Zusammenfassung

Thema der Dissertation: **Automatisierte Verfahren für die Themenanalyse nachrichtenorientierter Textquellen**

Verfasser: Dipl. Ing. Andreas Niekler

Im Bereich der medienwissenschaftlichen Inhaltsanalyse, stellt die Themenanalyse einen wichtigen Bestandteil dar. Für die Analyse großer digitaler Textbestände hinsichtlich thematischer Strukturen ist es deshalb wichtig, das Potential automatisierter computergestützter Methoden zu untersuchen. Dabei müssen die methodischen und analytischen Anforderungen der Inhaltsanalyse beachtet und abgebildet werden, welche auch für die Themenanalyse gelten. In dieser Arbeit werden die Möglichkeiten der Automatisierung der Themenanalyse und deren Anwendungsperspektiven untersucht. Dabei wird auf theoretische und methodische Grundlagen der Inhaltsanalyse und auf linguistische Theorien zu Themenstrukturen zurückgegriffen, um Anforderungen an eine automatische Analyse abzuleiten. Den wesentlichen Beitrag stellt dabei die Untersuchung der Potentiale und Werkzeuge aus den Bereichen des Data- und Text-Mining dar, die für die inhaltsanalytische Arbeit in Textdatenbanken hilfreich und gewinnbringend eingesetzt werden können. Weiterhin wird eine exemplarische Analyse durchgeführt, um die Anwendbarkeit automatischer Methoden für Themenanalysen, im methodischen Kontext der Inhaltsanalyse, zu zeigen. Die Arbeit demonstriert dabei auch Möglichkeiten der Nutzung interaktiver Oberflächen, formuliert die Idee und Umsetzung einer geeigneten Software und zeigt die Anwendung eines möglichen Arbeitsablaufs für die Themenanalyse auf. Die Darstellung der Potentiale automatisierter Themenuntersuchungen in großen digitalen Textkollektionen in dieser Arbeit leistet dabei einen Beitrag zur Erforschung der automatisierten Inhaltsanalyse.

Ausgehend von den Anforderungen, die an eine Themenanalyse gestellt werden, zeigt diese Arbeit, mit welchen Methoden und Automatismen des Text-Mining diesen Anforderungen nahe gekommen werden kann. Dabei sind zusammenfassend zwei Anforderungen herauszuheben, deren jeweilige Erfüllung die andere beeinflusst. Zum einen ist eine schnelle thematische Erfassung der Themen in einer komplexen Dokumentsammlung gefordert, um deren inhaltliche Struktur abzubilden und um Themen kontrastieren zu können. Zum anderen müssen die Themen in einem ausreichenden Detailgrad abbildbar sein, sodass eine Analyse des Sinns und der Bedeutung der Themeninhalte möglich ist. Beide Ansätze haben eine methodische Verankerung in den quantitativen und qualitativen Ansätzen der Inhaltsanalyse. Die Arbeit diskutiert diese Parallelen und setzt automatische Verfahren und Algorithmen mit den Anforderungen in Beziehung. Es können Methoden aufgezeigt werden, die

eine semantische und damit thematische Trennung der Daten erlauben und einen abstrahierten Überblick über große Dokumentmengen schaffen. Dies sind Verfahren wie Topic-Modelle oder clusternde Verfahren. Mit Hilfe dieser Algorithmen ist es möglich, thematisch kohärente Untermengen in Dokumentkollektion zu erzeugen und deren thematischen Gehalt für Zusammenfassungen bereitzustellen. Es kann gezeigt werden, dass die Themen trotz der distanzierten Betrachtung unterscheidbar sind und deren Häufigkeiten und Verteilungen in einer Textkollektion diachron dargestellt werden können. Diese Aufbereitung der Daten erlaubt die Analyse von thematischen Trends oder die Selektion bestimmter thematischer Aspekte aus einer Fülle von Dokumenten. Diachrone Betrachtungen thematisch kohärenter Dokumentmengen werden dadurch möglich und die temporären Häufigkeiten von Themen können analysiert werden. Für die detaillierte Interpretation und Zusammenfassung von Themen müssen weitere Darstellungen und Informationen aus den Inhalten zu den Themen erstellt werden. Es kann gezeigt werden, dass Bedeutungen, Aussagen und Kontexte über eine Kookkurrenzanalyse im Themenkontext stehender Dokumente sichtbar gemacht werden können. In einer Anwendungsform, welche die Leserichtung und Wortarten beachtet, können häufig auftretende Wortfolgen oder Aussagen innerhalb einer Thematisierung statistisch erfasst werden. Die so generierten Phrasen können zur Definition von Kategorien eingesetzt werden oder mit anderen Themen, Publikationen oder theoretischen Annahmen kontrastiert werden. Zudem sind diachrone Analysen einzelner Wörter, von Wortgruppen oder von Eigennamen in einem Thema geeignet, um Themenphasen, Schlüsselbegriffe oder Nachrichtenfaktoren zu identifizieren. Die so gewonnenen Informationen können mit einer detaillierten Analyse thematisch relevanter Dokumente ergänzt werden, was durch die thematische Trennung der Dokumentmengen möglich ist. Über diese methodischen Perspektiven lassen sich die automatisierten Analysen auch als empirische Messinstrumente im Kontext weiterer hier nicht besprochener kommunikationswissenschaftlicher Theorien einsetzen. Des Weiteren zeigt die Arbeit, dass grafische Oberflächen und Software-Frameworks für die Bearbeitung von automatisierten Themenanalysen realisierbar und praktikabel einsetzbar sind. Insofern zeigen die Ausführungen ebenfalls, wie die besprochenen Lösungen und Ansätze in die Praxis überführt werden können.

Wesentliche Beiträge liefert die Arbeit für die Erforschung der automatisierten Inhaltsanalyse. Dabei dokumentiert die Arbeit vor allem die wissenschaftliche Auseinandersetzung mit automatisierten Themenanalysen. Während der Arbeit an diesem Thema wurden geeignete Vorgehensweisen entwickelt, wie Verfahren des Text-Mining in der Praxis für Inhaltsanalysen einzusetzen sind. Unter anderem wurden Beiträge zur Visualisierung und einfachen Benutzung unterschiedlicher Verfahren geleistet. Verfahren aus dem Bereich des Topic Modelling, des Clustering und der Kookkurrenzanalyse mussten angepasst werden, sodass deren Anwendung in inhaltsanalytischen Anwendungen möglich ist. Weitere Beiträge entstanden im Rahmen der methodologischen Einordnung der

computergestützten Themenanalyse und in der Definition innovativer Anwendungen in diesem Bereich. Für die vorliegende Arbeit durchgeführte Experimente und Untersuchungen wurden komplett in einer eigens entwickelten Software durchgeführt, die auch in anderen Projekten erfolgreich eingesetzt wird. Um dieses System herum wurden Verarbeitungsketten, Datenhaltung, Visualisierung, grafische Oberflächen, Möglichkeiten der Dateninteraktion, maschinelle Lernverfahren und Komponenten für das Dokumentretrieval implementiert. Dadurch werden die komplexen Methoden und Verfahren für die automatische Themenanalyse einfach anwendbar und sind für künftige Projekte und Analysen benutzerfreundlich verfügbar. Sozialwissenschaftler, Politikwissenschaftler oder Kommunikationswissenschaftler können mit der Softwareumgebung arbeiten und Inhaltsanalysen durchführen, ohne die Details der Automatisierung und der Computerunterstützung durchdringen zu müssen.