Zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften (Dr. rer. pol.)
von der Fakultät für Wirtschaftswissenschaften
der Universität Fridericiana zu Karlsruhe
genehmigte Dissertation.

Clustern mit Hintergrundwissen

von
Dipl.-Wirtsch.-Inform. Andreas Hotho

18. August 2004

Tag der mündlichen Prüfung: 5. Mai 2004
Referent: Prof. Dr. Rudi Studer
Korreferent: Prof. Dr. Wolfgang Gaul

Gesetzt am 18. August 2004 um 15:25 Uhr.

Meiner Familie.

Geleitwort

Mit den rasant wachsenden Dokumentenbeständen, die sich einerseits durch die Entwicklung des World Wide Web, andererseits durch Realisierung von Intranets in Unternehmen ergeben, wächst zunehmend der Bedarf, diese großen Dokumentenbestände geeignet zu strukturieren, um sie den Benutzern besser zugänglich zu machen. Hierzu wurden in der Vergangenheit eine Vielzahl von Clusterverfahren entwickelt. Derartige Clusterverfahren finden des Weiteren auch Einsatz im Customer Relationship Management, um z.B. interessante Segmentierungen von Kunden zu finden. Je nach Anwendungskontext zeigen gängige Clusterverfahren allerdings Schwächen im Hinblick auf die Güte der gefundenen Cluster sowie die Bereitstellung flexibler, benutzerbezogener Sichten und Visualisierungen.

Hier setzt die vorliegende Dissertation an, indem sie Clusterverfahren durch den Einsatz von Ontologien als Hintergrundwissen weiter entwickelt sowie Ontologien und Verfahren der formalen Begriffsanalyse zur Berechnung und Visualisierung benutzerbezogener Cluster verwendet.

Ein wesentlicher Beitrag der Dissertation ist der Ansatz des subjektiven Clustern. Er zielt darauf ab, auf die Bedürfnisse des Anwenders zurechtgeschnittene Cluster zu berechnen, die erzeugten Clusterergebnisse in einer für den Anwender verständlichen Form zu präsentieren und bei hochdimensionalen Datensätzen eine systematische Reduktion der Dimensionalität zu erreichen. Hierzu wird die Methodik COSA (Concept Selection and Aggregation) eingeführt, die zum einen die Abbildung von Objekten der realen Welt auf Konzepte einer Ontologie unterstützt und die zum anderen die sichtenspezifische Auswahl von Konzepten beinhaltet. Es zeigt sich, dass der COSA-Ansatz in vielen Fällen deutliche Verbesserungen der Ergebnisse liefert. Gleichzeitig wird jedoch auch erkennbar, dass bei der Ontologie-basierten Definition von Sichten sorgfältig vorgegangen werden muss, um gute Ergebnisse zu erzielen.

Den methodischen Kern der Dissertation bildet die Integration von Hintergrundwissen zur Steigerung der Güte der Clusterergebnisse für Textdokumente. Dazu wird ein Ansatz eingeführt, die “Bag of Words”-Repräsentation von Dokumenten mit Hintergrundwissen zu erweitern. Zusätzlich werden verschiedene Strategien definiert, die das Problem der Mehrdeutigkeit von Wörtern adressieren. Schließlich werden verschiedene Strategien zur Integration von Konzepten und Oberkonzepten aus der Ontologie spezifiziert. Der durch diese Strategien definierte Variantenraum wird anschließend einer systematischen Evaluierung auf der Basis der Reuters-Dokumente unterzogen. Dabei zeigt es sich, dass der entwickelte Ansatz bei einer sorgfältigen Abstimmung der verschiedenen Strategien aufeinander zu signifikanten Verbesserungen der Clusterergebnisse führt.

Einen weiteren Schwerpunkt der Dissertation bildet die Analyse und Verbesserung von Ansätzen der Formalen Begriffsanalyse zum Clustern. Es wird aufgezeigt, wie die Formale Begriffsanalyse zur Strukturierung und Visualisierung von Clusterresultaten eingesetzt werden kann. Dies führt zu einer für den Anwender besseren Erklärung von Gemeinsamkeiten und Unterschieden der erzeugten Cluster. Darauf aufbauend kann dann auch eine explorative Analyse des Dokumentenbestandes erreicht werden - durch die Bestimmung und Visualisierung interessanter Teilverbände.

Abgerundet wird die Dissertation durch die Anwendung des Ansatzes des subjektiven Clusterns auf Telekommunikationsdaten. Ausgehend von einer Diskussion der Problematik des Clusterns in hochdimensionalen Merkmalsräumen werden Ansätze zur Reduktion der Dimensionalität betrachtet und aufgezeigt, wie eine geeignet strukturierte Telekommunikations-Ontologie zur Dimensionsre-

duktion eingesetzt werden kann. Dabei zeigt es sich, dass Ontologie-basierte Sichten zu einer Verbesserung der Clustergüte führen und die Sichten sehr flexibel auf die Bedürfnisse verschiedener Anwender ausgerichtet werden können.

Die vorliegende Dissertation beinhaltet interessante neue methodische Ansätze zum Clustern. Besonders hervorzuheben ist, dass nicht nur neue Methoden entwickelt, sondern diese auch einer systematischen Evaluierung anhand realer Datensätze unterzogen werden. Diese Forschungsarbeiten sind auch als Beitrag zum gerade entstehenden Gebiet des Semantic Web Mining zu sehen, bei dem u.a. die Fragestellung betrachtet wird, wie durch den Einsatz von Hintergrundwissen die Ergebnisse von Lernverfahren, hier von Clusterverfahren, verbessert werden können.

Prof. Dr. Rudi Studer
Karlsruhe, August 2004

Vorwort

Das Schreiben einer Dissertation erstreckt sich über viele Monate, in denen man immer wieder mit zahlreichen Fragen konfrontiert wird, während man in seinem manchmal doch zu warmen Büro sitzt und über die Lösung dieser sinniert. Häufig hilft dann ein klärendes Gespräch mit netten Kollegen, die man mehr oder minder zufällig an der Kaffeemaschine trifft, die als eine der besten Wissensmanagementlösungen gilt. Als "eingefleischter" Teetrinker tummelt man sich nur selten an der Kaffeemaschine. So bin ich dann auch mit der Anschaffung des Espresso-Automaten am AIFB zu den Kaffeetrinkern gewechselt und habe häufig dort mit dem einen oder anderen Kollegen angeregt aktuelle Probleme diskutiert.

An dieser Stelle möchte ich mich als erstes für die sehr schönen, spannenden aber auch arbeitsreichen letzten fünf Jahre bei meinem Mentor Rudi Studer bedanken, der die Gruppe Wissensmanagement am Institut für Angewandte Informatik und Formale Beschreibungsverfahren der Universität Karlsruhe leitet. Er gab mir die Chance und die Möglichkeit, in einem Projekt bei der Deutschen Telekom AG mit meiner wissenschaftlichen Arbeit zu starten. Die Arbeit vor Ort bei der Telekom aber auch am Institut erlaubte es mir, wertvolle Erfahrungen sowohl in der Projektarbeit als auch in der Forschung und Lehre zu sammeln. Obwohl Rudi als Mentor maßgeblich zum Erfolg der Arbeit beigetragen hat, möchte ich mich an dieser Stelle auch bei Prof. Dr. Wolfgang Gaul, meinem Zweitgutachter, sowie Prof. Dr. Hartmut Schmeck und Prof. Dr. Jan Kowalski für die Teilnahme an der mündlichen Prüfung bedanken.

Dr. Gutsche hat als Projektleiter und Initiator bei der Deutschen Telekom AG nicht nur die Arbeit erst möglich gemacht - er stand auch jederzeit für ein fachliches Gespräch zur Verfügung, wofür ich mich ganz herzlich bedanken möchte. Bedanken möchte ich mich auch bei Dr. Jäger und Heiko Zimmermann, die mir bei der Arbeit vor Ort bei der Deutschen Telekom AG in Bruchsal immer eine große Hilfe waren.

In den ersten Jahren am AIFB war mein Bürokollege Alexander Mädche stets und jederzeit für ein Statement zu meiner Arbeit bzw. zu einer Diskussion zu haben. Er beeindruckte mich immer wieder durch seine stete Anwesenheit. Auch konnte man mit ihn spät abends noch ein Bier am Institut trinken. Auch Steffen Staab und Gerd Stumme hatten immer ein offenes Ohr für meine Fragen und diskutierten mit mir gern erste Ideen. In den letzten Jahren teilte ich mir nicht nur die Unterstützung bei der Administration der Institutsrechner am AIFB, sondern auch den Raum mit Christoph Schmitz. Zusammen mit ihm wie auch vorher mit Daniel Merkle lösten wir so einige Probleme der Institutsrechner. Auch einige meiner Studenten sollen an dieser Stelle nicht unerwähnt bleiben. So haben Philipp Sorg und die anderen "Liwis" in ganz erheblichen Maße dazu beigetragen, die Administration der Rechner zu vereinfachen.

An dieser Stelle möchte ich mich auch bei meiner Familie in Leipzig bedanken, die mich nicht nur während meine Dissertation unterstützt hat, sondern die mir auch in den Jahren davor während meines Studiums jeder Zeit mit Rat und Tat zu Seite stand. Meiner Frau Dagmar, die mich gerade in der letzten Phase der Arbeit sehr unterstützt hat und mir die Ruhe und Kraft gegeben hat, möchte ich an dieser Stelle ganz herzlich danken. Ohne sie wäre die Arbeit wohl nicht fertig geworden.

Andreas Hotho
Kassel, August 2004

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einführung

1.1 Motivation

Die Clusteranalyse teilt Objekte in aussagefähige, bedeutungsvolle und nützliche Gruppen (Cluster) ein. Heute hat sie sich ihren Platz in vielen Anwendungsbereichen gesichert. Eingesetzt wird die Clusteranalyse z.B. in der Biologie, um Gene und Proteine mit ähnlicher Funkionalität zu finden. Den gemeinsamen Zugriff auf ähnliche Objekte einer Datenbank kann man durch ihren Einsatz beschleunigen. Sie wird auch zur Buchstabenerkennung in der Bildverarbeitung eingesetzt. Das Gruppieren von Kunden im Marketing oder die Unterstützung des Browsens bzw. Blätterns im World Wide Web sind weitere bekannte Anwendungsfelder.

Der Benutzer kann beim Browsen im Internet auf eine enorme Menge an Dokumenten und damit auf sehr viele Informationen zugreifen. Dies birgt aber auch die Gefahr, sich in dieser riesigen Menge an Information zu verirren und die gesuchte Information nicht finden zu können. Portale wie Yahoo oder Web.de[1] versuchen, manuell die Informationen zu strukturieren und den Anwender so bei der Suche zu unterstützen. Hierbei gruppieren sie Dokumente und weisen den Gruppen Themen zu, die in einer Hierarchie angeordnet sind. Clusterverfahren können sowohl bei der automatischen Erstellung der Gruppen aber auch der Hierarchien eingesetzt werden. Einen anderen, aber ähnlichen Weg geht Vivisimo[2]. Die Metasuchmaschine gruppiert die Ergebnisse von herkömmlichen Suchmaschinen zur schnelleren und verständlicheren Präsentation automatisch mit Hilfe der Clusteranalyse. Für die immer größer werdende Menge von Textdokumenten vor allem im World Wide Web, aber auch in Dokument-Management-Systemen in internen Firmennetzen, stellt das automatische und effiziente Berechnen von Clustern ein immer wichtigeres Mittel zur *erstmaligen und automatischen Strukturierung* von sehr großen Dokumentsammlungen oder zur ad hoc Gruppierung von kleineren Dokumentmengen dar.

Des Weiteren finden Clusterverfahren auch Anwendung im Customer Relationship Management bzw. im Marketing zur Segmentierung von Kunden. Hier sammeln Unternehmen zunehmend Informationen über Millionen von Kunden. Marketingmaßnahmen können dabei meist nicht auf jeden Kunden individuell abgestimmt werden. Mittels der Clusteranalyse werden Kunden gruppiert und Marketingmaßnahmen gezielt auf homogene Kundengruppen zugeschnitten.

Bei der Durchführung einer Clusteranalyse arbeiten Spezialisten aus dem Bereich der Statistik oder des Data Minings typischerweise mit Experten aus dem Anwendungsgebiet zusammen. So wird sichergestellt, dass die Ergebnisse auch zu der jeweiligen Aufgabe aus der Praxis passen. Während der Lösung der Aufgabe fließen in diesen Prozess auch viele anwendungsspezifische Informationen ein, die den Erfolg garantieren sollen. Sehr häufig steuert das Wissen der Experten z.B. die Auswahl oder Kombination der eingesetzten Merkmale. Für das Clustern oder die Segmentierung ist die Auswahl und Aufbereitung der verwendeten Merkmale sowie ein entsprechendes Domänenwissen essentiell [58]. So schreiben die Autoren in [58] S. 12: “[...] As with segmentation, the task of feature extraction is much more problem- and domain-dependent [...] Although the pattern classification techniques presented in this book cannot substitute for domain knowledge, [...]” und

[1] http://www.yahoo.com/ bzw. http://web.de/
[2] http://vivisimo.com/

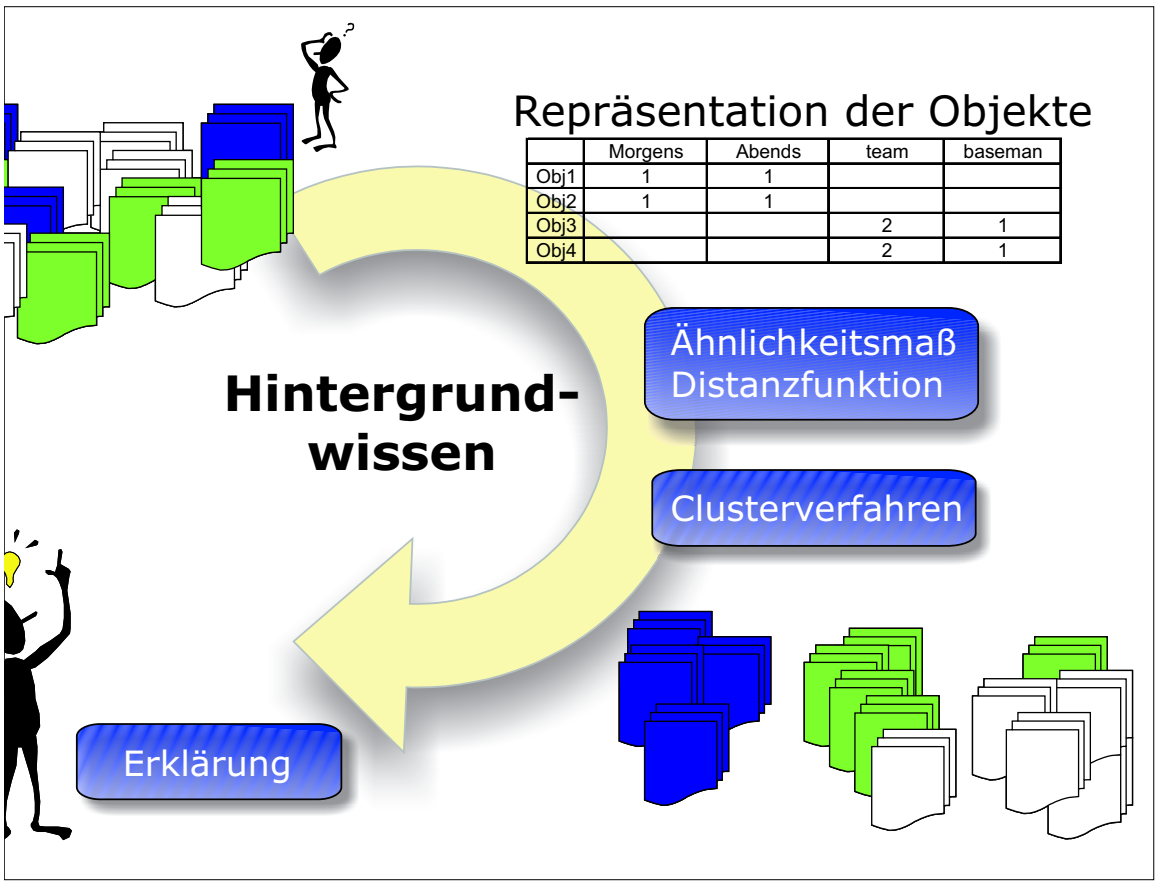

Abbildung 1.1: Der Clusterprozess

machen damit in diesem Zusammenhang klar, dass Wissen über die Domäne bei der Segmentierung helfen kann. Sie geben allerdings nicht an, wie dieses Wissen in den Prozess einfließen soll. Neben der trivialen Alternative, auf den Domänenexperten mit seinem Wissen zurückzugreifen und ihn bei jedem Schritt der Analyse zu befragen, wird in dieser Arbeit das Wissen bzw. Hintergrundwissen mittels *formaler Repräsentation* in Form von Ontologien automatisch in den Prozess integriert. Damit kann bisher der Benutzer nicht ersetzt werden, aber es wird ein Weg zur Integration des Wissens gezeigt. Wie in dieser Arbeit gezeigt wird, ist Domänenwissen ein wichtiger Faktor, um erfolgreich Clusterverfahren einsetzen zu können.

1.2 Problemstellung

Immer wieder kommt es vor, dass bei der Bildung von Gruppen nicht alle wichtigen Merkmale beachtet werden. Auch können zwischen einzelnen Merkmalen so komplexe Beziehungen existieren, dass deren Einfluss auf die Bildung von Gruppen nicht immer von den zu Grunde liegenden mathematischen Modellen korrekt erfasst werden kann. Andere Ursachen sind die Repräsentation der Objekte oder die Funktionen zur Berechnung der Ähnlichkeiten oder Distanzen, die die Beziehung zwischen den Objekten nicht immer korrekt ermitteln.

Im Folgenden wollen wir die Begriffe “Bilden von Gruppen”, “Gruppierung”, “Segmentierung” und “Clustern” synonym verwenden. Eine genauere Definition zum Begriff “Cluster” findet man in Kapitel 5.1.

Wir werden in Kapitel 3.2 den Knowledge Discovery Prozess einführen, der die Grundlage für anwendungsgetriebene Analysen darstellt. Daraus leitet sich der Clusterprozess aus Abbildung 1.1 ab. Ausgangspunkt bildet eine Menge von Objekten, die in Gruppen einzuteilen sind. Diese Aufgabe ergibt sich aus der Analyse des Geschäftsproblems. Die Anwender der Ergebnisse haben sehr häufig eine Vorstellung, wie (nach welchen Kriterien) die Cluster zu bilden sind. Diese müssen in den Clusterprozess einfließen, um das Ziel, den Vorstellungen des Anwenders entsprechend, zu erreichen. Ebenso begleitet der Anwender den gesamten Clusterprozess bis zum Schluss. Die einzelnen Schritte des Prozesses sind in Abbildung 1.1 zu finden und werden im Folgenden erläutert.

Die in der linken oberen Ecke symbolisierten Dokumente stellen die Menge der Objekte dar, die in Gruppen einzuteilen sind. Für die Durchführung dieser Aufgabe benötigt man neben einer geeigneten Repräsentation auch eine Maß für die Ähnlichkeit bzw. die Distanz zweier solcher Objekte. Die Tabelle rechts oben in Abbildung 1.1 repräsentiert die Objekte durch eine Menge von Merkmalen (Spalten), wie z.B. “Morgens” oder “team”. Die Merkmale bilden die Grundlage für ein Ähnlichkeitsmaß oder eine Distanzfunktion. Diese Funktionen setzen die Objekte in Beziehung zueinander und geben dafür einen numerischen Wert an. Auf dieser Basis können nun ganz unterschiedliche Verfahren zur Berechnung von Clustern angewendet werden. Ein solches Clusterverfahren liefert die gesuchte Gruppierung entsprechend der gegeben Repräsentation und des Ähnlichkeitsmaßes bzw. der Distanzfunktion der Objekte. Der Prozess endet mit der anschaulichen Präsentation der berechneten Cluster, die gleichzeitig dem Benutzer eine Erklärung der Clusterinhalte durch eine passende Visualisierung liefert.

Um die Ergebnisse eines solchen unüberwachten Verfahrens überprüfen zu können, wird eine genaue Beschreibung des Zieles benötigt. Eine Variante, die auch zur Berechnung von Maßzahlen einer solchen Evaluierung benutzt werden kann, ist in Abbildung 1.1 durch die unterschiedlichen Farben/Grautöne der Objekte (links oben und rechts unten) angedeutet. Die Objekte – in diesem Fall die Dokumente – sollen hier in Gruppen gleichen Inhaltes eingeteilt werden. Der Mensch ist in der Lage, die Einteilung in Gruppen vorzunehmen. Weiterhin ist es Ziel des Clusterprozesses, die Gruppen auch entsprechend dem Inhalt der Dokumente zu bilden. Es liegt daher nahe, die Einteilung eines Menschen (links oben in der Abbildung) als Basis für Vergleiche mit der Clusterung zu verwenden. Dabei wird als Grundannahme vorausgesetzt, dass die Clusterung möglichst mit der “menschlichen” Einteilung übereinstimmt. Sind beide Gruppierungen identisch, dann ist das Ziel der Clusterung erreicht.

Wie in Abbildung 1.1 rechts unten an den Farben/Grautönen zu erkennen ist, wird jedoch eine perfekte Übereinstimmung der beiden Gruppierungen nicht immer vorkommen. Es ist auch die berechtigte Frage zu stellen, ob dieses Idealziel überhaupt zu erreichen ist. Vergleicht man dazu mehrere manuelle Gruppierungen, so findet man sehr schnell auch hier unterschiedlichen Einteilungen (vgl. [33, 38]). Bis zu einem gewissen Grad kann man die automatische Berechnung von Clustern verbessern gemäß einer gegebenen Einteilung. Eine “perfekte” Lösung ist aber nicht zu erwarten.

Wichtiger ist daher, Benutzern die berechnete Clusterlösung in geeigneter Art und Weise zu präsentieren. Dabei sind Verfahren einzusetzen, die dem intuitiven Verständnis des Benutzers entgegen kommen und gleichzeitig die Terminologie des Benutzers beachten. Bei solch einer Präsentation der Ergebnisse erhält der Benutzer sehr schnell einen Einblick in die berechneten Cluster, versteht die Art und Weise der Clusterung und findet gegebenenfalls sehr leicht “fehlerhafte” Zuordnungen von Objekten.

Der Clusterprozess baut auf ganz unterschiedlichen Methoden auf und bietet daher verschiedenste Ansatzpunkte zur Verbesserung der Ergebnisse. Viele Arbeiten präsentieren verbesserte Ergebnisse durch die Modifikation vorhandener oder die Entwicklung neuer Clusterverfahren (vgl. z.B. [109, 107, 182, 199, 198]. Weitere Ansatzpunkte sind die Ähnlichkeitsmaße und die Distanzfunktio-

nen (vgl. z.B. [209]). Die hier vorgestellte Arbeit setzt am dritten möglichen Punkt an, nämlich der Repräsentation der Objekte. *Hintergrundwissen* wird an dieser Stelle in den Prozess eingebracht. Die veränderte Repräsentation führt sowohl zur *Steigerung der Verständlichkeit* als auch zur *Verbesserung der Güte* der Ergebnisse. Hintergrundwissen stellt damit in dieser Arbeit einen ganz zentralen Bestandteil dar und beeinflusst durch die Integration in die Repräsentation der Objekte die Bewertung durch Ähnlichkeitsmaße und Distanzfunktionen sowie die berechnete Gruppierung der Clusterverfahren. Dabei können unterschiedlichste Clusterverfahren und Maße mit dieser neuen Repräsentation verwendet werden.

Der Clusterprozess muss für eine erfolgreiche Anwendung gleichzeitig mit ganz verschiedenen Problemstellungen zurechtkommen. Aus diesem Grund wurde immer noch nicht die “ultimative” Methode des Clusterprozesses gefunden, welche allen Clusteraufgaben gleich gute Ergebnisse liefert. Identifiziert wurden die folgenden Problemstellungen:

1. *Effizienz*: Die immer größer werdende Menge an Daten verlangt immer effizientere Verfahren für das Clustern in sehr kurzer Zeit. Gerade im Bereich des Internets wird die Forderung nach einer sehr kurzen Laufzeit für den gesamten Clusterprozess gestellt. Ziel ist dabei z.B. eine Menge von Dokumenten, die man als Resultat einer Anfrage bei einer Suchmaschine erhält, noch vor der Präsentation in Gruppen einzuteilen, um so eine Aufwertung des Suchergebnisses zu erhalten.

2. *Effektivität*: Mit der Effektivität des Clusterns wird die Frage nach der Art und Weise – wie die Gruppen gebildet werden – angesprochen. Clustern ist nur dann von Interesse, wenn es effektiv in Bezug auf die Anwendung ist. Zum Beispiel beim Text-Dokument-Clustern sollten die Dokumente ähnlichen Inhaltes in gleiche Gruppen eingeteilt werden. Beim Clustern von Kunden der Deutschen Telekom AG möchte man in erster Linie Kunden mit gleichem Kommunikationsverhalten zusammenfassen.

3. *Erklärungsfähigkeit*: Nach der Berechnung der Cluster wird vom Benutzer häufig die Frage nach einer Begründung für die gebildeten Cluster und nach einer verständlichen Präsentation gestellt. Eine intuitive Präsentation der Ergebnisse ist in den meisten Fällen wesentlich wichtiger als eine um wenige Prozentpunkte gesteigerte Clustergüte.

4. *Benutzerinteraktion und Subjektivität*: Cluster werden mit Hilfe statistischer Größen berechnet. Es ist in den meisten Fällen das Ziel, der Vorstellung eines Benutzers entsprechend zu clustern. Da “nur” durch die Wahl der Datenvorverarbeitung und des Verfahrens die Ergebnisse beeinflussbar sind, ist das Ziel an dieser Stelle nur schwer zu erreichen. Wünschenswert wäre eine Methode mit einer verbesserten Integration der Interessen des Anwenders in den Clusterprozess.

Für die beschriebenen Probleme werden im nächsten Unterkapitel Lösungsvorschläge präsentiert, die dann sukzessive im weiteren Verlauf der Arbeit ausführlich behandelt werden. Dabei spielt Hintergrundwissen eine zentrale Rolle.

1.3 Lösungsansätze der Arbeit

In der Arbeit werden drei neu entwickelte Methoden zur Verwendung von Hintergrundwissen beim Clustern vorgestellt. Dies spiegelt sich in der Struktur der Arbeit wider:

- Subjektives Clustern berechnet benutzerbezogene Cluster bei gleichzeitiger Dimensionsreduktion. Damit wird u.a. die Verständlichkeit der Ergebnisse steigert (vgl. Abschnitt 1.3.1).
- Hintergrundwissen kann während der Vorverarbeitung der Dokumente erfolgreich in den Clusterprozess integriert werden (vgl. Abschnitt 1.3.2).
- Erstmals werden auch Verfahren der Formalen Begriffsanalyse zur Präsentation von Clustern verwendet, die für Menschen leicht verständliche Beschreibungen der berechneten Cluster liefern (vgl. Abschnitt 1.3.3).

Die entwickelten Methoden werden in zwei Anwendungsgebieten eingesetzt und evaluiert. Einerseits werden die Kunden der Deutschen Telekom AG anhand ihrer Verbindungsdaten, andererseits Textdokumente aus unterschiedlichen Domänen geclustert (vgl. Kapitel 2 und Teil III).

1.3.1 Subjektives Clustern

"Subjektives Clustern" verfolgt zwei Ziele. Auf der einen Seite soll dem Benutzer die Möglichkeit eingeräumt werden, mehr Einfluss auf den Clusterprozess zu nehmen. Auf der anderen Seite wird die Dimensionalität des Merkmalsraumes durch die Auswahl von geeigneten Merkmalen und der Aggregation gemäß einer Ontologie reduziert. Die dazu entwickelten Algorithmen werden in Kapitel 7 beschrieben. Im Folgenden wird auf die Notwendigkeit von Subjektivem Clustern eingegangen und anhand eines Beispiels die prinzipielle Idee erläutert.

1.3.1.1 Ausgangspunkt für Subjektives Clustern

Gegeben ist eine Menge von Objekten, die geclustert werden soll. Stellen wir uns außerdem auf den Standpunkt, dass per se jedes Objekt einmalig ist, können wir nicht Clustern. Die vom Benutzer zu beantwortende Frage lautet dann: "Welche Gemeinsamkeiten bzw. welche Unterschiede sind wichtig (für ihn oder für die Anwendung)?" Dies kann zu ganz unterschiedlichen Antworten für die gleichen Objekte führen. Die relevanten Merkmale wählt der Benutzer normalerweise aus, formt sie in geeigneter Weise um und bringt sie in den Clusterprozess ein. Auf diese Weise hilft er dem Verfahren, die Gruppen an den "richtigen" Stellen zu suchen. Der Auswahl und Transformation der Merkmale kommt hier entscheidende Bedeutung für die Ergebnisse der Clusterung zu. Wählt der Benutzer die "falschen" Merkmale oder transformiert er sie in ungeeigneter Weise, werden Objekte gleich behandelt, die gar nicht gleich sind, und Unterschiede als wichtig herausgehoben, die als unwichtig erachtet werden. Eine Clusterung wird so nicht erfolgreich sein.

Die Anwender spielen an dieser Stelle eine zentrale Rolle. Verschiedene Anwender können unterschiedliche Positionen einnehmen und so die Vorverarbeitung in unterschiedliche Richtungen lenken. Im Extremfall führt dies zu völlig disjunkten Clusterungen. Z.B. wird einerseits ein Manager, der seine geschäftlichen Interessen in den Vordergrund stellt, nicht an technischen Details interessiert sein. Andererseits wird ein Techniker gerade die technischen Details als primäres Ziel haben (vgl. [152]). Die beiden Sichtweisen und die unterschiedlichen Interessen werden sich typischerweise in unterschiedlichen Clusterungen widerspiegeln. Standardclusterverfahren berücksichtigen solche Interessen nur unzureichend, da sie auf der Basis der vorhandenen Merkmale die "objektiv" beste Clusterung berechnen. Der subjektive Standpunkt des Anwenders fließt in die Clusterung nicht ein, da es prinzipbedingt auch keine solche Clusterung gibt. Nur eine Menge von Clusterungen auf der Basis von verschiedenen Merkmalen kann dieses Problem lösen. Wir nennen im weiteren Verlauf eine Menge von ausgewählten Merkmalen eine "Sicht" (im Englischen "View").

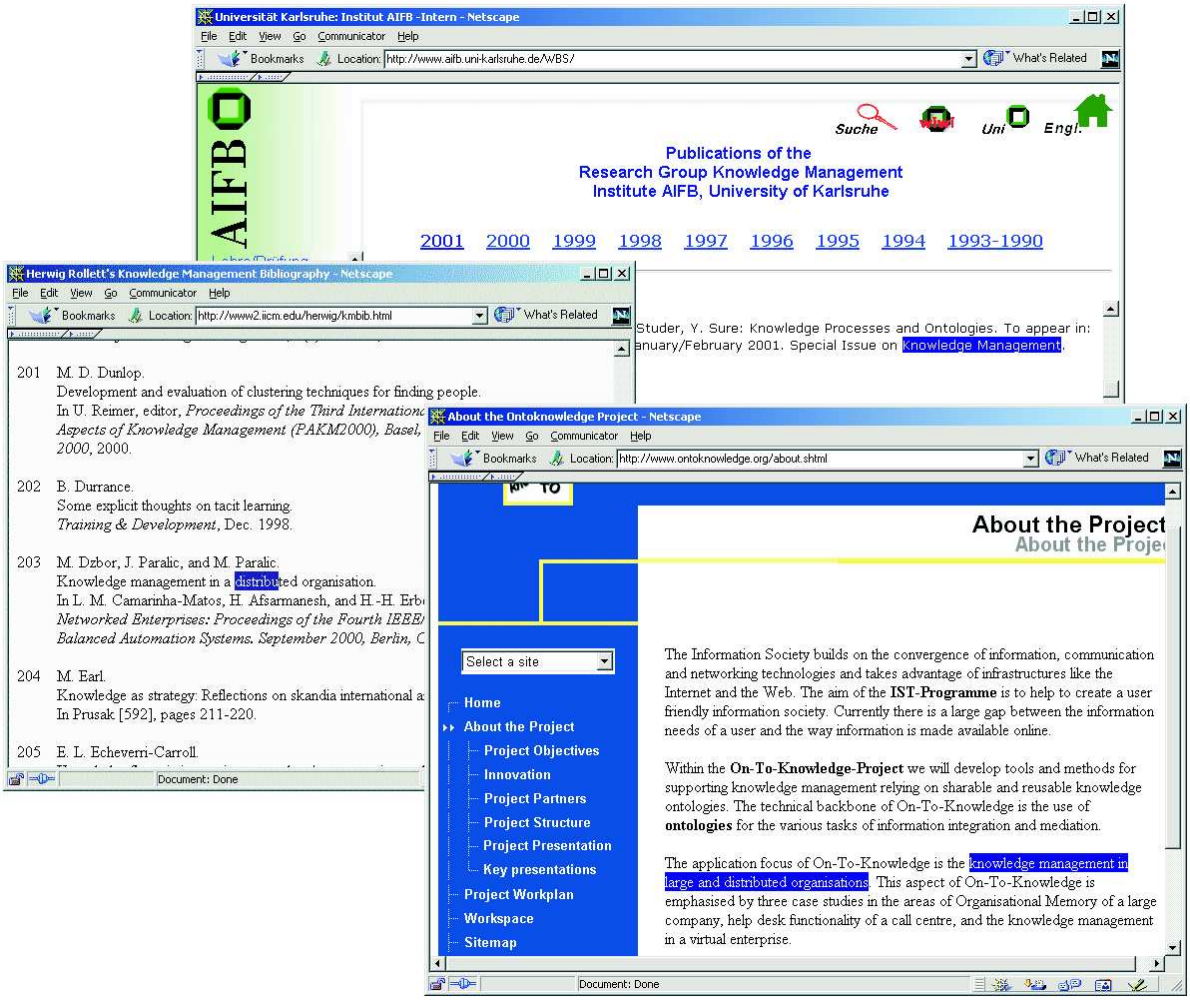

Abbildung 1.2: Beispiel Web-Seiten (von hinten nach vorn: AIFB Publikation(1), IICM Publikation(2) und OTK(3))

Das folgende Beispiel soll die Problematik etwas besser verdeutlichen. Die Web-Seiten der Abbildung 1.2 stellen die Dokumentenmenge dar.[3] Wir möchten wissen, wie ähnlich sich diese Seiten sind. Eine mögliche Antwort könnte wie folgt lauten: Zwei Seiten präsentieren Veröffentlichungen von Institutionen. Die dritte Seite bietet eine Überblick zu einem europäischen Forschungsprojekt. Aus diesem Grund sind sich die beiden ersten Seiten ähnlicher als die erste und die dritte sowie die zweite und die dritte. Eine ganz andere Aussage in Bezug auf die Ähnlichkeit erhält man, wenn die Institutionen, die in Beziehung zu den Web-Seiten stehen, wichtig sind. Das AIFB taucht sowohl auf der “AIFB Publikation”- als auch auf der “OTK”- Seite auf, hat jedoch nichts mit der “IICM Publikation”-Seite zu tun. Eine ähnliche Aussage wie die letzte erhält man, wenn man folgende gestalterischen Elemente der Seiten als Basis betrachtet. Die erste und die dritte Seite enthalten beide Rahmen (links und oben). Die zweite Webseite besteht nur aus einer Aufzählung. Allerdings enthält auch die erste Webseite (“AIFB Publikation”) eine solche Aufzählung. Wir könnten dieses Beispiel beliebig fortsetzen. Menschen würde wahrscheinlich die Seiten aus immer neuen Gründen in Gruppen einteilen, wobei es auch Gruppen von Personen geben wird, die die gleiche Basis nutzen, um die Web-Seiten in Gruppen einzuteilen.

[3] Zwei Seiten geben die Publikationen der beiden Forschungseinrichtungen AIFB und IICM wieder. Die dritte Seite gehört zum europäischen Forschungsprojekt On-To-Knowledge (OTK).

Beim Vergleich der verfügbaren Merkmale zur Bestimmung der Ähnlichkeit der drei Web-Seiten wird sehr schnell der Einfluss der verwendeten und die Auswahl der "richtigen" Merkmale sichtbar. Die Merkmale lassen sich an dieser Stelle nicht objektiv bestimmen, da nur die subjektive Aussage des Benutzers eine Auswahl ermöglicht. Dies führt auch zum Namen der Methode "Subjektives Clustern". Während der Vorverarbeitung der Daten werden die zur Clusterung verwendeten Merkmale ausgewählt und die Cluster werden auf der Basis dieser Merkmale berechnet. Die in Kapitel 7 vorgestellte Methode erzeugt dafür nicht nur eine Merkmalsmenge, sondern eine Menge von Merkmalsmengen. Sie erlaubt es, unterschiedlichen Anwendern vorberechnete Clusterungen nach ihren Gesichtspunkten auszuwählen. Dazu wird eine Strukturierung des Merkmalsraumes benötigt, die in dieser Arbeit durch eine Ontologie (siehe Kapitel 6) bereit gestellt wird.

Die Auswahl von Merkmalen führt nicht nur zur Fokussierung auf ein Themengebiet, sondern auch zur Reduktion der Anzahl. Der Aufwand zur Berechnung der Cluster reduziert sich damit ebenfalls drastisch. Clustern von Texten erfolgt typischerweise im hochdimensionalen Raum und stößt dort auf prinzipielle Probleme, die in [25] erstmals auch mit empirischen Ergebnissen belegt wurden. Jedes Objekt im hochdimensionalen Raum ist in der Tendenz ungefähr gleich weit von den anderen Objekten entfernt (siehe Kapitel 10.1.3.1), d.h. es ist sehr schwierig, Gruppen zu finden. Subjektives Clustern bietet hierfür einen Lösungsansatz.

1.3.1.2 Idee des Subjektiven Clusterns

Im folgenden Unterkapitel wollen wir die Idee des Subjektiven Clusterns an einem einfachen Beispiel verdeutlichen. Wir möchten die Web-Seiten anhand der vorkommenden Worte clustern mit der Idee, dass die Worte den Inhalt der Web-Seiten repräsentieren. Kommen wir dazu zurück zu den drei Web-Seiten aus Abbildung 1.2 und betrachten wir den Ausschnitt einer sehr einfachen Ontologie aus Abbildung 1.3. Sie besteht nur aus Konzepten[4] wie PUBLICATION oder TOPIC und den zugehörigen taxonomischen Beziehungen. Jedes Wort im Text kann auf ein Konzept abgebildet werden unter der Annahme, dass jedes Wort genau wie der lexikalische Eintrag des Konzeptes geschrieben wird. Tabelle 1.1 gibt für jede Webseite an, wie häufig die Konzepte PUBLICATION, KNOWLEDGE MANAGEMENT und DISTRIBUTED ORGANIZATION vorkommen. Nehmen wir weiterhin für dieses Beispiel diese Worte (und damit die Konzepte) als die einzigen wichtigen Worte der Texte an. Für die Berechnung der Ähnlichkeit verwenden wir die bekannte euklidische Metrik (vgl. 5.2). Damit ergibt sich für Objekt "OTK" der Wortvektor $\vec{t_d} = (0, 2, 1)$.

Tabelle 1.1: Beispiel für eine Konzept Vektor Repräsentation für die drei Web-Seiten aus Abbildung 1.2

Document #	PUBLICATION	KNOWLEDGE MANAGEMENT	DISTRIBUTED ORGANIZATION
1 ("AIFB Publ.")	1	2	0
2 ("IICM Publ.")	1	1	1
3 ("OTK")	0	2	1

Berechnen wir die Abstände der Web-Seiten auf der Basis dieser Konzepte, so beträgt die quadrierte euklidische Distanz zwischen jeweils zwei Seiten 2. Ein Clustern der drei Seiten ist nicht möglich. Erinnern wir uns an das einleitende Beispiel aus Kapitel 1.3.1.1, in dem unterschiedliche Merkmale zum Gruppieren der Seiten verwendet wurden. Die einfachste Möglichkeit, diese Idee zu übertragen, besteht im Streichen eines Konzeptes, z.B. DISTRIBUTED ORGANIZATION, aus der Tabelle 1.1. Die Distanz ergibt sich dann zu 1 für $d(1, 2)$ und $d(1, 3)$ und zu 2 für $d(2, 3)$. Der

[4] In der Arbeit wird das aus dem englischen übernommene Wort "Konzept" für die Konzepte einer Ontologie und nicht das deutsche Wort "Begriff" verwendet. Das Wort "Begriff" wird im Kontext der Formalen Begriffsanalyse eingesetzt. So sollen Konzepte und (Formale) Begriffe eindeutig unterschieden werden.

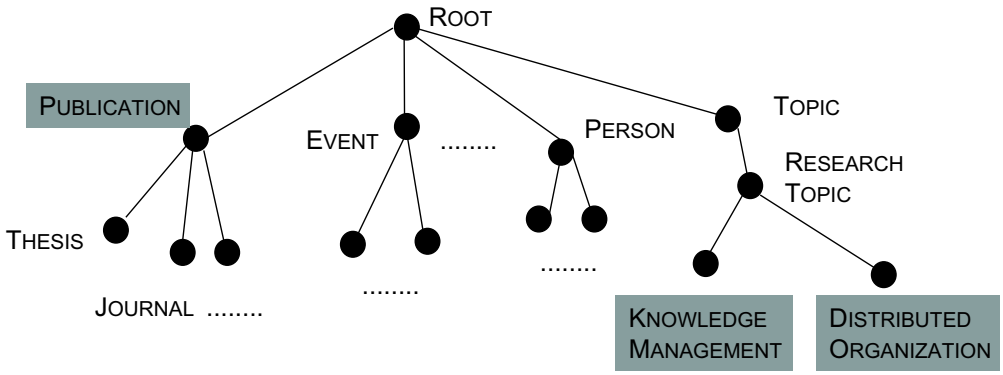

Abbildung 1.3: Beispiel Ontologie

Abstand zwischen den beiden Publikationsseiten, aber auch der Abstand zwischen der "OTK" und der "AIFB-Publikationsseite" wird kleiner. Andersherum werden 2 und 3 durch den Wegfall der verteilten Organisationen nicht mehr als so ähnlich betrachtet.

Auch jetzt ist ein Clustern der Dokumente noch schwierig. Nutzen wir die Ontologie aus Abbildung 1.3 und löschen nicht einfach das Konzept DISTRIBUTED ORGANIZATION, sondern verallgemeinern die beiden Forschungsgebiete zum Konzept RESEARCH TOPIC, dann ergeben sich die folgenden Distanzen: $d(1,2) = 0$, $d(1,3) = d(2,3) = 2$. Die beiden Publikationsseiten, die beide Veröffentlichungen zu verschiedenen Forschungsthemen enthalten, haben nun die gleiche Repräsentation und unterscheiden sich deutlich von der "OTK"-Seite. Die Ontologie wurde an dieser Stelle genutzt, um die für einen Benutzer wesentlichen Informationen als Basis einer Clusterung zu nutzen. Vorstellbar wäre auch der umgekehrte Schritt, d.h. die Auswahl von spezielleren Unterkonzepten von PUBLICATION, wie z.B. JOURNAL oder THESIS. Diese Informationen erlauben die Clusterung der beiden Publikationsseiten anhand völlig anderer Konzepte, also anhand unterschiedlicher Präferenzen.

