

Dominique Vincent

Evolution von Relationen in
temporalen partiten Themen-Graphen



Scientific
Publishing

Dominique Vincent

EVOLUTION VON RELATIONEN IN
TEMPORALEN PARTITEN THEMEN-GRAPHEN

Evolution von Relationen in temporalen partiten Themen-Graphen

von
Dominique Vincent

Dissertation, Karlsruher Institut für Technologie (KIT)
Fakultät für Wirtschaftswissenschaften, 2014

Tag der mündlichen Prüfung: 18. Dezember 2014

Referent: Prof. Dr. Wolfgang Gaul

Korreferent: Prof. Dr. Karl-Heinz Waldmann

Impressum



Karlsruher Institut für Technologie (KIT)
KIT Scientific Publishing
Straße am Forum 2
D-76131 Karlsruhe

KIT Scientific Publishing is a registered trademark of Karlsruhe
Institute of Technology. Reprint using the book cover is not allowed.

www.ksp.kit.edu



*This document – excluding the cover – is licensed under the
Creative Commons Attribution-Share Alike 3.0 DE License
(CC BY-SA 3.0 DE): <http://creativecommons.org/licenses/by-sa/3.0/de/>*



*The cover page is licensed under the Creative Commons
Attribution-No Derivatives 3.0 DE License (CC BY-ND 3.0 DE):
<http://creativecommons.org/licenses/by-nd/3.0/de/>*

Print on Demand 2015

ISBN 978-3-7315-0340-8

DOI 10.5445/KSP/1000045521

Evolution von Relationen in temporalen partiten Themen-Graphen

Zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften

(Dr.-Ing.)

von der Fakultät für
Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von

Dipl.-Wi.-Ing. Dominique Vincent

Tag der mündlichen Prüfung: 18. Dezember 2014
Referent: Prof. Dr. Wolfgang Gaul
Korreferent: Prof. Dr. Karl-Heinz Waldmann
(2014) Karlsruhe

Danksagung

Mein besonderer Dank gilt meinem Doktorvater, Herrn Prof. Dr. Wolfgang Gaul, für die Betreuung der Arbeit. In zahlreichen Arbeitsgesprächen unterstützte er den Fortschritt meiner Forschungstätigkeiten mit seinen fachlichen Ratschlägen, weiterführenden Anregungen und kreativen Ideen.

Weiterhin danke ich Herrn Prof. Dr. Karl-Heinz Waldmann für die Übernahme des Korreferats. Bei Herrn Prof. Dr. Martin Klarmann und Herrn Prof. Dr. Philipp Reiss bedanke ich mich, dass sie sich als Prüfer bzw. als Vorsitzender an meiner mündlichen Promotionsprüfung beteiligt haben.

Meinen ehemaligen Kollegen am Institut für Entscheidungstheorie und Unternehmensforschung danke ich für die vielen Fachgespräche, Ratschläge und die freundschaftliche Atmosphäre, dies gilt insbesondere für Herrn Dr. Dominic Gastes, Frau Dr. Rebecca Klages und Herrn Dr. Christoph Winkler sowie Frau Bayer und Frau Nickel vom Sekretariat.

Weiterer Dank gilt dem gesamten Team des Instituts für Informationswirtschaft und Marketing, Forschergruppe Marketing & Vertrieb, für die Unterstützung während der letzten Phase meiner Promotion.

Ganz besonders danke ich meinen Eltern für die jahrelange Unterstützung während des Studiums und der Promotion. Sie standen mir stets mit Rat und Tat zur Seite.

Karlsruhe im Dezember 2014
Dominique Vincent

Inhaltsverzeichnis

1	Einleitung	1
1.1	Motivation und Zielsetzung der Arbeit	1
1.2	Aufbau der Arbeit	3
2	Stand der Wissenschaft	5
3	Modellgrundlagen	17
3.1	Notationen	17
3.1.1	Referenzkorpus	17
3.1.2	Lokales Wörterbuch & Referenzwerte	18
3.1.3	Merkmalsextraktion & Vector Space Modell	18
3.1.4	Akkuratesse des Lokalen Wörterbuchs	20
3.2	(Un-) Ähnlichkeitsmaße	21
3.3	Clusteranalyse zur Themendetektion	24
3.3.1	Gütekriterium	25
3.3.2	Partitionierende Verfahren	25
3.3.3	Hierarchische Verfahren	27
3.3.4	Clustering-Kardinalität	30
4	Relationen in temporalen Themen-Graphen	33
4.1	Relationen & Schranken	33
4.2	Multi-Themen-Graph	35
4.3	Mono-Themen-Graph	36
4.4	Strukturelle Eigenschaften in Themen-Graphen	37
4.5	Evolution von temporalen Themen-Graphen	41
5	Evaluation	45
5.1	Implementierung in MATLAB und MSSQL	45
5.2	Rechner-Konfigurationen	46
5.3	Testdaten	47
5.3.1	IDS-Mannheim Datensatz	48
5.3.2	Spiegel Online (SPON) Datensatz	51
5.4	Testreihen	54
5.4.1	Testreihe I.	54
5.4.2	Testreihe II. (GfKl & DAGM & IFCS (2011))	58
5.4.3	Testreihe III. (GfKl & SFC (2013))	64
5.4.4	Testreihe IV.	81
6	Zusammenfassung und Ausblick	109

A Zusatz zu 5.4.4	113
Abbildungsverzeichnis	123
Tabellenverzeichnis	125
Literaturverzeichnis	127

Kapitel 1

Einleitung

1.1 Motivation und Zielsetzung der Arbeit

Die Fülle an Informationen, die heutzutage in digitaler Form verfügbar ist, macht eine manuelle Sichtung und Bewertung so gut wie unmöglich. Dies wird besonders deutlich durch die phänomenale Entwicklung des Internets in den letzten dreißig Jahren. Im Internet veranschlagte man schon in 2008 einen Zuwachs von ca. einer Million neuer Dokumente pro Tag (vgl. HEYER ET AL. (2008)). Während in den Anfangsjahren (1984) erst rund Tausend Rechner miteinander verbunden waren, erreichte das Internet im Jahr 2013 bereits 2,7 Milliarden Menschen weltweit (vgl. Abbildung 1.1). Die Zahl der Webseiten wuchs von 130 im Jahr 1993 auf über 634 Millionen im Jahr 2012. Der globale Informationsbedarf spiegelt sich auch in der Zahl der Suchanfragen wider, die von 9800 Anfragen pro Tag im Jahr 1998, beim Marktführer Google, auf inzwischen (Stand 2012) über 3 Milliarden täglich (1,2 Billionen pro Jahr) explodiert sind.

Diese Informationsflut wird sich in naher Zukunft noch verstärken. Die Fähigkeit, wichtige Schlüsselinformationen effektiv und effizient aus dem Informationsstrom zu filtern, ist daher eine entscheidende Kernkompetenz, um im informationsintensiven Umfeld zu bestehen (vgl. HUANG (2003), WAGNER ET AL. (2009)). Relevante Themen zu identifizieren und ihre Entstehungsgeschichte sowie Entwicklung im Zeitverlauf zu verfolgen, ist eine wichtige Voraussetzung für kompetente Entscheidungen z.B. in Politik, Wirtschaft und Gesellschaft. Aktuell aufkommende Themen können völlig neue Sachverhalte behandeln, oder auch von Ereignissen oder Entwicklungen aus der Vergangenheit beeinflusst sein. Für eine große Zahl anstehender Entscheidungen ist es von besonderer Wichtigkeit, die Beziehungen eines Themas zu früheren Themen oder Ereignissen zu kennen und analysieren zu können. So können Erfahrungen, Chancen und Risiken aus früheren Entscheidungen bzw. Entwicklungen bei der Beurteilung eines Sachverhalts und der Entscheidungsfindung eine wertvolle Hilfe sein. Damit wird eine sachdienliche und verantwortungsvolle Steuerung künftiger Entwicklungen unterstützt.

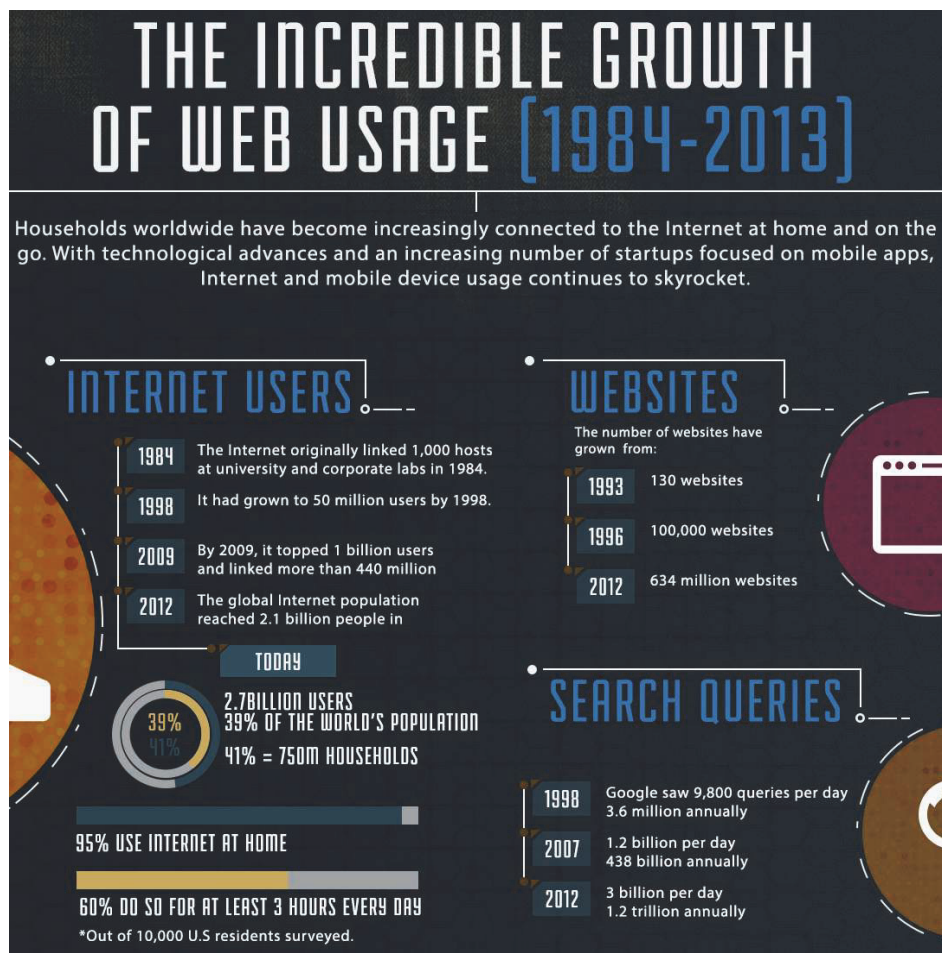


Abbildung 1.1: Entwicklung des Internet (Quelle: WhoIsHostingThis.com), abgerufen Sept. 2014.

Selbstverständlich wurde bereits in früheren Zeiten auf zurückliegende Erfahrungswerte bei einer Entscheidungsfindung aufgebaut. Jedoch waren die anstehenden Aufgaben und die dafür zugänglichen Informationen längst nicht so komplex und umfangreich wie in der heutigen global vernetzten Welt.

Die manuelle Sichtung und Bewertung war zwar anspruchsvoll aber möglich. Im Internetzeitalter sind Informationen global und aktuell für jedermann verfügbar. Dies erfordert allein wegen der Fülle an Informationen eine automatisierte effektive und effiziente Aufbereitung und Darstellung für den Nutzer.

Ziel dieser Arbeit ist es, ein automatisiertes System zu entwerfen und zu implementieren, dass einem interessierten Nutzer aggregierte und komprimierte Informationen zu relevanten Themen, unter Einbeziehung ihrer zeitlichen Entwicklungen, übersichtlich zugänglich macht.

1.2 Aufbau der Arbeit

Die vorliegende Arbeit gliedert sich in sechs Kapitel, die in Abbildung 1.2 graphisch dargestellt sind.

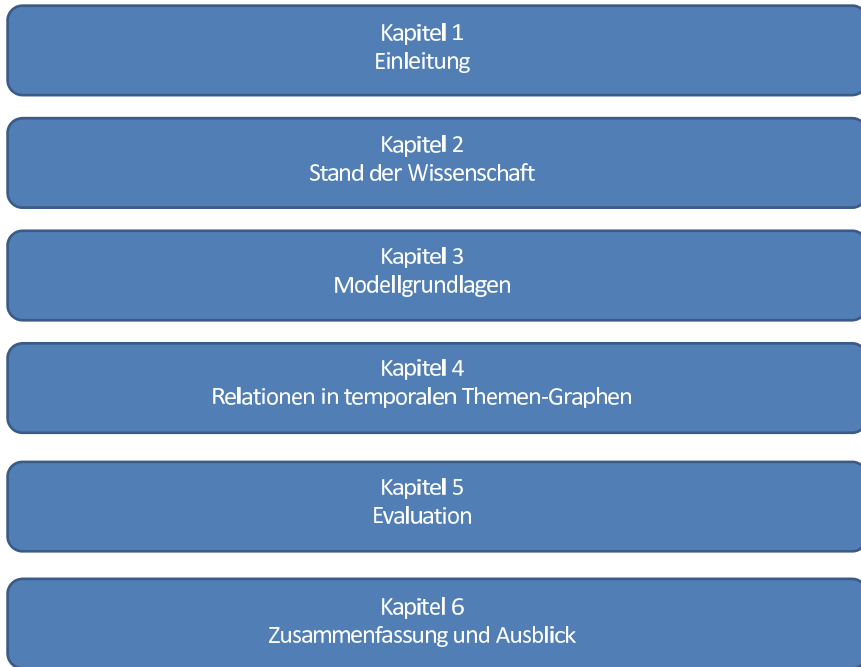


Abbildung 1.2: Gliederung der Arbeit.

In Kapitel 2 wird das dieser Arbeit zugrunde liegende Forschungsgebiet näher beleuchtet und Entwicklungen der letzten Jahre überblicksmäßig aufgezeigt.

In Kapitel 3 werden die für die Erstellung eines Modells nötigen formalen wie auch algorithmischen Grundlagen erläutert. Hierzu werden Grundbegriffe, Basiswerkzeuge, Verfahren und Algorithmen vorgestellt, die in der Literatur zur Verfügung stehen und für die vorliegende Arbeit angepasst werden.

Auf der Basis der vorgestellten Modellgrundlagen wird in Kapitel 4 ein Modell entwickelt, das erlaubt, korrelierte Themen in ihrer zeitlichen Evolution zu verfolgen und zu analysieren.

In Kapitel 5 wird das vorgestellte Modell anhand eines implementierten Prototypen überprüft. Es kommen zwei unterschiedliche Datensätze zur Anwendung. Die bei variierenden Test- und Parameterkonfigurationen erzeugten Testergebnisse werden in Grafiken visualisiert und erläutert.

Abschließend wird in Kapitel 6 eine Zusammenfassung gegeben und Erweiterungsmöglichkeiten des Modells bzw. der Implementierung werden aufgezeigt.

Kapitel 2

Stand der Wissenschaft

Der Forschungsschwerpunkt dieser Arbeit ist dem Forschungsgebiet 'Topic Detection and Tracking' (TDT) zuzuordnen, welches um die Jahrtausendwende durch gezielte Forschungsförderung seitens amerikanischer Interessensgruppen initiiert und finanziert wurde. Dabei spielte die 'Defense Advanced Research Projects Agency' (DARPA), sowie die 'National Science Foundation' (NSF) bei der Finanzierung eine maßgebliche Rolle (vgl. WAYNE (1998)). Neben diesen beiden Organisationen waren auch das 'National Institute of Standards and Technology' (NIST) als Standardisierungsinstitution (vgl. WAYNE (1998)), sowie das 'Linguistic Data Consortium' (LDC) zur Organisation bestimmter Testkorpora (vgl. WALLS ET AL. (1999)), beteiligt.

Das Forschungsfeld des TDT bezieht sich auf automatisierte Verfahren und Techniken zum Auffinden in thematischer Relation stehenden Materials innerhalb von Datenströmen. Dies wird unter anderem von der 'National Security Agency' (NSA) als interessante Herausforderung angesehen (vgl. WAYNE (1998)). ALLAN ET AL. (1998) fassen technische Herangehensweisen sowie Ergebnisse eines ersten 'Topic Detection and Tracking' Workshops, abgehalten an der University of Maryland im Jahr 1997, in einer Pilotstudie zusammen. Die Studie klärt, ob 'Topic Detection und Tracking' überhaupt technisch möglich ist und welche Probleme zu lösen sind. Sie grenzt den Begriff Topic (Themengebiet) von einem zweiten Begriff Event (Ereignis) ab. Danach ist ein Event ein einmaliges Ereignis, wie z.B. der Ausbruch des Mount Pinatubo am 15. Juni 1991, während Vulkanausbrüche im allgemeinen als Klasse von Ereignissen definiert werden. Topics (Themengebiete) beinhalten also mehrere verwandte Ereignisse, die ein gemeinsames Thema zum Gegenstand haben. Die Studie macht deutlich, dass, bis auf sporadische Ereignisse, die retrospektive Detektion von Ereignissen durch Clusterverfahren verlässlich möglich ist. TDT konzentriert sich auf fünf Aufgaben bzw. Tätigkeitsbereiche. Die 'Story Segmentation' hat zum Ziel, ein Transkript von Nachrichten in einen Strom individueller Nachrichtentexte aufzuteilen. Die 'First Story Detection' behandelt das Problem der Erkennung eines neuen Themas innerhalb eines Stroms von Nachrichten. Die 'Cluster Detection' gruppiert alle ankommenden Nachrichtendokumente nach Themenschwerpunkten. Das 'Tracking', also das kontinuierliche Beobachten des Nachrichtenstroms, dient der Auffindung zusätzlicher

Dokumente eines schon bekannten Themas. Die letzte der fünf Aufgabenstellungen des TDT ist die 'Story Link Detection', durch die entschieden wird, ob zwei gewählte Dokumente demselben Thema angehören (vgl. ALLAN ET AL. (2002), ALLAN (2002a)).

Im Folgenden werden wichtige Entwicklungen innerhalb des Forschungsgebiets anhand ausgewählter Beiträge vorgestellt.

OARD (1999) zeigt die Machbarkeit eines 'Topic Tracking' Systems. Dazu wird ein frei verfügbares 'Vector Space Text Retrieval' System der NIST, genannt 'Prototype Indexing and Search Engines' (PRISE), verwendet. Eine vereinfachte Version des Rocchio-Algorithmus zur Profilbildung eines Topics erzielt in Kombination mit einem 'Threshold' (Schranke) und einem harten temporalen Cutoff gute Ergebnisse.

WALLS ET AL. (1999) nutzen einen Incremental k -means Algorithmus, um Dokumente zu gruppieren. Der Vergleich von Dokumenten wird mit einem probabilistischen Ähnlichkeitsmaß und einer traditionellen 'Vector Space' Metrik durchgeführt. Das Inkrementelle k -means Verfahren verarbeitet Dokumente nacheinander in sequentieller Form. Dabei werden zwei Schritte durchlaufen: Erstens der 'Selection' Schritt, das heißt die Auswahl des Clusters, das dem Nachrichtentext am ähnlichsten ist. Zweitens der Schritt 'Thresholding', der auf Grund eines 'Threshold'-Wertes das Dokument mit dem Cluster vergleicht und entscheidet, ob das Dokument mit diesem Cluster vereinigt wird, also einem bestehenden Thema zugeordnet werden kann, oder ob das Dokument der Beginn eines neuen Themas ist. Für die Schritte 'Selection' und 'Thresholding' wenden WALLS ET AL. (1999) unterschiedliche Metriken an. 'Selection' nutzt eine sogenannte 'BBN Topic Spotting Metric' der US-Firma 'Raytheon BBN Technologies', während für das 'Thresholding' Problem ein Hybrid aus 'BBN Topic Spotting Metric' und einer konventionellen Cosinus Distanzmetrik eingesetzt wird.

Der Detection Algorithmus von ALLAN ET AL. (2000) unterstützt zwei Verfahren, das Centroid-Verfahren mit einem agglomerativen Clustering und einem 'Nearest Neighbor' Vergleich basierend auf dem bekannten k -NN Verfahren. Beide Verfahren setzen auf einen 'Threshold' beim Ähnlichkeitsvergleich, bei dessen Überschreiten ein ankommendes Dokument einem schon bekannten Thema zugeordnet wird, oder im Falle eines Unterschreitens dieses Dokument als Start eines neuen Clusters (Themas) dient. Als Ähnlichkeitsfunktion im 'Vector Space' dient unter anderem das Cosinusmaß. Dieses wird im vorgestellten TDT System ebenfalls für das Tracking angewendet.

RAJARAMAN/TAN (2001) wenden zur Dokumentenrepräsentation das weit verbreitete 'Vector Space' Modell an, kombiniert mit einem 'Self-Organizing Neural Network'. Als Neuronales Netz wird das 'Adaptive Resonance Theory' (ART) Netzwerk von CARPENTER ET AL. (1991) angewendet. Jedes Thema in einem solchen Netzwerk wird durch einen Knoten repräsentiert. Mittels Dokumentenvektoren und einem

Fuzzy Learning Algorithmus wird das Neuronale Netz trainiert und die vorliegende Clusterstruktur für einen folgenden Tracking Schritt gespeichert. Ankommende Dokumente können mit einem solchen trainierten Neuronalen Netz einem bestehenden Thema zugeordnet werden. Das Plotten der Dokumentenanzahl pro Thema über die Zeit hinweg ermöglicht das Verfolgen der Evolution eines Themas.

BARON/SPILIOPOULOU (2001) untersuchen 'Web Usage Data', um aufkommende Trends zu bestimmen. Dazu nutzen sie Veränderungen im Surfverhalten von Webnutzern. Durch eine Assoziationsanalyse können Regeln und Muster abgeleitet werden, wobei Änderungen der Regeln und Muster im Zeitverlauf auf eine Evolution des Nutzerverhaltens schließen lassen. Verschiedene Muster von Änderungen einer Assoziationsregel sind möglich und können zur Bestimmung von Trends herangezogen werden. Dabei ergibt sich die fundamentale Frage, ab wann eine Veränderung in den gemessenen Daten und somit einer Assoziationsregel als Mutation, bzw. als neue Assoziationsregel angesehen werden kann, die die alte Regel ersetzt.

In einer weiteren Studie vergleicht ALLAN (2002b) die Aufgabenstellung des 'Topic Tracking' mit dem Problem der 'Topic Detection'. Im Unterschied zum Verfolgen (Tracking) von Dokumenten eines interessierenden Themas ist das Ziel der 'Topic Detection' das Gruppieren aller ankommenden Dokumente in Themen, ungeachtet dessen, ob diese von Interesse sind oder nicht. Außerdem sollen neue unbekannte Themen, die vorher noch nicht antizipiert wurden, erkannt werden. Dabei kann 'Topic Detection' als eine Art paralleles Multi-Themen Tracking angesehen werden. Unterschiede zwischen den beiden Aufgabenbereichen können deutlich gemacht und Vor- und Nachteile gegeneinander abgewogen werden.

KLEINBERG (2003) nimmt an, dass ein Topic in einem Dokumentenstrom durch einen Ausbruch von Aktivität signalisiert wird, wobei bestimmte Merkmale scharf in ihrer Frequenz ansteigen, während das Topic erscheint. Um solche Ausbrüche zuverlässig zu bestimmen modelliert KLEINBERG (2003) den Strom mit Hilfe eines 'Finite-State' Automaten, in dem Ausbrüche als natürliche Zustandsübergänge dargestellt werden. Zur Messung eines Ausbruchs werden für jedes Wort innerhalb der Kollektion alle Häufungen im Nachrichtenstrom berechnet. Durch Berechnung eines Gewichts für jeden Ausbruch ist es möglich, diese nach ihrer Intensität zu ordnen. Dies ermöglicht ein Auffinden von Worten, die die stärksten auf- und abfallenden Muster für eine bestimmte Zeitperiode zeigen. Das System kann beispielweise für 'Web Usage Data', wie zum Beispiel 'Clickstreams' und 'Search Engine Query Logs', angewendet werden. Ausbrüche zeigen eine erhöhte kollektive Aufmerksamkeit für ein bestimmtes Ereignis oder Thema an.

Die im Folgenden zitierten Arbeiten nutzen neben dem Veröffentlichungsdatum zusätzliche temporale Informationen.

Eine Arbeit, die (neben Dokumentenähnlichkeiten) temporale Informationen innerhalb der Dokumente verwendet, findet sich z.B. bei PONS-PORRATA ET AL. (2002).

Um die Effektivität bei der Lösung von TDT Aufgaben zu verbessern, extrahieren KIM/MYAENG (2004) temporale Informationen aus Nachrichtentexten zusätzlich zu den bekannten Zeitstempeln (Datum der Veröffentlichung) der Dokumente. Dazu nutzen sie einen 'Finite-State' Automaten und ein Lexikon mit zeitbezogenen Termen. Die extrahierten temporalen Informationen werden in ein vordefiniertes Format umgewandelt, um Zeitpunkte bzw. Zeiträume wiederzugeben. Durch Nutzung dieser temporalen Informationen in den Dokumenten kann die Effektivität zeitsensitiver Klassifikation von Dokumenten in vordefinierte Ereigniskategorien gesteigert werden.

Laut LI ET AL. (2006) sind temporale Informationen wichtige Attribute eines Themas, wobei ein Topic normalerweise nur eine gewisse Zeitspanne existiert. Viele Forscher haben die Möglichkeit untersucht, temporale Informationen für die TDT Aufgaben zu nutzen. Dabei konzentrieren sie sich meist entweder auf das Datum der Veröffentlichung oder auf temporale Ausdrücke in Texten, um temporale Zusammenhänge abzuleiten. Dabei wird außer Acht gelassen, dass Menschen dazu neigen, zeitliche Angaben zu Ereignissen in unterschiedlicher Detailgenauigkeit zu formulieren, wenn die Zeit fortschreitet. Daher schlagen LI ET AL. (2006) eine neue Strategie, ein sogenanntes 'Time Granularity Reasoning' vor, um temporale Informationen für das Topic Tracking zu nutzen. Die Topic-Time eines Themas wird durch alle Zeitangaben der diesem Thema zugehörigen Dokumente definiert. Der temporale Zusammenhang zwischen einem Dokument und einem Topic wird bestimmt durch den höchsten Koreferenz-Grad zwischen allen Zeitangaben eines Dokuments und dem betrachteten Topic. Hierbei wird der Koreferenz-Grad zwischen einem Dokument und einem Topic abgeleitet von einer betrachteten Zeitangabe innerhalb eines Dokuments und einer Topic-Time, deren Granularität, sowie dem Zeitabstand zwischen der Topic-Time und dem Veröffentlichungsdatum des Textes selbst.

TU/SENG (2012) nutzen, um neue Forschungsfelder in wissenschaftlichen Datenbanken zu identifizieren, neben Termfrequenzen von Forschungs-Stichworten einen Neuigkeits-Index. Neuigkeits-Indizes basieren z.B. auf Veröffentlichungsdatum und Volumen-Nr. (Ausgabe.Nr.) einer wissenschaftlichen Zeitschrift.

Im Verlauf der Weiterentwicklung des Forschungsgebiets ergeben sich eine Reihe interessanter Aufgabenstellungen. Aus der Vielfalt der Arbeiten mit unterschiedlichen Schwerpunkten werden im Folgenden einige Beispiele vorgestellt.

FENG/ALLAN (2005) untersuchen die Problematik der Granularität von Topic-Hierarchien. Nachrichten-Topics können in unterschiedlicher Größe beschrieben werden, dabei ist die 'korrekte' Granularität schwer zu definieren. In hierarchischen

Strukturen entsprechen nicht alle Einheiten einer Hierarchie der Definition eines Topics. Die Übergänge zwischen Topics, Sub-topics bzw. Ereignissen bis hin zu den individuellen Dokumenten sind fließend. Eine Einführung einer Ereignisstruktur kann zu einem besseren Verständnis von Themen führen.

Das Topic Tracking System von BUN/ISHIZUKA (2006) beobachtet interessante Webseiten. Diese werden durch vordefinierte Keywords über eine kommerzielle Suchmaschine bestimmt. Die den Schlüsselworten ähnlichsten Webseiten bilden das Informationsareal. Ein 'Web Spider' scannt anschließend regelmäßig diese Webseiten auf Änderungen bzw. neu hinzukommende Artikel. Für diese Änderungen werden die einzelnen Termgewichte bestimmt. Die Sätze mit der höchsten mittleren Gewichtung dienen der Beschreibung und Zusammenfassung von Änderungen der beobachteten Topics.

Zum Auffinden von neu entstehenden und beständigen Themen fokussieren sich SCHULT/SPILIOPOULOU (2006) auf die Identifikation von Cluster-Labeln. Die gefundenen Cluster-Label müssen Änderungen der Dokumentenpopulation bei einer mit der Zeit anwachsenden Dokumentensammlung gewisse Zeitperioden 'überleben', um als Themen angesehen zu werden. Auch Änderungen im Merkmalsraum dominanter Wortformen, die sich auf Grund einer über die Zeit ändernden Terminologie im Dokumentenarchiv ergeben, müssen solche Cluster-Label überstehen.

MEI ET AL. (2006) suchen nach Subtopics von Weblogs und analysieren sie in Bezug auf raum-zeitliche Muster über einen probabilistischen Ansatz (vgl. dazu u.a. HOFMANN (1999), BLEI ET AL. (2003)). Weblogs bestehen aus einer Mischung an Subthemen, die mit einem raum-zeitlichen Muster ausgesendet werden. Das von MEI ET AL. (2006) entwickelte 'Probabilistic Mixture Model' erklärt die Erzeugung von Themen, sowie deren raum-zeitliches Muster. Dazu werden in einem ersten Schritt geläufige Themen der Blogs extrahiert, in einem zweiten Schritt werden Themen-Lebenszyklen für jeden Schauplatz erstellt, in einem dritten Schritt werden Momentaufnahmen eines Themas für jede Zeitperiode generiert. Durch eine vergleichende Analyse der Themen-Lebenszyklen und Momentaufnahmen der Themen können Evolutionsmuster aufgedeckt werden. Dazu nutzen MEI ET AL. (2006) Dokumente mit Time-Stamp sowie Ortsangaben. Zur Modellierung von semantisch kohärenten Themen und Subthemen werden Wahrscheinlichkeitsverteilungen von Wortformen genutzt.

Die Arbeit von HE ET AL. (2007) analysiert Wortform-Trajektorien bezüglich der Dimensionen Zeit und Frequenz. Mit bekannten Techniken aus der Signalverarbeitung identifiziert sie Korrelationen zwischen Merkmalen durch Spektralanalyse. Eine Menge von Wörtern mit identischen Trends kann gruppiert werden, um so ein Ereignis zu rekonstruieren. Die repräsentativen Wörter des gleichen Ereignisses zeigen dieselbe Entwicklung über die Zeit und sind hoch korreliert.

FUKUMOTO/SUZUKI (2007) behandeln das Problem unausgewogener Trainingsdaten im Topic Tracking: Einer großen Anzahl Dokumente, deren Themen nicht getrackt werden sollen, steht eine kleine Anzahl Dokumente eines zu trackenden Themas gegenüber. FUKUMOTO/SUZUKI (2007) schlagen eine neue Methode zur effektiven Generierung passender Trainingsdokumente vor. Für die kleine Anzahl positiv gelabelter Dokumente werden bilinguale Vergleiche von z.B. englischen und japanischen Korpora durchgeführt. Mittels eines zweisprachigen Wörterbuchs wird beispielsweise ein kleiner englischsprachiger Trainingsdatensatz um Trainingsdokumente aus dem japanischsprachigen Raum erweitert. FUKUMOTO/SUZUKI (2007) nehmen an, dass viele Nachrichtenquellen eines Landes mit einer höheren Frequenz und detaillierter über lokale Ereignisse berichten als Nachrichtenquellen weit entfernter ausländischer Nachrichtenstationen, selbst wenn die Ereignisse internationale Beachtung erlangt haben. Weitere Ausführungen zum Vergleich bilingualer Textkorpora finden sich zum Beispiel in DAGAN/CHURCH (1997), COLLIER ET AL. (1998) und UTSURO ET AL. (2003).

LANDMANN/ZUELL (2008) identifizieren Events innerhalb von Event-Texten anhand sogenannter Eventwörter. Für alle Wortformen innerhalb eines Event-Text-Korpus werden relative Termfrequenzen berechnet und mit relativen Termfrequenzen eines Referenzkorpus verglichen. Wörter des Eventkorpus mit den größten Abweichungen zu den Termfrequenzen der Wortformen aus dem Referenzkorpus gelten als Eventwörter und werden einer explorativen Faktoranalyse unterzogen.

KHY ET AL. (2008) clustern Dokumente auf Grund ihrer Ähnlichkeit und Neuigkeit (Alter), wobei jüngeren Dokumenten ein höheres Gewicht zugewiesen wird als älteren Dokumenten. Durch eine Erweiterung des konventionellen Cosinusmaßes im 'Vector Space' um eine 'Forgetting Function', wird der Fokus auf neuere Dokumente gelegt. Bei der Vergessensfunktion nimmt das initiale Gewicht Eins eines Dokuments zum Erscheinungszeitpunkt exponentiell ab. Ein sogenannter 'Forgetting Factor' bestimmt die Stärke des exponentiellen Verfalls. Durch Kombination des definierten neuigkeitsbasierten Ähnlichkeitsmaßes mit einer Variation des k -means Algorithmus werden sukzessiv alte Dokumente eliminiert und der Fokus beim Clustern auf aktuelle Dokumente gelegt. Somit werden eher Themen-Cluster gebildet, die aktuelle Themen sowie Trends behandeln.

LI/CROFT (2008) stellen einen Ansatz zur 'Novelty Detection' vor, der auf der Satzebene Informationsmuster verwendet um auf Suchanfragen von Nutzern spezifische Antworten geben zu können. Suchanfragen oder Themen werden anhand von Wortmustern automatisch klassifiziert in spezifische und allgemeine Themen. Suchanfragen-bezogene Informationsmuster, namentlich Satzlängen, Kombinationen von Namen (Personen, Orte, Organisationen, etc.) und Muster von Meinungen werden identifiziert und untersucht. Dabei werden für spezifische Topics bessere Ergebnisse erzielt als für allgemeine.

MATHIOUDAKIS/KOUDAS (2010) nutzen zur Trend Detektion den Kurznachrichtendienst Twitter. Jeder Tweet eines Twitter Nutzers ist durch ein Nutzerprofil charakterisiert. Somit sind jedem Tweet wertvolle Metadaten zugeordnet, wie zum Beispiel persönliche Informationen des Twitterers, Name, Ort, biographische Details des Nutzers, etc. MATHIOUDAKIS/KOUDAS (2010) identifizieren Keywords, die in einer außergewöhnlichen Häufung in den erscheinenden Tweets (Kurzmeldungen) auftreten. Solche aufkommenden Schlüsselwörter werden, basierend auf ihren Konkurrenzen, in Trends gruppiert. Neben den Keywords der Tweets zur Beschreibung gefundener Trends werden Verlinkungen zu weiterführenden Artikeln innerhalb der Tweets der Nutzer berücksichtigt. Solche Verlinkungen können beispielsweise auf professionelle Nachrichtenportale wie Reuters, New York Times, etc. verweisen. Aufkommende Ereignisse, Eilmeldungen, Sondermeldungen und allgemeine Themen, die die Aufmerksamkeit der Nutzer auf sich ziehen und über Twitter verbreitet werden, sind somit Grundlage für die Trendbestimmung. Auch die geographische Verortung von Twittermeldungen, die einem Trend angehören, wird identifiziert, so dass stark assoziierte Verortungen zu bestimmten Trends erkannt werden können. JIN ET AL. (2007) bedienen sich ebenfalls aus Dokumenten extrahierter Ortsangaben für ein 'Event Tracking'. Ein weiterer Ansatz zur Auswertung von Twittermeldungen findet sich bei BENHARDUS (2010).

Aus der breiten Palette der Fragestellungen lassen sich auch miteinander verbundene Problemstellungen finden, die teilweise mit zeitlichem Abstand bearbeitet werden. Ein Beispiel ist die Untersuchung synchroner und nicht synchroner Dokumentenströme.

Zur Bestimmung aufkommender Themenmuster werten WANG ET AL. (2007) multiple Dokumentenströme aus, die von unterschiedlichen Quellen ausgesendet werden. Dabei können auch Quellen verschiedener Sprachen berücksichtigt werden. Die untersuchten koordinierten Textströme unterschiedlicher Quellen müssen dabei Dokumentenmengen gleicher Zeiträume enthalten. Die aufkommenden Themenmuster werden über ein Topic Model textstromübergreifend bestimmt. Auch hier wird ein Thema als Wahrscheinlichkeitsverteilung von Wortformen eines Vokabulars beschrieben. Über Thresholds wird die Mindestdauer und Mindestintensität eines Themas festgelegt, um als aufkommendes Topic gemeldet zu werden. Entlang der Zeitlinie kann ein globales Themenmuster identifiziert werden. Auf Grund der Verortung der einzelnen Quellen können darüber hinaus lokale Themenmuster gefunden werden. Die Arbeit von WANG ET AL. (2007) untersucht zeitlich synchron verlaufende und somit koordinierte Dokumentenströme unterschiedlicher Quellen, um korrelierte Themenmuster zu entdecken.

Im Gegensatz dazu zielen WANG ET AL. (2009) darauf ab, semantisch ähnliche Themen unterschiedlicher Nachrichtenströme zu finden, die nicht zeitlich synchron verteilt sind, also in unterschiedlichen Quellen zu unterschiedlichen Zeiten auftreten. Die Korrelationen zwischen semantischen und temporalen Informationen in den

Nachrichtenströmen können genutzt werden, um Asynchronitäten zwischen den Strömen auszugleichen.

Ereignisepisodes, Themen-Transitionen und Lebensprofile sind die Schwerpunkte der folgenden Arbeiten:

WEI/CHANG (2007) suchen nach Ereignisepisodes anhand von Dokumentensequenzen. Häufig auftretende temporale Beziehungen zwischen den Ereignisepisodes dienen als Evolutionsmuster.

Dokumente sind nach Veröffentlichungsdatum geordnet und bilden Dokumentensequenzen. Dokumente eines Themas bilden innerhalb eines Zeitfensters eine sogenannte Intrasequenz-Episode eines Ereignisses. Es ist daher ausschlaggebend, Dokumente zu identifizieren, die dasselbe Ereignis beschreiben. Dazu werden die Dokumente als Dokumentenvektoren mit Termgewichten und Veröffentlichungsdatum dargestellt und geclustert. Als Repräsentanten gefundener Ereignisepisodes können die einem Ereignis zugeordneten Dokumentenvektoren genutzt werden. Die Dauer einer Ereignisepisode innerhalb einer Dokumentensequenz wird definiert durch die Veröffentlichungsdaten des ersten sowie letzten zugehörigen Dokuments einer Ereignisepisode. Ähnliche Intrasequenz Ereignisepisodes unterschiedlicher Dokumentensequenzen, also unterschiedlicher Zeitfenster, bilden Intersequenz Ereignisepisodes.

Das Analysieren von Themen-Transitionen in großen Dokumentenmengen ist laut ZENG/ZHANG (2007) mit den verbreiteten Topic Models schwer durchzuführen. Um solche Transitionen zu analysieren stellen ZENG/ZHANG (2007) das 'Variable Space Hidden Markov Model' (VSHMM) vor, eine Erweiterung des bekannten 'Hidden Markov Model' (HMM). Mit einer variablen Speicherzuordnung, die eine individuelle Repräsentation eines Dokuments je nach seiner Länge erlaubt, kann die Speicherallokation wie auch die Rechenzeit beim Ausführen des Modells verringert werden. Wortformen sind die sichtbaren Merkmale eines Dokuments, während Themen in dem Dokument abgeleitet werden müssen. Dabei repräsentieren Zustände (Hidden States) in dem 'Hidden Markov Model' die Themen, die durch eine Wahrscheinlichkeitsverteilung für alle Wortform beschrieben werden. Übergangswahrscheinlichkeiten zwischen den Hidden States bilden mögliche Transitionen der Themen in Dokumentensammlungen ab.

CHEN ET AL. (2009) modellieren Ereignisentwicklungen über ein Konzept von Lebensprofilen. Dazu nutzen sie ein 'Hidden Markov Model' (HMM) (vgl. MARKOV (2006)). Jedes Lebensprofil hat eine eindeutige charakteristische Aktivitätsentwicklung, die von Ereignissen mit ähnlichen Mustern abgeleitet werden kann. Die Ereignisentwicklung kann verschiedene Aktivitätszustände durchlaufen, wie zum Beispiel 'sehr aktiv', 'normal', 'inaktiv' etc. Zusammen mit charakteristischen Transitionen zwischen diesen Aktivitätszuständen können adäquate Lebensprofile der Ereignisse erstellt werden.

Eine Studie von KONTOSTATHIS ET AL. (2004) untersucht verschiedene weitere semi- und vollautomatisierte Ansätze zum 'Topic Detection and Tracking' ('Emerging Trend Detection'). KONTOSTATHIS ET AL. (2004) kommen zu dem Schluß, dass die vorgestellten Projekte dazu neigen, entweder auf reine Machine Learning Verfahren oder auf Visualisierungstechniken zu setzen. Isoliert genutzt erweisen sich beide Verfahrenstypen als inadäquat, in Kombination jedoch als erfolgversprechender.

Ansätze, die Machine Learning Verfahren mit Visualisierungen kombinieren, um Evolutionen von Themen beziehungsweise Relationen von Themen im Zeitverlauf zu bestimmen, werden im Folgenden näher beschrieben.

So stellen MEI/ZHAI (2005) eine probabilistische Methode vor, die latente Themen von Texten findet, evolutionäre Beziehungen zwischen diesen aufdeckt und einen Evolutionsgraphen der Themen erstellt. Dazu werden die Intensitäten der Themen über die Zeit modelliert und Lebenszyklen der Themen analysiert. In einer vorliegenden Dokumentensammlung wird ein semantisch kohärentes Thema als Wahrscheinlichkeitsverteilung von Wortformen modelliert. Dieses sogenannte Topic Model ist ein gebräuchliches Verfahren, vgl. dazu auch HOFMANN (1999), BLEI ET AL. (2003) und CROFT/LAFFERTY (2003). Aus Wortformen, die einen hohen Wahrscheinlichkeitswert in einer solchen Verteilung aufweisen, lässt sich oft der Gegenstand eines Themas erschließen. Die Themenspanne eines Themas wird definiert durch einen Start- und Endzeitpunkt. Themen mit einer Themenspanne über den kompletten Textstrom werden als Transkolektions-Themen bezeichnet. Evolutionäre Transitionen zwischen zwei Themen werden durch ihre Ähnlichkeit über die Kullback-Leibler Divergenz bestimmt. Eine evolutionäre Transition zwischen zwei Themen besteht dann, wenn die Ähnlichkeit zwischen ihnen eine gewisse Schranke überschreitet. Beide Themen müssen dabei zeitlich aufeinander folgen, das heißt, der Startzeitpunkt des späteren Themas muss nach dem Endzeitpunkt des vorangehenden Themas liegen. Eine Sequenz aufeinander folgender Themenspannen bilden eine Themen-Evolutionskette. Graphisch beschrieben werden solche temporalen Beziehungen durch einen gerichteten gewichteten Graphen, wobei dessen Knoten die Themen repräsentieren. Kanten beschreiben die evolutionäre Transition, die Gewichte der Kanten beschreiben die evolutionäre Distanz. Jeder Pfad innerhalb eines Evolutionsgraphen repräsentiert eine Themen-Evolutionskette. Für eine Menge von Transkolektions-Themen wird ein Themenlebenszyklus für jedes Thema als Stärkenverteilung des entsprechenden Themas über die komplette Zeitlinie definiert. Die Stärke eines Themas für jede Zeitperiode wird angegeben durch die Anzahl Worte, die durch dieses Thema in den Dokumenten dieser Zeitperiode generiert wurden. Durch Normalisieren dieser Anzahl Worte, entweder durch die Anzahl Zeitpunkte oder die totale Anzahl Worte in der Periode, erhält man die absolute oder relative Stärke. Die absolute Stärke misst die absolute Anzahl Texte, die ein Thema erklären kann, während die relative Stärke ein Indikator für die relative Stärke eines Themas innerhalb einer Zeitperiode darstellt.

Die Beschreibung möglicher Themenverschiebungen innerhalb von Transkolektions-Themen wird von MEI/ZHAI (2005) durch ein 'Hidden Markov Model' modelliert. Dadurch können für jedes Transkolektions-Thema sein Beginn, sein Ende und seine Änderungen über die Zeit hinweg analysiert werden.

Das Transition Tracking System von SPILIOPOULOU ET AL. (2006) legt den Fokus nicht auf topologische Eigenschaften der Cluster, sondern auf die Inhalte des zu Grunde liegenden Datenstroms. Zur Unterstützung strategischer Entscheidungen ist das Überwachen und Verstehen von Clustern und deren Änderungen über den Zeitverlauf eine entscheidende Aufgabe und stellt eine große Herausforderung dar. Während MEI/ZHAI (2005) Topic-Evolutionen in Textströmen über die Label der Topics betrachten, untersuchen SPILIOPOULOU ET AL. (2006) die Änderungen der Cluster selbst. Dieser Ansatz ist auch dann anwendbar, wenn kein Cluster-Label erstellt werden kann. Mittels einer Alterungsfunktion werden gleitende Zeitfenster abgebildet. Beobachtungen außerhalb dieses Fensters erhalten das Gewicht Null und sind somit irrelevant. Clusterzuordnungen werden durch Clusterüberlappungen unterschiedlicher Clusterings aus unterschiedlichen Zeitpunkten bestimmt. Cluster-Transitionen zugeordneter Cluster im Zeitverlauf können 'interner' Natur sein, d.h. den Inhalt wie auch die Form eines Clusters betreffen. Interne Transitionen beschreiben zum Beispiel die Größenveränderung eines Clusters, d.h. die Anzahl zugeordneter Objekte. Neben einer möglichen Verschiebung eines Clustermittelpunkts wird angegeben, ob ein Cluster kompakter oder diffuser wird. 'Externe Transitionen' beziehen sich auf die Beziehung eines Clusters zu den restlichen Clustern des Clusterings. Zu den externen Transitionen gehören zum Beispiel das Erscheinen eines neuen Clusters, das Überleben eines bestehenden Clusters, das Aufteilen eines Clusters in mehrere Cluster, das Absorbieren des Clusters durch ein anderes Cluster und das komplette Verschwinden eines Clusters.

