unterschiedliche thematische Bezüge haben und zu verschiedenen Zeiten oder an unterschiedlichen Orten veröffentlicht werden. Beispielsweise kann die Veränderung von Inhalten gezeigt werden, wenn Inhalte veröffentlichter Dokumente unterschiedlicher Zeiträume analysiert werden. Vorstellbar ist die Beobachtung der lokalen Unterschiede, wenn die Inhalte zwischen verschiedenen Regionen oder Ländern verglichen werden sollen. Diese inhaltsanalytische Aufgabe steht im Gegensatz zu reinen Retrieval-Aufgaben, also der gezielten Suche nach definierten Inhalten. Durch eine umfassende wissenschaftliche Diskussion und die damit verbundene Entwicklung der Inhaltsanalyse ist eine mächtige Analysemethode zur Bearbeitung von Fragestellungen entstanden. Anwendungen der Inhaltsanalyse lassen sich in der Kommunikations- und Medienwissenschaft, im Journalismus, den Geschichtswissenschaften, der Betriebswirtschaft, den Politikwissenschaften, dem Marketing u. v. a. aufzeigen. Die Fragestellungen reichen dabei von Themenanalysen oder Argumentationsanalysen und dem Wissensmanagement bis zur Auswertung der Kundenzufriedenheit eines Unternehmens.

Eine besondere Rolle für verschiedene Fragestellungen spielen dabei Inhalte, die aus einem redaktionellen Kontext stammen, wie Nachrichtenartikel aus Tageszeitungen und Nachrichtenmagazinen. Diese Inhalte referenzieren auf eine angenommene Realität, die in der medialen Öffentlichkeit journalistisch interpretiert und kommentiert wird. Solche Quellen enthalten demnach Ereignisse, Diskurse, Themen oder Personen, die zu gegebenen Zeiten in einer sozialen Öffentlichkeit stattfinden. Diese Inhalte können durch eine Inhaltsanalyse zugänglich gemacht werden und beispielsweise für die Beurteilung gesellschaftlicher Prozesse herangezogen werden. So ist es besonders diese Textsorte, welche manifestiertes Wissen über längere Zeit und tagesaktuell sichtbar macht. Aus diesem Grund spielt die Untersuchung geeigneter Analysemethoden für diese Textsorte eine zentrale Rolle in der vorliegenden Arbeit.

Bei der Analyse von Nachrichten, Kommentaren oder anderen Meldungen sind verschiedene Herangehensweisen von Bedeutung, die zu unterschiedlichen Lösungen und Problemen führen. Grundsätzlich können Inhaltsanalysen textuell vorliegender Nachrichten retrospektiv oder prospektiv untersucht werden. Bei einer retrospektiven Analyse werden alle zur Verfügung stehenden Materialien untersucht, die für die Generierung oder die Überprüfung von Fragestellungen von Belang sind. Das bedeutet insbesondere, dass in Nachrichtenarchiven die Daten selektiert werden, die bereits vorliegen und geeignet für die Überprüfung der Fragestellung sind. Im Gegensatz dazu sind prospektive Studien begleitend und das auszuwertende Material muss erst erhoben werden. Beispielsweise wäre die begleitende Analyse der Themenstruktur von Tageszeitungen eine prospektive Analyse, da ständig neue Daten erhoben werden müssen. In diesem Zusammenhang müssen bereits vorhandene Daten aus bestehenden Textkollektionen selektiert werden und neue Daten ständig in die Kollektion eingebracht und für die weitere Verarbeitung und Selektion indiziert werden.