1.3.1.3 Reduktion der Dimensionalität

Die Veränderung der Repräsentation kann für einen weiteren Effekt genutzt werden. Der Anwender überblickt bei der Präsentation von Ergebnissen nur eine geringe Anzahl an Merkmalen. Nutzt man die Ontologie, um eine veränderte Repräsentation der Web-Seiten zu erzeugen, erfolgt gleichzeitig eine Dimensionsreduktion. Dabei können sowohl sehr wenige allgemeine Konzepte als auch jede denkbare andere Kombination aus allgemeinen und speziellen Konzepten ausgewählt werden. Dies ist prinzipiell auch ohne Ontologie möglich, hat dann aber neben dem Verlust an Informationen auch den entscheidenden Nachteil, dass keinerlei Wissen über die Beziehungen der ausgewählten Konzepte existiert. Diese durch Ontologien bereitgestellten strukturellen Informationen helfen dem Anwender bei der Interpretation der Ergebnisse und können zur Verfeinerung herangezogen werden. Details findet man in Kapitel 10.

1.3.2 Clustern mit Hintergrundwissen

Das Clustern von Objekten basiert im Allgemeinen auf statistischen Maßen. In dieser Arbeit wird während der Vorverarbeitung der Daten formal repräsentiertes Hintergundwissen in die Repräsentation der Daten integriert und während der Clusterung der Objekte genutzt. Für die Clusterung der Objekte werden bekannte Maße und Verfahren aus der Statistik und dem Maschinellen Lernen eingesetzt. Neben der empirischen Evaluierung (vgl. Kapitel 8) wurde mittels Varianzanalyse

die Integration des Hintergrundwissens in die vorhandenen und in Klassen eingeteilten Dokumente anhand von mehreren Datensätzen untersucht.

Im Kapitel 1.3.1 wurde die Idee erläutert, wie eine Ontologie zur Struktuierung des Merkmalsraumes, zur Auswahl und Generierung von Merkmalen und zur Dimensionreduktion eingesetzt werden kann. Von Interesse ist nun der Schritt vom Merkmal im Allgemeinen zum Konzept der Ontologie. Wir verwenden als Objekte Dokumente. Bei Dokumenten bestehen die Merkmale aus Worten oder allgemeiner aus Termen, d.h. dass die Abbildung des Terms eines Dokumentes auf das Konzept einer Ontologie von Interesse ist. Im Folgenden werden der Weg, die auftretenden Probleme beim Abbilden und die Ideen zu deren Lösungen skizziert.

1.3.2.1 Abbildung von Worten auf Konzepte

Das Abbilden oder auch Mapping der Worte auf die Konzepte der Ontologie ist ein zentraler Punkt. Jedes Wort hat im Kontext eines Text-Dokumentes eine bestimmte Bedeutung. Wählt man Konzepte mit einer anderen Bedeutung beim Abbilden aus, so greifen die nachfolgenden Schritte zur Nutzung des formal repräsentierten Hintergrundwissens nicht. Erweitern wir unser Ontologiebeispiel aus Abbildung 1.3 und hängen zwei Konzepte mit den Namen KNOWLEDGE MANAGEMENT und DISTRIBUTED ORGANIZATION an das Konzept EVENT[5].

Die Wortsinnerkennung (im Englischen "word sense disambigation") beschäftigt sich mit der Identifizierung des verwendeten Sinnes eines Wortes im gegebenen Kontext (vgl. Kapitel 8.2.3.3). Durch die Erweiterung der Ontologie haben wir jetzt das Problem, das Wort im Text auf das richtige Konzept abbilden zu müssen. Das richtige Konzept bedeutet in diesem Zusammenhang dasjenige Konzept, welches den Sinn des Wortes im gegebenen Kontext richtig wiedergibt.

Im unserem Beispiel können wir "Knowledge Management" auf zwei Konzepte mappen, einmal das Konzept unter EVENT und einmal unter RESEARCH TOPIC. Im Folgenden wollen wir der Frage nachgehen, welche Fehler man bei Mapping machen kann und welche Auswirkungen diese auf die Ähnlichkeitsbeziehungen unseres Beispiels haben. Nehmen wir dazu an, dass "Knowledge Management" ein "Research Topic" in allen drei Dokumenten ist. Mappen wir jetzt fälschlicher Weise immer auf das alternative Konzept unter EVENT, so verändert sich unsere Situation vorerst nicht. Beziehen wir aber die Generalisierung in Form der Konzepte RESEARCH TOPIC und nun auch EVENT mit ein, würde das fehlerhafte Mapping zum Verlust des gemeinsamen Oberkonzeptes zwischen KNOWLEDGE MANAGEMENT und DISTRIBUTED ORGANIZATION führen. Erst das Konzept ROOT, also das allgemeinste Konzept, würde wieder eine Brücke schlagen, wobei zu beachten ist, dass ROOT alle Konzepte miteinander in Beziehung setzt und aus diesem Grund seine Nutzung wenig Sinn macht.

Verändern wir unser Beispiel und unterstellen für das "OTK"-Dokument, dass "Knowledge Management" jeweils einmal auf das Konzept unter EVENT und unter RESEARCH TOPIC abgebildet wurde. Damit ergibt sich folgende Tabelle für die Vektorrepräsentation der Dokumente:[6]

Die Distanzen ergeben sich zu: $d(1,2) = 2$, $d(1,3) = 4$ und $d(2,3) = 4$. Das Beispiel illustriert sehr anschaulich den Einfluss des Mappings auf die entsprechende Wortbedeutung.

Bisher haben wir angenommen, wir könnten die Bedeutung von Worten eines Textes herausfin-

[5] In Abbildung 1.3 sind nicht die Identifikatoren der Konzepte, sondern die Namen dargestellt. Sollten die Namen eindeutig sein, so kann man den Identifikator und den Namen auch gleich setzen. Dies kommt aber vor allen Dingen bei großen Ontologien selten vor. In unserem Beispiel gibt es daher zwei Konzepte, die den gleichen Namen haben, aber unterschiedliche Identifikatoren.

[6] Die Konzepte wurden aus Platzgründen wie folgt abgekürzt: PUBL. = PUBLICATION, KM (RT) = KNOWLEDGE MANAGEMENT unter RESEARCH TOPIC, DO = DISTRIBUTED ORGANIZATION und KM (EVENT) = KNOWLEDGE MANAGEMENT unter EVENT

Tabelle 1.2: Modifizierte Vektorrepräsentation aus Tabelle 1.1

Document #	PUBL.	KM (RT)	DO	KM (EVENT)
1 (“AIFB Publications”)	1	2	0	0
2 (“IICM Publications”)	1	1	1	0
3 (“OTK”)	0	1	1	1

den und sie dann auf die entsprechenden Konzepte mappen. Abschließend für dieses Kapitel wollen wir den Fall betrachten, uns stünde diese Information nicht zur Verfügung. Dann hätten wir zwei Alternativen. Wir könnten raten oder wir mappen das Wort einfach auf alle vorhandenen Bedeutungen. Beide Fälle können für empirische Untersuchungen als Grundlage für die Bedeutung von Wortsinnerkennungen dienen. Wir schauen uns den zweiten Fall ein wenig genauer an.

Tabelle 1.3: Modifizierte Vektorrepräsentation aus Tabelle 1.1, mapping von “Knowledge Management” auf alle Konzepte KNOWLEDGE MANAGEMENT

Document #	PUBL.	KM (RT)	DO	KM (EVENT)
1 (“AIFB Publications”)	1	2	0	2
2 (“IICM Publications”)	1	1	1	1
3 (“OTK”)	0	2	1	2

Tabelle 1.3 enthält die neue Vektorrepräsentation. Außer der Redundanz in zwei Spalten ist keine wesentliche Veränderung der Ähnlichkeitsbeziehung zwischen den Dokumenten festzustellen. Dies wird erst deutlich, wenn man die entsprechenden Oberkonzepte mit in Betracht zieht. Sie stellen, wie schon besprochen, die Beziehung zwischen den Dokumenten trotz unterschiedlicher Worte für ein Thema her. Dies kann auch im Hinzufügen von Rauschen enden. In unserem Beispiel wird “Knowledge Management” nun in allen Fällen nicht nur als “Research Topic” betrachtet, sondern man muss auch davon ausgehen, dass etwas über Ereignisse auf den Seiten zu finden ist, was bei unseren Seiten nicht ganz ausgeschlossen werden kann, aber nicht den primären Inhalt wiedergibt. Hätte das Wort “Knowledge Management” weitere Bedeutungen (das Wort “Bank” hat im Englischen laut WordNet 10 verschiedene Bedeutungen) und wir würden nach der “all-Strategie” (siehe Kapitel 8.2.3.3) vorgehen, so würden wir die Dokumente, in denen das Wort mindestens einmal vorkommt, mit allen Dokumenten in Beziehung setzen, die mit einer dieser Bedeutungen in Beziehung stehen. Damit würden wir den Inhalt des Dokumentes nicht genauer einem Thema zuschreiben, sondern Rauschen einfügen.

Erwähnt sei an dieser Stelle noch, dass große lexikalische Ressourcen wie WordNet neben einer Vielzahl von Bedeutungen unterschiedlicher Worte auch eine Reihe von Synonymen enthalten. Durch das Mapping von unterschiedlichen Worten mit gleicher Bedeutung auf ein Konzept werden ebenfalls erste linguistische Informationen in den Prozess integriert. Auch dies wirkt sich – wie schon bei den Konzepten diskutiert – auf die Ähnlichkeitsbeziehung der Dokumente aus. Domänenspezifische Ontologien enthalten kaum lexikalische Einträge mit mehreren Bedeutungen, so dass die beschriebenen Probleme nur in abgeschwächter Form auftreten.

1.3.2.2 Erweiterung der Konzeptvektorrepräsentation

Das folgende Beispiel illustriert wie und warum die Integration von Hintergundwissen funktionieren kann. Betrachten wir dazu die Beispieltexte aus Abbildung 1.2. Die dazugehörige Vektorrepräsentation, auch “Bag of Words” genannt, wurde im Kapitel 1.3.1 schon vorgestellt. Nutzen wir die Ontologie aus Kapitel 1.3 und verändern die Repräsentation der Web-Seiten nicht wie in Kapitel

1.3.1.2 beschrieben, indem wir Konzepte durch deren Generalisierung zusammenfassen bzw. ersetzen, sondern indem wir die generelleren Konzepte zur Erweiterung der Dokumentrepräsentation nutzen.

Als Ausgangssituation unserer Beispielrechnung bei quadrierter euklidischer Distanz bestehen die gleichen Abstände (2) zwischen den drei Web-Seiten. Anstatt die beiden Konzepte KNOWLEDGE MANAGEMENT und DISTRIBUTED ORGANIZATION durch RESEARCH TOPIC zu ersetzen, fügen wir diese Information in Form eines weiteren Attributes hinzu. Zum Beispiel erhält man dann als Termvektor für das "OTK" Dokument $\vec{t_d} = (0, 2, 1, 3)$. Die Distanzen zwischen den Dokumenten sind $d(1, 2) = 2$, $d(1, 3) = 3$ und $d(2, 3) = 3$. Durch die Erweiterung des Vektors um ein gemeinsames Attribut verändert man die Ähnlichkeit zwischen den Dokumenten. Im übertragenen Sinne stellt das gemeinsame Oberkonzept RESEARCH TOPIC eine Verbindung zwischen den beiden Unterkonzepten KNOWLEDGE MANAGEMENT und DISTRIBUTED ORGANIZATION her und setzt damit auch die Dokumente, die ausschließlich diese Unterkonzepte enthalten, in Beziehung zueinander. Der Anwender ist durch die Angabe von Beziehungen (z.B. gemeinsamer Oberkonzepte) in der Ontologie in der Lage, explizit Konzeptbeziehungen zu modellieren. Die Übersetzung der Worte eines Textes in Konzepte erlaubt es dann, diese Beziehungen auch für das Clustern zu nutzen und Ähnlichkeitsbeziehungen im Sinne des Anwenders zu verändern.

Mit Hilfe der beschriebenen ontologiebasierten Repräsentation von Dokumenten ist man nun in der Lage, die Dokumente gleichen Inhaltes *besser* im Hinblick auf die Bedürfnisse des Anwenders in Gruppen zusammenzufassen.

1.3.3 Beschreibung der gefundenen Cluster

Die um Hintergrundwissen erweiterte Repräsentation der Dokumente führt nicht nur zu besseren Ergebnissen beim partitionierenden Clusterverfahren (vgl. Kapitel 5), sondern bildet auch die Basis für eine intuitive verständliche Erklärung der gebildeten Cluster. Im Folgenden wird die Wirkung des Hintergrundwissens durch die Repräsentationsveränderung auf die Erklärung der Ergebnisse an einem Beispiel erläutert.

Um die Ideen zur Beschreibung von Clusterergebnissen erläutern zu können, benötigen wir ein neues Beispiel. Abbildung 1.4 stellt den Begriffsverband (siehe Kapitel 5.5) passend zur Clusterung mit 50 Clustern aus 1015 Texten des Reuters-Korpus dar und hebt einen Teilverband hervor. Als Hintergrundwissen wurde WordNet benutzt. Der Verband wurde mit Cernato (einer Software der Firma NaviCon AG[7]) visualisiert. Details zur Formalen Begriffsanalyse findet man in Kapitel 5.5, zum Reuters-Korpus in Kapitel 2.1 und zu WordNet in Kapitel 6.3.3.1. Um die Idee der Beschreibung von Clustern zu verstehen, wollen wir an dieser Stelle den Inhalt der Abbildung 1.4 und der dort dargestellten Cluster 1 und 3 analysieren und so gleichzeitig die Idee der Analysemethode vorstellen.

Jeder Kreis des in Abbildung 1.4 gezeichneten Begriffsverbandes stellt ein formales Konzept dar. Der Verband wurde aus technischen Gründen gegenüber der gewöhnlichen Darstellung auf den Kopf gestellt, d.h. Objekte und Merkmale werden getauscht. Die dementsprechend angepasste Leserichtung des Verbandes ergibt die generellen formalen Begriffe im unteren Teil des Bildes, die dann aufsteigend immer spezifischer werden.

Betrachten wir den formalen Begriff mit der Bezeichnung 'refiner' in Abbildung 1.4, so besteht der Umfang aus den Elementen {CL1, CL3} und der Inhalt aus {(h)refiner, (h)oil, ..., (h)compound, chemical compound}, d.h. die Cluster CL1 und CL3 werden durch die Synsets REFINER, OIL usw. beschrieben (vgl. Kapitel 5.5 zum Lesen des Verbandes). Das (h) zeigt, dass dieses Label

[7] http://www.navicon.de

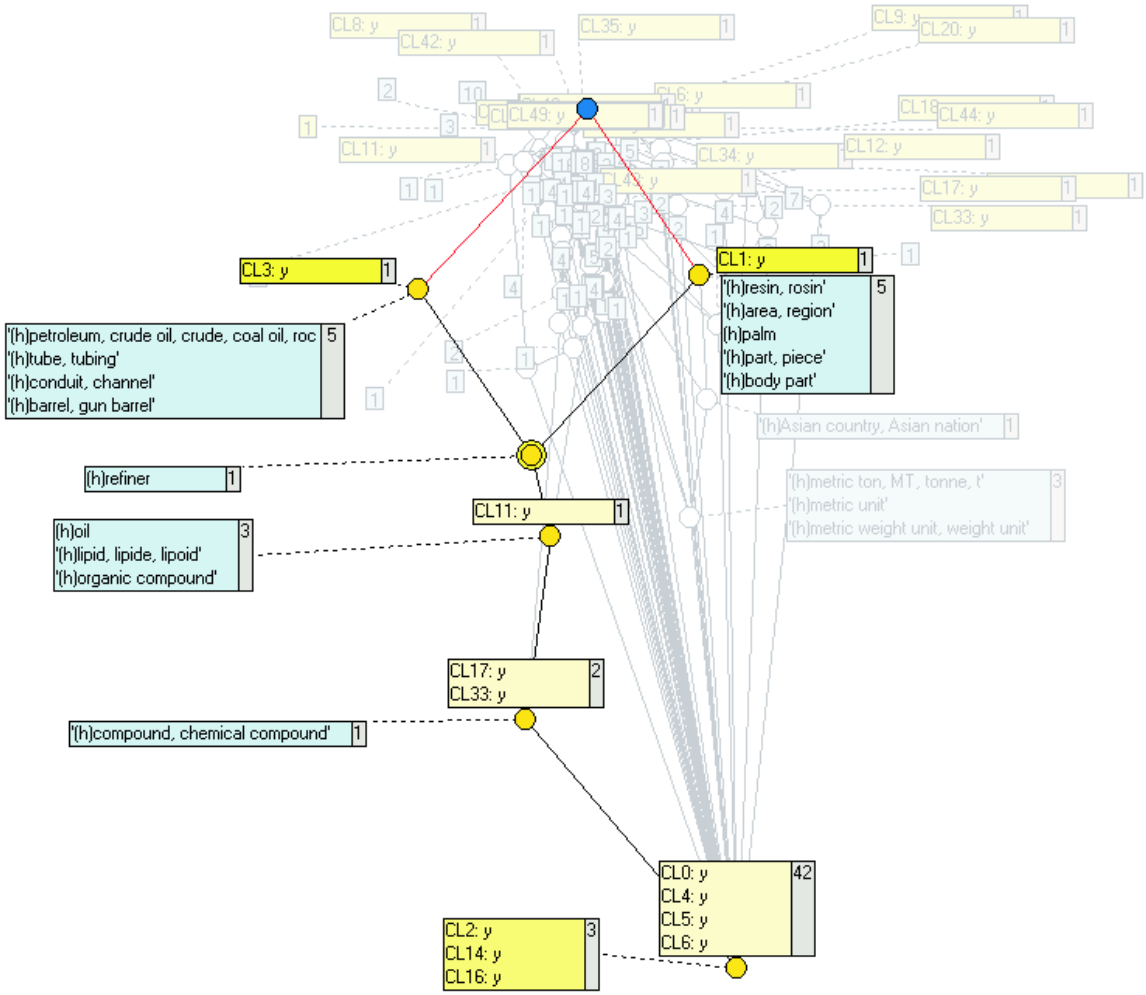

Abbildung 1.4: Einführendes Beispiel FCA, Verband mit zwei Clustern

ein WordNet-Synset ist und während der Vorverarbeitung hinzugefügt wurde. In der Abbildung ist weiterhin eine Kette vom Begriffen mit steigender Spezialisierung zu finden. Die Kette startet mit dem allgemeinsten formalen Begriff (außer dem Top-Begriff) mit der Bezeichnung ‘compound, chemical compound’. Der formale Begriff enthält im Umfang die Cluster CL1, CL3, CL11, CL17, CL33, deren Dokumente unter anderem über chemische Verbindungen sprechen. Der nächste formale Begriff schränkt die Inhalte auf ‘oil’ ein und enthält noch die Cluster CL1, CL3 und CL11. Um ‘refiner oil’ – also raffiniertes Öl – geht es in den Dokumenten der Cluster CL1 und CL3, die durch den nächstspezifischeren formalen Begriff gekennzeichnet sind.

Die bis jetzt genutzten formalen Begriffe und deren sehr allgemeine Bezeichner waren zur Eingrenzung des Themengebietes sehr nützlich. Wir wissen nun nicht nur, dass es in den Texten der Cluster um chemische Verbindungen geht, sondern auch, dass es sich um Öl handelt. Damit ist grob klar, um was es in den Clustern CL1 und CL3 geht und dass die Cluster CL11, CL17 und CL33 auch Informationen zum Thema “chemische Verbindungen” enthalten. Schauen wir uns nun noch die Bezeichner der formalen Begriffe von CL1 und CL3 im Detail an. Beide Cluster scheinen tatsächlich das Verfeinern von Öl anzusprechen, wobei es in Cluster 1 um Rohöl (crude oil) und in

Cluster 3 um Pflanzenöl (plant (resin, palm) oil) geht.

Die Bezeichner der formalen Begriffe helfen bei der Exploration der Clusterergebnisse. Würde die Präsentation der beschreibenden Terme mittels einer ungeordneten Menge erfolgen, wäre nicht nur die Information, die ein Cluster enthält, schwerer zu erfassen, sondern auch die Verbindung zu anderen Clustern, die verwandte Themen adressieren, ginge verloren.

Bezeichner der formalen Begriffe wie 'chemical compound' kommen normalerweise im Text selten oder gar nicht vor. Durch ihren allgemeinen Charakter und mit Hilfe des Begriffsverbandes lassen sich Dokumente – wie in unserem Beispiel die über Öle – leicht einer wesentlich allgemeineren Kategorie bzw. Thema zuordnen. Die Beziehungen der Cluster untereinander und die allgemeinen Bezeichner helfen während der Exploration der Ergebnisse und erleichtern so das Verständnis des gesamten Clusters. Da allgemeine Bezeichner nur selten oder gar nicht im Text vorkommen, müssen sie der Repräsentation hinzugefügt werden. Die Generalisierungstaxonomie von z.B. WordNet– aber auch jeder anderen Ontologie – kann an dieser Stelle ausgenutzt werden und liefert die benötigten allgemeinen Konzepte, die dann als Bezeichner im Verband auftauchen.

1.4 Gliederung der Arbeit

Die Arbeit ist in drei Teile gegliedert. Im ersten Teil werden die Grundlagen behandelt. Der zweite Teil stellt die entwickelten Ansätze und Methoden zur Integration von Hintergrundwissen und zur Präsentation der Clusterergebnisse vor. Der dritte Teil befasst sich mit der Anwendung der eingeführten Methoden anhand verschiedener Praxisfragestellungen. Diese Fragestellungen werden wir in Kapitel 2 entlang der zu Evaluierungszwecken genutzten Datensätze einführen und diskutieren.

1. Teil Der erste Teil beschäftigt sich mit den Grundlagen dieser Arbeit und führt die verwendeten Data-Mining-Verfahren ein. Aufbauend auf dem KDD-Prozess aus Kapitel 3 werden die für die Anwendungen wichtigen Vorverarbeitungsschritte sowie die Datenrepräsentationen in Kapitel 4 vorgestellt und genauer analysiert. Kapitel 5 führt die Begriffe Cluster und Clusterverfahren ein, gibt eine Übersicht über bestehende Clusterverfahren und geht im Detail auf die zwei zentral verwendeten Verfahren KMeans und Formale Begriffsanalyse ein. Die Einführung in den Bereich der Ontologien, deren formale Definition sowie Quellen zur Akquisition von Ontologien erfolgt in Kapitel 6.

2. Teil Der zweite Teil gliedert sich in drei große Abschnitte. Als erstes wird in Kapitel 7 die Methode des Subjektiven Clusterns anhand von Textdokumenten eingeführt und deren Güte mit Hilfe von statistischen Maßen evaluiert. Kapitel 8 stellt eine Methode zur Änderung der Repräsentation von Textdokumenten mittels Hintergrundwissen vor. Das Hintergrundwissen wird in Form von Ontologien in den Prozess integriert. Es werden verschiedene Wege untersucht und auch evaluiert, um die neuen Dokumentenrepräsentationen abzuleiten. Dabei wird die neue Dokumentenrepräsentation detailliert analysiert. Abschliessend wird in Kapitel 9 gezeigt, wie Hintergrundwissen zur explorativen Analyse und zur Beschreibung der gefunden Cluster eingesetzt werden kann.

3. Teil Der Anwendungsteil gliedert sich in zwei Teile. Das erste Kapitel dieses Teiles, Kapitel 10, wendet die Methode Subjektives Clustern auf die Kommunikationsdaten bei der Deutschen Telekom AG an, stellt die verwendete Ontologie vor und gibt die Ergebnisse wieder. Zudem führen wir eine Architektur für ein Wissensportal ein, die auf das Subjektive Clustern zur Informationsaufbereitung zurückgreift. In Kapitel 11 wird die Methode zum Clustern mit

Hintergrundwissen sowie die Visualisierung der Textcluster mit Hilfe der Formalen Begriffsanalyse für weitere Anwendungsdomänen aus den Bereichen eLearning, Landwirtschaft und Tourismus genutzt.

Kapitel 12 schließt die Arbeit mit einer Zusammenfassung und einem Ausblick ab. Die Arbeit enthält eine Reihe von verschiedenen Methoden zu denen man unterschiedliche verwandte Ansätze findet. Die verwandten Ansätze der einzelnen Methoden werden jeweils bei ihrer Einführung angegeben und sind daher am Ende der entsprechenden Kapitel zu finden.

2 Motivation aus der Anwendung

In diesem Kapitel gehen wir auf verschiedene Anwendungsgebiete für Clustern ein. Wir motivieren anhand unterschiedlicher Datensätze die Notwendigkeit für die Nutzung von Hintergrundwissen im Clusterprozess. Wir werden die Datensätze im Rest dieser Arbeit zur empirische Evaluierung nutzen. Weiterhin gehen wir auf die unterschiedlichen Charakteristika der Datensätze und die sich daraus ergebenden Anforderungen für das Clustern ein und beschreiben sie anhand statistischer Kennzahlen.

Die Datensätze lassen sich prinzipiell in zwei Gruppen einteilen. Auf der einen Seite handelt es sich um Textdokumente und auf der anderen Seite um Kommunikationsdaten der Deutschen Telekom AG. Die Datensätze haben eine Gemeinsamkeit. Sie spannen nach der Vorverarbeitung einen großen Merkmalsraum auf. Dieser Merkmalsraum erschwert nicht nur per se das Berechnen der Cluster, sondern ist wegen der puren Menge schwer verständlich. So beschränkte man sich bei den Telekomdaten auf die Ableitung von relativ kleinen Merkmalsräumen von 80 bis 100 Merkmalen (mehr zur Vorverarbeitung siehe Kapitel 4.3). Schon bei dieser Merkmalsmenge waren die Anwender nicht mehr in der Lage, die berechneten Cluster zu verstehen. Bei der Anwendung im Bereich des Textclusterns steigt die Anzahl der Merkmale schnell auf einige Tausend. Diese ohne Unterstützung zu interpretieren, erweist sich als sehr schwierig. Außerdem sind die Merkmale meistens aus dem Zusammenhang gerissen, so dass ihre Bedeutung nur im Kontext mit anderen Merkmalen klar wird. Mit Hilfe der entwickelten Methoden lassen sich die Merkmalsräume durch das akquirierte Anwenderwissen strukturieren und die Ergebnisse können angepasst an die Aufgabenstellung präsentiert werden.

Im Bereich der Textdokumente gliedern sich die insgesamt vier Datensätze entlang verschiedener Domänen. Beim Reuters-Korpus, der mit mehr als 21000 Dokumenten sehr umfangreich ist, handelt es sich um Nachrichtentexte. Sowohl beim eLearning-Korpus als auch beim Getess-Datensatz, die deutlich kleiner sind, handelt es sich um Web-Seiten, die entsprechende Kursmaterialien bzw. Tourismusbeschreibungen wiedergeben. Der FAO-Datensatz umfasst Beiträge zu Fachzeitschriften aus dem Bereich der Landwirtschaft.

Die manuelle Kategorisierung solcher Datensätze ist sehr aufwendig. Gerade die Startphase, in der keinerlei Einteilung vorhanden ist, gestaltet sich extrem schwierig. Ziel unserer Ansätze ist es, die erstmalige Erstellung einer Struktur mittels Clustern zu unterstützen. Dabei geht es nicht nur um den Prozess der Zuweisung der Kategorien zu den Dokumenten, sondern auch der entsprechenden Präsentation der Clusterergebnisse. Die Ontologie, die wir in diesen Prozess integrieren wollen, kann vorab manuell [214] oder semiautomatisch [153], und dabei völlig losgelöst von den Dokumenten, erarbeitet werden. Sie wird dann mit Hilfe unserer Methoden den automatischen Strukturierungsprozess leiten.

Ähnliches gilt auch für die Kommunikationsdaten bei der Telekom. Hier existiert sehr viel Wissen in den Köpfen der Mitarbeiter. Dieses in den Clusterprozess zu integrieren, erhöht die Güte und Verständlichkeit der Ergebnisse. Bei den Telekom Daten handelt es sich um die anonymisierte Aufzeichnung der Kommunikationsdaten von 10 % aller Telekom Kunden. Das entspricht ca. 130Gb Rohdaten pro Monat. Weiterhin stehen auch Kommunikations- und Befragungsdaten aus einem Kommunikationspanel zur Verfügung.

Anhand der Datensätze werden wir im Verlauf der Arbeit zeigen, dass die Integration von Hin-

tergrundwissen nicht nur in den ganz unterschiedlichen Domänen für mehr Verständlichkeit von Clusterergebnissen sorgt, sondern dass auch die Clustergüte steigt. Dabei wurden u.a. spezielle Anpassungen von Algorithmen im Bereich des Subjektiven Clusterns auf die unterschiedlichen Domänen, wie z.B. auf die Kommunikationsdaten, vorgenommen. Aber auch der Einsatz von Ontologien in Kombination mit Clusterverfahren in Wissensportalen [156, 114, 211] erleichtert die Strukturierung der Informationen erheblich. Im Folgenden werden wir die einzelnen Datensätze vorstellen. Bis auf den Reuters-Datensatz werden im Teil III der Arbeit die Ergebnisse auf der Basis dieser Datensätze mit Blick auf die Anwendung präsentiert.

Im Weiteren stellen wir die einzelnen Datensätze und Korpora detailliert vor.

2.1 Reuters Nachrichtentexte

Die Reuters-21578 Text Dokument Sammlung [149][1] besteht aus 21578 Nachrichtentexten, die 1987 bei der Nachrichtenagentur Reuters erschienen. Nachträglich wurden diese Texte von Gutachtern in 135 vorgegebene Kategorien eingeteilt bzw. indexiert und dann 1990 von Reuters der Forschergemeinschaft für Klassifikationsaufgaben zur Verfügung gestellt. Nachdem die erste Version des Reuters Korpus zu Problemen bei der Vergleichbarkeit der Ergebnisse führte, wurde 1996 die auch hier verwendete Reuters-21578 Version der Dokumentsammlung fertiggestellt. Die Dokumente sind seitdem mit SGML TAGS[2] versehen und Fehler in den Labels und den Texten wurden bereinigt.

Der Inhalt der Dokumente beschäftigt sich vorrangig mit Börsennachrichten. Dabei geht es u.a. um den Kauf und Verkauf von Unternehmen bzw. Aktien. Der Handel mit Rohstoffen wie z.B. Zucker oder Weizen, aber auch Vorhersagen auf dem Geldmarkt sind Themen. Zwei Beispieltexte sind im Anhang D zu finden.

Jedes Dokument im Datensatz kann einer oder mehreren vorgegebenen Kategorien angehören. Der Reuters-Datensatz bietet sich durch die vorhandenen Kategorien zur Evaluierung für Text-Klassifikations- und -Clusteraufgaben an. Er wird aus diesem Grund häufig in der Literatur referenziert, so dass Berechnungen auch mit den Ergebnissen anderer Autoren verglichen werden können. Wir werden einen großen Teil unserer Ergebnisse auf der Basis dieses Datensatzes vorstellen. Im Folgenden wird genauer auf den Aufbau des Reutersdatensatzes eingegangen und Besonderheiten werden beschrieben.

2.1.1 Details des Reuters-Korpus

Der Reuters-Korpus besteht aus insgesamt 21 SGML-Files. Jedes SGML-File enthält 1000 Artikel. Die Artikel starten mit der Zeile:

```
<REUTERS TOPICS=?? LEWISSPLIT=?? CGISPLIT=?? OLDID=?? NEWID=??>
```

und endet mit

```
</REUTERS>.
```

Der <REUTER> Tag wird durch fünf Attribute weiter beschrieben, wobei nur zwei an dieser Stelle genauer erläutert werden.

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/
[2] http://www.w3.org/MarkUp/SGML/

Das "TOPICS" Attribut beschreibt den Begutachtungsstatus des Dokumentes. Enthält das Attribut den Wert "YES", zeigt dies an, dass das Dokument von einem Gutachter gelesen und klassifiziert wurde, auch wenn im <TOPICS> Tag kein Eintrag vorhanden ist (nähere Erläuterungen zum <TOPICS> Tag siehe unten). Es gibt vier weitere Mengen von Kategorien, unabhängig von Topics, nämlich "Exchanges", "Org", "People" und "Places". Diese wurden für die Forschung aber bisher kaum verwendet und sind auch nicht Gegenstand dieser Arbeit. Die Größe der Dokumente schwankt zwischen 46 Bytes und ca. 6 kb.

Für den Reutersdatensatz existieren verschiedene Teilmengen, die im Laufe der Jahre für unterschiedliche Analysen verwendet wurden. Durch das Attribut "LEWISSPLIT" wird eine solche Teilmenge erzeugt. Bei Klassifikationsaufgaben ist eine typische Vorgehensweise, den vorhandenen Datensatz in Trainings- und Testmenge zu splitten. Das Modell wird dann auf der Trainingsmenge berechnet und die Güte auf der Testmenge geprüft. Um die Vergleichbarkeit für Arbeiten auf diesem Datensatz zu erhöhen, berechnet nicht jeder Forscher eigenständig einen neuen Splitt, sondern nutzt die durch das Attribut LEWISSPLIT gegebene Zerlegung in Trainings- und Testmenge. Die Ergebnisse unterschiedlicher Algorithmen und Verfahren sind damit vergleichbar. Da für Clusterverfahren keine Klassenzugehörigkeit der Dokumente benötigt wird, kann der gesamte mit Labels versehene Datensatz zur Modellbildung verwendet werden. Trotzdem nutzen einige Forscher z.B. nur den TEST Teil des Datensatzes, um Clustermodelle zu evaluieren. Insgesamt gehören 6188 Dokumente zur Test- und 13625 zur Trainingsmenge. Weiterhin wurden von den 135 vorgegebenen Topic-Labels[3] nur 120 mindestens einem Dokument im Datensatz zugewiesen, 15 wurden bisher nicht genutzt.

Für die Arbeit mit dem Reutersdatensatz wurden die Tag's <TOPIC> und <BODY> verwendet. Das <TITLE> Tag wurde in dieser Arbeit nicht in die Auswertungen einbezogen. Die beiden Tag's <TOPICS> und </TOPICS> umschliessen die von den Gutachtern vergebenen Kategorien. Die Anzahl kann zwischen keinem und beliebig vielen Kategorien variieren. Ist keine Kategorie vorhanden und das Attribut TOPICS hat den Wert "YES", so passt keine der vorgegebenen Kategorien auf dieses Dokument. Ist hingegen der Wert "NO" oder "BYPASS" und keine Kategorie angegeben, hat auch kein Gutachter diese Dokumente begutachtet. Beachtet man für die Trainings- und Testmengen das Attribut TOPICS und nutzt nur die kategorisierten Dokumente, so ergeben sich 9603 Dokumente für die Trainings- und 3299 für Testmenge, also 12902 Dokumente, die mit Kategorien versehen sind. Wir haben für diese Arbeit alle Dokumente, bei denen TOPICS auf "YES" steht, aber kein Label existiert, in der Klasse "defnoclass" zusammengefasst.

Die Tags <BODY> und </BODY> umschließen den ursprünglich veröffentlichten Text bereinigt[4] um unverständliche Sonderzeichen. Leider enthalten nicht alle Artikel ein solches Tag. Diese Dokumente wurden ebenfalls ignoriert, wodurch sich die Anzahl an Dokumenten von 12902 auf 12344 reduziert. In der Testmenge verbleiben 3009 Dokumente.

Für das partitionierende Clustern mit nichtüberlappenden Partitionen darf ein Dokument immer nur ein Label haben. Daher wurde von allen Dokumenten immer nur das *erste Label* verwendet. Die Anzahl der Topics reduziert sich damit von 120 auf 82. In der Testmenge sind noch 63 Topics enthalten.

Den Datensatz mit allen 12344 Dokumenten in 82 Klassen nennen wir PRC[5]. Dem PRC unterliegt die in Abbildung 2.1 dargestellte Verteilung der Dokumente über die vorhandenen Kategorien. Die Verteilung entspricht einer typischen Zip Verteilung [234, 147]. Man erkennt in der Abbildung leicht

[3] Im Reuters Datensatz werden die Kategorien mit "Topic" bezeichnet. Um die Beschreibung allgemeiner zu halten, bezeichnen wir diese im Folgenden als Kategorien.

[4] Die Sonderzeichen stehen in einem gesonderten Tag zur Verfügung.

[5] PRC heißt Preprocessed Reuters-21578 Corpus, da wir den PRC nur durch die Anwendung einer Reihe von "Vorverarbeitungsschritten" zur Extraktion der relevanten Dokumente erhalten haben.

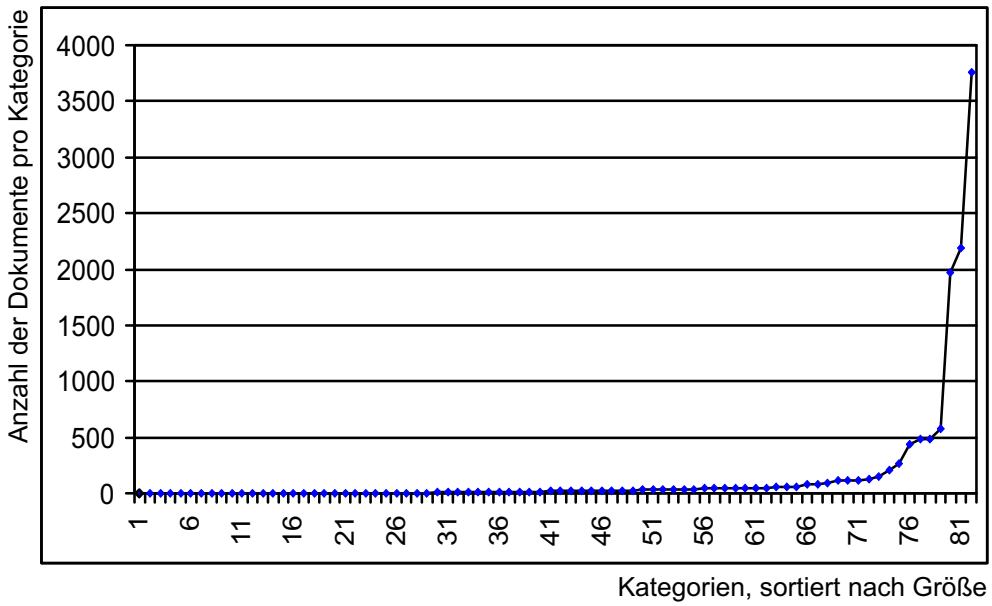

Abbildung 2.1: Häufigkeitsverteilung der Dokumente über die Reuterskategorien des ersten Labels

die große Anzahl an Kategorien, die nur sehr wenige Dokumente enthalten. Nur rund 1 % aller Dokumente liegen in 35 Kategorien, wobei ca. 85 % aller Dokumente in 10 Kategorien enthalten sind bzw. die größte Kategorie umfasst 3760 Dokumente. Zwei Probleme wurden im Rahmen der Arbeit identifiziert:

1. Auf der einen Seite gibt es Klassen mit sehr vielen Dokumenten, d.h. die meisten Dokumente gehören ein paar wenigen Klassen an. Das Purity-Maß (siehe Kapitel 5.3.3.2) liefert in einer solchen Situation für ungefähr gleich große Cluster (wie sie Bi-Sec-KMeans (vgl. [206]) liefert) immer sehr gute Ergebnisse. Die Fehlklassifikation einiger weniger Dokumente spielt bei diesem Ergebnis daher keine wesentliche Rolle. Selbst bei einem verbesserten Ergebnis wäre dies durch die guten Vorabergebnisse mit dem Purity-Maß kaum zu beobachten.

2. Auf der anderen Seite gibt es Kategorien mit sehr wenigen Dokumenten, die zum Teil sogar nur ein Dokument enthalten. Clusterverfahren wie KMeans oder Bi-Sec-KMeans haben mit dem Erkennen solcher Ausreißer Schwierigkeiten (vgl.[64]). Häufig findet man im Ergebnis der Clusterung die Dokumente der Kategorien mit wenigen Dokumenten - zusammen - in einem Cluster mit vielen Dokumenten einer großen Kategorie.

Um die Probleme genauer zu untersuchen haben wir systematisch Teilmengen an Dokumenten aus dem PRC-Datensatz entnommen. Diese Teildatensätze, die wir später zur Evaluierung herangezogen haben, werden im folgenden Abschnitt eingeführt. Sie enthalten z.B. (i) keine Kategorien mit sehr wenigen Dokumenten oder (ii) die maximal erlaubte Anzahl an Dokumenten wird beschränkt.

2.1.2 Reuters-Teildatensätze

Das folgende Kapitel beschreibt die aus dem Reuters-21578 Datensatz abgeleiteten Teildatensätze. Die ausgewählten Datensätze bilden auf der einen Seite Datensätze mit sehr wenigen Dokumenten pro Kategorie und auf der anderen Seite Datensätze mit ungefähr gleicher Anzahl an Dokumenten pro Kategorie. Sollte eine Kategorie über mehr als die gewünschte Anzahl an Dokumenten verfügen, so wählen wir zufällig die entsprechende Anzahl an Dokumenten aus. Enthält eine Kategorie nicht

die minimale Anzahl an Dokumenten, wird sie aus dem Datensatz ausgeschlossen und die Anzahl der Kategorien reduziert sich entsprechend.

PRC-max20 enthält nur Kategorien mit sehr wenigen Dokumenten (maximal 20). Anhand dieses Datensatzes mit 1035 Dokumenten kann man Experimente mit sehr wenigen Dokumenten pro Kategorie durchführen und das Verhalten der Verfahren untersuchen. Er enthält 82 Kategorien mit durchschnittlich 12.62 Dokumenten pro Kategorie (Standardabweichung: 8.18).

PRC-min15-max20 ist ein sehr homogener Korpus. Alle Kategorien enthalten fast die gleiche Menge an Dokumenten. Insgesamt umfasst der Datensatz 899 Dokumente. Minimal sind 15 und maximal 20 Dokumente in 46 Kategorien[6] enthalten, wobei der Durchschnitt bei 19.54 liegt (Standardabweichung: 1.15).

PRC-max100 besteht aus 82 Kategorien, deren 2755 Dokumente weniger gleichmäßig über die Kategorien verteilt sind. Die Anzahl pro Kategorie ist auf maximal 100 beschränkt. Gleichzeitig wird der Datensatz nicht mit zu vielen Dokumenten der extrem großen Kategorien geflutet. Die durchschnittliche Anzahl an Dokumenten pro Kategorie beträgt 33.59 (Standardabweichung: 36.28). Abbildung 2.2 zeigt analog zu Abbildung 2.1 die Verteilung der Dokumente über den Kategorien. Sie ist wesentlich homogener.

PRC-min15-max100 ist dem Korpus PRC-max100 sehr ähnlich, aber die Kategorien mit den sehr wenigen Dokumenten wurden durch die untere Schranke mit der minimalen Anzahl von 15 ausgeschlossen. Dadurch reduziert sich die Anzahl der Kategorien auf 46 und der Dokumente auf 2619 mit einem Durchschnittswert von 56.93 Dokumenten (Standardabweichung: 33.12).

PRC-min15 ist mit einer Dokumentanzahl von 12208 dem Gesamtkorpus PRC am ähnlichsten. Er besteht wiederum aus nur 46 Kategorien, wobei auch hier die Ausreißerkategorien durch die untere Schranke ausgeschlossen wurden. Im Durchschnitt enthält jede Kategorie 672.7 Dokumente (Standardabweichung 265.39).

PRC Der Vollständigkeit halber sei an dieser Stelle noch einmal der gesamte Korpus PRC erwähnt, der 12344 Dokumente umfasst. Die durchschnittliche Anzahl der Dokumente pro Kategorie beträgt 150.54 (Standardabweichung 520.3, siehe Abbildung 2.1).