WAGNER ET AL. (2009) bearbeiten Dokumente mit einem hierarchischen System, einer sogenannten 'Hierarchically Growing Hyperbolic Map' für große Datenmengen. Die oberste Hierarchieebene des Netzwerks wird mit dem kompletten Dokumentensatz trainiert, ähnlich wie eine traditionelle 'Self-Organizing-Map' (SOM) (vgl. KOHONEN (2001)). Nach dem Trainieren der obersten Ebene der Hierarchie des Netzwerks mit dem kompletten Dokumentensatz teilt die initiale Karte die Kollektion in eine Menge von Subcluster. Bestehende Knoten werden erweitert und die nächste Strukturebene wird trainiert, indem nach den am besten passenden Knoten entlang der bereits bestehenden Hierarchie gesucht wird. Diese Vorgehensweise teilt die Dokumentenmenge in Subcluster abnehmender Größe auf und ordnet sie auf dem hierarchischen Gitter der hyperbolischen Ebene an. Nachdem die Hierarchie ihre zuvor definierte maximale Tiefe erreicht hat, werden alle Knoten des Netzwerks anhand ihrer jeweiligen Prototypenvektoren gelabelt. Neu ankommende Informationen regen die am besten passenden Neuronen des Netzwerks an, ihre Aktivität steigt. Ein kontinuierlicher Informationsfluß erzeugt so eine Art dynamisches Muster der Nachrichtenaktivitäten.

Mit Momentaufnahmen aus einem solchen 'Film' können Einblicke in die bestehende Nachrichtenstruktur gewonnen werden.

YANG ET AL. (2009) identifizieren, ebenso wie MEI/ZHAI (2005), evolutionäre Beziehungen zwischen Nachrichtenereignissen. Sie erstellen dabei einen Evolutionsgraphen manuell über annotierte Nachrichtenereignisse. Modelliert wird der Evolutionsgraph als azyklischer Graph. Für einen Übergang von einem Ereignis A zu einem Ereignis B müssen drei Bedingungen erfüllt sein: A muss temporär vor B liegen, die Ereignisse A und B müssen ein ähnliches Vokabular besitzen und die beiden Ereignisse dürfen zeitlich nicht zu weit auseinander liegen, sonst sinkt die Wahrscheinlichkeit einer evolutionären Beziehung zwischen ihnen zu sehr. Von einem älteren Ereignis führt eine direkte gerichtete Kante zu einem in Relation stehenden jüngeren Ereignis. Ein Aufspalten von Ereignissen, genannt 'Event Threading' ist ebenso möglich wie Vereinigung von Ereignissen, genannt 'Event Joining'. Bei einem 'Event Threading' ist die Anzahl der ausgehenden Kanten größer als eins. Für ein 'Event Joining' ist die Anzahl eingehender Kanten größer als eins.

OLIVEIRA/GAMA (2010) stellen ein System zur Beobachtung von Cluster-Evolutionsprozessen vor. Ihr System beinhaltet eine Taxonomie für mögliche Transitionen, eine Tracking Methode auf Basis der Graphentheorie sowie einen Transitionsdetektions-Algorithmus. Transitionen sind Änderungen, die das ganze Clustering betreffen, wie z.B.: Birth, Death, Split, Merge und Survival von Clustern und werden über einen bipartiten Graphen bestimmt und klassifiziert. Das Konzept der Detektion basiert auf exakten Zuordnungen zwischen Clustern unterschiedlicher aufeinander folgender Zeitpunkte. Für jedes Paar möglicher Verbindungen zwischen Clustern unterschiedlicher Zeitpunkte werden bedingte Wahrscheinlichkeiten berechnet und im bipartiten Graphen als gewichtete Kanten dargestellt.

Die in diesem Kapitel vorgestellten Arbeiten stellen eine Auswahl an Veröffentlichungen aus dem Forschungsbereich des 'Topic Detection and Tracking' dar. Die dabei genannten Verfahren und Algorithmen werden, sofern sie für diese Arbeit von Relevanz sind, in Kapitel 3 näher beschrieben und erläutert.

Kapitel 3

Modellgrundlagen

Ziel der folgenden Modellüberlegungen ist es, ein kohärentes System zu erstellen, das erlaubt, korrelierte Themen, die in aufeinander folgenden Veröffentlichungen behandelt werden, in ihrer zeitlichen Evolution zu verfolgen und zu analysieren. Hierzu werden im Folgenden Grundbegriffe, Basiswerkzeuge, Verfahren und Algorithmen vorgestellt, die in der Literatur zur Verfügung stehen und für die vorliegende Arbeit angepasst werden.

3.1 Notationen

3.1.1 Referenzkorpus

Für die Bearbeitung der Dokumententexte wird zunächst ein Referenzkorpus \mathcal{R} benötigt. Das Referenzkorpus $\mathcal{R} = \{d_1, d_2, \dots, d_{|\mathcal{R}|}\}$ besteht aus einer Anzahl $|\mathcal{R}|$ von Dokumenten d_s , $s = 1, \dots, |\mathcal{R}|$. Ein Dokument $d_s = (w_{s_1}, \dots, w_{s_{R_s}})$ wiederum besteht aus R_s Wortformen w_s . Eine Wortform (Term) ist ein Wort beziehungsweise eine von diesem Wort flexivisch abgeleitete Form, zum Beispiel: gehen, gegangen, gingen etc. Auf eine Grundwortreduktion (auch Stemming genannt, vergleiche dazu auch PORTER (1997)), d.h. ein Zurückführen einer Wortform auf seinen lexikalischen Wortstamm, wird verzichtet, um insbesondere sprachliche Sonderheiten wie auch zeitliche Aspekte (Vergangenheit, Gegenwart, Zukunft, etc.) zu behalten, welche für spätere Beschreibungen der einzelnen identifizierten Themen (Topics) von Bedeutung sein können und für ein besseres Verständnis sorgen. Aggressive Stemmingalgorithmen, wie zum Beispiel der Porter-Stemmer (vgl. FRAKES (1992), PORTER (1997)), führen zu teils starken Sinnänderungen eines Textes. Ein Stemming ist zudem sprachenabhängig und z.B. im Deutschen oder Arabischen schwerer durchführbar (vgl. BAEZA-YATES/CASTILLO (2006)).

3.1.2 Lokales Wörterbuch & Referenzwerte

Aus dem Referenzkorpus \mathcal{R} kann ein Lokales Wörterbuch $\mathcal{L} = \{x \mid \exists w_{s_b} \in d_s : x = w_{s_b}\}$ abgeleitet werden, welches alle Wortformen w des Referenzkorpus \mathcal{R} beinhaltet. Alle weltweit existierenden Wortformen in allen Sprachen bilden das Globale Wörterbuch \mathcal{G} , somit gilt $\mathcal{L} \subseteq \mathcal{G}$.

Bezogen auf das Referenzkorpus \mathcal{R} wird für jede Wortform x des Lokalen Wörterbuchs \mathcal{L} die Generelle Termfrequenz

$$tf_x = \frac{\sum_{s=1}^{|\mathcal{R}|} \sum_{b=1}^{R_s} \delta_{\{w_{s_b} = x\}}}{\sum_{s=1}^{|\mathcal{R}|} R_s} \quad (3.1)$$

mit δ als Kronecker-Delta und $|M|$ als Kardinalität einer Menge M , sowie die Inverse Dokumentenfrequenz

$$idf_x = \log \frac{|\mathcal{R}|}{|\{d_s \mid \exists b : w_{s_b} = x\}|} \quad (3.2)$$

berechnet. Die Verwendung der Inversen Dokumentenfrequenz findet man u.a. bei SALTON ET AL. (1975), SALTON (1989), BERRY/BROWNE (2005), HE ET AL. (2007), BAEZA-YATES/RIBEIRO-NETO (2011).

Die Inverse Dokumentenfrequenz idf_x ist ein statistisches Maß zur Messung der generellen Wichtigkeit einer Wortform x innerhalb des Referenzkorpus \mathcal{R} .

Sie berücksichtigt eine Forderung an charakteristische Merkmale (dass sie in wenigen Dokumenten besonders häufig, im Allgemeinen jedoch eher selten auftreten (vgl. HEYER ET AL. (2008))) und ist somit ein Maß für die Themenspezifität einer Wortform x .

3.1.3 Merkmalsextraktion & Vector Space Modell

Für die so genannte Merkmalsextraktion, d.h. für die maschinenverarbeitbare Darstellung eines Dokuments d mit R Wortformen, wird neben der Generellen Termfrequenz tf_x und der Inversen Dokumentenfrequenz idf_x die Spezifische Termfrequenz

$$tf_{x,d} = \sum_{b=1}^R \delta_{\{w_b = x\}} \quad (3.3)$$

einer Wortform x bezüglich eines Dokuments d benötigt. Sie gibt die Häufigkeit einer Wortform x innerhalb eines Dokuments d wieder.

Eine Normalisierte Termfrequenz

$$\overline{tf}_{x,d} = \frac{tf_{x,d}}{\max_x \{tf_{x,d}\}} \quad (3.4)$$

bezogen auf ein Dokument d erhält man beispielsweise durch Division einer Spezifischen Termfrequenz $tf_{x,d}$ durch die Spezifische Termfrequenz des häufigsten Terms (der häufigsten Wortform) $\max_x \{tf_{x,d}\}$ eines betrachteten Dokuments d . Andere Normalisierungen wie zum Beispiel die Division durch die Gesamtanzahl erkannter Wortformen eines Dokuments d sind denkbar

$$\overline{tf}_{x,d} = \frac{tf_{x,d}}{\sum_{x \in \mathcal{L}} \sum_{b=1}^R \delta_{\{w_b=x\}}} \quad (3.5)$$

Dadurch wird erreicht, dass Dokumente verschiedener Längen, also mit unterschiedlicher Anzahl an Wortformen, miteinander verglichen werden können.

Durch Multiplikation dieser Normalisierten Termfrequenzen $\overline{tf}_{x,d}$ mit der entsprechenden Inversen Dokumentenfrequenz idf_x einer Wortform x erhält man eine Gewichtung

$$g_{x,d} = idf_x \cdot \overline{tf}_{x,d} \quad (3.6)$$

einer Wortform x in einem Dokument d . Weitere mögliche Termgewichtungen werden unter anderem in SALTON/BUCKLEY (1988), BERRY/BROWNE (2005), BAEZA-YATES/RIBEIRO-NETO (2011) besprochen. Die Termgewichte dienen zur Abbildung eines Dokuments d als Dokumentenvektor v^d im so genannten Vector Space (vgl. SALTON (1971), SALTON ET AL. (1975), SALTON (1989), LEE ET AL. (1997), KOBAYASHI/AONO (2008)) mit seinen Komponenten $v_z^d = g_{z,d}$

$$v^d = \begin{pmatrix} v_1^d \\ \vdots \\ v_z^d \\ \vdots \\ v_Z^d \end{pmatrix}$$

$$z = 1, \dots, Z,$$

wobei für eine effiziente Durchführung der Bearbeitung der Dokumententexte mit aussagekräftigem Ergebnis anstatt aller $|\mathcal{L}|$ bekannten Wortformen x aus dem Lokalen Wörterbuch \mathcal{L} eine Untermenge $\mathcal{L}' \subseteq \mathcal{L}$ aus den Z generell häufigsten Wortformen x des Lokalen Wörterbuchs \mathcal{L} verwendet wird. Dabei ist $|\mathcal{L}'| = Z \leq |\mathcal{L}|$.

Diese Untermenge \mathcal{L}' , auch Verkleinertes Lokales Wörterbuch genannt, wird nach einer vorausgehenden sogenannten Stoppwortentfernung generiert. Stoppworte sind sehr häufig vorkommende Wortformen, wie zum Beispiel *der*, *die*, *das* etc., mit im Allgemeinen sehr niedrigen Inversen Dokumentenfrequenzen. Da sie themenübergreifend relativ häufig auftreten, sind sie zum Differenzieren von Dokumenten in Themen

(Topics) nur von sehr geringem Nutzen. Zur Entfernung dieser Stoppworte werden allgemein verfügbare Stoppwortlisten eingesetzt (vgl. FOX (1992)).

Eine Hauptkomponentenanalyse ist eine weitere Möglichkeit zur Dimensionsreduktion, die z.B. in ZHUKOV/GLEICH (2004) zur Themendetektion angewendet wird. Der Rechenaufwand zur Bestimmung der Hauptkomponenten für große Dimensionen ist jedoch oftmals zu aufwendig. Verfahren, die z.B. auch in der Bild- bzw. Gesichtserkennung eingesetzt werden, versuchen diese Problematik zu umgehen (vgl. GÓMEZ/PESQUET-POPESCU (2007)). Ansätze mit mehreren Wortformen als Merkmale, wie z.B. Bi-Gramme, Tri-Gramme, etc., wurden ebenfalls examiniert (vgl. PAPKA/ALLAN (1998), ZHANG ET AL. (2007)). Weitere Arbeiten z.B. zur Auswahl adäquater Merkmale, zur Dimensionsreduktion und zur Termgewichtung im 'Vector Space' finden sich z.B. in DHILLON ET AL. (2004), HOWLAND/PARK (2004), SOUCY/MINEAU (2005).

3.1.4 Akkuratessse des Lokalen Wörterbuchs

Die Güte des Verkleinerten Lokalen Wörterbuchs \mathcal{L}' in Bezug auf ein (neues nicht) im Referenzkorpus \mathcal{R} enthaltenes Dokument d_s kann durch die Prozentuale Worterkennungsrates für d_s

$$per_{d_s} = \frac{\sum_{x \in \mathcal{L}'} \sum_{b=1}^{R_s} \delta_{\{w_{s_b} = x \mid x \in \mathcal{L}'\}}}{R_s} \quad (3.7)$$

ermittelt werden. Sie ist somit ein Maß zur Bestimmung der Akkuratessse des Verkleinerten Lokalen Wörterbuchs \mathcal{L}' und repräsentiert den prozentualen Anteil erkannter Wörter w_s des Dokuments d_s .

Aktualität und Größe des Lokalen Wörterbuchs \mathcal{L} haben einen direkten Einfluss auf das Verkleinerte Lokale Wörterbuch \mathcal{L}' und somit auch auf die Prozentuale Erkennungsrates per_{d_s} eines Dokuments d_s . Eine mit der Zeit τ abfallende Prozentuale Worterkennungsrates $per_{d_s}^\tau$, bei zeitlich aufeinander folgenden Dokumenten d^τ , deutet auf zeitliche Entwicklungen einer Sprache hin. Die Verschlechterung der Worterkennungsrates kann durch eine Aktualisierung des Lokalen Wörterbuchs \mathcal{L} auf \mathcal{L}_a , gebildet aus einem zeitgemäßerem Referenzkorpus \mathcal{R}_a , behoben werden. Durch das Lokale Wörterbuch \mathcal{L} bzw. Verkleinerte Lokale Wörterbuch \mathcal{L}' nicht erkannte häufig vorkommende Wortformen w aus aktuellen Analysedokumenten d^τ können auf neu aufgetretene Trend-Schlagwörter hindeuten, die sich themenspezifisch gebildet haben und sollten in das Lokale Wörterbuch \mathcal{L} bzw. \mathcal{L}' mit aufgenommen werden. Ein Beispiel dazu wird im Kapitel 5 gegeben.

Neben der Prozentualen Worterkennungsrate kann auch die Worterkennungsrate für d_s aus Mengensicht

$$mer_{d_s} = \sum_{x \in \mathcal{L}'} \frac{|ER_{x,d_s}|}{|ER_{x,d_s}| + |NER_{x,d_s}|} \quad (3.8)$$

mit

$$ER_{x,d_s} = \{w_{s_b} \mid \exists s_b \in \{1, \dots, R_s\}, w_{s_b} = x, \wedge x \in \mathcal{L}'\} \quad (3.9)$$

$$NER_{x,d_s} = \{w_{s_b} \mid \exists s_b \in \{1, \dots, R_s\}, w_{s_b} = x, \wedge x \notin \mathcal{L}'\} \quad (3.10)$$

von Interesse sein, welche ebenfalls ein Maß zur Bestimmung der Akkuratesses des Lokalen Wörterbuchs \mathcal{L} bzw. \mathcal{L}' ist, wobei hierbei speziell Doppelzählungen von Wortformen w_s innerhalb eines Dokuments d_s ausgeschlossen werden.

Vgl. dazu auch GONG ET AL. (2011).

3.2 (Un-) Ähnlichkeitsmaße

Zur Analyse von Dokumenten muss deren Ähnlichkeit zueinander ermittelt werden. Eine Einführung in Bezug auf zahlreiche Ähnlichkeitsmaße findet sich z.B. bei COR-MACK (1971), SNEATH/SOKAL (1973), BOCK (1974), SINT (1975), BOCK (1980) und ist Grundlage für die nachfolgenden Erläuterungen.

Dokumente d_i, d_j einer Dokumentenmenge D können als Vektoren v^{d_i} bzw. v^{d_j} im 'Vector Space' dargestellt werden. Diese Vektoren v^d bilden ihre entsprechenden Dokumente als Punkte im Z -dimensionalen Raum \mathbb{R}^Z ab. Eine mittels dieser Dokumentenvektoren v^d gebildete $|D| \times Z$ Datenmatrix liefert die Grundlage für weitere Datenanalyseverfahren. Solche Verfahren setzen voraus, dass Ähnlichkeiten zwischen zu analysierenden Objekten, in unserem Fall Dokumenten, numerisch erfassbar sind.

Mittels eines Ähnlichkeitsmaßes $sim(d_i, d_j)$ mit $0 \leq sim(d_i, d_j) = sim(d_j, d_i) \leq 1 = sim(d_i, d_i)$ (mit Normierung auf den Wert 1) bzw. Unähnlichkeitsmaßes $dis(d_i, d_j)$ mit $dis(d_i, d_i) = 0 \leq dis(d_i, d_j) = dis(d_j, d_i) \leq 1$ ist somit für jedes der $\frac{|D|(|D|-1)}{2}$ Dokumentenpaare d_i, d_j der Dokumentenmenge D eine Ähnlichkeit bzw. Unähnlichkeit bestimmbar. dis wird im Folgenden als Distanzmaß bezeichnet.

Wie allgemein bekannt, werden folgende Eigenschaften von einem Distanzmaß dis verlangt: Gefordert wird mindestens die Eigenschaft der Reflexivität $dis(d_i, d_i) = 0, \forall d_i \in D$, sowie der Symmetrie $dis(d_i, d_j) = dis(d_j, d_i), \forall d_i, d_j \in D$.

Erfüllt ein Distanzmaß weitere Eigenschaften wie zum Beispiel die Äquivalenz $dis(d_i, d_j) = 0 \Rightarrow d_i = d_j, \forall d_i, d_j \in D$ und die Dreiecksungleichung $dis(d_n, d_j) \leq dis(d_n, d_i) + dis(d_i, d_j), \forall n, i, j = 1, \dots, |D|$, heißt dis speziell ein metrisches Distanzmaß.

Die Art der untersuchten Merkmale der zu vergleichenden Objekte spielt bei der Wahl geeigneter (Un-) Ähnlichkeitsmaße eine wesentliche Rolle.

Für quantitative Datenmatrizen beispielsweise, wird häufig für $r \in \mathbb{N}^+$ die L_r -Distanz (auch Minkowski Distanz genannt)

$$dis(d_i, d_j) = \left(\sum_{z=1}^Z |v_z^{d_i} - v_z^{d_j}|^r \right)^{\frac{1}{r}} \quad (3.11)$$

verwendet, wobei sich für $r = 1$ die sogenannte City-Block Distanz (auch Manhattan Distanz genannt), für $r = 2$ der euklidische Abstand $dis(d_i, d_j) = \|v^{d_i} - v^{d_j}\|$ und für $r \rightarrow \infty$ die Tchebychev Distanz (auch Maximum-Distanz genannt) ergeben.

Für die L_r -Distanz ist auch eine Gewichtung

$$dis(d_i, d_j) = \left(\sum_{z=1}^Z \alpha_z |v_z^{d_i} - v_z^{d_j}|^r \right)^{\frac{1}{r}} \quad (3.12)$$

mit $\alpha_1, \dots, \alpha_Z \geq 0$ für jedes Merkmal spezifisch möglich.

Ein weiteres und vor allem im Text Mining weit verbreitetes und erprobtes Distanzmaß ist das Cosinus-Maß (vgl. u.a. STREHL ET AL. (2000))

$$\cos(v^{d_i}, v^{d_j}) = \frac{\sum_{z=1}^Z v_z^{d_i} \cdot v_z^{d_j}}{\sqrt{\sum_{z=1}^Z (v_z^{d_i})^2} \cdot \sqrt{\sum_{z=1}^Z (v_z^{d_j})^2}} \quad (3.13)$$

als Ähnlichkeitsmaß

$$sim(d_i, d_j) \cong \cos(v^{d_i}, v^{d_j})$$

beziehungsweise als Unähnlichkeitsmaß

$$dis(d_i, d_j) \cong 1 - \cos(v^{d_i}, v^{d_j})$$

zwischen Dokumenten d_i und d_j mit

$$sim(d_i, d_j), dis(d_i, d_j) \in \mathbb{R} \cap [0, 1].$$

Je kleiner der Winkel zwischen zwei betrachteten Dokumentenvektoren v^{d_i} und v^{d_j} ist, desto ähnlicher sind sich die Dokumente d_i und d_j . Da ein Dokument d_i durch den entsprechenden Dokumentenvektor v^{d_i} im 'Vector Space' dargestellt wird und dessen Einträge $v_z^{d_i}$ auf Grund der Merkmalsextraktion (vgl. Kapitel 3.1.3) nur positive Werte annehmen, befinden sich alle Dokumentenvektoren v^d im 1. Orthanten, somit kann der Winkel zwischen zwei Dokumentenvektoren Werte aus $[0^\circ, 90^\circ]$ (bzw. $[0, \frac{1}{2}\pi]$) annehmen.

Da die Bedingung der Äquivalenz $dis(d_i, d_j) = 0 \Rightarrow (d_i = d_j), \forall d_i, d_j \in D$, für $dis(d_i, d_j) = 1 - \cos(v^{d_i}, v^{d_j})$ nicht erfüllt wird, spricht man in diesem Fall (genauer) von einer Quasimetrik.

Ein weiteres bekanntes Maß ist die Mahalanobis-Distanz (vgl. MAHALANOBIS (1936))

$$dis(d_i, d_j) = (v^{d_i} - v^{d_j})' \sum^{-1} (v^{d_i} - v^{d_j}) \quad (3.14)$$

wobei \sum eine empirische $Z \times Z$ -Kovarianzmatrix der Merkmale ist.

Bei binären Daten, also $v_z^{d_i} = 0$ oder 1 (ja/nein), betrachtet man die Anzahlen $a = \sum_z \min\{v_z^{d_i}, v_z^{d_j}\}$ und $b = \sum_z \min\{1 - v_z^{d_i}, 1 - v_z^{d_j}\}$ für übereinstimmende 1- bzw. 0-Komponenten von v^{d_i} und v^{d_j} . Als Ähnlichkeitsmaß kann dann zum Beispiel

$$sim(d_i, d_j) = \frac{a + b}{Z} \quad (3.15)$$

als sogenannter M-Koeffizient oder

$$sim(d_i, d_j) = \frac{a}{Z - b} \quad (3.16)$$

als S-Koeffizient von Jaccard (Tanimoto) dienen, wobei sich aus dem Ähnlichkeitsmaß das Distanzmaß $dis(d_i, d_j) = 1 - sim(d_i, d_j)$ als metrisches Distanzmaß ergibt (vgl. JACCARD (1901), BOCK (1974) und SPÄTH (1977)).

Bei qualitativen Daten ist im Gegensatz zur quantitativen Analyse eine weitere Unterteilung in nominale und ordinale Merkmale notwendig. Beispielsweise ist als geeignetes Maß bei nominalen Daten die (gewichtete) Anzahl von übereinstimmenden Komponenten $sim(d_i, d_j) = \sum_z \alpha_z \delta_{\{v_z^{d_i} = v_z^{d_j}\}}$ denkbar. Ein mögliches Gewicht für jedes Merkmal z ist zum Beispiel $\alpha_z = 1$ (M-Koeffizient) (vgl. BOCK (1974), SPÄTH (1977)).

Bei probabilistischen Ähnlichkeitsmaßen gibt $sim(d_i, d_j)$ die Wahrscheinlichkeit wieder, mit der ein gegebenes Paar $d_i, d_j \in D$ höhere Merkmalsunterschiede aufweist als ein zufällig gewähltes Objektpaar (vgl. BOCK (1980)).

Probleme treten auf, wenn gleichzeitig verschiedene Merkmalstypen (quantitative, qualitative) zu bearbeiten sind. Lösungsansätze sind z.B. die Reduzierung auf einen einzigen Typ oder Mittelung bzw. Gewichtung der für jeden Merkmalstyp getrennt berechneten Ähnlichkeiten bzw. Unähnlichkeiten. Dabei geht im Allgemeinen jedoch Information verloren.

Der Wahl eines Ähnlichkeit- oder Distanzmaßes liegen keine grundsätzlich objektivierbaren Regeln zu Grunde. Sie stellt somit immer ein subjektives Element der Analyse dar.

3.3 Clusteranalyse zur Themendetektion

Große Dokumentensammlungen $D = \{d_1, d_2, \dots, d_{|D|}\}$ von Dokumenten d , wie sie von Online Portalen wie z.B. 'Spiegel Online', 'Die Welt', 'Die FAZ' u.a. geliefert werden, können durch das Gruppieren in eine relativ geringe Anzahl von Themen übersichtlich dargestellt und damit für weitere Analysen (vgl. Kapitel 4) nutzbar gemacht werden. Um Themen zu identifizieren und Dokumente nach Themen zu gruppieren (clustern), bieten sich eine Reihe von Clusteranalyse-Verfahren an. Eine Übersicht über gängige Verfahren findet sich u.a. bei BOCK (1974), BOCK (1980), ARABIE ET AL. (1996), JAIN ET AL. (1999) und ist Grundlage für die folgenden Ausführungen.

Der zu analysierende Datentyp bestimmt die Wahl des Ähnlichkeits- bzw. Distanzmaßes (vgl. Kapitel 3.2), sowie die Art des Clusterverfahrens. Dabei ist es ratsam, verschiedene Clusterverfahren zu erproben und gefundene Klassen hinsichtlich ihrer praktischen Zweckmäßigkeit und Interpretierbarkeit zu prüfen. Neben einer Untersuchung von Homogenitäts- und Heterogenitätsindizes ist die visuelle Inspektion einer Indizierten Hierarchie, auch bekannt als Dendrogramm, ein wesentliches Instrument. Ferner ist die (Nicht-) Übereinstimmung zweier Klassifikationen (Clusterings) auch quantifizierbar (weitere Ausführungen dazu siehe MIRKIN/CHERNYI (1970), RAND (1971), ARABIE/BOORMAN (1973), FOWLKES/MALLOWS (1983), HUBERT/ARABIE (1985)).

Die Clusteranalyse soll eine mittels der Merkmalsextraktion (vgl. Kapitel 3.1.3) aus der Dokumentenmenge D erstellte $|D| \times Z$ -Datenmatrix $(v_z^{d_i})$, deren i -te Zeile $v^{d_i} = (v_1^{d_i}, \dots, v_Z^{d_i})$ das Dokument $d_i \in D$ charakterisiert, in ein System $\mathcal{K} = \{C_1, \dots, C_k, \dots, C_{|\mathcal{K}|}\}$ nichtleerer Teilmengen $C_k \subseteq D$ überführen.

Falls eine solche Klassifikation \mathcal{K} $|\mathcal{K}|$ disjunkte Klassen (Cluster) $C_1, \dots, C_{|\mathcal{K}|}$ erzeugt, bei dem jedes Dokument $d_i \in D$ genau einer Klasse $C_k \in \mathcal{K}$ angehört, spricht man von einer Partition (disjunkte Klassifikation) der Dokumentenmenge D . Bei einer nicht disjunkten Klassifikation \mathcal{K} wird ein Dokument $d_i \in D$ nicht eindeutig einer Klasse $C_k \in \mathcal{K}$ zugeordnet. Stattdessen wird der Grad der Zugehörigkeit eines Dokuments $d_i \in D$ zu einer Klasse $C_k \in \mathcal{K}$ über Anteilswerte angegeben.

Ein weiterer Klassifikationstyp ist das hierarchische Clustering, bei dem Familien von Klassifikationen in einer stammbaumähnlichen Anordnung organisiert werden (vergleiche dazu Abbildung 3.1).

Intention der Clusteranalyse-Verfahren ist, eine gegebene Datenstruktur durch eine Klassenstruktur mit möglichst homogenen Klassen, die zueinander möglichst heterogen sind, darzustellen.

3.3.1 Gütekriterium

Zur Bestimmung, wie gut ein Clustering bzw. eine Klassifikation \mathcal{K} eine gegebene Datenstruktur wiedergibt, benötigt man ein numerisches Gütekriterium $g(\mathcal{K})$, das dem vorliegenden Daten- und Klassifikationstyp angepasst sein muss. Mit diesem Kriterium bestimmt man unter möglichen gefundenen Partitionen diejenige mit minimalstem Gütewert $g(\mathcal{K})$.

Ein häufig angewendetes Gütekriterium ist das sogenannte Varianzkriterium

$$g(\mathcal{K}) := \sum_{k=1}^{|\mathcal{K}|} \sum_{d_i \in C_k} \|v^{d_i} - c_k\|^2 \rightarrow \min_{\mathcal{K}} \quad (3.17)$$

auch L_2 -Kriterium oder Abstandsquadratensummenkriterium (vgl. BOCK (1974), SPÄTH (1977)) genannt.

Hierbei bezeichnet c_k den Klassenrepräsentanten (Centroid) von C_k mit seinen Komponenten

$$c_{k_z} = \frac{1}{|C_k|} \cdot \sum_{d_i \in C_k} v_z^{d_i} \quad (3.18)$$

also den Mittelwertvektor der Vektoren v^{d_i} aus der Klasse C_k . Dabei wird angenommen, dass thematisch ähnliche Dokumente $d_i \in D$, dargestellt durch die entsprechenden Dokumentenvektoren v^{d_i} im \mathbb{R}^Z , natürliche 'Punktwolken' bilden. Das Zentrum einer um diese 'Punktwolke' gebildeten Klasse C_k , wird durch den Centroid c_k charakterisiert.

3.3.2 Partitionierende Verfahren

Die Potenzmenge $\wp(D)$ mit ihrer großen Zahl möglicher Partitionen lässt eine Überprüfung sämtlicher Klassifikationen $\mathcal{K} = \{C_1, \dots, C_{|\mathcal{K}|}\} \subset \wp(D)$ mit $\emptyset \neq C_k \subset D, \forall C_k \in \mathcal{K}$, kaum zu. Man geht üblicherweise so vor, dass man eine zufällig oder geschickt gewählte Ausgangspartition $\mathcal{K}^{(0)}$ iterativ in Bezug auf das Gütekriterium optimiert.

Für solche partitionierende Verfahren sind zwei Vorgehensweisen üblich.

Beim **Austauschverfahren** (vgl. RUBIN (1966), MACQUEEN (1967)) wird ausgehend von einer Anfangsklassifikation $\mathcal{K}^{(0)}$ für jedes einzelne Dokument $d_i \in D$ getestet, ob durch Verschieben in eine andere Klasse eine Verbesserung hinsichtlich des Gütekriteriums erfolgt. Dieser Austausch von Dokumenten wird solange durchgeführt, bis keine weitere Verbesserung mehr möglich ist. Die Lösung entspricht nicht zwangsläufig einem globalen Optimum, also einer optimalen Klassifikation, jedoch einem lokalen Optimum bzgl. der Ausgangsklassifikation $\mathcal{K}^{(0)}$. Das Austauschverfahren erreicht üblicherweise bessere (kleinere) Gütewerte als ein im Folgenden beschriebenes Minimal-Distanz-Verfahren, benötigt dafür aber längere Rechenlaufzeit (vgl. dazu auch BOCK (1980)).

Während das Austauschverfahren ein gegebenes Gütekriterium direkt verbessert, wird bei einem **Minimal-Distanz-Verfahren** (vgl. FORGY (1965), MACQUEEN (1967)) wie dem k -means Verfahren das Gütekriterium indirekt verbessert (vgl. RUBIN (1966), BOCK (1974), SPÄTH (1977)).

Das **k-means Verfahren** ist ein auch im Text Mining weit verbreitetes und erprobtes Clusterverfahren. Als Distanzfunktion für metrische Datenmatrizen wird standardmäßig die euklidische Distanz benutzt. Andere Distanzmaße, wie das in dieser Arbeit zumeist eingesetzte Cosinus-Maß (vgl. Kapitel 3.2), haben sich im Text Mining jedoch dem euklidischen Maß als überlegen erwiesen (vgl. Kapitel 5.4.1, siehe auch STREHL ET AL. (2000)). Generell sind auch andere Distanzmaße denkbar. Beim k -means Verfahren (vgl. MACQUEEN (1967), SPÄTH (1977), MACKAY (2002)) muss die Anzahl zu erstellender Cluster als Eingabe vorgegeben werden. Vereinfacht durchläuft der k -means Algorithmus folgende Schritte (vgl. dazu auch BOCK (1974), HEYER ET AL. (2008), BAEZA-YATES/RIBEIRO-NETO (2011)). Nach zufällig gewählten Dokumenten als Initial-Centroiden werden die verbleibenden Dokumente demjenigen Cluster zugewiesen, dessen Ähnlichkeit durch Bestimmung der Distanz (vgl. Kapitel 3.2) zum entsprechenden Centroid am größten ist. BOCK (1974) zeigt, dass bei Berücksichtigung des Varianzkriteriums (vgl. Formel (3.17)) als Gütekriterium $g(\mathcal{K})$ jedes Objekt, in unserem Fall $d_i \in D$, zwangsläufig der Klasse zugeordnet wird, dessen Centroid es am ähnlichsten ist. Dies führt zu einem Initial-Clustering $\mathcal{K}^{(0)}$. Nach Neuberechnung der Zentren (Centroide) erfolgt eine Neuordnung aller Dokumente. Der Schritt der Neuberechnung einschließlich nachfolgender Neuordnung wird solange durchgeführt, bis sich die Centroide nicht mehr ändern. Es kann gezeigt werden (vgl. BOCK (1974)), dass die Partitionen $\mathcal{K}^{(\xi)}$ mit $\xi = 0, 1, 2, \dots$ hinsichtlich des Varianzkriteriums immer bessere Gütewerte $g(\mathcal{K}^{(0)}) \geq g(\mathcal{K}^{(1)}) \geq g(\mathcal{K}^{(2)}) \geq \dots$ erreichen und in der Regel eine stationäre Partition (Clustering) $g(\mathcal{K}) := g(\mathcal{K}^{(\xi)}) = g(\mathcal{K}^{(\xi+1)}) = \dots$ existiert.

Verschiedene Abwandlungen des k -means Algorithmus sind bekannt (vgl. BAEZA-YATES/RIBEIRO-NETO (2011)). Beim sogenannten Batch-Mode des k -means Verfahrens werden die Centroide nach der Zuordnung aller Dokumente neu berechnet, während beim sogenannten Online-Mode (auch 'incremental k -means' genannt) des k -means Algorithmus die Centroide bereits nach jedem einzelnen zugewiesenen Dokument neu berechnet werden (vgl. STEINBACH ET AL. (2000)). Laut STEINBACH ET AL. (2000) erreicht der Online-Mode bei allgemeinen Textmengen bessere Ergebnisse als der Batch-Mode. In Kapitel 5 wird der Online-Mode des k -means Algorithmus angewendet.

Kritisch am k -means Verfahren anzumerken ist, dass die Klassenanzahl $|\mathcal{K}|$ im Voraus zu wählen ist, sowie die Tatsache, dass verschiedene Durchläufe des k -means Verfahrens zu unterschiedlichen Ergebnissen führen, bedingt durch die jeweils zufällige Wahl der Initial-Centroiden (vgl. BAEZA-YATES/RIBEIRO-NETO (2011)).

3.3.3 Hierarchische Verfahren

Für zahlreiche Anwendungen, so auch in der vorliegenden Arbeit, werden hierarchische Verfahren eingesetzt. Detaillierte Ausführungen zu den hierarchischen Clusterverfahren, von denen die folgenden Erläuterungen abgeleitet sind, finden sich u.a. bei BOCK (1974), BOCK (1980), GORDON (1987), MARKOV/LAROSE (2007), MANNING ET AL. (2008), BAEZA-YATES/RIBEIRO-NETO (2011).

Während der k -means Algorithmus ein partitionierendes Verfahren ist, das eine Dokumentenmenge D in $|\mathcal{K}|$ Cluster einteilt, erstellen hierarchische Verfahren Hierarchien von Klassen (Clustern). Ihre Vorgehensweise ist entweder divisiver Natur, indem große Cluster in kleinere aufgeteilt werden, oder agglomerativer Art durch Vergrößerung vordefinierter Cluster in größere (vgl. JAIN ET AL. (1999), BAEZA-YATES/RIBEIRO-NETO (2011)). Divisive Verfahren sind konzeptionell komplexer als agglomerative Verfahren, da sie im Allgemeinen eine 'Subroutine' wie z.B. ein partitionierendes k -means Verfahren benötigen (auch bekannt unter 'bisecting k -means') (vgl. STEINBACH ET AL. (2000), MANNING ET AL. (2008)). Die übliche Vorgehensweise ist das agglomerative hierarchische Clustering.

Bei einem agglomerativen hierarchischen Clustering werden einelementige Klassen $\{d_1\}, \dots, \{d_{|D|}\}$ so zusammengefasst, dass daraus neue größere Partitionen entstehen. Durch weiterfolgende Fusionen von Klassen der neu entstandenen Partitionen erhält man schließlich eine Hierarchie $\mathcal{H} = \{C_1, \dots, C_{|\mathcal{K}|}\}$ bzgl. D , bei der sowohl die zu analysierende Dokumentenmenge D , als auch die daraus gebildeten einelementigen Mengen $\{d_1\}, \dots, \{d_{|D|}\}$ enthalten sind.

Solche Hierarchien können graphisch als Dendrogramm (vgl. Abbildung 3.1) dargestellt werden, wobei die Verschiedenheit $\mathcal{V}(C_k, C_l)$ zweier Cluster C_k und C_l , die zur Fusion anstehen, eine Rolle spielt und jeweils die am wenigsten verschiedenen Cluster fusioniert werden.

Im Dendrogramm wird jede neue Klasse C , entstanden durch Fusion der Klassen C_k und C_l , auf der Höhe $h(C) := \mathcal{V}(C_k, C_l)$ eingezeichnet. $\mathcal{V}(C_k, C_l)$ ist ein Verschiedenheitsindex, der je nach Verfahren (vgl. die Formeln (3.20) bis (3.24)) die Verschiedenheit zwischen den zwei Clustern C_k und C_l bestimmt. $h(C)$ kann als 'Heterogenitätsmaß' einer Klasse $C \in \mathcal{H}$ aufgefasst werden. Dabei gilt $h(C_k) \leq h(C)$ für eine Unterklasse $C_k \subset C$. Für jedes h der Indizierten Hierarchie (\mathcal{H}, h) existiert eine disjunkte Klassifikation $\mathcal{K}(h) = \{C_1, \dots, C_{|\mathcal{K}(h)|}\}$ mit $C \in \mathcal{H}$. Dabei ist noch zu entscheiden, welche Höhe h^* eine beste Lösung $\mathcal{K}(h^*)$ für die zu Grunde liegende Datenmenge darstellt.

Im Vergleich zu partitionierenden Verfahren, wie beispielsweise dem k -means Clusterverfahren, erlaubt ein hierarchisches Verfahren einen besseren Einblick in die vorliegende Datenstruktur. Oft wird auch argumentiert, dass hierarchische Clusteralgorithmen bessere Ergebnisse erzielen als ein 'flacher' Clusteralgorithmus (vgl. CUTTING ET AL. (1992), JAIN ET AL. (1999) und LARSEN/AONE (1999)). Jedoch haben ZHAO/KARYPIS (2002) experimentelle Ergebnisse geliefert, die einen anderen Schluß zulassen (vgl. MANNING ET AL. (2008)).

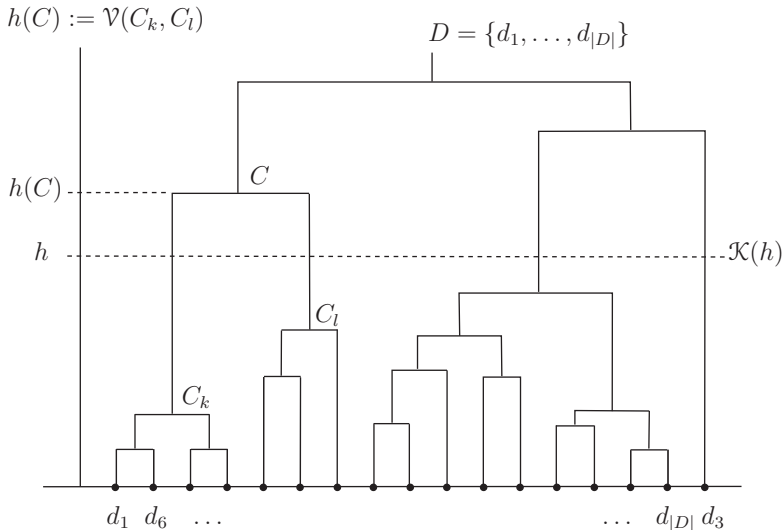


Abbildung 3.1: Dendrogramm einer Indizierten Hierarchie (\mathcal{H}, h) (in Anlehnung an BOCK (1980)).

Die Erstellung eines Dendrogramms über ein agglomeratives Verfahren wird nachfolgend kurz beschrieben.

Man beginnt mit einer Anfangspartition $\mathcal{K}^{(0)} = \{C_1^{(0)}, \dots, C_{|D|}^{(0)}\}$, bei der jede Klasse mit $C_i^{(0)} = \{d_i\}$, $\forall d_i \in D$, aus einelementigen Mengen der Dokumentenmenge D besteht. Somit entspricht $|\mathcal{K}^{(0)}| = |D|$.

Anschließend werden sukzessive die sich ähnlichsten Cluster $C_{l_1}^{(\xi)}$ und $C_{l_2}^{(\xi)}$ mit

$$\min_{\substack{l_1, l_2 \in \{1, \dots, |D| - \xi\}, \\ C_{l_1}^{(\xi)} \neq C_{l_2}^{(\xi)}}} \mathcal{V}(C_{l_1}^{(\xi)}, C_{l_2}^{(\xi)}) = \mathcal{V}(C_{l_1^*}^{(\xi)}, C_{l_2^*}^{(\xi)}) \quad (3.19)$$

durch Iteration für $\xi = 0, 1, \dots, |D| - 1$ zu einem neuen Cluster zusammengefasst. Aus $\mathcal{K}^{(\xi)} = \{C_1^{(\xi)}, \dots, C_{|D| - \xi}^{(\xi)}\}$ erhält man somit jeweils ein nachfolgendes Clustering $\mathcal{K}^{(\xi+1)} = \{C_1^{(\xi+1)}, \dots, C_{|D| - \xi - 1}^{(\xi+1)}\}$. Am Ende des Iterationsprozesses $\xi = |D| - 1$ erhält man, nachdem alle $|D|$ einelementigen Klassen zu einer einzigen Klasse vereinigt sind, $\mathcal{K}^{(|D|-1)} = \{D\}$ als Clustering.

Die hierarchischen Clusterverfahren arbeiten mit unterschiedlichen Verschiedenheitsindizes \mathcal{V} (vgl. JOHNSON (1967)). Danach lassen sich die hierarchischen Verfahren einteilen in 'Single-Linkage' (vgl. FLOREK ET AL. (1951), SNEATH (1957), SNEATH/SOKAL (1973), CROFT (1977)), 'Complete-Linkage' (vgl. MCQUITTY (1957), LANCE/WILLIAMS (1966), LANCE/WILLIAMS (1967)), 'Average-Linkage' (vgl. SOKAL/MICHENER (1958)) und 'Ward'-Verfahren (vgl. WARD (1963), EL-HAMDOUCHI/WILLETT (1986)).

Formal lauten sie wie folgt:

'Single-Linkage':

$$\mathcal{V}(C_k, C_l) := \min_{d_i \in C_k, d_j \in C_l} \{dis(d_i, d_j)\} \quad (3.20)$$

'Complete-Linkage':

$$\mathcal{V}(C_k, C_l) := \max_{d_i \in C_k, d_j \in C_l} \{dis(d_i, d_j)\} \quad (3.21)$$

'Average-Linkage':

$$\mathcal{V}(C_k, C_l) := \frac{1}{|C_k| \cdot |C_l|} \sum_{d_i \in C_k, d_j \in C_l} dis(d_i, d_j) \quad (3.22)$$

'Ward':

$$\mathcal{V}(C_k, C_l) := \sqrt{\frac{2 \cdot |C_k| \cdot |C_l|}{|C_k| + |C_l|}} \|c_k - c_l\|_2 \quad (3.23)$$

Ein modifiziertes Ward-Verfahren für Dokumente, das den euklidischen Teil durch das Cosinusmaß ersetzt, lautet wie folgt:

$$\begin{aligned} \mathcal{V}(C_k, C_l) &:= \sqrt{\frac{2 \cdot |C_k| \cdot |C_l|}{|C_k| + |C_l|}} dis(C_k, C_l) \\ &= \sqrt{\frac{2 \cdot |C_k| \cdot |C_l|}{|C_k| + |C_l|}} (1 - \cos(c_k, c_l)) \end{aligned} \quad (3.24)$$

wobei sich der entsprechende Centroid c_k eines Clusters C_k mit seinen Komponenten

$$c_{k_z} = \frac{1}{|C_k|} \cdot \sum_{d_i \in C_k} v_z^{d_i} \quad (3.25)$$

mit $k = 1, \dots, |\mathcal{K}|$ und $z = 1, \dots, Z$ ergibt.

Das 'Single-Linkage'-Verfahren bildet vorzugsweise wenige große Klassen. Elemente mit großer Unähnlichkeit zu allen anderen Elementen bleiben lange isoliert (Ausreißer) und werden erst spät im Iterationsverfahren mit einer Klasse fusioniert. Ein Problem des 'Single-Linkage'-Verfahren ist deshalb die sogenannte 'Kettenbildung', bei der getrennte Klassen über ein Brückenelement miteinander verknüpft werden (vgl. z.B. MARKOV/LAROSE (2007)).

Das 'Complete-Linkage'-Verfahren bildet eher kleine Klassen. Ausreißer entdeckt dieses Verfahren seltener.

Wird beim 'Ward'-Verfahren die euklidische Distanz angewendet, werden die beiden Klassen fusioniert, bei denen das Gütekriterium, in diesem Fall das Varianzkriterium, sich am wenigsten verschlechtert. Auch dieses Verfahren eignet sich nicht zur Erkennung von Ausreißern. Die Ergebnisse der vorliegenden Arbeit basieren im Wesentlichen auf dem modifizierten 'Ward'-Verfahren.