Da Kollektionen mit Nachrichtenartikeln sehr große Mengen an Texten enthalten, muss bei der Selektion auf technische Hilfsmittel zurückgegriffen werden, die innerhalb der Textdaten eine sinnvolle Selektion oder Erweiterung erlauben. Als weit verbreitetes Hilfsmittel wird hier die Volltextsuche und Indizierung genutzt, um einen Zugang zu ermöglichen, der über den Inhalt der Texte funktioniert. Weiterhin können, falls vorhanden, Metadaten einzelner Dokumente genutzt werden, um Selektionen über Ressorts, Personen oder Publikationen zu ermöglichen. Die valide Selektion relevanter Inhalte und eine nachfolgende Analyse ist mit reinen Volltextdaten durchaus möglich, aber mit Problemen behaftet. Durch die reine Selektion anhand von definierten Schlüsselwörtern und der damit verbundenen Ambiguitäten kann nicht sicher entschieden werden, ob eine so gefundene Menge von Dokumenten wirklich zur Fragestellung passt oder andere Zusammenhänge enthält. Der Volltextindex bildet als reine Such- und Indizierungsfunktion nicht ab, wie sich beispielsweise Thematisierungen inhaltlich ändern. Werden innerhalb einer Thematisierung Akteure oder Referenzen auf Ereignisse verändert, so läuft eine reine Volltextsuche Gefahr, diese Änderungen bei der Selektion der Untersuchungsmenge oder Grundgesamtheit nicht zu berücksichtigen und unvollständig zu sein. Aus diesem Grund muss die analytische Arbeit in großen Textkollektionen weitere Methoden berücksichtigen, die

- inhaltliche Strukturen analysieren,
- inhaltsanalytische Fragestellungen abbilden können,
- zeitliche Zusammenhänge bilden können,
- referenzierte Orte, Personen oder Institutionen integrieren und
- mit der Textmenge umgehen können.

Sollen unbekannte bzw. neue Phänomene oder Zusammenhänge, bei denen noch keine Theoriebildung erfolgt ist, untersucht werden, so ist oft nicht klar, wie diese in den Dokumenten identifiziert werden können. Durch diese Unklarheit ist es nicht möglich, geeignete Texte zu selektieren. Die Auswahl von Schlüsselwörtern fällt schwer, da die Bildung von Hypothesen und einer Theorie noch nicht abgeschlossen ist. Aus diesem Grund muss es möglich sein, die Inhalte nach verschiedenen Kriterien zu explorieren. Strukturen müssen sichtbar gemacht werden, die nicht im Voraus definiert sind. Dies dient der Theoriebildung und die Analysten können sich mit dem Datenbestand und dessen Inhalt vertraut machen. Innerhalb von Nachrichtenmeldungen sind dabei Themenstrukturen von Interesse, da diese das öffentliche Interesse und das öffentliche Geschehen abbilden. Im Gegensatz dazu werden Diskurse und Argumentationen in den textuellen Meldungen und Kommentaren mitgeführt, die oft für Inhaltsanalysen die eigentliche Aussage über ein bestimmtes Phänomen liefern. So führt Früh aus, dass die Analyse von Diskursen und Argumenten hohe Anforderungen an eine Inhaltsanalyse stellt und mehrere Indikatoren innerhalb der Texte untersucht werden müssen (vgl. FRÜH 2007: 85). Durch diese Komplexität ist es umso wichtiger, dass die Auswahl der zu analysierenden Dokumente aus einer Textkollektion sehr genau und vollständig passiert, da diese Vorarbeit eine wichtige Voraussetzung für die Qualität der Analyse ist. Deshalb ist es wichtig, Möglichkeiten der Exploration zu haben, um Relevantes von nicht Relevantem zu trennen. Dies kann beispielsweise die thematische Einschränkung einer Dokumentmenge sein. Aus diesem Grund müssen für hypothesengeleitete Analysen in großen Textkollektionen Methoden zur Verfügung stehen, die Textstrukturen oder Themenstrukturen erkennen und anhand derer weitere Analysen und Dokumentmengen vorbereitet werden können.