Die beiden folgenden Datensätze stellen spezielle Teilmengen aus allen Dokumenten des Reuters Korpus dar. In der Literatur sind sie als Varianten zu finden (vgl. z.B. [182, 42]).

PRC-testonly besteht nur aus den Dokumenten, die in der LEWIS Aufteilung mit “TEST” gekennzeichnet sind. Der Datensatz beinhaltet 3009 Dokumenten mit 63 Kategorien bei einer durchschnittlichen Anzahl von 153.7 Dokumenten pro Kategorie (Standardabweichung: 47.8).

PRC-single8654 enthält 8654 Dokumente. Diese wurden genau einer Kategorie zugewiesen.

Die ersten sechs Datensätze werden in dieser Arbeit zu Evaluierungszwecken verwendet. Dabei wird auch untersucht, in wie weit sich die unterschiedliche Verteilung der Dokumente über die Kategorien auf die Güte der Clusterergebnisse mit und ohne Hintergrundwissen auswirkt.

[6] Die restlichen 36 Kategorien enthalten weniger als 15 Dokumente pro Kategorie und können aus diesem Grund nicht mehr berücksichtigt werden.

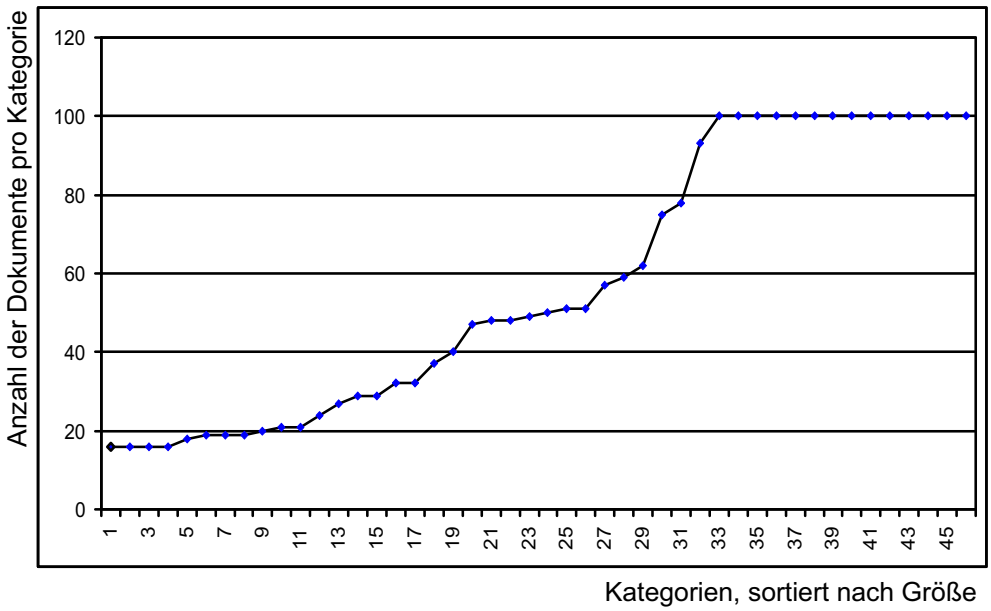

Abbildung 2.2: Verteilung der Dokumente auf die Kategorien des Datensatzes PRC-min15-max100

2.2 Java-eLearning-Datensatz

Der Java-eLearning-Datensatz ist ein relativ kleiner Datensatz, bestehend aus Web-Seiten eines eLearning-Kurses über die Programmiersprache Java (siehe [96]). Es handelt sich hier um einen über das Internet abrufbaren Kurs zum Erlernen von Java. Der Kurs besteht aus 224 Web-Seiten verteilt auf 36 Kategorien, die beim Erstellen des Kurses themenspezifisch angelegt wurden. Da in vielen Kategorien nur ein Dokument vorhanden ist, sind wir [96] gefolgt und haben nur die Kategorien mit mehr als 10 Dokumenten betrachtet. Damit verbleiben 94 Dokumente, die 2013 verschiedene Wortstämme und insgesamt 20394 Worte, die über acht Kategorien verteilt sind, enthalten. Die minimale Anzahl an Dokumenten pro Kategorie beträgt 10 und die maximale 19. Die Größe der Dokumente schwankt zwischen 495 Bytes und 35 kB. Die zu diesem Datensatz für die Experimente verwendete Ontologie wird in Kapitel 6.3.2.3 beschrieben.

Das Tutorial führt u.a. in Themen wie die objektorientierte Programmierung ein. Neben der detaillierten Erläuterung der Befehle von Java anhand von Beispielen werden auch technische Details vorgestellt. Die Themen der acht Kategorien sind:

- Applets
- Access to Applets in HTML
- Arrays
- Classes
- Control Structures
- JDK
- Operators
- Strings

Mit Hilfe unserer Methoden sind wir in der Lage, dieses Kursmaterial automatisch zu strukturieren und entsprechende Ergebnisse zu visualisieren. Wir werden zeigen, dass wir mit Hilfe des Hintergrundwissens die ursprünglich vorgegebene Einteilung besser wiederentdecken können.

2.3 Landwirtschaftliche Texte der FAO

Die “Food and Agriculture Organisation” kurz FAO[7] ist eine Teilorganisation der Vereinten Nationen. Die FAO soll durch die Förderung der landwirtschaftlichen Entwicklung gegen Hunger und Armut arbeiten. In ihrer strategischen Ausrichtung hat die FAO die Vision, ein so genanntes “Center of Excellence” zu sein. Dazu sammelt, analysiert, interpretiert und verteilt sie Informationen, die in Beziehung zu den Themen Ernährung, Lebensmittel, Landwirtschaft, Forstwirtschaft und Fischerei stehen. Um diese Aufgabe erfüllen zu können, unterhält die FAO primär zwei große Informations Managment Systeme, nämlich das Electronic Information Management System (EIMS) und das FAO Document Online Catalog (FAODOC). Ein großer Teil der elektronisch verfügbaren Dokumente[8] sind im “FAO Corporate Document Repository” (DocRep)[9] abgelegt, der ein Teil des EIMS ist und können über den FAO Information Finder[10] zugegriffen werden. Der größte Teil der FAODOC Dokumente ist nicht in elektronischer Form verfügbar. Ein Katalog dazu kann aber online angefragt werden. Wir wollen an dieser Stelle nicht genauer auf den Aufbau und Zugriff der Informationssysteme der FAO eingehen. Auf alle Informationen der FAO kann man über das WWW in fünf Sprachen zugreifen. In Kooperation mit der FAO in Rom war es möglich, einen Korpus an landwirtschaftlichen Texten und deren Kategorien im WWW zu sammeln. Die Kategorien stammen hauptsächlich aus dem vor Ort entwickelten Thesaurus AGROVOC, der genauer in Kapitel 6.3.2.1 vorgestellt wird. Im Folgenden werden wir den Korpus beschreiben. Weitere Informationen zum Korpus findet man in [145] und [146].

Alle Dokumente werden von einer Gruppen von Leuten vor Ort mit Metadaten versehen. Dabei wird eine erweiterte Version des bekannten Dublin Core Standards[11] verwendet (siehe AgMES-Projekt[12]). Die Dokumente werden unter Verwendung des kontrollierten Vokabulars aus dem AGROVOC-Thesaurus katalogisiert. Es kann aus 16607 Schlagworten ausgewählt werden. Während des Katalogisierens wird jedem Dokument eine Menge von Schlagworten (kurz Desc) aus dem kontrollierten Vokabular zugewiesen, wobei maximal sechs primäre und beliebig viele sekundäre Schlagworte pro Dokument vergeben werden können. Im Sinne dieser Arbeit kann man jedes Schlagwort ganz allgemein als eine Kategorie betrachten, wobei jedes Dokument mindestens einer dieser Kategorien zugeordnet wurde.

Jedes Dokument unseres Datensatzes enthält nicht nur Schlagworte aus dem AGROVOC, sondern gehört auch noch maximal drei von AGROVOC unabhängigen Kategorien (kurz Cat) an. Diese Kategorien werden aus einer Menge von insgesamt 115 ausgewählt. Die komplette Liste findet man in [146]. Alle Daten sind in den drei Sprachen Englisch, Französisch und Spanisch abgelegt. Wir fassen die Schlagworte des AGROVOC und die unabhängigen Kategorien in den folgenden Tabellen unter dem Begriff Label zusammen.

Der Aufwand der manuellen Vergabe von Kategorien ist extrem groß. Zusätzlich ist die Einteilung der Dokumente historisch gewachsen. Hier kann das Clustern mit Hintergrundwissen ansetzen, um neue Strukturen passend zum aktuellen Thesaurus zu berechnen. Ein weiterer Vorteil des ontologiebasierten Clusterns liegt in der Möglichkeit auch mehrsprachige Ressourcen verarbeiten zu können. Möglich wird dies durch die Übersetzung der Worte in die sprachunabhängige ontologiebasierte

[7] http://www.fao.org/

[8] Die FAO erfasst nicht nur Dokumente sondern so genannte Ressourcen. Diese umfassen Bücher, Zeitungen, Zeitschriften, Artikel, Web-Seiten, Fotos, Pressemeldungen, Veröffentlichungen (gedruckte nicht veränderbare Ressourcen). Wir konzentrieren uns für unseren Datensatz auf elektronisch verfügbare Texte im Web und fassen diese unter dem Begriff Dokumente zusammen.

[9] http://www4.fao.org/faobib/index.html

[10] http://www.fao.org/waicent/search/default.asp

[11] http://dublincore.org/

[12] http://www.fao.org/agris/agMES/default.htm

Tabelle 2.1: Dokumentverteilung aller FAO-Dokumente auf Labels (Schlagworte oder Kategorien), sowie die Anzahl der Labels pro Dokument

Statistik des Datensatz AG_{raw}		**Sprache**					
		Englisch (en)		**Französisch (fr)**		**Spanisch (es)**	
		Desc	**Cat**	**Desc**	**Cat**	**Desc**	**Cat**
Total	# Dokumente	1708	1879	481	897	519	769
	# Label	1185	115	503	86	511	93
	# zugew. Label	5072	3328	1494	1620	1574	1434
Label-Ebene	Max (#Dok/Label)	96	315	67	214	71	179
	Min (#Dok/Label)	1	1	1	1	1	1
	Avg (#Dok/Label)	1,44	16,34	0,95	10,43	1,02	8,27
Dok.-Ebene	Max (#zugew. Label/Dok)	8	3	7	4	7	7
	Min (#zugew. Label/Dok)	1	1	1	1	1	1
	Avg (#zugew. Label/Dok)	2,97	1,77	3,11	1,81	3,03	1,86

Repräsentation.

Wir können sowohl die Schlagworte (die wir im Folgenden mit Desc abkürzen) als auch die Kategorien (Cat) der Dokumente zum Zusammenstellen von Datensätzen nutzen. Wir beschränken uns bei den Schlagworten auf die primären. Die von der FAO zur Verfügung gestellten Informationen erlaubten uns, die in Tabelle 2.1 wiedergegebene Anzahl an Dokumenten in der jeweiligen Sprache aus dem WWW herunterzuladen. Leider existierten nicht alle Dokumente in jeder der fünf Sprachen und nicht jedes Dokument ist mit Schlagworten und Kategorien versehen. Wegen technischer Probleme konnten auch nicht alle Dokumente, die uns von der FAO benannt wurden, heruntergeladen werden. Dies lag vor allem an dem Informationssystem der FAO.

In Tabelle 2.1 gibt die Spalte “Desc” in der Zeile “Total #Label” die Anzahl der insgesamt verwendeten unterschiedlichen Schlagworte des AGROVOC bzw. bei “Cat” die Anzahl der unterschiedlichen (unabhängigen) Kategorien wieder. Der Zeile “Label-Ebene” entnimmt man, wie sich die Dokumente über die Kategorien verteilen. Die geringen Durchschnittswerte zeigen an, dass es schwierig ist, eine größere Menge an Dokumenten für ein Label (Schlagwort oder Kategorie) zu beschaffen. Wir mussten gerade aus diesem Grund beim Erstellen der Datensätze für das Clustern viele der Label ausschließen. In der letzten Zeile mit dem Namen “Dokumenten-Ebene” erhält man einen Überblick über die Verteilung der Label pro Dokument. Man erkennt die deutlich höhere durchschnittliche Anzahl an Label bei den AGROVOC-Schlagworten.

Die größte Menge an Dokumenten steht mit 1708 in englischer Sprache zur Verfügung — für Französisch (481) und Spanisch (519) gibt es nur rund ein Drittel. Die heruntergeladenen Dokumente unterscheiden sich in Länge und Stil stark voneinander. Die Größe schwankt zwischen 1.5kb und 600kb und bedeutet damit für automatische Verfahren eine sehr große Herausforderung. Dies stellt auch einen substantiellen Unterschied zu den Dokumenten des Reuters- und eLearning-Korpus (siehe Kapitel 2.1 und Kapitel 2.2) dar. Außerdem ist die Anzahl der möglichen Kategorien des Reuters-Korpus mit 135 gegenüber der Anzahl der AGROVOC-Schlagworte deutlich geringer. Zwar verwendeten die Katalogisierer in dem in dieser Arbeit zu Grunde liegenden Datensatz nicht alle AGROVOC-Schlagworte, mit 1185 verschiedenen ist die Anzahl aber immer noch deutlich höher als beim Reuters-Datensatz.

Für einen ersten Test beschränken wir uns wie beim Reuters-Datensatz auch für den AGROVOC-Datensatz auf die erste Kategorie. Tabelle 2.2 fasst die Eigenschaften der verbliebenen Dokumente zusammen. Wir bezeichnen (i) den englischen Datensatz basierend auf den AGROVOC-

Schlagworten mit AGeD, (ii) den auf den Kategorien basierenden mit AGeC, und (iii) den französischen Datensatz basierend auf den AGROVOC-Schlagworten mit AGfD usw.

Tabelle 2.2: Dokumentverteilung der FAO Dokumente auf Labels (Schlagworte oder Kategorien) mit mindestens 50 Dokumenten, wobei nur das erste Label berücksichtigt wurde

Statistik des Datensatz AG_{single}		**Language**					
		English (en)		**French (fr)**		**Spanish (es)**	
		Desc	**Cat**	**Desc**	**Cat**	**Desc**	**Cat**
Total	# Dokumente	374	1016	117	612	188	563
	# Label	6	7	3	7	6	7
	# zugew. Label	374	1016	117	612	188	563
Label-Ebene	Max (#Dok/Label)	86	271	55	171	56	158
	Min (#Dok/Label)	51	102	30	50	21	50
	Avg (#Dok/Label)	62,33	145,14	39	87,43	31,33	80,43
Dok.-Ebene	Max (#zugew. Label/Dok)	1	1	1	1	1	1
	Min (#zugew. Label/Dok)	1	1	1	1	1	1
	Avg (#zugew. Label/Dok)	1	1	1	1	1	1

Bei der Analyse der Kategoriennamen[13] in Tabelle 2.3 fällt die geringe Überlappung der AGROVOC-Schlagworte zwischen den einzelnen Sprachen auf. Leider konnte uns die FAO in Rom im Rahmen dieser Arbeit nicht mit der gleichen Anzahl an Dokumenten per Kategorie und Sprache ausstatten.

Tabelle 2.3: Namen der in Tabelle 2.2 verwendeten FAO-Schlagworte oder FAO-Kategorien

Sprache					
Englisch (en)		**Französisch (fr)**		**Spanisch (es)**	
Desc	**Cat**	**Desc**	**Cat**	**Desc**	**Cat**
EXTENSION ACTIVITIES	E14	FOREST MANAGEMENT	E10	FOREST MANAGEMENT	E10
FOREST MANAGEMENT	E50	FORESTRY	E14	FOREST RESOURCES	E14
FOREST RESOURCES	E70	FORESTRY DEVELOPMENT	E50	FORESTRY	E50
FORESTRY DEVELOPMENT	K01		M11	FORESTRY DEVELOPMENT	E71
SUSTAINABILITY	K10		K01	FORESTRY POLICIES	K01
TRIFOLIUM REPENS	M11		K10	NONWOOD FOREST PRODUCTS	K10
	P01		P01		P01

2.4 Der Getess-Tourismus-Korpus

Das Getess-Projekt[14] beschäftigte sich mit dem Bau eines neuartigen Informationssystems. Es kombiniert Techniken aus dem Bereich der natürlichsprachlichen Anfragebearbeitung mit modernen Dialogsystemen und Datenbanken. Mit Hilfe von Ontologien wird die Domäne des Informationssystems beschrieben. Sie bilden auch die Brücke zwischen den Anfragen für die Datenbank und den natürlichsprachlichen Anfragen des Benutzers. Dazu nimmt das System die natürlichsprachlichen Fragen des Benutzers entgegen, verarbeitet sie mit Techniken aus dem Bereich NLP (vgl. Kapitel 3.1.3) gemäß einer gegeben Domänenontologie und erzeugt daraus eine Datenbankanfrage.

[13] Die Bezeichner der Kategorien in den einzelnen Spalten sind unabhängig von allen Nachbarspalten.

[14] Die Web-Seite des Projektes lautet: http://www.getess.de/, unter der auch ein Prototyp eines neuen Informationssystems zur Verfügung steht.

Das System kann durch den Austausch der Domänenontologie leicht an neue Anwendungsgebiete angepasst werden.

Eine der aufbereiteten Domänen des Informationssystems ist der Tourismusbereich (siehe [137]). Im Projekt wurden die Webseiten dem Web-Portals “All-In-All” [15], einem Anbieter für Tourismusinformationen in Mecklenburg-Vorpommern, entnommen. So sammelte man mit einem Web-Crawler 2234 HTML-Dokumente mit insgesamt über 16 Millionen Worten vom Anbieter ein. Die Dokumente beschreiben Orte, Unterkünfte, Ausstattungen von Unterkünften, administrative Informationen oder kulturelle Ereignisse. Diese Informationen werden normalerweise Touristen zur Präsentation der Region Mecklenburg-Vorpommern zur Verfügung gestellt. Das Informationssystem des Getees-Projektes nutzt diese Web-Seiten zur Beantwortung von Anfragen.

Die Web-Seiten werden wir in dieser Arbeit als Datensatz unter dem Namen Getess-Datensatz nutzen. Weiterhin wurde im Verlaufe des Projektes eine umfangreiche Ontologie für den Tourismusbereich entwickelt. Dieses schon modellierte domänenspezifische Hintergrundwissen bietet eine ideale Grundlage für das Clustern mit Hintergrundwissen. Leider existiert bisher für diesen Korpus keine manuelle Einteilung der Dokumente in Kategorien. Die berechneten Cluster können daher nicht apriori überprüft werden. Es bietet sich aber an, die Clusterergebnisse mit Hilfe der in den Kapiteln 7 und 9 entwickelten Methoden zu präsentieren und zu visualisieren. Wir werden in der Arbeit für diesen Datensatz erste Ergebnisse zeigen.

Setzt man unsere Methoden wieder im Portal ein, könnten mit Hilfe von Benutzerpräferenzen individuelle Sichten auf die im Portal verfügbaren Daten erzeugt werden bzw. Benutzer können sich berechnete Sichten auswählen (Subjektives Clustern in Kapitel 7). Die Ontologie zusammen mit der Benutzerpräferenz steuert dann die Informationsbereitstellung. Alternativ könnte eine Informationsvisualisierung durch die Begriffsverbände erfolgen. Die visualisierten Informationen lassen ein einfaches Browsen in den Webdokumenten zu.

2.5 Telekomdatensatz

Die Deutsche Telekom AG zeichnet zu Analysezwecken die Kommunikation ihrer Kunden in anonymisierter Form auf. Zum einen erlauben die Analysen ein besseres Verständnis der Kunden. Zum anderen werden sie zur Dimensionierung der Netzkapazitäten eingesetzt. Durch den steigenden Wettbewerb ist die Deutsche Telekom AG gezwungen, neue und attraktive Tarife für große Kundengruppen anzubieten. Dazu ist es notwendig, diese Gruppen zu identifizieren. Wir stellen im Folgenden den Panel-Datensatz und die uns zur Verfügung stehende 10 % Stichprobe vor. Sie bilden u.a. die Grundlage der Analysen bei der Telekom. Wir werden die entwickelte Methode des Subjektiven Clusterns in dieser Arbeit in einer erweiterten Form auf die Daten der 10 % Stichprobe anwenden und erste, leicht verständliche, Ergebnisse präsentieren. Nachfolgend stellen wir die beiden Datensätze vor.

2.5.1 Panel-Datensatz

Das bei der Deutschen Telekom AG vorhandene Telekommunikationspanel “PAS” beschreibt und speichert das Kommunikationsverhalten der Kunden, um es besser erforschen zu können. Mit Einverständnis von ca. 5000 privaten Haushalten und ca. 6000 Arbeitsstätten werden deren Telefonanschlüsse überwacht und das Verhalten protokolliert, um wichtige Informationen über die Anzahl der aufkommenden Verbindungen, die Dauer der Verbindungen, die Art der Verbindungen (zum Bei-

[15] http://www.all-in-all.de

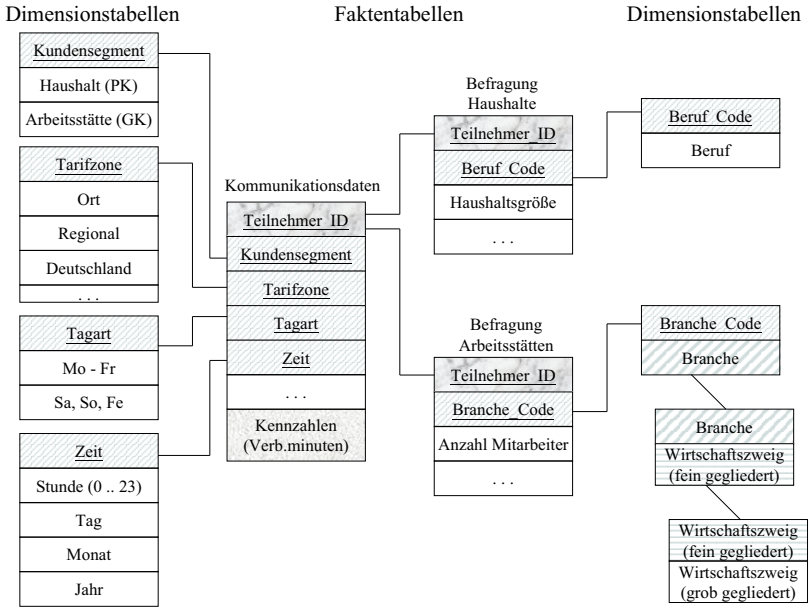

Abbildung 2.3: Auszug aus dem “PAS”-Sternschema

spiel Orts-, Regional-, Deutschland-, Auslandsgespräch oder Internetverbindung, etc.), der Wochentag, die Tageszeit (aufgeteilt in Stundenfenster) sowie einige weitere Kennzahlen in anonymisierter Form zu speichern. Zusätzlich werden von zwei renommierten Marktforschungsinstituten regelmäßig Befragungen bei den Haushalten und Arbeitsstätten durchgeführt, um mit Hilfe der erhobenen soziodemographischen Merkmale Kundengruppen besser beschreiben zu können. Damit hofft man, gerade in der Zeit des äußerst dynamischen Wettbewerbs ein Instrumentarium zu besitzen, mit dem gezieltere Marketingaktionen und eine innovative Preisgestaltung möglich werden. Dieses komplexe Datenmaterial bietet eine gute Grundlage, um das Hintergrundwissen für die späteren Analysen zu akquirieren. Wir stellen die Struktur des “PAS” im Folgenden vor.

Das “PAS” ist in einer relationalen Datenbank abgelegt, auf dem ein OLAP (On-Line Analytical Processing) Tool aufsetzt (vgl. [35]). Sämtliche Daten (Kommunikations- und Befragungsdaten) werden in der Datenbank abgelegt. Um auf die Daten mittels OLAP Tool zugreifen zu können, müssen diese einem konzeptuellen mehrdimensionalen Datenmodell entsprechen. Dieses Modell, auch Sternschema genannt, speichert in der Datenbank zwei Arten von Tabellen, die Fakten- und Dimensionstabellen. Die durch das mehrdimensionale Datenmodell beschriebenen Verknüpfungen werden dann im OLAP Tool modelliert. Abbildung 2.3 zeigt den wichtigsten Ausschnitt des “PAS”-Sternschema. Dabei sieht man die Faktentabellen “Befragung Haushalte”, “Befragung Arbeitsstätten” und “Kommunikationsdaten” (die Struktur der Kommunikationsdaten ist denen der 10 % Stichprobe, die wir später auch für die Analysen einsetzen werden, sehr ähnlich) sowie Teile der über 60 Dimensionstabellen. Die teilweise vorhandene Redundanz von beschreibenden Merkmalen auch in den Kommunikationsdaten wurde zur Performanzsteigerung bewusst in Kauf genommen. Damit wurde auch die Speicherung in der 3. Normalform ausgeschlossen. Die Modellierung dieses Sternschemas ist der Modellierung einer Ontologie sehr ähnlich. Daher können viele Zusammenhänge relativ einfach in die Ontologie übernommen werden. So bekommen wir eine gute Arbeitsgrundlage für die Ontologieakquisition in Kapitel 10.

2.5.2 Zehn Prozent Stichprobe

Bei der 10 % Stichprobe handelt es sich um Kommunikationsdaten, die, wie der Name schon sagt, 10 % aller Kunden der Deutschen Telekom AG enthalten. Bei der Ziehung der Stichprobe wurde darauf geachtet, dass diese repräsentativ gezogen wurde. Zum Zeitpunkt der Ziehung wurde auch festgelegt, für welche Kunden die Daten zu sammeln sind. Jedes Gespräch eines solchen Kunden wird in eine bestimmten Anzahl von Datensätzen zerlegt und in einer Datenbank abgespeichert. Die vorverarbeiteten Datensätze stehen aber auch als ASCII Datei zur Verfügung. Die Zerlegung der Gespräche ist notwendig, um eine spätere Auswertung der Daten zu erleichtern. Die Kunden sind unterteilt nach Privat- und Geschäftskunden, wobei sich unter den Privatkunden auch kleinere Geschäftskunden, die z.B. zu Hause arbeiten, befinden. Die gesammelten Kommunikationsdaten enthalten unter anderem Informationen über den Zeitpunkt, die Dauer, die Tarifeinheiten, die Tarifzone, die Tagart, die Tarifart, die Stunde des Beginns, das Quell- und Zielortsnetz und eine eindeutige anonymisierte TeilnehmerID. Mit Hilfe dieser Daten lässt sich sowohl das Kommunikationsverhalten jedes einzelnen Kunden als auch das Kommunikationsverhalten auf unterschiedlichstem Aggregationsniveau analysieren. Diese Daten bilden auch die Grundlage für weitere Analysen. Leider stehen über die Kunden keine beschreibenden Informationen wie beim Panel zur Verfügung. Mit Hilfe von Befragungsdaten aus der Marktforschung könnte man die fehlenden Informationen ergänzen, falls diese Rechnungsinformationen enthalten.

Mit den zur Verfügung stehenden Datensätzen sind wir in der Lage, anwendungorientiert unsere Methoden zu evaluieren. Der nächste Teil der Arbeit behandelt die Grundlagen. Im folgenden Kapitel werden wir uns mit dem KDD-Prozessmodell beschäftigen sowie die Begriffe Data Mining und Text Mining in Beziehung zum Knowledge Discovery in Databases setzen.

Teil I

Grundlagen

3 Wissensentdeckungsprozess

In der heutigen Zeit steigt die automatisch gesammelte Menge an Daten kontinuierlich. Die in den Daten versteckten Zusammenhänge und die so verborgenen Informationen möchte man mit Hilfe von Knowledge Discovery Methoden erschließen. Komplexe Vorverarbeitungs- und Modellierungsschritte verhindern eine einfache ad hoc Analyse der Daten und machen eine Methodologie notwendig, um eine kontrollierte Durchführung und eine systematische Anwendung der einzelnen Schritte zu ermöglichen. Dazu wurde in den letzten Jahren der Wissensentdeckungsprozess (Knowledge Discovery Prozess) entwickelt.

In diesem Kapitel werden wir erst die Begriffe wie Data und Text Mining definieren bzw. in das Themengebiet des Knowledge Discoveries in Abschnitt 3.1 einordnen. Speziell interessiert uns das Gebiet des Text Mining (siehe Abschnitt 3.1.3). Wir beschäftigen uns in dieser Arbeit u.a. mit dem Clustern von Text-Dokumenten, das als Teil des Text Mining betrachtet wird. Man findet das Textclustern aber auch in benachbarten Forschungsgebieten. Aus diesem Grund werden wir uns nicht nur die Verbindung von Text Mining zum KDD-Prozess, sondern auch zu Forschungsbereichen wie Information Retrieval oder Informationsextraktion ansehen. Der KDD-Prozess wird in Kapitel 3.2 als Methodologie zum Lösen von KDD-Aufgaben eingeführt.

3.1 Knowledge Discovery und Data Mining

3.1.1 Knowledge Discovery

In der Literatur findet man unterschiedlichste Definitionen der Begriffe Knowledge Discovery (Wissensentdeckung) oder Knowledge Discovery in Databases (KDD) (Wissensgewinnung aus Datenbanken) und Data Mining. Zur Abgrenzung von Data Mining und KDD definieren wir KDD nach Fayyad u.a. wie folgt [68]:

> "Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"

Die Analyse der Daten im KDD zielt auf das Finden verborgener Muster und Zusammenhänge in diesen Daten. Unter Daten versteht man eine Menge von Fakten, die sich z.B. in einer Datenbank, aber auch in einer einfachen Datei befinden können. Eigenschaften der gefunden Muster sind Verständlichkeit für den Menschen, Gültigkeit im Rahmen von gegebenen statistischen Maßen, Neuheit und Nützlichkeit. Verschiedene Verfahren sind außerdem in der Lage, nicht nur neue Muster zu entdecken, sondern gleichzeitig generalisierte Modelle, die die gefundenen Zusammenhänge beschreiben, zu erzeugen. Der zu Grunde liegende Prozess besteht aus nichttrivialen Schritten, d.h es werden nicht einfach nur Maße wie Mittelwert oder Varianz berechnet. Der Ausdruck “potentiell nützlich” beschreibt, dass die zu findenden Muster für eine Anwendung einen Mehrwert generieren. Damit koppelt die Definition das Knowledge Discovery mit der Anwendung.

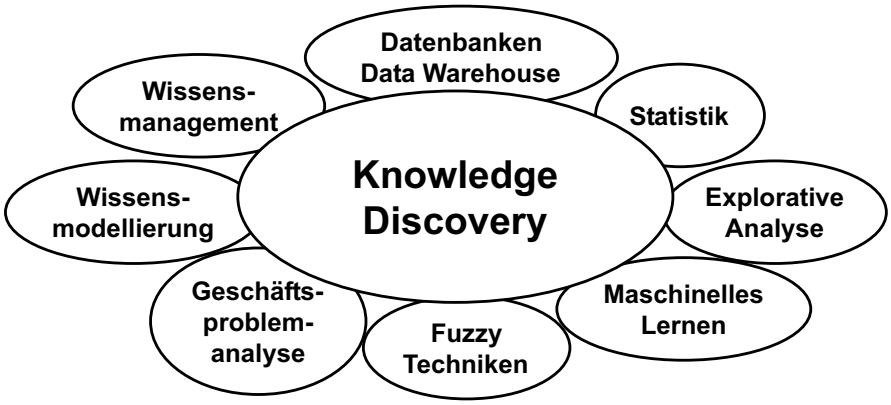

Abbildung 3.1: Benachbarte Forschungsgebiete

3.1.2 Data Mining

Die folgenden zwei unterschiedlichen Bedeutungen von Data Mining zeigen, in welchem Fluss sich das Gebiet noch befindet. Auf der einen Seite steht *Data Mining als Synonym für KDD* und beinhaltet alle Aspekte des Wissensgewinnungsprozesses. Diese Bedeutung ist insbesondere in der Praxis verbreitet und führt häufig zu Problemen, die Begriffe sauber voneinander zu trennen. Die zweite Betrachtungsweise sieht *Data Mining als Teil des KDD-Prozesses* (vgl. [68]) und umschreibt die Phasen Mustergewinnung und Modellierung, d.h. die Anwendung von Algorithmen und Verfahren zur Berechnung der gesuchten Muster bzw. Modelle (vgl. Abschnitt 3.2). Andere Autoren wie z.B. Kumar und Joshi [141] sehen Data Mining zusätzlich als die Suche nach wertvollen Informationen in *großen Datenmengen*. Für den Rest der Arbeit setzen wir Data Mining mit der Modellierungsphase des KDD-Prozesses gleich.

Die Wurzeln des Data Mining liegen in verschiedensten Fachgebieten. Damit wird der interdisziplinäre Charakter unterstrichen. Abbildung 3.1 zeigt wichtige Forschungsgebiete, aus deren Bereich die im Data Mining angewendeten Verfahren stammen.

Auf drei der angesprochenen Fachgebiete wollen wir im Folgenden eingehen. *Datenbanken* sind notwendig, um große Mengen an Daten effizient analysieren zu können. Dabei stellt die Datenbank nicht nur das Medium zum konsistenten Speichern und Zugreifen dar, sondern rückt ins nähere Forschungsinteresse, da die Analyse der Daten mit Data Mining Verfahren durch Datenbanken unterstützt werden kann. Eine Verknüpfung oder die Nutzung von Datenbanktechnologie im Data Mining Verfahren ist sinnvoll. Einen Überblick findet man in [36].

Maschinelles Lernen (ML) hat viele der im heutigen Data Mining verwendeten Verfahren hervorgebracht. Vorrangig werden in diesem Fachgebiet Suchverfahren auf symbolischen Daten entwickelt. Mitchell stellt in [169] viele der ML-Verfahren vor.

Die *Statistik* beschäftigt sich mit der Analyse von Daten. Viele Methoden der Statistik werden heute im Bereich KDD eingesetzt, wobei der Fokus bei der Statistik auf der Modellbildung der den Daten zugrundeliegenden Phänomene liegt. Einen guten Überblick des Data Mining aus Sicht der Statistik geben [24, 159].

3.1.3 Text Mining

Im Folgenden wollen wir die verschiedenen Blickwinkel auf dieses Forschungsgebiet zusammentragen und systematisieren. Text Mining oder Knowledge Discovery from Text (KDT) – erstmals erwähnt in Feldman u.a. [69] – beschäftigt sich mit der Analyse von Texten. Es nutzt Techniken des

Information Retrieval, der Informationsextraktion sowie der Sprachverarbeitung (NLP) und verbindet sie mit den Verfahren und Methoden des Data Mining, Maschinellen Lernens und der Statistik. Im Ergebnis wählt man ein ähnliches Vorgehen wie beim KDD-Prozess, wobei nicht mehr Daten im allgemeinen, sondern Texte im speziellen im Vordergrund der Analyse stehen. Daraus ergeben sich neue Fragen für die verwendeten Data Mining Verfahren.

Wir starten mit den Gebieten Information Retrieval (IR), Natural Language Processing (NLP) und Informationsextraktion (IE), die eng mit Text Mining verbunden sind und bei der Lösung ihrer Aufgaben auch Data Mining und statistische Verfahren einsetzen:

Information Retrieval (IR) ist das Finden von Dokumenten, die Antworten auf Fragen enthalten und nicht das Finden von Antworten an sich [103]. Zur Erreichung des Ziels werden statistische Maße und Methoden zur automatischen Verarbeitung von Textdaten verwendet. Information Retrieval im weiteren Sinne befasst sich mit der gesamten Breite der Informationsverarbeitung, angefangen von Data Retrieval bis zum Knowledge Retrieval (siehe [200]). Einen Überblick findet man in [200].

Natural Language Processing (NLP) Das generelle Ziel von NLP ist es, ein besseres Verständnis der natürlichen Sprache durch die Nutzung von Computern zu erlangen [138]. Andere verstehen unter NLP auch den Einsatz einfacher und robuster Techniken zur schnellen Verarbeitung von Text, wie sie z.B. in [2] vorgestellt werden. Das Spektrum der eingesetzten Techniken reicht von der einfachen Manipulation von Strings bis zur automatischen Verarbeitung von natürlichsprachlichen Anfragen. Dazu werden u.a. linguistische Analysetechniken zur Verarbeitung von Text eingesetzt.

Informationsextraktion (IE) Das Ziel von IE ist die Extraktion von spezifischen Informationen aus Text-Dokumenten. Diese werden in datenbankartigen Schemata abgelegt (vgl. [228]) und stehen dann für die Nutzung zur Verfügung.

Vergleichen wir nun die folgenden Definitionen für Text Mining mit den eben vorgestellten Forschungsgebieten:

Text Mining = Informations-Extraktion Der erste Ansatz geht davon aus, dass Text Mining im Wesentlichen der Information-Extraktion – dem Gewinnen von Fakten aus Texten – entspricht.

Text Mining = Text Data Mining Text Mining kann wie bei Data Mining auch das Anwenden von Algorithmen und Verfahren aus den Bereichen ML und Statistik auf Texten bedeuten. Dazu ist es notwendig, die Texte entsprechend vorzuverarbeiten. Viele Autoren nutzen Informations-Extraktions-Methoden, um Daten aus den Texten zu extrahieren. Auf den extrahierten Daten können dann Data Mining Algorithmen angewendet werden (vgl. [176, 78]).

Text Mining = KDD-Prozess Angelehnt an das Prozessmodell aus dem Knowledge Discovery findet man in der Literatur häufig Text Mining als Prozess mit einer Reihe von Teilschritten, unter anderem auch Informations-Extraktion sowie die Anwendung von Data Mining oder statistische Verfahren. Hearst fasst dies in [103] sinngemäß als die Extraktion von bis dahin nicht entdeckten Informationen in großen Textsammlungen zusammen. Auch Kodratoff in [138] und Gomez in [105] sehen Text Mining als prozessorientierten Ansatz auf Texten.

In der aktuellen Text Mining Forschung werden u.a. Fragen zu den Themen Text-Repräsentation, -Klassifikation, -Clustern oder der Suche nach Auffälligkeiten untersucht. Dabei spielen die Merkmalsauswahl aber auch der Einfluss von Domänenwissen und domänenspezifische Verfahren eine

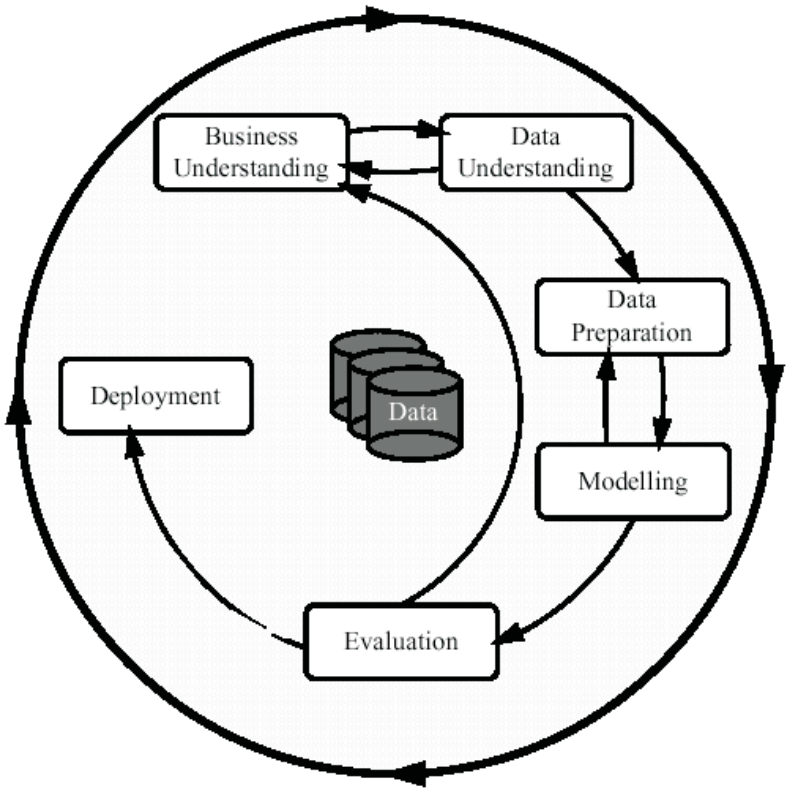

Abbildung 3.2: Schematische Darstellung des zyklischen Crisp-DM Prozessmodells

Rolle. Deswegen ist eine Anpassung der Algorithmen an die Textdaten erforderlich. Hierzu wird häufig auf die Erfahrung aus den Bereichen des IR, NLP und IE bei der Verarbeitung von Texten aufgebaut.

3.2 Der KDD-Prozess

Bei der Lösung von Geschäftsproblemen mit Data Mining bzw. Knowledge Discovery ist ein strukturiertes und zielgerichtetes Vorgehen notwendig. In der Literatur sind verschiedene Ansätze, so genannte Prozessmodelle, zur Strukturierung zu finden. Im Wesentlichen geben sie Anhaltspunkte für das Lösen der Problemstellung, indem sie die Aufgabe in verschiedene Phasen zerlegen. Grundsätzlich beinhalten alle Prozessmodelle die gleiche Idee, unterscheiden sich aber in der Anzahl der Phasen sowie in der Verteilung der Aufgaben auf die Phasen. Beispiele findet man bei Brachman/Anand [28] oder Engels [62]. Stellvertretend soll hier ein weiteres Modell – das CRISP-DM (CRoss-Industry Standard Process for Data Mining) Modell – vorgestellt werden [40]. Hinter der Gruppe, die das Crisp-DM Modell entwickelt hat, verbirgt sich eine Interessengemeinschaft aus verschiedenen Industrieunternehmen, die ein standardisiertes Vorgehen im Bereich Data Mining etabliert haben.

Das CRISP-DM Modell unterscheidet sechs Phasen, “Business Understanding”, “Data Understanding”, “Data Preparation”, “Modelling”, “Evaluation” und “Deployment” (siehe Abbildung 3.2). In der “Business Understanding” Phase werden aus Sicht des Unternehmens gemeinsam mit dem Analysten der geschäftliche Hintergrund, die geschäftlichen Erfolgsfaktoren und daraus abgeleitet die Ziele und die Erfolgsfaktoren des Knowledge Discovery Prozesses festgelegt. In der Doku-

mentation, die alle Phasen begleitet, werden neben dem Projektplan die Werkzeuge und Techniken festgehalten, die in dieser Phase in Betracht gezogen werden. “Data Understanding” beschäftigt sich mit dem Sammeln, Beschreiben und Kennenlernen der Daten. Um gute Ergebnisse sicherzustellen, wird schon in dieser Phase die Qualität der Daten geprüft und die Grundlage für die nächste Phase “Data Preparation” geschaffen. Die Ergebnisse der dieser Phase ermöglichen die Entscheidung, welche Daten in den Knowledge Discovery Prozess einfließen und wie diese vorverarbeitet werden müssen. Das Säubern der Daten, das Ableiten von neuen Attributen oder das Zusammenführen von unterschiedlichen Datenbeständen sind mögliche Vorverarbeitungsschritte (auch Preprocessing genannt). Diese haben entscheidenden Einfluss auf die Güte der Ergebnisse, wobei der Aufwand im Vergleich zu allen anderen Phasen extrem hoch ist. Die Phase “Modelling” beschäftigt sich mit der Anwendung von Verfahren zur eigentlichen Modellbildung. Die in der Begriffsdefinition erwähnten Muster werden in dieser Phase entdeckt bzw. die Modelle abgeleitet. Dazu werden die entsprechenden Data Mining Techniken ausgewählt, eine Umgebung zur Evaluierung der generierten Modelle wird erstellt und das beste Modell sowie die Parameter dafür werden ermittelt. Die Ergebnisse der Modellbildungsphase sind in der Phase “Evaluation” zu interpretieren und mit den geschäftlichen Erfolgsfaktoren abzustimmen. Anhand der erzielten Ergebnisse lassen sich nun weitere Schritte ableiten, die dann in der Anwendung des erzeugten Modells enden. In der “Deployment” Phase wird ein Plan zur Installation der Anwendung erarbeitet und die Anwendung wird in den produktiven Betrieb überführt. Die Ergebnisse des gesamten Prozesses werden als Erfahrung in einem Report abgelegt.