3.3.4 Clustering-Kardinalität

In der Praxis muss eine möglichst sinnvolle Anzahl $|\mathcal{K}|$ an Klassen gewählt werden. Während beim k -means Verfahren die Klassenanzahl $|\mathcal{K}|$ im Voraus angegeben werden muss, stellt sich bei hierarchischen Verfahren die Frage, auf welcher 'Höhe' einer Hierarchie \mathcal{H} der sogenannte 'cut' gemacht werden soll (vgl. $\mathcal{K}(h)$ Abbildung 3.1). Die Beantwortung dieser Frage ist nicht immer einfach und naheliegend. Auch eine subjektive Beurteilung durch Inspektion kann auf Grund unterschiedlicher Präferenzen eines Betrachters, hinsichtlich der Granularität einer möglichen Klassifikation, differieren. Der Versuch einer Objektivierung dieses Problems ist in der Literatur als Ellbogenkriterium bekannt (vgl. u.a. BOCK (1980), MANNING ET AL. (2008)). Dabei wird die Anzahl $|\mathcal{K}|$ Klassen so gewählt, dass eine Verfeinerung der Klassenstruktur einen Güteindex (z.B. das Varianzkriterium, vgl. dazu Kapitel 3.3.1) nur noch unwesentlich verbessern würde.

Für $|\mathcal{K}| = 2, 3, \dots$ ermittelt man den zugehörigen Güteindex $g_{|\mathcal{K}|} := \min\{g(\mathcal{K}) \mid \mathcal{K} \text{ mit } |\mathcal{K}| \text{ Klassen}\}$. Bei einem für steigende Anzahl $|\mathcal{K}|$ monoton fallenden Güteindex wählt man als 'optimale' Anzahl $|\mathcal{K}|$ den Wert, bei dem die Abnahme $g_{|\mathcal{K}|-1} - g_{|\mathcal{K}|}$ bzw. $\frac{g_{|\mathcal{K}|-1} - g_{|\mathcal{K}|}}{g_{|\mathcal{K}|-1}}$ relativ groß ist (vgl. BOCK (1980)). In der Abbildung 3.2 ist diese Stelle als Knick für $|\mathcal{K}| = 4$ zu erkennen, der an einen Ellbogen erinnert.

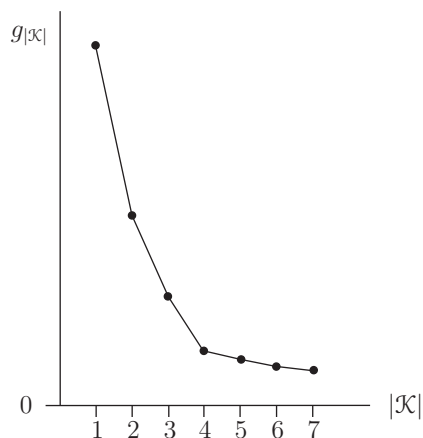


Abbildung 3.2: Beispiel Ellbogen.

Ein weit verbreitetes Gütekriterium ist das bereits beschriebene Varianzkriterium (vgl. Kapitel 3.3.1).

$$g(\mathcal{K}) = g(\{C_1, \dots, C_{|\mathcal{K}|}\}) = \sum_{k=1}^{|\mathcal{K}|} \sum_{d_i \in C_k} \sum_{z=1}^Z (v_z^{d_i} - c_{kz})^2 \quad (3.26)$$

In der Arbeit wird neben diesem Kriterium, eine abgewandelte Form, welche den euklidischen Teil durch das für das Text Mining weit verbreitete Cosinusmaß ersetzt, angewendet:

$$\begin{aligned} g(\mathcal{K}) &= g(\{C_1, \dots, C_{|\mathcal{K}|}\}) \\ &= \sum_{k=1}^{|\mathcal{K}|} \sum_{d_i \in C_k} \text{dis}(d_i, C_k) \end{aligned} \quad (3.27)$$

$$= \sum_{k=1}^{|\mathcal{K}|} \sum_{d_i \in C_k} (1 - \cos(v^{d_i}, c_k)) \quad (3.28)$$

Analog zur Ellbogendarstellung ist es bei hierarchischen Verfahren auch möglich, den 'cut' so zu legen, dass er den größten Abstand zwischen zwei 'Höhen' im Dendrogramm schneidet. Dies kann als 'natürliche' Klassenanzahl der vorliegenden Datenstruktur interpretiert werden (vgl. MANNING ET AL. (2008)).

Eine weitere Möglichkeit einer graphischen Unterstützung zur Bestimmung adäquater Clusteranzahlen wird z.B. in ROUSSEEUW (1987) beschrieben. Dabei wird jedes Cluster durch eine sogenannte Silhouette repräsentiert, die auf dem Vergleich von Kompaktheit und Separation basiert. Alle Silhouetten werden kombiniert in einem einzigen Silhouetten-Plot, das erlaubt die relative Güte der Cluster zu erfassen. Die durchschnittlichen Silhouetten ermöglichen somit eine Evaluation eines Clusterings und dessen Güte und können somit zur Bestimmung adäquater Clusteranzahlen genutzt werden.

In dieser Arbeit wird neben visueller Inspektion das Ellbogenkriterium nach Formel (3.26) bzw. (3.28) angewendet.

Kapitel 4

Relationen in temporalen Themen-Graphen

Um Themen (Topics) zu identifizieren, ist es üblich, eine Sammlung $D = \{d_1, d_2, \dots, d_{|D|}\}$ von Dokumenten d zu analysieren (gruppieren) (vergleiche dazu auch Kapitel 3.3). Für das Bestimmen von Trends muss zusätzlich die zeitliche Entwicklung der Dokumenteninhalte d^τ für sich ändernde Zeitfenster τ betrachtet werden, was zu zeitlich aufeinander folgenden Dokumentenkorpora $D^\tau = \{d_1^\tau, d_2^\tau, \dots, d_{|D^\tau|}^\tau\}$ führt. Als Dokumentenquellen können Online Nachrichtenportale und Printmedien wie zum Beispiel Spiegel Online, Die Welt, Die FAZ u.a., deren Dokumente als Dokumentenstrom über die Zeit hinweg ausgesendet werden, dienen. Solche Informationsquellen stellen kontinuierlich Dokumente zu aktuellen Themen von allgemeinem Interesse bereit. Diese Dokumente besitzen sogenannte Zeitstempel (Datum der Veröffentlichung) und können somit in Zeitfenstern τ zusammengefasst werden. Dokumente eines Betrachtungsintervalls $[t_1, t_2]$ aufeinander folgender Zeitfenster $\tau \in [t_1, t_2]$ bilden dann die Textmengen D^τ für die Trendanalysen. Die Textmengen D^τ werden auch als Analysekorpora bezeichnet.

4.1 Relationen & Schranken

Für einen Betrachtungszeitraum $[t_1, t_2]$ erhält man für jedes $\tau \in [t_1, t_2]$ ein Clustering $\mathcal{K}^\tau = \{C_1^\tau, \dots, C_{|\mathcal{K}^\tau|}^\tau\}$.

Um Zusammenhänge zwischen den gefundenen Themen-Clustern verschiedener Zeitfenster $\tau \in [t_1, t_2]$ zu ermitteln, werden für alle Klassifikationen \mathcal{K}^τ paarweise Relationsmatrizen erstellt (vgl. Tabelle 4.1). Man erhält für $m_\tau = \tau - t_1 + 1$ Zeitfenster $m_\tau \times (m_\tau - 1)/2$ Relationsmatrizen.

	C_1^τ	\dots	$C_{k_\tau}^\tau$	\dots	$C_{ \mathcal{K}^\tau }^\tau$	\dots	$C_{K_{\tau',\tau}}^\tau$
$C_1^{\tau'}$			\vdots				
\vdots			\vdots				
$C_{k_{\tau'}}^{\tau'}$		\dots	$dis^{\tau',\tau}(C_{k_{\tau'}}^{\tau'}, C_{k_\tau}^\tau)$				
\vdots							
$C_{ \mathcal{K}^{\tau'} }^{\tau'}$							
\vdots							
$C_{K_{\tau',\tau}}^{\tau'}$							

Tabelle 4.1: Relationsmatrix mit $dis^{\tau',\tau}(C_{k_{\tau'}}^{\tau'}, C_{k_\tau}^\tau)$ für $\mathcal{K}^{\tau'}$ und \mathcal{K}^τ .

Mit den Clusterings $\mathcal{K}^{\tau'}$ und \mathcal{K}^τ mit $\tau', \tau \in [t_1, t_2]$ erstellt man eine Relationsmatrix (vgl. Tabelle 4.1), indem die Unähnlichkeiten $dis^{\tau',\tau}(C_{k_{\tau'}}^{\tau'}, C_{k_\tau}^\tau)$ (vgl. dazu Kapitel 3.2) zwischen den entsprechenden Clustermengen bestimmt werden. Man beachte, dass die Anzahl Themen-Cluster unterschiedlicher Zeitfenster τ' und τ nicht identisch sein muss. Die Matrix von Unähnlichkeiten hat die Dimension $K_{\tau',\tau} = \max\{|\mathcal{K}^{\tau'}|, |\mathcal{K}^\tau|\}$. Für den Fall, dass $|\mathcal{K}^\tau|$ kleiner ist als $|\mathcal{K}^{\tau'}|$, hat die Relationsmatrix für die Spalten $C_{|\mathcal{K}^\tau|+1}^\tau$ bis $C_{K_{\tau',\tau}}^\tau$ fehlende Werte. Für den umgekehrten Fall $|\mathcal{K}^\tau| > |\mathcal{K}^{\tau'}|$ hat die Relationsmatrix für die Zeilen $C_{|\mathcal{K}^{\tau'}|+1}^{\tau'}$ bis $C_{K_{\tau',\tau}}^{\tau'}$ ebenfalls fehlende Werte. Für $|\mathcal{K}^\tau| = |\mathcal{K}^{\tau'}|$ ergibt sich eine quadratische Relationsmatrix.

Falls die Unähnlichkeit $dis^{\tau',\tau}(C_{k_{\tau'}}^{\tau'}, C_{k_\tau}^\tau)$ eines Clusterpaares $C_{k_{\tau'}}^{\tau'}$ und $C_{k_\tau}^\tau$ in der Relationsmatrix 'klein' ist, geht man davon aus, dass das Thema des Clusters $C_{k_{\tau'}}^{\tau'}$ im Zeitfenster τ' sehr ähnlich zum Thema von Cluster $C_{k_\tau}^\tau$ in Zeitfenster τ ist.

Abbildung 4.1 verdeutlicht mögliche Eigenschaften von Themen-Cluster Relationen zwischen verschiedenen Zeitfenstern. Dabei werden für die Unähnlichkeit obere dis_{ub} und untere dis_{lb} Schranken ('thresholds') festgelegt. Diese Schranken hängen ab von den zu Grunde liegenden zu analysierenden Daten, dem benutzten (Un-)Ähnlichkeitsmaß und der Größe Z des 'Vector Space' mit seinen aus dem Verkleinerten Lokalen Wörterbuch \mathcal{L}' verwendeten Wortformen x . (Vgl. auch GAUL/VINCENT (2013))

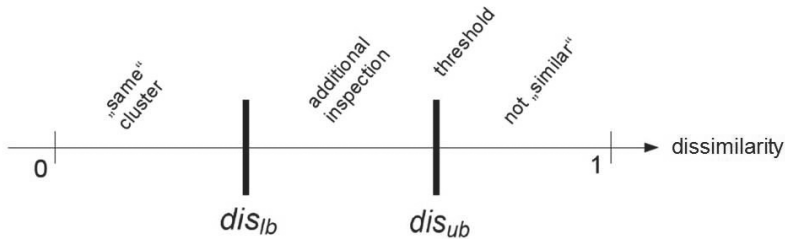


Abbildung 4.1: Schranken dis_{lb} und dis_{ub} .

Folgende Situationen sind denkbar (vgl. Abbildung 4.1). Befindet sich die Unähnlichkeit zweier Cluster (Themen) unter der vorgegebenen Schranke dis_{ub} , geht man davon aus, dass sich die zwei Themen-Cluster ähnlich sind ("same" cluster). Ist die Unähnlichkeit größer als die vorgegebene obere Schranke dis_{ub} , werden die zwei Themen-Cluster als zueinander unähnlich (not "similar") angesehen, sie behandeln somit unterschiedliche Themen. Liegen die Unähnlichkeitswerte zwischen Clustern innerhalb dieser beiden Schranken ist eine zusätzliche Inspektion durch einen Experten erforderlich, mit der entschieden wird, ob die Themen-Cluster noch ähnlich sind, d.h. die Unähnlichkeit noch 'klein genug' ist.

Wenn das Zeilenminimum aller Unähnlichkeiten eines Themen-Clusters $C_{k_{\tau'}}^{\tau'}$ aus dem Zeitfenster τ' größer ist als die vordefinierte obere Schranke dis_{ub} , kann man daraus folgern, dass das Themen-Cluster keine Entsprechung, also kein ähnliches Themen-Cluster im Zeitfenster τ hat. Das Thema $C_{k_{\tau'}}^{\tau'}$ aus Zeitfenster τ' existiert in τ nicht.

Falls alle Unähnlichkeiten in der Spalte von $C_{k_{\tau}}^{\tau}$ diese obere Schranke dis_{ub} überschreiten, beschreibt Thema $C_{k_{\tau}}^{\tau}$ ein neues Thema in Zeitfenster τ in Bezug auf Zeitfenster τ' für $\tau' < \tau$.

4.2 Multi-Themen-Graph

Die Themen-Cluster $C_{k_{\tau}}^{\tau}$ aller Zeitfenster τ eines Betrachtungshorizonts $[t_1, t_2]$ können in einem Graphen als Knoten (τ, k_{τ}) in einem durch eine horizontale Zeitfensterachse sowie eine vertikale Achse aufgespannten Diagramm dargestellt werden, wobei die Knoten entsprechend ihrer zeitlichen Komponente τ entlang der Zeitfensterachse ausgerichtet werden.

Jeder Knoten (τ, k_{τ}) eines Themas $C_{k_{\tau}}^{\tau}$ besitzt neben seiner zeitlichen Komponente τ eine spezifische Themen-Frequenz

$$f((\tau, k_{\tau})) = \frac{|C_{k_{\tau}}^{\tau}|}{|D^{\tau}|} \quad (4.1)$$

die in der vertikalen Richtung wiedergegeben wird. Die Themen-Frequenz gibt den Anteil der einem Themen-Cluster $C_{k_{\tau}}^{\tau}$ zugeordneten Dokumente eines Zeitfensters τ an und spiegelt somit das allgemeine Interesse an diesem Thema wider.

Zeitfenster τ haben in dieser Arbeit immer dieselbe Größe. Bei unterschiedlich großen Zeitfenstern muss die Themen-Frequenz bezüglich der Größe des jeweilig betrachteten Zeitfensters τ normalisiert werden.

Durch den Vergleich von Clusterings aller Zeitfenster τ innerhalb eines Betrachtungsintervalls $[t_1, t_2]$ kann die Relation $Rel(t_1, t_2)$ als sogenannter Multi-Themen-Graph erstellt werden.

In den folgenden Beispielgraphen werden exemplarisch die Zeitfenster $t_1, \tau', \tau, \tau + r, t_2$ aus der Relation $Rel(t_1, t_2)$ ausgewählt, wobei weitere innerhalb des Betrachtungszeitraums $[t_1, t_2]$ der Relation $Rel(t_1, t_2)$ vorhandene Zeitfenster aus Übersichtlichkeitsgründen weggelassen werden.

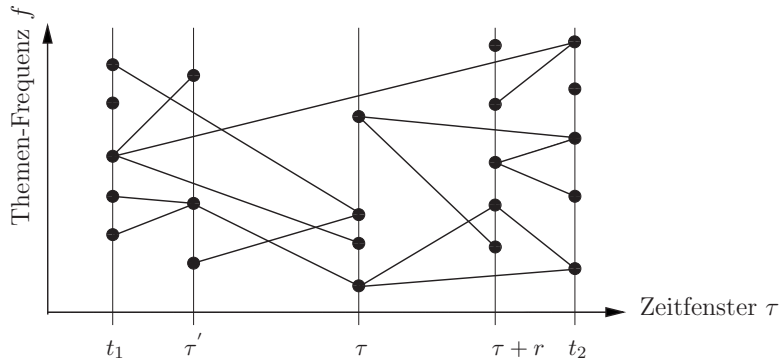


Abbildung 4.2: Multi-Themen-Graph der Relation $Rel(t_1, t_2)$.

Die Knoten verschiedener Zeitfenster τ werden durch Kanten verbunden, wenn auf Grund der Relationsschranken-Überlegung ihre Unähnlichkeit 'klein genug' ist (vgl. Kapitel 4.1). Dadurch entsteht der maximal m_{t_2} -partite Multi-Themen-Graph $G = (N, A)$, bestehend aus $N = \{(\tau, k_\tau) \mid \tau = t_1, \dots, t_2, k_\tau = 1, \dots, |\mathcal{K}^\tau|\}$ Knoten und $A = \{((\tau', k_{\tau'}), (\tau'', k_{\tau''})) \mid dis^{\tau', \tau''}(C_{k_{\tau'}}^{\tau'}, C_{k_{\tau''}}^{\tau''}) \text{ ist 'klein genug'}\}$ Kanten zwischen den Knoten N , wobei 'klein genug' über dis_{ib} und evtl. Expertenentscheidungen festgelegt wird. Ein solcher Multi-Themen-Graph bildet definitionsgemäß thematische Ähnlichkeiten zwischen gefundenen Themen-Clustern unterschiedlicher Zeitfenster τ ab. Der Praktiker erhält so einen Einblick in thematische wie auch zeitliche Zusammenhänge größerer Themenkomplexe.

4.3 Mono-Themen-Graph

Der Multi-Themen-Graph G der Relation $Rel(t_1, t_2)$ eines Betrachtungshorizonts $[t_1, t_2]$ lässt sich in verschiedene Subgraphen aufgliedern, u.a. in sogenannte Mono-Themen-Graphen, die einen bestimmten zusammenhängenden Themenkomplex innerhalb des Multi-Themen-Graphen G der Relation $Rel(t_1, t_2)$ beschreiben.

Für jeden Knoten $(\tilde{\tau}, k_{\tilde{\tau}})$ (Themen-Cluster $C_{k_{\tilde{\tau}}}^{\tilde{\tau}}$) kann ein zugehöriger Mono-Themen-Graph $G_{(\tilde{\tau}, k_{\tilde{\tau}})}$ erstellt werden.

Dieser Subgraph $G_{(\tilde{\tau}, k_{\tilde{\tau}})} = (N_{(\tilde{\tau}, k_{\tilde{\tau}})}, A_{(\tilde{\tau}, k_{\tilde{\tau}})})$ besteht aus dem Knoten $(\tilde{\tau}, k_{\tilde{\tau}})$ und allen Knoten (τ, k_τ) , die über einen Pfad, bezeichnet als $(\tilde{\tau}, k_{\tilde{\tau}}) \xrightarrow{*} (\tau, k_\tau)$, von dem betrachteten 'Ausgangs'-Knoten $(\tilde{\tau}, k_{\tilde{\tau}})$ aus erreichbar sind, sowie allen dabei involvierten Kanten. Formal schreibt man

$N_{(\tilde{\tau}, k_{\tilde{\tau}})} = \{(\tilde{\tau}, k_{\tilde{\tau}})\} \cup \{(\tau, k_{\tau}) \mid \tau = t_1, \dots, t_2, k_{\tau} = 1, \dots, |\mathcal{K}^{\tau}|, (\tilde{\tau}, k_{\tilde{\tau}}) \xrightarrow{*} (\tau, k_{\tau})\}$
 und $A_{(\tilde{\tau}, k_{\tilde{\tau}})} = \{((\tau', k_{\tau'}), (\tau'', k_{\tau''})) \mid dis^{\tau', \tau''}(C_{k_{\tau'}}^{\tau'}, C_{k_{\tau''}}^{\tau''}) \text{ ist 'klein genug', } (\tau', k_{\tau'}) \in N_{(\tilde{\tau}, k_{\tilde{\tau}})}, (\tau'', k_{\tau''}) \in N_{(\tilde{\tau}, k_{\tilde{\tau}})}\}$.

Ein so erstellter Mono-Themen-Graph $G_{(\tilde{\tau}, k_{\tilde{\tau}})}$ entspricht einer Subrelation $Rel(t_1, t_2)_{(\tilde{\tau}, k_{\tilde{\tau}})}$ von $Rel(t_1, t_2)$ und beschreibt das Beziehungsgeflecht eines Themas $C_{k_{\tilde{\tau}}}^{\tilde{\tau}}$ zum Zeitpunkt t_2 mit einer im Allgemeinen rückwärtsgewandten Betrachtungsspanne mit $m_{t_2} = t_2 - t_1 + 1$ Zeitfenstern.

Für alle Knoten $(\tau, k_{\tau}) \in N_{(\tilde{\tau}, k_{\tilde{\tau}})}$ sind die Mono-Themen-Graphen identisch, da sie alle demselben Themenkomplex angehören.

Um Gesichtspunkte, dass sich mit fortschreitender Zeit Betrachtungszeiträume verschieben und somit die Berücksichtigung von Zeitfenstern verändern können, besonders zu betonen, spricht man auch von temporalen Themen-Graphen (s.a. Abbildung 4.7). Der Multi-Themen-Graph in Abbildung 4.2 besteht aus mindestens vier Mono-Themen-Graphen mit Knotenmengen von mehr als einem Knoten, die jeweils einen eigenen Themenkomplex darstellen und die entsprechende Subrelationen innerhalb dieses Themengeflechts widerspiegeln. Diese vier Graphen werden im Kapitel 4.4 mit ihren spezifischen strukturellen Eigenschaften detaillierter erläutert.

Graphen, die nur aus einem Knoten ohne Kanten bestehen, werden hier nicht näher betrachtet. Solche unvernetzten Knoten können als sogenannte 'Eintagsfliegen' angesehen werden, die für unsere Betrachtungen nicht weiter von Interesse sind. Unvernetzte Knoten an den Rändern des Betrachtungszeitraums t_1 und t_2 können Eintagsfliegen sein, aber auch das Ende bzw. der Beginn eines Themas. Eine Aussage dazu kann nur bei Vergrößerung bzw. Verschiebung des Betrachtungszeitraums getroffen werden.

Durch die Beschränkung auf vernetzte Knoten erhält man **mindestens bi-partite** und auf Grund der Betrachtungsspanne $m_{t_2} = t_2 - t_1 + 1$ **maximal m_{t_2} -partite Themen-Graphen**.

Bei der Beschränkung auf einzelne Mono-Themen-Graphen des Multi-Themen-Graphen wird anhand der exemplarischen Abbildungen 4.3, 4.4, 4.5, 4.6 (siehe Kapitel 4.4) deutlich, dass eine solche Fokussierung eine Komplexitätsreduktion zur Folge hat und einem Nutzer eine auf ein Thema bzw. Themengeflecht reduzierte vereinfachte Sichtweise erlaubt. Dadurch werden Zusammenhänge innerhalb eines Themenkomplexes besser und schneller erfassbar, auch solche, die einer manuellen Sichtung verborgen geblieben wären.

4.4 Strukturelle Eigenschaften in Themen-Graphen

Die Knoten (τ, k_{τ}) eines Themen-Graphen, sowohl eines Multi- (vgl. Kapitel 4.2) als auch eines Mono-Themen-Graphen (vgl. Kapitel 4.3), weisen in Bezug auf mit ihnen in Beziehung stehende Relationen im Zeitverlauf des Betrachtungshorizonts $[t_1, t_2]$ verschiedene charakteristische Eigenschaften auf.

Von **Vereinigung** (Merger) in einem Knoten (τ, k_τ) spricht man, wenn sich Kanten verschiedener Knoten $(\tau', k_{\tau'})$ aus der Vergangenheit (mit $\tau' < \tau$) im Knoten (τ, k_τ) vereinigen. Für die Anzahl aus der Vergangenheit eingehender Kanten in (τ, k_τ) gilt dann $|\{((\tau', k_{\tau'}), (\tau, k_\tau)) \mid \tau' < \tau\}| > 1$.

Der umgekehrte Fall heißt **Aufspaltung** (Split) und tritt auf, wenn ein Knoten (τ, k_τ) mit mehreren Knoten $(\tau'', k_{\tau''})$ in der Zukunft (mit $\tau'' > \tau$) Kanten besitzt. Für die Anzahl in die Zukunft weisender und von (τ, k_τ) ausgehender Kanten gilt dann $|\{((\tau, k_\tau), (\tau'', k_{\tau''})) \mid \tau'' > \tau\}| > 1$.

Als **absolut neu** innerhalb der Relation $Rel(t_1, t_2)$ bezeichnet man einen Knoten (τ, k_τ) , wenn kein Pfad von einem Knoten $(\tau', k_{\tau'})$ aus einem zeitlich zurückliegenden Zeitfenster $\tau' < \tau$ zum Knoten (τ, k_τ) existiert. Dann gilt $\{(\tau', k_{\tau'}) \xrightarrow{*} (\tau, k_\tau) \mid \tau' < \tau\} = \emptyset$.

Ein Knoten (τ, k_τ) gilt als **r-temporär neu** innerhalb der Relation $Rel(t_1, \tau + r)$, wenn keine Kante von einem Knoten aus einem zeitlich vorangehenden Zeitfenster $\tau' < \tau$ zum aktuell betrachteten Knoten (τ, k_τ) existiert, in $Rel(t_1, t_2)$ aber ein Pfad $(\tau', k_{\tau'}) \xrightarrow{*} (\tau, k_\tau)$ zwischen (τ, k_τ) über einen Knoten $(\tau'', k_{\tau''})$ in der Zukunft $\tau'' > \tau + r$ bzgl. $Rel(t_1, \tau + r)$ mit $\tau'' \leq t_2$ und einem Knoten $(\tau', k_{\tau'})$ in der Vergangenheit $\tau' < \tau$ gefunden werden kann, d.h. es existiert ein Pfad nach (τ, k_τ) aus der Vergangenheit τ' bezüglich τ über die Zukunft $\tau'' > \tau + r$ zur Gegenwart τ .

In der Relation $Rel(t_1, t_2)$ wird ein Knoten (τ, k_τ) **direkt alt** genannt, falls mindestens eine Kante zu einem bezüglich τ älteren Knoten $(\tau', k_{\tau'})$ in $\tau' < \tau$ existiert. Es gilt $|\{((\tau', k_{\tau'}), (\tau, k_\tau)) \mid \tau' < \tau\}| \geq 1$.

Ein **indirekt alter** Knoten ist äquivalent zu einem 'r-temporär neuen' Knoten.

Der zum Thema $C_{k_\tau}^\tau$ gehörige Knoten (τ, k_τ) ist **r-direkt erloschen**, falls für die Relation $Rel(t_1, \tau + r)$ keine Kante $((\tau, k_\tau), (\tau'', k_{\tau''}))$ mit $\tau < \tau'' = \tau + r \leq t_2$ existiert.

Ein Knoten (τ, k_τ) wird **r-temporär erloschen** genannt, falls für die Relation $Rel(t_1, \tau + r)$ kein Pfad $(\tau, k_\tau) \xrightarrow{*} (\tau'', k_{\tau''})$ von dem Knoten (τ, k_τ) zu einem Knoten $(\tau'', k_{\tau''})$ in der Zukunft $\tau'' = \tau + r \leq t_2$ existiert.

Ist ein Knoten (τ, k_τ) 'r-temporär erloschen' und gilt $r = t_2 - \tau$, so sprechen wir für die Relation $Rel(t_1, t_2)$ von einem **absolut erloschenen** Knoten bzw. Thema.

Die bereits angesprochenen Subgraphen des Multi-Themen-Graphen aus Abbildung 4.2 werden im Folgenden in den separaten Abbildungen 4.3, 4.4, 4.5, 4.6 einzeln vorgestellt und ihre strukturalen Eigenschaften erläutert.

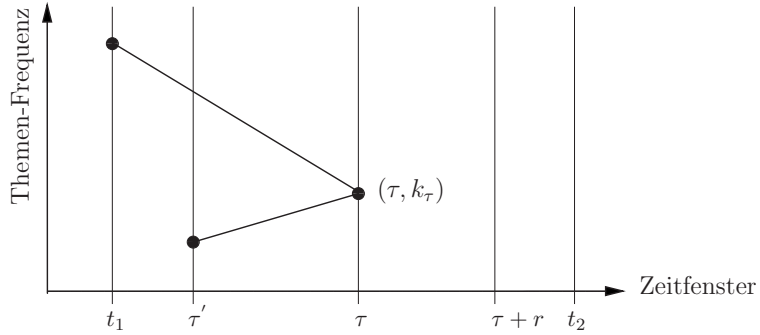


Abbildung 4.3: Mono-Themen-Graph.

Abbildung 4.3 zeigt, dass der Knoten (τ, k_τ) aus der Vereinigung der Knoten (t_1, k_{t_1}) und $(\tau', k_{\tau'})$ aus unterschiedlichen, zurückliegenden Zeitfenstern t_1 und τ' entsteht. Für den Knoten (τ, k_τ) existieren keine Knoten in der Zukunft $\tau + r \leq t_2$ bezogen auf τ , zu denen ein Pfad existiert. Der Knoten (τ, k_τ) ist damit 'r-temporär erloschen'. Da auch kein Pfad in die Zukunft bezüglich τ für $r = t_2 - \tau$ existiert, ist das Thema $C_{k_\tau}^\tau$ des Knoten (τ, k_τ) 'absolut erloschen'.

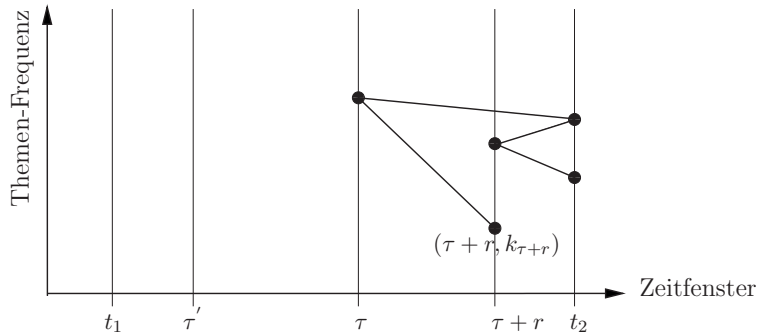


Abbildung 4.4: Mono-Themen-Graph.

Ein weiterer Subgraph des Multi-Themen-Graph aus Abbildung 4.2 ist in Abbildung 4.4 dargestellt. Dieser Mono-Themen-Graph hat seinen zeitlichen Ursprung in τ . Das Thema $C_{k_\tau}^\tau$ ist bezüglich unseres Betrachtungshorizonts $[t_1, t_2]$ 'absolut neu'. Es existieren keine Knoten $(\tau', k_{\tau'})$ in der Vergangenheit τ' , bezogen auf die Gegenwart τ des Knoten (τ, k_τ) , zu denen ein Pfad existiert.

Weiterhin ist zu erkennen, dass der Knoten $(\tau + r, k_{\tau+r})$ 'r-direkt erloschen' ist, da keine Kante von diesem Knoten aus in die Zukunft weist. Allerdings ist er weder 'r-temporär', noch 'absolut erloschen', da ein Pfad aus der Zukunft t_2 über die Vergangenheit τ zurück in die Gegenwart $\tau + r$ existiert.

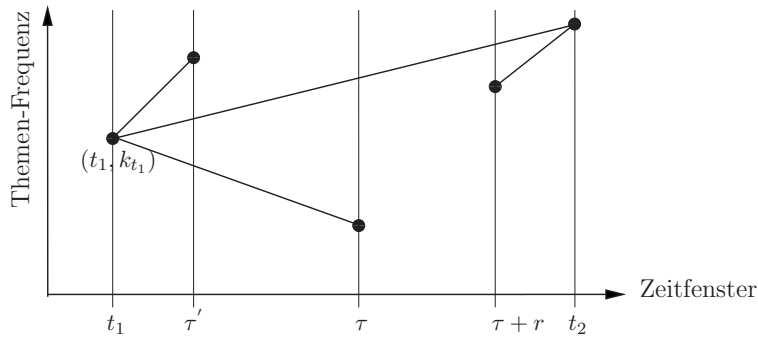


Abbildung 4.5: Mono-Themen-Graph.

In Abbildung 4.5 ist ein Themen-Cluster $C_{k_{t_1}}^{t_1}$, dargestellt als Knoten (t_1, k_{t_1}) , im Zeitfenster t_1 zu erkennen, der durch Aufspaltung in Relation zu drei zeitlich unterschiedlichen Clustern in den Zeitfenstern τ' , τ und t_2 steht. Der dem Mono-Themen-Graphen zugeordnete Knoten in $\tau + r$ ist als 'r-temporär neu' erkennbar, da keine, bezüglich seiner Gegenwart $\tau + r$, eingehenden Kanten zu älteren Knoten zu Zeitfenstern $\tau < \tau + r$ existieren. Allerdings ist Knoten $(\tau + r, k_{\tau+r})$ über einen bezüglich $\tau + r$ in der Zukunft liegenden Knoten in t_2 mit älteren Knoten in t_1 sowie τ' und τ verbunden.

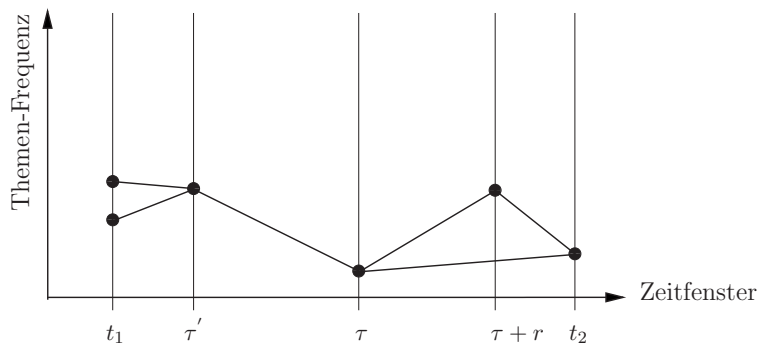


Abbildung 4.6: Mono-Themen-Graph.

Die Abbildung 4.6 zeigt, dass auch zyklische Strukturen im Relationsgeflecht eines Themen-Graphen vorkommen können.

Die vorgestellten strukturalen Eigenschaften in temporalen Themen-Graphen werden im Teil Evaluation in Kapitel 5.4.4 anhand eines ausgewählten realen Mono-Themen-Graphen (vgl. Abbildung 5.35) konkretisiert.

4.5 Evolution von temporalen Themen-Graphen

Die nachfolgende Abbildung 4.7 zeigt, wie ein Multi-Themen-Graph für die Ausgangsrelation $Rel(t_1, t_2)$ schrittweise in der Zeit 'weiterwandert'.

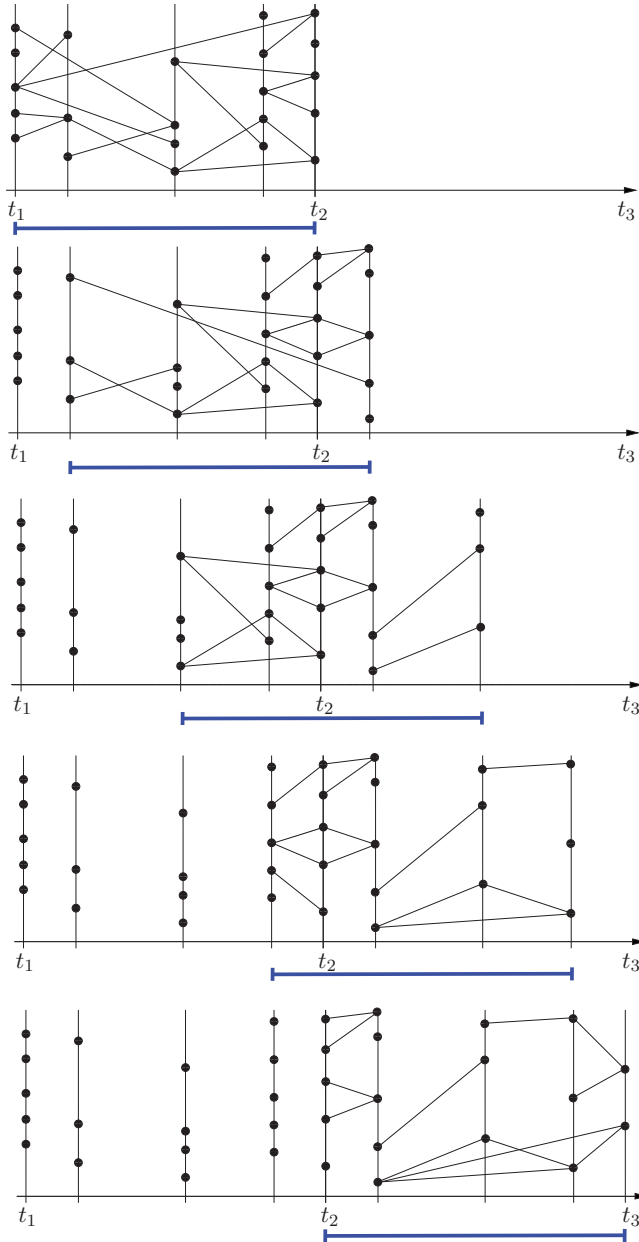


Abbildung 4.7: Evolution eines temporalen Multi-Themen-Graphen im Zeitverlauf.

Dabei entfallen 'alte' Kanten zu Knoten aus den jeweils 'ältesten' Zeitfenstern des Betrachtungszeitraums und neu erscheinende Knoten werden durch neue Kanten mit dem Graphen vernetzt.

Diese bzgl. der Zeitfenster fortschreitende Vorgehensweise ist mit einem vertretbaren Rechenaufwand verbunden, da für 'wandernde' Relationen $Rel(t_x, t_y)$ mit $x < y$ iterativ die zugehörigen Multi- wie auch Mono-Themen-Graphen erstellt werden. Man kann so für beliebig lange Zeiträume die Graphenevolution Stück für Stück verfolgen.

Eine vereinfachte Variante, sich einen Überblick über größere Betrachtungszeiträume zu verschaffen, ist die Konkatenation aufeinander folgender Relationen $Rel(t_x, t_y), Rel(t_y, t_z)$ mit $x < y < z$, die zu entsprechend größeren Themen-Graphen führt. Vergleiche dazu den Multi-Themen-Graphen in Abbildung 4.8, der durch Konkatenation des Multi-Themen-Graphen der Relation $Rel(t_1, t_2)$ und des Multi-Themen-Graphen der Relation $Rel(t_2, t_3)$ gebildet wird (vgl. Abbildung 4.7).

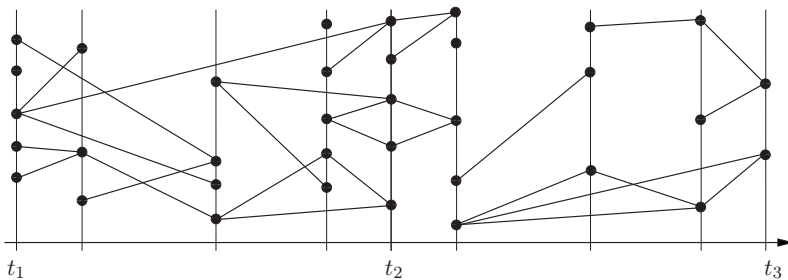


Abbildung 4.8: Multi-Themen-Graph einer Konkatenation der Relation $Rel(t_1, t_2)$ mit der sich ihr anschließenden Relation $Rel(t_2, t_3)$.

Ein Nachteil dieser Vorgehensweise ist allerdings, dass Relationsgrenzen überschreitende Kanten nicht erscheinen. Diese Kanten werden nur erfasst, wenn man stattdessen die Betrachtungsspanne $m_{t_y} = t_y - t_x + 1$, also den Betrachtungszeitraum $[t_x, t_y]$ entsprechend vergrößert (siehe Abbildung 4.9). Die durch die Erweiterung (Vergrößerung der Betrachtungszeitspanne) hinzukommenden Kanten sind in der Abbildung blau gestrichelt markiert.

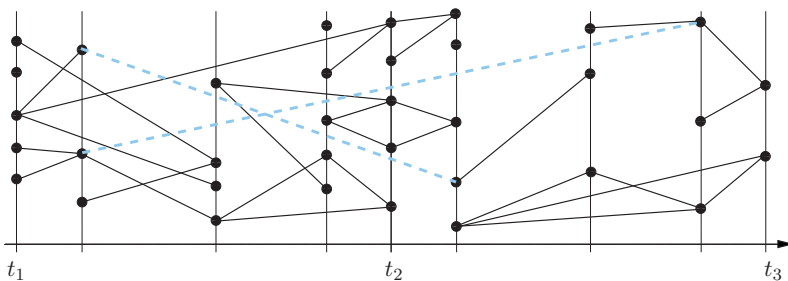


Abbildung 4.9: Multi-Themen-Graph der Relation $Rel(t_1, t_3)$ mit auf $m_{t_3} = t_3 - t_1 + 1$ vergrößerter Betrachtungszeitspanne.

Eine Erweiterung des Betrachtungshorizonts $[t_1, t_2]$ vergrößert tendenziell natürlich einen Multi- wie auch seine Mono-Themen-Graphen um weitere Relationen. Dies kann bei großen Betrachtungszeiträumen wegen der zusätzlichen Kanten zu Unübersichtlichkeiten führen, aber auch zu einem besseren Verständnis der temporalen Zusammenhänge innerhalb der Themenkomplexe.

Die Vergrößerung des Betrachtungshorizonts $[t_y - m_{t_y} + 1, t_y]$ wird allerdings durch den steigenden Rechenaufwand begrenzt. Für die Relation $Rel(t_y - m_{t_y} + 1, t_y)$ müssen $m_{t_y} \times (m_{t_y} - 1)/2$ Relationsmatrizen berechnet werden.

Abschließend wird an zwei Subgraphen der vergrößerten Relation $Rel(t_1, t_3)$ aus Abbildung 4.9 der Unterschied zwischen einer Konkatenation und der entsprechenden Vergrößerung der Betrachtungsspanne gezeigt.

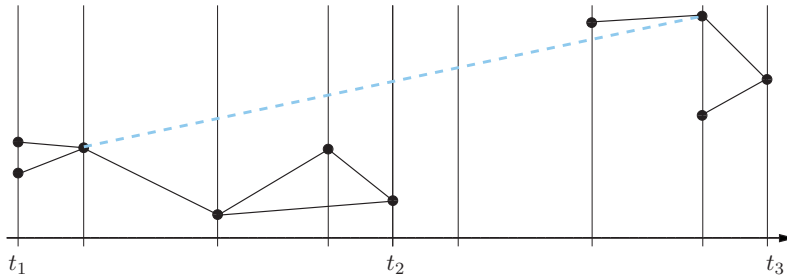


Abbildung 4.10: Mono-Themen-Graph der vergrößerten Relation $Rel(t_1, t_3)$.

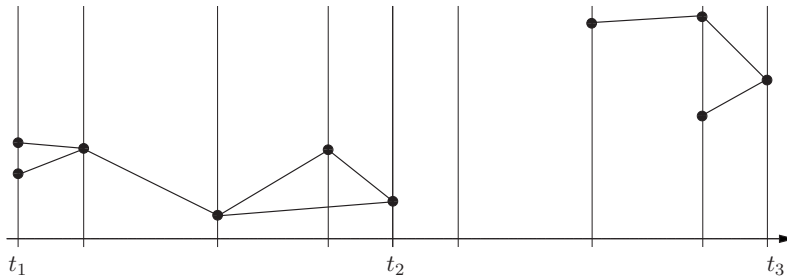


Abbildung 4.11: Zerfall eines Mono-Themen-Graph bei Konkatenation.

Die Mono-Themen-Graphen in den Abbildungen 4.10 und 4.12 des Betrachtungshorizonts $[t_1, t_3]$ zerfallen bei Konkatenation der Relationen $Rel(t_1, t_2)$ und $Rel(t_2, t_3)$ in jeweils zwei Mono-Themen-Graphen (vgl. Abbildungen 4.11 und 4.13), da der temporale Zusammenhang dieser zwei Themenkomplexe über die Relationsgrenze t_2 hinweg nicht bekannt ist.

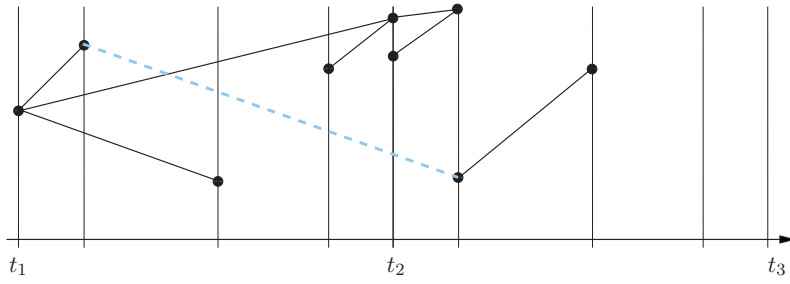


Abbildung 4.12: Mono-Themen-Graph der vergrößerten Relation $Rel(t_1, t_3)$.

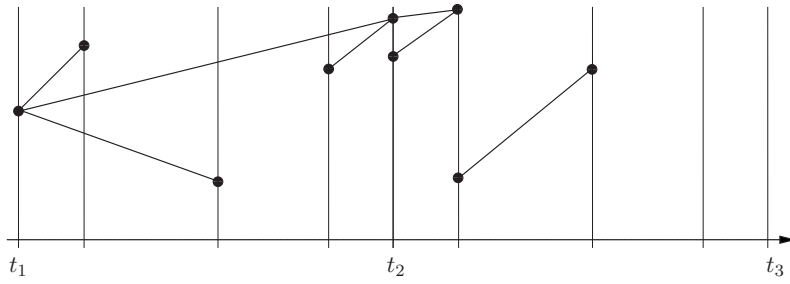


Abbildung 4.13: Zerfall eines Mono-Themen-Graph bei Konkatination.

Kapitel 5

Evaluation

5.1 Implementierung in MATLAB und MSSQL

Das in den Kapiteln 3 und 4 vorgestellte Modell wurde implementiert in MATLAB® (R2012b) von MathWorks.

Dabei bietet MATLAB mit seiner höheren Programmiersprache eine interaktive Umgebung für numerische Berechnungen, zur Visualisierung und Programmierung. MATLAB dient in Wissenschaft und Forschung zur Datenanalyse, Algorithmen-Entwicklung und zur Erstellung von Modellen und Anwendungen. Mit der Programmiersprache, den Tools und den integrierten mathematischen Funktionen lassen sich verschiedene Ansätze ausprobieren, die schneller zu einer Lösung führen als herkömmliche Programmiersprachen wie C/C++ oder Java™. Vergleiche dazu <http://www.mathworks.de/products/matlab> (abgerufen Sept. 2014).

Der mittels MATLAB modular aufgebaute Prototyp ruft der Reihe nach von dem Modell abgeleitete Unterfunktionen auf. Dabei stehen unter anderem die in MATLAB integrierten Algorithmen k -means sowie die hierarchischen Clusterverfahren aus der *Statistic Toolbox* zur Verfügung. Alle übrigen in den Kapiteln 3 und 4 vorgestellten Berechnungen und Verfahren wurden im Rahmen dieser Arbeit für die Modellimplementierung in MATLAB programmiert.