Der analytische Umgang mit Daten wird unter dem Begriff ›Data-Mining‹ zusammengefasst. Data-Mining ist eine Disziplin, bei der große Datenmengen zur Wissensgewinnung durch Algorithmen analysiert werden. Dabei konzentriert sich Data-Mining auf numerische Datenbankdaten, wie beispielsweise den Verkaufszahlen in einer Region. Im Data-Mining

werden solche Rohdaten genutzt, um Muster, Trends oder Abweichungen zu messen. Die Methoden des Data-Minings generieren demnach Ergebnisse, die angewandt auf die analytische Arbeit in Textarchiven eine Arbeitsoptimierung versprechen. Die Übertragung der Data-Mining-Methoden auf Textdatenbanken wird unter dem Begriff ›Text-Mining‹ zusammengefasst (vgl. HEYER 2006: 4). Unter Zunahme der Methoden der Verarbeitung natürlicher Sprache aus der Informatik ergibt sich hier ein wissenschaftliches Untersuchungsfeld, das bereits enormes Potenzial entwickelt hat. Dennoch stützt sich die inhaltliche Arbeit und die Forschung in Textarchiven, seien sie digital oder nicht, oft auf den manuellen Umgang mit den Dokumenten. Dabei wird versucht, einen repräsentativen Querschnitt der Grundgesamtheit aller Dokumente in einem Archiv manuell zu untersuchen. Die Unterstützung durch die elektronische Datenverarbeitung beschränkt sich oft auf Such- und Verwaltungsfunktionen der Dokumente. Die Analyse selbst wird mit einer konkreten Operationalisierung meist manuell durchgeführt. Oft ist es für die zu untersuchenden Textmengen aus ökonomischen Gründen nicht möglich, mehrere Operationalisierungen oder Untersuchungsmethoden zu erproben. Dennoch gibt es hierfür in der Informatik Ansätze, die vielversprechende Ergebnisse liefern, um die digitalen Inhalte für Untersuchungen in verschiedenen Disziplinen zugänglich zu machen und die manuelle Arbeit mit Dokumenten zu entlasten.[1]

1.1 Ausgangslage

Die Analyse von Textdatenbanken kann unter zwei verschiedenen Bedingungen erfolgen. Sind neue Themen oder Änderungen an einer Textdatenbank bei einer prospektiven Analyse signifikant, so sollen sie gemessen und beurteilt werden können. Sowohl bei prospektiven und retrospektiven Untersuchungen kann bereits bestehender Inhalt mit einem zeitlichen Bezug betrachtet und analysiert werden, um die Dynamik des in den Texten enthaltenen Wissens zu untersuchen. Untersuchungen die einen Zeitpunkt

1 In Scharkow (2012) werden Methoden des Data-Minings und des maschinellen Lernens für die Anwendung in der Inhaltsanalyse untersucht. Weiterhin zeigen die Aufsatzsammlungen in West (2001) und Sommer (2014) das Potenzial von computergestützten Ansätzen.

mit einem anderen vergleichen, werden ›diachron‹ genannt und stehen im Gegensatz zu Untersuchungen, die nicht auf zeitliche Bezüge beschränkt sind. In zwei Beispielen soll die praktische Relevanz dieser Unterscheidungen für unterschiedliche Disziplinen erläutert werden.

In der Kommunikations- und Medienwissenschaft, genauer in der empirischen Kommunikationsforschung, stellt die Beurteilung der Berichterstattung über ein Thema eine besondere Aufgabe dar. In verschiedenen Quellen werden Belegstellen für ein Thema gesucht, um zu messen, wie sich das Thema in einer bestimmten zeitlichen Abfolge in dem untersuchten Medium verhält. Dabei stellt die zeitliche Einordnung einen wesentlichen Bestandteil der Untersuchung dar, da mitunter herausgefunden werden soll, welche Ereignisse und Entwicklungen Einfluss auf die Berichterstattung oder die Themenaufmerksamkeit hatten. Inhaltsanalysen mit einem Fokus auf Thematisierungen in textuellen Nachrichtenmedien sind demnach diachrone Textanalysen. So kann beispielsweise die nachhaltige Wirkung von Ereignissen auf Themen durch retrospektive Analysen in Nachrichtentexten gemessen werden. Auf der anderen Seite können mit prospektiven Analysen Trends für Themen abgeleitet werden, ohne die Ereignisse oder Einflussfaktoren zu analysieren, die dafür verantwortlich sind.