Die beschriebenen Phasen werden nicht strikt nacheinander angewendet, sondern man versucht sich der Lösung der einzelnen Teilprobleme in einem zyklischen und iterativen Prozess zu nähern. Dabei steht am Anfang immer die “Business Understanding” Phase, die eindeutig die Ziele aus Sicht des Unternehmens für den Knowledge Discovery Prozess festlegt und damit ein planvolles und zielgerichtetes Vorgehen garantiert. Die Pfeile in Abbildung 3.2 zeigen die möglichen Sprünge zwischen den einzelnen Phasen und symbolisieren damit den iterativen Prozesscharakter.

Die Aufgaben der abstrakten Phasen müssen im Folgenden in konkrete Teilaufgaben zerlegt werden. Abbildung 3.3 zeigt das schematische Vorgehen der Aufgabenzerlegung, ausgehend von der Unterteilung in Phasen bis zu den konkreten Prozessinstanzen über so genannte generische und spezialisierte Aufgaben. Dabei versucht man in jedem Zwischenschritt die zuvor festgelegten Aufgaben einer Phase zu präzisieren, indem man erst generische Aufgaben definiert und diese dann auf den Kontext bezogen spezialisiert, um letztendlich die Aufgabe wirklich durchzuführen.

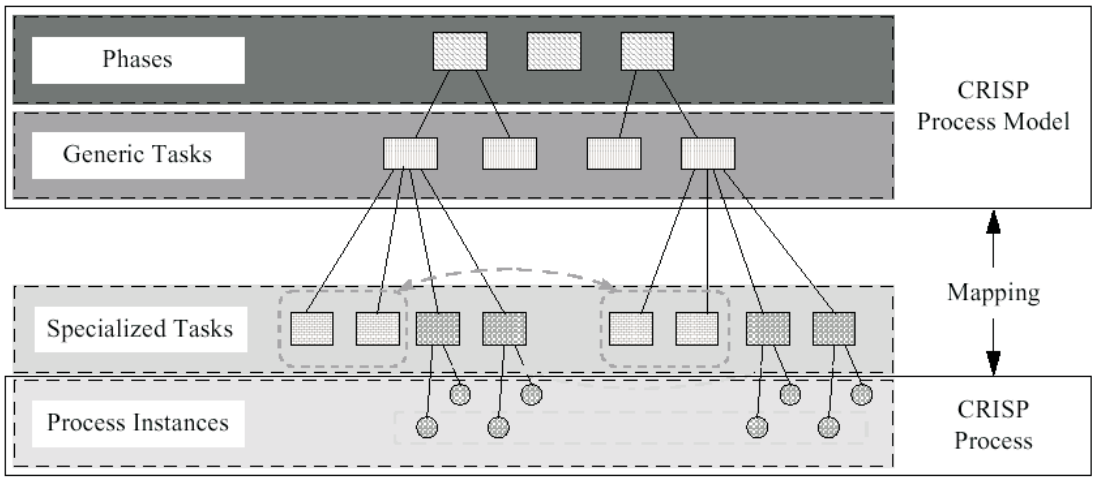

Abbildung 3.3: Crisp-DM Prozess Modell und die unterschiedlichen Stufen der Aufgabenzerlegung

Das Prozessmodell liefert uns durch seine Phasen eine grobe Richtlinie, wie KDD-Aufgaben in der Praxis systematisch gelöst werden können. Die Strukturierung hilft, die gesamte Aufgabe in übersichtliche und lösbare Teilaufgaben zu zerlegen. Diese Schritte konnten aber bis heute nicht automatisiert werden. Experten und Anwender müssen sie zusammen vollziehen.

Der KDD-Prozess liefert in dieser Arbeit die Gliederung für das Clustern von Objekten. Um die einzelnen Schritte und ihren Einfluss auf das Gruppieren von Objekten besser verstehen zu können, wurde in der Einleitung in Kapitel 1 der vom KDD-Prozess abgeleitete Clusterprozess skizziert. Dieser verdeutlicht die notwendigen Schritte.

4 Datenvorverarbeitung

In diesem Kapitel steht die Vorverarbeitung der Daten im Vordergrund. Dieser Teil des Knowledge Discovery Prozesses aus Abschnitt 3.2 hat sich als einer der wichtigen Aufgaben bei der Analyse von Daten herauskristallisiert. Wir starten das Kapitel mit der Klärung verschiedener Begriffe und der Fixierung der Notation. Anschließend beschreiben wir Vorverarbeitungsschritte zum Clustern von Textdokumenten in Abschnitt 4.2 und zum Clustern von Kunden anhand von Kommunikationsdaten in Abschnitt 4.3. Latent Semantic Indexing (LSI), das wir in dieser Arbeit als weiteren Vorverarbeitungsschritt zum Clustern von Textdokumenten einsetzen, wird in Abschnitt 4.4 eingeführt.

4.1 Notation

Die zentralen Bezeichner dieser Arbeit werden in diesem Abschnitt eingeführt. Ein Ziel der Arbeit ist das Clustern von Objekten. Objekte werden mit d bezeichnet[1]. Der zugehörige Vektor zur Beschreibung der Eigenschaften des Objektes besteht aus Merkmalen. Wir verwenden die Worte Attribut und Feature synonym zu Merkmal.

Im Bereich des Text-Clustern handelt es sich bei den Merkmalen um Worte oder allgemeiner um Terme. Worte eines Textes haben eine Entsprechung in der natürlichen Sprache. Terme sind allgemeiner gefasst und bestehen aus einer Menge von Zeichen (Strings). Die Menge der Worte ist in der Menge der Terme enthalten (mehr dazu siehe Kapitel 6). Für jedes Wort oder jeden Term zählt man die Häufigkeit des Vorkommens im Dokument und erhält so eine Objektbeschreibung. Man nennt diese Repräsentation auch “Bag of Words” oder “Bag of terms” einem “Sack”, bestehend aus Worten oder Termen (mehr dazu in Abschnitt 4.2).

Im Bereich der Kommunikationsdaten kann ein Kunde anhand verschiedener Größen – wie Verbindungsdauer oder Anzahl der Verbindungen – beschrieben werden. Das zu beschreibende Objekt ist der Kunde, den wir anhand seiner Kommunikationsmerkmale beschreiben (mehr dazu in Abschnitt 4.3).

Es sollen an dieser Stelle weitere Bezeichner bzw. Konventionen eingeführt werden.

Für den Bereich des Text-Clusterns sind unsere Objekte Dokumente. Mit D bezeichnen wir im Folgenden die Menge der Dokumente und jedes einzelne Dokument mit d, in Analogie zur der Bezeichnung von Objekten. T sei die Menge aller Terme (Merkmale) und t ein Term der Termmenge T.

Mit $\mathrm{tf}(d, t)$ bezeichnen wir die absolute Häufigkeit von Term $t \in T$ in Dokument $d \in D$, wobei $D = \{d_1, \ldots, d_n | n \in \mathbb{N}\}$ die Menge aller Dokumente und $T = \{t_1, \ldots, t_m | m \in \mathbb{N}\}$ die Menge aller unterschiedlichen Terme aus D darstellt. Weiterhin gibt $\mathrm{df}(t)$ die absolute Häufigkeit des Terms t an. Man zählt dazu in wie vielen Dokumenten aus D Term t vorkommt. Wir schreiben für den resultierenden Vektor bestehend aus allen Termen eines Dokumentes wie folgt:

$$\vec{t_d} := (tf(d, t_1), \ldots, tf(d, t_m)). \tag{4.1}$$

[1] Wir unterscheiden nicht zwischen speziellen Objekttypen und geben daher z.B. Dokumenten auch das Zeichen d.

Die Termhäufigkeit eines Terms $\text{tf}(t)$ über alle Dokumente ergibt sich dann zu:

$$\text{tf}(t) = \sum_{d \in D} \text{tf}(d, t). \tag{4.2}$$

$\text{cf}(d, c)$ gibt analog zu $\text{tf}(d, t)$ die Häufigkeit an, mit der ein Konzept c in Dokument d vorkommt (die Definition von Konzept findet man in Kapitel 6). Die Konzepthäufigkeit wird sowohl im Bereich des Dokumentclusterns als auch im Bereich des Clusterns von Kunden verwendet. Im ersten Fall handelt es sich um die Auftretenshäufigkeit der Konzepte, die zu Termen des Dokumentes korrespondieren. Im zweiten Fall geben die Konzepte z.B. Verbindungsminuten oder die Anzahl der Verbindungen wieder.

Für die mit $tfidf$ gewichteten Termvektoren (siehe Abschnitt 4.2.5.1) ersetzen wir $\text{tf}(d, t)$ durch $tfidf(d, t)$. Dies geschieht für die Konzeptvektoren analog.

Weiterhin benötigen wir eine Schreibweise für den Mittelwert der Merkmale einer Menge von Objekten. Der Mittelwert wird bei der Berechnung der Cluster mit KMeans (siehe Abschnitt 5.4.1) benötigt. Der Mittelwert für eine gegebene Menge an Objekten D bei gegebener Repräsentation $\vec{t_d}$ berechnet sich als:

$$\vec{t_D} = \frac{1}{|D|} \sum_{d \in D} \vec{t_d} \tag{4.3}$$

wobei merkmalsweise gemittelt wird. Man nennt den Mittelwert beim Clustern auch Zentroid oder Zentroidvektor.

Um die Termhäufigkeit für Mengen von Dokumenten $D' \subseteq D$ und Termen $T' \subseteq T$ angeben zu können, sei

$$\text{tf}(D', t) = \sum_{d \in D'} \text{tf}(d, t) \tag{4.4}$$

die Termhäufigkeit des Terms t in der Menge D' und

$$\text{tf}(d, T') = \sum_{t \in T'} \text{tf}(d, t) \tag{4.5}$$

die Termhäufigkeit über alle Terme der Menge T' des Dokumentes d .

4.2 Vorverarbeiten von Textdokumenten

Die Vorverarbeitung von Textdokumenten ist ein Teil des in Kapitel 3 eingeführten Wissensentdeckungsprozesses. Die einzelnen Vorverarbeitungsschritte sowie die Repräsentation der Dokumente üben maßgeblichen Einfluss auf die Güte der Text-Clusterung aus. Neben der Überführung der Dokumente in eine Vektorrepräsentation steigern einfache linguistische Techniken wie “Stemming” oder das Löschen von Stoppworten, aber auch die Gewichtung der Häufigkeitsvektoren die Clustergüte.

4.2.1 Das Vektorraummodell

Das Vektorraummodell wird im Bereich des Information Retrieval zur Repräsentation von Text-Dokumenten verwendet (vgl. [191, 160, 72]). In der Literatur findet man auch die Bezeichnung

"Bag of Words"-Modell. Es handelt sich um eine Term-Dokument-Matrix. Im Information Retrieval nutzt man diese Repräsentation um Anfragen nach Dokumenten zu beantworten. Dazu fasst man die Query-Terme auch als Vektor auf und vergleicht sie anhand eines Ähnlichkeitsmaßes mit den Dokumenten. Das ähnlichste Dokument wird als Ergebnis auf die Anfrage zurückgeliefert. Die Dokumente werden als Wort- oder Termvektoren $\vec{t_d}$ repräsentiert, wobei der Termvektor die Häufigkeit des Terms t im Dokument d angibt (siehe 4.1). Alle Dokumentvektoren zusammen ergeben die Term-Dokument-Matrix.

Zu den bekannten Eigenschaften der Termvektoren zählt deren dünne Besetzung. Jedes Dokument enthält häufig nur einen Bruchteil aller im Korpus vorkommenden Terme. Viele Terme werden im Vektor eines Dokumentes überhaupt nicht referenziert (und sind daher gleich Null). Insgesamt enthält ein typischer Korpus mehr als 10000 verschiedene Worte. Der Reuters-Korpus enthält z.B. 20574 Worte. Die Häufigkeitsverteilung der Terme im Korpus folgt dem Zipf'schen Gesetz (vgl. [147] [234]).

Auffällig am Vektorraummodell ist die erfolgreiche Anwendung in der Praxis bei gleichzeitig recht "schwacher" Vorverarbeitung der Dokumente. Dabei bietet sich das Vektorraummodell für eine schnelle Verarbeitung auch großer Dokumentmengen an. Man verschenkt durch die "Bag of Words" Betrachtung sehr viel an Informationen, die in der Anordnung der Worte und zum Teil auch in der Formatierung der Dokumente kodiert sind. Diese Informationen könnten mit Techniken aus der Linguistik, NLP oder IE extrahiert werden. Eine Kombination der verschiedenen Techniken erscheint vielversprechend aber nicht trivial (vgl. [218, 66, 97, 77, 230]).

Schauen wir uns in den nächsten Abschnitten ein paar einfache Vorverarbeitungsschritte an.

4.2.2 Stemming

Ein Vorverarbeitungsschritt beim Textclustern ist das Reduzieren der Worte auf ihre Wortstämme mit Hilfe von Heuristiken. In der Literatur wird der Prozess der Wortreduktion, also das Abschneiden von Affixen, auf die Stammform als Stemming bezeichnet. Ein mit dem Stemming sehr verwandter Prozess ist die so genannte Lemmatization. Lemmatization ist im Unterschied zum Stemming das Finden der Lexeme von gebeugten Worten und nicht das pure Abschneiden der Endungen zu unverständlichen Wortstücken (vgl. [160]). Der Vorverarbeitungsschritt des Stemmings ist nicht ganz unumstritten, da die in den Wortformen zusätzlich erhaltene Information nützlich für die Anwendung sein kann.

In [160] (vgl. S. 132) wird argumentiert, dass Stemming an sich intuitiv sinnvoll ist und man mit den Wortstämmen wahrscheinlich bessere Ergebnisse erhalten wird. Mit Hilfe der empirischen Forschung konnten Schwächen von Stemming im Information Retrieval herausgearbeitet werden. [160] geben drei Gründe für das Scheitern von Stemming an. Erstens ist es wichtig, Worte einer Wortgruppe ohne Stemming als Suchworte zu verwenden, um den Sinn zu erhalten, z.B. wird die Suche nach "Operating System" mit "operat" und "system" nicht wesentlich besser funktionieren als mit den ursprünglichen Worten. Zweitens kann das Zerlegen eines Token, das eine Wortgruppe darstellt, zu Problemen führen, da die Information gerade in der Gruppierung besteht. Der dritte Grund ist die englische Sprache. Sie enthält nur sehr wenig Morphologie und eine intelligente morphologische Analyse ist daher nicht nötig.

Beim Textclustern wird Stemming üblicherweise verwendet [18, 182, 206] und hat sich positiv auf die Ergebnisse ausgewirkt. Wir haben in unseren Experimenten auf den bekannten Porter-Stemmer [185] zurückgegriffen.

4.2.3 Stoppworte

Das Führen einer Stoppwort-Liste ist ebenfalls ein gebräuchlicher Ansatz im Bereich Text Mining und Information Retrieval. Die Stoppwort-Liste enthält Worte, die in der Sprache bekanntermaßen sehr häufig vorkommen, wie z.B. “der”, “die” oder “das” im Deutschen. Es existieren für die verschiedenen Sprachen Standard-Listen. Eine Liste mit Stoppworten für verschiedene Sprachen findet man auf der CLEF-Webseite[2] (Cross-Language Evaluation Forum). Im Information Retrieval wird für das Englische sehr häufig die Stoppwortliste[3] des SMART Systems eingesetzt [190].

4.2.4 Löschen seltener Worte (Pruning)

Das Löschen seltener Worte ist durch die Tatsache motiviert, dass seltene Terme bei der Identifizierung von Clustern kaum helfen. Salton und Buckley beschreiben in [192] das Phänomen, dass die sehr und mittel häufigen Terme die meisten Informationen enthalten. Dies wird im Bereich Information Retrieval auch als “Gesetz” der IR bezeichnet (siehe [195]). Im Bereich Text-Klassifikation wird die Annahme durch empirische Studien untermauert. Sebastiani [195] findet in der Literatur zwei gängige Arten um seltene Terme zu löschen: Auf der einen Seite werden alle Terme gelöscht, die in weniger als δ Dokumenten vorkommen (Dokument-Pruning). Auf der anderen Seite wird die Häufigkeit der Terme im gesamten Korpus genommen und bei Unterschreiten der Schranke δ der Term gelöscht. In diesem Fall handelt es sich um Term-Pruning, wobei wir die Schranke mit “Prunethreshold” δ bezeichnen. Beim Dokument-Pruning liegen die Werte meist im Bereich zwischen eins und drei. Beim Term-Pruning wird meist eine Schranke von eins bis fünf gewählt (siehe [195]).

Worte, die nur einmal im gesamten Datenbestand auftauchen, sind für den Menschen meistens nur bei Kenntnis des Wortes von Bedeutung. Diese Worte helfen aber nicht bei Clusterverfahren, die auf der Basis von wiederholtem Auftreten der Worte in mehreren Dokumenten die Ähnlichkeit zueinander bestimmen. Ohne eine Wiederholung ist ein Vergleich zweier Dokumente anhand dieses Wortes nicht möglich. Im Gegensatz zu den Clusterverfahren ist die Situation im Information Retrieval, dass nur ein Dokument das angefragte Wort enthält, besonders gut, weil dann dieses Dokument bestimmt das einzig relevante Dokument ist. Außerdem wird in dieser Situation kein Ranking benötigt. Man wird diese Worte daher nicht löschen.

Formal lässt sich das Term-Pruning wie folgt aufschreiben: Alle Terme $t \in T$, die eine Termhäufigkeit kleiner als der Prunethreshold δ aufweisen, werden aus der Menge der Terme gelöscht. Daraus ergibt sich die reduzierte Termmenge $T := \{t \in T \mid \text{tf}(t) > \delta\}$, welche dann die Grundlage für das Clustern bildet. Beim Dokument-Pruning wird die Schranke Prunethreshold δ mit der Dokumenthäufigkeit des Terms verglichen $\text{df}(t)$. Die neue Termmenge ergibt sich dann zu: $T := \{t \in T \mid \text{df}(t) > \delta\}$

In Kapitel 8.2.1 werden wir die Auswirkungen des Term-Pruning auf die Clustergüte untersuchen.

4.2.5 Gewichtung von Termvektoren

4.2.5.1 tfidf

Das tfidf Maß $(\text{term frequency} - \text{inverted document frequency})$[4] gewichtet die Häufigkeiten von Termen (tf) eines Dokumentes mit einem Faktor (idf), der die Wichtigkeit entsprechend der Anzahl der

[2] http://www.unine.ch/Info/clef/

[3] ftp://ftp.cs.cornell.edu/pub/smart

[4] In der Literatur verwenden verschiedene Autoren die gleiche Abkürzung “tfidf” für verschiedene Gewichtungsschemata (vgl. [195]).

Dokumente, in denen der Term vorkommt, anpasst. Terme, die sehr selten oder sehr oft vorkommen, erhalten daher ein geringeres Gewicht als Terme, welche die Balance zwischen den beiden Extremen halten. Die Gewichtung geschieht unter der Annahme, dass die Terme mit den beiden extremen Auftretenshäufigkeiten nicht viel zum Clusterergebnis beitragen können. Beispielsweise kommt der Term "Reuters" am Ende jedes Dokumentes im Reuters-Korpus vor. Damit entspricht df(t) der Anzahl aller Dokumente im Korpus und das Gewicht des Terms ergibt sich zu 0. tfidf ist wie folgt definiert [197]:

Definition 1 (tfidf). *tfidf von Term t in Dokument d ist definiert als:*

$$tfidf(d,t) := \log(tf(d,t)+1) * \log\left(\frac{|D|}{df(t)}\right) \tag{4.6}$$

wobei df(t) die Dokumentenhäufigkeit von Term t ist, die angibt, in wie vielen Dokumenten Term t vorkommt.

Wenn wir die tfidf Gewichtung anwenden, dann ersetzen wir den Termvektor $\vec{t_d} := (\text{tf}(d,t_1),\ldots,\text{tf}(d,t_m))$ durch $\vec{t_d} := (tfidf(d,t_1),\ldots,tfidf(d,t_m))$.

In der Literatur existieren ausgeklügeltere Maße als tfidf (siehe z.B., [9]). Wir wollen im nächsten Abschnitt einige dieser Maße vorstellen. Prinzipiell können sie tfidf ersetzen. Es wäre dann zu zeigen, dass auch diese Maße die Ergebnisse mit Hintergrundwissen positiv beeinflussen. Ein Ziel der Arbeit war herauszufinden, ob und wie sich Gewichtungsmaße auf die Integration von Hintergrundwissen auswirken. Daher wurde das tfidf-Standardmaß in dieser Arbeit verwendet.

4.2.5.2 Verwandte Gewichtungen

Die tfidf Gewichtung kann z.B. durch "Mutual Information" oder "BM25" ersetzt werden. Pantel u.a. nutzen in [182] "Mutual Information" (MI) zur Gewichtung der Termvektoren. MI ist wie folgt definiert:

$$MI(d,t) = log\frac{P(t,d)}{P(t)\cdot P(d)}, \tag{4.7}$$

wobei $P(t,d)$ die Wahrscheinlichkeit ist, dass Term t und Dokument d gemeinsam auftreten und $P(t)$ und $P(d)$ die Wahrscheinlichkeiten des Terms bzw. des Dokumentes sind.

Eine weitere Alternative ist die Näherung der bekannten BM25-Gewichtung [189] von Amit u.a. in [10] die prinzipiell nach dem gleichen Schema wie tfidf-Gewichtung funktioniert:

$$BM25(d,t) = \frac{\text{tf}(d,t)\cdot log(\frac{|D|-\text{df}_{(t)+0.5}}{\text{df}_{+0.5}})}{2\cdot(0.25+0.75\cdot\frac{dl}{avdl})\text{tf}(d,t)} \tag{4.8}$$

dl gibt die Länge der Dokumente in Bytes und avdl die durchschnittliche Länge der Dokumente im Korpus an. Die restlichen Konstanten dienen zur besseren Gewichtung der Vektoren. BM25 unterscheidet sich bei der Berechnung des idf-Wertes und bei der Verknüpfung von tf mit idf von tfidf.

Einen Rahmen für Termgewichtungen spannen Amati u.a. und [9, 34] auf. Weiterhin geben sie einen Überblick der verschiedenen Varianten von Termgewichtungen und vergleichen die Maße empirisch mit Hilfe des TREC-10 Datensatzes. Der folgende Abschnitt diskutiert noch die einfache Möglichkeit die Termhäufigkeiten zu logarithmieren.

4.2.6 Absolute vs. logarithmierte Werte

Dokumente können unter anderem durch die absolute Häufigkeit der Terme oder auch Konzepte repräsentiert werden. Weiterhin entspricht die Häufigkeitsverteilung der Terme im Korpus einer hyperbolischen Verteilung, in der die meisten Terme nur sehr selten auftreten und einige wenige Terme sehr oft vorkommen. Für Clusterverfahren, die nicht für spezielle Verteilungsfunktionen entwickelt wurden, wirkt sich diese Verteilung negativ aus. Logarithmiert man die absoluten Häufigkeiten tf mit der Funktion $\log(\text{tf} + 1)$, führt dies meist zu einer deutlichen Steigerung der Clusterergebnisse. Wir werden daher bei einigen Experimenten auf die logarithmierten Häufigkeiten zurückgreifen.

Dies trifft nicht nur auf die Text-Dokumente zu, sondern auch auf die Verteilung der Verbindungsdauer im Bereich der Telekommunikation. Aus diesem Grund wurden auch die kundenbeschreibenden Merkmale logarithmiert (siehe Kapitel 4.3.2).

4.2.7 Zusammenfassung

In diesem Abschnitt haben wir die typische Repräsentation von Textdokumenten, das Vektorraummodell, sowie gängige Vorverarbeitungsmethoden eingeführt. Wie schon eingangs erwähnt, stellt die Vorverarbeitung einen zentralen Punkt für die erfolgreiche Berechnung von Data Mining Modellen dar. Daher existieren eine Reihe von Ansätzen zur Verbesserung der Vorverarbeitungsschritte z.B. mittels Linguistik. Auch die Auswahl der richtigen Merkmale spielt eine wichtige Rolle [171] bei der Vorverarbeitung im Bereich Text Mining.

4.3 Vorverarbeitung von Kommunikationsdaten

In diesem Abschnitt beschäftigen wir uns mit den Vorverarbeitungsschritten für Kommunikationsdaten[5]. Wir behandeln das typische Vorgehen zum Ableiten von kundenbeschreibenden Merkmalen aus den Kommunikationsdatensätzen und beschreiben die Eigenschaften des resultierenden Datensatzes.

4.3.1 Ableiten von Merkmalen aus Kommunikationsdaten

Die erste Aufgabe zur Analyse der Kunden anhand ihrer Kommunikationsdaten ist die Vorverarbeitung der Daten, um kundenbezogene Merkmale zu generieren. Die Kommunikationsdatensätze in der Originalform müssen dazu in kundenbeschreibende Merkmale transformiert werden. Die Merkmale müssen so beschaffen sein, dass man sie für alle Kunden generieren kann. Außerdem sollten die Merkmale den Kunden möglichst gut charakterisieren, um eine Clusterung überhaupt zu ermöglichen. Um diese Aufgabe zu lösen, wurden die Kommunikationsdatensätze der 10 % Stichprobe (siehe Abschnitt 2.5) zusammengefasst (aggregiert) und kundenbezogen repräsentiert. Abbildung 4.1 zeigt die vier wesentlichen Dimensionen Tarifzone (Dim. 1), Uhrzeit (Dim. 2), Tagart (Dim. 3) und Verbindungsnetzbetreiber (Dim. 4), die jedes Gespräch charakterisieren. Man ist nun in der Lage, jedes Gespräch in genau eine Kombination der Ausprägungen dieser vier Dimensionen einzusortieren, z.B. “ein Ortsgespräch zwischen 9.00 und 18.00 Uhr an einem Werktag über den Anbieter Telekom”. Das neue Merkmal wird durch die angegebenen Kombinationen der Kommunikationseigenschaften festgelegt. Insgesamt ergeben sich 84 Merkmale. Zum Zeitpunkt der Datenerhebung

[5] Die Kommunikationsdaten wurden von der Deutschen Telekom AG zur Verfügung gestellt und sind in Abschnitt 2.5 beschrieben.

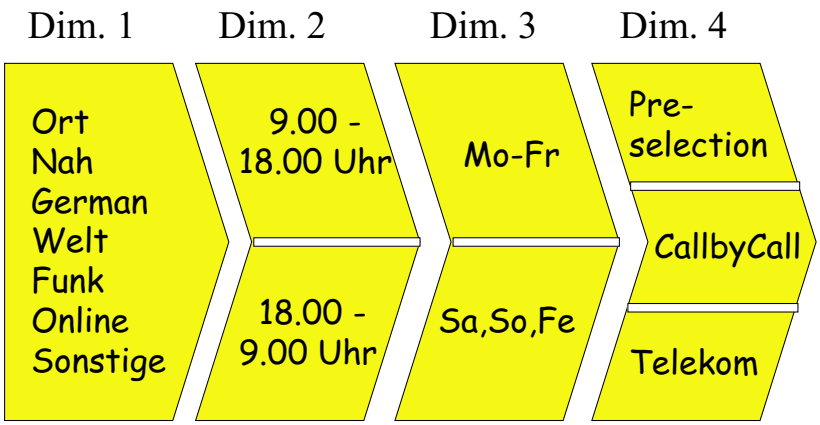

Abbildung 4.1: Dimensionen für die Merkmalsgenerierung

waren bei allen Kunden keine Ortsgespräche über Call by Call und Preselection möglich. Mit Hilfe der übrig bleibenden 76 Merkmale kann jeder Kunde beschrieben werden.

Ermittelt man alle Gespräche eines Kunden, die in ein Merkmal fallen, so bieten sich verschiedene Möglichkeiten, diese Gespräche zu einem Wert zusammenzufassen. Eine Variante, die auch exemplarisch in den Analysen in Kapitel 10 zum Einsatz kam, ist die Summe aller Verbindungsminuten zu berechnen. Man könnte sich aber auch vorstellen, nur die Anzahl der Gespräche zu zählen. Weitere Varianten findet man in Kapitel 10.1.5.1 bei der Befragung der Telekomexperten. Die Auswahl des Maßes hängt von der zu lösenden Aufgabe ab. Für das Generieren der Merkmale benötigt man ferner noch einen Referenzzeitraum. Dieser muss am Anfang der Analyse festgelegt werden.

Das Ergebnis der Transformation ist in unserem Beispiel ein 76-dimensionaler Datensatz mit der Summe aller Verbindungensminuten pro Merkmal, wobei jedes Merkmal nur die Kommunikationsdatensätze entsprechend der Merkmalsbeschreibung berücksichtigt. Jeder Kunde erhält durch den 76-dimensionalen Merkmalsraum ein Profil. Das Profil fasst sein Verhalten für den Analysezeitraum zusammen. Die berechnete Repräsentation erlaubt es, Kunden miteinander zu vergleichen. Die Merkmale spiegeln das Kommunikationsverhalten der Kunden wieder. Den Schritt der Merkmalserzeugung nennt man auch Reverse-Pivoting [186], [158].

Möchte man Analysen unterschiedlicher Zeiträume vergleichen, so ist zu beachten, dass die Änderung des Referenzzeitraumes den Vergleich der Ergebnisse erschwert. So sind unterschiedliche Monate unterschiedlich lang und saisonale Effekte sind nicht zu unterschätzen. Auch ist eine Auswertung über mehr als einen Monat denkbar und wünschenswert.

4.3.2 Eigenschaften der Telekom-Merkmale

Visualisiert man die Verteilungsfunktion der berechneten Merkmale, so erhält man eine linksschief normalverteilte Funktion (eine ähnliche Verteilungsfunktion wurde für die Gesprächsdauerverteilung in [124] beschrieben). Für die Clusterung der Daten mit z.B. KMeans erweist sich eine linksschiefe Normalverteilung als sehr nachteilig. Die Ergebnisse einer Clusterung mit KMeans sind deswegen ohne weitere Vorverarbeitung unbrauchbar. Bei einem Clusterlauf erhält man sehr viele Cluster, die nur einen oder sehr wenige Kunden enthalten und meist ein oder zwei Cluster mit sehr vielen Kunden. Um diesen Effekt zu vermeiden, sollte die linksschiefe Verteilung der Daten näherungsweise in eine Normalverteilung transformiert werden. Hierfür bietet sich das Logarithmieren der Merkmale (siehe Abschnitt 4.2.6) an. Eine Transformation mit der log Funktion führt hier zu fast normalverteilten Daten.

Bei der genaueren Betrachtung der Verteilungsfunktion fällt ein weiteres Phänomen auf. Neben den Daten, die linksschief normalverteilt sind, findet man in jedem Merkmal sehr viele Kunden, die gar keine Gespräche über eines der Merkmale führen. Damit teilen sich die Kunden typischerweise in jedem Merkmal in Kunden, die kommuniziert haben, und Kunden, die keine Gespräche in diesem Merkmal besitzen. Diese Eigenschaft weist auch die Vektorrepräsentation der Text-Dokumente auf. Es handelt sich in beiden Fällen um eine so genannte dünn besetzte hochdimensionale Matrix.

Die Auswahl der berechneten Merkmale erfolgt in dieser Arbeit in gewisser Weise willkürlich, was uns zu der Frage führt: “Welche Merkmale beschreiben den Kunden am besten bzw. welche Aggregate sollten gebildet werden?”. Leider können wir auf diese Frage keine endgültige Antwort geben. Die Auswahl der Merkmale hängt von der zu analysierenden geschäftlichen Fragestellung ab. Weiterhin stellen die 76 Merkmale gerade im Vergleich zur Text-Dokument-Repräsentation mit mehr als 10000 Merkmalen noch keine “große” Anzahl an Merkmalen dar. Es wäre leicht vorstellbar, die Uhrzeit statt in zwei Zeitfenster in vier oder mehr einzuteilen, um so den Verkehr der Kunden detaillierter analysieren zu können. Auf diese Weise lässt sich die Anzahl der Merkmale leicht auf 10000 steigern und der Merkmalsraum wird dem der Text-Dokumente immer ähnlicher. Abschnitt 10.1.3.1 wird sich mit dem Problem des hochdimensionalen Raumes für das Clustern von Kunden auseinander setzen. Wir werden zeigen, dass die hohe Merkmalszahl nicht nur zu unverständlichen sondern auch zu schlechten Clusterergebnissen führt und dass die Anzahl der Dimensionen reduziert werden muss. Um verständliche Strukturen in den Kundendaten zu entdecken, wenden wir das Subjektive Clustern auf den Kommunikationsdaten an (siehe Kapitel 7).

4.4 Latent Semantic Indexing (LSI)

Latent Semantic Indexing ist eine wichtige Erweiterung des Vektorraummodelles aus der Sicht des Information Retrieval [57]. Um die Güte von Information Retrieval Ergebnissen zu steigern, nutzt LSI die implizite Struktur zwischen den Termen und Dokumenten aus [48]. Eine gestutzte Singulär-Wert-Zerlegung (singular value decomposition, SVD) wird zur Schätzung der verborgenen Struktur verwendet. LSI kann außerdem zur Dimensionalitätsreduktion eingesetzt werden. Der berechnete Konzept-Raum besteht meist aus deutlich weniger Merkmalen als der Originalraum [160].

LSI ist in der Lage, Terme und Dokumente anhand ihres gemeinsamen Auftretens im Korpus in Verbindung zu bringen und sowohl Terme als auch Dokumente, die in inhaltlicher Beziehung zueinander stehen, in einem projizierten Raum in die gleiche Region, also räumlich nahe zueinander, abzubilden. Stichwortbasierte Anfragen liefern auf der Basis des Konzept-Raumes nicht nur Dokumente als Antwort, die das Stichwort enthalten, sondern auch Dokumente aus der inhaltlichen Nachbarschaft, d.h. Dokumente zum gleichen Thema. LSI ist in der Lage zwei Kernprobleme des Information Retrieval zu lösen: Polysemie und Synonymie. Die durchschnittliche Verbesserung gegenüber herkömmlichen vektorbasierten Ansätzen beträgt bis zu 30 % (vgl. [23]).

Wir wiederholen die wichtigsten Ideen von LSI und setzen LSI in Abschnitt 8.4 als Vorverarbeitungsschritt für das Clustern von Dokumenten ein. Dabei wenden wir den Clusterschritt nicht mehr auf dem Orginalvektorraummodell sondern auf dem LSI-Konzept-Raum an und clustern auf der Basis der LSI-Konzepte. Schütze a.u. berichten in [194] über die Anwendung von LSI zur Dimensionsreduktion beim Clustern von Dokumenten. Lerman untersucht in [148] den Einfluss der Dimensionsreduktion auf die Ergebnisse von hierarchisch-aglomerativen Clusterverfahren. Wir folgen bei der Einführung von LSI Dowling [57] und Berry u.a. [23].

Es sei $A = (\vec{t}_1, \ldots, \vec{t}_{|D|})^T$ die $n \times m$-Dokument-Term-Matrix und ohne Beschränkung der Allgemeinheit sei $m \geq n$. Dann ist die Singulär-Wert-Zerlegung, bezeichnet mit $SVD(A)$, definiert als:

$$A = U\Sigma V^T \tag{4.9}$$

wobei U eine $m \times r$ Matrix, V eine $n \times r$ Matrix, r der Rang von A und Σ eine Diagonalmatrix ist, die die Singulärwerte enthält. Behält man nun nur die k größten Singulärwerte in Σ und die passenden Spalten in U und V, lässt sich A folgendermaßen annähern:

$$A_k = U_k\Sigma_k V_k^T \tag{4.10}$$

wobei A_k die wesentliche Struktur ohne Rauschen, das durch die Verwendung unterschiedlicher Worte für den gleichen Sachverhalt entsteht, wiedergibt (vgl. [48]). Stichwortbasierte Anfragen werden mit Hilfe des Kosinus-Maßes zwischen Dokument und Anfrage im Konzept-Raum berechnet [140].

4.5 Merkmalsextraktion zur Clusterbeschreibung

Merkmalsextraktion (im Englischen “feature section”) spielt eine wichtige Rolle bei der Modellbildung im überwachten Lernen. Man setzt die Methoden erfolgreich zur Vermeidung von Overfitting[6] ein (vgl. [171]). Beim unüberwachten Lernen kann die Merkmalsextraktion vor und nach dem Clustern zum Einsatz kommen. Es zeigt sich, dass die Auswahl der geeigneten Merkmale vor dem Clustern nicht trivial ist und häufig in einem “Trial and Error” Prozess ([123] S. 271) endet. Unterschiedliche Teilmengen der Merkmale werden zum Clustern der Objekte ausprobiert und die Ergebnisse analysiert und bewertet. Für die Merkmalsauswahl *vor* dem Clustern bieten wir die Lösung des Subjektiven Clusterns an, welche Hintergrundwissen zur systematischen Strukturierung des Merkmalsraumes nutzt (siehe Kapitel 7). In diesem Abschnitt steht die Merkmalsextraktion *nach* dem Clustern im Vordergrund. Speziell sind wir an Merkmalen interessiert, die die Ergebnisses des Clusterns in verständlicher Form beschreiben.

Durch das Clustern stehen uns für die Merkmalsextraktion Klassen zur Steuerung der Extraktion zur Verfügung. Es handelt sich daher um einen überwachten Prozess. Wir können neben den vom Clusterverfahren selbst gelieferten Beschreibungen auch auf bekannte Merkmalsextraktionsverfahren aus dem Bereich des überwachten Lernens zurückgreifen.

Der folgende Abschnitt motiviert verschiedene Merkmalstypen, die zur Beschreibung von Clustern herangezogen werden können. In Abschnitt 4.5.2 wird die Auswahl der wichtigsten Merkmale aus Zentroidvektoren des KMeans-Clusterverfahrens eingeführt. Da die Auswahl beschreibender Merkmale eng mit der überwachten Merkmalsextraktion verbunden ist, gehen wir in Abschnitt 4.5.3 auf gängige Merkmalsextraktionsmethoden ein.

4.5.1 Motivation

Um die Cluster eines Clustermodelles inhaltlich und für Menschen verständlich beschreiben zu können, benötigen wir entsprechende Merkmale. Jedes Objekt wird anhand von Merkmalen während des Clusterprozesses beschrieben. Zum Beispiel bieten sich die Terme des Termvektoren eines Dokumentes zur Beschreibung eines Dokumentes an. Sie repräsentieren in gewisser Art und Weise den Inhalt der Dokumente. Leider ist der Termvektor einer Dokumentmenge normalerweise sehr groß. Nicht alle Terme transportieren die gleiche Menge an Informationen, die zum Verständnis des Clusters benötigt werden. Die folgenden Methoden versuchen Terme zu extrahieren, die am wichtigsten

[6] Unter Overfitting versteht man die Überanpassung der Modelle an den Datensatz.

für die *Beschreibung* bzw. *Abgrenzung* des Inhaltes eines Clusters sind, und so eine möglichst große Informationsmenge transportieren (vgl. [126, 127, 209]. Wir unterscheiden zwei Kategorien:

Beschreibende Merkmale sind Merkmale, die den Inhalt einer gegebenen Menge von Objekten unabhängig von allen anderen Objekten so prägnant wie möglich wiedergeben.

Unterscheidende Merkmale sind Merkmale, die den Inhalt einer gegebenen Menge von Objekten in Abhängigkeit von allen übrigen Objekten so prägnant wie möglich wiedergeben. Diese Merkmale grenzen die gegebene Objektmenge vom Rest des Datensatzes ab.

Vorstellbar ist auch eine Kombination beider Merkmalstypen. Oft kommen Merkmale in beiden Merkmalsmengen vor. Es besteht also ein großer Zusammenhang zwischen den genannten Merkmalstypen.

A. Strehl u.a. unterscheiden bei der Merkmalsextraktion für Clusterergebnisse in [209] ebenfalls diese beiden Gruppen. Sie nutzen für die beschreibenden Merkmale die Auftretenshäufigkeit der Terme im Cluster und für die unterscheidenden Merkmale den Unterschied der Auftretenshäufigkeit der Worte im Cluster gegenüber einem durchschnittlichen Dokument. Auch Karypis u.a. gehen in [126] ähnlich vor. Wir beschreiben ihr Vorgehen für die beschreibenden Merkmale im nächsten Abschnitt im Detail.

4.5.2 Merkmalsextraktion aus Zentroidvektoren

Die Extraktion beschreibender Merkmale aus Zentroidvektoren einer KMeans Clusterung wird in [126, 127] vorgestellt. Dort wird auch die Aussagekraft der Clusterbeschreibung anhand von realen Beispieldatensätzen demonstriert. In dieser Arbeit wird eine modifizierte Variante von Karypis u.a. verwendet. Sie liefert nicht eine fixe Anzahl von Merkmalen, sondern alle Merkmale der Zentroide, deren Gewichte über einer festgelegten Schranke liegen. Die Schranke legt die Bedeutung des Merkmales zur Beschreibung eines Clusters fest und erlaubt die indirekte Kontrolle der Gesamtanzahl der zur Beschreibung herangezogenen Merkmale.

Gegeben sei eine Clusterung $\mathbb{P}$ der Objekte D und die Repräsentation jedes Objektes $d \in D$ durch den entsprechenden Vektor $\vec{t_d}$. Weiterhin benötigen wir einen Wert für die Schranke θ. Die Schranke θ wird als Anteil des Maximalwertes im Zentroid angegeben. Ziel ist es für jeden Cluster $P \in \mathbb{P}$ eine Menge von "wichtigen" Merkmalen zu spezifizieren, die zur Beschreibung verwendet werden können.

Der Zentroidvektor (vgl. Gleichung 4.3) ist der Vektor

$$\vec{t_P} := (g(P, t_1), \ldots, g(P, t_{|T|})) \tag{4.11}$$

eines jeden Clusters $P \in \mathbb{P}$ mit dem Gewicht $g(P,t) = \frac{1}{|P|} \sum_{d \in P} tfidf(d,t)$ des Merkmales t im Zentroidvektor. Das Gewicht pro Term entspricht dem Mittelwert über der Objektmenge. tfidf kann durch tf oder cf ersetzt werden. Für die Berechnung der wichtigen Merkmale normalisieren wir jeden Zentroidvektor auf die Länge eins: $||\vec{t_P}||_2 = 1$. Der Maximalwert über alle Cluster und Merkmale ergibt sich zu

$$zmax = max_{P \in \mathbb{P}, t \in T}(g(P,t)).$$

Die beschreibenden Merkmale des Clusters erhält man, indem man in die Ergebnismenge jedes Clusters alle Merkmale aufnimmt, die ein Gewicht g größer $(\theta \cdot zmax)$ haben. Der Vektor ist wie folgt definiert:

$$\forall_{P\in\mathbb{P},t\in T}(g(P,t)=1 : g(P,t) \geq \theta \cdot zmax) \text{ und } (g(P,t)=0 : g(P,t) < \theta \cdot zmax) \quad (4.12)$$

Die Menge der Merkmale T kann auf die im Zentroiden verbleibende Menge $\mathcal{T}$ wie folgt reduziert werden:

$$\mathcal{T} := \{t : \exists g(P,t) = 1 \text{ mit } P \in \mathbb{P} \text{ und } t \in T\}. \quad (4.13)$$

Die Methode kann sowohl auf Cluster als auch auf Dokumente angewendet werden.[7] Es gibt auch die Möglichkeit, mehrere Schranken θ_1, θ_2 festzulegen. Auf diesem Weg kann man mehr Informationen aus der Clusterung in die Beschreibung übernehmen, wobei man die Balance zwischen Informationsmenge (mehr Schranken) und Verständlichkeit (weniger Schranken) halten muss.