Die Analysedaten können vom Prototypen entweder direkt von der Festplatte eingelesen oder über eine Schnittstelle (Java Database Connectivity (JDBC)) aus einer MSSQL-Datenbank abgerufen werden. Der Microsoft SQL Server (MSSQLServer) ist ein relationales Datenbankmanagementsystem von Microsoft für große Datenmengen.

5.2 Rechner-Konfigurationen

Zur Evaluation mittels MATLAB von MathWorks standen drei verschiedene zeitgemäße Rechner-Konfigurationen zur Verfügung:

- 1) Standard Desktop PC
Intel Core 2 Quad Q6600 2,4 GHz (4-Kern Prozessor)
8 GB Ram
64 Bit-Betriebssystem
Windows Vista Business SP1

- 2) Standard Desktop PC
AMD Phenom(tm) X4 945 Processor 3.01 GHz (4-Kern Prozessor)
16 GB Ram
64 Bit-Betriebssystem
Windows 7 Professional SP1

- 3) Standard Laptop PC
Intel Core i5 2.67 GHz (4-Kern Prozessor (davon 2 simuliert))
8 GB Arbeitsspeicher
64 Bit-Betriebssystem
Windows 7 Professional SP1

Die gesamte Evaluierung des erstellten Modells ist auf gängigen Rechnersystemen mit akzeptablem Zeitaufwand durchführbar. Dabei ist es möglich, eine Fülle an relevanten Dokumenten über eine längere Zeitspanne zu monitoren.

5.3 Testdaten

Es gibt in der Literatur keinen einheitlichen Datensatz, der als Referenz im Text Mining dient. So setzen zum Beispiel RAJARAMAN/TAN (2001) Nachrichtenartikel der Webseiten CNET und ZDNet ein, während bei TAI ET AL. (2002) die Dokumentensammlungen Medlin und Cranfield Anwendung finden. KOGAN ET AL. (2003) arbeiten mit medizinischen Abstracts der Medlars Collection, Information Science Abstracts der CISI Collection und Aerodynamics Abstracts des Cranfield Corpus. WEI/LEE (2004) nutzen Artikel der Seite excite.com. MOONEY/BUNESCU (2005) wiederum nutzen unter anderem biomedizinische Abstracts (Medlin Corpus), Stellenausschreibungen und Produktbeschreibungen (Amazon Buchbeschreibungen). TERACHI ET AL. (2006) testen mittels japanischsprachigen Artikeln des Journal of the Japanese Society for Quality Control für die Jahre 1995 bis 2000. CHEN ET AL. (2007) setzen Dokumente von Nachrichtenseiten ein, die der Washington Post, Reuters und CNN entstammen. Bei HE ET AL. (2007) werden Newsdokumente des Reuters Corpus mit einem Jahr Spanne genutzt. PONS-PORRATA ET AL. (2007) werten einen für den Forschungsbereich Topic Detection and Tracking (TDT) erstellten TDT Corpus des Linguistic Data Consortium aus, der allerdings, außer für Mitglieder, kostenpflichtig ist. Die Arbeit von LI ET AL. (2008) basiert hauptsächlich auf dem Reuters Corpus, dem Classic Dataset (Smart) und dem Corpus of the Text Retrieval Conference (TREC). Ein weiterer verwendeter Datensatz ist zum Beispiel die sogenannte Vegemite Database bei CAI ET AL. (2008). Artikel eines internetbasierten Newsletters des Hotel- und Gastgewerbes dienen WAGNER ET AL. (2009) als Testgrundlage. RAJAN ET AL. (2009) setzen auf einen tamilischen Datensatz und GANG ET AL. (2011) auf die English Gigaword Fourth Edition from the Linguistic Data Consortium. SEGEV/KANTOLA (2012) testen anhand Patenten des USA Patent and Trademark Office.

Die vorangehend aufgeführten Beispiele zeigen, dass sich kein standardisierter Testdatensatz herausgebildet hat.

In der vorliegenden Arbeit werden für die Evaluation des erstellten Modells Stichproben von zwei verschiedenen Datensätzen verwendet, die im Folgenden einzeln detailliert vorgestellt werden.

5.3.1 IDS-Mannheim Datensatz

Das Institut für Deutsche Sprache (IDS) ist die zentrale Einrichtung zur Erforschung und Dokumentation der deutschen Sprache mit Sitz in Mannheim. Sie ist Mitglied der Leibniz-Gemeinschaft und hat als Ziel, eine empirische Grundlage für germanistisch-sprachwissenschaftliche Forschung zu schaffen. Für wissenschaftliche Recherchen stellt sie unter anderem das Deutsche Referenzkorpus DeReKo zur Verfügung (vgl. u.a. KUPIETZ (2005), KUPIETZ/KEIBEL (2009), KUPIETZ ET AL. (2010)).

Das Deutsche Referenzkorpus DeReKo ist eine umfangreiche Sammlung deutschsprachiger Texte aus der Gegenwart und jüngeren Vergangenheit.

Inzwischen (Stand 2013) besteht DeReKo aus über 6 Milliarden Wörtern und stellt somit laut IDS die größte deutschsprachige Dokumentensammlung weltweit dar. Ein Großteil der Dokumente stammt von Zeitungsartikeln (vgl. Abbildung 5.1).



Abbildung 5.1: Quellen für DeReKo (Quelle: IDS Institut für Deutsche Sprache).

DeReKo besteht ausschließlich aus Copyright (urheberrechtlich) geschütztem Textmaterial. Die Nutzungsmöglichkeit, auch für wissenschaftliche Zwecke, ist daher teilweise eingeschränkt, da das IDS nicht Rechteinhaber der Texte in DeReKo ist und nur begrenzte Nutzungsrechte der über 150 Lizenzgeber bekommen hat (vgl. KUPIETZ/KEIBEL (2009)).

Dieser Arbeit liegt als Testdatensatz eine Stichprobe des DeReKo vor. Sie besteht aus vorkategorisierten Dokumenten aus fünf Themengebieten (Politik-Inland, Sport-Fußball, Staat-Gesellschaft & Familie-Geschlecht, Technik-Industrie & Transport-Verkehr, Wirtschaft-Finanzen & Öffentliche-Finanzen). Sie erstreckt sich über die Jahre 2004 bis 2008, also einen Zeitraum von 5 Jahren, mit jeweils ca. 20 Dokumenten pro Jahr für die genannten Themen.

Auf Grund der Copyright Bestimmungen werden die Dokumente nicht in ihrer Originalfassung, sondern als Frequenzlisten vom IDS zur Verfügung gestellt.

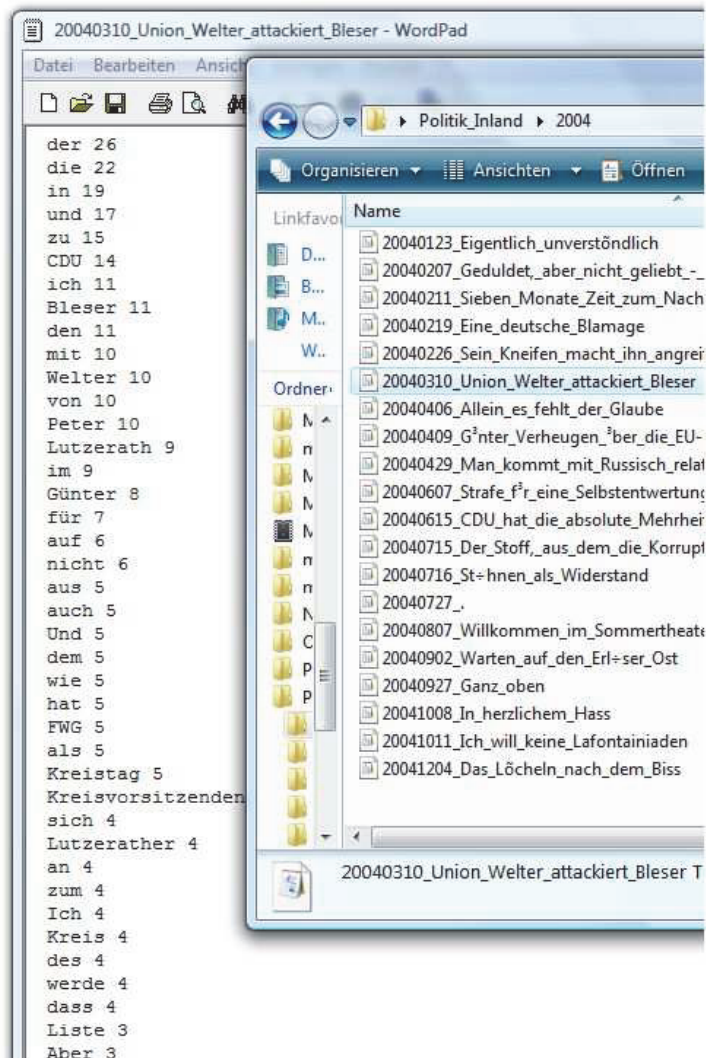


Abbildung 5.2: Beispiel Testdokumente aus dem Bereich Politik.

Abbildung 5.2 zeigt ein Beispiel aus dem Themenbereich Politik-Inland. Das rechte Fenster (Dateiexplorer) zeigt vorklassifizierte Zeitungsartikel aus dem Bereich Politik für das Jahr 2004. Im linken Fenster (WordPad) ist ein Ausschnitt aus der Frequenzliste eines Zeitungsartikels (hier: "Union Welter attackiert Bleser" vom 10. März 2004) geöffnet.

5. Evaluation

Zusätzlich zu den Testdokumenten wurde vom IDS auch ein Wörterbuch mit 2 Millionen Wortformen einschließlich der Generellen Termfrequenz ihres Auftretens und ihrer Inversen Dokumentenfrequenz zur Verfügung gestellt. Das Wörterbuch, ebenso die angegebenen Referenzwerte wurden aus dem Deutschen Referenzkorpus DeReKo erstellt. Für die Stoppwortbereinigung wird eine im Internet allgemein verfügbare Stoppwortliste eingesetzt.

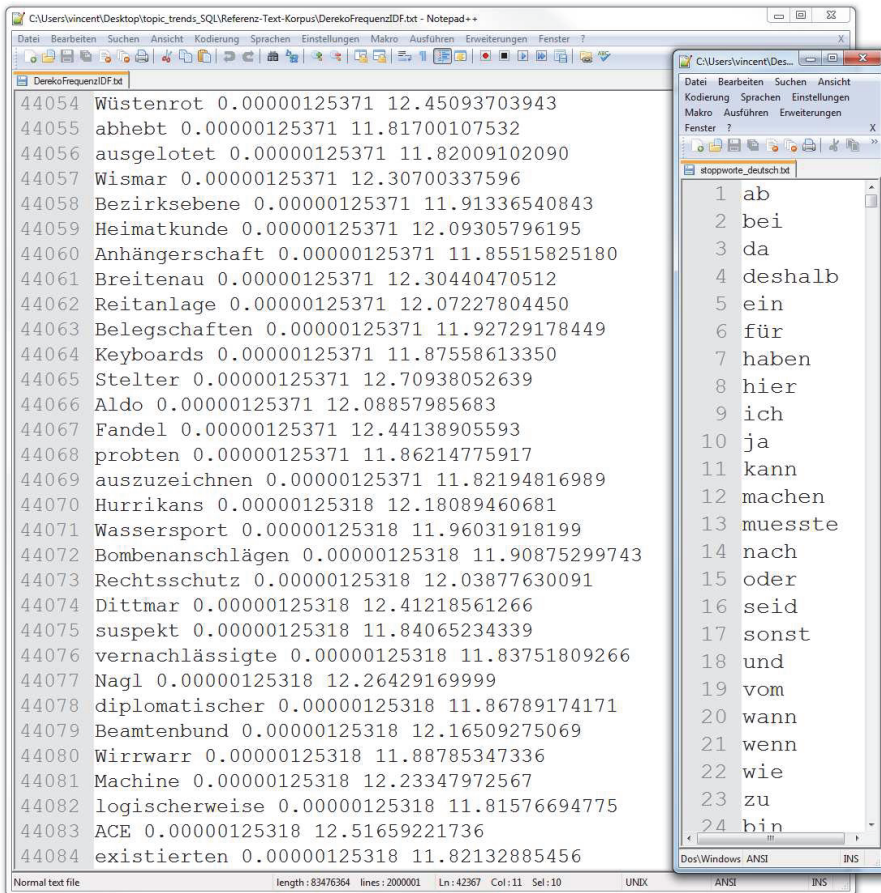


Abbildung 5.3: Deutsches Wörterbuch (mit Genereller Termfrequenz und Inverser Dokumentenfrequenz) & Stoppwortliste.

Das linke Fenster (DerekoFrequenzIDF.txt) zeigt einen Ausschnitt aus dem Wörterbuch. Von links nach rechts aufgeführt sind: Position einer Wortform im Wörterbuch (abhängig von der Generellen Termfrequenz), Wortform, Generelle Termfrequenz und Inverse Dokumentenfrequenz. Das rechte Fenster (stoppworte_deutsch.txt) zeigt einen Ausschnitt aus einer Stoppwortliste.

5.3.2 Spiegel Online (SPON) Datensatz



Abbildung 5.4: Spiegel Online (Quelle: spiegel.de).

Der zweite Datensatz, der dieser Arbeit zu Grunde liegt, besteht aus Spiegel Online Dokumenten. Die Artikel wurden über einen längeren Zeitraum gecrawlt, um einer Überlastung des Internetauftritts des Spiegel vorzubeugen.

Der Crawler erfasst dabei nur solche Dokumente, die auch bei manueller Sichtung gefunden werden können. Das heißt, ausgehend von der Startseite Spiegel.de folgt der Crawler allen verlinkten Unterseiten der Domain Spiegel.de. Online verfügbare Artikel, zu denen es keine Verlinkungen gibt, bleiben unberücksichtigt. Somit stellt die Menge aller vom Crawler erfassten Nachrichtenartikel eine Stichprobe des Spiegel Online Datensatzes dar. Die Dokumente einschließlich ihrer Metadaten wurden in eine MSSQL-Datenbank abgespeichert. Metadaten aus dem HTML-Code (Hypertext Markup Language) der jeweiligen Dokumentenseite sind beispielsweise time-stamp (Datum der Veröffentlichung), URL (Uniform Resource Locator), Headline1 (Titel), Headline2 (Untertitel), Topic1 (Rubrik, z.B. Politik), Topic2 (Unterrubrik, z.B. Politik → Inland/Ausland), etc. (vgl. Abbildung 5.5).

Über eine Schnittstelle werden die so erlangten Dokumente aus der MSSQL-Datenbank zur Analyse für MATLAB (vgl. Kapitel 5.1) bereitgestellt.

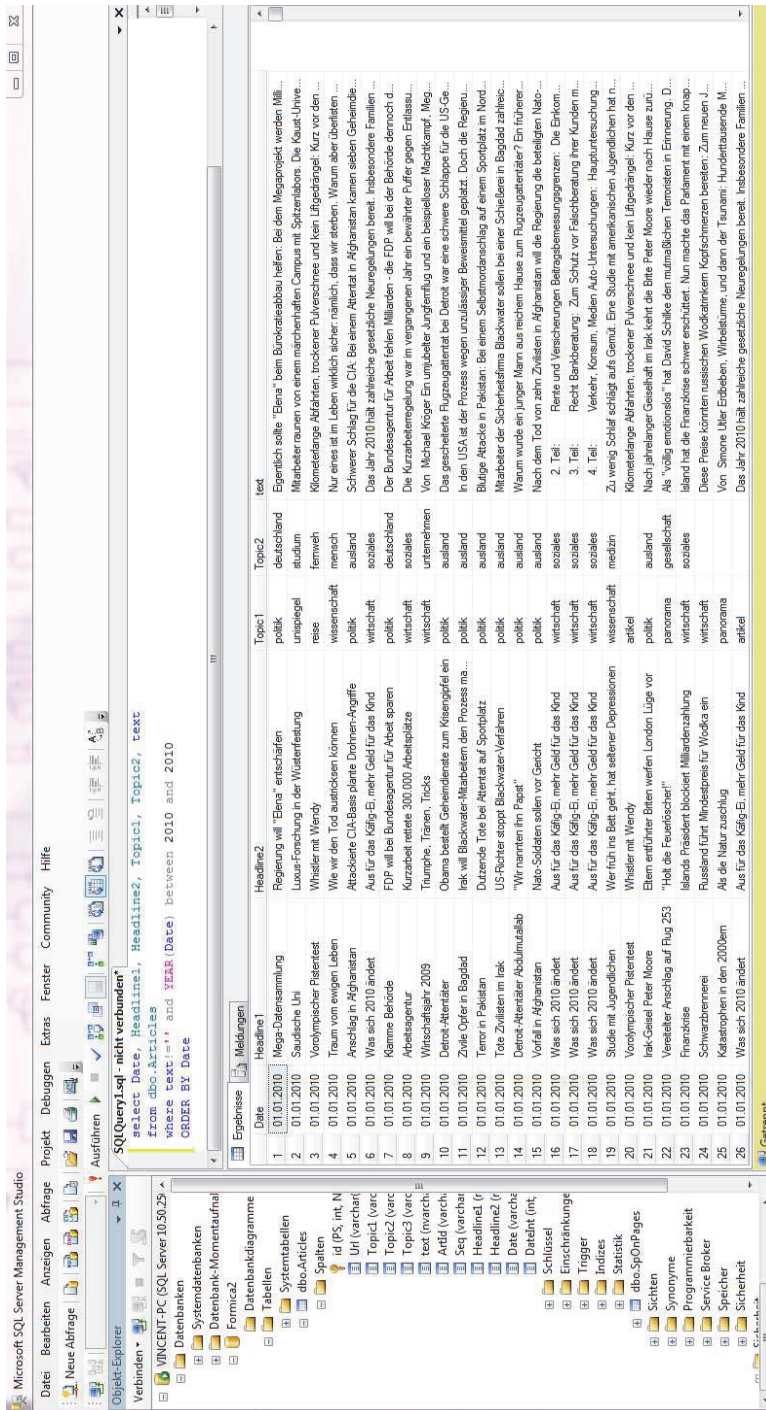


Abbildung 5.5: Ausschnitt SQL Daten von Spiegel Online.

Abbildung 5.6 zeigt die zeitliche Entwicklung (Anzahl Dokumente) der in der MSSQL-Datenbank enthaltenen Dokumente der Jahre 1990 bis 2010.

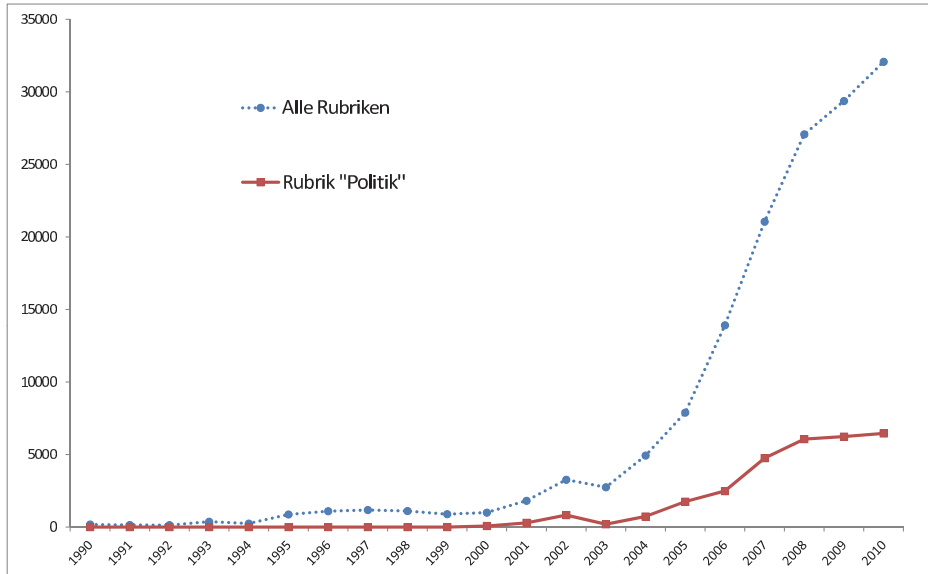


Abbildung 5.6: Spiegel Online - Zeitliche Entwicklung.

Das Gros der online verfügbaren Dokumente bis zur Jahrtausendwende (vgl. die blaue Kurve in Abbildung 5.6 für die Jahre 1994 bis 2000) besteht hauptsächlich aus für den online Bereich Spiegel.de (nach-) digitalisierten Artikeln des 'Spiegel Print' sowie des 'Spiegel Spezial' Mediums. Laut Spiegel Gruppe zeichnet sich das 'Spiegel Print' Medium, also die Wochenausgabe des Spiegel durch fundierteren tiefgründig recherchierten Journalismus aus. Besondere Themen, die der Spiegel meist als Serie erfolgreich behandelt hat, werden mit Zusatzmaterial als 'Spiegel Spezial' veröffentlicht.

Ab der Jahrtausendwende stieg die Zahl der Dokumente für den online Bereich zunächst zögerlich und ab etwa 2003 rasant an. Die Zahl, der über die Spiegelredaktion verarbeiteten Dokumente, lässt die noch junge Entwicklung des Internets mit ihrer in den letzten Jahren stark beschleunigten Zunahme an Veröffentlichungen unterschiedlichen Inhalts sichtbar werden. Daraus lässt sich unmittelbar die enorm gestiegene Wichtigkeit des Mediums Internet ableiten (vgl. Abbildung 1.1).

Etwa ab der Jahrtausendwende wird für den online Bereich des Spiegel extra generierter Content (Artikel) einzelnen Rubriken, wie z.B. Politik, gesondert zugeordnet (vgl. Abbildung 5.6 rote Kurve).

5.4 Testreihen

In diesem Kapitel werden die bisher beschriebenen Modellüberlegungen an mehreren konkreten Fallbeispielen validiert. In jeder der aufeinander folgenden Testreihen wird dabei ein anderer Aspekt der Graphenentwicklung besonders beleuchtet.

5.4.1 Testreihe I.

Grundlage für die in diesem Kapitel vorgestellten Tests ist eine Datenstichprobe des DeReKo. Als Testdokumente dienen Zeitungsartikel aus den drei Themengebieten Politik-Inland, Sport-Fußball, Technik-Industrie & Transport-Verkehr. Der Zeitraum der erfassten Zeitungsartikel beläuft sich auf die Jahre 2004 bis 2008. Pro Jahr und Themengebiet ergeben sich ungefähr 20 datierte Artikel. Die Zeitungsartikel liegen auf Grund von Copyright Bestimmungen als Frequenzlisten vor, die durch das IDS vorkategorisiert sind. Diese vorkategorisierten Texte werden in einem Pool zusammengeführt, das ein Analysekorpus von 276 Texten ergibt.

Außer den eigentlichen Testdokumenten (Zeitungsartikeln) des Analysekorpus wird die Generelle Termfrequenz und die Inverse Dokumentenfrequenz, berechnet aus dem DeReKo, für die Testläufe verwendet.

Die Tests überprüfen Modellgrundlagen und vergleichen verschiedene Ansätze aus Kapitel 3. Die Ergebnisvergleiche mit den Vorkategorisierungen des IDS erlauben eine objektive und nachvollziehbare Bewertung. Durchgeführt wurde der Test mit der Rechner-Konfiguration 1 (vgl. Kapitel 5.2).

Im Einzelnen werden folgende Parameterkonfigurationen getestet:

- Verkleinertes Lokales Wörterbuch \mathcal{L}' ohne Stoppwortentfernung mit $|\mathcal{L}'| = Z \in \{2.000, 20.000, 200.000, 2.000.000\}$
- Distanzmaß (euklidisches Maß vs. Cosinusmaß) (vgl. Kapitel 3.2)
- Merkmalsextraktion (vgl. Formel (3.5))
- Clusterverfahren (k -means mit $|\mathcal{K}| = 3$, vgl. Kapitel 3.3.2)

Testergebnisse

Einfluss des Parameters Z :

Der Parameter Z hat wesentlichen Einfluss sowohl auf die sogenannte Prozentuale Worterkennungsrate (vgl. Formel (3.7) in Kapitel 3.1.4), als auch auf die Programmlaufzeit. Je größer die Dimension Z ist, desto höher fällt die Worterkennungsrate aus. Bereits bei der Dimension 200.000 wird eine sehr gute Worterkennungsrate von 96 % erreicht. Die Vergrößerung der Dimension um das 10-fache ergibt lediglich einen Zuwachs der Worterkennungsrate von 3,5 % (siehe Abbildung 5.7).

Eine hohe Worterkennungsrate gewährleistet, dass bei der Bearbeitung der Texte durch das System kaum Informationsverluste auftreten.

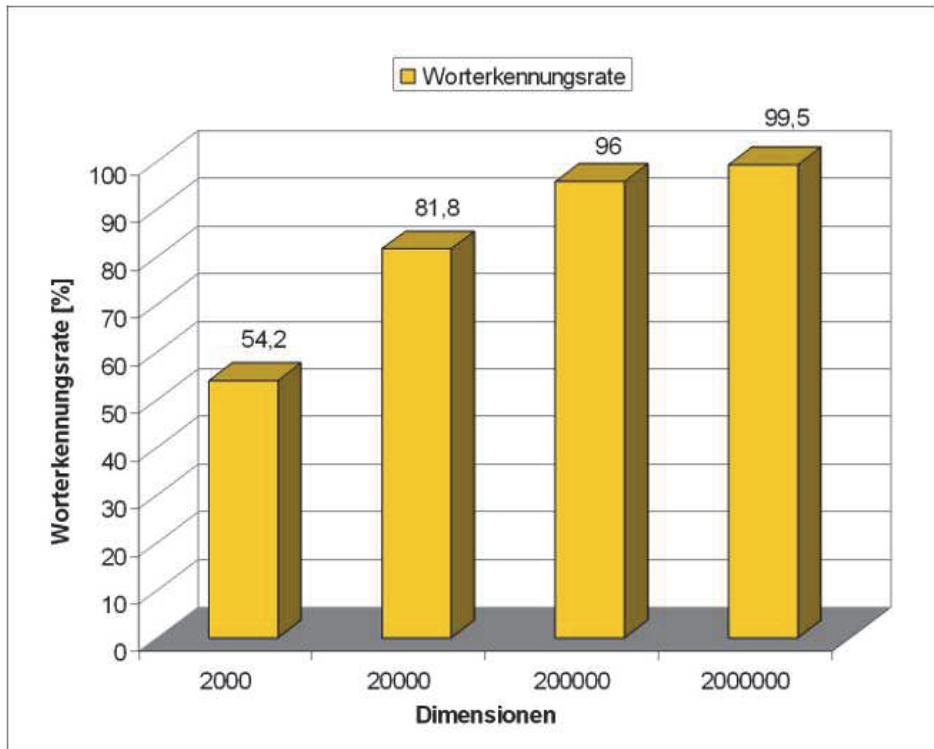


Abbildung 5.7: Worterkennungsrate nach Formel (3.8).

Die Programmlaufzeit ist ebenfalls stark von der Dimension abhängig. Bei der Dimension 200.000 braucht das Programm 15 Minuten. Für 2.000.000 benötigt das Programm bereits 150 Minuten, also die 10-fache Zeit (siehe Abbildung 5.8) bezogen auf eine Dokumentenanzahl von 276 Texten. Bei einer Vergrößerung bzw. Verringerung der Dokumentenanzahl vergrößert bzw. verkürzt sich entsprechend auch die Programmlaufzeit.

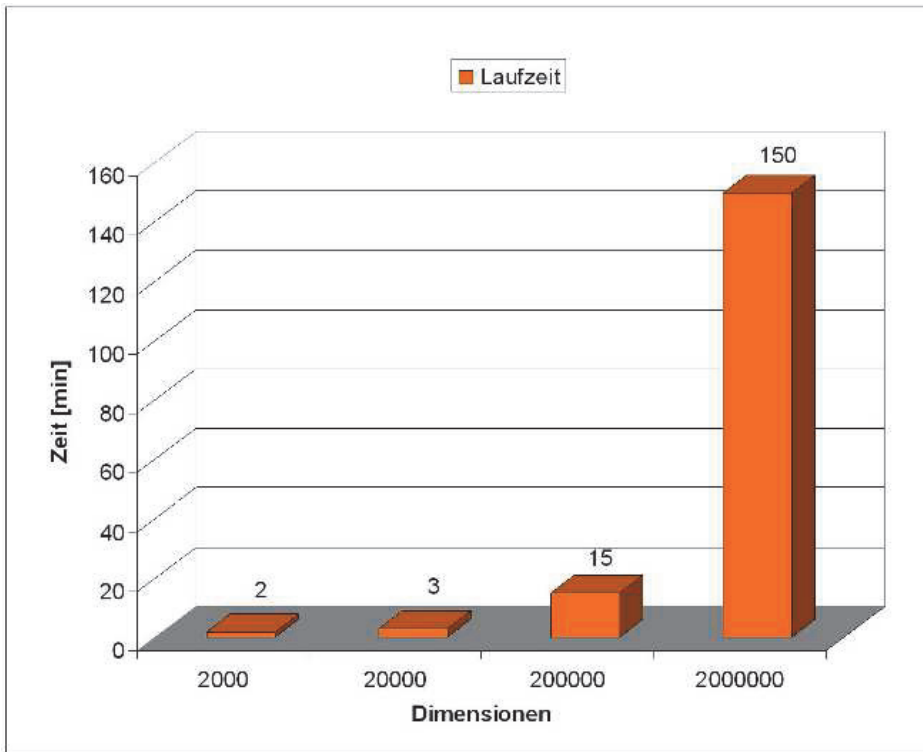


Abbildung 5.8: Laufzeit.

Für eine brauchbare Worterkennungsrate mit angemessenem Zeitaufwand (Rechenaufwand), eignet sich die Dimension $Z = 20.000$.

Einfluss des Distanzmaßes:

Die Testdokumente sind den drei Themengebieten Politik-Inland, Sport-Fußball, Technik-Industrie & Transport-Verkehr eindeutig vom IDS zugeordnet worden. Durch das Zusammenfassen aller Texte in nur einem gemeinsamen Pool und das anschließende Clustern durch den Prototypen, werden die Texte neu zu $|\mathcal{K}| = 3$ Themengebieten gruppiert.

Der Vergleich der Clusterergebnisse mit den Ausgangsgruppierungen des IDS ergibt die Cluster-Zuordnungsrate. Sie gibt den Grad der Übereinstimmung des Clusterergebnisses mit den Ausgangsgruppierungen des IDS wieder.

Je höher die Cluster-Zuordnungsrate ist, desto besser hat der Prototyp die Themengebiete (Topics) erkannt (vgl. dazu Abbildung 5.9).

In der folgenden Abbildung 5.9 ist der Einfluss des Distanzmaßes in Abhängigkeit der Dimension Z für die euklidische Distanz und das Cosinusmaß, als gängige Distanzen, aufgeführt. Man kann erkennen, dass das Cosinusmaß, wie aus der Literatur bekannt, der euklidischen Distanz im Text Mining weit überlegen ist (vgl. Kapitel 3.2).

Dimensionen	Maß	Laufzeit	Cluster-Zuordnungsrate
2.000	Cosinus-Maß	2 min	91,7%
20.000	Cosinus-Maß	5 min	95 %
200.000	Cosinus-Maß	17 min	90 %
2.000.000	Cosinus-Maß	155 min	93,4 %
2.000	Euklidische Distanz	2 min	46,7%
20.000	Euklidische Distanz	4 min	kein sinnvolles Ergebnis erkennbar.
200.000	Euklidische Distanz	18 min	kein sinnvolles Ergebnis erkennbar.
2.000.000	Euklidische Distanz	-	nicht getestet.

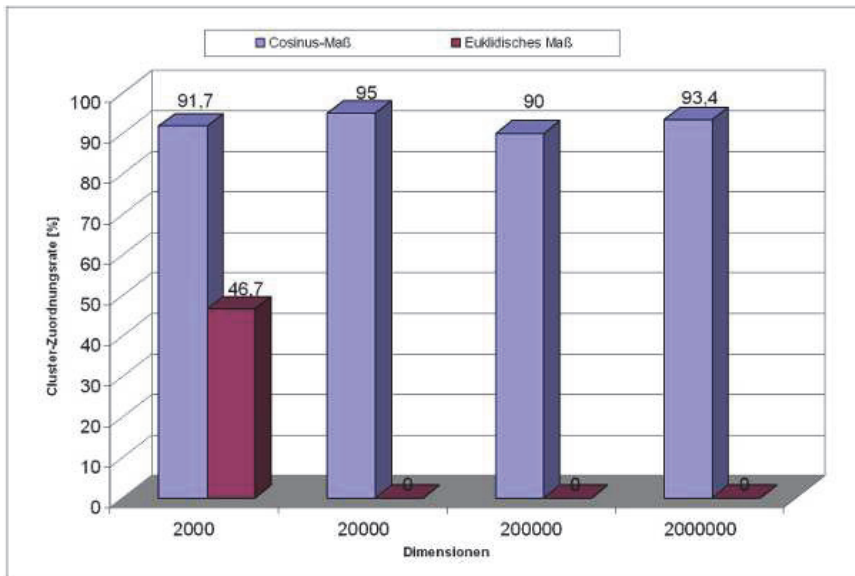


Abbildung 5.9: Cluster-Zuordnungsrate (euklidische Distanz vs. Cosinusmaß).

5.4.2 Testreihe II. (GfKI & DAGM & IFCS (2011))

In der Testreihe II. wird überprüft, ob ein Tracking von Themen mittels Relationsmatrizen möglich ist. Dabei werden Relationen nur für direkt aufeinander folgende Clusterings berechnet und nur eindeutige Matchings von Themen zugelassen.

Mehrfach-Zuordnungen von Clustern, wie auch die Veränderungen der Betrachtungsspannen und die Erstellung von Relationsgraphen werden in den nachfolgenden Testreihen III. und IV. detailliert beschrieben.

Der Testdatensatz ist eine Stichprobe des Datensatzes DeReKo. Die Testdokumente, bestehend aus Artikeln namhafter Printmedien einschließlich ihrer Zeitstempel, sind durch das Institut für Deutsche Sprache in die vier Themen Politik-Inland (P), Sport-Fußball (S), Technik-Industrie & Transport-Verkehr (TIT), und Wirtschaft & Finanzen (WF) kategorisiert, und können fünf aufeinander folgenden Zeitfenstern zugeordnet werden.

Die Testkonfiguration mit 254 Dokumenten ist in Tabelle 5.1 gegeben. Einer Unterstichprobe für das Zeitfenster $\tau = 1$, bestehend aus 52 Dokumenten, können drei Themen des IDS zugeordnet werden. Zeitfenster $\tau = 2$ und $\tau = 3$ enthalten Teilstichproben aus 65 und 50 Dokumenten aus vier bzw. drei Themen. Den Zeitfenstern $\tau = 4$ und $\tau = 5$ können zwei beziehungsweise drei Themen zugewiesen werden.

$ D^\tau $	52	65	50	35	52
Zeitfenster τ	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$
Cluster 1	C_1^1	C_1^2	C_1^3	C_1^4	C_1^5
Cluster 2	C_2^1	C_2^2	C_2^3	C_2^4	C_2^5
Cluster 3	C_3^1	C_3^2	C_3^3	-	C_3^5
Cluster 4	-	C_4^2		-	-

Tabelle 5.1: Testkonfiguration

Am Ende der nachfolgenden Ausführungen wird geklärt, welche konkreten Themen durch die $C_{k_\tau}^\tau$ -Notationen aus Tabelle 5.1 codiert werden.

Dem Test liegt ein Lokales Wörterbuch \mathcal{L} mit $|\mathcal{L}| = 2.000.000$ Wortformen zu Grunde. Die Testdurchläufe beschränken sich auf ein Verkleinertes Lokales Wörterbuch \mathcal{L}' mit $|\mathcal{L}'| = Z$ mit den $Z \in \{200, 2.000, 20.000\}$ häufigsten Wortformen des Lokalen Wörterbuchs \mathcal{L} als Dimensionen für den 'Vector Space'.

Mit folgenden Parameterkonfigurationen werden die Tests durchgeführt:

- Verkleinertes Lokales Wörterbuch \mathcal{L}' ohne Stopwortentfernung mit $|\mathcal{L}'| = Z \in \{200, 2.000, 20.000\}$
- Cosinusmaß als Distanzmaß (vgl. Kapitel 3.2)
- Merkmalsextraktion (vgl. Formel (3.4))
- Hierarchisches Clusterverfahren (Ward-Verfahren, vgl. Formel (3.24) Kapitel 3.3.3)
- Relationsmatrizen mit Relationsschranken $dis_{lb} = 0,4$ und $dis_{ub} = 0,55$

Die Testdokumente sind in Zeitfenster nach ihrer Datierung eingeteilt. Ein hierarchisches Clusterverfahren (Ward-Verfahren) erstellt für jedes Zeitfenster eine Indizierte Hierarchie. Durch visuelle Inspektion des daraus resultierenden Dendrogramms wird die Anzahl Themen innerhalb eines Zeitfensters bestimmt.

Mit (Un-)Ähnlichkeitsmatrizen zwischen gefundenen Themen aufeinander folgender Zeitfenster und der Schranken-Überlegung (vgl. Kapitel 4.1) wird ein Tracking durchgeführt. Ziel ist, über ein Monitoring die zeitliche Entwicklung gefundener Themen zu verfolgen und neu auftkommene Themen bzw. verschwundene Themen zu bestimmen.

Übergang zwischen den Zeitfenstern $1 \rightarrow 2$

Die Grafiken 5.10 (a) und (b) zeigen die Dendrogramme der beiden Clusterings in $\tau = 1$ und $\tau = 2$. In Zeitfenster $\tau = 1$ werden, über visuelle Inspektion (vgl. Kapitel 3.3.4) drei und in Zeitfenster $\tau = 2$ vier relevante Cluster gefunden (markiert mit blauen Ellipsen).

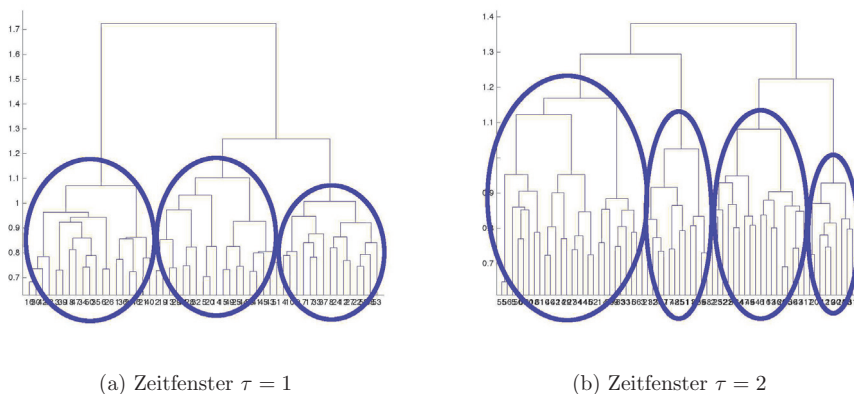


Abbildung 5.10: Dendrogramme für den Übergang von Zeitfenster $1 \rightarrow 2$.

Tabelle 5.2 zeigt die Unähnlichkeitsmatrix der beiden Clusterings \mathcal{K}^1 und \mathcal{K}^2 .

	C_1^2	C_2^2	C_3^2	C_4^2
C_1^1	0.4713	0.2807	0.3961	0.2866
C_2^1	0.5696	0.4144	0.4275	0.2210
C_3^1	0.2587	0.4170	0.5375	0.4909
C_4^1	missing values			

Tabelle 5.2: Relationsmatrix für \mathcal{K}^1 und \mathcal{K}^2 .

Die gelb markierten Felder mit den niedrigsten Unähnlichkeitswerten in der Matrix weisen auf Dokumentencluster hin, die zueinander am ähnlichsten sind ($C_2^1 \leftrightarrow C_4^2$, $C_3^1 \leftrightarrow C_1^2$, $C_1^1 \leftrightarrow C_2^2$ obwohl C_1^1 und C_4^2 ebenfalls einen niedrigen Unähnlichkeitswert aufweist). C_3^2 wird keinem Cluster aus $\tau = 1$ zugeordnet. C_1^1 ist bereits eindeutig dem Cluster C_2^2 zugeordnet und die übrigen Werte in der Spalte von C_3^2 sind höher als die Relationsschranke $dis_{lb} = 0,4$, jedoch unter der Relationsschranke $dis_{ub} = 0,55$. Damit wäre eine 'Expert Inspection' für Cluster C_3^2 nötig, jedoch sind die beiden anderen Cluster C_2^1 und C_3^1 bereits eindeutig vergeben. C_3^2 kann somit als ein neu auftretendes Thema (grün markiert) angesehen werden. Für die Zeile 4 hat die Matrix zudem fehlende Werte.

Übergang zwischen den Zeitfenstern $2 \rightarrow 3$

Der Dendrogrammvergleich zwischen den Zeitfenstern 2 und 3 zeigt eine Reduktion der Clusteranzahl von 4 auf 3 Cluster.

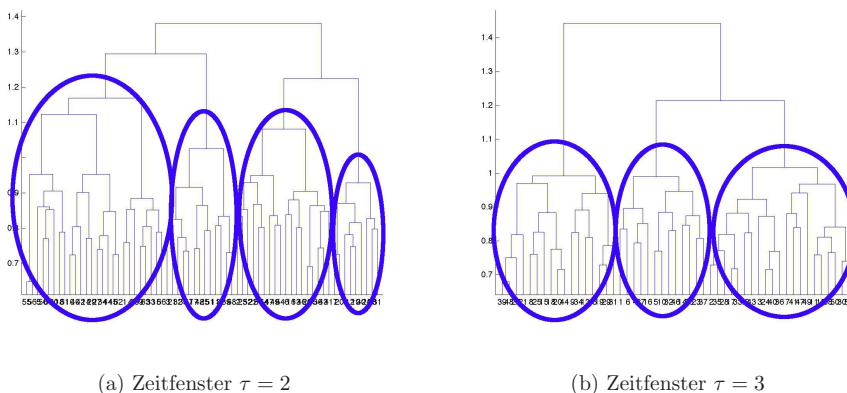


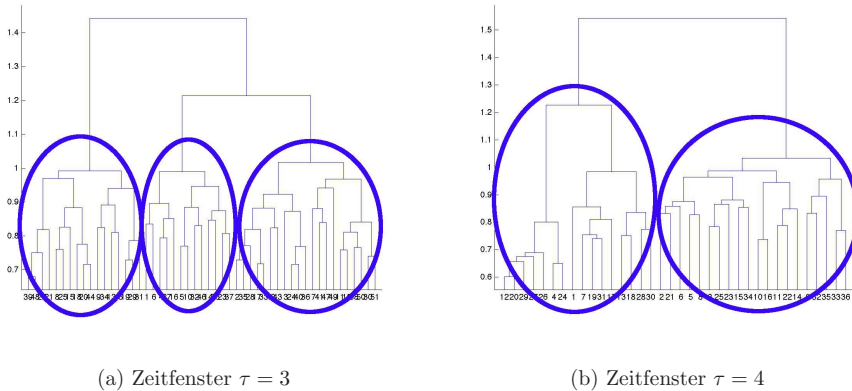
Abbildung 5.11: Dendrogramme für den Übergang von Zeitfenster $2 \rightarrow 3$.

	C_1^3	C_2^3	C_3^3	C_4^3
C_1^2	0.5189	0.4678	0.5177	missing values
C_2^2	0.3865	0.2745	0.3675	
C_3^2	0.4128	0.4510	0.2692	
C_4^2	0.2538	0.2820	0.3513	

Tabelle 5.3: Relationsmatrix für \mathcal{K}^2 und \mathcal{K}^3 .

Beim Übergang von Zeitfenster $\tau = 2$ auf $\tau = 3$ verschwindet ein Cluster. Das Thema C_1^2 , markiert in Rot, existiert in Zeitfenster $\tau = 3$ nicht länger. Wegen $dis^{2,3}(C_1^2, *) \geq dis_{ub}$ in der Zeile von C_1^2 überschreiten alle Unähnlichkeiten die vordefinierte untere Schranke, somit ist das Cluster thematisch nicht ähnlich zu irgendeinem Cluster (Thema) im darauf folgenden Zeitfenster $\tau = 3$. Die Werte für C_1^2 liegen zwar noch unter der oberen Schranke $dis_{ub} = 0,55$, jedoch sind alle neuen Cluster in $\tau = 3$ bereits vergeben, C_1^2 kann keinem neuen Cluster in $\tau = 3$ zugeordnet werden, das Thema C_1^2 stirbt.

Übergang zwischen den Zeitfenstern $3 \rightarrow 4$

Abbildung 5.12: Dendrogramme für den Übergang von Zeitfenster $3 \rightarrow 4$.

	C_1^4	C_2^4	C_3^4
C_1^3	0.3597	0.5690	missing values
C_2^3	0.2634	0.5062	
C_3^3	0.3994	0.5574	

Tabelle 5.4: Relationsmatrix für \mathcal{K}^3 und \mathcal{K}^4 .

Abbildung 5.12 (a) und (b) zeigen die Dendrogramme für die Zeitfenster $\tau = 3$ und $\tau = 4$. Durch die Markierung mit den blauen Ellipsen ist zu erkennen, dass eine Drei-Cluster-Lösung \mathcal{K}^3 und eine Zwei-Cluster-Lösung \mathcal{K}^4 gewählt wird. Diesmal ist Cluster C_2^3 dem Cluster C_1^4 zugeordnet ($C_2^3 \leftrightarrow C_1^4$). Alle Werte in der Spalte von C_2^4 sind groß. Die Unähnlichkeiten $dis^{3,4}(C_1^3, C_2^4)$ und $dis^{3,4}(C_3^3, C_2^4)$ sind größer als die obere Schranke dis_{ub} . Diese Cluster können dem Cluster C_2^4 nicht zugeordnet werden. Die Unähnlichkeit $dis^{3,4}(C_2^3, C_2^4)$ liegt zwar zwischen dis_{lb} und dis_{ub} , jedoch ist das Cluster C_2^3 schon eindeutig dem Cluster C_1^4 zugeordnet. Das heißt, C_2^4 ist ein neu aufkommendes Cluster, zudem hat die dritte Spalte fehlende Werte wegen $|\mathcal{K}^4| = 2$. Zusätzlich ist erkennbar, dass C_1^3 und C_3^3 in $\tau = 3$ erloschen sind (vgl. Tabelle 5.4).