Ein weiteres Beispiel stellt die Analyseanforderungen im journalistischen Umfeld heraus. Eine Redaktion befindet sich immer im Spannungsfeld zwischen aktuellen Informationen, eigenen Inhalten, der öffentlichen Meinung und der ständigen Beobachtung der Medienrezipienten. Durch die mittlerweile etablierte Beteiligung der Rezipienten am Inhalt, sei es in Form von Rückkanälen bzw. Interaktivität oder von eigenen Beiträgen, können diese direkt und unmittelbar beobachtet werden. Um das Umfeld, die Wirkungen und den Erfolg einer Redaktion zu verstehen, müssen ständige Medienresonanzanalysen durchgeführt werden. Redaktionen arbeiten mit rechnergestützten Systemen, weshalb heute ein Großteil der Inhalte digital vorliegt. Wichtig für die Arbeit der Redaktionen ist in diesem Szenario, dass Trends in der Berichterstattung oder aufkommende Themen rechtzeitig aufgegriffen werden können und ein Überblick über die Inhalte bewahrt werden kann. Das Widerspiegeln der sogenannten ›Medienagenda‹ und der Erwartungen der Rezipienten ist in diesem Fall ein ständiger Prozess, der darauf angewiesen ist, aktuelle Entwicklungen in den Textdatenbanken, hier also in den Nachrichtentexten, hervorzuheben und sichtbar zu machen.

Zusammenfassend kann gesagt werden, dass die Themenanalyse einen wichtigen Bestandteil fast aller Inhaltsanalysen darstellt, um die Themen

selbst eingehender zu untersuchen oder um die Textanalysen an einem bestimmten Thema zu orientieren. Dabei spielt es eine Rolle, ob die analysierte Textmenge ständig wächst, also veränderlich ist, und prospektiv oder als abgeschlossene Menge retrospektiv untersucht wird. Zusätzlich können bei geeigneten Textquellen zeitliche Unterschiede untersucht werden. Existiert in einer Textsammlung keine diachrone Unterscheidung, beispielsweise über Zeitstempel einzelner Dokumente, so ist die Textmenge nur als Ganzes zu untersuchen und die Untersuchung von temporären Trends ist nicht möglich.

1.2 Problemstellung und Ziele

Die vorliegende Arbeit untersucht, wie Verfahren des Text-Minings die inhaltlich-thematische Auswertung in Textarchiven unterstützt und welchen Anforderungen diese Analyse genügen muss. Dabei konzentrieren sich die Untersuchungen auf publizierte Inhalte von Redaktionen (redaktioneller Inhalt), wie Nachrichtenartikel aus Tageszeitungen, Journalen und Magazinen. In den Inhalten der Quellen müssen Themen erkannt und über längere Zeiträume beobachtet werden. Die diachrone Analyse und die nichtdiachrone Analyse werden im Zusammenhang mit verschiedenen Textsorten und maschinellen Lernverfahren untersucht. Dabei ist die Eignung der Verfahren für prospektive und retrospektive Analysen über eine geeignete Repräsentation und Verarbeitung der Daten herstellbar. Es sollen Verfahren verwendet und evaluiert werden, welche die Relevanz, Aktivität, Kontexte und Veränderungen von Themen über die Zeit verfolgen. Besonders aufschlussreich sind dabei die Vorgänge und Merkmale, die das öffentliche Interesse eines Themas begleiten. In einer exemplarischen Untersuchung wird deshalb untersucht, wie diese Veränderungen am sprachlichen Kontext und der Häufigkeit eines Themas beobachtet werden können.