4.5.3 Verwandte Ansätze zur Merkmalsextraktion

Das Festlegen von Schwellwerten oder Schranken stellt einen Weg zur Erzeugung von Klasseneinteilungen (binning) dar. Man nennt die Umwandlung von numerischen Werten in kategorische auch Diskretisierung. Verschiedene Methoden und Verfahren findet man z.B. in [231]. Auf die Bedeutung im Allgemeinen wird in [186] und als wichtiger Vorverarbeitungsschritt für z.B. Klassifikationsaufgaben z.B. in [56] eingegangen. Die unterschiedlichen Methoden zur Diskretisierung bieten Ansatzpunkte, um mehr von den numerisch kodierten Informationen als kategorische Werte abzulegen und so die Cluster noch besser beschreiben zu können und zu verstehen.

In [13] wird ein Ansatz zur Extraktion von aussagekräftigen Bezeichnungen (meaningful labels) basierend auf Self-Organizing Maps vorgestellt. Im Artikel wird das G-Maß eingeführt, welches eine automatische Extraktion der Bezeichnungen erlaubt. Die extrahierten Bezeichnungen werden mittels des so genannten “z-value” mit anderen Verfahren verglichen. So wird auch die Güte der Methode bestimmt.

[125] gibt einen Überblick über Maße für die Gewichtung von Termen/Worten aus Texten. Neben der Gewichtung von einzelnen Termen werden auch Maße für Bi-Gramme vorgestellt. Durch die Zusammenfassung von Termen ließen sich aussagekräftigere Merkmale extrahieren.

Vergleiche von Ansätzen zur Merkmalsextraktion sowie einen Überblick über den damals aktuellen Stand der Forschung geben Blum und Langley in [26]. Mit der Kombination von Klassifikation und Merkmalsextraktion im Allgemeinen befassen sich Molina u.a. in [173] und von Textdokumenten im Besonderen D. Mladenic in [171]. Einen Überblick über die aktuelle Literatur im Bereich Text-Klassifikation und Merkmalsextraktion gibt [170].

Nachdem wir in diesem Kapitel die gängigen Vorverarbeitungstechniken vorgestellt haben, wird im nächsten Kapitel der Begriff Cluster definiert und verschiedene Clusterverfahren werden eingeführt. Das Kapitel wird weiterhin einen Überblick über ausgewählte Themen der aktuellen Forschung im Bereich des Clusterns beinhalten.

[7] Jeder Cluster könnte genau aus einem Dokument bestehen.

5 Clusteranalyse

Die Clusteranalyse zählt im Bereich des Maschinellen Lernens bzw. Data Mining zu den unüberwachten Lernverfahren. Viele Methoden wurden schon früh im Bereich der multivariaten Statistik entwickelt. In beiden Bereichen existieren viele Veröffentlichungen zum Thema Clustern. Ein häufig zitierter Überblicksartikel ist von Jain [123] und ein neuerer von Berkhin [20]. Kaufman und Rousseeuw stellen in ihrem Buch [129] eine Reihe von Clusterverfahren sowie Evaluierungsmaße für Clusterverfahren vor. Einige Herausforderungen für die Zukunft aus dem Bereich des Clusterns hochdimensionaler Daten beschreiben Steinbach u.a. in [207]. Die häufige Anwendung in vielen Bereichen der Wirtschaft und der Wissenschaft sowie die große Anzahl an Veröffentlichungen zu diesem Thema verdeutlichen die Bedeutung der Clusteranalyse.

Unter der Clusteranalyse oder auch kurz Clustern versteht man im Allgemeinen das (automatische) Gruppieren von (homogenen) Objekten auf der Basis bekannter Informationen über und Beziehungen zwischen Objekten. Die Gruppen bezeichnet man als Cluster. Anders ausgedrückt teilt man die Objekte anhand von Daten in bedeutsame und nützliche Gruppen. Automatische Clusterverfahren wurden in Gebieten wie z.B. Biologie und Psychologie entwickelt und werden heute u.a. zum Gruppieren von Textdokumenten oder zum Finden von Kundengruppen im Marketing erfolgreich eingesetzt.

Wir werden im folgenden Abschnitt den Begriff Clusterung definieren und verschiedene Gütekriterien für Cluster herausarbeiten. Abschnitt 5.2 stellt Distanz- und Ähnlichkeitsmaße vor, die Anwendung in den Clusterverfahren finden. Evaluierungsmaße für Clusterverfahren bilden den Kern von Abschnitt 5.3. Die Abschnitte 5.4 und 5.5 geben die in der Arbeit angewendeten Clusterverfahren KMeans, Bi-Sec-KMeans und Formale Begriffsanalyse wieder. Wir schließen das Kapitel mit Abschnitt 5.6 – einem Überblick an bekannten Clusterverfahren – und setzten diese Verfahren in Beziehung zu KMeans, Bi-Sec-KMeans und Formale Begriffsanalyse. Weiterhin werden die unterschiedlichen Eigenschaften der Verfahren herausgearbeitet.

5.1 Cluster und Clusterung

Clustern beschreibt das (automatische) Gruppieren ähnlicher Objekte. Das Ergebnis des Clusterns ist eine Clusterung, die wir mit $\mathbb{P}$ bezeichnen. Sie besteht aus einer Menge von Clustern P. Jeder Cluster besteht aus einer Menge von Objekten D. Man bezeichnet Cluster auch als Segmente und den Vorgang des Clusterns auch als Segmentierung. Objekte eines Clusters sollten sich ähnlich und unähnlich zu Objekten anderer Cluster sein. Üblicherweise sind Clusterungen besser, wenn die Objekte innerhalb eines Clusters ähnlicher und zwischen den Clustern unähnlicher sind. Auf Maße zur Berechnung der Clustergüte gehen wir in Kapitel 5.3 ein.

Wir definieren Cluster und Clusterung unabhängig von den Beziehungen der Objekte zueinander wie folgt:

Definition 2. *Ein Cluster P ist eine Teilmenge der Objektmenge D. Eine Clusterung $\mathbb{P}$ ist eine Menge von Clustern.*

Für eine nicht überlappende Clusterung $\mathbb{P}$ einer Objektmenge D gilt: $\bigcup_{P \in \mathbb{P}} P = D$ und $\bigcap_{P \in \mathbb{P}} P = \emptyset$, d.h., dass alle Objekte mindestens einem und nur einem Cluster zugeordnet werden dürfen. KMeans und Bi-Sec-KMeans berechnen nicht überlappende Cluster.

Die formale Definition beschreibt die Mengenbeziehung zwischen den Objekten, dem Cluster und der Clusterung. Sie geht nicht auf die Berechnung der Cluster anhand von Objekteigenschaften ein. Eine automatische Berechnung von Clustern erfolgt im Allgemeinen durch einen Clusteralgorithmus. Die Algorithmen bauen zur Berechnung der Cluster auf objektbeschreibenden Merkmalen auf. Auf deren Basis lassen sich Beziehungen zwischen den Objekten berechnen. Die Objektbeziehungen, typischerweise Ähnlichkeiten zwischen den Objekten, können auch direkt angegeben werden[1]. Das Distanz- oder Ähnlichkeitsmaß berechnet aus den Merkmalen die Objektbeziehung. Alle Informationen fließen in den Algorithmus zur Berechnung der Clusterung ein. Im Ergebnis erhält man eine Clusterung, wobei die Cluster bestimmte Eigenschaften im Merkmalsraum aufweisen. Auf die Eigenschaften der Cluster nehmen während der Berechnung die Merkmale, Maße und der Algorithmus Einfluss und verändern auf diesem Wege die resultierende Clusterung.

Die Clusterverfahren haben die Aufgabe, mit Hilfe der Merkmale und der Ähnlichkeitsmaße die Cluster zu berechnen. Allerdings variiert die Vorstellung eines Clusterergebnisses von Anwender zu Anwender. Man kann mit Hilfe der Merkmale und der Ähnlichkeitsmaße Einfluss auf die Ergebnisse der Clusteralgorithmen nehmen und so die Clusterung steuern. Inwieweit das Ergebnis des Clusterverfahrens sich mit den Vorstellungen des Anwenders deckt, kann man mit Evaluierungsmaßen berechnen. Die verschiedenen Verfahren können Cluster mit unterschiedlichen Eigenschaften bezüglich des Merkmalsraumes unterschiedlich gut im Sinne des Anwenders berechnen. Steinbach u.a. illustrieren dies in [207]. Sie stellen verschiedene Gütekriterien zur Berechnung von Clustern vor, die ihrerseits die verschiedenen Clusterverfahren und unterschiedliche Clusterformen (im Merkmalsraum) nach sich ziehen.

Wir folgen [207] und stellen verschiedene Sichtweisen der Cluster im Merkmalsraum vor. Dabei wollen wir keine formale Definition angeben, sondern umgangsprachlich die verschiedenen Clustertypen erläutern. Wir gehen von einer gegebenen Menge numerischer Merkmale aus. Weiterhin setzen wir voraus, dass die Beziehung zwischen den Objekten über Ähnlichkeits- oder Distanzmaße auf der Basis der gegebenen Merkmale berechnet wird. Wir unterscheiden die folgenden fünf Clustertypen:

Gut getrennte Cluster: Ein Cluster ist eine Menge von Objekten im Raum, so dass *jedes Objekt* im Cluster dichter zu jedem anderen Objekt in seinem Cluster ist als zu jedem Objekt in jedem anderen Cluster.

Die Cluster sind damit klar voneinander getrennt. Diese Herangehensweise an das Clustern ist eher eine idealtypische. Häufig wird man in der Praxis damit keine Lösung finden, da die Objekte im Raum so angeordnet sind, dass keine Clusterung diese Bedingung erfüllen kann. Viele Verfahren folgen daher der zentrumsbasierten Clustersicht.

Zentrumsbasierte Cluster: Ein Cluster ist eine Menge von Objekten im Raum, so dass *jedes Objekt* innerhalb eines Clusters dichter zu seinem Zentrum ist als zum Zentrum jedes anderen Clusters. Meistens nimmt man den so genannten Zentroid (den Durchschnitt über alle Objekte des Clusters, siehe 4.5.2) oder den Median (das repräsentativste Objekt) des Clusters als Zentrum.

[1] Nicht jeder Algorithmus kann mit direkten Distanz- bzw. Ähnlichkeitsmaßen arbeiten. Daher ist dies nicht immer sinnvoll. In dieser Arbeit gehen wir von einer Menge von Merkmalen für jedes Objekt aus.

Bei dieser Definition können Objekte eines Clusters dichter zu den Objekten eines anderen Clusters liegen als zu Objekten des eigenen Clusters. Dies kommt durch den neuen Bezugspunkt (Zentroid) zustande.

Kontinuierliche Cluster: Ein Cluster ist eine Menge von Objekten im Raum, so dass ein Objekt in einem Cluster dichter zu einem oder mehreren Objekten des eigenen Clusters ist als zu jedem Objekt, das nicht im Cluster liegt.

Damit schafft man die Möglichkeit, auch nichtkonvexe Strukturen mittels Clusterverfahren zu entdecken.

Dichte-basierte Cluster: Ein Cluster besteht aus einer Menge von Objekten, die eine dichte Region im Raum bilden und durch Regionen mit geringerer Dichte von anderen Regionen mit hoher Dichte getrennt werden.

Im Unterschied zur kontinuierlichen Clusterdefinition können dichte-basierte Clusterfahren auch mit Ausreißern und Rauschen umgehen. Verfahren, die der kontinuierlichen Clusterdefinition folgen, wären bei Anwesenheit von Rauschen nicht in der Lage, Cluster zu identifizieren.

Ähnlichkeitsbasierte Cluster: Ein Cluster ist eine Menge von "ähnlichen" Objekten. Die Objekte der anderen Cluster sind nicht "ähnlich". Diese Definition zielt auf lokale Eigenschaften, die bei jedem Cluster hervorgehoben werden, ab.

Was man unter einem Cluster versteht, hängt im Endeffekt vom Anwender ab. Die Fülle der Clustersichten spiegelt die unterschiedlichen Sichten der Anwender wieder. Welcher Clustertyp in einem Datensatz enthalten ist, kann man vom Anwender nur bei bis zu 3-dimensionalen Merkmalsräumen, die leicht zu visualisieren sind, erfahren bzw. erfragen.

Die in der Arbeit verwendeten Verfahren KMeans und Bi-Sec-KMeans berechnen zentrumsbasierte Cluster. Beide Verfahren nutzen eine Distanzfunktion zur Berechnung der Objektbeziehung. Wir stellen zwei häufig verwendete Maße im nächsten Abschnitt vor.

5.2 Distanz- und Ähnlichkeitsmaße

Wir werden in diesem Kapitel zwei gängige Ansätze zur Berechnung von Distanzen und Ähnlichkeiten einführen, so wie man sie auch in vielen Lehrbüchern der Statistik und des Data Mining findet (vgl. [129, 58]). Wir starten mit der Minkowski Metrik und leiten aus ihr die euklidische Metrik ab. Dieses Distanzmaß bildet die Basis für das KMeans Clusterverfahren. Bei den Ähnlichkeitsmaßen gehen wir auf das Kosinus-Maß ein. Beide Maße werden in Kombination mit KMeans verwendet. Während die euklidische Distanz im Bereich des Data Mining eingesetzt wird, findet das Kosinus-Maß vorwiegend Anwendung im Bereich Text Mining.

5.2.1 Minkowski-Metrik

Eine der bekanntesten Distanzmaße ist die Minkowski-Metrik. Sie ist definiert als:

$$dist(\vec{x}, \vec{y}) = \sqrt[r]{\sum_{i=1}^{n} |x_i - y_i|^r} \tag{5.1}$$

wobei $\vec{x}$ und $\vec{y}$ zwei Vektoren in einem n-dimensionalen Raum sind und $r \in R^+$. Jede Metrik muss nicht negativ, reflexiv und symmetrisch sein und die Dreiecksungleichung erfüllen. Die aus der Minkowski-Metrik abgeleiteten Metriken bezeichnet man häufig als L_r-Norm. Die zwei bekanntesten Metriken sind:

1. für $r = 1$ die L_1-Norm oder auch Manhattan- oder City-Block-Distanz:

$$dist(\vec{x}, \vec{y}) = \sum_{i=1}^{n} |x_i - y_i|$$

und

2. für $r = 2$ die L_2-Norm oder euklidische Distanz:

$$dist(\vec{x}, \vec{y}) = \sqrt[2]{\sum_{i=1}^{n} |x_i - y_i|^2} \ .$$

Die L_1-Norm oder auch Manhattan-Metrik bezieht ihren Namen aus der Analogie zu Manhattan. Die Straßen in Manhattan verlaufen nur in Nord-Süd oder in Ost-West Richtung. Um von einer Ecke eines Häuserblockes zu dessen diagonal gelegenen Ecke zu gelangen, muss man den Straßen folgen und kann nicht den kürzesten Weg entlang der Diagonalen wählen. Die L_1-Norm berechnet die Entfernung zwischen zwei Punkten auf die gleiche Weise.

5.2.2 Kosinus-Maß

Im Folgenden wollen wir auf das im Bereich Text Mining häufig verwendete Kosinus-Maß ("cosine similarity") eingehen. Zur Berechnung der Ähnlichkeit zweier Vektoren $\vec{x}, \vec{y}$ bestimmt man das normalisierte innere Produkt, welches dem Kosinus des Winkels zwischen den beiden Vektoren entspricht:

$$\cos(\sphericalangle(\vec{x}, \vec{y})) = \frac{\vec{x} \cdot \vec{y}}{\parallel \vec{x} \parallel \cdot \parallel \vec{y} \parallel} = \frac{\sum_{i=1}^{n} (x_i \cdot y_i)}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}} \ . \tag{5.2}$$

Zur Veranschaulichung geben wir das Kosinus-Maß in entsprechender Textvektor-Schreibweise wieder. Zusätzlich sind wir nicht an einem Ähnlichkeits- sondern an einem Distanzmaß interessiert, das wir für das KMeans-Verfahren benötigen. Die Kosinus-Distanz zweier Termvektoren $\vec{t_{d1}}, \vec{t_{d2}}$ der Dokumente d_1, d_2 ergibt sich zu:

$$dist(\vec{t_{d1}}, \vec{t_{d2}}) = 1 - \cos(\sphericalangle(\vec{t_{d1}}, \vec{t_{d2}})) = 1 - \frac{\sum_{t \in \mathcal{T}} (\text{tf}(d_1, t) \cdot \text{tf}(d_2, t))}{\sqrt{\sum_{t \in \mathcal{T}} \text{tf}(d_1, t)^2} \cdot \sqrt{\sum_{t \in \mathcal{T}} \text{tf}(d_2, t)^2}} \tag{5.3}$$

wobei $\text{tf}(d, t)$, wie in Kapitel 4.1 ausgeführt, der Häufigkeit des Terms t im Dokument d entspricht. An dieser Stelle wollen wir noch auf den Zusammenhang zwischen Euklid-Metrik und Kosinus-Maß hinweisen.

Unter der Voraussetzung, dass die Länge der Vektoren $\vec{t_{d1}}$ und $\vec{t_{d2}}$ auf eins normiert wird, gilt:

$$dist_{cos}(\vec{t_{d1}}, \vec{t_{d2}}) = \frac{dist_{Euklid}(\vec{t_{d1}}, \vec{t_{d2}})^2}{2} \tag{5.4}$$

wobei $dist_{cos}$ und $dist_{Euklid}$ die Kosinus bzw. Euklid-Distanz berechnen.

5.3 Evaluierung von Clusterergebnissen — Clustergüte, Clusteranzahl und Clustervergleich

In diesem Abschnitt wollen wir die Evaluierung von Clusterergebnissen diskutieren. Dazu gehen wir als Erstes auf prinzipielle Fragen in Abschnitt 5.3.1 ein. Abschnitt 5.3.2 beschreibt die Bestimmung der Clusteranzahl und in Abschnitt 5.3.3 und Abschnitt 5.3.4 diskutieren wir vergleichende und statistische Maßzahlen zur Beurteilung der Clustergüte. Wir beenden das Kapitel mit einer Zusammenfassung in Abschnitt 5.3.5.

5.3.1 Methodik

Die Evaluation von Clusterergebnissen gestaltet sich prinzipiell schwierig. Die Aufgabe eines Clusterverfahren ist es, ohne apriori Wissen mit Hilfe von Abstandsmaßen Gruppen von Objekten zu bilden. Die Objekte einer Gruppen sollen gemäß des gewählten Maßes ähnlich/gleich sein und sich von den Objekten der anderen Gruppen unterscheiden. Neben der Wahl des richtigen Maßes ist das Berechnen der Gruppen ein zentrales Problem des Clusterns. In dieser Arbeit liegt der Schwerpunkt auf der Auswahl einer geeigneten Repräsentation der Objekte. Die Evaluierung soll uns bei der Beurteilung der Clusterergebnisse und damit bei der Auswahl der besten Repräsentation gemäß des gewählten Maßen unterstützen.

Dem Clusterverfahren stehen, wie gesagt, keine Informationen des Anwenders über die als ähnlich angesehenen Objekte zur Verfügung. Der Anwender erwartet aber trotzdem, dass die Gruppen entsprechend der eigenen Vorstellung gebildet werden. Das Clusterverfahren kann diese Aufgabe prinzipbedingt nicht “perfekt” erfüllen, da es unsere Vorstellungen nicht erraten kann. Man kann dem Clusterverfahren nun helfen, indem man die Repräsentation der Objekte geeignet wählt. An dieser Stelle setzt auch der hier vorgestellte Ansatz, Hintergrundwissen in die Repräsentation der Objekte zu integrieren, an.

Durch die fehlenden Informationen über die Ziele des Clusterns, bzw. deren sehr generischen Beschreibung, ergeben sich auch Probleme, die Güte der gefundenen Cluster zu bestimmen. Damit wird auch die Evaluierung schwierig. Stehen mehr Informationen für die Gruppierungsaufgabe zur Verfügung, d.h. der Anwender gibt z.B. Einteilungen bekannt, so ändert sich die Aufgabe und man spricht nicht mehr vom Clustern sondern vom überwachten Lernen bzw. vom Klassifizieren.

Für die Evaluierung von Clusterergebnissen findet man in der Literatur zwei Ansätze. Man kann eine vom Menschen gegebene Gruppierung mit der Clusterlösung vergleichen. Dabei nimmt man an, dass die zum Test herangezogenen Datensätze und deren Gruppierung im Allgemeinen eine gute Einteilung widerspiegeln. Die Alternative bilden statistische Maßzahlen, die beurteilen, wie gut bestimmte statistische Eigenschaften erfüllt werden. Diesem Ansatz unterliegt die Annahme, dass die statistische Maßzahl die Anforderungen aus der Anwendung gut widerspiegelt.

Der erste Ansatz setzt bekannte apriori Gruppen zur Evaluierung voraus. Man kann diese nutzen, um die Güte des Clusterverfahrens zu bestimmen. Unter der Annahme, dass das Clusterverfahren die vorgegebenen Gruppen berechnen sollte, ist eine solche Evaluierungsstrategie sinnvoll. Dazu

nutzt man die Informationen über die Klassenzugehörigkeit nicht zum Clustern, sondern nur zur Überprüfung der Clustergüte. Berechnet man die gleiche Anzahl von Clustern wie auch vorgegebene Klassen vorhanden sind, so würde idealer Weise jede Klasse genau einem Cluster entsprechen. Um zu ermitteln, welcher Cluster welcher Klasse entspricht, schaut man, welcher Klasse die Objekte eines Clusters angehören und benennt den Cluster mit dem Namen der Klasse, die am häufigsten vorkommt. Man ermittelt für jeden Cluster das so genannte "Label" oder den Bezeichner [123]. Auf die korrekte Bestimmung der Clusteranzahl gehen wir im nächsten Abschnitt 5.3.2 ein. Abschnitt 5.3.3 beschäftigt sich mit vergleichenden Maßen zur Evaluierung von Clusterergebnissen.

Die zweite Alternative besteht in der Berechnung von statistischen Maßzahlen, die eine Aussage über das gefundene Clusterergebnis zulassen. Fickel stellt in [73] für verschiedene Clusteralgorithmen entsprechende statistische Gütemaße zur Verfügung. Damit kann man die Güte der vorhandenen Clusterlösung sowie den Einfluss der einzelnen Variablen messen. Weitere Beispiele für statistische Maßzahlen findet man z.B. in [226] oder [129]. In gewisser Weise beurteilen alle diese Maßzahlen nur die Güte der Clusteralgorithmen. Leider wird die Annahme, dass dies auch eng mit der Anwendung korreliert ist, nicht immer erfüllt. Dies kann dazu führen, dass Clusterergebnisse laut statistischem Maß besser sind, aber aus Sicht der Anwendung schlechter. Die Maße sind statistisch gesehen begründet, erlauben aber keine Rückschlüsse auf die Güte der Lösung aus Sicht der betriebswirtschaftlichen Anwendungen. Auf Maße aus diesem Bereich gehen wir in Abschnitt 5.3.4 ein.

Ein alternativer Ansatz, der auch auf bekannten manuell erstellten Labels aufbaut, sieht die Clusteraufgabe als Vorverarbeitungsschritt einer Klassifikationsaufgabe. In diesem Fall kann man die Verbesserung der Klassifikationsgüte des zweiten Schrittes zur Evaluierung des Clusterverfahrens nutzen. Berechnet die Clusterung neue und nützliche Merkmale für den Klassifikationsschritt, so steigt die Güte des gesamten Prozesses. Man evaluiert so indirekt das Clusterverfahren [127].

Wir werden uns im folgenden Abschnitt mit der Bestimmung der Clusteranzahl auseinander setzen. Dieses Problem ist eng verbunden mit der Wahl des richtigen Evaluierungsmaßes. Hätten wir ein solches Maß zur Verfügung, wäre es leicht, die korrekte Anzahl zu berechnen. Leider gibt es so ein Maß nicht, so dass man auch nicht die optimale Anzahl an Clustern berechnen kann.

5.3.2 Clusteranzahl

Die Bestimmung der Clusteranzahl ist ein bisher ungelöstes Problem aus Sicht der Praxis. In der Literatur existiert eine Reihe von Maßen, so genannte Indizes, die zur Berechnung der Clustergüte und damit zur Bestimmung der Clusteranzahl herangezogen werden können [166]. Diese haben häufig den Nachteil, dass sie nicht auf große Datenmengen angewendet werden können und nicht unbedingt die Güte aus Sicht des Anwenders beurteilen. Man ist zwar in der Lage, mit Hilfe von verfahrensspezifischen Zielfunktionen bei einem gegeben Clusterverfahren und entsprechenden Parametern eine Clusteranzahl zu berechnen, damit wird das Problem aber nur auf die Angabe einer adäquaten Zielfunktion bzw. die Schätzung der korrekten Parameter verlagert. Z.B. berechnen dichte-basierte Verfahren (siehe Abschnitt 5.6.7) automatisch die Anzahl der Cluster. Dafür muss die Dichte vorgegeben werden. Weiterhin spielt die Auswahl des passenden Clusterverfahrens für die gegebene Aufgabe eine wesentliche Rolle. Die letztendliche Entscheidung über die korrekt bestimmte Anzahl der Cluster obliegt dem Anwender. Aus diesem Grund ist man nicht in der Lage, ohne zusätzliche Informationen des Anwenders die Clusteranzahl korrekt automatisch zu ermitteln.

In der Literatur findet man eine große Anzahl an Maßen zur Berechnung der Clustergüte, die sich auch zur Abschätzung der Clusteranzahl eignen (vgl. [129, 166, 73]. Der in Abschnitt 5.3.4.2 eingeführte Silhouetten-Koeffizient eignet sich ebenfalls zur Bestimmung der Clusteranzahl, da der Silhouetten-Koeffizient unabhängig von der Clusteranzahl k ist. Man geht dazu wie folgt vor und

berechnet:

$$\overline{SC} = \max_{k=2,3,\ldots,n-1} SC(\mathbb{P}_k) \tag{5.5}$$

wobei $\mathbb{P}_k$ die Clusterung mit k Clustern ist. Gleichung 5.5 berechnet für alle $n-2$ möglichen Clusterungen den Silhouetten-Koeffizienten. Anschließend wählt man die Clusterung mit dem größten Silhouetten-Koeffizienten. Sie hat laut Maßzahl die Struktur am besten bestimmt. Im Bereich der Hierarchischen Clusterverfahren findet die Bestimmung der Clusteranzahl oft mittels des Ellenbogenkriteriums statt. Dazu vergleicht man z.B. die Innnerklassenvarianz der einzelnen Clusterungen und wählt die Clusterung, ab der die Innerklassenvarianz (im agglomerativen Fall) deutlich ansteigt (vgl. [15]).

In dieser Arbeit verfolgen wir nicht das Ziel, eine optimale Anzahl an Clustern mittels Indizes oder anderen Zielfunktionen automatisch zu bestimmen. Vielmehr fordern wir vom Anwender die Angabe einer für ihn sinnvollen Anzahl an Clustern und präsentieren ihm die Ergebnisse durch die Nutzung von Hintergrundwissen und entsprechender Visualisierungstechniken in verständlicher Form.

5.3.3 Vergleichende Maßzahlen

Zur Evaluierung der Clüstergüte existieren zwei prinzipiell unterschiedliche Ansätze (siehe Abschnitt 5.3.1). In diesem Abschnitt stellen wir Maße vor, die die Güte der Clusterung anhand einer gegebenen Klassifikation berechnen. Diese Maße werden häufig im Bereich Information Retrieval eingesetzt. Wir beschreiben die Maße Precision, Recall, Purity, F-Measure und Entropy.

5.3.3.1 Precision und Recall

Die klassischen Maße des Information Retrieval sind Precision π, die Präzision oder Genauigkeit und Recall ρ, die Vollständigkeit. Sie dienen zur Schätzung der Effektivität des verwendeten Klassifikations- oder Clustermodells. Da keine objektiven Maße zur Verfügung stehen, vergleicht man zur Beurteilung der Güte von Klassifikationsmodellen eine Klassifikation mit einer anderen. Eine der beiden Klassifikationen, die wir im folgenden mit $\mathbb{L}$ abkürzen, bildet die Basis des Vergleiches. Sie repäsentiert im Normalfall die Meinung eines Experten und wird oft manuell erstellt (vgl. [195, 232]). Die andere Klassifikation, die in dieser Arbeit i.A. einer Clusterung entspricht, kürzen wir mit $\mathbb{P}$ ab.

Gegeben sei die Klasse L und eine Menge von Objekten, z.B. Dokumente. Die Aufgabe des Experten und des Modells (in unserem Fall des Clusterverfahrens) ist es, Dokumente dieser Menge der Klasse L zuzuordnen oder nicht. Tabelle 5.1 gibt die möglichen Fälle dieses Zweiklassenproblems wieder, die bei der Zuordnung der Dokumente zu den Klassen auftreten können. TP_L (true positives) entspricht der Menge an Dokumenten, die das Modell und der Experte der gleichen Klasse L zugeordnet haben. FP_L (false positives) gibt die Menge an Dokumenten an, die vom Modell fälschlicher Weise der Klasse L zugeordnet wurden, FN_L (false negatives) die fälschlicher Weise nicht der Klasse L zugeordnet wurden und TN_L (true negatives) die korrekter Weise nicht der Klasse L zugeordnet wurden. Tabelle 5.1 vergleicht auf diese Weise die Meinung des Experten mit dem Modell.

Precision π und Recall ρ in Bezug zur Klasse L berechnen sich wie folgt:

$$\pi(L) := \frac{|TP_L|}{|TP_L| + |FP_L|}, \tag{5.6}$$

Klasse L		Experten-Urteil	
		YES	NO
Modell-Urteil	YES	TP_L	FP_L
	NO	FN_L	TN_L

Tabelle 5.1: Kontingenztabelle für Klasse L

$$\rho(L) := \frac{|TP_L|}{|TP_L| + |FN_L|}. \tag{5.7}$$

Die Precision in Bezug auf die Klasse L ist definiert als die bedingte Wahrscheinlichkeit, dass die Entscheidung, ein zufällig gewähltes Dokument d in Klasse L zu klassifizieren, korrekt ist. Analog lässt sich der Recall als bedingte Wahrscheinlichkeit definieren. Der Recall gibt die Wahrscheinlichkeit an, mit der ein zufällig gewähltes Dokument, das zur Klasse L gehören sollte, auch in diese klassifiziert wird. Die Wahrscheinlichkeiten für Precision und Recall kann man mit Hilfe der Kontingenztabelle 5.1 nach Gleichung 5.6 bzw. 5.7 schätzen. Die klassenbezogenen Werte einer Klassifikation $\mathbb{L}$ können in folgender Weise gemittelt werden.

Mikrodurchschnitt (microaveraging): Bei der Mikrodurchschnittsbildung ("microaveraging") summiert man über die jeweiligen Einzelentscheidungen:

$$\pi^{\mu}(\mathbb{L}) := \frac{|TP|}{|TP| + |FP|} = \frac{\sum\limits_{L \in \mathbb{L}} |TP_L|}{\sum\limits_{L \in \mathbb{L}} (|TP_L| + |FP_L|)} \tag{5.8}$$

$$\rho^{\mu}(\mathbb{L}) := \frac{TP}{TP + FN} = \frac{\sum\limits_{L \in \mathbb{L}} |TP_L|}{\sum\limits_{L \in \mathbb{L}} (|TP_L| + |FN_L|)} \tag{5.9}$$

wobei das "μ" für Mikrodurchschnitt steht.

Makrodurchschnitt (makroaveraging): In diesem Fall werden erst für jede Klasse die Precision und Recall Werte berechnet, bevor die Durchschnittsbildung erfolgt:

$$\pi^{M}(\mathbb{L}) := \frac{\sum\limits_{L \in \mathbb{L}} \pi(L)}{|\mathbb{L}|} \tag{5.10}$$

$$\rho^{M}(\mathbb{L}) := \frac{\sum\limits_{L \in \mathbb{L}} \rho(L)}{|\mathbb{L}|} \tag{5.11}$$

wobei das "M" für Makrodurchschnitt steht.

Man kann die Mikrodurchschnitts- und die Makrodurchschnittsbildung jeweils als gewichtetes bzw. und ungewichtetes Mittel betrachten. Dies führt zu Ergebnissen, die einer unterschiedlichen Interpretation bedürfen. Bei der Mikrodurchschnittsbildung steht die korrekte Klassifikation eines jeden Dokumentes im Vordergrund. Die Makrodurchschnittsbildung bewertet die Gesamtgüte des Modells klassenbezogen und unabhängig von der Klassengröße. Dies wirkt sich besonders bei sehr unterschiedlich großen Klassen aus. Welche der beiden Durchschnitte man nutzt, hängt ganz von

der Anwendung ab. Die Nutzung von Precision und Recall wird im Bereich der Text-Klassifikation für nicht sinnvoll erachtet (vgl. [195]). Wir stellen das Kombinationsmaß F-Measure, welches sich auch für das Clustering adaptieren lässt, in Abschnitt 5.3.3.3 vor. Vergeben sowohl Experte als auch Modell nur ein Label pro Dokument, dann gilt $\pi^{\mu} = \rho^{\mu}$ (vgl. [198]). Die Ergebnisse hängen von der Clusteranzahl ab, können aber zum Vergleich verschiedener Verfahren bei konstanter Clusteranzahl eingesetzt werden. Im nächsten Kapitel stellen wir das aus der Precision abgeleitet Purity-Maß vor und definieren auch den Gegenspieler – die InversePurity.

5.3.3.2 Purity und InversePurity

Das *Purity* Maß basiert auf dem aus dem Information Retrieval bekannten Precision Maß (vgl. [206]). Wir folgen bei der Definition des Maßes Steinbach [206] und vergleichen es anschließend mit der Precision aus Abschnitt 5.3.3.1.

Gegeben seien die beiden Partitionierungen $\mathbb{P}$ und $\mathbb{L}$, wobei $\mathbb{P}$ die Partitionierung des Clusterverfahrens und $\mathbb{L}$ die zum Vergleich zur Verfügung stehende Partitionierung ist. Letztere wird typischerweise von Experten erstellt. Die Precision $\pi(P, L)$ eines Clusters $P \in \mathbb{P}$ für eine gegebene Kategorie $L \in \mathbb{L}$ berechnet man folgendermaßen:

$$\pi(P, L) := \frac{|P \cap L|}{|P|} \tag{5.12}$$

wobei $\pi(L) = \pi(P, L)$ gilt. Die Menge $P \cap L$ entspricht dabei der Menge TP_L aus Formel (5.6), wobei die Klasse L aus Abschnitt 5.3.3.1 dem Cluster P entspricht. Der Recall wird wie folgt berechnet:

$$\rho(P, L) := \frac{|L \cap P|}{|L|} \tag{5.13}$$

Es sei noch angemerkt, dass $\pi(P, L) = \rho(L, P)$ gilt. Der Purity-Wert für die gesamte Clusterung $\mathbb{P}$ wird mit Hilfe der gewichteten Summe aller Precision-Werte berechnet:

$$\text{Purity}(\mathbb{P}, \mathbb{L}) := \sum_{P \in \mathbb{P}} \frac{|P|}{|D|} \max_{L \in \mathbb{L}} \pi(P, L) \tag{5.14}$$

und bildet durch die Summation über die Cluster ein ergänzendes Maß zur Mikrodurchschnittsbildung, bei der über die gegebenen Klassen summiert wird.

Es sei an dieser Stelle noch auf zwei Dinge hingewiesen. Der Purity-Wert für die Clusterung wird in [231] auch mit “accuracy” (Genauigkeit) bezeichnet. Das Purity Maß bevorzugt Clusterungen mit vielen Klassen. Eine perfekte Clusterung erhält man, wenn die Anzahl der Cluster gleich der Anzahl der Dokumente ist.

In Analogie zu den beiden Maßen Precision und Recall, die Gegenspieler darstellen, definieren wir analog zu Purity die *InversePurity* wie folgt:

$$\text{InversePurity}(\mathbb{P}, \mathbb{L}) := \sum_{L \in \mathbb{L}} \frac{|L|}{|D|} \max_{P \in \mathbb{P}} \pi(L, P) \tag{5.15}$$

Im Unterschied zur Purity summiert die InversePurity nicht über die Cluster, sondern über die vorgegebenen Kategorien. Damit ist sie identisch mit der Mikrodurchschnittsbildung für Recall aus Abschnitt 5.3.3.1. Die folgenden Überlegungen sollen die Gleichheit der beiden Maße zeigen. Wir vereinfachen Gleichung 5.15 wie folgt:

$$\text{InversePurity}(\mathbb{P},\mathbb{L}) := \sum_{L\in\mathbb{L}} \frac{|L|}{|D|} \max_{P\in\mathbb{P}} \frac{|L\cap P|}{|L|} = \frac{\sum_{L\in\mathbb{L}} \max_{P\in\mathbb{P}} |L\cap P|}{|D|} = \frac{\sum_{L\in\mathbb{L}} TP_L}{\sum_{L\in\mathbb{L}}(TP_L + FN_L)} \tag{5.16}$$

Der Nenner ist wegen $TP_L + FN_L = L$ identisch und entspricht der Anzahl der Dokumente $|D|$. Weiterhin muss für die Gleichheit des Zählers gelten: $\max_{P\in\mathbb{P}} |L\cap P| = TP_L$. Die Idee der Inverse-Purity ist dem Recall-Maß ähnlich und bewertet den Cluster am besten, der die meisten Dokumente einer vorgegeben Kategorie enthält. Wir unterstellen, dass der Anwender bei guter Clusterung in der Lage ist, den Cluster mit den meisten Dokumenten einer vorgegeben Kategorie L zu identifizieren und so ein Label für den Cluster P zu vergeben. Die resultierende Schnittmenge $P \cap L$ für den gewählten Cluster entspricht genau TP_P was wiederum in diesem Fall gleich TP_L ist.

5.3.3.3 F-Measure

Die Maße Precision und Recall sollten für die Evaluierung gemeinsam benutzt werden. Dazu werden Verknüpfungen vorgeschlagen. Die F_β Funktion von [188] ist die wohl bekannteste. Sie ist wie folgt definiert:

$$F_\beta(P,L) = \frac{(\beta^2+1)\pi(P,L)\rho(P,L)}{\beta^2\pi(P,L) + \rho(P,L)}. \tag{5.17}$$

und liefert eine ganze Klasse von Verknüpfungsfunktionen. Der Parameter β gewichtet den Einfluss der beiden Maße Precision und Recall zueinander. Wählt man $\beta = 0$, dann entspricht F_β der Precision. Bei $\beta = +\infty$ ist F_β gleich dem Recall. Normalerweise wird $\beta = 1$ gewählt. Precision und Recall sind dann gleich gewichtet.

Für die Bewertung von Clusterergebnissen kommt es wieder auf die Art der Summierung an. Gewöhnlich berechnet man das F-Measure (in diesem Fall das F_1) der Clusterung als gewichtetes Mittel:

$$F_1(\mathbb{P},\mathbb{L}) := \sum_{L\in\mathbb{L}} \frac{|L|}{|D|} \max_{P\in\mathbb{P}} \frac{2\cdot\rho(P,L)\cdot\pi(P,L)}{\rho(P,L)+\pi(P,L)} \tag{5.18}$$

5.3.3.4 Die Entropie als Evaluationsmaß

Während die (Inverse) Purity und F-Measure die "besten" Treffer zwischen Cluster und manuell definierten Kategorien berücksichtigen, berechnet die Entropie, wie groß der Informationsgehalt der Clusterung ist.

Zur Berechnung der Entropy muss man die bedingte Wahrscheinlichkeit $Prob(L|P)$, dass ein Objekt des Clusters P zur Kategorie L gehört, schätzen. Die Entropie des Clusters P berechnet sich wie folgt:

$$E(P,\mathbb{L}) = -\sum_{L\in\mathbb{L}} Prob(L|P)\cdot log(Prob(L|P)) \tag{5.19}$$

Die Gesamtentropie der Clusterung $\mathbb{P}$ in Bezug auf $\mathbb{L}$ ergibt sich zu:

$$E(\mathbb{P},\mathbb{L}) = \sum_{P\in\mathbb{P}} Prob(P)\cdot E(P,\mathbb{L}), \tag{5.20}$$

wobei $Prob(L|P)$ mit der Precision $\pi(P, L)$ geschätzt wird und $Prob(P) = \frac{|P|}{|D|}$ ist. Eine Entropie von Null zeigt den besten Wert an (vgl. [206, 30].

5.3.4 Statistische Maßzahlen

Die folgenden Abschnitte betrachten Maße, die nicht auf eine vorgegebene Klassifizierung $\mathbb{L}$ der Objekte zurückgreifen können. Sie werden in der Literatur zur Statistik auch Indizes genannt. Sie bewerten anhand statistischer Zusammenhänge die Güte einer Clusterung. Man findet in der Literatur eine große Anzahl an Indizes (vgl. [166, 15, 73, 58]). Eines der bekanntesten Maße ist der mittlere quadratische Fehler. Er erlaubt, Aussagen über die Güte der gefundenen Cluster in Abhängigkeit von der Clusteranzahl zu machen, wobei die Ergebnisse besser werden, je höher die Anzahl ist. In [129] wird ein alternatives Maß, der Silhouetten-Koeffizient, der unabhängig von der Clusteranzahl ist, vorgestellt. Beide Maße führen wir im Folgenden ein.

5.3.4.1 Mittlerer quadratischer Fehler

Hält man die Anzahl der Dimensionen und die Anzahl der Cluster konstant, so kann man den mittleren quadratischen Fehler (Mean Square Error, MSE) ebenfalls zur Beurteilung der Güte von Clusterungen heranziehen. Der mittlere quadratische Fehler ist ein Maß für die Kompaktheit der Clusterung und ist wie folgt definiert:

Definition 3 (MSE). *Der gesamte mittlere quadratische Fehler (MSE) für eine gegebene Clusterung $\mathbb{P}$ ist definiert als*

$$MSE(\mathbb{P}) = \sum_{P \in \mathbb{P}} MSE(P), \tag{5.21}$$

wobei der mittlere quadratische Fehler für einen Cluster P wie folgt berechnet wird:

$$MSE(P) = \sum_{d \in P} dist(d, \mu_P)^2, \tag{5.22}$$

und μ_P der Zentroid (siehe Abschnitt 4.5.2) des Clusters P ist.

5.3.4.2 Der Silhouetten-Koeffizient

Eines der wenigen von der Anzahl der Cluster unabhängigen Maße zur Beurteilung der Clustergüte ist der Silhouetten-Koeffizient. Wir folgen bei der Darstellung des Koeffizienten [129], Seite 87ff:

Definition 4 (Silhouetten-Koeffizient). *Sei $\mathbb{P}$ eine Clusterung einer Menge von Objekten D z.B. Dokumenten. Die Distanz zwischen einem Objekt $d \in D$ und einem Cluster $P \in \mathbb{P}$ wird wie folgt berechnet:*

$$dist(d, P) = \frac{\sum_{p \in P} dist(d, p)}{\mid P \mid}. \tag{5.23}$$

Weiterhin sei $a(d, \mathbb{P}) = dist(d, P)$ die Distanz von Objekt d zu seinem Cluster P $(d \in P)$ und $b(d, \mathbb{P}) = \min_{P \in \mathbb{P}, d \notin P} dist(d, P)$ die Distanz des Dokumentes d zum nächsten Cluster.