Übergang zwischen den Zeitfenstern $4 \rightarrow 5$

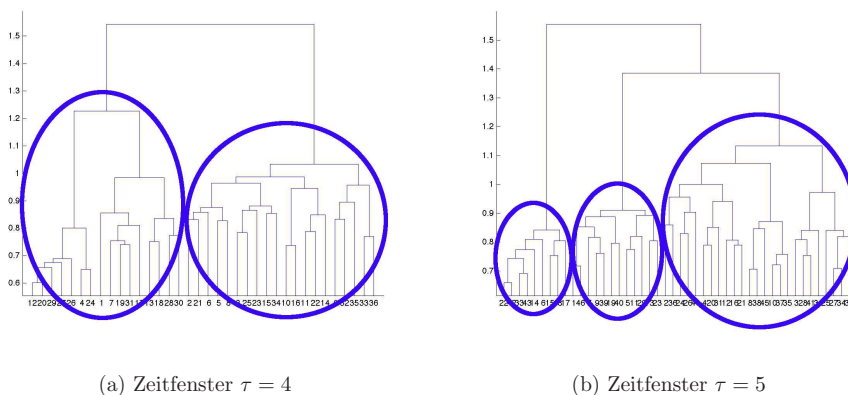


Abbildung 5.13: Dendrogramme für den Übergang von Zeitfenster $4 \rightarrow 5$.

	C_1^5	C_2^5	C_3^5
C_1^4	0.4535	0.2581	0.5551
C_2^4	0.6309	0.3928	0.2601
C_3^4	missing values		

Tabelle 5.5: Relationsmatrix für \mathcal{K}^4 und \mathcal{K}^5 .

Zwei bestehende Themen, Cluster C_1^4 und Cluster C_2^4 in Zeitfenster $\tau = 4$, können jeweils einem Cluster C_2^5 und Cluster C_3^5 in Zeitfenster $\tau = 5$ zugeordnet werden (gelb markiert). Das Thema C_1^5 ist ein neues Thema in $\tau = 5$, da alle Werte in der Spalte die untere Schranke dis_{lb} überschreiten, zudem sind beide bestehenden Cluster schon zugeordnet.

Ergebnisüberblick

Die Testkonfiguration in Tabelle 5.1 spiegelt sich auch in der grafischen Darstellung der Indizierten Hierarchien (Dendrogramme) der jeweiligen Zeitfenster wider.

Durch visuelle Inspektion der Clusterrepräsentanten (Centroiden) erhält man folgende in Tabelle 5.6 gezeigten Ergebnisse (vgl. dazu auch Tabelle 5.1). Diese bestätigen durch die auf Basis der hierarchischen Clusterings erzeugten Relationsmatrizen die gefundenen Tracking Ergebnisse.

time window τ	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$
cluster 1	P	S	WF	P	WF
cluster 2	WF	P	P	S	P
cluster 3	S	TIT	TIT		S
cluster 4		WF			

Tabelle 5.6: Ergebnisüberblick.

Das Cluster C_1^1 kann durch visuelle Inspektion des entsprechenden Centroiden als Politik-Inland (P) gelabelt werden. Dabei erhält man über die Relationsmatrizen von $\tau = 1$ bis $\tau = 4$ folgende Zuordnungen: $C_1^1 \leftrightarrow C_2^2 \leftrightarrow C_3^3 \leftrightarrow C_4^4 \leftrightarrow C_5^5$. Das Thema Politik-Inland (P) existiert also in allen zeitfensterabhängigen Clusterings \mathcal{K}^τ .

Für das Thema Wirtschaft & Finanzen (WF) erhält man folgende Zuordnungen im Zeitverlauf: $C_2^2 \leftrightarrow C_4^4 \leftrightarrow C_1^3 \leftrightarrow \emptyset \leftrightarrow C_1^5$.

Das Cluster C_1^5 wird durch den Vergleich mit dem direkt vorangehenden Clustering \mathcal{K}^4 als neu gelabelt. Um herauszufinden, ob es im Betrachtungshorizont schon früher aufkam, muss man das Cluster C_1^5 mit allen Clustern aller vorangehenden Zeitfenster $\tau' < \tau - 1$ vergleichen.

Das Thema Sport-Fußball (S) ist in Zeitfenster $\tau = 3$ verschwunden, erscheint jedoch in $\tau = 4$ erneut.

Das Thema Technik-Industrie & Transport-Verkehr (TIT) ist neu im Zeitfenster $\tau = 2$, verschwindet aber wieder ab Zeitfenster $\tau = 4$.

Wird ein Cluster in einem früheren Zeitfenster gefunden, dessen Unähnlichkeit zu dem betrachteten Cluster kleiner ist als die untere Schranke dis_b , nimmt man an, dass das betrachtete Cluster nicht neu ist. Überschreitet sie jedoch die obere Schranke dis_{ub} , geht man von einem neu aufkommenden Thema aus.

Das vorliegende überschaubare Beispiel zeigt, wie Themen über die Zeit hinweg verfolgt werden können. Dabei muss geklärt werden, ob ein in einem Zeitfenster neu aufkommendes Cluster möglicherweise schon in einem früheren Zeitfenster in Erscheinung getreten ist. Das heißt, nicht nur das unmittelbar vorangehenden Zeitfenster muss überprüft werden, sondern auch alle Cluster in weiter zurückliegenden Zeitfenstern.

Von besonderer Bedeutung für das Tracking von Themen, bzw. die Relationsbestimmung zwischen Clustern unterschiedlicher Zeitfenster, ist die Wahl adäquater oberer und unterer Schranken, zwischen denen zusätzliche Inspektionen erfolgen müssen. Diese Parameter werden von der Größe des 'Vector Space' beeinflusst. Je größer die Dimension Z , also das genutzte Wörterbuch \mathcal{L}' gewählt wird, desto größer ist die Anzahl an weniger häufigen Generellen Termfrequenzen, die zur Dokumentenrepräsentation im 'Vector Space' verwendet werden, und desto größer müssen auch die Schranken gewählt werden.

Die Ergebnisse in diesem Kapitel wurden in Frankfurt im Jahre 2011 auf der gemeinsamen Konferenz der Gesellschaft für Klassifikation (GfKl) und der Deutschen Arbeitsgemeinschaft für Mustererkennung (DAGM), an die sich auch ein Symposium der International Federation of Classification Societies (IFCS) anschloss, vorgestellt. Weitere Erläuterungen finden sich unter anderem in GAUL/VINCENT (2013).

5.4.3 Testreihe III. (GfKl & SFC (2013))

Im folgenden Kapitel wird eine weitere Testkonfiguration beschrieben. Als Testdatensatz dient eine Stichprobe von Dokumenten des Nachrichtenportals Spiegel Online (vgl. Kapitel 5.3.2). Um überschaubare und gleichzeitig interessante Ergebnisse zu erhalten, werden nur Dokumente aus der Rubrik 'Politik' ausgewählt.

Basis der Tests ist ein Verkleinertes Lokales Wörterbuch \mathcal{L}' mit einer Größe von $|\mathcal{L}'| = Z = 20.000$, erstellt aus dem Lokalen Wörterbuch \mathcal{L} des IDS (vgl. Kapitel 5.4.1).

Der Betrachtungszeitraum $[t_1, t_2]$ erstreckt sich vom 01.01.2011 bis zum 31.03.2011. Jeder Monat wird in 3 Zeitfenster mit ungefähr 10 Tagen Spanne eingeteilt. Insgesamt umfasst die Stichprobe 1952 ausgewählte Dokumente aus der Rubrik 'Politik'.

Die Programmlaufzeit des Prototypen beläuft sich für die Verarbeitung aller Berechnungsschritte für 1952 Dokumente bei Formel (3.28) auf 37 min mit Rechner-Konfiguration 3 und auf 22 min bei Rechner-Konfiguration 2. Für Formel (3.26) werden entsprechend 35 min bzw. 20 min benötigt. Die Laufzeiten beziehen sich auf die Rechner-Konfiguration 3 (2) (vgl. Kapitel 5.2).

Einen Überblick über die Testkonfiguration gibt Tabelle 5.7.

Zeitfenster τ		$ D^\tau $	\overline{mer}^τ	\overline{per}^τ	$ \mathcal{K}^\tau $ Formel (3.26)	$ \mathcal{K}^\tau $ Formel (3.28)
1	01.01.2011 – 10.01.2011	151	0,4541	0,3581	13	14
2	11.01.2011 – 21.01.2011	184	0,4523	0,3572	3	9
3	22.01.2011 – 31.01.2011	213	0,4492	0,3529	7	14
4	01.02.2011 – 10.02.2011	253	0,4496	0,3565	12	12
5	11.02.2011 – 20.02.2011	216	0,4484	0,3517	10	9
6	21.02.2011 – 28.02.2011	201	0,4594	0,3531	6	10
7	01.03.2011 – 10.03.2011	237	0,4543	0,3552	5	12
8	11.03.2011 – 20.03.2011	200	0,4601	0,3584	5	13
9	21.03.2011 – 31.03.2011	297	0,4640	0,3626	11	11

Tabelle 5.7: Testkonfiguration & Ergebnisse für 1. Quartal 2011.

Die Dokumente des Betrachtungszeitraums $[t_1, t_2]$ sind nach ihrer Datierung in die Dokumentenmengen D^τ für das jeweilige Zeitfenster τ eingeteilt (vgl. Tabelle 5.7).

Die durchschnittliche Worterkennungsrate für Wortform-Mengen $\overline{mer}^\tau = \sum_{d_s^\tau \in D^\tau} mer_{d_s^\tau}$

für ein Zeitfenster τ beträgt im Durchschnitt ca. 45% (vgl. Kapitel 3.1.4). Für \overline{per}^τ werden ca. 36% erreicht. Die relativ niedrige Erkennungsrate im Vergleich zu den Ergebnissen in Kapitel 5.4.1 wird von mehreren Faktoren beeinflusst. So ist es möglich, dass nicht alle Wortformen der Spiegel Dokumente im Wörterbuch \mathcal{L} des DeReKo enthalten sind, bzw. dass verwendete Wörterbuch \mathcal{L} neuere Wortformen nicht beinhaltet. Einen entscheidenden Einfluss hat die Stoppwortentfernung (siehe Kapitel 3.1.3). Ohne Stoppwortentfernung, bei sonst gleichen Parametern, liegt die Worterkennungsrate \overline{mer}^τ mit ca. 81% wesentlich höher. Für die Prozentuale Worterkennungsrate \overline{per}^τ sind es ca. 86%.

Mit folgenden Parameterkonfigurationen werden die Tests durchgeführt:

- Verkleinertes Lokales Wörterbuch \mathcal{L}' mit Stopwortentfernung mit $|\mathcal{L}'| = Z = 20.000$
- Cosinusmaß als Distanzmaß (vgl. Kapitel 3.2)
- Merkmalsextraktion (vgl. Formel (3.4))
- Hierarchisches Clusterverfahren (Ward-Verfahren, vgl. Formel (3.24) Kapitel 3.3.3)
- Ellbogenkriterium nach Formel (3.26) und Formel (3.28) (vgl. Kapitel 3.3.4) mit Beschränkung von $|\mathcal{K}|$ auf $\mathbb{N} \cap [2, 15]$
- Relationsmatrizen mit Relationsschranken $dis_{lb} = 0,47$ und $dis_{ub} = 0,56$

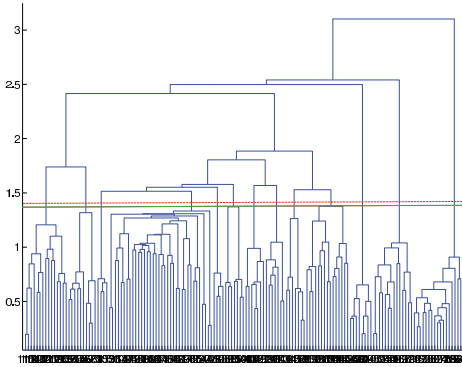
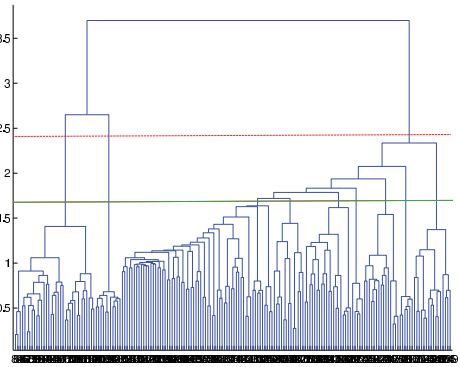
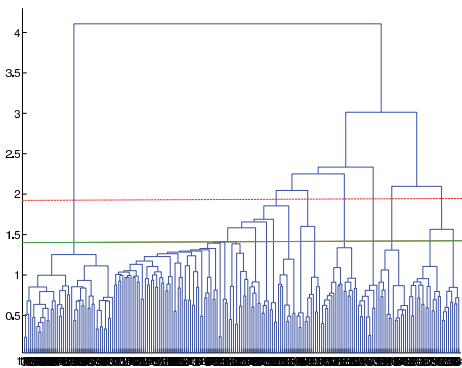
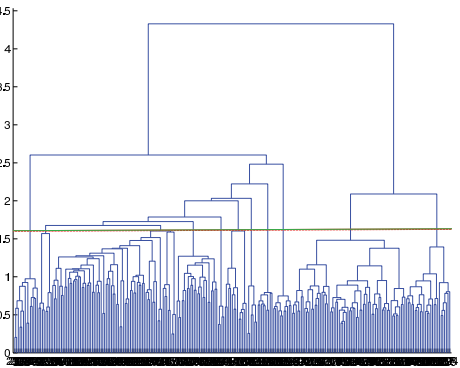
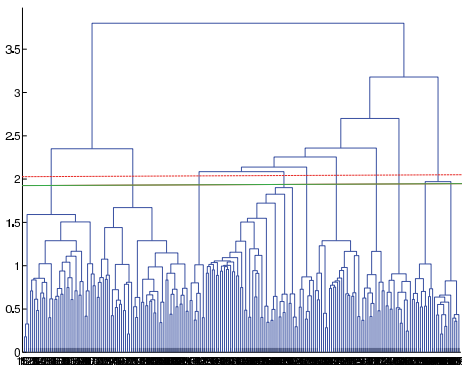
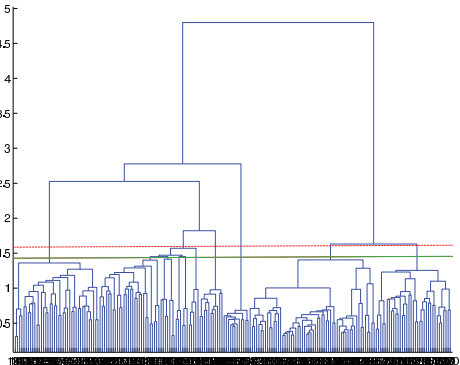
Durch das Hierarchische Clusterverfahren Ward (vgl. Kapitel 3.3.3) ergibt sich für jedes Zeitfenster τ eine Indizierte Hierarchie mit zugehörigem Dendrogramm. Die für das jeweilige Zeitfenster τ zu bestimmende Clusteranzahl (Anzahl der Themen) wird jedoch nicht durch visuelle Inspektion (wie in Kapitel 5.4.2), sondern mittels Ellbogenkriterium (vgl. Kapitel 3.3.4) bestimmt. Dabei werden die beiden Gütekriterien nach Formel (3.26) bzw. (3.28) berechnet. Das Gütekriterium nach Formel (3.28) führt in diesem Test zu tendenziell feineren differenzierteren Clusterings als nach Formel (3.26).

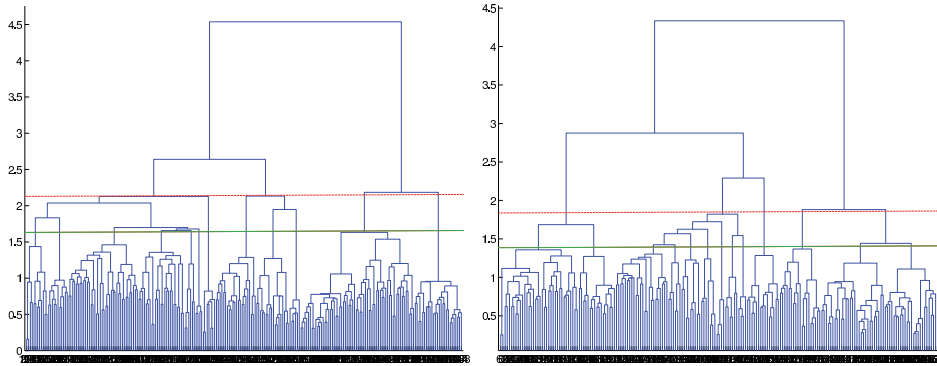
Mit den so gefundenen Clusterings \mathcal{K}^τ verschiedener Zeitfenster $\tau \in [t_1, t_2]$ wird mit Hilfe der $m_\tau \times (m_\tau - 1)/2$ Relationsmatrizen mit $m_\tau = \tau - t_1 + 1$ und der definierten Relationsschranken für jedes Thema (Cluster) sein zugehöriger Mono-Themen-Graph erstellt (vgl. Kapitel 4).

Für eine schnelle Erfassbarkeit eines Clusters (Thema) innerhalb eines Themen-Graphen wird für den entsprechenden Centroid eine sogenannte Tag-Cloud generiert. Auf der Webseite www.wordle.net steht dafür ein entsprechendes Skript zur Verfügung. Für jede Tag-Cloud werden die 100 höchst-gewichteten Terme des Centroids als Wortwolke dargestellt. Die Schriftgröße eines Terms innerhalb der Wolke entspricht seiner Gewichtung im Centroid (vgl. Abbildung 5.15). Zur Relationsbestimmung zwischen Clustern (Ähnlichkeitsvergleich der Centroiden) könnten auch Ähnlichkeiten zwischen den Tag-Clouds bestimmt werden (vgl. PARK ET AL. (2010)).

Dendrogramme für die Zeitfenster 1 bis 9

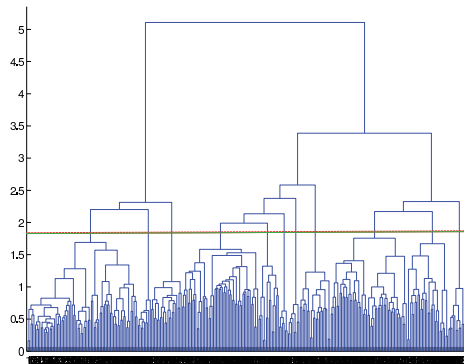
Die Grafiken 5.14 (a) bis (i) zeigen die Dendrogramme der Ward-Clusterings in $\tau = 1$ bis $\tau = 9$. Die mit dem Güterkriterium nach Formel (3.26) gefundenen Cuts werden mit einer (gestrichelten) roten Linie, die nach Formel (3.28) mit einer grünen Cut-Linie dargestellt.

(a) Zeitfenster $\tau = 1$ (b) Zeitfenster $\tau = 2$ (c) Zeitfenster $\tau = 3$ (d) Zeitfenster $\tau = 4$ (e) Zeitfenster $\tau = 5$ (f) Zeitfenster $\tau = 6$



(g) Zeitfenster $\tau = 7$

(h) Zeitfenster $\tau = 8$



(i) Zeitfenster $\tau = 9$

Abbildung 5.14: Dendrogramme für die Zeitfenster 1 bis 9.

Das Zeitfenster $\tau = 8$

Im Folgenden werden anhand des Zeitfensters $\tau = 8$ die gefundenen Ergebnisse näher erläutert.

Nach Formel (3.26) werden 5 Cluster (Themen) identifiziert (vgl. Abbildung 5.14 (h), rote gestrichelte Cut-Linie), deren Centroide (Ausschnitt mit 20 der 100 häufigsten Terme) mit ihren entsprechenden Tag-Clouds in den Abbildungen 5.15 (a) bis (j) dargestellt sind.

Der Clusteralgorithmus weist jedem Cluster $C_{k_\tau}^\tau$ eines Clustering \mathcal{K}^τ eine eindeutige Kennzeichnung $k_\tau \in \{1, \dots, |\mathcal{K}^\tau|\}$ für jedes Zeitfenster τ zu. Die Nummerierung erfolgt unabhängig von der Themen-Frequenz des jeweiligen Clusters.

	1	2
1	'Rebellen'	3.7815
2	'al'	3.3121
3	'Libyen'	2.9211
4	'Truppen'	2.3766
5	'Uno'	1.4610
6	'Soldaten'	1.3527
7	'Regime'	1.3143
8	'Reporter'	0.9914
9	'Stock'	0.8319
10	'Panzer'	0.8054
11	'Machthaber'	0.7909
12	'Mustafa'	0.7633
13	'Reuters'	0.7237
14	'Resolution'	0.6979
15	'Qaida'	0.6699
16	'Flugzeuge'	0.6554
17	'Zivilisten'	0.6493
18	'Kämpfe'	0.6478
19	'Nato'	0.6308
20	'Jonathan'	0.6221

(a)



(b) $f((8, 1)) = 9,5 \times 10^{-2}$

	1	2
1	'Libyen'	7.2877
2	'Uno'	3.5320
3	'Rebellen'	1.9723
4	'al'	1.8933
5	'Resolution'	1.8609
6	'Truppen'	1.7760
7	'Sicherheitsrat'	1.5192
8	'Westerwelle'	1.1351
9	'Außenminister'	1.1069
10	'Eingreifen'	1.0250
11	'Gemeinschaft'	1.0238
12	'Zivilisten'	0.9952
13	'Diktator'	0.9489
14	'Arabischen'	0.9218
15	'Sarkozy'	0.9209
16	'Nato'	0.8461
17	'internationale'	0.8339
18	'Machthaber'	0.8288
19	'militärische'	0.8163
20	'militärischen'	0.7836

(c)



(d) $f((8, 2)) = 26,5 \times 10^{-2}$

	1	2
1	'Demonstranten'	5.0163
2	'Sicherheitskräfte'	3.2880
3	'Saudi'	2.8835
4	'Arabien'	2.5777
5	'arabischen'	2.2749
6	'Proteste'	2.0735
7	'Hauptstadt'	1.8340
8	'Schiiten'	1.7055
9	'schiitischen'	1.6538
10	'Opposition'	1.4416
11	'Augenzeugen'	1.4068
12	'Reuters'	1.2948
13	'Demonstrationen'	1.2641
14	'König'	1.2115
15	'al'	1.1198
16	'Abdullah'	1.0884
17	'Golf'	1.0142
18	'Soldaten'	0.9896
19	'Truppen'	0.9563
20	'Gewalt'	0.9490

(e)



(f) $f((8, 3)) = 7,5 \times 10^{-2}$

	1	2
1	'Marine'	0.8528
2	'Davis'	0.5515
3	'Nato'	0.4509
4	'Afghanistan'	0.4170
5	'Soldaten'	0.4007
6	'Bericht'	0.3951
7	'pakistanischen'	0.3928
8	'Bord'	0.3871
9	'Armee'	0.3641
10	'Gaza'	0.3641
11	'rot'	0.3629
12	'getötet'	0.3551
13	'Vorwürfe'	0.3474
14	'Anhalt'	0.3451
15	'Westjordanland'	0.3437
16	'Bundeswehr'	0.3355
17	'Ermittler'	0.3349
18	'Taliban'	0.3304
19	'Neuwahlen'	0.3218
20	'Sachsen'	0.3077

(g)



(h) $f((8, 4)) = 30 \times 10^{-2}$

Das vierte Cluster C_4^8 (Abbildung 5.15 g & (h)) ist ein so genanntes 'Miscellaneous' (Vermischtes) Cluster aus Dokumenten verschiedener Themen und damit in sich relativ heterogen. Die einzelnen Sub-Themen des 'Miscellaneous' Clusters besitzen auf der gewählten Cut-Höhe nicht ausreichend viele homogene Dokumente, um eigene homogene Cluster zu bilden. Für den End-User ist ein solches 'Miscellaneous' Cluster leicht durch seine Tag-Cloud (vgl. Abbildung 5.15 (h)) zu erkennen. Die entsprechende Wortwolke zeigt eine relativ gleichmäßige Größenverteilung der Wortformen. Es gibt weniger Wortformen, die als Schlagworte besonders hervorstechen.

Das fünfte Cluster C_5^8 (Abbildung 5.15 (i) & (j)) firmiert unter dem Schlagwort 'Japan'. Die in der Tag-Cloud genutzten Wortformen legen nahe, dass es sich hierbei um die Atomkatastrophe von 'Fukushima' handelt, zumal deren zeitliches Auftreten (11. März 2011) mit dem betrachteten Zeitfenster $\tau = 8$ übereinstimmt (vgl. Tabelle 5.7). Es fällt auf, dass der Begriff 'Fukushima' in der Tag-Cloud nicht vertreten ist. Die Wortform 'Fukushima' ist zu diesem Zeitpunkt zwar im Lokalen Wörterbuch \mathcal{L} enthalten (Position 316.322 (von 2 Mio.)), nicht aber im Verkleinerten Lokalen Wörterbuch \mathcal{L}' mit $Z = 20.000$. Solche zeitlich aufkommenden Schlagworte (vgl. Kapitel 3.1.4) sollten durch ein aktualisiertes Referenzkorpus \mathcal{R}_a (über ein aktualisiertes Lokales Wörterbuch \mathcal{L}_a) in ein aktualisiertes Verkleinertes Lokales Wörterbuch \mathcal{L}'_a aufgenommen werden. Dennoch ist das Thema für einen End-User mit den schon in \mathcal{L}' enthaltenen Wortformen ausreichend beschrieben.

Themenkomplex 'Libyen'

Jede der fünf Themen (Cluster) aus Zeitfenster 8, also C_1^8 bis C_5^8 , sind einem Mono-Themen-Graphen zugeordnet. Dabei können mehrere Cluster dem gleichen Mono-Themen-Graphen angehören (vgl. Kapitel 4.3). Im Folgenden wird der den Clustern C_1^8 und C_2^8 zugehörige Mono-Themen-Graph erläutert und evaluiert (vgl. Abbildung 5.17).

Zunächst wird für alle Cluster $C_{k_\tau}^\tau$ die Themen-Frequenz $f((\tau, k_\tau))$ ermittelt (vgl. Kapitel 4.2). Tabelle 5.8 zeigt die Berechnung der Themen-Frequenz der Cluster C_1^8 bis C_5^8 in Zeitfenster $\tau = 8$. Ein Knoten (τ, k_τ) (als Repräsentant eines Clusters $C_{k_\tau}^\tau$) im Themen-Graphen wird durch seine zeitliche Komponente τ und die Themen-Frequenz $f((\tau, k_\tau))$ des Clusters $C_{k_\tau}^\tau$ bestimmt. Für Knoten $(8, 1)$, also Cluster C_1^8 ergibt sich als Themen-Frequenz $f((8, 1)) = 9,5 \times 10^{-2}$. Für C_2^8 ergibt sich $f((8, 2)) = 26,5 \times 10^{-2}$.

Abbildungen	Cluster $C_{k_\tau}^\tau$	Themen-Frequenz $f((\tau, k_\tau)) = \frac{ C_{k_\tau}^\tau }{ D^\tau }$
Abbildung 5.15 (a) & (b)	C_1^8	$f((8, 1)) = \frac{ C_1^8 }{ D^8 } = \frac{19}{200} = 9,5 \times 10^{-2}$
Abbildung 5.15 (c) & (d)	C_2^8	$f((8, 2)) = \frac{ C_2^8 }{ D^8 } = \frac{53}{200} = 26,5 \times 10^{-2}$
Abbildung 5.15 (e) & (f)	C_3^8	$f((8, 3)) = \frac{ C_3^8 }{ D^8 } = \frac{15}{200} = 7,5 \times 10^{-2}$
Abbildung 5.15 (g) & (h)	C_4^8	$f((8, 4)) = \frac{ C_4^8 }{ D^8 } = \frac{60}{200} = 30 \times 10^{-2}$
Abbildung 5.15 (i) & (j)	C_5^8	$f((8, 5)) = \frac{ C_5^8 }{ D^8 } = \frac{53}{200} = 26,5 \times 10^{-2}$

Tabelle 5.8: Themen-Frequenzen $f((\tau, k_\tau))$ für Cluster C_1^8 bis C_5^8 .

Die Kanten des Themen-Graphen von C_1^8 werden über Relationsmatrizen für die verschiedenen Zeitfenster $t_1 \leq \tau' < \tau = 8$ berechnet (vgl. auch Kapitel 4.1). Tabelle 5.9 zeigt die Relationsmatrix für die Clusterings $\mathcal{K}^{\tau=7}$ und $\mathcal{K}^{\tau=8}$. Diese Matrix ist eine exemplarisch ausgewählte Relationsmatrix der $m_\tau \times (m_\tau - 1)/2$ Relationsmatrizen, wobei in diesem Fall (für die Gegenwart τ) für $m_\tau = \tau - t_1 + 1$ gilt.

	C_1^8	C_2^8	C_3^8	C_4^8	C_5^8
C_1^7	0,9127	0,9163	0,8977	0,7503	0,8145
C_2^7	0,9234	0,9348	0,9305	0,8025	0,8472
C_3^7	0,2001	0,2929	0,7174	0,7730	0,9229
C_4^7	0,3484	0,1213	0,7615	0,7564	0,8886
C_5^7	0,6873	0,7594	0,5699	0,4806	0,7753

Tabelle 5.9: Relationsmatrix mit $dis^{7,8}(C_{k_7}^7, C_{k_8}^8)$ für \mathcal{K}^7 und \mathcal{K}^8 .

Bei einer unteren Relationsschranke $dis_b = 0,47$ liefert die Relationsmatrix für die Clusterings \mathcal{K}^7 und \mathcal{K}^8 vier Kanten $((7, 3), (8, 1))$, $((7, 3), (8, 2))$, $((7, 4), (8, 1))$ und $((7, 4), (8, 2))$ zwischen den Clustern C_3^7, C_4^7 aus Zeitfenster 7 und C_1^8, C_2^8 aus Zeitfenster 8 (gelbe Markierungen in Tabelle 5.9).

Die weiteren Kanten innerhalb des Mono-Themen-Graphen ergeben sich durch Berechnung der restlichen, hier nicht aufgeführten 27 Relationsmatrizen.

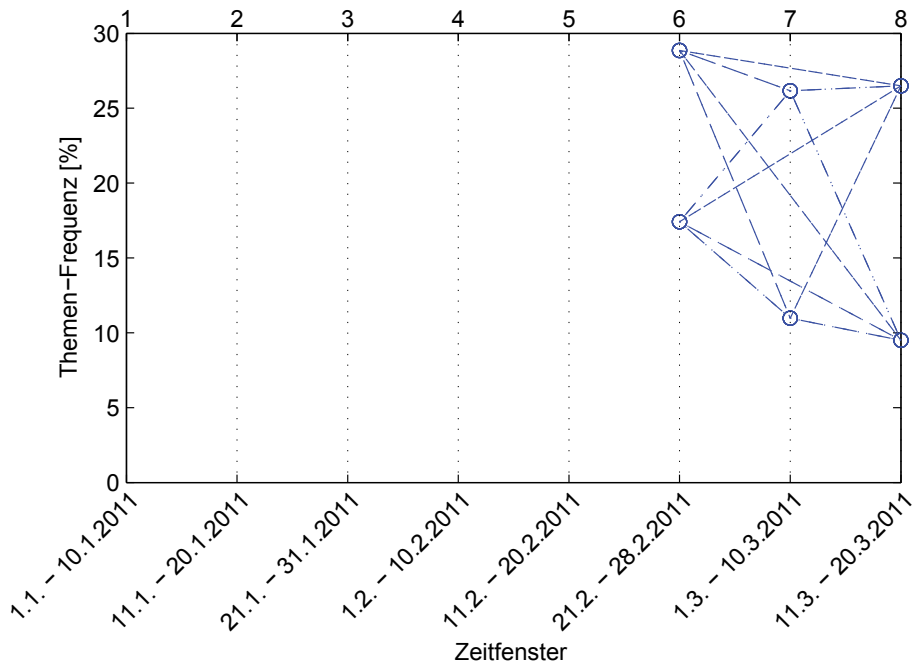


Abbildung 5.16: Mono-Themen-Graph $G_{(8,1)}$ der $Rel(1,8)_{(8,1)}$ (nach Formel (3.26)).

Der für die Cluster C_1^8 und C_2^8 dargestellte Mono-Themen-Graph in Abbildung 5.16 weist vier weitere in der Vergangenheit (Zeitfenster 6 und 7) bezüglich $\tau = 8$ liegende Knoten (Cluster) auf. Bei allen sechs Clustern des Themen-Graphen handelt es sich um den Themenkomplex 'Libyen'. Abbildung 5.17 (a) bis (d) zeigt die Tag-Clouds einschließlich der jeweiligen Themen-Frequenzen der vier zurückliegenden Cluster.

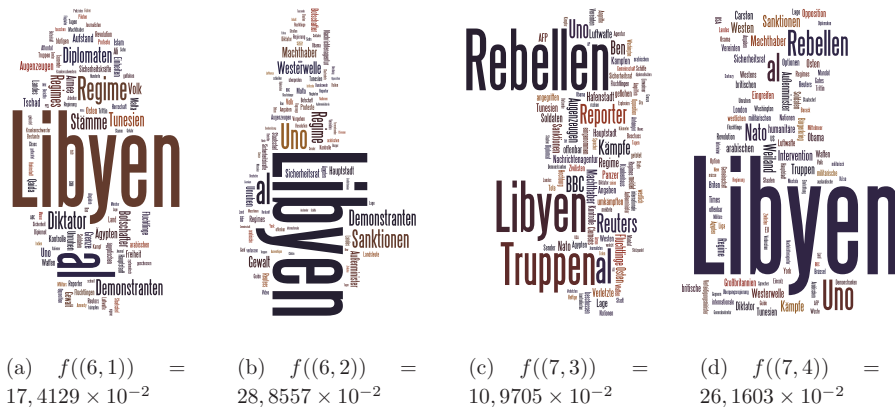


Abbildung 5.17: Tag-Clouds für Cluster in den Zeitfenstern 6 & 7 im Mono-Themen-Graphen 'Libyen'.

Der Clusteralgorithmus hat die vorliegenden Dokumente zum Thema 'Libyen' im Zeitfenster $\tau = 8$ auf Grund ihrer Homogenität bzw. Heterogenität bei gegebenem Ellbogenkriterium (Formel (3.26)) in die Cluster C_1^8 und C_2^8 aufgeteilt. Der Mono-Themen-Graph zeigt, dass beide Cluster C_1^8 und C_2^8 über Relationen zu Clustern zurückliegender Zeitfenster indirekt miteinander in Beziehung stehen.

Die Erweiterung des Mono-Themen-Graphen um ein zusätzliches Zeitfenster $\tau = 9$ ist in Abbildung 5.18 dargestellt. In dem neuen Zeitfenster $\tau = 9$ wird der Mono-Themen-Graph um drei weitere Knoten ergänzt (zugehörige Tag-Clouds und Themen-Frequenzen siehe Abbildung 5.19 (a) bis (c)). Diese neu entstandenen Knoten gehören ebenfalls dem Themenkomplex 'Libyen' an.

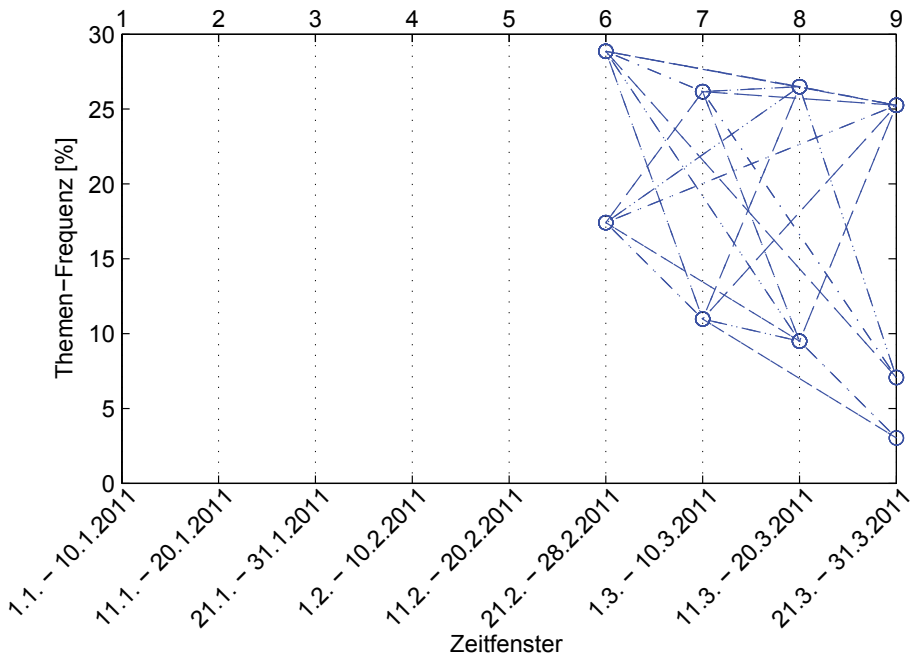


Abbildung 5.18: Mono-Themen-Graph $G_{(9,6)}$ der $Rel(1,9)_{(9,6)}$ (nach Formel (3.26)).

Bei der Berechnung mit dem Ellbogenkriterium nach Formel (3.28) ergeben sich selbstverständlich für die Cluster des neuen Mono-Themen-Graphen eigene Nummerierungen, die von den bisherigen abweichen können. Cluster, die sich in beiden Graphen entsprechen, sind anhand der charakteristischen Merkmalsausprägungen ihrer Centroide bzw. Tag-Clouds, sowie ihrer Position und Vernetzung im Themen-Graphen, auffindbar.

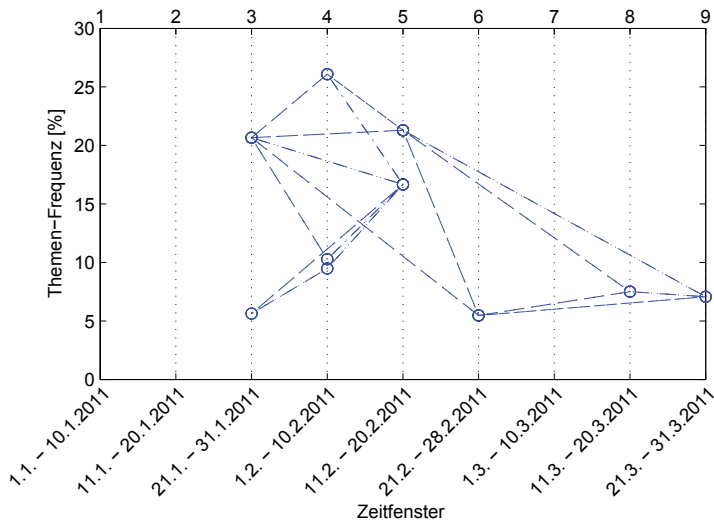
Der Vergleich beider Graphen (Abbildungen 5.18 und 5.20) zeigt, dass die Cluster, bis auf eine Ausnahme, ihre Position im jeweiligen Mono-Themen-Graphen beibehalten. Das Cluster mit $f((8,2)) = 26,5 \times 10^{-2}$ in Abbildung 5.18 (vgl. auch Abbildungen 5.15 (c) & (d)) teilt sich auf und wird zu den Clustern mit $f((8,3)) = 12 \times 10^{-2}$ und $f((8,4)) = 14,5 \times 10^{-2}$ in Abbildung 5.20 (vgl. dazu auch die zugehörigen Tag-Clouds in Abbildung 5.21 (a) und (b)). Dies erklärt sich durch die granularere Klassifikation, die sich durch das Ellbogenkriterium nach Formel (3.28) ergibt. Die Aufspaltung ermöglicht differenziertere Einblicke in den betrachteten Themenkomplex.

(a) $f((8,3)) = 12 \times 10^{-2}$ (b) $f((8,4)) = 14,5 \times 10^{-2}$

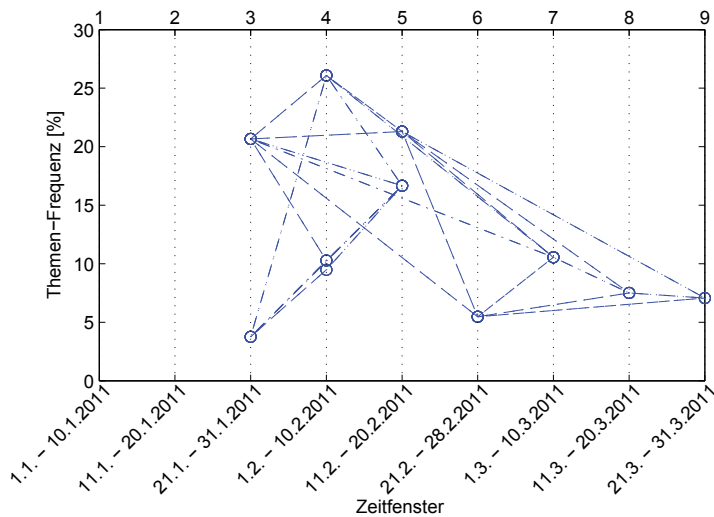
Abbildung 5.21: Tag Clouds für Cluster bei Themen-Frequenz $f((8,3)) = 12 \times 10^{-2}$ & $f((8,4)) = 14,5 \times 10^{-2}$ in dem Zeitfenster $\tau = 8$ im Mono-Themen-Graphen 'Libyen' zum Zeitpunkt $\tau = 9$ (nach Formel (3.28)).

Themenkomplex 'Arabischer Frühling'

Am Themenkomplex 'Arabischer Frühling' mit seinen beiden Mono-Themen-Graphen (Abbildung 5.22 (a) und (b)) lassen sich einige weitere Auswirkungen des jeweiligen Ellbogenkriteriums auf die Graphenstruktur verdeutlichen.



(a) Mono-Themen-Graph $G_{(9,11)}$ der $Rel(1,9)_{(9,11)}$ nach Formel (3.26)



(b) Mono-Themen-Graph $G_{(9,11)}$ der $Rel(1,9)_{(9,11)}$ nach Formel (3.28)

Abbildung 5.22: Mono-Themen-Graphen für den Themenkomplex 'Arabischer Frühling' (nach Formel (3.26) bzw. (3.28)).

Bei einem granulareren Clustering (Abbildung 5.22 (b)) entsteht im siebten Zeitfenster ein neuer Knoten mit einer Themen-Frequenz $f((7, 2)) = 10,5 \times 10^{-2}$. Dieser Knoten $(7, 2)$ ist fünffach vernetzt:

Mit Knoten $(3, 14)$ in Zeitfenster 3 mit $f((3, 14)) = 20,7 \times 10^{-2}$, mit Knoten $(4, 9)$ in Zeitfenster 4 mit $f((4, 9)) = 26,1 \times 10^{-2}$, mit Knoten $(5, 6)$ in Zeitfenster 5 mit $f((5, 6)) = 21,3 \times 10^{-2}$, mit Knoten $(6, 8)$ mit $f((6, 8)) = 5,5 \times 10^{-2}$ in Zeitfenster 6, sowie mit Knoten $(8, 13)$ mit $f((8, 13)) = 7,5 \times 10^{-2}$ im Zeitfenster 8.

Das neue Cluster vereint offensichtlich Dokumente, deren Themen ihren Ursprung in Clustern früherer Zeitfenster haben und auf Grund ihrer Ähnlichkeit nun neu gruppiert werden. Es besteht darüber hinaus eine Beziehung zu einem späteren Cluster in Zeitfenster 8. Bei einer groberen Klassifikation werden die entsprechenden Dokumente anderen Clustern zugeordnet.

Ein weiterer Unterschied zwischen den beiden Mono-Themen-Graphen ist eine zusätzliche Kante zwischen den Knoten $(3, 14)$ mit $f((3, 14)) = 3,8 \times 10^{-2}$ und dem Knoten $(4, 9)$ mit $f((4, 9)) = 26,1 \times 10^{-2}$ bei feinerer Klassifikation. Der Knoten $(3, 1)$ in Zeitfenster 3 mit der Themen-Frequenz $f((3, 1)) = 5,6 \times 10^{-2}$ (Abbildung 5.22 (a)) verliert bei einer feineren Klassifikation Dokumente, die Themen-Frequenz nimmt ab auf $f((3, 14)) = 3,8 \times 10^{-2}$ (Abbildung 5.22 (b)). Dadurch steigt seine Homogenität an, und damit seine Ähnlichkeit zu dem Knoten $(4, 9)$ in Zeitfenster 4 mit $f((4, 9)) = 26,1 \times 10^{-2}$, so dass sich eine neue Kante von Knoten $(3, 14)$ nach Knoten $(4, 9)$ bildet.

Die Wahl des Ellbogenkriteriums hat, wie diese Beispiele zeigen, direkten Einfluss auf die Beziehungen im Themen-Graphen. Die daraus abzuleitenden Beurteilungen müssen je nach Informationsbedarf des Betrachters vorgenommen werden.

Ausgewählte Ergebnisse dieser Testreihe wurden auf der von der Société Francophone de Classification (SFC) und der Gesellschaft für Klassifikation (GfKl) ausgerichteten European Conference on Data Analysis (2013) in Luxembourg vorgestellt.

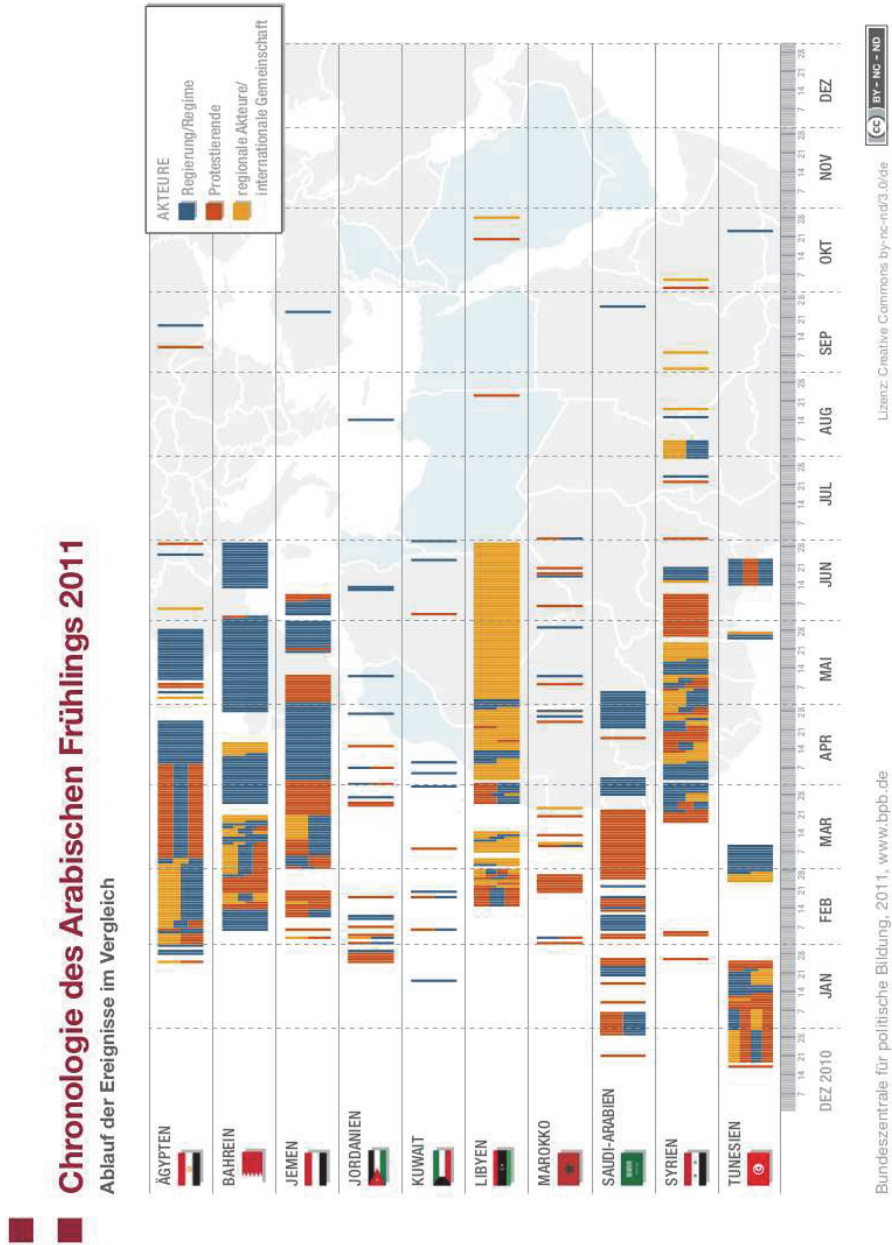


Abbildung 5.23: Chronologie des Arabischen Frühlings (Quelle: Bundeszentrale für politische Bildung, 2011, www.bpb.de).