Die Aufnahme bestimmter Inhalte und die Relevanz von Artikeln spiegelt die Medienöffentlichkeit wider. An der Analyse und an Rückschlüssen auf reale Phänomene anhand der Medienöffentlichkeit sind viele Disziplinen interessiert. Ob ein Thema inhaltlich interessant ist, wird von verschiedenen Faktoren bestimmt. Dazu gehören die sogenannten ›Nachrichtenfaktoren‹ und die Verfügbarkeit der Informationen. Die Daten, welche mit in dieser Arbeit untersuchten Verfahren erzeugt werden, sollen solche Faktoren bestimmbar und bewertbar machen. Aus der Gesamtlage der so

entstehenden Daten können Redaktionen oder Forschergruppen Aussagen und Interpretation über Thematisierungen erstellen. Dies erhöht die Effektivität und die Qualität einer Analyse von Textdaten.

Die Möglichkeit, Kontextveränderungen zu bestimmen, die zu einem erhöhten Informationsinteresse der Öffentlichkeit führen, muss bei einer Untersuchung der Thematisierungen vorhanden sein. Demnach ergeben sich für die Arbeit Anforderungen wie

- die Identifikation von Verfahren, mit denen themenbasierte Inhaltsanalysen möglich sind,
- die Evaluation und Beurteilung dieser Verfahren,
- die Identifikation von Merkmalen eines Themas, die durch die Verfahren messbar werden wie
  - die Beobachtung über die Zeit,
  - die Beobachtung des Themenanteils am Textarchiv und
  - die Beobachtung von weiteren Informationen in einem Thema, wie beispielsweise Nennungen von Personen und
- die Identifikation von Wortkontexten innerhalb eines Themas, die zur Beurteilung desselben Themas beitragen.

1.3 Aufbau der Arbeit

In einem Grundlagenteil, der sich in Kapitel 2 anschließt, wird zunächst erläutert, welche Anforderungen beachtet werden müssen, um methodisch korrekte Inhaltsanalysen durchzuführen. Diese Einführung dient der Vorstellung der Methoden der Inhaltsanalyse, welche in den Kommunikations- und Sozialwissenschaften etabliert sind. Diese Einführung soll eine Schnittstelle zur kommunikationswissenschaftlichen Anwendung der Inhaltsanalyse herstellen. Weiterhin werden in diesem Kapitel Methoden für die computergestützte Verarbeitung von Textdokumenten vorgestellt, die eine Grundlage für die automatische Durchführung von Inhaltsanalysen darstellen. Aus den Ausführungen in Kapitel 2 ergeben sich damit Anforderungen und Fragestellungen für die Umsetzung automatisierter Inhalts- bzw. Themenanalysen. In Abschnitt 2.4 werden die Forschungsfragen und Anforderungen der Arbeit noch einmal hinsichtlich der Grundlagen konkretisiert.

Kapitel 3 spricht verschiedene Verfahren an, die sich für die Themenanalyse in elektronischen Dokumenten eignen. Die Arbeit konzentriert sich

dabei auf die Verwendung von Verfahren, die ohne Training und Vorwissen arbeiten können. Die manuelle Erstellung von Trainingsdaten ist aufwendig. Da die Inhalte der nachrichtenorientierten Quellen sehr veränderlich sind, ist eine Anpassung der Trainingsdaten in kurzen Abständen nötig.

In Kapitel 4 werden die untersuchten Verfahren in exemplarischen Analysen implementiert, erprobt und evaluiert. Dabei werden sowohl die Analyse inhaltlicher Aspekte als auch die Auswertung von Zeitreihen durchgeführt. Die Evaluierung der Verfahren spielt eine wichtige Rolle, um deren Eignung und Anwendbarkeit für Inhaltsanalysen zu zeigen. In Kapitel 5 schließt sich eine Diskussion der Ergebnisse und Erkenntnisse an. Dabei werden konkrete Anforderungen und Forschungsfragen, die hinsichtlich der methodischen Grundlagen der Inhaltsanalyse in Abschnitt 2.4 aufgestellt werden, diskutiert und beantwortet.