*Die **Silhouette** $s(d, \mathbb{P})$ eines Dokumentes $d \in D$ ist dann definiert als:*

$$s(d, \mathbb{P}) = \frac{b(d, \mathbb{P}) - a(d, \mathbb{P})}{\max\{a(d, \mathbb{P}), b(d, \mathbb{P})\}}. \tag{5.24}$$

*Der **Silhouetten-Koeffizient** $SC_P(\mathbb{P})$ eines Clusters $P \in \mathbb{P}$ ergibt sich zu:*

$$SC_P(\mathbb{P}) = \frac{\sum_{d \in P} s(d, \mathbb{P})}{\mid P \mid}. \tag{5.25}$$

*Der **Silhouetten-Koeffizient** $SC(\mathbb{P})$ der gesamten Clusterung ergibt sich zu:*

$$SC(\mathbb{P}) = \frac{\sum_{d \in D} s(d, \mathbb{P})}{\mid D \mid}. \tag{5.26}$$

In [129] wurde die euklidische Distanz (siehe Kapitel 5.2.1) für $dist(d, p)$ gewählt.

Mit Hilfe des Silhouetten-Koeffizienten ist man in der Lage, die Güte eines Clusters bzw. der gesamten Clusterung zu beurteilen (Details findet man in [129]). [129] nennt charakteristische Werte des Silhouetten-Koeffizienten zur Bewertung der Clusterqualität. Ein Wert für $SC(\mathbb{P})$ zwischen 0.7 und 1.0 signalisiert exzellente Separation zwischen den gefundenen Clustern, d.h. die Objekte innerhalb eines Clusters sind sehr dicht beieinander und liegen weit entfernt von anderen Clustern. Die Struktur wurde durch das Clusterverfahren sehr gut identifiziert. Für den Bereich von 0.5 bis 0.7 sind die Objekte klar den entsprechenden Clustern zugeordnet. Eine Menge Rauschen ist im Datensatz vorhanden, wenn der Silhouetten-Koeffizient im Bereich von 0.25 bis 0.5 liegt, wobei auch hier noch Cluster identifizierbar sind. Viele Objekte konnten in diesem Fall durch das Clusterverfahren nicht eindeutig einem Cluster zugeordnet werden. Bei Werten unter 0.25 ist es praktisch unmöglich, eine Clusterstruktur zu identifizieren und sinnvolle (aus Sicht der Anwendung) Clusterzentren zu berechnen. Das Clusterverfahren hat die Clusterung mehr oder weniger “erraten”.

5.3.5 Zusammenfassung

In Abschnitt 5.3 wurde das prinzipielle Vorgehen beim Evaluieren von Clusterung vorgestellt. Dabei wurde neben der Bestimmung der Clusteranzahl auch die vergleichende und statistische Evaluierung diskutiert. Grundsätzlich versucht man, mit Hilfe der vorgestellten Maße die Güte der Clusterverfahren zu beurteilen bzw. auch die Clusteranzahl abzuschätzen. Ein wichtiger Punkt ist die Einbeziehung des Anwenders in den Prozess. Nur der Anwender ist in der Lage, die Güte einer Clusterung zu beurteilen. Im folgenden Abschnitt werden wir das KMeans und Bi-Sec-KMeans Clusterverfahren einführen.

5.4 KMeans und Bi-Sec-KMeans

5.4.1 KMeans

KMeans ist eines der in der Praxis am häufigsten verwendeten Clusterverfahren im Bereich Data Mining und Statistik (vgl. [101]). Das ursprünglich aus der Statistik stammende Verfahren ist einfach zu implementieren und kann auch auf große Datenmengen angewendet werden. Es hat sich gezeigt, dass gerade im Bereich des Clusterns von Texten KMeans gute Ergebnisse erzielt. Ausgehend von einer Startlösung, in der alle Objekte auf eine vorgegebene Anzahl von Clustern verteilt werden, versucht man durch gezieltes Ändern der Zuordnung von Objekten zu den Clustern die Lösung zu verbessern. Mittlerweile existieren eine Reihe von Varianten, wobei das Grundprinzip auf Forgy 1965 [75] bzw. MacQueen 1967 [151] zurückgeht. In der Literatur zur Vektorquantisierung ist das

Verfahren auch unter dem Namen Lloyd-Max-Algorithmus bekannt ([82]).[2] Das Grundprinzip ist im Algorithmus 5.1 wiedergegeben.

Algorithmus 5.1 Der KMeans Algorithmus

Input: Menge D, Abstandsmaß $dist$, Anzahl k an Cluster
Output: Eine Partitionierung $\mathbb{P}$ der Menge D (wobei für Menge $\mathbb{P}$ mit k disjunkten Teilmengen aus D gilt: $\bigcup_{P\in\mathbb{P}} P = D$).

1: Wähle zufällig k Datenpunkte aus D als Ausgangszentroide $\vec{t_{P_1}} \ldots \vec{t_{P_k}}$.
2: **repeat**
3: Weise jedem Element aus D seinem nächsten Zentroid gemäß $dist$ zu.
4: Berechne die Clusterzentroide $\vec{t_{P_1}} \ldots \vec{t_{P_k}}$ der Cluster $P_1 \ldots P_k$ (erneut).
5: **until** Clusterzentroide $\vec{t_{P_1}} \ldots \vec{t_{P_k}}$ stabil.
6: **return** $\mathbb{P} := \{P_1, \ldots, P_k\}$, die Menge der Cluster.

KMeans besteht im Wesentlichen aus den Schritten drei und vier im Algorithmus, wobei man die Anzahl der Cluster k vorgeben muss. In Schritt drei werden die Objekte ihrem nächsten der k Zentroide zugeordnet. Schritt vier berechnet auf der Basis der neuen Zuordnungen die Zentroide neu. Wir wiederholen die beiden Schritte in einer Schleife (Schritt fünf), bis sich die Clusterzentroide nicht mehr ändern.

Der Algorithmus 5.1 entspricht einer einfachen Hill-Climbing-Prozedur, die typischerweise in einem lokalen Optimum stecken bleibt (das Finden des globalen Optimums ist ein NP-vollständiges Problem). Neben einer geeigneten Methode, die Startlösung zu bestimmmen (Schritt eins), benötigen wir ein Maß zur Berechnung der Distanz oder Ähnlichkeit in Schritt drei. Weiterhin kann das Abbruchkriterium der Schleife in Schritt fünf unterschiedlich gewählt werden.

Üblicherweise wird die quadrierte euklidische Distanz zur Berechnung der Abstände eingesetzt (siehe Kapitel 5.2). In der Literatur findet man auch häufig im Bereich des Text-Clusterns das Kosinus-Maß zur Berechnung der Ähnlichkeit der Objekte (z.B. [206] oder [144]). Die Clusterzentroide in Schritt vier berechnen wir nach Gleichung 4.3, welches dem Mittelwert über alle Dokumente pro Term entspricht. Man kann für die L_2-Norm zeigen, dass die durch die Iteration der Schritte drei und vier entstehende Folge die SQ in den Clustern minimiert:

$$SQ = \sum_{P\in\mathbb{P}} \sum_{d\in P} |\vec{t_d} - \vec{t_P}|^2 \quad . \tag{5.27}$$

Beim Abbruchkriterium existieren verschiedene Möglichkeiten. So stoppt der Algorithmus, wenn sich die Zentroide nicht mehr verändern oder die Zuordnung der Objekte zu den Clustern konstant bleibt. Beide Kriterien sind äquivalent und führen nach einer kleinen Anzahl von Iterationen (wesentlich kleiner als die Anzahl der Dokumente) zur Beendigung des Algorithmus. Zusätzlich kann man die Anzahl der Iterationen zählen und nach einer vorzugebenden Maximalzahl abbrechen. Häufig ist es sinnvoll, beide Kriterien zu kombinieren, da es Fälle gibt, in denen die Clusterlösung oszilliert und das Verfahren zwischen zwei oder mehr Lösungen variiert.

Kritisch auf die Lösungsgüte von KMeans wirkt sich die Startlösung aus, die u.a. bestimmt, in welchem lokalen Minimum der Algorithmus endet. Schon Duda und Hart führen in [59] aus, dass dies ein zentrales Problem von Hill-Climbing-Verfahren ist. Der bekannteste und defacto Standardansatz ist das wiederholte Starten des Verfahrens mit zufälligen Startlösungen (vgl. [59], [129]) und die Auswahl der Lösung mit dem kleinsten MSE. Hat man eine Vorstellung über die Lage der Cluster, kann man die Startlösung auch vorgeben, um eine bessere Clusterlösung zu erhalten.

[2] [88] liefert einen umfangreichen Überblick über Vektorquantisierung und deren Statistik.

Ansätze mit einem systematischen Vorgehen, wie dem Berechnen von Clusterlösungen auf kleinen Stichproben mittels agglomerativen hierarchischen Clusterverfahren, die dann als Startlösung für KMeans verwendet werden, untersuchte Milligan in [168]. Fortschrittlichere Ansätze findet man z.B. in [162, 219, 30, 67].

In der Literatur gibt es unzählige Varianten des KMeans. Auf der einen Seite wurde das Verfahren auf sehr große Datenmengen skaliert [29, 184, 183] und eine parallelisierte Variante [53] entwickelt. Auf der anderen Seite existieren zahlreiche Modifikationen inhaltlicher Art, die andere Clusterergebnisse bewirken. Eine der ältesten stammt von MacQueens 1967 [151]. Im Unterschied zum Forgy-Verfahren werden die Clustermittelwerte nach jeder Neuzuweisung eines Punktes neu berechnet. In verschiedenen Artikeln wurden sie auch zum Clustern von Texten eingesetzt [144, 206, 51]. Die Variante von PAM in [129] basiert statt auf dem Mittelwert auf dem Median. ISOData [16, 83] basiert zwar weiterhin auf dem Mittelwert, erweitert aber jede Iteration um die Anpassung der Clusteranzahl.

Im nächsten Abschnitt schauen wir uns nun den so genannten Bi-Sec-KMeans-Algorithmus an. Er kann prinzipiell auf allen KMeans-Varianten aufbauen. Wir nutzen im Folgenden den Forgy-Algorithmus mit dem Kosinus-Maß.

5.4.2 Bi-Sec-KMeans

Bi-Sec-KMeans ist wie KMeans ein sehr schnelles und effizientes Verfahren. Es ist in der Lage, große Datensätze wie z.B. den Reuters Datensatz zu verarbeiten. In [206] wird neben der hohen Geschwindigkeit auch die hohe Qualität der Ergebnisse hervorgehoben. Die Ergebnisse sind zum Teil besser als die von KMeans und von bekannten agglomerativen hierarchischen Clusterverfahren.

Nicht geklärt ist der Ursprung des Verfahrens. Während [206] das Verfahren nur beschreibt, aber keine Quelle angibt, referenziert [193] auf Forgy's Artikel [75]:

> "[...] This bisecting algorithm has been recently discussed and emphasized. [...] It is here worth noting that the algorithm above recalled is the very classical and basic version of K-means (except for a slightly modified initialization step), also known as Forgy's algorithm [...]"[193] .

Die angegebene Referenz [75] ist aber nur eine Zusammenfassung und lässt keine Schlüsse auf das eigentliche Verfahren zu. [193] zitiert weiterhin Gose u.a. [87]. Gose u.a. erwähnen zwar die Möglichkeit, Hierarchien von Clustern durch mehrfaches Anwenden von partitionierenden Verfahren zu erzeugen, zitieren aber auch nur die Kurzfassung [75] und beschreiben das Verfahren nicht als Bi-Sec-KMeans. MacQueen in [151] referenziert in seiner Arbeit auch Forgy's Arbeit von 1965. Seine Beschreibung lässt nicht den Schluss zu, dass Forgy den Bi-Sec-KMeans-Algorithmus entwickelt hat.

Kommen wir nun zum eigentlichen Algorithmus. Der Bi-Sec-KMeans-Algorithmus 5.2 basiert auf dem KMeans-Algorithmus. Er splittet wiederholt mit Hilfe von KMeans einen Cluster in zwei Teile, solange bis die gewünschte Clusteranzahl erreicht ist.

Bi-Sec-KMeans erbt die Eigenschaften von KMeans. Er ist abhängig von der gewählten Startlösung, dem Distanzmaß und konvergiert nur in ein lokales Optimum. Auch könnte man jede Variante des KMeans zum Splitten der Cluster einsetzen. Zusätzlich muss beim Bi-Sec-KMeans Verfahren immer ein Cluster zum Splitten ausgewählt werden. Die offensichtliche Variante ist die Wahl des größten Clusters, d.h. den Cluster mit der größten Menge an Objekten (siehe Schritt drei in Algorithmus 5.2). Das führt zu einer Clusterung mit ungefähr gleich großen Clustern. Möglich ist aber auch die Wahl des Clusters mit der größten Varianz oder eine Kombination aus beiden Ansätzen.

Algorithmus 5.2 Der Bi-Sec-KMeans Algorithmus

Input: Menge D mit Abstandsmaß, Anzahl k an Cluster
Output: Eine Partitionierung $\mathbb{P}$ der Menge D wobei für die Menge $\mathbb{P}$ mit k disjunkten Teilmengen aus D gilt: $\bigcup_{P \in \mathbb{P}} P = D$).

1: $\mathbb{P} := \{D\}$.
2: **for** $i := 1$ to $k - 1$ **do**
3: Wähle $P \in \mathbb{P}$ mit maximaler Kardinalität.
4: Wähle zufällig zwei Datenpunkte aus P als Ausgangszentroide $\vec{t_{P_1}}$ und $\vec{t_{P_2}}$.
5: **repeat**
6: Weise jeden Punkt aus P dem nächsten Zentroid zu.
7: Berechne die Clusterzentroide $\vec{t_{P_1}}$ und $\vec{t_{P_2}}$ der P_1 und P_2.
8: **until** Zentroide stabil.
9: $\mathbb{P} := (\mathbb{P} \setminus \{P\}) \cup \{P_1, P_2\}$.
10: **end for**

Steinbach u.a. haben in [206] die Auswirkungen der unterschiedlichen Strategien zur Auswahl der Cluster beim Clustern von Textdokumenten untersucht. Im Ergebnis war keine Strategie wirklich besser. Dies deckt sich mit den Tests, die wir im Rahmen unserer empirischen Studien durchgeführt haben. Daher verwendeten wir für unsere empirischen Untersuchungen immer den größten Cluster zum Splitten.

Der nächste Abschnitt führt die Formale Begriffsanalyse ein, die wir in der Arbeit zum konzeptuellen Clustern und zur Visualisierung von Bi-Sec-KMeans-Clusterergebnissen verwenden.

5.5 Einführung in die Formale Begriffsanalyse

Die Formale Begriffsanalyse ist ein Gebiet der angewandten Mathematik und der Informatik und wurde 1982 erstmals von Wille in [229] eingeführt. Wir werden in diesem Abschnitt die Teile der Theorie wiederholen, die für das Verständnis der Arbeiten aus Kapitel 8.5 notwendig sind. Mehr Details sind in [79, 80] zu finden. Im Folgenden führen wir die Begriffe formaler Kontext, formaler Begriff und Begriffsverband ein.[3] Techniken zur Visualiserung von Begriffsverbänden findet man in Abschnitt 5.5.3.

5.5.1 Formaler Kontext, Begriff, Begriffsverband

Die Formale Begriffsanalyse (FBA) wurde als mathematische Theorie eingeführt und modelliert den Begriff des "Begriffes" mittels der Verbandstheorie. Um die mathematische Beschreibung von Begriffen in Form von Intensionen (Inhalt) und Extensionen (Umfang) zu ermöglichen, benötigt die Formale Begriffsanalyse einen *formalen Kontext*, der wie folgt definiert ist:

Definition 5. *Ein* formaler Kontext *ist ein Tripel* $\mathbb{K} := (G, M, I)$*, wobei* G *eine Menge von* Gegenständen*,* M *eine Menge von* Merkmalen *und* I *eine binäre Relation zwischen* G *und* M *(d.h.* $I \subseteq G \times M$*) ist.* $(g, m) \in I$ *liest man "*Gegenstand g hat Merkmal m*".*

Abbildung 5.1 ist ein Beispiel für einen einfachen Kontext. Die Menge G ist eine Menge von Web-Seiten. Der Bezeichner der Gegenstände in Abbildung 5.1 zeigt das Thema (Finanzen und Sport) der Webseite an. Die Merkmale des Kontextes sind die Worte "bank, financ, market, american, team,

[3] Eine genaue Unterscheidung von Begriffen und Konzepten findet man in Kapitel 6.1.3.

	bank	financ	market	american	team	baseman	season
FinanceText1	X	X	X	X			
FinanceText2	X	X	X				
SportText1					X	X	X
SportText2				X	X	X	X

Abbildung 5.1: Einfacher formaler Kontext mit sieben Wortenstämmen aus vier Texten

baseman, season" der Web-Seiten. Die binäre Relation I wird durch die Tabelle aus Abbildung 5.1 anhand der Kreuze gegeben. Jedes Kreuz ist genau dann gesetzt, wenn der Text das Wort mindestens einmal enthält.

Aus dem formalen Kontext lässt sich die Begriffshierarchie, die man auch *Begriffsverband* nennt, wie folgt ableiten:

Definition 6. *Für $A \subseteq G$ definieren wir*

$$A^I := \{m \in M \mid \forall g \in A\colon (g,m) \in I\}$$

und für $B \subseteq M$ definieren wir

$$B^I := \{g \in G \mid \forall m \in B\colon (g,m) \in I\}.$$

Ein formaler Begriff *eines formalen Kontextes (G, M, I) ist definiert als ein Paar (A, B) mit $A \subseteq G$, $B \subseteq M$, $A^I = B$ und $B^I = A$. Die Mengen A und B nennt man* Umfang *und* Inhalt *des formalen Begriffes (A, B). Die Unterbegriff-Oberbegriff-Relation ist definiert als:*

$$(A_1, B_1) \leq (A_2, B_2) :\Longleftrightarrow A_1 \subseteq A_2 \quad (\Longleftrightarrow B_1 \supseteq B_2)\ .$$

Die Menge aller formalen Begriffe eines Kontextes $\mathbb{K}$ zusammen mit der partiellen Ordnung $\leq$ ist immer ein vollständiger Verband,[4] *den man* Begriffsverband *von $\mathbb{K}$ nennt. Bezeichnet wird der Begriffsverband mit $\underline{\mathfrak{B}}(\mathbb{K})$.*

Abbildung 5.2 zeigt ein Liniendiagramm des Begriffsverbandes zum Kontext aus Abbildung 5.1. Die Darstellung des Liniendiagramms folgt den Konventionen zur Visualisierung von Begriffshierarchien – dem internationalen Standard ISO 704 [122] oder den deutschen Standards DIN 2331 [55] und DIN 2330 [54]. In einem Liniendiagramm repräsentiert jeder Knoten einen formalen Begriff. Ein Begriff $\mathfrak{c}_1$ ist ein Unterbegriff eines Begriffes $\mathfrak{c}_2$ genau dann, wenn eine absteigende Kante vom Knoten, der $\mathfrak{c}_2$ repräsentiert, zum Knoten, der $\mathfrak{c}_1$ repräsentiert, existiert. Den Namen des Gegenstandes g findet man immer am Knoten des kleinsten Begriffes, der g im Umfang hat. Dual wird der Name des Merkmales m immer dem Knoten des größten Begriffes, der m als Inhalt hat, zugeordnet. Die Relation des Kontextes lässt sich direkt aus dem Liniendiagramm ablesen, da jeder Gegenstand g genau dann ein Merkmal m hat, wenn der Begriff mit der Bezeichnung g ein Unterbegriff des Begriffes mit der Bezeichnung m ist. Der Umfang eines Begriffes besteht aus allen Gegenständen, deren Bezeichnung an einen Unterbegriff angefügt ist. Dual ergibt sich der Inhalt aus allen Merkmalen der Oberbegriffe.

Aus Abbildung 5.2 lässt sich Folgendes ablesen: Der mit "american" bezeichnete Begriff hat {FinanceText1, SportText2} als Umfang und {american} als Inhalt. Eine Ober-Unterbegriffbeziehung

[4] d.h. für jede Menge von formalen Begriffen existiert immer ein kleinster gemeinsamer Oberbegriff und ein größter gemeinsamer Unterbegriff.

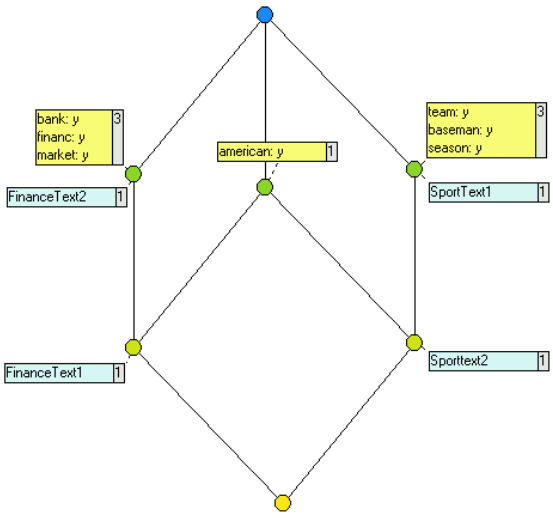

Abbildung 5.2: Begriffsverband für Kontext aus Abbildung 5.1

findet man im Beispiel zwischen den Begriffen ({FinanceText1, FinanceText2}, {bank, financ, market}) und ({FinanceText1}, {bank, financ, market, american}), wobei ({FinanceText1}, {bank, financ, market, american}) der Unterbegriff zu ({FinanceText1, FinanceText2}, {bank, financ, market}) ist.

Abbildung 5.4 visualisiert den Begriffsverband des Kontextes in Abbildung 5.3. Dieser Kontext ist umfangreicher als der erste und enthält weitere interessante Eigenschaften. Der Kontext stammt aus einem realen Beispiel. Es wurden 21 Dokumente im Internet gesammelt und manuell in drei Klassen eingeteilt. Wir nennen diesen Datensatz im Folgenden “DS1”. Jeweils sieben der Dokumente gehören zu den Klassen Finanzwirtschaft, Software und Fußball. 1419 verschiedene Wortstämme wurden extrahiert, wobei 253 als Stoppworte identifiziert und entfernt wurden. Weiterhin wurden alle Wortstämme, die in nur einem Dokument enthalten sind und alle Terme, die seltener als fünfmal vorkommen, entfernt. Der resultierende Termvektor besteht aus 117 Worten. Die Dokumente wurden mittels KMeans zu zehn Clustern zusammengefasst. Die Schranke für die Aufnahme eines Terms in die Clusterrepräsentation des Kontextes wurde auf 15 % des maximalen Wertes des Clusters festgelegt (mehr dazu siehe Kapitel 4.5). Alle 117 Merkmale finden sich auch im Kontext in Abbildung C.1 in Anhang C wieder. Der Kontext in Abbildung 5.3 enthält acht der 117 Merkmale. Durch die geringe Anzahl an Merkmalen ermöglicht der resultierende Begriffsverband die Diskussion weiterer interessanter Eigenschaften von Begriffsverbänden.

Abhängigkeiten zwischen Merkmalen können durch Implikationen beschrieben werden. Für die Merkmalsmengen $X, Y \subseteq M$ findet man im Kontext genau dann eine Implikation $X \rightarrow Y$, wenn jeder Gegenstand, der alle Merkmale in X hat, auch alle Merkmale in Y hat. Anders ausgedrückt gilt eine Implikation $X \rightarrow Y$ in einem Begriffsverband (G,M,I) immer dann, wenn $Y \subseteq X^{II}$ ist. Eine sehr einfache Implikation in Abbildung 5.4 ist z.B. {financi}→{base} oder {service}→{financi, base}. Dem entnimmt man die Implikation, indem man den größten gemeinsamen Unterbegriff beider Begriffe, im ersten Fall ist dies trivialer Weise “financi”, im zweiten “service”, lokalisiert.

Wenden wir dieses Vorgehen auf weitere Begriffe in Abbildung 5.4 an, so findet man im Beispiel unter anderem die Implikation: {end, service} → {develop,software}. Der größte Begriff, den

	end	european	cup	base	financi	develop	softwar	servic
Finance - 0 (3)				X	X	X	X	
Soccer - 1 (3)		X	X					
Soccer - 2 (1)			X					
Finance - 3 (3)	X			X	X			
Finance - 4 (1)				X	X		X	X
Software - 5 (4)				X				
CL6 (0)	X	X	X	X	X	X	X	X
Software - 7 (3)	X			X	X	X	X	X
Soccer - 8 (3)	X		X					
CL9 (0)	X	X	X	X	X	X	X	X

Abbildung 5.3: Kontext zum DS1-Datensatz

“end” und “service” gemeinsam haben, ist der Begriff mit dem Gegenstand “Software - 7 (3)”. Betrachtet man den Inhalt dieses Begriffes, so enthält die Menge die Merkmale: {end, servic, financi, base, software, develop}. Die Menge entspricht der Menge {end, service}^{II}. Da {develop,software} in dieser Menge enthalten ist, haben wir auf diese Weise die genannte Implikation gefunden. Ein weiteres Beispiel stellt die Implikation {european, end} $\rightarrow M$ dar.

Im Folgenden werden wir die begrifflichen Skalen einführen. Die begrifflichen Skalen bieten die Möglichkeit, auch für nicht binäre Merkmale einen Verband und eine entsprechende Visualisierung zu berechnen. Wir werden so in die Lage versetzt, mehr Informationen aus den numerischen Attributen in den Verband zu übernehmen.

5.5.2 Begriffliches Skalieren

In den meisten Anwendungen kommen nicht nur binäre Merkmale vor. Vielmehr besitzen die Merkmale häufig mehr als eine Ausprägung, wie z.B. Farben, Studienrichtung oder die Häufigkeit von Worten in Dokumenten. In der Begriffsanalyse bezeichnet man Kontexte mit nicht-binären Merkmalen entsprechend als *mehrwertige Kontexte*. Ein mehrwertiger Kontext entspricht einer Relation, wie sie im Bereich der Datenbanken verwendet wird. Einfach ausgedrückt, handelt es sich dabei um eine Tabelle. Die Tabelle darf nur ein Schlüsselmerkmal enthalten, welches dann die Menge G der Gegenstände repräsentiert.

Um aus einem mehrwertigen Kontext einen Begriffsverband ableiten zu können, muss der Kontext in einen einwertigen Kontext überführt werden. Dieser Übersetzungsprozess erfolgt durch begriffliches Skalieren. Auf dem resultierenden einwertigen Kontext können die bekannten Techniken der Formalen Begriffsanalyse angewendet werden.

Definition 7. *Ein* mehrwertiger Kontext *ist ein Tupel* $(G, M, (W_m)_{m\in M}, I)$*, wobei* G *eine Menge von Gegenständen und* M *eine Menge von Merkmalen ist.* W_m *ist eine Menge von Werten für jedes* $m \in M$*, und* $I \subseteq G \times \cup_{m\in M}(\{m\} \times W_m)$ *ist eine Relation für die gilt, dass aus* $(g, m, w_1) \in I$ *und* $(g, m, w_2) \in I$ *stets* $w_1 = w_2$ *folgt.*

Eine begriffliche Skala *eines Merkmales* $m \in M$ *ist ein einwertiger Kontext* $\mathbb{S}_m := (G_m, M_m, I_m)$ *mit* $W_m \subseteq G_m$*. Der Kontext* $\mathbb{R}_m := (G, M_m, J_m)$ *mit* $gJ_mn :\iff \exists w{\in}W_m\colon\ (g, m, w){\in}I \wedge (w, n){\in}I_m$ *wird als* realisierte Skala *des Merkmales* $m \in M$ *bezeichnet.*

Die Menge M_m enthält die Werte eines Merkmales, die zur Transformation des mehrwertigen in den einwertigen Kontext genutzt werden. Im Prinzip gibt es keinen Unterschied zwischen einem “normalen” Kontext und einer Skala. Die besondere Aufgabe, die dem Kontext einer Skala zukommt, macht eine separate Benennung sinnvoll.

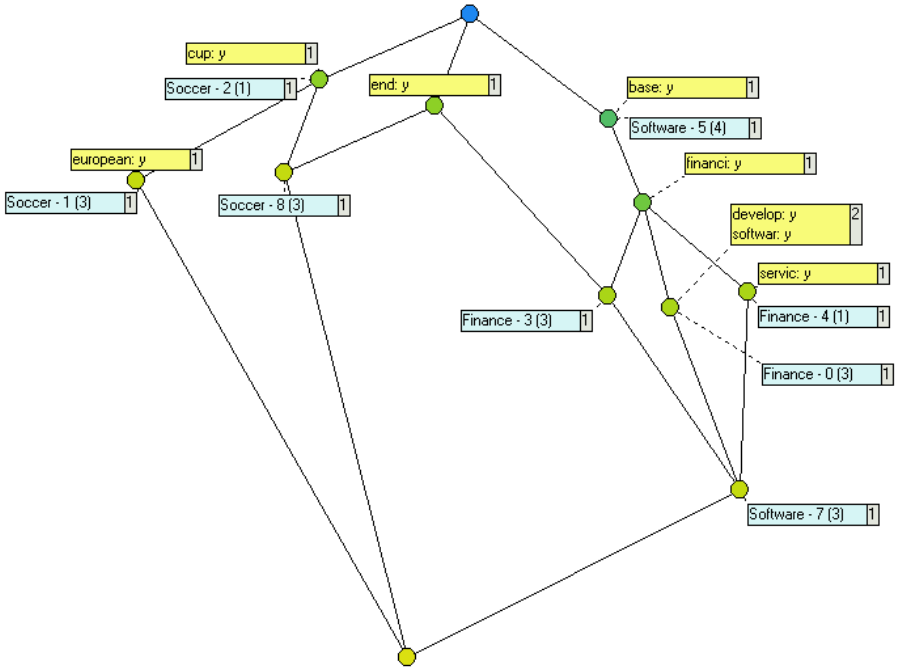

Abbildung 5.4: Begriffsverband zu 21 Texten mit zehn KMeans-Clustern aus den Bereichen Finanzwirtschaft, Fußball und Software (Die Gegenstände sind die KMeans-Cluster, wobei die Clusternummer nach dem Bindestrich zu finden ist. Der Eintrag in Klammern gibt die Anzahl der Dokumente an.)

In der Arbeit nutzen wir die Skalen für die verbesserte Repräsentation der Texte. Bisher konnten wir nur binäre Informationen der Form: “Wort kommt im Text vor oder Wort kommt nicht im Text vor”, im Kontext kodieren. Worte kommen häufiger mehr als einmal im Text vor. Gegenüber diesem numerische Wert, der meistens noch gewichtet wird (z.B. mit tfidf, siehe Kapitel 4.2.5.1), verliert man bei der binären Kodierung für den formalen Kontext sehr viele Informationen. Mit Hilfe der begrifflichen Skalen ist man in der Lage, nicht nur vorkommende Worte in den Kontext aufzunehmen, sondern man kann zwischen unterschiedlich wichtigen Worthäufigkeiten unterscheiden. Das Diskretisieren in mehr als zwei Klassen reduziert den Informationsverlust und macht den resultierenden Begriffsverband aussagekräftiger.

5.5.3 Visualisierung von “gedrehten” Begriffsverbänden

Zur Visualisierung von Begriffsverbänden lassen sich *Hasse-Diagramme* verwenden. Im letzten Abschnitt wurden die Abbildungen 5.2 und 5.4 mit dieser Technik visualisiert und das Lesen der Diagramme erläutert. Als Merkmale wurden z.B. Worte oder Wortstämme und als Gegenstände z.B. Dokumente bzw. Dokumentcluster verwendet. Die Anzahl der Worte ist im Allgemeinen höher als die Anzahl der Dokumentcluster (und meist auch höher als die Anzahl der Dokumente). Zur Visualisierung haben wir auf die Software Cernato der Firma Navicon GmbH zurückgegriffen. Sie ist in der Lage, Verbände für eine größere Anzahl von Gegenständen, aber nur eine kleine Anzahl von Merkmalen zu berechnen. Um die Software trotzdem einsetzen zu können, invertieren wir die gewöhnliche Leserichtung des visualisierten Begriffsverbandes. Die Knoten des dargestellten Graphen

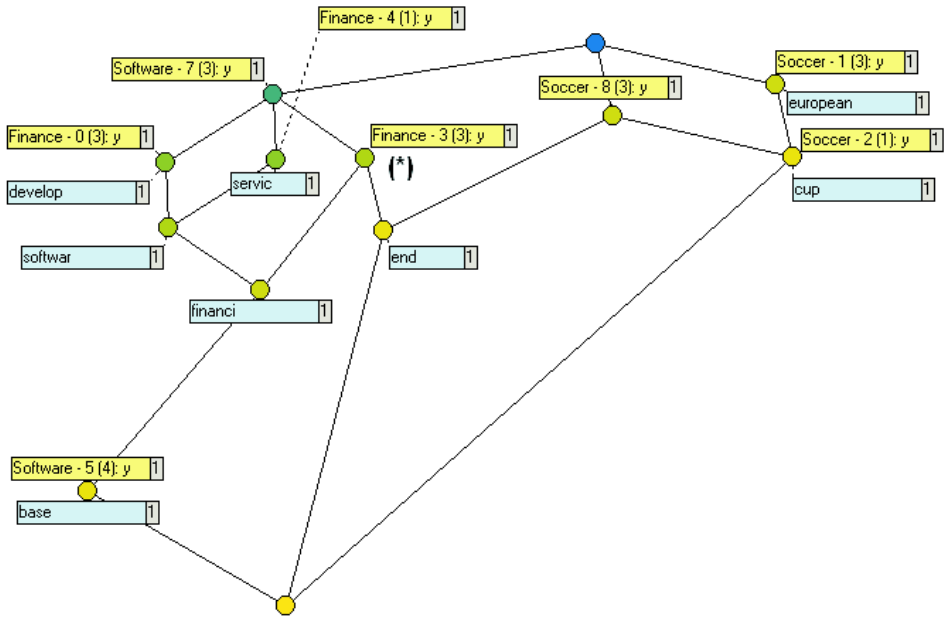

Abbildung 5.5: Gedrehter Begriffsverband zum Kontext in Abbildung 5.3

sind weiterhin die Formalen Begriffe. Ein Begriff $\mathfrak{c}_1 \in \mathfrak{B}(\mathbb{K})$ ist aber genau dann ein Unterbegriff von $\mathfrak{c}_2 \in \mathfrak{B}(\mathbb{K})$, wenn es einen Pfad aus absteigenden (!) Kanten vom Knoten, der $\mathfrak{c}_1$ repräsentiert, zum Knoten, der $\mathfrak{c}_2$ repräsentiert, gibt. Ein Beispiel zeigt Abbildung 5.5 mit dem "auf den Kopf gestellten" Verband zu Abbildung 5.4.

Der Name eines Gegenstandes g wird immer mit dem spezifischsten Knoten verbunden (z.B. das kleinste Konzept in Bezug auf $\leq$), welcher g noch im Umfang (z.B. ist es in Abbildung 5.5 für den formalen Begriff (*) der Gegenstand {Finance - 3}) hat. Analog findet man den Namen des Merkmales m immer am allgemeinsten Knoten, der m noch im Inhalt hat (z.B. hat (*) u.a. das Merkmal "end" im Inhalt; vgl. Abbildung 5.5). Weiterhin können wir immer die Beziehungen des formalen Kontextes aus dem Diagramm lesen. Jeder Gegenstand g hat ein Merkmal m genau dann, wenn ein Begriff, der mit g ausgezeichnet ist, ein Unterkonzept des Begriffes mit der Bezeichnung m ist. Der Umfang eines Begriffes besteht aus den Bezeichnern aller Gegenstände der korrespondierenden Unterbegriffe und analog besteht der Inhalt aus den Bezeichnern aller Merkmale der korrespondierenden Oberbegriffe.

In dieser Arbeit werden wir soweit möglich auf die gewöhnliche Darstellung der Liniendiagramme zurückgreifen. Falls die Anzahl der Worte zu groß wird, werden wir die gedrehten Liniendiagramme verwenden und dies an gegebener Stelle explizit erwähnen.

5.6 Clusterverfahren

Im folgenden Abschnitt werden ausgewählte Clusterverfahren als Beispiel einer Verfahrensklasse vorgestellt. Ziel ist es, einen Überblick über die aktuell verfügbaren Verfahren und Ansätze im Bereich des Clusterns zu geben. Einige Verfahren sind dabei gut geeignet für das Clustern von Texten und andere für das Clustern von Kommunikationsdaten (siehe Kapitel 2.5). Wieder andere

erlauben eine anschauliche Beschreibung der Clusterergebnisse. Wir arbeiten die Vor- und Nachteile der verschiedenen Verfahrensklassen heraus und vergleichen sie mit den in der Arbeit angewendeten Verfahren (siehe Abschnitte 5.4 und 5.5) bzw. Verfahrenskombinationen.

Beim Einsatz der Clusteranalyse ist ein entscheidender Faktor das Skalenniveau der untersuchten Daten. Es gibt nominal, ordinal, Intervall- und metrisch skalierte Daten [15], [208], die auch eine unterschiedliche Behandlung der Daten während des Clusterns und so auch unterschiedliche Verfahren nach sich ziehen. Wir konzentrieren uns vorrangig auf Verfahren für die metrische Skala und beschränken uns beim Überblick auf ausgewählte und mit der Arbeit verwandte Verfahren. Einen sehr umfangreichen Überblick gibt der Artikel von Berkhin [20].

Wir starten den Abschnitt mit den hierarchischen Clusterverfahren, die nur auf sehr kleine Datenmengen anwendbar sind. Wir gehen dann auf Co-Clustering, den EM-Algorithmus und auf RDBC als eine Verallgemeinerung bzw. relationale Erweiterung des KMeans, auf SOM's (Self Organizing Maps) aus dem Bereich des Maschinellen Lernens und auf Subspace-Clustering ein. Den Abschluss der Verfahren auf numerischen Werten bilden die dichte-basierten Verfahren wie z.B. Optics. Bevor wir das Kapitel mit einer Zusammenfassung beenden, gehen wir noch auf den Bereich der konzeptuellen Clusterverfahren und hier insbesondere auf das bekannte COBWEB-Verfahren als ein Vertreter diese Klasse ein.

5.6.1 Hierarchische Clusterverfahren

Hierarchische Verfahren bekamen ihren Namen durch das Bilden einer Folge von Gruppierungen bzw. Clustern. Die Folge kann in einer Hierarchie von Clustern dargestellt werden. Diese Hierarchie lässt sich zum einen durch stufenweises Verfeinern, ausgehend von einer einzigen, alle Objekte umfassenden Menge, erstellen. Man spricht in diesem Fall auch von so genannten "divisiven" Verfahren. Die "agglomerativen" Verfahren hingegen fügen die einelementigen Cluster (feinste Stufe) schrittweise zusammen, bis alle Objekte in einem Cluster enthalten sind. In der Praxis kommt den divisiven Verfahren fast keine Bedeutung zu. Daher soll im Folgenden nur der agglomerative Algorithmus skizziert werden.

Beim agglomerativen Verfahren bildet initial jedes Objekt d der Menge D einen Cluster. Es handelt sich um die erste Clusterlösung. Es kann nicht vorkommen, dass Objekte anteilig mehreren Clustern angehören. Man bestimmt die Ähnlichkeit zwischen den Clustern auf der Basis der ersten Clusterung und wählt die beiden Cluster p, q der Clusterung $\mathbb{P}$ mit der minimalen Distanz $dist(p, q)$ aus. Beide Cluster werden fusioniert und man erhält eine neue Clusterung. Diesen Vorgang setzt man fort und berechnet die Abstände zwischen dem neuen Cluster und allen übrigen neu, um dann wieder die beiden Cluster mit der minimalen Distanz $dist(p, q)$ zusammenzufügen. Der Algorithmus bricht ab, wenn nur noch ein Cluster übrig ist. Die Distanz kann nach Gleichung 5.1 oder Gleichung 5.2 berechnet werden. Es ist auch möglich, die Cluster direkt auf der Basis der Ähnlichkeitsbeziehung, gegeben durch eine Matrix, zu berechnen. Die Berechnung der Ähnlichkeit zwischen den Clustern mit mehr als einem Element bestimmt das Ergebnis. Die Verfahren Single Linkage, Complete Linkage oder Ward Verfahren sind die gebräuchlichsten. Details findet man z.B. in [15].

Mittels so genannter Dendrogramme kann man die Hierarchie der Cluster darstellen, die sich durch das wiederholte Verschmelzen der Cluster bei der Verfahrensanwendung ergibt. Gleichzeitig erlauben die Dendrogramme die Abschätzung der richtigen Clusteranzahl. Die unterschiedlichen Varianten des Hierarchischen Clustern berechnen nicht zwingend die inhärente Klassenstruktur, d.h. zusätzliches Wissen ist für die Auswahl der richtigen Variante notwendig. Jede Variante zur Berechnung der Ähnlichkeit hat ihre Vor- und Nachteile, wobei wir auf die Details an dieser Stelle nicht eingehen wollen (mehr in [15]). Wesentlich problematischer für die Anwendung der Verfahren auf große Datenmengen ist die Speicherung der Ähnlichkeitsmatrix. Diese benötigt $d(d-1)/2$

Speicherplätze. Auch das Laufzeitverhalten mit $O(n^2)$ ist gegenüber dem linearen Verhalten von KMeans schlechter.

5.6.2 Co-Clustering

Co-Clustering-Verfahren bezeichnen das simultane Clustern von Objekten und Merkmalen, wobei in [52] die Objekte Textdokumente und die Merkmale Worte sind. Sie folgen damit einem anderen Paradigma als die “klassischen” Clusterverfahren wie KMeans, die nur Elemente der einen Dimension anhand ihrer Ähnlichkeit bezüglich der zweiten clustern. Co-clustering-Verfahren sind in anderen Anwendungsgebieten auch unter den Namen “Biclustering” [37] oder “Two-Mode Clustering” [81] bekannt und werden z.B. bei der Analyse von Genom-, Marktforschungs- oder Web-Log-Daten eingesetzt und weiterentwickelt.

Co-Clustering geht wie KMeans auch von einer Vektorrepräsentation der Objekte aus. Die in dieser Arbeit eingeführte ontologiebasierte Repräsentation der Objekte (für Dokumente oder Kommunikationsdaten) erlaubt die Anwendung von Co-Clustering-Verfahren. Die sich ergebende Matrix wird als zweidimensionale Kontingenztabelle aufgefasst und typischerweise werden iterativ abwechselnd Objekt- und Merkmalscluster berechnet (vgl. [52, 81]).

Eine interessante offene Forschungsfrage ist die Anreicherung von Co-Clustering-Verfahren mit Hintergrundwissen. Denkbar wäre insbesondere eine enge Verzahnung von Co-Clustering und ontologiebasierter Repräsentation zur Steigerung der Clustergüte bzw. zum gezielten Ableiten von themenbezogenen Clustern.

5.6.3 SOM

Self Organizing Maps (SOM) sind Vertreter der unüberwachten Lernverfahren basierend auf neuronalen Netzen. Es erfolgt eine exakte Zuordnung der Objekte zu den Clustern. Die Neuronen werden in einem regelmäßigen ein- oder zweidimensionalen Gitter angeordnet. Die SOMs projizieren hochdimensionale Daten auf das zweidimensionale Neuronengitter und erlauben auf diese Weise die Visualisierung der enthaltenen Zusammenhänge. Ein Beispiel für den Bereich des Text Mining findet man in [143].