5.4.4 Testreihe IV.

Im vorigen Kapitel 5.4.3 ist die Erstellung der Mono-Themen-Graphen für einen kleinen Zeitraum von drei Monaten dargelegt worden.

Das folgende Kapitel zeigt die Evolution von Mono-Themen-Graphen über einen Zeitraum von 52 Kalenderwochen (KW), beginnend mit der 27. Kalenderwoche des Jahres 2010 bis einschließlich der 26. Kalenderwoche des Jahres 2011. Aus der Testkonfiguration (Tabelle 5.11) ist abzulesen, dass die gewählten Zeitfenster unserer natürlichen Zeiteinteilung von Kalenderwochen entsprechen.

Zeitfenster τ		$ D^\tau $	$\overline{mer^\tau}$	$\overline{per^\tau}$	$ \mathcal{K}^\tau $ Formel (3.26)	$ \mathcal{K}^\tau $ Formel (3.28)
1	27 KW 2010	104	0,4575	0,3699	13	11
2	28 KW 2010	113	0,4517	0,3632	10	13
3	29 KW 2010	103	0,4500	0,3632	10	7
4	30 KW 2010	98	0,4586	0,3729	10	11
5	31 KW 2010	99	0,4606	0,3688	9	6
6	32 KW 2010	103	0,4507	0,3571	12	14
7	33 KW 2010	121	0,4570	0,3634	4	9
8	34 KW 2010	122	0,4537	0,3583	11	6
9	35 KW 2010	147	0,4537	0,3622	4	13
10	36 KW 2010	135	0,4578	0,3598	11	10
11	37 KW 2010	167	0,4511	0,3618	14	9
12	38 KW 2010	128	0,4538	0,3654	14	14
13	39 KW 2010	149	0,4470	0,3633	14	9
14	40 KW 2010	129	0,4438	0,3596	8	9
15	41 KW 2010	122	0,4513	0,3587	13	11
16	42 KW 2010	164	0,4460	0,3606	12	10
17	43 KW 2010	156	0,4422	0,3557	14	14
18	44 KW 2010	171	0,4414	0,3562	6	14
19	45 KW 2010	136	0,4477	0,3578	8	13

Zeitfenster τ		$ D^\tau $	$\overline{mer^\tau}$	$\overline{per^\tau}$	$ \mathcal{K}^\tau $ Formel (3.26)	$ \mathcal{K}^\tau $ Formel (3.28)
20	46 KW 2010	131	0,4538	0,3626	14	5
21	47 KW 2010	132	0,4462	0,3643	2	14
22	48 KW 2010	113	0,4535	0,3612	12	8
23	49 KW 2010	107	0,4580	0,3656	5	9
24	50 KW 2010	117	0,4588	0,3659	3	9
25	51 KW 2010	94	0,4475	0,3559	9	8
26	52 KW 2010	88	0,4514	0,3677	13	13
27	1 KW 2011	112	0,4544	0,3555	8	6
28	2 KW 2011	117	0,4559	0,3592	10	9
29	3 KW 2011	131	0,4466	0,3538	14	14
30	4 KW 2011	146	0,4503	0,3553	13	8
31	5 KW 2011	175	0,4540	0,3551	5	12
32	6 KW 2011	170	0,4449	0,3549	11	13
33	7 KW 2011	148	0,4474	0,3499	10	14
34	8 KW 2011	184	0,4584	0,3532	5	9
35	9 KW 2011	161	0,4552	0,3535	12	6
36	10 KW 2011	146	0,4579	0,3603	14	11
37	11 KW 2011	147	0,4594	0,3559	4	8
38	12 KW 2011	185	0,4683	0,3678	11	9
39	13 KW 2011	168	0,4560	0,3561	11	11
40	14 KW 2011	139	0,4500	0,3540	12	13
41	15 KW 2011	120	0,4516	0,3554	12	6
42	16 KW 2011	107	0,4513	0,3586	7	14
43	17 KW 2011	110	0,4550	0,3583	8	13
44	18 KW 2011	173	0,4459	0,3563	7	8

Zeitfenster τ		$ D^\tau $	$\overline{mer^\tau}$	$\overline{per^\tau}$	$ \mathcal{K}^\tau $ Formel (3.26)	$ \mathcal{K}^\tau $ Formel (3.28)
45	19 KW 2011	123	0,4537	0,3630	8	14
46	20 KW 2011	126	0,4585	0,3624	11	14
47	21 KW 2011	129	0,4481	0,3552	13	8
48	22 KW 2011	126	0,4410	0,3487	8	11
49	23 KW 2011	122	0,4540	0,3584	8	9
50	24 KW 2011	117	0,4531	0,3575	9	8
51	25 KW 2011	158	0,4500	0,3564	9	11
52	26 KW 2011	163	0,4485	0,3571	14	14

Tabelle 5.11: Testkonfiguration & Ergebnisse für 3. Quartal 2010 bis 2. Quartal 2011 (einschließlich).

Als Testdatensatz dient auch in diesem Test eine Stichprobe des Nachrichtenportals Spiegel Online (vgl. Kapitel 5.3.2) aus der Rubrik 'Politik' mit insgesamt 6952 Dokumenten. Für die Rechner-Konfiguration 3 (2) beläuft sich die Rechenlaufzeit für Formel (3.28) auf 3h 1 min bzw. 1h 29 min. Dies entspricht ungefähr der 4-fachen Laufzeit des in Kapitel 5.4.3 beschriebenen Tests.

Folgende Parameterkonfigurationen werden verwendet:

- Verkleinertes Lokales Wörterbuch \mathcal{L}' mit Stoppwortentfernung mit $|\mathcal{L}'| = Z = 20.000$
- Cosinusmaß als Distanzmaß (vgl. Kapitel 3.2)
- Merkmalsextraktion (vgl. Formel (3.4))
- Hierarchisches Clusterverfahren (Ward-Verfahren, vgl. Formel (3.24) Kapitel 3.3.3)
- Ellbogenkriterium nach Formel (3.26) und Formel (3.28) (vgl. Kapitel 3.3.4) mit Beschränkung von $|\mathcal{K}|$ auf $\mathbb{N} \cap [2, 15]$

- Relationsmatrizen mit Relationsschranken $dis_{lb} = 0,47$ und $dis_{ub} = 0,56$
- Relevante Betrachtungszeiträume mit Betrachtungsspannen $\tau - t_x + 1 \leq m_\tau \in \{6, 11, 26, 27, 45, 49, 50\}$ zur Bildung der Mono-Themen-Graphen

Im Anhang A sind die Dendrogramme für die 52 Zeitfenster zu finden.

Die Evolution der Mono-Themen-Graphen wird an den ausgewählten Themenkomplexen 'Taliban', 'Bin Laden', 'Atomausstieg' und 'FDP' gezeigt.

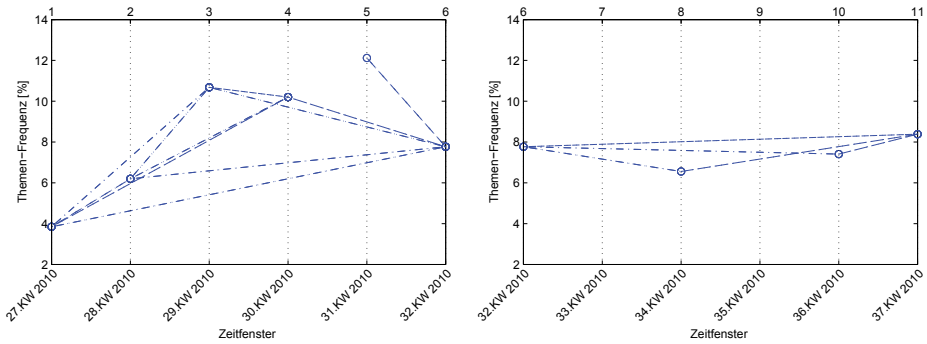
Themenkomplex 'Taliban'

Die Mono-Themen-Graphen des Themenkomplexes 'Taliban' sind nach Formel (3.26) erstellt worden. Ausgewählte Tag-Clouds dieser Mono-Themen-Graphen sind in Abbildung 5.24 dargestellt.



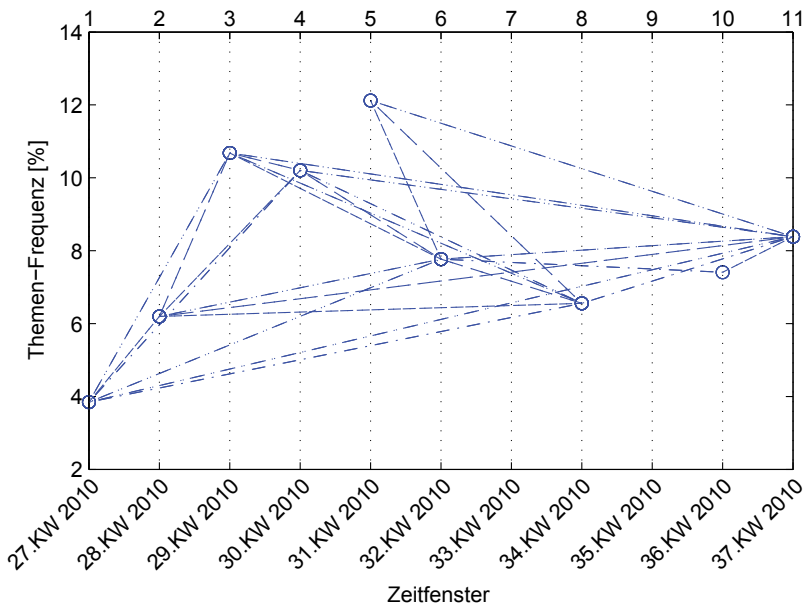
Abbildung 5.24: Ausgewählte Tag-Clouds für die Mono-Themen-Graphen in den Abbildungen 5.25 und 5.26.

Ein Mono-Themen-Graph einer Relation kann aufgebaut werden durch abschnittsweise Aneinanderreihung kleinerer aufeinander folgender Relationen (Konkatenation) oder durch Vergrößerung der Betrachtungsspanne m_{t_y} einer Relation $Rel(t_y - m_{t_y} + 1, t_y)$. Beim Vergleich beider Methoden zeigen sich deutliche Unterschiede in Bezug auf die Vernetzung, während Clusterpositionen gleich bleiben. Es können lediglich zusätzliche Knoten mit Kanten bei einer Vergrößerung von m_{t_y} hinzukommen (vgl. Kapitel 4.5).



(a) Mono-Themen-Graph $G_{(6,6)}$ der $Rel(1,6)_{(6,6)}$

(b) Mono-Themen-Graph $G_{(11,12)}$ der $Rel(6,11)_{(11,12)}$



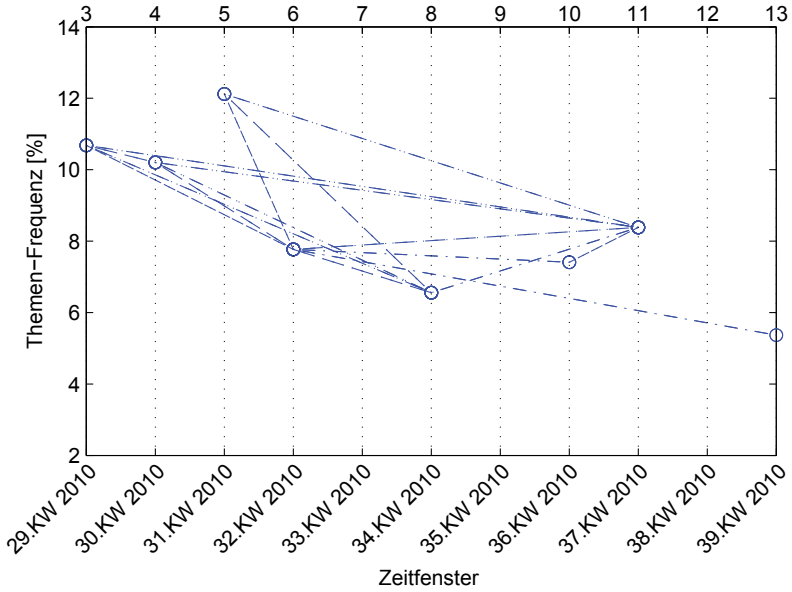
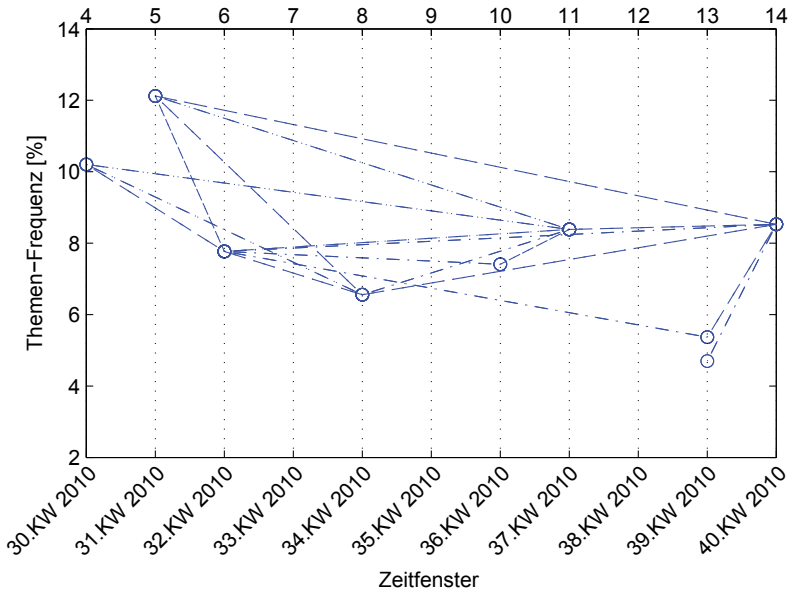
(c) Mono-Themen-Graph $G_{(11,12)}$ der $Rel(1,11)_{(11,12)}$

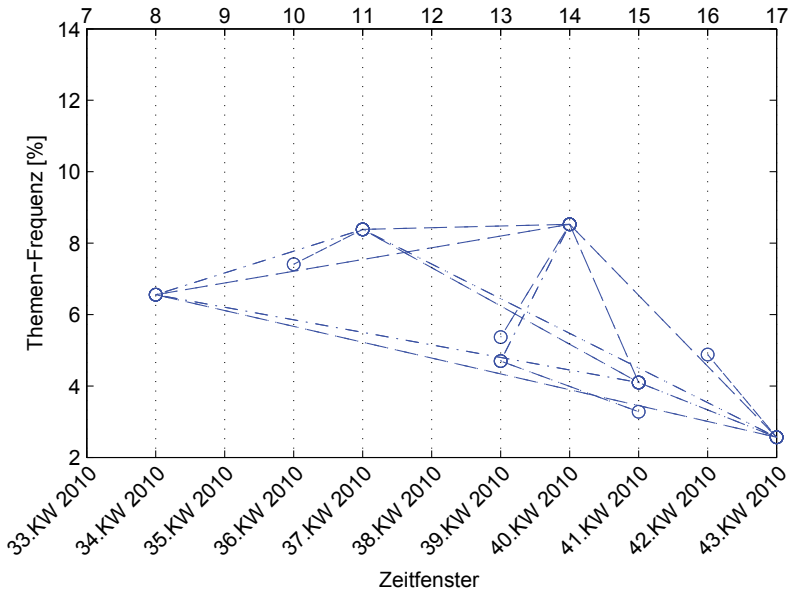
Abbildung 5.25: Mono-Themen-Graphen für den Themenkomplex 'Taliban'.

In Abbildung 5.25 (a) und (b) sind die beiden aufeinander folgenden Mono-Themen-Graphen der Relationen $Rel(1, 6)_{(6,6)}$ und $Rel(6, 11)_{(11,12)}$ mit einer Betrachtungsspanne von $m_{\{6,11\}} = 6$ für den Themenkomplex 'Taliban' gegeben. Es ist zu erkennen, dass eine Konkatenation der beiden aufeinander folgenden Mono-Themen-Graphen nicht dem Mono-Themen-Graphen der Relation $Rel(1, 11)_{(11,12)}$ (mit Betrachtungsspanne $m_{11} = 11$, Abbildung 5.25 (c)) dieses Themenkomplexes entspricht. Kanten zwischen gefundenen Themen-Clustern, die die 'Konkatenationsgrenze', also die Grenze $t_2 = 6$ des Betrachtungszeitraums $[t_1 = 1, t_2 = 6]$ überschreiten, werden bei einfacher Verknüpfung nicht gefunden. Dies erklärt sich durch die fehlenden Relationsberechnungen zwischen Clusterings der beiden Relationen $Rel(1, 6)$ und $Rel(6, 11)$. Die Relationsmatrizen für zwei Clusterings unterschiedlicher Relationen müssten erst bereitgestellt werden, was Rechenzeitprobleme erzeugen kann. Die Verknüpfung zweier aufeinander folgender Relationen $Rel(1, 6)$ und $Rel(6, 11)$ erfolgt an der Schnittstelle beider Relationen ausschließlich durch gemeinsame Knoten an der Konkatenationsgrenze.

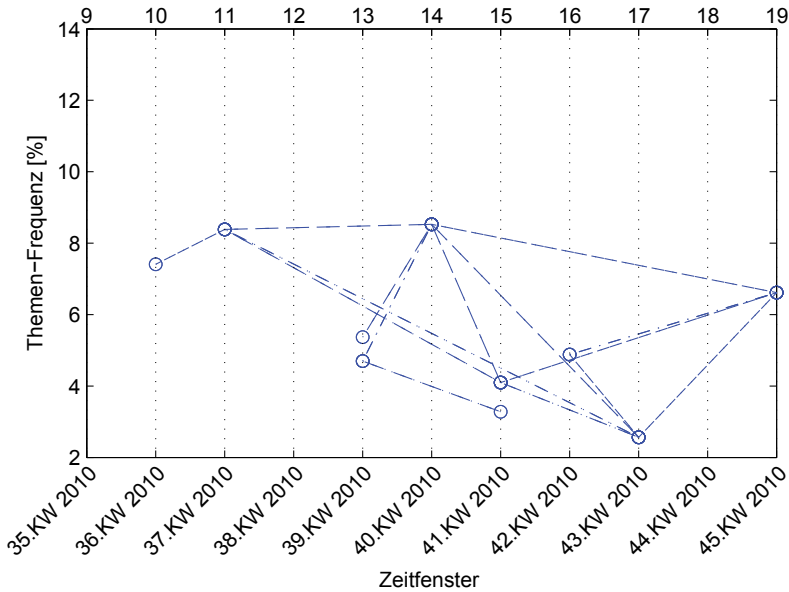
Der Mono-Themen-Graph aus Abbildung 5.25 (c) entwickelt sich im Zeitverlauf weiter. Die folgenden Abbildungen 5.26 (a) bis (c) zeigen diese Entwicklung an einigen ausgewählten Relationen: Relation $Rel(3, 13)_{(13,5)}$ in Abb. 5.26 (a), Relation $Rel(4, 14)_{(14,1)}$ in Abb. 5.26 (b), Relation $Rel(7, 17)_{(17,7)}$ in Abb. 5.26 (c), Relation $Rel(9, 19)_{(19,3)}$ in Abb. 5.26 (d).

Das 'Graphenmuster' des Mono-Themen-Graphen ändert sich bei der Transition des Betrachtungszeitraums $[t_x, t_y]$ sukzessive. Dabei bleibt das Muster relativ konstant, wenn sich die Knoten mit ihren zugehörigen Kanten innerhalb des Betrachtungszeitraums 'bewegen'. Wenn Knoten, bedingt durch die Verschiebung des Betrachtungszeitraums $[t_x, t_y]$ nicht mehr erscheinen, bzw. neu hinzu kommen, kann sich das Muster erheblich verändern, da Kanten zu 'alten' Knoten wegfallen und/oder Kanten zu 'neuen' Knoten entstehen. Vielfältig vernetzte Knoten, die dem verschobenen Betrachtungszeitraum nicht mehr angehören, bzw. neu erscheinende Knoten, können das Muster besonders stark beeinflussen.

(a) Mono-Themen-Graph $G_{(13,5)}$ der $Rel(3,13)_{(13,5)}$ (b) Mono-Themen-Graph $G_{(14,1)}$ der $Rel(4,14)_{(14,1)}$



(c) Mono-Themen-Graph $G_{(17,7)}$ der $Rel(7,17)_{(17,7)}$



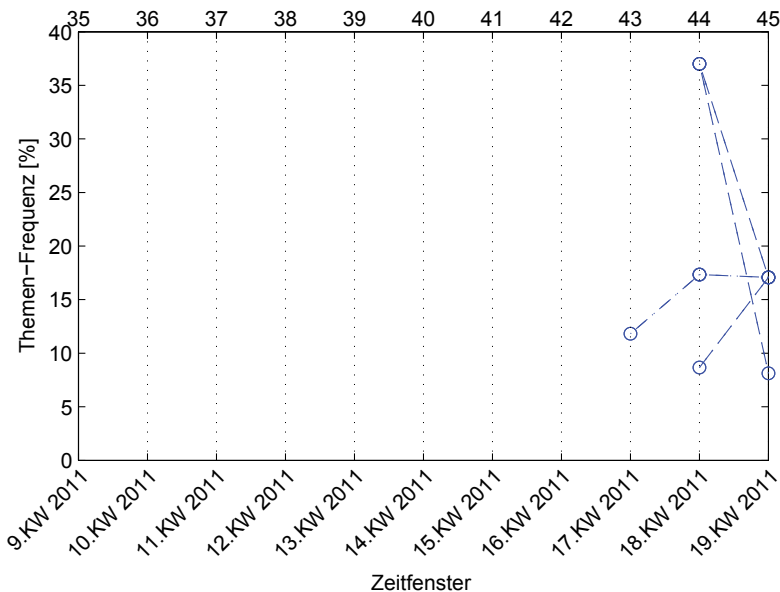
(d) Mono-Themen-Graph $G_{(19,3)}$ der $Rel(9,19)_{(19,3)}$

Abbildung 5.26: Evolution der Mono-Themen-Graphen für den Themenkomplex 'Taliban'.

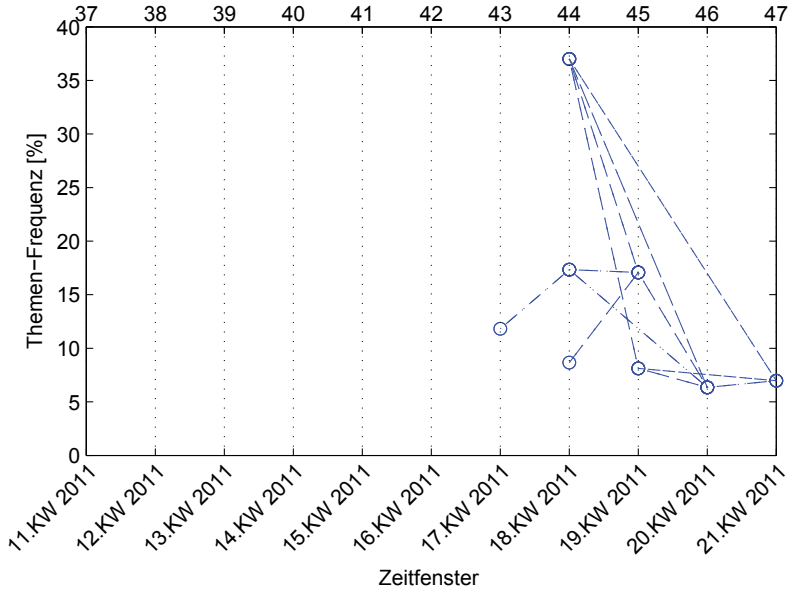
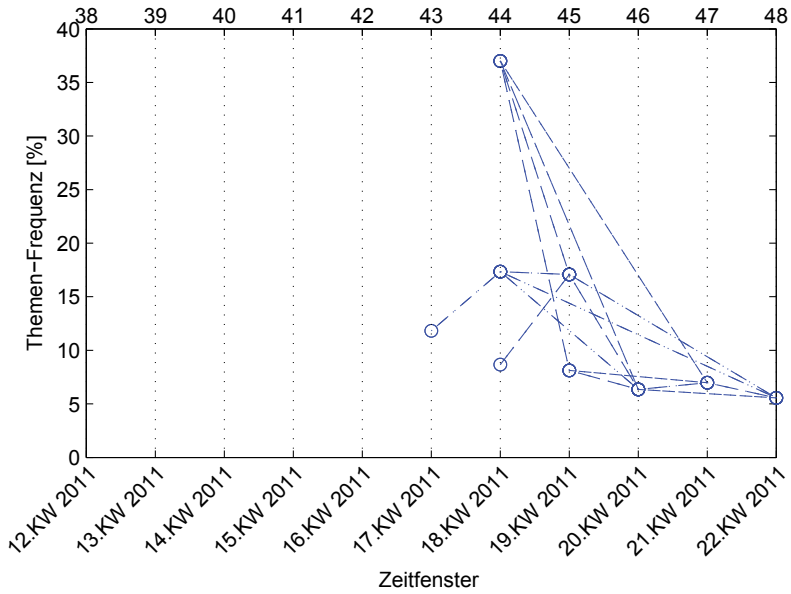
Die Betrachtung der ausgewählten Tag-Clouds dieses Themen-Graphen zeigt die Person Bin Laden und die Tötung Bin Ladens im Spannungsfeld zwischen der Großmacht USA, repräsentiert durch den Präsidenten Barack Obama, und der Terrororganisation Al-Qaida. Auch die geographische Verortung der Geschehnisse mit ihrem zeitlichen Auftreten ist aus den Schlagworten der Tag-Clouds abzulesen (z.B.: Pakistan, Islamabad).

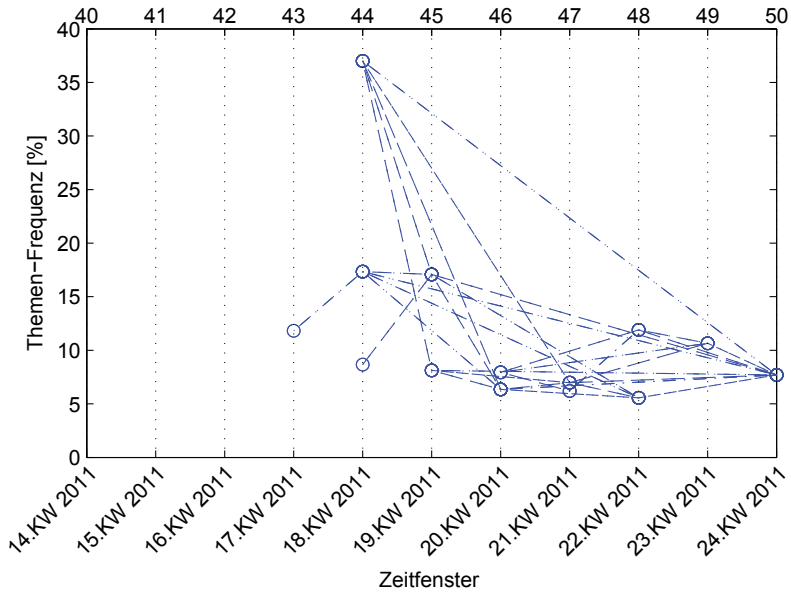
Bemerkenswert ist die thematische Wichtigkeit des Themenkomplexes 'Bin Laden' in dem Zeitfenster $\tau = 44$. Bei Betrachtung der Themen-Frequenzen f für die drei Cluster $f((44, 1)) = 8,6705 \times 10^{-2}$, $f((44, 2)) = 36,9942 \times 10^{-2}$ und $f((44, 7)) = 17,3410 \times 10^{-2}$ des zugehörigen Mono-Themen-Graphen wird deutlich, dass über 60% der Dokumente von D^{44} diesem Themenkomplex angehören. Dieses Thema hat alle anderen Themen in dieser Woche journalistisch in den Hintergrund rücken lassen.

Auch das schnelle Desinteresse an gehypten Themen wird hier deutlich: Innerhalb weniger Wochen fällt der Themenkomplex von einem Cluster mit Themen-Frequenz von ungefähr 36% ($f((44, 2)) = 36,9942 \times 10^{-2}$) auf unter 8% ($f((50, 2)) = 7,6923 \times 10^{-2}$) ab.

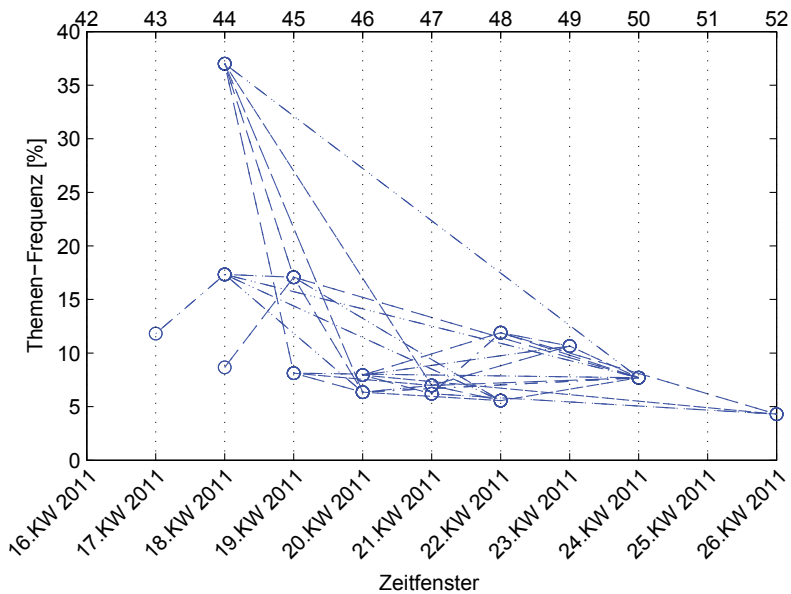


(a) Mono-Themen-Graph $G_{(45,3)}$ der $Rel(35, 45)_{(45,3)}$

(b) Mono-Themen-Graph $G_{(47,10)}$ der $Rel(37, 47)_{(47,10)}$ (c) Mono-Themen-Graph $G_{(48,1)}$ der $Rel(38, 48)_{(48,1)}$



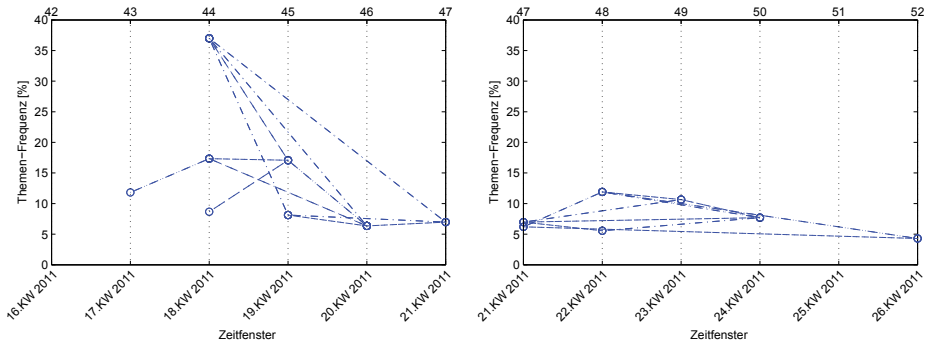
(d) Mono-Themen-Graph $G_{(50,2)}$ der $Rel(40, 50)_{(50,2)}$



(e) Mono-Themen-Graph $G_{(52,13)}$ der $Rel(42, 52)_{(52,13)}$

Abbildung 5.28: Evolution der Mono-Themen-Graphen für den Themenkomplex 'Bin Laden'.

Abbildung 5.29 (a) und (b) zeigt die Aufspaltung des Mono-Themen-Graphen der Relation $Rel(42, 52)_{(52,13)}$ aus Abbildung 5.28 (e) in zwei Teil-Relationen $Rel(42, 47)_{(47,10)}$ und $Rel(47, 52)_{(52,13)}$. Dabei steigt zwar die Übersichtlichkeit, jedoch mit einem einhergehenden Informationsverlust. Bei einer Konkatenation dieser beiden Teil-Relationen gibt es auch hier keine die Konkatenationsgrenze überschreitenden Kanten. Somit gilt: $Rel(42, 52)_{(52,13)} \neq Rel(42, 47)_{(47,10)} \cup Rel(47, 52)_{(52,13)}$.



(a) Mono-Themen-Graph $G_{(47,10)}$ der $Rel(42, 47)_{(47,10)}$

(b) Mono-Themen-Graph $G_{(52,13)}$ der $Rel(47, 52)_{(52,13)}$

Abbildung 5.29: Mono-Themen-Graphen für Teil-Relationen $Rel(42, 47)_{(47,10)}$ und $Rel(47, 52)_{(52,13)}$.

Zum Vergleich ist in Abbildung 5.30 der Mono-Themen-Graph $G_{(52,13)}$ für die gleiche Relation $Rel(42, 52)_{(52,13)}$ berechnet nach Formel (3.28) gegeben. Das Graphenmuster ähnelt auch hier dem entsprechenden Graphenmuster nach Formel (3.26) (vgl. Abbildung 5.28 (e)).

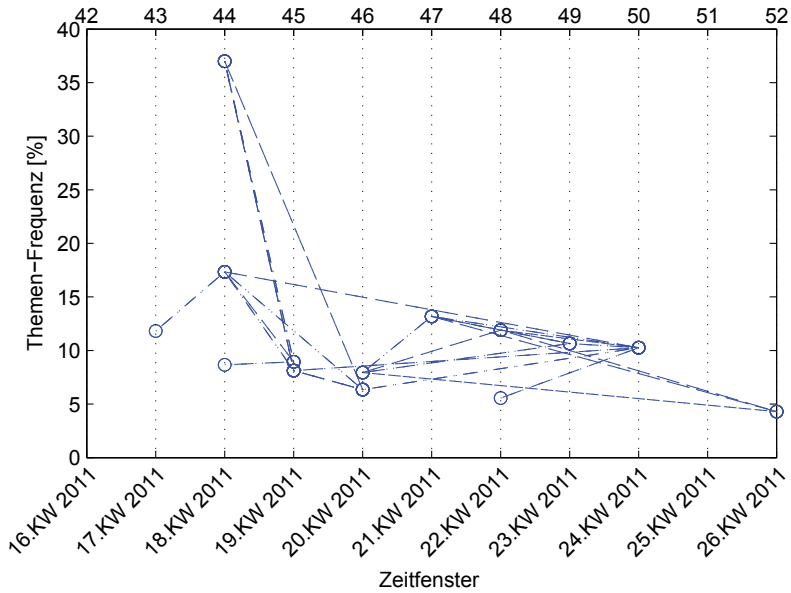


Abbildung 5.30: Mono-Themen-Graph $G_{(52,13)}$ der $Rel(42, 52)_{(52,13)}$ nach Formel (3.28).

Themenkomplex 'Atomausstieg'

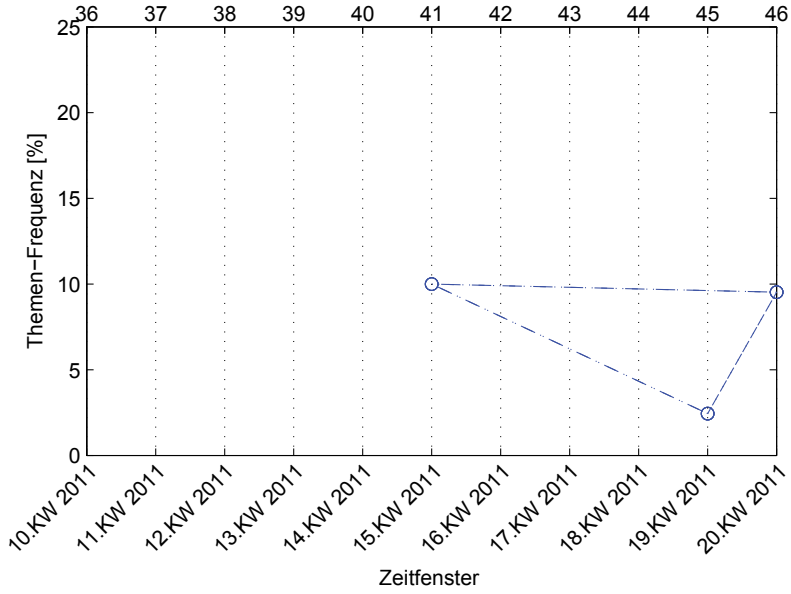
Im Gegensatz zu den Graphen der Themenkomplexe 'Taliban' und 'Bin Laden' wurden die Mono-Themen-Graphen des Themenkomplexes 'Atomausstieg' nach Formel (3.28) erstellt. Einige ausgewählte Tag-Clouds dieser Mono-Themen-Graphen sind in den Abbildungen 5.31 (und im Weiteren in 5.34 und 5.36) dargestellt.



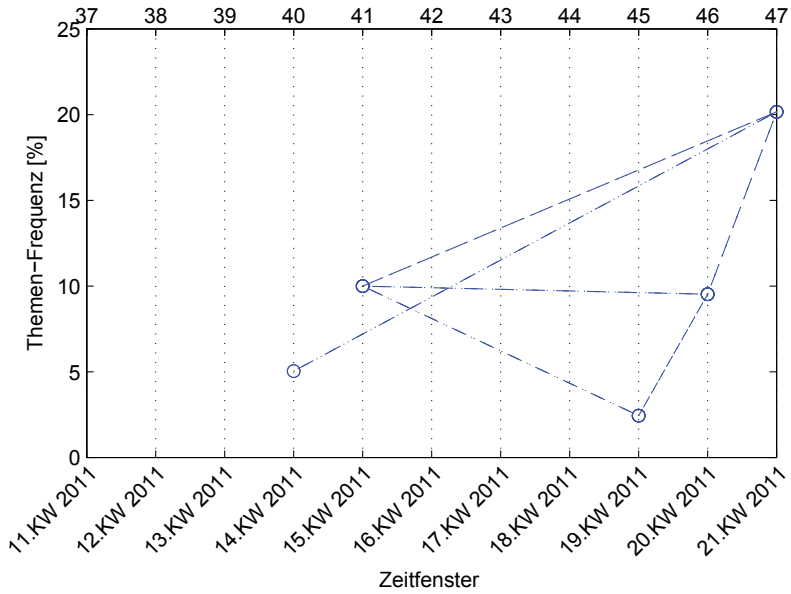
Abbildung 5.31: Ausgewählte Tag-Clouds für die Mono-Themen-Graphen.

Der Beginn der thematischen Entwicklung des Mono-Themen-Graphen (mit einer Betrachtungsspanne $m_{t_y} = 11$) ist abhängig von der Lage des Betrachtungszeitraums. Liegt der Anfang des Betrachtungszeitraums in Zeitfenster 36, dann erscheint erstmalig ein Knoten mit dem Thema 'Atomausstieg' in Zeitfenster 41 (vgl. Abbildung 5.32 (a), $Rel(36, 46)_{(46,10)}$). Wandert der Betrachtungszeitraum weiter, erscheint das Thema bereits in Zeitfenster 40 mit einem dem Mono-Themen-Graphen über eine Kante neu hinzugefügten Knoten (vgl. Abbildung 5.32 (b), $Rel(37, 47)_{(47,8)}$).

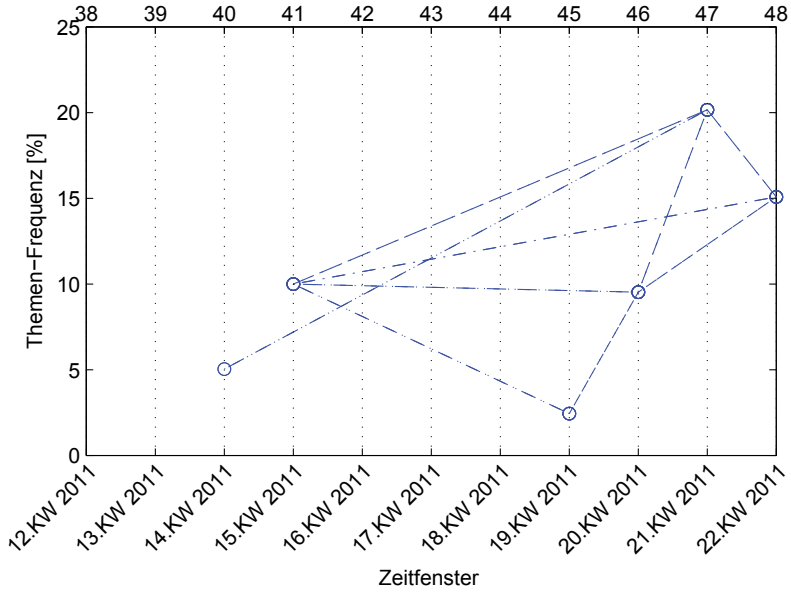
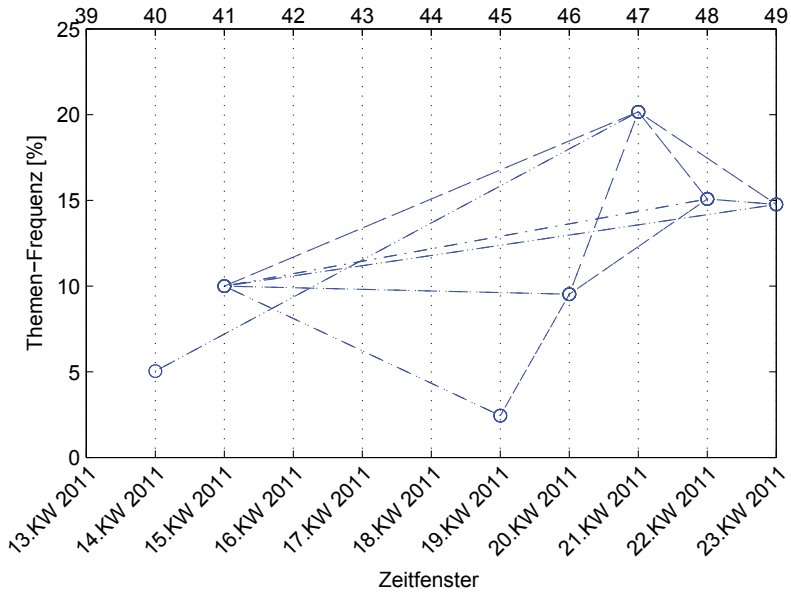
Die weitere Evolution des Mono-Themen-Graphen bis Zeitfenster 52 lässt sich über Abbildung 5.32 verfolgen.

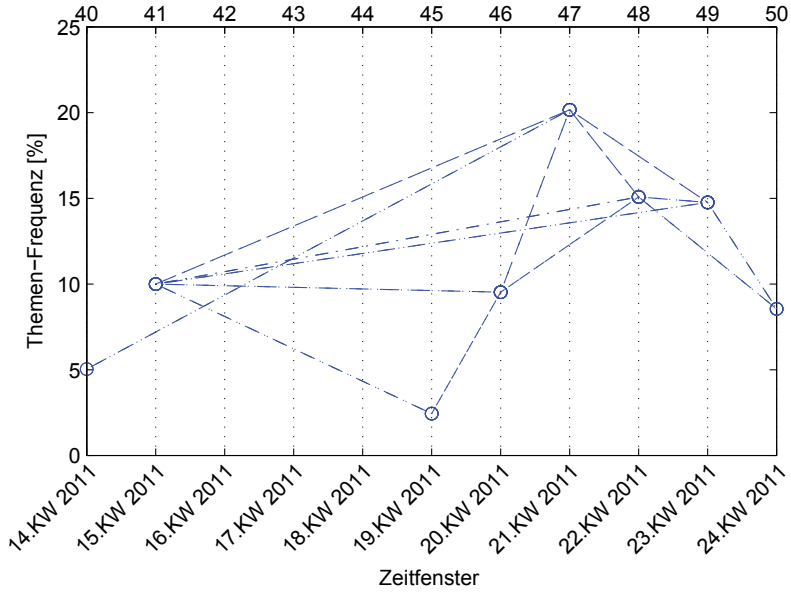


(a) Mono-Themen-Graph $G_{(46,10)}$ der $Rel(36, 46)_{(46,10)}$

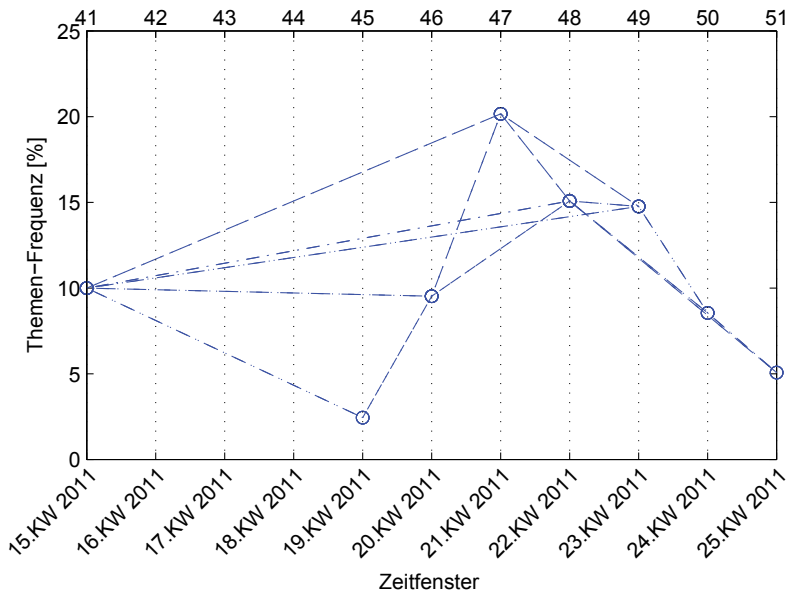


(b) Mono-Themen-Graph $G_{(47,8)}$ der $Rel(37, 47)_{(47,8)}$

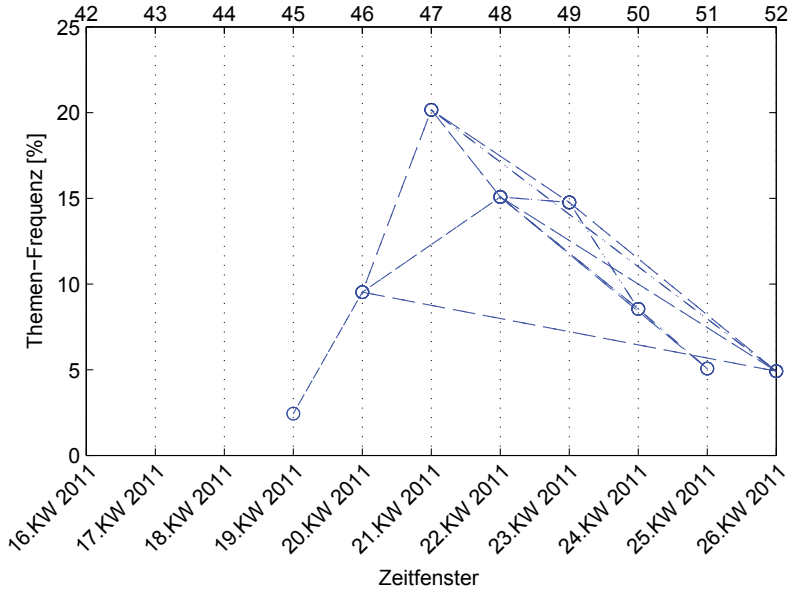
(c) Mono-Themen-Graph $G_{(48,8)}$ der $Rel(38,48)_{(48,8)}$ (d) Mono-Themen-Graph $G_{(49,5)}$ der $Rel(39,49)_{(49,5)}$



(e) Mono-Themen-Graph $G_{(50,5)}$ der $Rel(40, 50)_{(50,5)}$



(f) Mono-Themen-Graph $G_{(51,10)}$ der $Rel(41, 51)_{(51,10)}$



(g) Mono-Themen-Graph $G_{(52,4)}$ der $Rel(42, 52)_{(52,4)}$

Abbildung 5.32: Evolution der Mono-Themen-Graphen für den Themenkomplex 'Atomausstieg' bei $m_{\{46,47,48,49,50,51,52\}} = 11$.