SOMs arbeiten nach folgendem Prinzip: Jedes Neuron wird durch einen n-dimensionalen Gewichtsvektor repräsentiert, wobei n gleich der Anzahl der Attribute im Datensatz ist. Jedes Neuron ist mit einer bestimmten Menge benachbarter Neuronen verbunden. Man präsentiert nacheinander die Elemente des Datensatzes und ermittelt jedes Mal ein "Gewinnerneuron"(Best Matching Unit, BMU). Das Gewinnerneuron besitzt die höchste Ähnlichkeit zum präsentierten Datum. Der Abstand wird über die, z.B. euklidische Distanz (Gleichung 5.1), berechnet. Die Kohonen-Lern-Regel beschreibt die Veränderung der Gewichtsvektoren so, dass man dieses Neuron und die in seiner Umgebung befindlichen Neuronen noch besser an das Datum anpasst. Um den Lern- bzw. Anpassungsprozess besser zu steuern, wird die Nachbarschaft der BMU über eine Nachbarschaftsfunktion und eine Lernrate in Abhängigkeit von der Zeit gesteuert. Der am Anfang sehr groß gewählte Radius bewirkt eine globale Suche und verhindert damit das schnelle Konvergieren in ein lokales Optimum. Im Verlaufe des Trainings reduziert man die Lernrate und damit die Größe der Nachbarschaft, bis alle Objekte eines Clusters auf ein Neuron abgebildet werden, und vollzieht damit die Feinjustierung der SOM. Durch dieses Vorgehen werden nicht nur Daten, die in den gleichen Cluster fallen, zusammengefasst, sondern auch die benachbarten Cluster auf dem Gitter sind sich ähnlicher als weiter auseinander liegende.

Der KMeans Algorithmus (Abschnitt 5.4) und die SOMs sind zwei ähnliche Verfahren. Man kann zeigen, dass sich die SOMs genau wie eine bestimmte Variante des KMeans verhalten, wenn man

als Lernrate 1/(n+1) wählt [139]. Sie könnten alternativ zu KMeans eingesetzt werden.

5.6.4 EM-Algorithmus

Der EM-Algorithmus [49, 169] besteht aus den zwei Schritten “expectation” und “maximization” und kann zur Lösung einer verallgemeinerten Variante des KMeans eingesetzt werden. Dabei lässt man gegenüber KMeans die Annahme der deterministischen Zuordnung der Klassen fallen. Der EM-Algorithmus ist ein probabilistisches Clusterverfahren, d.h. die Objekte werden nicht wie bei KMeans exakt einem Cluster zugeordnet, sondern mit einer bestimmten Wahrscheinlichkeit. Weiterhin nimmt man an, dass die Daten durch die Mischung von z.B. k Gaußverteilungen entstanden sind, die man nun schätzen will. Der EM-Algorithmus geht zur Schätzung in zwei sich wiederholenden Schritten vor:

E-step Berechne die Zuordnungswahrscheinlichkeit zum Cluster für jedes Objekt.

M-step Schätze die Verteilungsparameter basierend auf den Zugehörigkeitswahrscheinlichkeiten der Objekte zu den Clustern.

Typischerweise werden die Zugehörigkeitswahrscheinlichkeiten der Objekte zu den Clustern als Gewichte der Objekte gespeichert.

Der EM-Algorithmus zeigt ein schlechtes Konvergenzverhalten und benötigt einen großen Datensatz. Weiterhin muss vor der Anwendung des Verfahrens das zu schätzende Modell identifiziert werden.

5.6.5 Relational Distance-Based Clustering

Aus dem Bereich des so genannten First-Order-Clusterns, das sich mit dem Anwenden von Clusterverfahren auf relationalen Daten beschäftigt, kommen Ansätze von Kirsten und Wrobel in [134, 135]. Sie stellen sowohl das Relational Distance-Based Clustering (RDBC) – eine Variante numerischer First-Order-Clusterer auf der Basis hierarchischer Clusterverfahren – sowie ein Verfahren als Erweiterung des schnellen KMeans-Verfahren vor. Zentrale Idee ist die Berechnung numerischer Werte für die Ähnlichkeiten relationaler Daten sowie die Nutzung von Distanz-Metriken aus dem Gebiet des relationalen Lernens.

5.6.6 Subspace-Clustering

In der Literatur existieren eine Reihe von Ansätzen aus dem Bereich Clustern in Unterräumen, auch Subspace-Clustering genannt. Dabei existieren zwei Vorgehensweisen. Auf der einen Seite kann man die Dimensionsreduktion vor dem eigentlichen Clusterlauf durchführen. Hier kommen bekannte Verfahren aus der Statistik wie die Hauptkomponentenanalyse zum Einsatz [59]. Auf der anderen Seite wird die Dimensionsreduktion direkt mit dem Clusterverfahren verknüpft. In [7, 5] werden zwei Ansätze zur Kombination von Cluster- und Projektionsverfahren vorgestellt.

Die Gruppe um Charu Aggrawal nutzt in ihrem Verfahren ORCLUS die Singulärwertzerlegung, um simultan Unterräume und Cluster zu berechnen. Um für jeden Cluster die spezifisch wichtigen Dimensionen berechnen zu können, wird dynamisch während des Clusterns für jeden Cluster eine neue Projektion berechnet. Die Projektion kann dann von Cluster zu Cluster variieren. Vereinfacht gesprochen fokussiert man die Sicht auf die Objekte eines Clusters und hebt die relevanten Merkmale hervor. Das Verfahren baut auf KMeans auf und erweitert es um die Projektion der Cluster

in die Unterräume. Die Laufzeit ist höher als bei KMeans (mehr in [5]). Dies macht das Verfahren nicht anwendbar auf große Datenmengen.

Im Gegensatz zu ORCLUS basiert der Ansatz von Rakesh Agrawal u.a. [7] auf so genannten dichte-basierten Clusterverfahren. Wir gehen im nächsten Unterkapitel auf diese Verfahrensklasse genauer ein. Während ORCLUS zur Projektion Linearkombinationen der ursprünglichen Merkmale einsetzt, nutzt CLIQUE die Merkmale direkt und erhält so die Interpretierbarkeit. Dies wird auch bei der Beschreibung der berechneten Cluster durch DNF-Ausdrücke (Ausdrücke in disjunktiver Normalform (DNF)) genutzt. Der Ansatz ist der Idee des Subjektiven Clusterns sehr ähnlich (mehr dazu in Abschnitt 7.5).

Beide Ansätze zeigen erfolgreich die Berechnung von Clustern auch in hochdimensionalen Räumen.

Der folgende Abschnitt stellt dichte-basierte Clusterverfahren vor.

5.6.7 Dichte-basierte Clusterverfahren

Dichte-basierte Verfahren stellen eine weitere Klasse der Clusterverfahren dar. Sie basieren primär auf der Nächsten-Nachbar-Suche und nutzen zur Identifikation der Cluster die so genannte Dichte, die vorab durch Parameter spezifiziert werden muss. Einen guten Überblick geben Ester und Sander in [64].

Die Grundidee eines dichte-basierten Clusterns basiert auf der lokalen Punktedichte, die um jedes Objekt innerhalb des Clusters oberhalb eines gegebenen Grenzwertes liegen muss. Die lokale Punktedichte für ein Objekt d ergibt sich durch die Anzahl der Objekte, die in einer festgelegten Umgebung um das Objekt d liegen. Die Punktmenge eines Clusters besteht aus den Punkten eines dichten Gebietes, die wiederum zusammenhängen. Durch das Festlegen der Dichte ist man nicht nur in der Lage, einzelne Cluster zu bestimmen, sondern man kann Punkte auch als Rauschen im Raum identifizieren. Diese Idee kann formal präzisiert werden und führt so zu verschiedenen dichte-basierten Clusteralgorithmen.

Ein einfacher dichte-basierter Algorithmus ist DBSCAN - Density-Based Clustering of Applications with Noise [65]. Nachteil von DBSCAN ist der starke Einfluss der Parameter auf die Clusterergebnisse. Weiterentwicklungen wie der Optics-Algorithmus berücksichtigen auch unterschiedlich dichte Regionen und sind wesentlich unempfindlicher gegenüber der Parameterwahl.

Um Cluster auf unterschiedlich dichten Regionen berechnen zu können, bietet sich eine Kombination aus hierarchischem und dichte-basiertem Clustern an. Durch die festgelegte Dichte im DBSCAN-Verfahren werden nur Cluster dieser Dichte bestimmt. Baut man eine Hierarchie von Clustern auf, so ist man in der Lage, auch in einem Cluster weitere Gebiete zu identifizieren, in denen die Punkte dichter liegen. Dafür muss man die Punkte in eine Ordnung bringen, so dass die Punkte eines Clusters zusammenhängen. Dieser Idee folgt Optics (Ordering Points to Identify the Clustering Structure), beschrieben in [11]. Die effiziente Umsetzung verdankt man der Eigenschaft, dass dichte-basierte Cluster bzgl. einer höheren Dichte vollständig in dichte-basierten Clustern mit niedrigerer Dichte enthalten sind. Wählt man immer den dichtesten Punkt, der noch dichteerreichbar ist, so werden die Punkte gemäß ihrer Dichte-Erreichbarkeit sortiert. Diesen Teil erledigt Optics. Danach muss die Clusterhierarchie aus dem sortierten Clusterdatenbestand berechnet werden. Für die Details sei wieder auf [64] verwiesen.

5.6.8 Konzeptuelles Clustern — COBWEB

Bisher haben wir vorrangig Verfahren für numerische Daten behandelt. Eine erste Brücke zu den konzeptuellen Clusterverfahren bildeten die relationalen Verfahren aus Abschnitt 5.6.5. Michalski

[164, 165] definiert das konzeptuelle Clustern (conceptual clustering) als eine Aufgabe des Maschinellen Lernens. Dabei versteht er unter konzeptuellem Clustern das Lernen von Beobachtungen "learning by observation" und löst diese Aufgabe durch die Konstruktion einer Klassifikationshierarchie der beobachteten Objekte. Die Cluster dieser Hierarchie lassen sich durch logische Ausdrücke beschreiben. Michalski entwickelte den Cluster/2-Algorithmus (vgl.[165]), den wir an dieser Stelle nicht weiter beschreiben wollen. Abschnitt 5.5 beschäftigt sich mit der Formalen Begriffsanalyse zum konzeptuellen Clustern. Eines der schnelleren konzeptuellen Clusterverfahren ist COBWEB, dessen Idee wir im weiteren Verlauf dieses Abschnittes erläutern wollen.

COBWEB [74] ist eines der bekanntesten konzeptuellen Clusterverfahren, welches automatisch eine Klassifikationshierarchie lernt. Es handelt sich um ein inkrementelles und damit für große Datensätze geeignetes Verfahren, welches eine Hill-Climbing Suche durch den Raum der Klassifikationshierarchien durchführt. Das Verfahren nutzt als Maß zur Steuerung der Suche die so genannte "category utility", die die Innerklassenähnlichkeit und die Zwischenklassenunähnlichkeit mittels bedingter Wahrscheinlichkeiten beschreibt. Das Einfügen neuer Instanzen startet immer beim Wurzel-Knoten. Die Instanz kann dann entweder durch das Erstellen neuer Knoten oder das Mergen mit anderen Knoten in die Hierarchie integriert werden. Auch werden weitere Heuristiken angegeben, um Knoten im Baum zwecks besserer Beschreibung der Daten zu splitten oder zusammenzufassen.

Die Vorteile der konzeptuellen Clusterverfahren liegen in der prinzipbedingten Beschreibung der gefunden Cluster. Die Beschreibung der Cluster wird immer automatisch berechnet. Problematisch sind häufig die Laufzeitverhalten bzw. bei schnelleren Verfahren der Einfluss der Heuristik auf die Ergebnisse.

5.6.9 Zusammenfassung und Ausblick

Insbesondere zeigt der Überblick, dass keines der Clusterverfahren für alle Arten von Daten geeignet ist. Die vorgestellten Verfahren decken nur einen kleinen Teil der bekannten Clusterverfahren ab, wobei wir auf unterschiedliche Verfahrenklassen eingegangen sind. Neben den konzeptuellen und hierarchischen Clusterverfahren, die für kleine Datensätze entwickelt wurden und mehr Verständlichkeit für den Menschen bieten, sind wir auf die modellbasierten Verfahren wie den EM-Algorithmus aber auch auf die dichte-basierten Verfahren, die in der Lage sind, Cluster ganz unterschiedlicher Form zu erkennen, eingegangen. Nicht näher erläutert haben wir gemischtskalierte Verfahren wie sie in [73] beschrieben werden und das Gebiet der Fuzzy-Clusteranalyse [120].

Die heutige Entwicklung zielt immer mehr auf Verfahren, die sehr große Datenmengen verarbeiten können. Ziel ist es, möglichst nur einmal den Datenbestand lesen zu müssen bzw. nur Teilmengen der Daten überhaupt zu analysieren. Die aktuelle Forschung untersucht zur Zeit vor allen Dingen das Problem der Hochdimensionalität des Merkmalsraumes, das eng korreliert ist mit dem Problem der großen Datenmengen. Zu diesem Thema findet man viele Veröffentlichung gerade aus der Datenbank-Community, wie z.B. die Verfahren CLIQUE [7] oder OptiGrid [107]. Ganz andere Ansätze stammen aus der Wavelet-Theorie [196], wobei man Wavelet-Transformationen zur Vorverarbeitung einsetzt. Diese Ansätze bieten auch die Chance, das Problem der Hochdimensionalität der Datensätze zu überwinden [175].

Die Anwendbarkeit auf den Bereich des Text-Clusterns oder zum Clustern von Kommunikationsdaten sowie die Entwicklung neuer und sehr schneller Verfahren sind weitere Aufgaben für die Zukunft. Der Ansatz dieser Arbeit zielt auf die Kombination bestehender Verfahren sowie die Integration von Hintergrundwissen. Die Synergien der kombinierten Verfahren sowie das Hintergrundwissen tragen zu besseren und verständlicheren Clusterergebnissen bei.

6 Ontologien

Das Kapitel gliedert sich in drei Teile. Abschnitt 6.1 beschreibt die Herkunft von Ontologien und die Verbindung zum Text Mining. Eine informelle und eine formale Definition des Begriffs Ontologie findet man in Abschnitt 6.2. Abschnitt 6.3 schließt das Kapitel mit der Vorstellung verschiedener Quellen für die Akquisition von Hintergrundwissen bzw. beschäftigt sich mit der Erstellung und Akquisition von Ontologien.

6.1 Grundlagen und Geschichte

6.1.1 Die Wurzeln der Ontologien

Ontologien erhielten in den vergangenen Jahren im Bereich der Informatik und insbesondere im Bereich der Künstlichen Intelligenz immer mehr Aufmerksamkeit [91, 92, 221, 155, 71]. Im aufkommenden Semantic Web dienen Ontologien als Backbone und bilden eine zentrale Schicht zur Repräsentation des Wissens [22, 21, 19]. Ursprünglich stammt der Begriff “Ontologie” aus der Philosophie und umschreibt eine philosophische Disziplin, die sich mit der Natur und der Organisation des Seins beschäftigt. Philosophen suchen dabei Antworten auf Fragen der Art: “Was ist Sein?” und “Was sind die gemeinsamen Merkmale allen Seins?”. Der Term “Ontologie” wurde von Aristoteles in Metaphysics, IV, 1 eingeführt. Im Folgenden wollen wir Ontologien nicht aus der Sicht der Philosophie, sondern der Informatik betrachten. In der Informatik handelt es sich dabei um ein technisches Artefakt, bestehend aus Konzepten und Beziehungen zwischen diesen, um Teile der realen Welt zu beschreiben [153]. In Abschnitt 6.2 werden wir eine ausführliche Definition geben.

Verschiedene Forscher haben Klassifikationsschemata für Ontologien entwickelt (vgl. [210]). Guarino schlägt in [94] vor, die Ontologien anhand ihrer Generalität zu unterscheiden und teilt die Ontologien in die Bereiche Top-Level-, Domänen-, Aufgaben- und Anwendungsontologien ein. Die Top-Level-Ontologien formalisieren allgemeine Dinge wie Raum, Zeit oder Ereignisse und sind so unabhängig von einer konkreten Aufgabe. Sie können für die Modellierung spezifischer Ontologien wiederverwendet werden. Domänen- und Aufgabenontologien sind Ontologien, die speziell für eine Domäne entwickelt wurden, wobei erstere die Domäne im Allgemeinen beschreiben und zweitere mehr den Fokus auf die Aufgaben in einer Domäne legt. Anwendungsontologien stellen die spezifischste Form der Ontologie dar. Sie übernehmen spezielle Rollen in Anwendungen und bilden die Basis für Implementierungen (vgl. [214]).

Man findet in der Literatur eine Reihe von Beispielen für erfolgreiche ontologiebasierte Anwendungen (entnommen und erweitert aus [63, 214, 153]):

- Wissensbasierte Systeme (z.B. [70]),
- Sprachverarbeitung und maschinelle Übersetzung, (z.B. Wordnet [167], [202], [41]),
- Information Retrieval und Informationsintegration (z.B. [128], [163], [227]),
- Text Mining (z.B. [69],[113],[118],[115]),
- Webportale und Wissensportale (z.B. Yahoo [142], SEAL [157], OntoWeb [211])
- Intelligente Suchmaschinen (z.B. Getess [203], OntoSeek [95]),
- Digitale Bibliotheken (z.B. [8],[145]),

- Intelligente Benutzerschnittstellen (z.B. [130],[203], [3]),
- Software Agenten (z.B. [61], [235]),
- Geschäfts(prozess)modellierung (z.B. [46], [220], [222]),
- E-Business, Semantic Web Services (z.B. [71], [4], [12]),

Im folgenden Abschnitt skizzieren wir die Beziehung zwischen Ontologien, Texten und den Objekten der realen Welt.

6.1.2 Text Mining und Ontologien

Ontologien und Sprache sind eng miteinander verbunden. Die Analyse von Textdokumenten oder die Extraktion von Informationen erfolgt immer in einer bestimmten Sprache und bedarf des Verständnisses dieser Sprache. Ist das Verständnis nicht vorhanden und nutzt man “nur” Heuristiken, um Daten aus Texten zu extrahieren, führt dies zu vielen ungelösten Fragestellungen im Bereich des Text Mining. Die Frage nach der Wortsinnerkennung z.B. ist ein bis heute nicht zufriedenstellend gelöstes Problem (vgl. [121]). Wortsinnerkennung bedeutet, dass eine Maschine in der Lage sein soll, die Worte des Textes in einen Zusammenhang zu stellen und so den gemeinten Sinn, den Sinn, den der Autor beim Schreiben im Kopf hatte, im Text zu erkennen. Für die Erkennung des Sinnes muss die Maschine das Wort intern in einer Art Abbild erfassen, um die verschiedenen Bedeutungen eines Wortes trennen zu können. Aus diesem Abbild, dieser internen Repräsentation heraus lassen sich Schlüsse über die Bedeutung des Wortes ziehen, die wiederum auf den realen Gegenstand, der in diesem Zusammenhang mit dem Wort verbunden ist, referenzieren. Auf diese Weise wird die Bedeutung des Wortes erkannt. Wir identifizieren drei wesentliche Dinge, das Wort im Text, die Bedeutung des Wortes in der realen Welt und die interne Repräsentation der “verstehenden” Maschine.

Die Kommunikation des Menschen mit seiner Umwelt bzw. die damit verbundenen Prozesse im Gehirn unterliegen einem ähnlichen Beziehungsdreieck. Die Zusammenhänge wurden erstmals von Odgen und Richards in [180] unter dem Begriff “Meaning Triangle” bzw. “semiotisches Dreieck” zusammengefasst. Die Dreiecksbeziehung an sich ist aber schon viel älter. Abbildung 6.1 gibt die Beziehung zwischen Symbolen, Dingen und Konzepten wieder. Symbole sind Worte. Symbole referenzieren auf Konzepte - Gedanken, die mit dem Ding, dem realen Objekt verbunden sind. Es gibt keine direkte Beziehung zwischen Symbolen und Dingen, nur die indirekte. Das deutet auch die gestrichelte Linie an. Je besser die Worte einen Gedanken reflektieren und auf diesem Weg die Verbindung zu den Dingen der realen Welt herstellen, desto einfacher und klarer kann man sich ausdrücken und das reale Ding beschreiben.

Für die maschinelle Analyse bedeutet das semiotische Dreieck, dass Worte zwar eine bestimmte Bedeutung referenzieren, nicht aber direkt das gemeinte Objekt/Ding. Es ist auch nicht klar, welche Bedeutung und damit welche Dinge durch ein Wort im Text angesprochen werden. Die Erfassung der Bedeutung eines Wortes geschieht erst im Kopf des Menschen, d.h. für eine Maschine benötigt man ein Hilfsmittel, um die Bedeutungen von einfachen Worten in der Maschine ablegen und damit arbeiten zu können. Ontologien können für diese Aufgabe herangezogen werden. Sie bilden die Repräsentation, die Worte mit Konzepten und Konzepte mit den Gegenständen der realen Welt verbinden. Ontologien erlauben die Abstraktion der Konzepte von der Sprache und den Worten und referenzieren die realen Dinge/Objekte der Welt. Nutzen wir eine Ontologie zur Repräsentation unserer Objekte, können wir Objekte unterscheiden, die mit dem gleichen Wort verbunden sind.

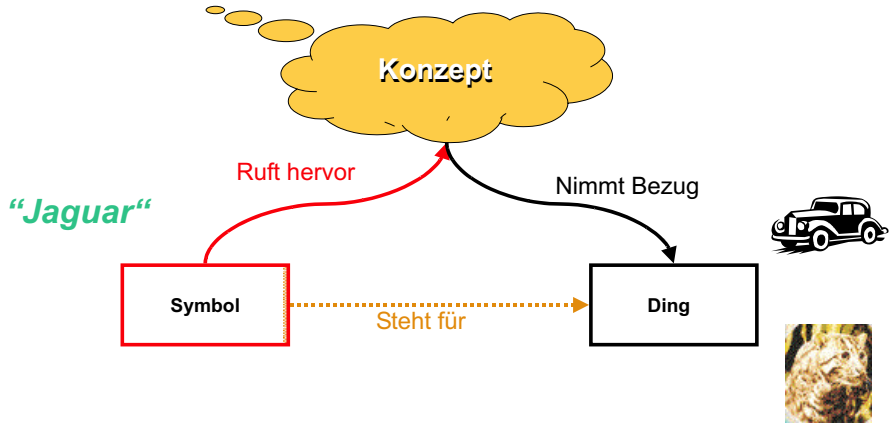

Abbildung 6.1: Das Dreieck von Ogden & Richards [180]

6.1.3 Begrifflichkeiten

Um in der Arbeit Konzepte und Begriff aus den Bereichen der Ontologien und der Formalen Begriffsanalyse (siehe Kapitel 5.5) besser auseinander halten zu können, greifen wir im Rahmen der Arbeit auf die folgende sprachliche Regelung zurück: In der englischen Literatur wird für den Begriff “Begriff” das Wort “Konzept” (engl. Concept) verwendet. In dieser Arbeit referenziert Konzept immer das Konzept einer Ontologie und der Begriff immer den formalen Begriff. Zur besseren Unterscheidung wird bei der Formalen Begriffsanalyse häufig der “formale Begriff” als Bezeichner und bei Ontologien das “Konzept einer Ontologie” verwendet, um die Zuordnung besser verständlich zu machen.

6.2 Definition einer Ontologie

Der Abschnitt beschreibt die “Karlsruher” Perspektive einer Ontologie, kurz KAON[1]. Wir folgen bei der Definition den Arbeiten von Stumme und Bozsak et.al. [213, 27]. In der Literatur existieren verschiedene “Definitionen” über das, was eine Ontologie sein sollte. Einige wurden in [93] diskutiert, wobei die wohl bekannteste die folgende von Gruber [92] ist:

> *“An ontology is an explicit specification of a conceptualization”.*

Gruber definiert eine “Ontologie als eine explizite Spezifikation einer Konzeptualisierung”. Diese Konzeptualisierung bildet ein abstraktes Modell eines Teiles unserer Welt, indem es die relevanten Konzepte einer Ontologie dieses Teiles identifiziert und benennt. Die Art und die möglichen Beschränkungen der verwendeten Konzepte werden “explizit” definiert. Die Definition von Gruber wird häufig erweitert um drei weitere Elemente:

> “An ontology is an explicit, *formal* specification of a *shared* conceptualization *of a domain of interest*”.

Mit “formal” wird das Problem angesprochen, dass eine Ontologie maschineninterpretierbar sein soll. Eine Ontologie und die damit verbundene Konzeptualisierung sollte – “shared” – sein, also

[1] Bei KAON-Framework handelt es sich zum einen um eine formale Ontologiedefinition (vgl. [213, 27]) und zum anderen, um passende Implementierungen als API und Oberfläche, die unter http://kaon.semanticweb.org/ zu finden sind.

etwas Gemeinsames darstellen und nicht nur das private Verständnis eines Individuums widerspiegeln. Dazu benötigt man ein gemeinsames Vokabular. Nur dann können Menschen miteinander kommunizieren und Wissen austauschen. Davenport fasst dies so zusammen:

> People can‘t share knowledge if they don‘t speak a common language. [44]

Mit “domain of interest” wird der Focus der verwendeten Ontologien eingeschränkt. Es steht nicht die Modellierung der Welt, sondern die Modellierung eines ausgewählten Bereiches — einer Domäne —, die von Interesse ist, im Vordergrund. Man bezeichnet Ontologien mit dieser Einschränkung als domänenspezifische Ontologien. Domänenunabhängige Ontologien konzentrieren sich nicht auf einen Bereich, sondern haben das Ziel, spezifische Zusammenhänge der Welt zu modellieren.

Alle Definitionen haben einen sehr hohen Grad der Generalisierung gemeinsam, der weit von einer präzisen mathematischen Definition entfernt ist. Der Grund für diese unpräzisen Definitionen ist der Versuch, die verschiedenen Arten von Ontologien zu erfassen. Sie zielen nicht auf eine bestimmte Methode der Wissensrepräsentation [223].

Für unsere Arbeiten benötigen wir eine präzise und detaillierte Definition einer Ontologie. Wir müssen uns daher für eine spezielle Art und Weise der Ontologierepräsentation entscheiden. Wir fassen die folgende Ontologiedefinition unter dem Akronym “KAON” (Karlsruher Ontologie) zusammen. Dazu werden wir erst den Kern einer Ontologie definieren und diesen um verschiedene Aspekte erweitern. Erste KAON-Implementierungen bestanden nur aus einer Kern-Ontologie und wurden aus diesem Grund in Kombination z. B. mit F-Logic [131], so wie es in Ontobroker [47] und OntoEdit [204] implementiert ist, aber auch mit anderen Sprachen, die logische Schlüsse erlauben, genutzt. Mittlerweile existiert auch eine erweiterte Version von KAON (vgl. [174]). Wir beziehen uns im folgenden auf [213, 27]:

Definition 8. *Eine* Kern-Ontologie (im engsten Sinne) *ist eine Struktur*

$$\mathcal{O} := (C, \leq_C, R, \sigma, \leq_R)$$

bestehend aus

- *zwei disjunkten Mengen C und R, deren Elemente als* Konzeptidentifizierer *und* Relationidentifizierer *bezeichnet werden,*
- *einer partiellen Ordnung $\leq_C$ auf C, genannt* Konzept Hierarchie *oder* Taxonomy,
- *einer Funktion $\sigma\colon R \to C^+$, genannt* Signatur,
- *einer partiellen Ordnung $\leq_R$ auf R, genannt* Relation Hierarchie. *$r_1 \leq_R r_2$ und $|\sigma(r_1)| = |\sigma(r_2)|$ impliziert $\pi_i(\sigma(r_1)) \leq_C \pi_i(\sigma(r_2))$,[2] für alle $1 \leq i \leq |\sigma(r_1)|$.*

Oft nennen wir Konzeptidentifizierer und Relationidentifizierer der Einfachheit halber *Konzepte* und *Relationen*. Für binäre Relationen definieren wir *Domain* und *Range* wie folgt:

Definition 9. *Für eine Relation $r \in R$ mit $|\sigma(r)| = 2$, definieren wir deren* Domain *und* Range *als* $\mathrm{dom}(r) := \pi_1(\sigma(r))$ *und* $\mathrm{range}(r) := \pi_2(\sigma(r))$.

Wenn $c_1 <_C c_2$, für $c_1, c_2 \in C$ gilt, dann ist c_1 ein Unterkonzept von c_2, *und c_2 ist ein* Oberkonzept von c_1. *Wenn $r_1 <_R r_2$, für $r_1, r_2 \in R$ gilt, dann ist r_1 eine* Unterrelation von r_2, *und r_2* ein Oberrelation von r_1.[3]

Wenn $c_1 <_C c_2$ und es existiert kein $c_3 \in C$ für das $c_1 <_C c_3 <_C c_2$ gilt, dann ist c_1 ein direktes Unterkonzept *von c_2, und c_2 ist ein* direktes Oberkonzept *von c_1. Wir schreiben das folgendermaßen: $c_1 \prec c_2$.* Direkte Oberrelationen *und* direkte Unterrelationen *werden analog definiert.*

[2] Mit π_i bezeichnet man die Projektion aus der Menge C^+ auf das i-te Element.

[3] Kleiner als $<_C$ stellt die verkürzte Schreibweise für $c_1 \leq_C c_2$ und $c_1 \neq c_2$ dar. Analog gilt dies auch für Relationen.

Beziehungen zwischen Konzepten und/oder Relationen, aber auch deren Beschränkungen, können in einer logischen Sprache ausgedrückt werden. Wir stellen im Folgenden eine allgemeine Definition zur Verfügung, die die Nutzung verschiedener Sprachen erlaubt.

Definition 10. *$\mathcal{L}$ bezeichnet die logische Sprache. Ein* $\mathcal{L}$-Axiomen System *für eine Ontologie $\mathcal{O} := (C, \leq_C, R, \sigma, \leq_R)$ ist ein Paar*

$$A := (AI, \alpha)$$

wobei

- *AI eine Menge ist, deren Elemente man* Axiomidentifizierer *nennt und*
- *$\alpha \colon AI \to \mathcal{L}$ eine Abbildung ist.*

Die Elemente von $A := \alpha(AI)$ bezeichnet man als Axiome.

Eine Ontologie mit $\mathcal{L}$-Axiomen *ist ein Paar*

$$(\mathcal{O}, A)$$

wobei $\mathcal{O}$ eine Ontologie und A ein $\mathcal{L}$-Axiomen System für $\mathcal{O}$ ist.

Im Folgenden steht *Ontologie* entweder für eine Kern Ontologie oder für eine Ontologie mit $\mathcal{L}$-Axiomen.

Gemäß internationalem Standard ISO 704 stellen wir für Konzepte und Relationen Namen zur Verfügung. Wir nennen sie allerdings nicht "Name" sondern Zeichen, um so allgemein wie möglich zu sein.

Definition 11. *Ein* Lexikon *für eine Ontologie $\mathcal{O} := (C, \leq_C, R, \sigma, \leq_R)$ ist eine Struktur*

$$Lex := (S_C, S_R, Ref_C, Ref_R)$$

bestehend aus

- *zwei Mengen S_C und S_R, deren Elemente man als* Zeichen für Konzepte *und* Relationen *bezeichnet,*
- *einer Relation $Ref_C \subseteq S_C \times C$, die man als* lexikalische Referenz von Konzepten *bezeichnet, wobei $(c, c) \in Ref_C$ für alle $c \in C \cap S_C$ gilt,*
- *einer Relation $Ref_R \subseteq S_R \times R$, die man als* lexikalische Referenz von Relationen *bezeichnet, wobei $(r, r) \in Ref_R$ für alle $r \in R \cap S_R$ gilt.*

Basierend auf Ref_C, definieren wir für $s \in S_C$,

$$Ref_C(s) := \{c \in C \mid (s, c) \in Ref_C\}$$

und für, $c \in C$,

$$Ref_C^{-1}(c) := \{s \in S_C \mid (s, c) \in Ref_C\} \ .$$

Ref_R und Ref_R^{-1} sind analog definiert.

Eine Ontologie mit einem Lexikon *ist ein Paar*

$$(\mathcal{O}, Lex)$$

wobei $\mathcal{O}$ die Ontologie und Lex das Lexikon für $\mathcal{O}$ ist.

Ontologien formalisieren die intensionalen Aspekte einer Domäne. Der extensionale Teil wird durch eine Wissensbasis (Knowledge␣Base) bereit gestellt. Sie enthält die Instanzen der Konzepte und Relationen.

Definition 12. *Eine* Wissensbasis *ist eine Struktur*

$$KB := (C_{KB}, R_{KB}, I, \iota_C, \iota_R)$$

bestehend aus

- *zwei Mengen C_{KB} und R_{KB},*
- *einer Menge I, deren Elemente man* Instanzbezeichner *(oder einfach* Instanzen *oder* Objekte*) nennt,*
- *einer Funktion $\iota_C \colon C_{KB} \to \mathfrak{P}(I)$, die man als* Konzeptinstanziierung *bezeichnet,*
- *einer Funktion $\iota_R \colon R_{KB} \to \mathfrak{P}(I^+)$, die man als* Relationinstanziierung *bezeichnet.*

Wie für Konzepte und Relationen stellen wir auch für Instanzen Namen zur Verfügung.

Definition 13. *Ein* Instanzlexikon *einer Wissensbasis $KB := (C_{KB}, R_{KB}, I, \iota_C, \iota_R)$ ist ein Paar*

$$IL := (S_I, R_I)$$

bestehend aus

- *einer Menge S_I deren Elemente man* Zeichen einer Instanz *nennt,*
- *einer Relation $R_I \subseteq S_I \times I$, die man als* lexikalische Referenz der Instanz *bezeichnet.*

Eine Wissensbasis mit Lexikon *ist ein Paar*

$$(KB, IL)$$

wobei KB die Wissenbasis und IL das Instanzlexikon für KB ist.

Für eine gegebene Wissensbasis kann man die Extension der Konzepte und Relationen einer Ontologie basierend auf der Konzeptinstanziierung und der Relationeninstanziierung ableiten.

Definition 14. *Sei $KB := (C_{KB}, R_{KB}, I, \iota_C, \iota_R)$ eine Wissensbasis. Die* Extension $[\![c]\!]_{KB} \subseteq I$ eines Konzeptes $c \in C$ *ist durch die folgenden rekursiven Regeln definiert:*

- $[\![c]\!]_{KB} \leftarrow \iota_C(c)$
- $[\![c]\!]_{KB} \leftarrow [\![c]\!]_{KB} \cup [\![c']\!]_{KB}$, *für* $c' < c$.
- *den Axiomen in A (falls $\mathcal{O}$ eine Ontologie mit $\mathcal{L}$-Axiomen ist).*

Die Extension $[\![r]\!]_{KB} \subseteq I^+$ einer Relation $r \in R$ *ist durch die folgende rekursive Regel definiert:*

- $[\![r]\!]_{KB} \leftarrow \iota_R(r)$
- $[\![r]\!]_{KB} \leftarrow [\![r]\!]_{KB} \cup [\![r']\!]_{KB}$, *für* $r' < r$.
- *den Axiomen in A (falls $\mathcal{O}$ eine Ontologie mit $\mathcal{L}$-Axiomen ist).*

Falls aus dem Kontext zu erkennen ist, um welche Wissensbasis es sich handelt, schreiben wir auch $[\![c]\!]$ und $[\![r]\!]$ an Stelle von $[\![c]\!]_{KB}$ und $[\![r]\!]_{KB}$.

Die folgende Definition erlaubt die Überprüfung der Konsistenz zwischen Ontologie und Wissensbasis.

Definition 15. *Eine Wissensbasis* $KB := (C_{KB}, R_{KB}, I, \iota_C, \iota_R)$ *ist* konsistent zu einer Ontologie $\mathcal{O}$*, falls alle der folgenden Bedingungen erfüllt sind:*

- $\mathcal{O}$ *ist konsistent (falls* $\mathcal{O}$ *eine Ontologie mit* $\mathcal{L}$*-Axiomen ist),*
- $C_{KB} \subseteq C$*,*
- $R_{KB} \subseteq R$*,*
- $[\![r]\!] \subseteq \prod_{c \in \sigma(r)} [\![c]\!]$*, für alle* $r \in R$*,*
- KB *ein Modell für* $A \cup \{c_1 \leq c_2 \rightarrow [\![c_1]\!] \subseteq [\![c_2]\!] \mid c_1, c_2 \in C\} \cup \{r_1 \leq r_2 \rightarrow [\![r_1]\!] \subseteq [\![r_2]\!] \mid r_1, r_2 \in R\}$ *ist.*

Die Definitionen für Ontologien und Wissensbasis bilden die Grundlage der erweiterten Repräsentation von Daten in dieser Arbeit. Die gängigen Modelle aus dem Bereich des Text Mining aber auch die Organisation von kundenbeschreibenden Kommunikationsmerkmalen lassen sich mit Hilfe von KAON modellieren.

Im nächsten Abschnitt stellen wir Vorgehensweisen zur Modellierung von Ontologien vor und gehen dabei auf vorhandene Ressourcen verschiedener Bereiche z.B. der Linguistik ein.

6.3 Modellierung von Ontologien

Die Akquisition von Ontologien bildet den Ausgangspunkt für die Nutzung als Hintergrundwissen im Clusterprozess. Dabei existieren eine Reihe von Vorgehensweisen zur manuellen und (semi) automatischen Erstellung von Ontologien, die wir im folgenden Abschnitt ansprechen werden. In Abschnitt 6.3.2 gehen wir auf domänenspezifische Ontologien und in Abschnitt 6.3.3 auf domänenunabhängige Ontologien ein. Wir verweisen auch auf Quellen, wie z.B. Thesauri, und überführen diese in eine Ontologie.

6.3.1 Manuelle und (semi-)automatische Ontologieerstellung

Die manuelle Erstellung einer Ontologie ist die einfachste aber auch aufwendigste Methode, eine Ontologie zu erstellen. Für die Unterstützung der Ontologieerstellung wurden in den letzten Jahren zahlreiche Werkzeuge entwickelt (vgl. [60, 84]). Eine Evaluierung wurde als Teil des "Evaluation of Ontology-based Tools" 2002 Workshops (EON 2002) durchgeführt [215]. APECKS, Chimaera, DOGMAModeler, KAON OImodeller, OilEd, OntoEdit, Ontosaurus, Protégé, WebODE und WEBOnto sind die bekanntesten Tools, die auf dem Workshop vorgestellt wurden. Die Werkzeuge lassen sich in drei Kategorien unterteilen: Frameorientierte [131], Bescheibungslogikorientierte [14] und Werkzeuge zur Verarbeitung natürlicher Sprache. Eine Zusammenfassung findet man in [214]. Alle Werkzeuge unterstützen die Modellierung einfacher Elemente wie Konzepte, Zeichen (lexikalische Elemente), Relationen sowie die Zuordnung von Instanzen zu Konzepten. Unterschiede gibt es bei der Repräsentation von Axiomen, die sehr abhängig von der darunterliegenden Logiksprache sind. Systeme aus dem Bereich der Beschreibungslogiken sowie einen Vergleich dieser findet man in [110]. Ein Beispiel für ein Frame-Logik basiertes System ist Ontobroker [47]. Die Unterstützung durch entsprechende Werkzeuge erlaubt die einfache Erstellung und Verarbeitung einer Ontologie. In dieser Arbeit wird die KAON-API und KAON-OImodeller passend zur KAON-Ontologie-Definition des letzten Abschnittes verwendet (vgl. [174] und Abschnitt 6.2).

Beim manuellen Erstellen einer Ontologie ist neben dem Werkzeug zur Unterstützung der Ontologieerstellung auch ein systematischen Vorgehen notwendig. Mit Hilfe einer Methodologie zur Ontologieerstellung [214] stellt man Eigenschaften wie Vollständigkeit und Konsistenz durch ein prozessorientiertes Vorgehen sicher. Nach [214] unterteilt man den Erstellungsprozess in die Phasen

Möglichkeitsstudie (Feasibility Study), Anfangsphase (Kickoff), Verfeinerungsphase (Refinement), Evaluationsphase und Anwendungs- und Evolutionsphase. Für die verschiedenen Phasen stehen zur Unterstützung des Anwenders wieder Werkzeuge zur Verfügung.

Für die Anwendung der in dieser Arbeit entwickelten Verfahren bei der Deutschen Telekom musste in einem ersten Schritt eine domänenspezifische Ontologie modelliert werden. Dabei wurde auf das Phasenmodell der OTK-Methodologie [214] zur Erstellung der Ontologie zurückgegriffen. In Kapitel 10 wird die Anwendung des Prozessmodells zur Akquisition einer kommunikationsdatenspezifischen Ontologie beschrieben.

Um den Aufwand bei der Erstellung einer Ontologie zu reduzieren, werden immer häufiger Techniken aus dem Bereich des Maschinellen Lernens, Data Mining und der Linguistik verwendet. Man fasst diesen Ansatz unter der Bezeichnung “Ontology Learning” zusammen [155]. Einen Überblick gibt [85]. Hauptziel ist die Unterstützung des Ontologieerstellers mit z.B. der automatischen Extraktion von Wortlisten aus Texten oder der Extraktion von Beziehungen zwischen Konzepten mittels Assoziationsregeln (vgl. [153]). Dazu unterscheiden [153] die Ansätze: Ontology Learning aus Texten, aus Wörterbüchern, aus Wissenbasen, aus semi-strukturierten Schemata und aus relationalen Schemata.

Allen Ansätzen ist gemein, dass sie den Anwender, der die Ontologie erstellt, nur unterstützen, die Ontologie aber nicht vollständig automatisch akquirieren. Dies erscheint vor dem Hintergrund, Ontologien zur Repräsentation von Hintergrundwissen im Knowledge Discovery zu verwenden, plausibel, da bei einem funktionierenden vollautomatischen Ansatz die Ontologien jederzeit aus den gegeben Ressourcen vollständig und korrekt zur Repräsentation des Hintergrundwissens rekonstruiert werden könnten. Die so berechneten Ontologien stellen dann nur noch eine interne Komponente eines “verstehenden” Systems dar und werden nicht mehr notwendigerweise explizit benötigt. Weitere Arbeiten diskutieren die Grenzen der automatischen Ontologieakquisition aus einer anderen Perspektive. Brewster u.a. [31] argumentieren, dass nur das Wissen, das im Fokus eines Textes steht, explizit abgelegt wird und dass Hintergrundwissen aus der Domäne zum Verständnis notwendig ist. Dabei findet man in Texten Wissen in unterschiedlichen Ebenen, von sehr allgemein bis sehr speziell, vor. Je spezifischer der Text ist, desto schwieriger wird es, den Text zu verstehen, und desto mehr Hintergrundwissen ist dafür notwendig. Bisher übernimmt der Anwender diesen Teil und verbindet die extrahierten “Wissensstücke” aus dem Text zu einer vollständigen Ontologie. Der Anwender stellt somit den entscheidenden Faktor bei der Erstellung einer Ontologie dar und wird heute ausschließlich durch Werkzeuge bei der Erstellung der Ontologie unterstützt und nicht ersetzt.

In den nächsten Abschnitten werden wir bestehende Wörterbücher und Thesauri, die wir in der Arbeit einsetzen, vorstellen. Man unterscheidet domänenspezifische und domänenunabhängige Ontologien.

6.3.2 Domänenspezifische Ontologien

Die folgenden drei Abschnitte erörtern jeweils eine domänenspezifische Ontologie. Die AGROVOC-Ontologie stammt aus dem Bereich der Landwirtschaft, die Getess-Ontologie aus dem Bereich des Tourismus und die Java-Ontologie aus dem Bereich des eLearning.