Die nachfolgende Abbildung 5.33 zeigt den Graphen bei einer Verbreiterung der Betrachtungsspanne auf $m_{52} = 27$ im Betrachtungszeitraum [26, 52].

Dabei ist zu beachten, dass dadurch der Graph in der entsprechenden Abbildung zeitlich gestaucht erscheint (vgl. Abbildung 5.32 und 5.33).

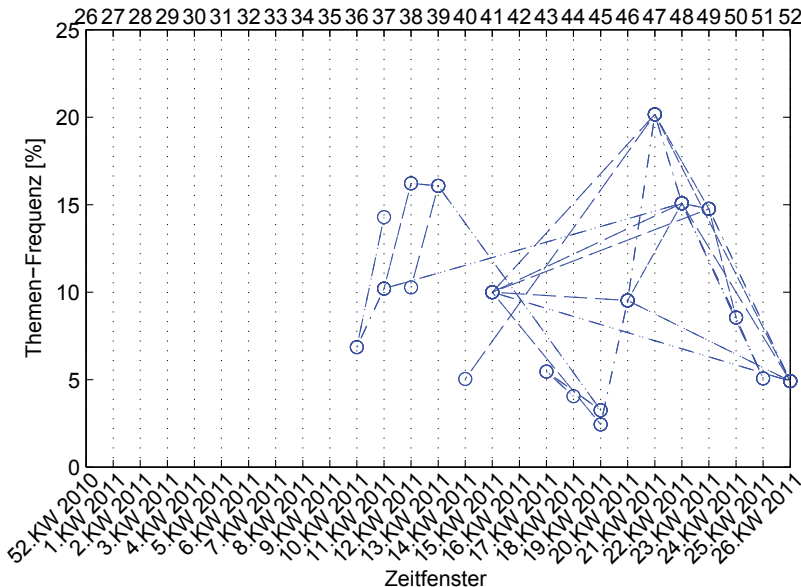


Abbildung 5.33: Verbreiterung der Betrachtungsspanne auf $m_{52} = 27$ des Mono-Themen-Graphen 'Atomausstieg' auf die Relation $Rel(26, 52)_{(52,4)}$.

Durch die Vergrößerung der Betrachtungsspanne werden dem Themenkomplex Cluster aus der Vergangenheit zugeordnet, die diesem zuvor unbekannt waren (vgl. Knoten in den Zeitfenstern 36 bis 39 in Abbildung 5.33).

Einige Tag-Clouds für diese neuen Cluster finden sich in Abbildung 5.34. Bei Betrachtung der Tag-Clouds mit den Schlagworten 'Japan', 'Erdbeben', 'Atomkraft', 'Katastrophe', 'Baden-Württemberg', 'Grünen', ergeben sich weitere thematische Zusammenhänge zum Themenkomplex 'Atomausstieg'. Dem Erdbeben in Japan am 11. März 2011 (KW 10) mit der darauf folgenden Atomkatastrophe in Fukushima folgte eine politische Diskussion in Deutschland, die in Baden-Württemberg bei der Landtagswahl am 27. März 2011 (KW 12 bzw. KW 13) zum Regierungswechsel führte. Unter dem Eindruck der dramatischen Ereignisse in Japan wurde erstmals eine CDU geführte Landesregierung durch eine den Ministerpräsidenten stellende 'Grüne' Landesregierung ersetzt.

Der Mono-Themen-Graph der Relation $Rel(1, 49)_{(49,5)}$ in Abbildung 5.35 zeigt eine Reihe strukturaler Eigenschaften, die bereits in Kapitel 4.4 dargelegt worden sind.

So findet sich z.B. in Knoten (17, 6) mit $f((17, 6)) = 3, 2051 \times 10^{-2}$ eine 'Vereinigung' (Merger) von zwei in der Vergangenheit bzgl. Knoten (17, 6) liegenden Knoten (10, 7) mit $f((10, 7)) = 8, 1481 \times 10^{-2}$ und Knoten (11, 1) mit $f((11, 1)) = 9, 5808 \times 10^{-2}$.

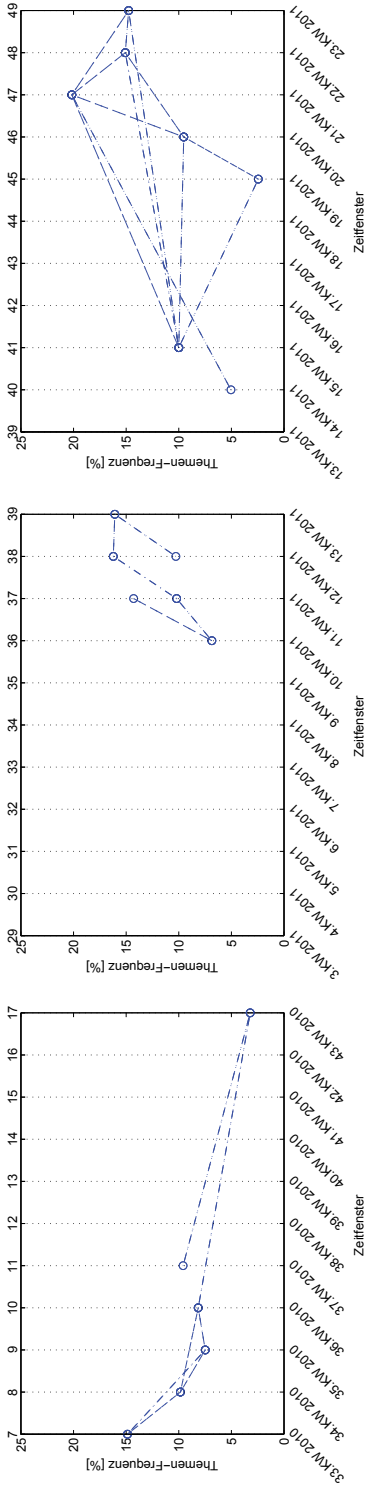
Der umgekehrte Fall, also eine 'Aufspaltung' (Split) tritt u.a. in Knoten (7, 5) mit $f((7, 5)) = 14, 8760 \times 10^{-2}$ auf. Dieser Knoten (7, 5) steht mit zwei in der Zukunft bzgl. Knoten (7, 5) liegenden Knoten (8, 5) mit $f((8, 5)) = 9, 8361 \times 10^{-2}$ und Knoten (9, 9) mit $f((9, 9)) = 7, 4830 \times 10^{-2}$ in Relation.

Der Knoten (7, 5) ist für die Relation $Rel(1, 49)_{(49,5)}$ 'absolut neu', da von ihm aus kein Pfad in die Vergangenheit für den Betrachtungszeitraum [1, 49] existiert.

Knoten (11, 1) mit $f((11, 1)) = 9, 5808 \times 10^{-2}$ ist für $r = 5$ innerhalb der Relation $Rel(1, 11 + 5)$ 'r-temporär neu', da Knoten (11, 1) keine Kante zu einem Knoten aus einem zeitlich vorangehenden Zeitfenster besitzt, in $Rel(1, 49)$ aber ein Pfad über den Knoten (17, 6) in der Zukunft von Knoten (11, 1) zu einem Knoten (z.B. Knoten (7, 5)) in die Vergangenheit von Knoten (11, 1) existiert. Zudem ist der Knoten (11, 1) r-direkt erloschen, da für $Rel(1, 11 + 5)$ keine Kante zu einem in der Zukunft bzgl. Knoten (11, 1) liegenden Knoten existiert. Knoten (17, 6) ist für $r = 19$ 'r-temporär erloschen', da für die Relation $Rel(1, 17 + 19)$ kein Pfad von dem Knoten (17, 6) zu einem Knoten in der Zukunft existiert. Knoten (17, 6) gilt für die Relation $Rel(1, 49)$ jedoch nicht als 'absolut erloschen', weil für $r > 19$ ein Pfad von Knoten (17, 6) in die Zukunft gefunden werden kann (z.B. zu Knoten (37, 3) mit $f((37, 3)) = 10, 2041 \times 10^{-2}$).

Der durch die nochmalige Vergrößerung auf den Betrachtungszeitraum [1, 49] entstandene Mono-Themen-Graph der Relation $Rel(1, 49)_{(49,5)}$ knüpft einen thematischen Zusammenhang zu einem temporal noch weiter in der Vergangenheit liegenden Teil-Themenkomplex, dessen Themen-Cluster in den Zeitfenstern 7 bis 17 liegen. Die neue Verknüpfung erfolgt über eine einzige Kante ((9, 9), (37, 3)) zwischen den Knoten in Zeitfenster 9 und 37.

Die Tag-Clouds dieser Zeitfenster werden in Abbildung 5.36 gezeigt. Der hinzugefügte Teil-Themenkomplex wird charakterisiert durch die Schlagworte seiner Tag-Clouds wie zum Beispiel 'Kanzlerin', 'Merkel', 'Atomkraftwerke', 'Verlängerung', 'erneuerbaren', 'Energien', 'Gorleben', 'Szenario'. Dieser Teil-Komplex spiegelt die zu dieser Zeit geführte politische Diskussion um die Laufzeitverlängerung bestehender Atomkraftwerke in der Bundesrepublik Deutschland wider.



(a) Mono-Themen-Graph $G_{(17,6)}$ der $Rel(7, 17)_{(17,6)}$ (b) Mono-Themen-Graph $G_{(39,10)}$ der $Rel(29, 39)_{(39,10)}$ (c) Mono-Themen-Graph $G_{(49,5)}$ der $Rel(39, 49)_{(49,5)}$

Abbildung 5.37: Die Teil-Themenkomplexe des Mono-Themen-Graphen 'Atomausstieg' der $Rel(1, 49)_{(49,5)}$.

Die Vergrößerung der Betrachtungsspanne m_{t_y} eines Themen-Graphen hat gegenüber einer einfachen Aneinanderreihung von Teil-Themen-Graphen den großen Vorteil, dass dadurch Zusammenhänge zwischen zeitlich weit auseinanderliegenden Themen aufgespürt werden können. Allerdings ist eine vorliegende Betrachtungsspanne m_{t_y} nicht beliebig erweiterbar. Der steigende Rechenaufwand für einen vergrößerten Betrachtungshorizont $[t_y - m_{t_y} + 1, t_y]$ wirkt sich auf Grund der zu berechnenden $m_{t_y} \times (m_{t_y} - 1)/2$ Relationsmatrizen, begrenzend aus (vgl. Kapitel 4.5).

Ein entscheidender Faktor für die Begrenzung der Betrachtungsspanne ist auch die Übersichtlichkeit der Mono-Themen-Graphen. So kann es bei einer relativ geringen Vergrößerung von m_{t_y} zu einer explosionsartigen Zunahme an Knoten und Kanten kommen (vgl. Abbildung 5.38).

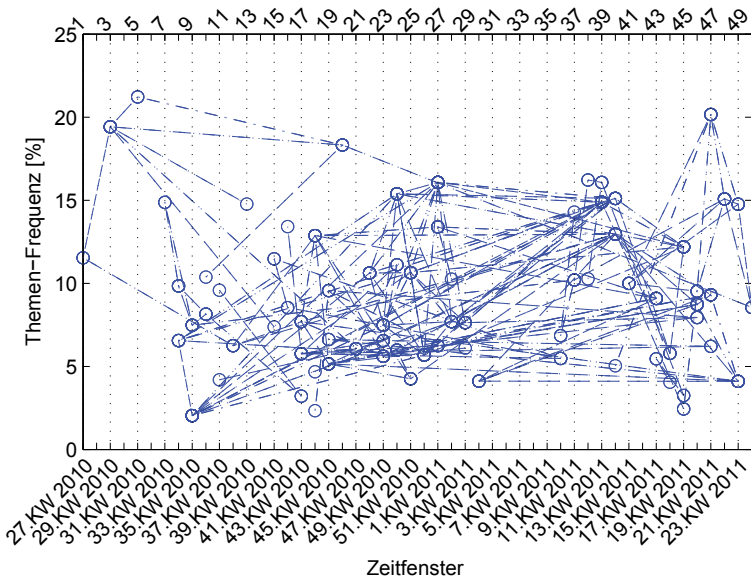


Abbildung 5.38: Mono-Themen-Graph $G_{(50,5)}$ der $Rel(1,50)_{(50,5)}$.

Im vorliegenden Beispiel Abbildung 5.38 erscheint bei der Vergrößerung um ein weiteres Zeitfenster ein einziger neuer Knoten in diesem neu hinzugefügten Zeitfenster. Dieser Knoten $(50, 5)$ mit $f((50, 5)) = 8,5470 \times 10^{-2}$ ist ausschlaggebend für die explosionsartige Zunahmen an Kanten und Knoten im Graphen. Er bildet eine Kante zu einem in der Vergangenheit liegenden Knoten $(20, 2)$ mit $f((20, 2)) = 18,3206 \times 10^{-2}$ in Zeitfenster 20. Diese Kante fungiert quasi als Brücke zu einem bereits bestehenden Mono-Themen-Graphen eines anderen Themenkomplexes. Wenn, wie im vorliegenden Beispiel, solche Brücken-Cluster komplexe Themen-Graphen miteinander kombinieren, geht nicht nur die Übersichtlichkeit der Darstellung verloren. Es können darüber hinaus auch Cluster in den Graphen eingebunden werden, die untereinander praktisch keine thematischen Ähnlichkeiten mehr aufweisen.

Themenkomplex 'FDP'

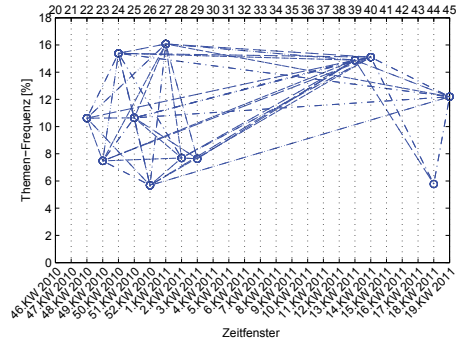
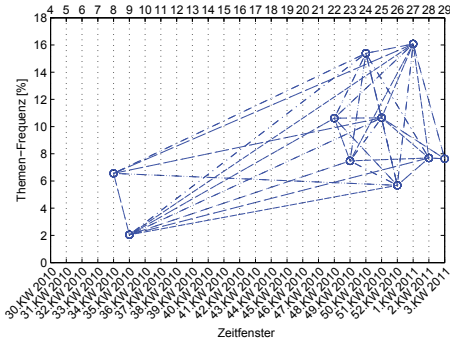
Ein weiteres Beispiel strukturell graphisch ansprechender Mono-Themen-Graphen gehört dem Themenkomplex 'FDP' an, dessen ausgewählte Tag-Clouds in Abbildung 5.39 gezeigt werden.



Abbildung 5.39: Ausgewählte Tag-Clouds für die Mono-Themen-Graphen in Abbildung 5.40.

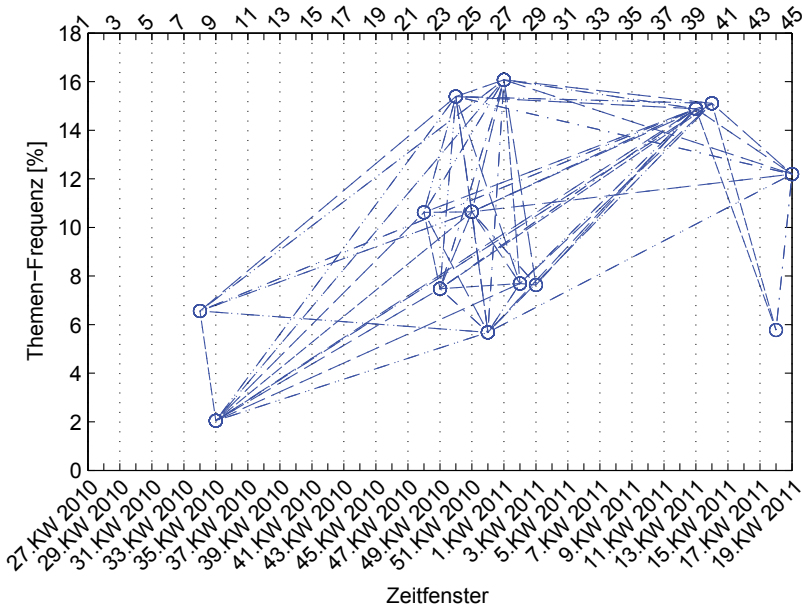
Die Mono-Themen-Graphen (vgl. Abbildung 5.40) wurden für dieses Beispiel ebenfalls nach Formel (3.28) generiert.

Sowohl die Tag-Clouds als auch das Graphenmuster beschreiben ein abgegrenztes politisches Thema, das zunächst mit geringer Intensität beginnt und nach längerer Pause von ca. drei Monaten rund zwei Monate lang mit schwankender Frequenz präsent ist. Nach einer weiteren Pause von ca. zehn Wochen fokussieren die Themen-Cluster auf die Zeitfenster 39 und 40.



(a) Mono-Themen-Graph $G_{(29,11)}$ der $Rel(4, 29)_{(29,11)}$

(b) Mono-Themen-Graph $G_{(45,14)}$ der $Rel(20, 45)_{(45,14)}$



(c) Mono-Themen-Graph $G_{(45,14)}$ der $Rel(1, 45)_{(45,14)}$

Abbildung 5.40: Mono-Themen-Graphen für den Themenkomplex 'FDP'.

Stellvertretend für die Cluster des Themen-Graphen werden in Abbildung 5.41 die Dokumente des fokussierten Clusters C_{13}^{40} aufgelistet (mit Titel und Untertitel). In Zeitfenster 40 werden von $|D^{40}| = 139$ Dokumenten 21 dem Cluster C_{13}^{40} zugeordnet, das entspricht einer Themen-Frequenz $f((40, 13)) = 15, 1079 \times 10^{-2}$.

13	'06.04.2011'	'Atomdebatte'	'Grüne erklimmen Rekordhoch, FDP stürzt ab'
13	'05.04.2011'	'Parteisaniierer Rösler'	'Krisen-FDP scharft sich um den Juniorchef'
13	'08.04.2011'	'Neue Koalitionsoptionen'	'Die Alles-ist-möglich-Republik'
13	'06.04.2011'	'Umfrage-Absturz'	'33 Weckrufe für Merkels Schlummerkoalition'
13	'06.04.2011'	'Liberale Krise'	'Genscher fordert radikaleren FDP-Umbau'
13	'06.04.2011'	'Niedergang eines Prestigepostens'	'Der halbierte Westerwelle'
13	'05.04.2011'	'Künftiger Chef-Liberaler'	'Wer ist Philipp Rösler?'
13	'05.04.2011'	'FDP in der Krise'	'Der Kuschel-Putsch'
13	'09.04.2011'	'FDP-Krise'	'Seehofer-Ultimatum erzürnt die Liberalen'
13	'05.04.2011'	'Machtwechsel'	'FDP-Spitze entscheidet sich für Rösler'
13	'05.04.2011'	'FDP-Führungswechsel'	'Rösler verspricht Neustart für lädierte Liberale'
13	'05.04.2011'	'FDP-Interna'	'Jungliberale küren Brüderle zum Lieblingsfeind'
13	'04.04.2011'	'FDP-Machtprobe'	'Brüderle rüstet zum Kampf gegen die Jungen'
13	'05.04.2011'	'Merkel und die FDP-Krise'	'Koalition außer Kontrolle'
13	'04.04.2011'	'FDP-Hoffnungsträger Rösler'	'Der junge Milde greift nach der Macht'
13	'04.04.2011'	'Liberale Krise'	'FDP-Veteran fordert Westerwelles Komplett-Rückzug'
13	'04.04.2011'	'Liberales Vakuum'	'FDP-Machtkampf bremst schnellen Chefwechsel'
13	'04.04.2011'	'FDP-Krise'	'Westerwelle gibt Amt des Vizekanzlers ab'
13	'04.04.2011'	'Vize-Kanzlerschaft'	'Ein bisschen Glanz, wenig Macht'
13	'04.04.2011'	'S.P.O.N. - Der Schwarze Kanal'	'Schmuseliberalismus, nein danke!'
13	'08.04.2011'	'Noch-FDP-Chef Westerwelle'	'Der Draußenminister'

Abbildung 5.41: Dokumente des Cluster C_{13}^{40} .

Kapitel 6

Zusammenfassung und Ausblick

Der Schwerpunkt der vorliegenden Arbeit ist die Entwicklung und Erprobung eines Modells zum Auffinden relevanter Themen und Trends innerhalb von online Dokumentenströmen. Das Modell ermöglicht die Darstellung von Relationen zwischen Themen verschiedener Zeitfenster als Themen-Graphen. Durch Variieren der Betrachtungszeitspannen können Beziehungen zwischen Themen verschiedener Zeitfenster in unterschiedlicher Komplexität abgebildet werden. Bei Verschiebung des Betrachtungszeitraums wird auch die Evolution eines Themen-Graphen (Themenkomplexes) verfolgbar. Neben den typischen Lebenszyklen wie Entstehen, Anwachsen, Abnehmen und Verschwinden eines Themas werden auch thematische Relationen zwischen Themen und deren Änderungen über die Zeit innerhalb eines Themenkomplexes sichtbar.

Nach einer kurzen Einleitung in Kapitel 1 wird in Kapitel 2 der Kontext der Arbeit innerhalb eines größeren Forschungsgebiets beleuchtet.

In Kapitel 3 werden die nötigen Modellgrundlagen definiert und erläutert. Zunächst wird unter anderem auf einen benötigten Referenzkorpus und die daraus erzeugten Referenzwerte, wie die Inverse Dokumentenfrequenz, eingegangen. Aus dem Referenzkorpus wird ein um Stoppworte bereinigtes Lokales Wörterbuch generiert, das für eine Dimensionsreduktion auf eine bestimmte Anzahl Wortformen begrenzt wird. Die Akkuratessse dieses Wörterbuchs kann mit verschiedenen Indizes überprüft werden. Dieses modifizierte Lokale Wörterbuch dient der Erstellung eines 'Vector Space' zur Dokumentenrepräsentation unter Einbeziehung der jeweiligen Termgewichte. Die Dokumentenvektoren werden mittels eines Unähnlichkeitsmaßes, hier dem Cosinus Maß, auf (Un-)Ähnlichkeit hin überprüft und ähnliche Dokumente zu Themen-Clustern gruppiert. Für das Clustern der Dokumente wird vornehmlich ein hierarchisches Clusterverfahren angewendet. Zur Bestimmung adäquater Clusteranzahlen, also Themen pro Zeitfenster, können verschiedene Kriterien genutzt werden. Die Arbeit verwendet das Ellbogenkriterium mit zwei verschiedenen Gütekriterien.

In Kapitel 4 wird auf Basis der beschriebenen Grundlagen ein Modell entwickelt, dass die Entdeckung von relevanten Themen, sowie deren Verfolgung über Zeitfenster hin-

weg erlaubt. Mit Hilfe von Relationsmatrizen zwischen Clusterings unterschiedlicher Zeitfenster eines Betrachtungshorizonts werden Beziehungen zwischen Themen-Clustern ermittelt. Durch eine geeignete Schranken-Überlegung werden thematisch ähnliche Themen-Cluster in einem Themen-Graphen über Kantenverbindungen bildlich dargestellt. Neben der zeitlichen Komponente eines Themen-Clusters kann über die Themen-Frequenz auch seine gegenwärtige Bedeutung in dem jeweiligen Zeitfenster angegeben werden. In dem Kapitel werden auch mögliche strukturelle Eigenschaften innerhalb von Themen-Graphen diskutiert, sowie die Auswirkung von Veränderungen der Betrachtungszeiträume und Betrachtungsspannen. Durch eine fortschreitende Verschiebung eines Betrachtungszeitraums ist eine kontinuierliche Themen- und Trend-Analyse möglich.

In Kapitel 5 erfolgt die Evaluation des implementierten Modells. Nach einer kurzen Beschreibung der Modellimplementierung und der verwendeten Rechnerkonfigurationen werden zwei für die Evaluation verfügbare Testdatensätze beschrieben. Dabei steht neben dem DeReKo Datensatz des Instituts für Deutsche Sprache auch ein Datensatz des online Nachrichtenportals Spiegel Online zur Verfügung, die in mehreren Testreihen untersucht werden.

Die erste Testreihe der Evaluation legt dabei das Hauptaugenmerk auf eine optimale 'Vector Space' Dimension in Bezug auf Worterkennungsrate, Berechnungslaufzeiten und Clusteringergebnisse im Vergleich mit vorklassifizierten Dokumenten des DeReKo. Die zweite Testreihe der Evaluation zeigt, dass ein Tracking, also Verfolgen von Themen über die Zeit hinweg, möglich ist und die mit einem hierarchischen Verfahren erhaltenen Dendrogramme den zu Grunde liegenden Testdaten und der definierten Testkonfiguration in ihrer Struktur entsprechen.

Die dritte und vierte Testreihe legen den Fokus auf die Erstellung der Themen-Graphen. Dabei werden mögliche Einflussparameter auf die Themen-Graphen, wie Zeitfenstergröße, verschiedene Gütekriterien für das Ellbogenkriterium, Betrachtungsspannenveränderungen wie auch das Fortschreiten von Betrachtungszeiträumen untersucht. Dabei kann gezeigt werden, dass die gefundenen Themenkomplexe in ihrer Struktur (Inhalt und Intensität) und ihrem zeitlichen Verlauf entsprechenden bekannten Ereignissen zugeordnet sind.

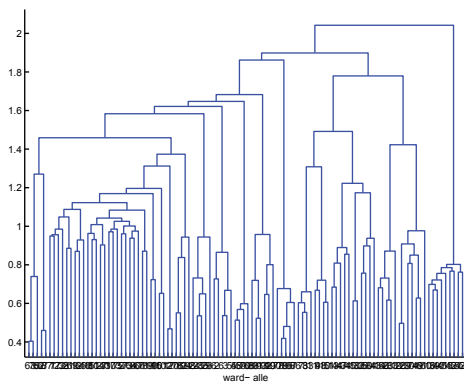
Mit gängigen Rechnerkonfigurationen (Stand 2014) lassen sich online Dokumente des Spiegel, z.B. aus der Rubrik 'Politik', in weniger als einer Stunde automatisch bearbeiten und in Form von Themen-Graphen einem End-Nutzer verfügbar machen.

Wie in Kapitel 5 gezeigt wurde, kann eine Erweiterung der Betrachtungsspanne zu einem besseren Verständnis temporaler Zusammenhänge eines Themenkomplexes führen. Als Problem erweist sich jedoch eine explosionartige Zunahme an Knoten und Kanten, die hierbei in einigen Fällen auftritt und die Überschaubarkeit für einen End-Nutzer erschwert. Als mögliche Lösung kann die Betrachtungsspanne dynamisch so festgelegt werden, dass die Anzahl der Knoten und Kanten eine bestimmte vordefinierte Obergrenze nicht überschreiten darf.

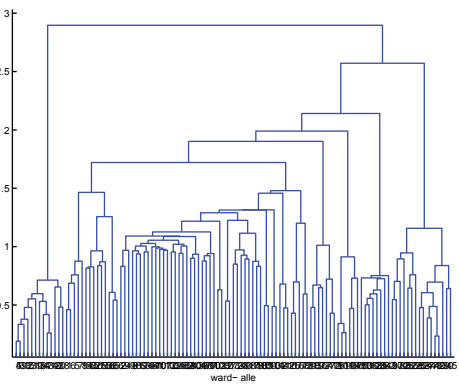
Eine andere Lösungsmöglichkeit für den Erhalt der Übersichtlichkeit ist die Fokussierung auf besonders interessierende Themen-Knoten. Ausgehend von diesen Knoten wird der bestehende Themen-Graph dahingehend eingeschränkt, dass nur Knoten in die Betrachtung mit eingehen, die über einen Pfad mit einer maximalen Anzahl an Kanten erreichbar sind (vgl. GAUL (2011)). Besonders interessierende Knoten können z.B. anhand von Suchworten bzw. Schlagworten vorgegeben werden. Dabei kann ein Ranking nach den ähnlichsten Centroiden erfolgen. Auch eine weitere Gewichtung, die das zeitliche Zurückliegen eines Knoten berücksichtigt, ist möglich.

Anhang A

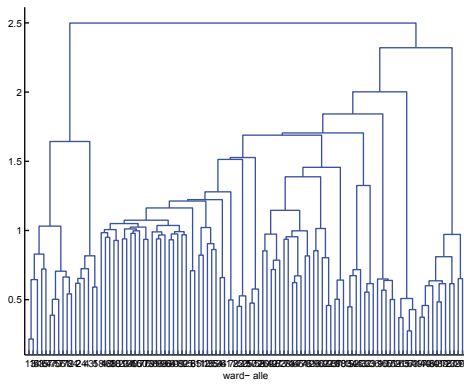
Zusatz zu 5.4.4



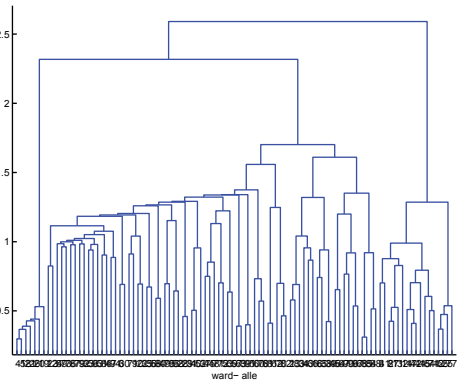
(1) Zeitfenster $\tau = 1$



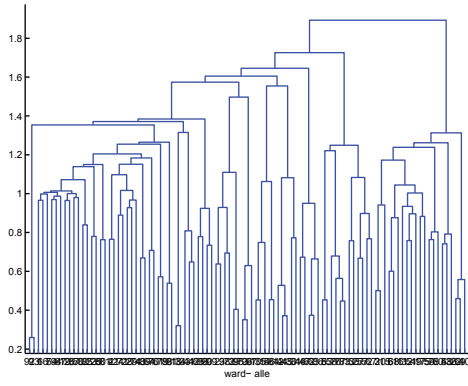
(2) Zeitfenster $\tau = 2$



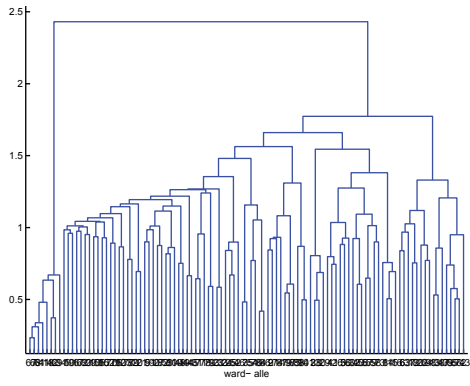
(3) Zeitfenster $\tau = 3$



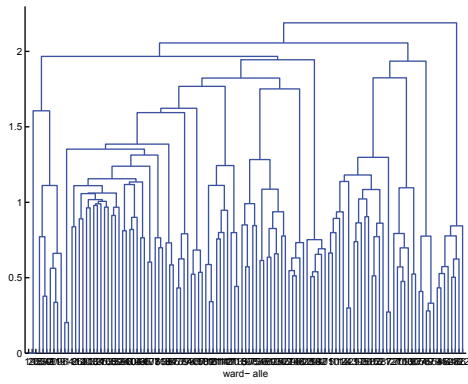
(4) Zeitfenster $\tau = 4$



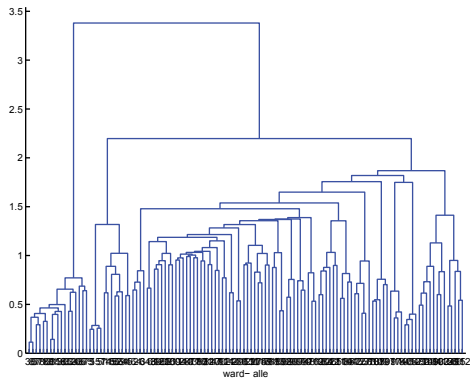
(5) Zeitfenster $\tau = 5$



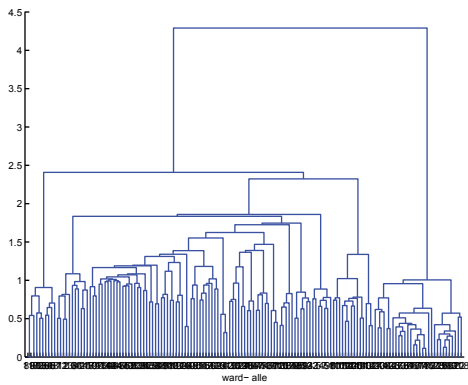
(6) Zeitfenster $\tau = 6$



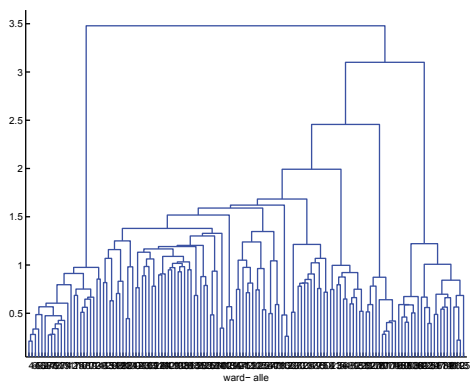
(7) Zeitfenster $\tau = 7$



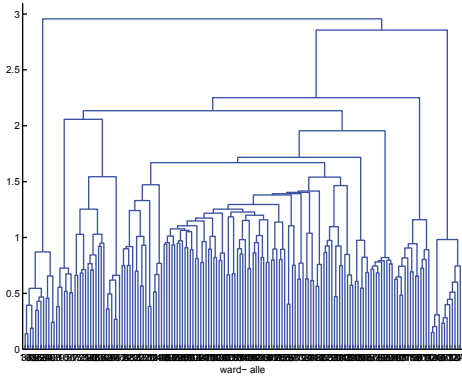
(8) Zeitfenster $\tau = 8$



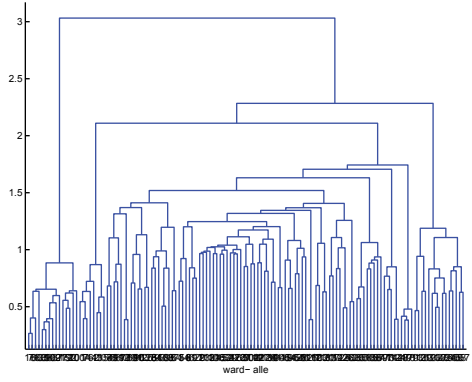
(9) Zeitfenster $\tau = 9$



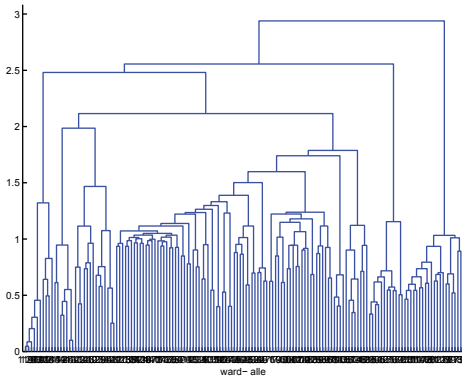
(10) Zeitfenster $\tau = 10$



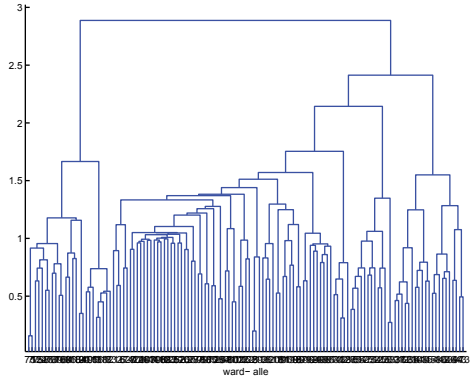
(11) Zeitfenster $\tau = 11$



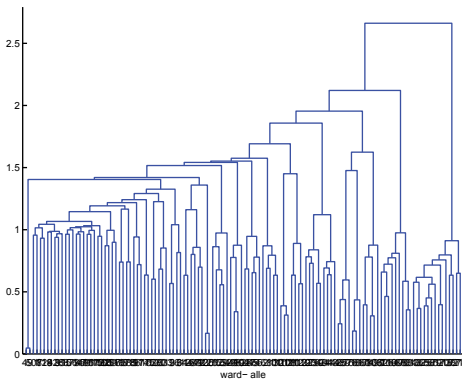
(12) Zeitfenster $\tau = 12$



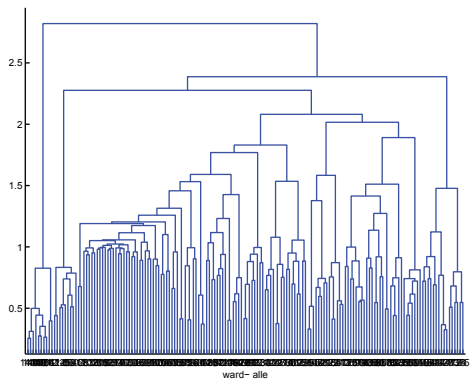
(13) Zeitfenster $\tau = 13$



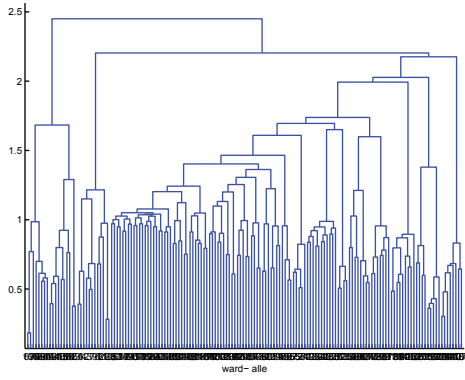
(14) Zeitfenster $\tau = 14$



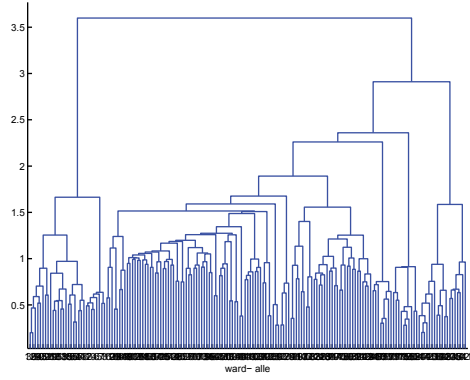
(15) Zeitfenster $\tau = 15$



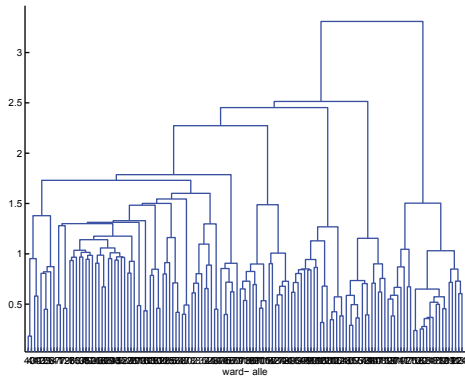
(16) Zeitfenster $\tau = 16$



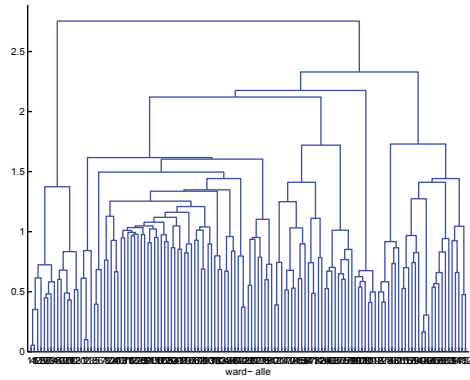
(17) Zeitfenster $\tau = 17$



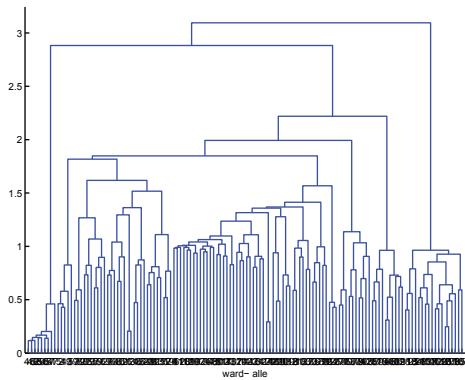
(18) Zeitfenster $\tau = 18$



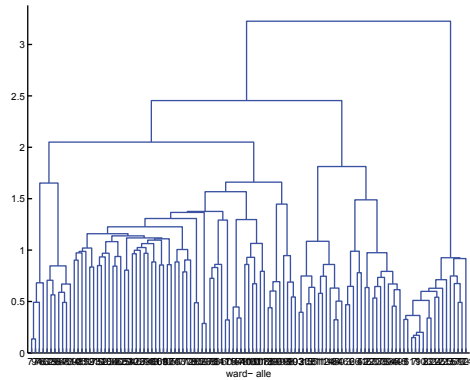
(19) Zeitfenster $\tau = 19$



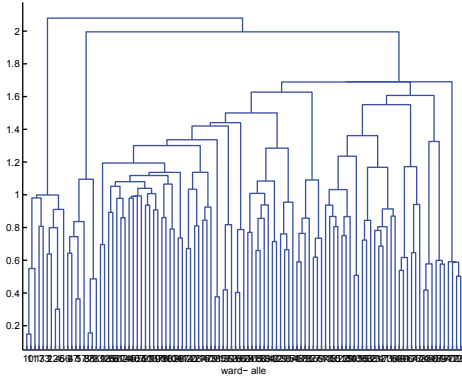
(20) Zeitfenster $\tau = 20$



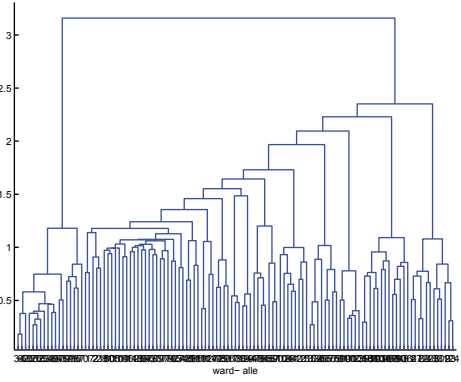
(21) Zeitfenster $\tau = 21$



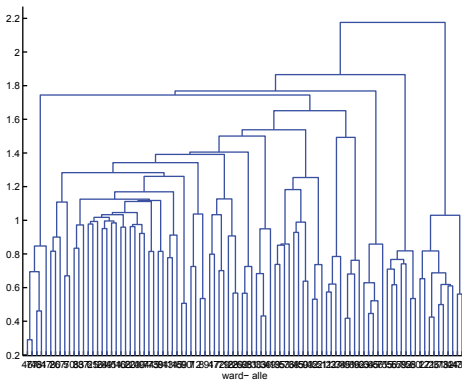
(22) Zeitfenster $\tau = 22$



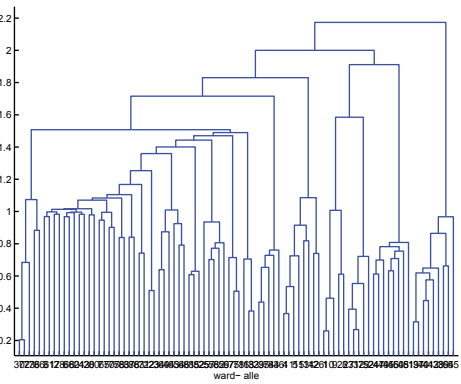
(23) Zeitfenster $\tau = 23$



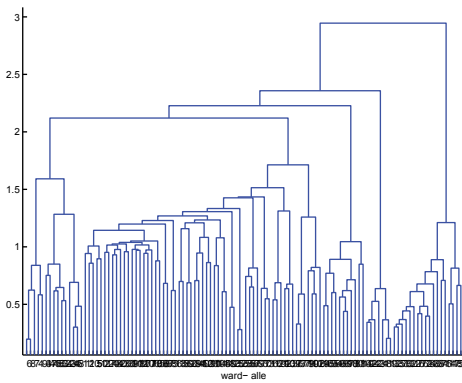
(24) Zeitfenster $\tau = 24$



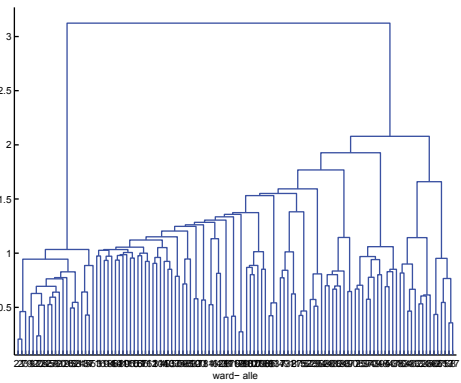
(25) Zeitfenster $\tau = 25$



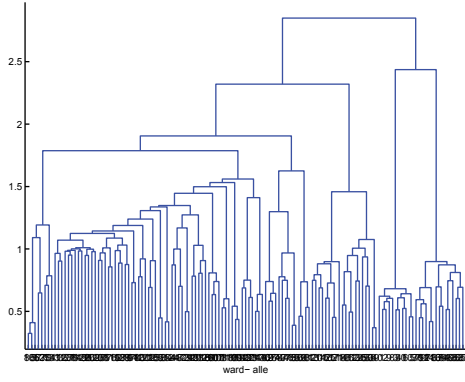
(26) Zeitfenster $\tau = 26$



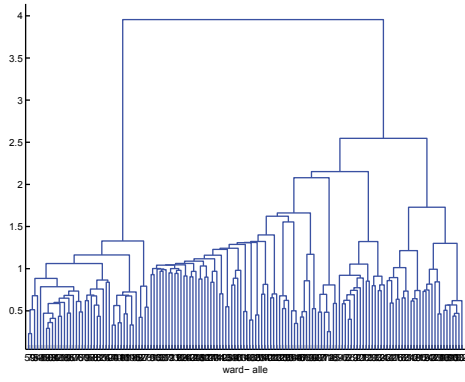
(27) Zeitfenster $\tau = 27$



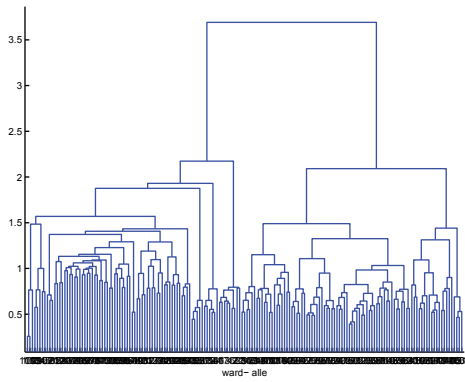
(28) Zeitfenster $\tau = 28$



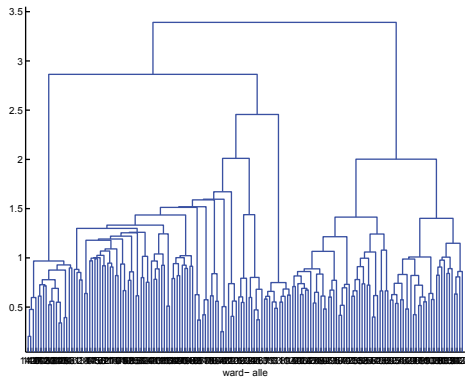
(29) Zeitfenster $\tau = 29$



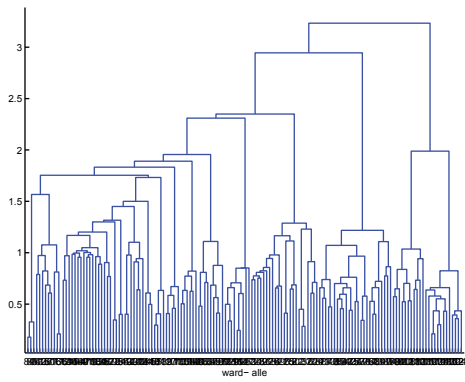
(30) Zeitfenster $\tau = 30$



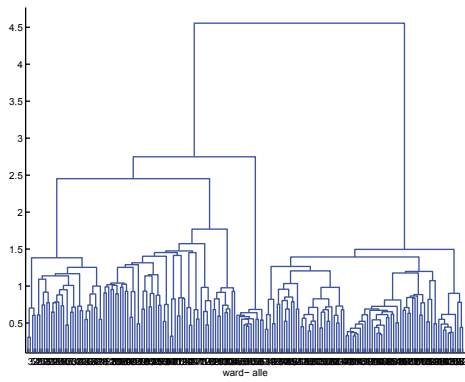
(31) Zeitfenster $\tau = 31$



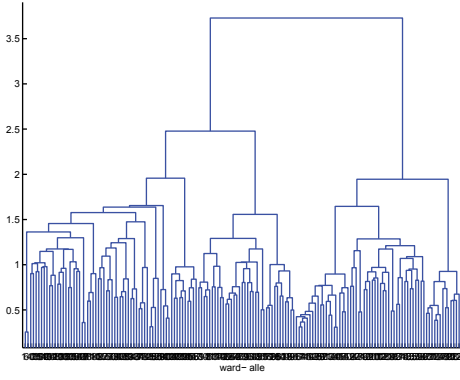
(32) Zeitfenster $\tau = 32$



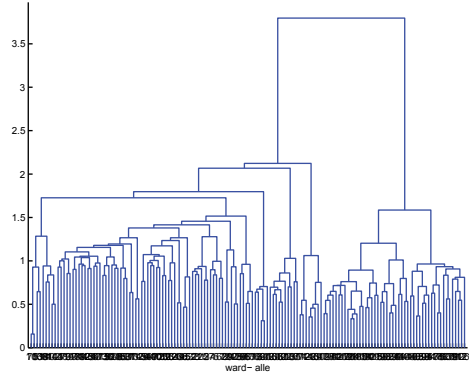
(33) Zeitfenster $\tau = 33$



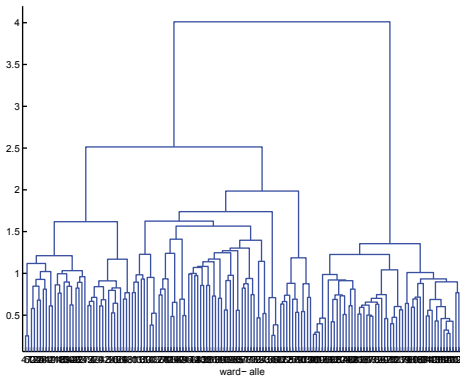
(34) Zeitfenster $\tau = 34$



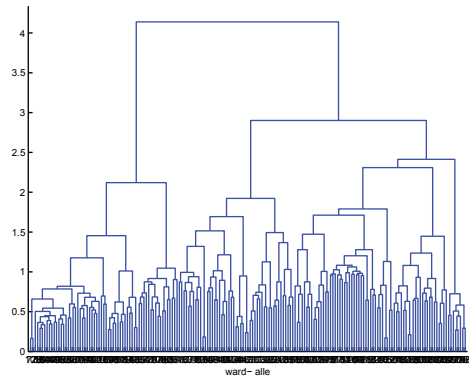
(35) Zeitfenster $\tau = 35$



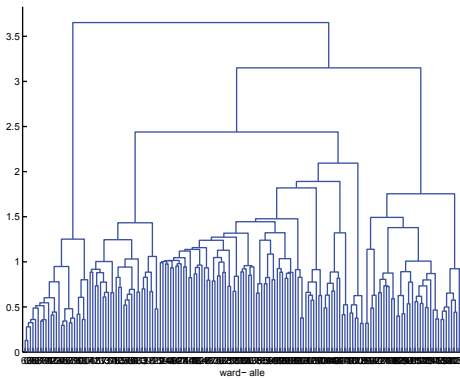
(36) Zeitfenster $\tau = 36$



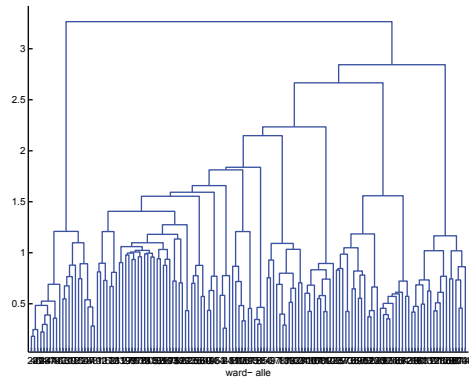
(37) Zeitfenster $\tau = 37$



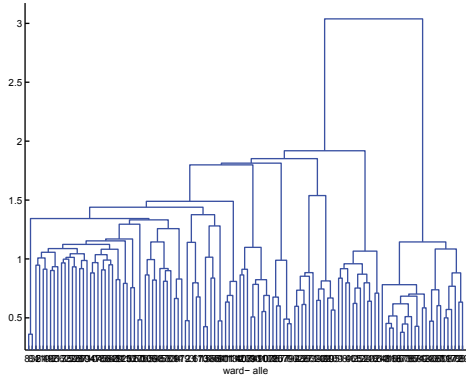
(38) Zeitfenster $\tau = 38$



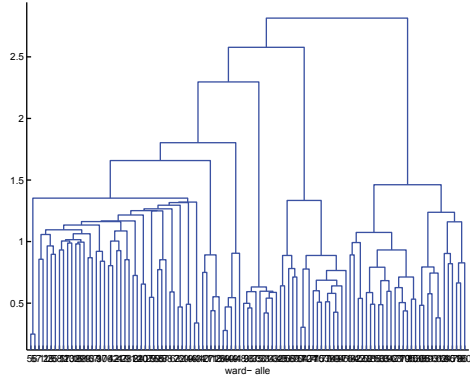
(39) Zeitfenster $\tau = 39$



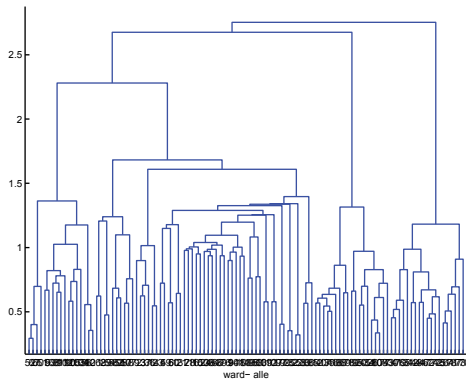
(40) Zeitfenster $\tau = 40$



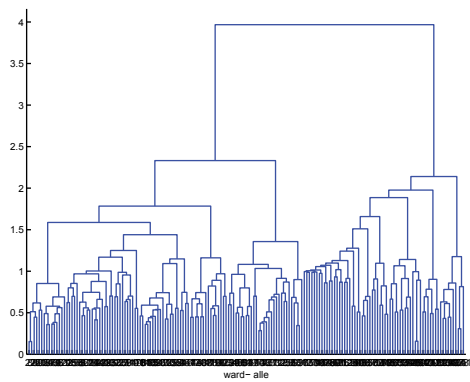
(41) Zeitfenster $\tau = 41$



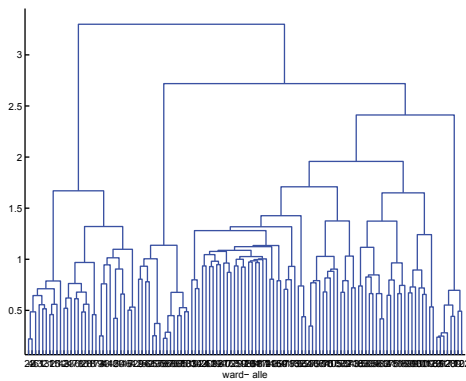
(42) Zeitfenster $\tau = 42$



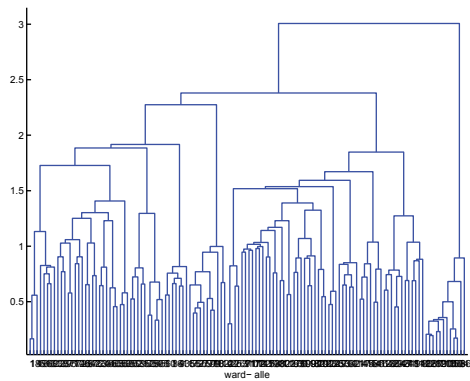
(43) Zeitfenster $\tau = 43$



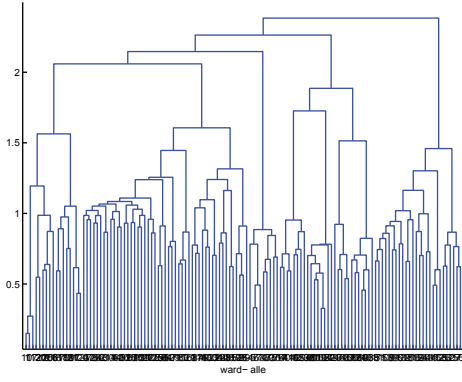
(44) Zeitfenster $\tau = 44$



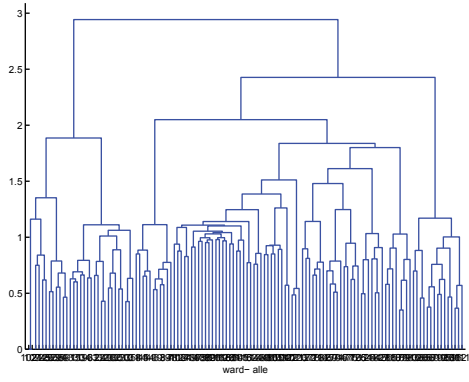
(45) Zeitfenster $\tau = 45$



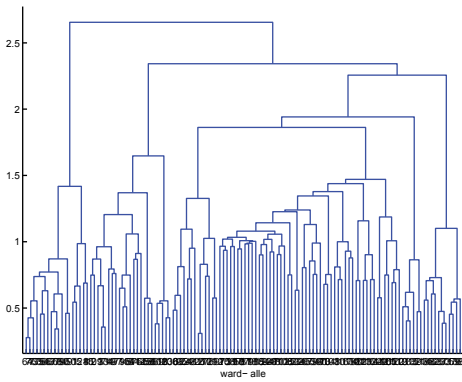
(46) Zeitfenster $\tau = 46$



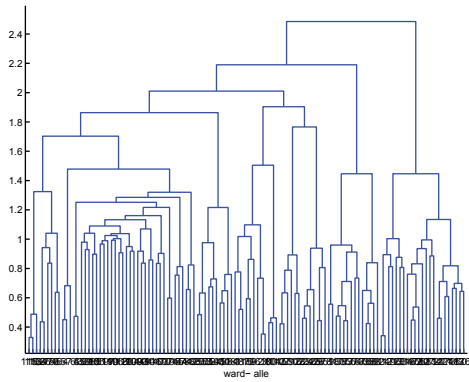
(47) Zeitfenster $\tau = 47$



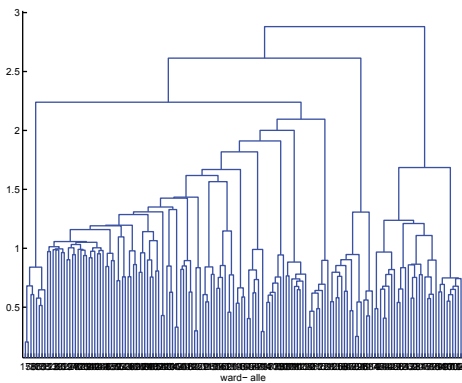
(48) Zeitfenster $\tau = 48$



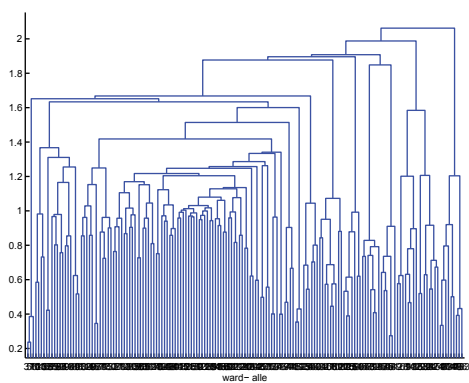
(49) Zeitfenster $\tau = 49$



(50) Zeitfenster $\tau = 50$



(51) Zeitfenster $\tau = 51$



(52) Zeitfenster $\tau = 52$

Abbildung A.1: Dendrogramme der Zeitfenster 1 bis 52.

Abbildungsverzeichnis

1.1	Entwicklung des Internet (Quelle: WhoIsHostingThis.com), abgerufen Sept. 2014	2
1.2	Gliederung der Arbeit	3
3.1	Dendrogramm einer Indizierten Hierarchie (\mathcal{H}, h) (in Anlehnung an BOCK (1980))	28
3.2	Beispiel Ellbogen	30
4.1	Schranken dis_{lb} und dis_{ub}	34
4.2	Multi-Themen-Graph der Relation $Rel(t_1, t_2)$	36
4.3	Mono-Themen-Graph	39
4.4	Mono-Themen-Graph	39
4.5	Mono-Themen-Graph	40
4.6	Mono-Themen-Graph	40
4.7	Evolution eines temporalen Multi-Themen-Graphen im Zeitverlauf	41
4.8	Multi-Themen-Graph einer Konkatenation der Relation $Rel(t_1, t_2)$ mit der sich anschließenden Relation $Rel(t_2, t_3)$	42
4.9	Multi-Themen-Graph der Relation $Rel(t_1, t_3)$ mit auf $m_{t_3} = t_3 - t_1 + 1$ vergrößerter Betrachtungszeitspanne	42
4.10	Mono-Themen-Graph der vergrößerten Relation $Rel(t_1, t_3)$	43
4.11	Zerfall eines Mono-Themen-Graph bei Konkatenation	43
4.12	Mono-Themen-Graph der vergrößerten Relation $Rel(t_1, t_3)$	44
4.13	Zerfall eines Mono-Themen-Graph bei Konkatenation	44
5.1	Quellen für DeReKo (Quelle: IDS Institut für Deutsche Sprache)	48
5.2	Beispiel Testdokumente aus dem Bereich Politik	49
5.3	Deutsches Wörterbuch (mit Genereller Termfrequenz und Inverser Dokumentenfrequenz) & Stoppwortliste	50
5.4	Spiegel Online (Quelle: spiegel.de)	51
5.5	Ausschnitt SQL Daten von Spiegel Online	52
5.6	Spiegel Online - Zeitliche Entwicklung	53
5.7	Worterkennungsrates nach Formel (3.8)	55
5.8	Laufzeit	56
5.9	Cluster-Zuordnungsrate (euklidische Distanz vs. Cosinusmaß)	57
5.10	Dendrogramme für den Übergang von Zeitfenster 1 \rightarrow 2.	59
5.11	Dendrogramme für den Übergang von Zeitfenster 2 \rightarrow 3.	60
5.12	Dendrogramme für den Übergang von Zeitfenster 3 \rightarrow 4.	61
5.13	Dendrogramme für den Übergang von Zeitfenster 4 \rightarrow 5.	62
5.14	Dendrogramme für die Zeitfenster 1 bis 9.	68

5.15	Centroide und Tag Clouds für die Cluster C_1^8 bis C_5^8 in Zeitfenster 8. . .	71
5.16	Mono-Themen-Graph $G_{(8,1)}$ der $Rel(1, 8)_{(8,1)}$ (nach Formel (3.26)) . . .	74
5.17	Tag-Clouds für Cluster in den Zeitfenstern 6 & 7 im Mono-Themen-Graphen 'Libyen'.	74
5.18	Mono-Themen-Graph $G_{(9,6)}$ der $Rel(1, 9)_{(9,6)}$ (nach Formel (3.26)) . . .	75
5.19	Tag Clouds für Cluster in dem Zeitfenster 9 im Mono-Themen-Graphen 'Libyen'.	76
5.20	Mono-Themen-Graph $G_{(9,6)}$ der $Rel(1, 9)_{(9,6)}$ (nach Formel (3.28)) . . .	76
5.21	Tag Clouds für Cluster bei Themen-Frequenz $f((8, 3)) = 12 \times 10^{-2}$ & $f((8, 4)) = 14, 5 \times 10^{-2}$ in dem Zeitfenster $\tau = 8$ im Mono-Themen-Graphen 'Libyen' zum Zeitpunkt $\tau = 9$ (nach Formel (3.28)).	77
5.22	Mono-Themen-Graphen für den Themenkomplex 'Arabischer Frühling' (nach Formel (3.26) bzw. (3.28))	78
5.23	Chronologie des Arabischen Frühlings (Quelle: Bundeszentrale für politische Bildung, 2011, www.bpb.de)	80
5.24	Ausgewählte Tag-Clouds für die Mono-Themen-Graphen in den Abbildungen 5.25 und 5.26.	84
5.25	Mono-Themen-Graphen für den Themenkomplex 'Taliban'	85
5.26	Evolution der Mono-Themen-Graphen für den Themenkomplex 'Taliban'	88
5.27	Tag-Clouds für ausgewählte Mono-Themen-Graphen in den Abbildungen 5.28.	89
5.28	Evolution der Mono-Themen-Graphen für den Themenkomplex 'Bin Laden'	92
5.29	Mono-Themen-Graphen der Teil-Relationen $Rel(42, 47)_{(47,10)}$ und $Rel(47, 52)_{(52,13)}$	93
5.30	Mono-Themen-Graph $G_{(52,13)}$ der $Rel(42, 52)_{(52,13)}$ (nach Formel (3.28))	94
5.31	Ausgewählte Tag-Clouds für die Mono-Themen-Graphen.	95
5.32	Evolution der Mono-Themen-Graphen für den Themenkomplex 'Atomausstieg' bei $m_{\{46,47,48,49,50,51,52\}} = 11$	99
5.33	Verbreiterung der Betrachtungsspanne auf $m_{52} = 27$ des Mono-Themen-Graphen 'Atomausstieg' auf die Relation $Rel(26, 52)_{(52,4)}$	100
5.34	Ausgewählte Tag-Clouds für die Mono-Themen-Graphen in den Abbildungen 5.33.	101
5.35	Verbreiterung der Betrachtungsspanne auf $m_{49} = 49$ des Mono-Themen-Graphen 'Atomausstieg' auf die Relation $Rel(1, 49)_{(49,5)}$	101
5.36	Ausgewählte Tag-Clouds für die Mono-Themen-Graphen in den Abbildungen 5.35.	103
5.37	Die Teil-Themenkomplexe des Mono-Themen-Graphen 'Atomausstieg' der $Rel(1, 49)_{(49,5)}$	104
5.38	Mono-Themen-Graph $G_{(50,5)}$ der $Rel(1, 50)_{(50,5)}$	105
5.39	Ausgewählte Tag-Clouds für die Mono-Themen-Graphen in Abbildung 5.40.	106
5.40	Mono-Themen-Graphen für den Themenkomplex 'FDP'	107
5.41	Dokumente des Cluster C_{13}^{40}	108
A.1	Dendrogramme der Zeitfenster 1 bis 52.	121

Tabellenverzeichnis

4.1	Relationsmatrix mit $dis^{\tau',\tau}(C_{k_{\tau'}}^{\tau'}, C_{k_{\tau}}^{\tau})$ für $\mathcal{K}^{\tau'}$ und \mathcal{K}^{τ}	34
5.1	Testkonfiguration	58
5.2	Relationsmatrix für \mathcal{K}^1 und \mathcal{K}^2	60
5.3	Relationsmatrix für \mathcal{K}^2 und \mathcal{K}^3	61
5.4	Relationsmatrix für \mathcal{K}^3 und \mathcal{K}^4	61
5.5	Relationsmatrix für \mathcal{K}^4 und \mathcal{K}^5	62
5.6	Ergebnisüberblick.	63
5.7	Testkonfiguration & Ergebnisse für 1. Quartal 2011.	65
5.8	Themen-Frequenzen $f((\tau, k_{\tau}))$ für Cluster C_1^8 bis C_5^8	73
5.9	Relationsmatrix mit $dis^{7,8}(C_{k_7}^7, C_{k_8}^8)$ für \mathcal{K}^7 und \mathcal{K}^8	73
5.11	Testkonfiguration & Ergebnisse für 3. Quartal 2010 bis 2. Quartal 2011 (einschließlich).	83

Literaturverzeichnis

- ALLAN, J. (2002a): Introduction to Topic Detection and Tracking. In: Allan, J. (ed.), *Topic Detection and Tracking*, pages 1–16. Kluwer Academic Publishers, Norwell, MA, USA.
- ALLAN, J. (2002b): Detection As Multi-Topic Tracking. *Information Retrieval*, 5(2-3): 139–157.
- ALLAN, J./CARBONELL, J./DODDINGTON, G./YAMRON, J./YANG, Y. (1998): Topic Detection and Tracking Pilot Study: Final Report. In: *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, Lansdowne, VA, USA.
- ALLAN, J./LAVRENKO, V./FREY, D./KHANDELWAL, V. (2000): UMass at TDT 2000. In: *Topic Detection and Tracking Workshop Notebook*, pages 109–115.
- ALLAN, J./LAVRENKO, V./SWAN, R. (2002): Exploration Within Topic Tracking and Detection. In: Allan, J. (ed.), *Topic Detection and Tracking*, pages 197–224. Kluwer Academic Publishers, Norwell, MA, USA.
- ARABIE, P./BOORMAN, S. A. (1973): Multidimensional Scaling of Measures of Distance Between Partitions. *Journal of Mathematical Psychology*, 10(2):148–203.
- ARABIE, P./HUBER, L./DE SOETE, G. (1996): *Clustering and Classification*. World Scientific Publishing, New Jersey, USA.
- BAEZA-YATES, R./CASTILLO, C. (2006): Web Searching. In: Brown, K. (ed.), *Encyclopedia of Language & Linguistics (Second Edition)*, pages 527–538. Elsevier, Oxford.
- BAEZA-YATES, R./RIBEIRO-NETO, B. (2011): *Modern Information Retrieval: The Concepts and Technology Behind Search*. Addison Wesley.
- BARON, S./SPILIOPOULOU, M. (2001): Monitoring Change in Mining Results. In: Kambayashi, Y./Winiwarter, W./Arikawa, M. (eds.), *Data Warehousing and Knowledge Discovery*, volume 2114 of *Lecture Notes in Computer Science*, pages 51–60. Springer Berlin Heidelberg.
- BENHARDUS, J. (2010): Streaming Trend Detection in Twitter. In: *UCCS REU for Artificial Intelligence, Natural Language Processing and Information Retrieval, Final Report*.

- BERRY, M. W./BROWNE, M. (2005): *Understanding Search Engines: Mathematical Modeling and Text Retrieval (Software, Environments, Tools), Second Edition*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.
- BLEI, D. M./NG, A. Y./JORDAN, M. I. (2003): Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.
- BOCK, H.-H. (1974): *Automatische Klassifikation. Theoretische und Praktische Methoden zu Gruppierung und Strukturierung von Daten (Cluster-Analyse)*. Vandenhoeck & Ruprecht, Göttingen.
- BOCK, H.-H. (1980): Clusteranalyse - Überblick und neuere Entwicklungen. *Operations-Research-Spektrum*, 1(4):211–232.
- BUN, K. K./ISHIZUKA, M. (2006): Emerging Topic Tracking System in WWW. *Knowledge-Based Systems*, 19(3):164 – 171.
- CAI, K./SPANGLER, S./CHEN, Y./ZHANG, L. (2008): Leveraging Sentiment Analysis for Topic Detection. In: *Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology. WI-IAT '08. IEEE/WIC/ACM*, volume 1, pages 265–271.
- CARPENTER, G. A./GROSSBERG, S./ROSEN, D. B. (1991): Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System. *Neural Networks*, 4(6):759–771.
- CHEN, C. C./CHEN, M. C./CHEN, M.-S. (2009): An Adaptive Threshold Framework for Event Detection Using HMM-Based Life Profiles. *ACM Transactions on Information Systems*, 27(2):9:1–9:35.
- CHEN, K.-Y./LUESUKPRASERT, L./CHOU, S.-C. T. (2007): Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling. *IEEE Transactions on Knowledge and Data Engineering*, 19(8):1016–1025.
- COLLIER, N./HIRAKAWA, H./KUMANO, A. (1998): Machine Translation vs. Dictionary Term Translation: A Comparison for English-Japanese News Article Alignment. In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 263–267, Stroudsburg, PA, USA. Association for Computational Linguistics.
- CORMACK, R. M. (1971): A Review of Classification. *Journal of the Royal Statistical Society. Series A (General)*, 134(3):321–367.
- CROFT, W. B. (1977): Clustering Large Files of Documents Using the Single-Link Method. *Journal of the American Society for Information Science*, 28(6):341–344.
- CROFT, W. B./LAFFERTY, J. (2003): *Language Modeling for Information Retrieval*. Kluwer Academic Publishers, Norwell, MA, USA.
- CUTTING, D. R./KARGER, D. R./PEDERSEN, J. O./TUKEY, J. W. (1992): Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections.

- In: *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '92, pages 318–329, New York, NY, USA. ACM.
- DAGAN, I./CHURCH, K. (1997): Termight: Coordinating Humans and Machines in Bilingual Terminology Acquisition. *Machine Translation*, 12(1-2):89–107.
- DHILLON, I./KOGAN, J./NICHOLAS, C. (2004): Feature Selection and Document Clustering. In: Berry, M. W. (ed.), *Survey of Text Mining: Clustering, Classification, and Retrieval*, pages 73–100. Springer, New York.
- EL-HAMDOUCHI, A./WILLETT, P. (1986): Hierarchic Document Classification Using Ward's Clustering Method. In: *Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '86, pages 149–156, New York, NY, USA. ACM.
- FENG, A./ALLAN, J. (2005): Hierarchical Topic Detection in TDT-2004. Technical report.
- FLOREK, K./LUKASZEWICZ, J./PERKAL, J./STEINHAUS, H./ZUBRZYCKI, S. (1951): Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicae*, 2:282–285.
- FORGY, E. (1965): Cluster Analysis of Multivariate Data: Efficiency Versus Interpretability of Classification. *Biometrics*, 21(3):768–769.
- FOWLKES, E. B./MALLOWS, C. L. (1983): A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383):553–569.
- FOX, C. (1992): Lexical Analysis and Stoplists. In: Frakes, W. B./Baeza-Yates, R. (eds.), *Information Retrieval: Data Structures & Algorithms*, pages 102–130. Prentice Hall, Inc. Upper Saddle River, NJ, USA.
- FRAKES, W. B. (1992): Stemming Algorithms. In: Frakes, W. B./Baeza-Yates, R. (eds.), *Information Retrieval: Data Structures & Algorithms*, pages 131–160. Prentice Hall, Inc. Upper Saddle River, NJ, USA.
- FUKUMOTO, F./SUZUKI, Y. (2007): Topic Tracking Based on Bilingual Comparable Corpora and Semisupervised Clustering. *ACM Transactions on Asian Language Information Processing*, 6(3):11:1–22.
- GANG, D./JUN, G./WEI-RAN, X. (2011): Burst Feature Detection Using Parameter Estimated Two-State Automaton. *The Journal of China Universities of Posts and Telecommunications*, 18, Supplement 1(0):90–96.
- GAUL, W. (2011): Web Page Importance Ranking. *Advances in Data Analysis and Classification*, 5:113–128.
- GAUL, W./VINCENT, D. (2013): An Approach for Topic Trend Detection. In: Lausen, B./Van den Poel, D./Ultsch, A. (eds.), *Algorithms from and for Nature and Life*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 347–354. Springer International Publishing.

- GÓMEZ, C./PESQUET-POPESCU, B. (2007): A Simple and Efficient Eigenfaces Method. In: *Proceedings of the 9th International Conference on Advanced Concepts for Intelligent Vision Systems*, ACIVS'07, pages 364–372, Berlin, Heidelberg. Springer-Verlag.
- GONG, L./ZENG, J./ZHANG, S. (2011): Text Stream Clustering Algorithm Based on Adaptive Feature Selection. *Expert Systems with Applications*, 38(3):1393–1399.
- GORDON, A. D. (1987): A Review of Hierarchical Classification. *Journal of the Royal Statistical Society. Series A (General)*, 150(2):119–137.
- HE, Q./CHANG, K./LIM, E.-P. (2007): Analyzing Feature Trajectories for Event Detection. In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 207–214, New York, NY, USA. ACM.
- HEYER, G./QUASTHOFF, U./WITTIG, T. (2008): *Text Mining: Wissensrohstoff Text. IT lernen*. W3L-Verl., Herdecke ; Bochum, 1. korr. nachdr. edition.
- HOFMANN, T. (1999): Probabilistic Latent Semantic Indexing. In: *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA. ACM.
- HOWLAND, P./PARK, H. (2004): Cluster-Preserving Dimension Reduction Methods for Efficient Classification of Text Data. In: Berry, M. W. (ed.), *Survey of Text Mining: Clustering, Classification, and Retrieval*, pages 3–23. Springer New York.
- HUANG, A. H. (2003): Effects of Multimedia on Document Browsing and Navigation: An Exploratory Empirical Investigation. *Information and Management*, 41(2):189–198.
- HUBERT, L./ARABIE, P. (1985): Comparing Partitions. *Journal of Classification*, 2: 193–218.
- JACCARD, P. (1901): Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37: 547–579.
- JAIN, A. K./MURTY, M. N./FLYNN, P. J. (1999): Data Clustering: A Review. *ACM Computing Surveys*, 31(3):264–323.
- JIN, Y./MYAENG, S. H./JUNG, Y. (2007): Use of Place Information for Improved Event Tracking. *Information Processing and Management*, 43(2):365–378.
- JOHNSON, S. (1967): Hierarchical Clustering Schemes. *Psychometrika*, 32:241–254.
- KHY, S./ISHIKAWA, Y./KITAGAWA, H. (2008): A Novelty-Based Clustering Method for On-line Documents. *World Wide Web*, 11(1):1–37.
- KIM, P./MYAENG, S. H. (2004): Usefulness of Temporal Information Automatically Extracted from News Articles for Topic Tracking. *ACM Transactions on Asian Language Information Processing*, 3(4):227–242.

- KLEINBERG, J. (2003): Bursty and Hierarchical Structure in Streams. *Data Mining and Knowledge Discovery*, 7(4):373–397.
- KOBAYASHI, M./AONO, M. (2008): Vector Space Models for Search and Cluster Mining. In: Berry, M. W./Castellanos, M. (eds.), *Survey of Text Mining II*, pages 109–127. Springer London.
- KOGAN, J./NICHOLAS, C./VOLKOVICH, V. (2003): Text Mining with Information-Theoretic Clustering. *Computing in Science and Engineering*, 5(6):52–59.
- KOHONEN, T. (2001): *Self-Organizing Maps*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 3rd edition.
- KONTOSTATHIS, A./GALITSKY, L. M./POTTENGER, W. M./ROY, S./PHELPS, D. J. (2004): A Survey of Emerging Trend Detection in Textual Data Mining. In: Berry, M. W. (ed.), *Survey of Text Mining: Clustering, Classification, and Retrieval*, pages 185–224. Springer, New York.
- KUPIETZ, M. (2005): Near-Duplicate Detection in the IDS Corpora of Written German. Technical report, Institut für Deutsche Sprache.
- KUPIETZ, M./KEIBEL, H. (2009): The Mannheim German Reference Corpus (DeReKo) as a Basis for Empirical Linguistic Research. In: Minegishi, M./Kawaguchi, Y. (eds.), *Working Papers in Corpus-Based Linguistics and Language Education*, number 3, pages 53–59. Tokyo University of Foreign Studies (TUFS).
- KUPIETZ, M./BELICA, C./KEIBEL, H./WITT, A. (2010): The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In: Calzolari, N./Choukri, K./Maegaard, B./Mariani, J./Odiijk, J./Piperidis, S./Rosner, M./Tapias, D. (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).
- LANCE, G. N./WILLIAMS, W. T. (1966): A Generalized Sorting Strategy for Computer Classifications. *Nature*, 212:218.
- LANCE, G. N./WILLIAMS, W. T. (1967): A General Theory of Classificatory Sorting Strategies 1. Hierarchical Systems. *The Computer Journal*, 9(4):373–380.
- LANDMANN, J./ZUELL, C. (2008): Identifying Events Using Computer-Assisted Text Analysis. *Social Science Computer Review*, 26(4):483–497.
- LARSEN, B./AONE, C. (1999): Fast and Effective Text Mining Using Linear-Time Document Clustering. In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '99*, pages 16–22, New York, NY, USA. ACM.
- LEE, D. L./CHUANG, H./SEAMONS, K. (1997): Document Ranking and the Vector-Space Model. *IEEE Software*, 14(2):67–75.
- LI, B./LI, W./LU, Q. (2006): Topic Tracking with Time Granularity Reasoning. *ACM Transactions on Asian Language Information Processing*, 5(4):388–412.


- LI, X./CROFT, W. B. (2008): An Information-Pattern-Based Approach to Novelty Detection. *Information Processing and Management*, 44(3):1159–1188.
- LI, Y./CHUNG, S. M./HOLT, J. D. (2008): Text Document Clustering Based on Frequent Word Meaning Sequences. *Data & Knowledge Engineering*, 64(1):381–404.
- MACKEY, D. J. C. (2002): *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA.
- MACQUEEN, J. B. (1967): Some Methods for Classification and Analysis of Multivariate Observations. In: Cam, L. M. L./Neyman, J. (eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.
- MAHALANOBIS, P. C. (1936): On the Generalised Distance in Statistics. In: *Proceedings National Institute of Science, India*, volume 2, pages 49–55.
- MANNING, C. D./RAGHAVAN, P./SCHÜTZE, H. (2008): *Introduction to Information Retrieval*. Cambridge University Press.
- MARKOV, A. A. (2006): An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains. *Science in Context*, 19: 591–600.
- MARKOV, Z./LAROSE, D. T. (2007): *Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*. Wiley-Interscience.
- MATHIOUDAKIS, M./KOUZAS, N. (2010): TwitterMonitor: Trend Detection over the Twitter Stream. In: *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 1155–1158, New York, NY, USA. ACM.
- MCQUITTY, L. L. (1957): Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies. *Educational and Psychological Measurement*, 17:207–229.
- MEI, Q./ZHAI, C. (2005): Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining. In: *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 198–207, New York, NY, USA. ACM.
- MEI, Q./LIU, C./SU, H./ZHAI, C. (2006): A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs. In: *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 533–542, New York, NY, USA. ACM.
- MIRKIN, B. G./CHERNYI, L. B. (1970): Measurement of the Distance Between Distinct Partitions of a Finite Set of Objects. *Automation and Remote Control*, 31: 786–792.

- MOONEY, R. J./BUNESCU, R. (2005): Mining Knowledge from Text Using Information Extraction. *ACM SIGKDD Explorations Newsletter - Natural Language Processing and Text Mining*, 7(1):3–10.
- OARD, D. W. (1999): Topic Tracking with the PRISE Information Retrieval System. In: *Proceedings of the DARPA Broadcast News Workshop*, pages 209–211.
- OLIVEIRA, M./GAMA, J. (2010): Bipartite Graphs for Monitoring Clusters Transitions. In: Cohen, P./Adams, N./Berthold, M. (eds.), *Advances in Intelligent Data Analysis IX*, volume 6065 of *Lecture Notes in Computer Science*, pages 114–124. Springer Berlin Heidelberg.
- PAPKA, R./ALLAN, J. (1998): Document Classification Using Multiword Features. In: *Proceedings of the Seventh International Conference on Information and Knowledge Management, CIKM '98*, pages 124–131, New York, NY, USA. ACM.
- PARK, J./CHOI, B.-C./KIM, K. (2010): A Vector Space Approach to Tag Cloud Similarity Ranking. *Information Processing Letters*, 110(12-13):489–496.
- PONS-PORRATA, A./BERLANGA-LLAVORI, R./RUIZ-SHULCLOPER, J. (2002): On-line Event and Topic Detection by Using the Compact Sets Clustering Algorithm. *Journal of Intelligent & Fuzzy Systems*, 12(3,4):185–194.
- PONS-PORRATA, A./BERLANGA-LLAVORI, R./RUIZ-SHULCLOPER, J. (2007): Topic Discovery Based on Text Mining Techniques. *Information Processing and Management*, 43(3):752–768.
- PORTER, M. F. (1997): An Algorithm for Suffix Stripping. In: Sparck Jones, K./Willett, P. (eds.), *Readings in Information Retrieval*, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- RAJAN, K./RAMALINGAM, V./GANESAN, M./PALANIVEL, S./PALANIAPPAN, B. (2009): Automatic Classification of Tamil Documents Using Vector Space Model and Artificial Neural Network. *Expert Systems with Applications*, 36(8):10914–10918.
- RAJARAMAN, K./TAN, A.-H. (2001): Topic Detection, Tracking, and Trend Analysis Using Self-Organizing Neural Networks. In: *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD '01*, pages 102–107, London, UK, UK. Springer-Verlag.
- RAND, W. M. (1971): Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850.
- ROUSSEEUW, P. (1987): Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65.
- RUBIN, J. (1966): An Approach to Organizing Data into Homogeneous Groups. *Systematic Biology*, 15(3):169–182.
- SALTON, G. (1971): *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

- SALTON, G. (1989): *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- SALTON, G./BUCKLEY, C. (1988): Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5):513–523.
- SALTON, G./WONG, A./YANG, C. S. (1975): A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620.
- SCHULT, R./SPILIOPOULOU, M. (2006): Discovering Emerging Topics in Unlabelled Text Collections. In: Manolopoulos, Y./Pokorný, J./Sellis, T. (eds.), *Advances in Databases and Information Systems*, volume 4152 of *Lecture Notes in Computer Science*, pages 353–366. Springer Berlin Heidelberg.
- SEGEV, A./KANTOLA, J. (2012): Identification of Trends from Patents Using Self-Organizing Maps. *Expert Systems with Applications*, 39(18):13235–13242.
- SINT, P. (1975): *Ähnlichkeitsstrukturen und Ähnlichkeitsmaße*. Schriftenreihe des Instituts für sozio-ökonomische Entwicklungsforschung der Österreichischen Akademie der Wissenschaften, Wien.
- SNEATH, P. H. A. (1957): The Application of Computers to Taxonomy. *Journal of General Microbiology*, 17(1):201–226.
- SNEATH, P. H. A./SOKAL, R. R. (1973): *Numerical Taxonomy: The Principles & Practice of Numerical Classification*. W. H. Freeman, San Francisco.
- SOKAL, R. R./MICHENER, C. D. (1958): A Statistical Method for Evaluating Systematic Relationships. *University of Kansas Science Bulletin*, 38:1409–1438.
- SOUICY, P./MINEAU, G. W. (2005): Beyond TFIDF Weighting for Text Categorization in the Vector Space Model. In: *Proceedings of the 19th International Joint Conference on Artificial Intelligence, IJCAI'05*, pages 1130–1135, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- SPÄTH, H. (1977): *Cluster-Analyse - Algorithmen zur Objektklassifizierung und Datenreduktion*. Oldenbourg Wissenschaftsverlag.
- SPILIOPOULOU, M./NTOUTSI, I./THEODORIDIS, Y./SCHULT, R. (2006): MONIC: Modeling and Monitoring Cluster Transitions. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06*, pages 706–711, New York, NY, USA. ACM.
- STEINBACH, M./KARYPIS, G./KUMAR, V. (2000): A Comparison of Document Clustering Techniques. In: *KDD Workshop on Text Mining*.
- STREHL, A./GHOSH, J./MOONEY, R. (2000): Impact of Similarity Measures on Web-page Clustering. In: *Workshop on Artificial Intelligence for Web Search*, pages 58–64. AAAI.
- TAI, X./REN, F./KITA, K. (2002): An Information Retrieval Model Based on Vector Space Method by Supervised Learning. *Information Processing & Management*, 38(6):749–764.

- TERACHI, M./SAGA, R./TSUJI, H. (2006): Trends Recognition in Journal Papers by Text Mining. In: *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pages 4784–4789, Taipei, Taiwan.
- TU, Y.-N./SENG, J.-L. (2012): Indices of Novelty for Emerging Topic Detection. *Information Processing & Management*, 48(2):303–325.
- UTSURO, T./HORIUCHI, T./HINO, K./HAMAMOTO, T./NAKAYAMA, T. (2003): Effect of Cross-Language IR in Bilingual Lexicon Acquisition from Comparable Corpora. In: *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1, EACL '03*, pages 355–362, Stroudsburg, PA, USA. Association for Computational Linguistics.
- WAGNER, R./ONTRUP, J./SCHOLZ, S. (2009): Event Detection in Environmental Scanning. In: Gaul, W./Bock, H.-H./Imaizumi, T./Okada, A. (eds.), *Cooperation in Classification and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 161–168. Springer Berlin Heidelberg.
- WALLS, F./JIN, H./SISTA, S./SCHWARTZ, R. (1999): Topic Detection in Broadcast News. In: *Proceedings of the DARPA Broadcast News Workshop*, pages 193–198. Morgan Kaufmann Publishers, Inc.
- WANG, X./ZHAI, C./HU, X./SPROAT, R. (2007): Mining Correlated Bursty Topic Patterns from Coordinated Text Streams. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '07*, pages 784–793, New York, NY, USA. ACM.
- WANG, X./ZHANG, K./JIN, X./SHEN, D. (2009): Mining Common Topics from Multiple Asynchronous Text Streams. In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining, WSDM '09*, pages 192–201, New York, NY, USA. ACM.
- WARD, J. H. (1963): Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244.
- WAYNE, C. L. (1998): Topic Detection and Tracking (TDT) - Overview and Perspective. In: *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne Conference Resort, Lansdowne Virginia.
- WEI, C.-P./CHANG, Y. (2007): Discovering Event Evolution Patterns From Document Sequences. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 37(2):273–283.
- WEI, C.-P./LEE, Y.-H. (2004): Event Detection from Online News Documents for Supporting Environmental Scanning. *Decision Support Systems*, 36(4):385–401.
- YANG, C./SHI, X./WEI, C.-P. (2009): Discovering Event Evolution Graphs From News Corpora. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 39(4):850–863.
- ZENG, J./ZHANG, S. (2007): Variable Space Hidden Markov Model for Topic Detection and Analysis. *Knowledge-Based Systems*, 20(7):607–613.

- ZHANG, W./YOSHIDA, T./TANG, X. (2007): Text Classification Using Multi-Word Features. In: *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pages 3519–3524, Montréal, Canada.
- ZHAO, Y./KARYPIS, G. (2002): Evaluation of Hierarchical Clustering Algorithms for Document Datasets. In: *Proceedings of the Eleventh International Conference on Information and Knowledge Management, CIKM '02*, pages 515–524, New York, NY, USA. ACM.
- ZHUKOV, L./GLEICH, D. F. (2004): Topic Identification in Soft Clustering Using PCA and ICA. Yahoo! Research Labs.

The background features a network graph with white nodes and lines on a blue and green bokeh background. A large, semi-transparent diamond shape is centered, containing a stylized mountain range with three peaks in shades of brown and grey. The text is overlaid on the lower-left portion of this diamond.

Das Auffinden relevanter Themen und Trends innerhalb von online Dokumentenströmen ist ein wichtiges Anliegen diverser Forschungsgebiete. In dieser Arbeit wird ein Modell entwickelt zur Darstellung von Relationen zwischen Themen-Clustern von Dokumenten in verschiedenen Zeitfenstern als Themen-Graphen. Durch Variieren der Betrachtungszeitspannen können Beziehungen zwischen Themen verschiedener Zeitfenster in unterschiedlicher Komplexität abgebildet werden. Bei Verschiebung des Betrachtungszeitraums wird die Evolution eines Themen-Graphen verfolgbar. Typische Lebenszyklen wie Entstehen, Anwachsen, Abnehmen und Verschwinden eines Themas, wie auch thematische Relationen zwischen Themen und deren Änderungen über die Zeit innerhalb eines Themenkomplexes werden sichtbar. Neben der zeitlichen Komponente eines Themen-Clusters kann über die Themen-Frequenz auch seine gegenwärtige Bedeutung in dem jeweiligen Zeitfenster angegeben werden. Mögliche Einflussparameter auf die Themen-Graphen, wie Zeitfenstergröße, verschiedene Gütekriterien für das beim Clustering verwendete Ellbogenkriterium, Betrachtungsspannenveränderungen wie auch das Fortschreiten von Betrachtungszeiträumen werden untersucht. Es kann gezeigt werden, dass die gefundenen Themenkomplexe in ihrer Struktur (Inhalte und Themen-Frequenzen) und ihrem zeitlichen Verlauf entsprechenden bekannten Ereignissen zugeordnet sind.

ISBN 978-3-7315-0340-8



9 783731 503408 >