6.3.2.1 AGROVOC-Thesaurus

Der AGROVOC Thesaurus[4] ist ein fünfsprachiger multilingualer Thesaurus, bestehend aus einem kontrollierten Vokabular und einer Hierarchie. Das Thema des Thesaurus ist die Landwirtschaft im

[4] http://www.fao.org/agrovoc/

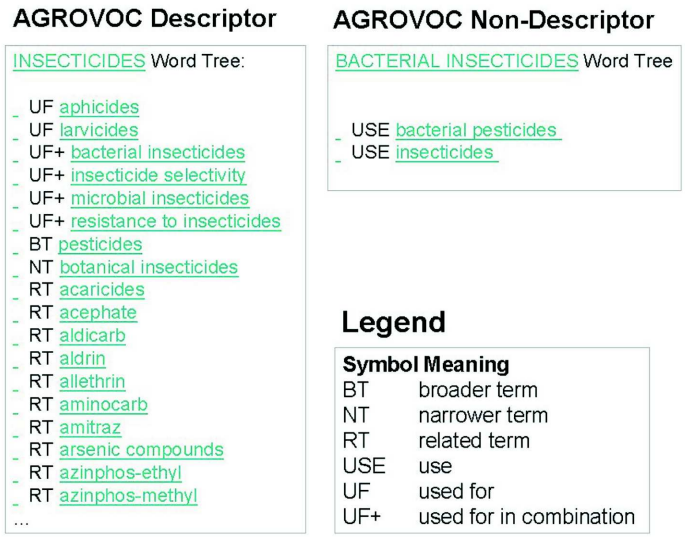

Abbildung 6.2: AGROVOC-Thesaurus: Ein Beispiel mit Descriptoren und no-Descriptoren

weitesten Sinne. Er wird zum Verschlagworten und zur Anfrage von Informationen bei der FAO (mehr zur FAO siehe Kapitel 2.3) eingesetzt. Die Hauptaufgabe des Thesaurus ist die Standardisierung des Katalogisierungsprozesses durch das bereitgestellte kontrollierte Vokabular. So informiert der Thesaurus Nutzer und Katalogisierer gleichermaßen über die Bedeutung von Schlagworten wie z.B. INSECTICIDES (man nennt dieses Wort Descriptor) und weist sie darauf hin, dieses Schlagwort an Stelle von LARVICIDES oder APHICIDES (man nennt diese Worte no-Descriptoren) zu nutzen. Descriptor und no-Descriptor stehen durch die Relationen “use” und “use for” in Beziehung miteinander (siehe Abbildung 6.2[5]).

Die Descriptoren sind in einer Taxonomie angeordnet, wobei jeder Descriptor auf einige speziellere und allgemeinere Terme verweisen kann. Mit z.B. “related term”, “use”, “used for” und “used for+” gibt es eine Reihe weiterer Beziehungen zwischen den Schlagworten des Thesaurus. Zum Beispiel zeigt “use” an, dass es sich bei dem Schlagwort um einen “non-descriptor” handelt und man den angegeben Descriptor zum Indexieren des Dokumentes verwenden soll (mehr in [146]). Weiterhin wurde jedes Schlagwort in verschiedene Sprachen übersetzt.

Der AGROVOC-Thesaurus bietet durch seinen Umfang und die vielen vorhandenen Beziehungen eine gute Ausgangsbasis für eine Ontologie. In einem ersten Schritt sind alle Informationen des Thesaurus in eine Ontologie zu konvertieren. Dazu werden alle Descriptoren zu Konzepten. Alle no-Descriptoren werden als Synonyme der Konzepte ihren entsprechenden Descriptoren zugeordnet. Sowohl Descriptoren als auch no-Descriptoren wurden in allen sechs Sprachen in die Ontologie übernommen. Bei der “used for in combination with” Relation kann man die no-Descriptoren nicht als Synonyme betrachten. Aus diesem Grund wurden sie als Konzepte der Ontologie hinzugefügt und über eine Relation “used for in combination with” (uf+) mit ihren Descriptoren in Beziehung gesetzt (Details in [146]). Für die Hierarchie wurden die Spezialisierungsbeziehungen “Broader Term” und “Narrower Term” genutzt. Auch hier ist die Übernahme aller Relationen in die Ontologie kritisch zu betrachten. In den meisten Fällen ist die Interpretation der Beziehung im Sinne der Ontologie als Hierarchie korrekt. Damit sind alle für diese Arbeit wichtigen Informationen aus dem Thesaurus in die Ontologie konvertiert. Für die Übernahme der noch fehlenden Beziehungen in die Ontologie verweisen wir auf [146].

[5] http://www.fao.org/agrovoc/

Im Ergebnis enthält die Ontologie 17513 Konzepte, die durch 177934 lexikalische Einträge beschrieben werden. Davon sind 117000 Konzeptbezeichner und 55285 Synonyme. Die maximale Tiefe der Taxonomie beträgt 8 bei einem Durchschnitt von 3.03 Konzeptebenen.

6.3.2.2 Tourismus-Ontologie

Im Getess Projekt wurde neben dem Text-Korpus (vgl. Abschnitt 2.4) auch eine umfangreiche Ontologie für den Tourismus-Bereich modelliert. Sie beschreibt neben den regionalen Zusammenhängen und Informationen über Mecklenburg-Vorpommern auch z.B. Hotels und deren Ausstattung (vgl. [137]). Die Ontologie existiert in deutscher und englischer Sprache. Wir nennen die Ontologie im Folgenden Getess-Ontologie.

Die Getess-Ontologie besteht aus $|C| = 1030$ Konzepten. Das Lexikon umfasst $Lex = 1950$ Wortstämme. Die durchschnittliche Tiefe der Taxonomie in der Ontologie ist 4.6 und der längste Pfad von der Wurzel zu einem Blatt ist 9.

6.3.2.3 Java eLearning-Ontologie

Nicola Henze beschreibt in [104] eine Ontologie für die Programmiersprache Java. Die Ontologie wurde für die Unterstützung eines offenen und adaptiven Hypermedia-Systems entwickelt und besteht aus 511 Konzepten und 12 nichttaxonomischen Relationen. Die maximale Tiefe der Taxonomie beträgt 8 bei einem Durchschnitt von 5.2. Sie enthält 505 Labels und 11 Synonyme.

6.3.3 Domänenunabhängige Ontologien

Es existieren eine Reihe domänenunabhängiger Ontologien. Neben den Top-Level Ontologien gibt es in diesem Bereich noch die so genannten Common Sense Ontologien [94], die domänenunabhängig modelliert sind. Wir stellen WordNet als eine sehr umfangreiche und mächtige Ressource der englischen Sprache vor und gehen in Abschnitt 6.3.3.2 auf das deutsche Äquivalent zur Verarbeitung deutscher Texte ein.

6.3.3.1 WordNet

WordNet[6] [167] ist eine frei verfügbare Ressource, die in jahrelanger manueller Arbeit erstellt wurde. Sie kann als Kern einer Ontologie mit Lexikon für die englische Sprache angesehen werden. Vorteil von WordNet ist das sehr umfangreiche Lexikon sowie sein genereller Charakter. Durch die Nutzung von WordNet konnten wir die sonst sehr aufwendige Modellierung für unsere Experimente vermeiden.

WordNet besteht aus so genannten Synsets, die in unserem Fall den Konzepten C der Ontologie entsprechen. Die Synsets werden von englischsprachigen Worten referenziert, die das Lexikon bilden. Wir können dieses Lexikon direkt als Lexikon Lex in die Ontologie übernehmen. Weiterhin existiert eine Hypernym/Hyponym Hierarchy. Wir nutzen diese Beziehung als IsA-Beziehung im Sinne der Ontologie. Eine genauere Beschreibung von WordNet und der enthaltenen Elemente findet man in [167][7]. WordNet, in der von uns verwendeten Version, umfasst 109377 Konzepte (Synsets) und 144684 lexikalische Einträge[8] (in WordNet Worte genannt).

[6] http://www.cogsci.princeton.edu/~wn/

[7] Auf der Webseite http://www.cogsci.princeton.edu/~wn/man1.7.1/wngloss.7WN.html stehen aktuelle Informationen zur Verfügung.

[8] Die Anzahl an lexikalischen Einträgen in WordNet ist höher, da auch morphologische Ableitungen in WordNet enthalten sind.

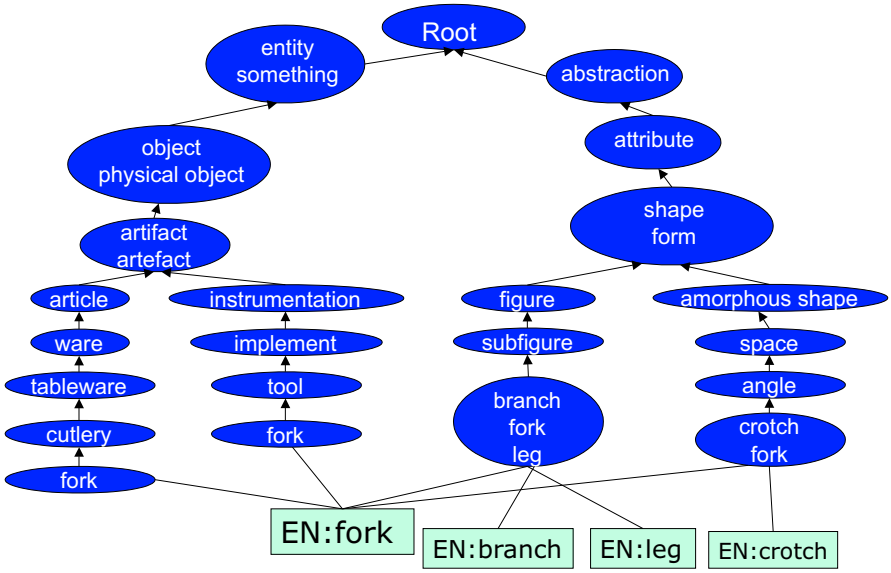

Abbildung 6.3: Auszug aus der WordNet-Taxonomie mit vier Bedeutungen des Wortes “fork”

Schauen wir uns ein kleines Beispiel genauer an. Abbildung 6.3 zeigt für das Wort “fork” (Gabel) vier mögliche Bedeutungen, sprich vier Verallgemeinerungszweige aus der Taxonomie. Die Übersetzung des Wortes “fork” erfolgte mit der in WordNet enthaltenen Ref_C-Funktion (siehe Abschnitt 6.2). In zwei Fällen handelt es sich um physikalische Objekte und in den anderen beiden Fällen um etwas Abstraktes. Der linke Ast des Baumes in Abbildung 6.3 reflektiert die Beziehung von Gabel mit Besteck. Das ist die am häufigsten mit Gabel assoziierte Bedeutung. Nur durch den Zusatz von z.B. Besteck, wird die Bedeutung des Wortes Gabel eindeutig bestimmt. Ein anderer Sinn von Gabel ist das Gabeln im Sinne von Verzweigen. In der Informatik “forked” (verzweigt/gabelt) man z.B. Prozesse. Durch die Einbettung der Konzepte in die Taxonomie kann man die Bedeutung der verwendeten Worte (lexikalischen Einträge) mittels WordNet ermitteln. Dies nutzen wir in Kapitel 8.2.3 für eine einfache Wortsinnerkennung aus.

Darüberhinaus bietet WordNet nicht nur die in Kapitel 8.2.3 benötigte einfache Übersetzungsfunktion Ref_C an. In der von WordNet angebotenen Variante liefert die Funktion nicht nur die Menge der Konzepte für einen Term, sondern eine geordnete Menge von Konzepten. Die Ordnung richtet sich nach der Auftretenshäufigkeit des angefragten Termes in der englischen Sprache. Konzepte, die für den Term alltäglicher sind, werden vor Konzepten genannt, die nicht so alltäglich sind.

6.3.3.2 GermaNet

GermaNet[9] ist das WordNet der deutschen Sprache (siehe [98]). Es besteht zur Zeit aus 41777 Konzepten (Synsets) mit 52251 lexikalischen Einträgen. GermaNet ist damit wesentlich kleiner als WordNet.

Analog zum Beispiel für WordNet aus Abschnitt 6.3.3.1 finden wir für das Wort “Gabel” in GermaNet zwei Bedeutungen. Neben der zu erwartenden Bedeutung in Beziehung zu Geschirr finden wir als weiteres Oberkonzept “nicht definite Raumeinheit” bzw. “Maßeinheit”. Die deutlich kleinere Ressource enthält leider keine Beziehung zum Substantiv “Verzweigen” oder “Teilen”. Gabeln als Verb ist hingegen enthalten und hat als Oberkonzept das Synset “teilen”.

[9] http://www.sfs.uni-tuebingen.de/lsd/

Ebenso wie WordNet bietet auch GermaNet eine erweiterte Version der Ref_C-Funktion mit Ordnung an.

Teil II

Nutzung von Hintergrundwissen

7 Subjektives Clustern

Dieses Kapitel führt die Methode “Subjektives Clustern” im Detail ein. Dabei handelt es sich um einen ontologiebasierten Vorverarbeitungsschritt zur Reduktion der Dimensionalität für das Clustern von z.B. Textdokumenten. Das folgende Kapitel gliedert sich in fünf Teile. Abschnitt 7.1 gibt eine Einführung in bzw. eine Motivation für das Subjektive Clustern. Der zentrale Abschnitt 7.2 dieses Kapitels stellt den Algorithmus “Concept Selection and Aggregation“ (COSA) zur Berechnung niedrigdimensionaler Sichten vor, wobei dessen Eigenschaften in Abschnitt 7.3 anhand von Textdokumenten diskutiert und evaluiert werden. In Abschnitt 7.4 wird eine Erweiterung des COSA-Algorithmus zur Handhabung von abhängigen Merkmalen eingeführt. Dadurch wird die Anwendung von COSA auf Kommunikationsdaten aus dem Bereich der Telekommunikation möglich. Abschließend vergleichen wir in Abschnitt 7.5 das Subjektive Clustern mit verwandten Ansätzen aus der Literatur. Wir folgen in diesem Kapitel der Arbeit [113].

7.1 Einführung

Im folgenden Abschnitt wird die Methode des Subjektiven Clusterns eingeführt und die Ziele werden beschrieben. Abschnitt 7.1.2 definiert die Begriffe rund um COSA, und in Abschnitt 7.1.3 werden aus den allgemeinen Textvorverarbeitungsschritten, die in Kapitel 4 vorgestellt werden, die für COSA spezifischen abgeleitet.

7.1.1 Ziele des Subjektiven Clusterns

Subjektives Clustern bezeichnet eine in dieser Arbeit neu entwickelte Methode, bei der es nicht nur um die Berechnung von Clustern mittels statistischer Größen geht, sondern die auch subjektive Informationen des Anwenders erfasst und bei der Clusterberechnung berücksichtigt. Die Clusterergebnisse sollen niedrigdimensional sein, um die Interpretierbarkeit und Verständlichkeit für den Anwender zu steigern. In Kapitel 1.3.1 wurde die Problemstellung aus Sicht eines praktischen Beispiels eingeführt und motiviert. Der Lösungsansatz Subjektives Clustern verfolgt die Ziele:

- Subjektivität der Clusterung,
- Verständlichkeit der Ergebnisse und
- Reduktion der hohen Dimensionsanzahl.

Mit *Subjektivität* wird ausgedrückt, dass jeder Anwender eine eigene Vorstellung über ein zu erzielendes Clusterergebnis hat. Auch die Bedeutung der Merkmale ist für jeden Anwender unterschiedlich. Daher wird es nicht nur *eine* korrekte Clusterung sondern meist mehrere Clusterungen geben, die verschiedene Merkmale zum Clustern verwenden. Die Güte jeder dieser Clusterungen wird von jedem Anwender unterschiedlich beurteilt werden. Aus diesem Grund ist es auch wichtig, dass die Präsentation der Clusterergebnisse für den Anwender leicht *verständlich* ist. Die Reduktion der *hochdimensionalen* (Text-) Vektoren auf eine überschaubare und verständliche Menge von Merkmalen stellt die Grundlage für leicht zu verstehende Clusterergebnisse dar. Dabei müssen die

Merkmale für den Menschen interpretierbar bleiben, so dass statistische Techniken zur Dimensionsreduktion, wie z.B. LSI (siehe Abschnitt 4.4), nicht angewendet werden können. Die bei LSI abgeleiteten Merkmale lassen sich nicht mit einfachen Worten charakterisieren und können so für die Interpretation von Clusterergebnissen nicht eingesetzt werden.

Das Subjektive Clustern setzt zur Dimensionsreduktion auf strukturelle Beziehungen zwischen den Merkmalen, die in einer Ontologie abgelegt sind. Gleichzeitig bietet die Ontologie die Grundlage für die Präsentation und Auswahl einer Clusterung, da nur Merkmale zum Clustern verwendet werden, die in der Ontologie vorkommen. Die Zusammenhänge der Merkmale bzw. Konzepte in der Ontologie erlauben eine einfache Navigation. Der Abschnitt 7.2 stellt den Algorithmus COSA zur Berechnung von Sichten (Details zu Sichten, siehe nächster Abschnitt) vor. Jede Sicht besteht aus einer Menge von Konzepten der Ontologie, wobei die Sichten durch unterschiedliche Konzepte repräsentiert werden, die gleichzeitig auch in den zu clusternden Objekten vorkommen. Jede Sicht wird mit Hilfe eines Standardclusterverfahrens wie z.B. KMeans geclustert. Nach der Anwendung von COSA und KMeans steht dem Anwender eine Menge von Sichten mit den entsprechenden Clusterergebnissen zur Verfügung. Der Anwender kann eine oder mehrere dieser Sichten auswählen und drückt dadurch implizit seine Interessen an der Sicht aus. Durch die Merkmale der ausgewählten Sicht berücksichtigt die Clusterung die Präferenzen des Anwenders. Eine gewählte Sicht erfüllt damit die drei gesetzten Ziele. COSA stellt den zentralen Mechanismus zur Erzeugung von subjektiven, leicht verständlichen und niedrigdimensionalen Clusterungen zur Verfügung.

Im nächsten Abschnitt erörtern wird die Begriffe Sicht und Aggregat und fixieren ihre Verwendung in der Arbeit.

7.1.2 Sicht und Aggregat

Der Begriff “Sicht” (engl. View) ist aus dem Bereich der Datenbanken entlehnt (siehe [1, 17]) und wird schon in der Architektur eines Datenbankmanagmentsystems erwähnt. Man unterscheidet bei der Architektur eines Datenbankmanagmentsystems drei Ebenen: die physikalische, die logische und die externe (konzeptuelle) Ebene. Die externe Ebene bietet dem Anwender die so genannten Sichten auf die logische Ebene an und bildet die benutzerbezogene Abstraktionsebene. In [1] wird alles das als Sicht bezeichnet, was sich mit einer Anfrage gegen die Datenbank berechnen lässt und im Ergebnis eine Relation hat. Diese Relation kann mit einem Namen versehen werden und man spricht dann von einer Sicht.

Mit Hilfe der Ontologie werden in dieser Arbeit Merkmale (jedes Merkmal hat ein passendes Attribut in einer Relation der Datenbank)[1] ausgewählt und zu einer niedrigdimensionalen Anfrage kombiniert. Diese Anfrage wird als Sicht in der Datenbank abgelegt. Die Sicht wird zur Berechnung der Clusterungen verwendet. Die Datenbank übernimmt dabei die Vorverarbeitung der Daten, d.h. sie berechnet die Zusammenfassung der Daten auf die niedrigdimensionalen Merkmalsvektoren. Dieser Schritt des Zusammenfassens der Daten bezeichnet man auch als *Aggregation*.

Ein Aggregat ist das Ergebnis der Anwendung einer Aggregationsfunktion wie z.B. Summe, Anzahl oder Durchschnitt für ein ausgewähltes Merkmal einer Relation auf einer Menge von Objekten. Es fasst die detaillierten Daten zusammen. Häufig werden diese Funktionen zur Analyse von Daten eines Data Warehouses eingesetzt (vgl. [17]). Wir werden in der Arbeit die Aggregationsfunktionen zur Berechnung des Supports im COSA-Algorithmus (siehe Abschnitt 7.2) verwenden und die Informationen der Dokumente zur Auswahl der “wichtigen” Merkmale zusammenfassen. Außerdem benötigt man die Aggregationsfunktion zur Analyse von Kommunikationsdaten in Abschnitt 7.4.

[1] Auch Textdokumente lassen sich nach der Vorverarbeitung zu Konzeptvektoren in einer Datenbank abspeichern.

Im folgenden Abschnitt werden die Vorverarbeitungsschritte speziell für die Evaluierung des Subjektiven Clusterns anhand von Textdokumenten vorgestellt.

7.1.3 Einfache Textvorverarbeitungsstrategien

Der erste Schritt zum Clustern von Textdokumenten ist die Überführung der Texte in eine geeignete Repräsentation. Dies geschieht im Allgemeinen während der Vorverarbeitungsphase. Für Textdokumente existiert eine große Anzahl an Repräsentationsmechanismen. Das einfachste und gebräuchlichste Modell ist das "Bag of Words"- oder allgemeiner "Bag of Terms"-Modell (siehe Abschnitt 4.2.1), das wir im Folgenden als Grundlage für Referenzclusterungen verwenden werden. Die Referenzclusterungen basieren auf Repräsentationen ohne Hintergrundwissen und bilden die Grundlage zum Vergleich mit dem Subjektiven Clustern. Die nachfolgenden Abschnitte beschreiben die beiden Repräsentationen SiVer und TES auf der Basis des "Bag of Words". Als Datensatz wird der Getess-Korpus (siehe Abschnitt 2.4) verwendet.

7.1.3.1 Einfache Vektorrepräsentation (SiVer)

Die einfache Vektorrepräsentation *SiVer* (Simple Vector Representation) entspricht dem "Bag of Words" Modell bestehend aus allen extrahierten Termen ohne weitere Vorverarbeitung (es erfolgte keine Gewichtung und auch keine Normalisierung der Vektoren). Beim Getess-Korpus handelt es sich um 46947 unterschiedliche Terme, d.h. um 46947-dimensionale Vektoren. Da diese Repräsentation per se einige Nachteile aufweist (z.B. hohe Anzahl an Merkmalen), wird im folgenden Abschnitt 7.1.3.2 mit TES eine deutlich verbesserte Repräsentation vorgestellt, die nur wenige "wichtige" Terme auf der Basis von bekannten Information-Retrieval-Maßen berücksichtigt und als niedrigdimensionale Repräsentationen deutliche Vorteile, wie eine bessere Verständlichkeit, gegenüber SiVer aufweist.

7.1.3.2 Term-Selektion (TES)

Die *Term-Selektion* (TES), der zweite Vorverarbeitungsansatz für Dokumente ohne Hintergrundwissen, basiert auf SiVer. Um die Vergleichbarkeit mit den niedrigdimensionalen Sichten von COSA zu verbessern, berücksichtigt diese Repräsentation aber nur die besten Terme. D.h., die Anzahl der Terme entspricht der Merkmalsanzahl $\dim$, die bei Start von COSA vorgegeben wird. Die resultierende Repräsentation besteht durch den Reduktionsschritt ebenfalls nur noch aus wenigen Dimensionen, was einen besseren Vergleich zulässt. Die Reihenfolge der Terme wird auf der Basis des Information-Retrieval-Maßes $tfidf$ (siehe Kapitel 4.2.5.1) berechnet.

Konkret wird zur Berechnung von TES die Menge aller Terme $t \in T$ für einen Korpus D bestimmt. Stoppworte (siehe Kapitel 4.2.3) aus einer gegebenen Liste werden in der Menge T nicht berücksichtigt. TES wählt die Teilmenge $\tilde{T} \subseteq T$ der Terme $t \in \tilde{T}$ mit den größten $W(t)$-Werten,

$$W(t) := \sum_{i:=1...|D|} tfidf(i,t), \tag{7.1}$$

und erzeugt so einen $|\tilde{T}|$-dimensionalen Termvektor für jedes Dokument d.

Die mit SiVer und TES abgeleiteten Termvektoren stellen wortbasierte Repräsentationsformen dar und bilden die Basis für den Vergleich der mit Hilfe von Hintergrundwissen und COSA berechneten und geclusterten konzeptbasierten Sichten.

7.2 Concept Selection and Aggregation (COSA)

Das Verfahren COSA umfasst zwei Phasen. Die erste Phase ist domänenspezifisch und beinhaltet die Abbildung von Objekten der realen Welt auf die passenden Konzepte einer Ontologie. Die zweite Phase beinhaltet die eigentliche Auswahl der Konzepte für die Menge von Sichten. Diese Phase ist domänenunabhängig und verwendet die Ontologie und die zugehörigen Daten zur Berechnung der Sichten.

Im folgenden Abschnitt wird das Prinzip der Abbildung von Objektmerkmalen auf Konzepte anhand der Abbildung von Worten auf Konzepte erläutert. Ein weiteres Beispiel findet man in Abschnitt 10.1.5.1, wo das Vorgehen anhand der Abbildung von Kommunikationsmerkmalen der Telekomkunden auf eine Ontologie erläutert wird. Für die Abbildung der Worte auf die Konzepte setzen wir flache und effiziente Verfahren zur Verarbeitung der natürlichen Sprache ein. Beim Telekombeispiel reicht dafür ein einfaches Lexikon.

7.2.1 Abbildung von Termen auf Konzepte

Die Abbildung von Termen aus Texten auf die Konzepte der Ontologie erfolgt durch das Modul SMES (Saarbrücken Message Extraction System). Dabei handelt es sich um ein System zur flachen Sprachverarbeitung aus Texten (siehe [177]). COSA nutzt von SMES den *Tokenizer* basierend auf regulären Ausdrücken und die *lexikalische Analyseeinheit*, welche ein allgemeines *Wörterbuch* und ein so genanntes *Domänen-Lexikon* $\mathcal{L}$ (der domänenspezifische Teil des Lexikons folgt der Definition 11) einschließt.

SMES geht wie folgt vor: Der Tokenizer analysiert den Text und identifiziert einfache Worte, komplexere Ausdrücke wie “$20.00” oder “United States of Amerika” und expandiert bekannte Abkürzungen. Das Wörterbuch enthält mehr als 120000 Wortstämme. Während der lexikalischen Analyse wird das Wörterbuch u.a. zur morphologischen Analyse der Terme z.B. zur Identifikation von zusammengesetzten Worten und zur Bestimmung von Eigennamen verwendet. Das Ergebnis dieses ersten Prozessschrittes, der einfachen linguistischen Analyse des Textes, liefert eine Menge von Wortstämmen passend zum Domänen-Lexikon $\mathcal{L}$. Das Domänen-Lexikon enthält die Abbildung der Wortstämme auf die entsprechenden Konzepte C aus der Ontologie $\mathcal{O}$ und legt so die Funktion Ref_C (siehe Definition 11) fest. Auf diese Weise kann der Ausdruck “Hotel Schwarzer Adler” mit dem Konzept HOTEL in Verbindung gebracht werden.

SMES extrahiert für jedes Dokument d einen Termvektor $\vec{t_d}$ und übersetzt ihn in einen Konzeptvektor $\vec{c_d}$. Die Häufigkeit eines Konzeptes $|C|$ entspricht der kumulierten Häufigkeit der assoziierten Terme $|t|$ spezifiziert durch $Ref_C(t)$, die in SMES implementiert ist. Im Ergebnis stellt dieser Schritt für jedes Dokument einen Konzeptvektor mit entsprechenden Häufigkeiten zur Verfügung.

7.2.2 Eine Heuristik zur Erzeugung “guter” Aggregate

Während der Abbildung von Worten/Termen eines Textes auf die Konzepte kommt es zu einer ersten Reduktion der Dimensionalität des Termvektors. Dies liegt an den in jeder Sprache vorhandenen Synonymen und der geringeren Anzahl an Konzepten gegenüber der Anzahl an Termen. Trotz der deutlich geringeren Anzahl an Konzepten gegenüber der SiVer-Repräsentation ist der resultierende Konzeptvektor noch immer sehr groß. Dies führt weiterhin zu Problemen beim Clustern (vgl. [25]).

Um sowohl Clusterstrukturen entdecken zu können als auch leicht interpretierbare Clusterergebnisse zu berechnen, benötigen wir eine Heuristik zur weiteren Reduktion der Dimensionalität des Merkmalsraumes. Dafür sollen Merkmale verwendet werden, die weder zu häufig noch zu selten vorkommen (diese Annahme liegt auch dem tfidf-Maß zu Grunde). Unser Ansatz im Rahmen von

COSA ist im Algorithmus 7.1 durch die Funktion `GenerateConceptViews` realisiert. Auf der Basis einer Ontologie und der Daten, wird eine Menge von Sichten mit der Dimensionalität dim berechnet. Neben diesen Eingabegrößen benötigt der Algorithmus ein Startkonzept. Das Startkonzept bildet den Ausgangspunkt der *Top-Down* gerichteten und datengetriebenen Navigation durch die Ontologie, im Speziellen entlang der Heterarchie. Die Top-Down-Navigation folgt der Idee, dass es sich lohnt, Konzepte mit hohem *Support* (Definition siehe unten) in ihre Unterkonzepte zu zerlegen und die Daten anhand dieser Unterkonzepte im Detail zu analysieren. Konzepte, die keinen oder nur sehr geringen Support haben, müssen nicht im Detail (d.h. durch ihre Unterkonzepte) repräsentiert werden. Konzepte mit sehr geringem Support werden aus der Repräsentation entfernt. Mit dem Ersetzen von Konzepten durch ihre Unterkonzepte erweitert man den Konzeptvektor und durch das Löschen von Konzepten mit geringem Support wird die Größe des Konzeptvektors reduziert. So ist man in der Lage, die Dimensionalität des Vektors zu steuern. Den kompletten Algorithmus in Pseudocodenotation findet man in Algorithmus 7.1 und Tabelle 7.1 spezifiziert die verwendeten Funktionen. Im Ergebnis erzeugt der Algorithmus Listen mit Konzepten, die nach Abschnitt 7.1.2 Sichten genannt werden. Die Konzepte einer Sicht kommen weder zu oft noch zu selten vor.

Algorithmus 7.1 GenerateConceptViews$(dim, \mathcal{O}, \text{ROOT}, D)$

Input: Dimensionalität dim, Ontologie $\mathcal{O}$, Startkonzept ROOT, Objektmenge D
Output: Menge von Sichten

```
1:  Agenda:= [ROOT];
2:  repeat
3:    Elem:= First(Agenda);
4:    Agenda:= Rest(Agenda);
5:    if Leaf(Elem) then
6:      continue := FALSE;
7:    else
8:      if Atom(Elem) then
9:        Elem:= Subconcepts(Elem);
10:     end if
11:     NewElem:= BestSupportElem(Elem);
12:     RestElem:= Elem \ NewElem;
13:     if ¬Empty(RestElem) then
14:       Agenda:= SortInto(RestElem, Agenda);
15:     end if
16:     Agenda:= SortInto(NewElem, Agenda);
17:     if Length(Agenda)> dim then
18:       Agenda:= Butlast(Agenda);
19:     end if
20:   end if
21:   if Length(Agenda) = dim then
22:     Output(Agenda);
23:   end if
24: until continue = FALSE
```

Bevor wir den Algorithmus im Detail erläutern, benötigen wir noch eine Definition für den Support eines Konzeptes c (Support(c)). Dazu definieren wir den *direkten Support* Support(d, c) eines Konzeptes c für ein Dokument d über dessen Konzepthäufigkeit cf (siehe Abschnitt 4.1).

Tabelle 7.1: Liste aller in Algorithmus 7.1 verwendeten Funktionen

Subconcepts(C)	liefert eine willkürlich geordnete Liste aller direkten Unterkonzepte von C.
Support(C)	vgl. Gleichung 7.4.
Support(*ListC*)	berechnet die Summe des Supports Support(C) über alle Konzepte C in *ListC*.
SortInto(*Element*, *List*)	sortiert das Konzept oder die Liste der Konzepte aus *Element* gemäß dem Support(*Element*) in die Konzeptliste *List* ein und entfernt alle doppelten Einträge.
BestSupportElem(*List*)	liefert das *Element* der Liste *List* mit dem maximalen Support(*Element*).
[*Element*]	erstellt eine Liste mit einem *Element*.
First(*List*), Rest(*List*)	liefert das erste Element bzw. alle Elemente bis auf das erste Element einer Liste.
Atom(*Element*)	liefert wahr, wenn *Element* keine Liste aus Konzepten ist.
Leaf(*Element*)	liefert wahr, wenn *Element* ein Konzept ohne Unterkonzepte ist.
List $\setminus$ *Element*	löscht das *Element* aus der *List*.
Length(*List*)	gibt die Länge von *List* zurück.
Butlast(*List*)	gibt *List* ohne das letzte Element gemäß der internen Sortierung zurück.

$$\text{Support}(d, c) := \sum_{b \in H(c,\infty)} \text{cf}(d, b), \tag{7.2}$$

wobei

$$H(c, r) := \{c' | \exists c_1, \ldots, c_i \in C \colon c' \prec c_1 \prec \cdots \prec c_i = c,\ 0 \leq i \leq r\} \tag{7.3}$$

für eine gegebenes Konzept c die r nächsten Unterkonzepte der Taxonomie liefert. Insbesondere liefert $H(c, \infty)$ alle Unterkonzepte von c.

$$\text{Support}(c) := \sum_{d \in D} \text{Support}(d, c) \tag{7.4}$$

Gleichung 7.4 berechnet den Support eines Konzeptes c in Bezug auf alle Dokumente.

Als Input benötigt `GenerateConceptViews` eine Menge von Dokumenten D, eine Ontologie $\mathcal{O}$ mit dem passenden Startkonzept, z.B. ROOT, sowie die gewünschte maximale Dimensionszahl dim. Als Ergebnis liefert `GenerateConceptViews` eine Menge von Sichten. Jede Sicht besteht aus einer Menge von Merkmalen, die Konzepte oder Mengen von Konzepten sein können. Besteht ein Merkmal aus Mengen von Konzepten, so berechnet sich das Merkmal durch Bildung der Summe der entsprechenden Häufigkeiten dieser Konzepte. Jede Sicht hat eine fest vorgegebene Anzahl von Merkmalen, nämlich die Anzahl der Dimensionen *dim* .

Die Variable *Agenda* enthält die Beschreibung der aktuellen Liste von Merkmalen/Konzepten, um die Sichten basierend auf der Dokumentmenge D zu erzeugen. Beim Aufruf wird der *Agenda* immer das Startkonzept (siehe Algorithmus 7.1 Zeile 1, z.B. ROOT) übergeben. Nehmen wir an, das

Konzept ROOT hätte die drei Unterkonzepte [UNTERKUNFT, URLAUB, STADTRUNDFAHRT]. Die Konzepte sind gemäß dem Support sortiert. Die aktuelle Liste der Konzepte wird verändert, indem man das erste Konzept mit dem höchsten Support aus der *Agenda* entfernt (Zeilen 3 und 4) und dieses, sofern es nicht ein Blattkonzept ist, in die Unterkonzepte verzweigt (Zeile 9). Es erfolgt nur eine binäre Verzweigung, um die Anzahl der hinzugefügten Konzepte pro Schritt zu beschränken.

Führen wir das Beispiel fort und nehmen an, dass das Konzept UNTERKUNFT die Unterkonzepte [HOTEL, GÄSTEHAUS, JUNGENDHERBERGE] hat. Für das Verzweigen, wählen wir das Konzept mit dem höchsten Support (Zeile 11) z.B. HOTEL und fassen die anderen beiden Konzepte in einem Merkmal zusammen (Zeile 12). Die Liste [GÄSTEHAUS, JUNGENDHERBERGE], die das neue Merkmal bildet, wird wie ein normales atomares Konzept behandelt.[2] HOTEL und [GÄSTEHAUS, JUNGENDHERBERGE] werden beide in die *Agenda* eingefügt. Die Ordnung der *Agenda* entsprechend dem Support wird dabei aufrecht erhalten (Zeile 14 und 16). Nun ist die *Agenda* folgendermaßen zusammengesetzt[3]: [UNTERKUNFT, [GÄSTEHAUS, JUNGENDHERBERGE], HOTEL, STADTRUNDFAHRT].

Besteht die *Agenda* nach der letzten Verzweigung aus mehr Merkmalen als in der Inputvariable dim spezifiziert (Zeile 17), wird das letzte Merkmal (Zeile 18) aus der *Agenda* entfernt. Entspricht die Anzahl der Merkmale in der *Agenda* der spezifizierten Dimensionsanzahl dim (Zeile 21), dann wird die aktuelle *Agenda* der Ausgabemenge (Zeile 22) hinzugefügt. Dadurch entsteht eine Sicht. Durch die fortschreitende Verfeinerung von Konzepten mit hohem Support und dem Löschen von Konzepten mit niedrigem Support wird die aktuelle Agenda geändert. Auf diesem Wege werden weitere Sichten erzeugt. Jede Sicht unterscheidet sich in mindestens einem Merkmal von allen anderen Sichten. Der Algorithmus 7.1 betrachtet solche Konzepte im Detail, die den stärksten Support aufweisen. Er liefert dabei nicht "die" eine dim-dimensionale Sicht auf die Objektmenge, sondern eine Menge von Sichten, die ihrerseits wieder unterschiedliche Blickwinkel auf die ursprünglichen Daten wiedergeben.

Die Vektoren jeder Sicht repräsentieren nur einen Teil der Informationen eines kompletten Konzeptvektors. Der Informationsverlust durch die Merkmalsreduktion wird sich nicht vermeiden lassen. Nicht immer ist die Nutzung aller Informationen sinnvoll und eine Fokussierung auf die wesentlichen und wichtigen Informationen bzw. Merkmale kann die Clusterergebnisse verbessern. Die Auswahl der "wichtigen" Merkmale übernimmt in diesem Ansatz der Algorithmus 7.1, wobei nicht nur eine sondern mehrere wichtige Merkmalsmengen ausgewählt werden. Durch die Präsentation mehrerer fokussierter Clusterergebnisse kann der gesamte Informationsverlust reduziert werden. Gleichzeitig ermöglicht man dem Anwender die Auswahl aus einer Vielzahl von relevanten Clusterungen.

Im Folgenden stellen wir die Ergebnisse einer vergleichenden empirischen Untersuchung vor. Dazu berechnen wir die Clusterergebnisse für Textdokumente eines realen Datensatzes auf der Basis von COSA und auf der Basis der beiden Referenzrepräsentationen SiVer und TES und analysieren die Ergebnisse mit Hilfe von statistischen Maßen.

7.3 Evaluierung von COSA auf Textdokumenten

Dieser Abschnitt beschreibt eine empirische Evaluierung von COSA und vergleicht dazu COSA mit den beiden Vorverarbeitungsstrategien SiVer und TES (siehe Abschnitt 7.1.3). Zum Vergleich der Clusterergebnisse nutzen wir das Silhouetten-Maß (siehe Abschnitt 5.3.4.2) und den mittleren

[2] Man könnte sich auch vorstellen, dass die Ontologie um ein "künstliches" Konzept erweitert wird. Dies enthält die Liste, die das neue Merkmal darstellt. Das Merkmal wird auf diese Weise wieder durch ein Konzept repräsentiert.

[3] Die Reihenfolge der Konzepte ist fiktiv.

quadratischen Fehler (siehe Abschnitt 5.3.4.1). Im nächsten Abschnitt gehen wir auf die Ziele der Evaluierung ein. Abschnitte 7.3.2, 7.3.3, 7.3.4 und 7.3.5 geben die Ergebnisse wieder. Wir fassen die Ergebnisse in Abschnitt 7.3.6 zusammen.

7.3.1 Ziele

COSA produziert eine Menge von niedrigdimensionalen Repräsentationen für einen Datensatz. Das führt zu einer Menge von Clusterergebnissen, nicht zu einem einzelnen Clusterergebnis. Ein Vergleich zu einer einzelnen vorgegebenen *objektiven* Klassifikation, wie z.B. den Klassenlabels beim Reuters-Korpus, widerspricht dem Ziel des Subjektiven Clusterns und kann die Ergebnisse nur schlecht bewerten. Daher wählen wir zwei statistische Maße, den Silhouetten-Koeffizienten (siehe Abschnitt 5.3.4.2) und den MSE (siehe Abschnitt 5.3.4.1), um die verschiedenen Clusterergebnisse miteinander zu vergleichen.

Als Evaluierung wird eine empirischen Studie auf einem realen Textkorpus durchgeführt. Man berechnet zwei Referenzclusterungen auf der Basis von SiVer und TES und vergleicht sie mit den Clusterergebnissen basierend auf COSA. Für die folgenden Untersuchungen verwenden wir den Getess-Korpus (siehe Abschnitt 2.4) und die Getess-Ontologie (siehe Abschnitt 6.3.2.2). Neben der linguistischen Vorverarbeitung durch SMES (siehe Abschnitt 7.2.1) wenden wir auf den Text-Korpus die bekannten Vorverarbeitungsschritte wie das Extrahieren von Wortstämme und das Entfernen von Stoppworten an (siehe Abschnitt 4.2). Außerdem werden alle Termhäufigkeiten logarithmiert (mehr dazu in Abschnitt 4.2.6).

Als Clusterverfahren setzen wir KMeans ein (siehe Abschnitt 5.4.1). In einigen Vorstudien untersuchten wir auch verschiedene Heuristiken, um gute Startlösungen für KMeans zu identifizieren. Für den hochdimensionalen Bereich sind Verbesserungen für die Startlösung, wie in [29] beschrieben, bekannt. Die Unterschiede sind im niedrigdimensionalen Raum laut Silhouetten-Koeffizient sehr gering, so dass wir auf die Verwendung von speziellen Verfahren verzichten.

Als Abstandsmaß kommt die Euklidmetrik (siehe Abschnitt 5.2) zum Einsatz.

7.3.2 Vergleich von SiVer, TES mit COSA

Die Ergebnisse der Evaluierung mit dem Silhouetten-Koeffizienten zeigen in den meisten Fällen eine Verbesserung von KMeans basierend auf COSA im Vergleich zum Standardansatz KMeans mit TES. KMeans basierend auf SiVer war durch den hochdimensionalen Raum extrem behindert und der Silhouetten-Koeffizient immer 0 — es wurde keine Clusterstruktur im Datensatz entdeckt.

Abbildung 7.1 zeigt den Silhouetten-Koeffizienten (SC) für eine feste Anzahl von Merkmalen (15) und Cluster (10) für alle Varianten, sprich KMeans angewendet auf SiVer, TES und die Sichten von COSA. Wie schon angedeutet sind die Ergebnisse für die SiVer-Vorverarbeitung extrem schlecht. TES schneidet mit $SC = 0.16$ etwas besser ab, aber die Interpretation des Silhouetten-Koeffizienten zeigt praktisch keine Struktur an. Die Cluster können nicht klar getrennt werden. Die Anwendung von COSA ergibt für die gegebene Ontologie 89 Sichten. In der Abbildung sind die Sichten nach dem Silhouetten-Koeffizienten sortiert. Die Sicht mit dem besten Ergebnis von $SC = 0.48$ ist deutlich besser als das Standardverfahren TES und erlaubt auch eine klare Unterscheidung der Cluster im Datensatz.

Wie eingangs erwähnt wollen wir anhand eines zweiten Gütemaßes die Clusterungen miteinander vergleichen. Wir nutzen dafür den mittleren quadratischen Fehler (MSE) aus Abschnitt 5.3.4.1. Da dieses Maß nur Vergleiche bei gleicher Clusteranzahl und gleicher Dimension zulässt, können nur die beiden Varianten COSA und TES miteinander verglichen werden.

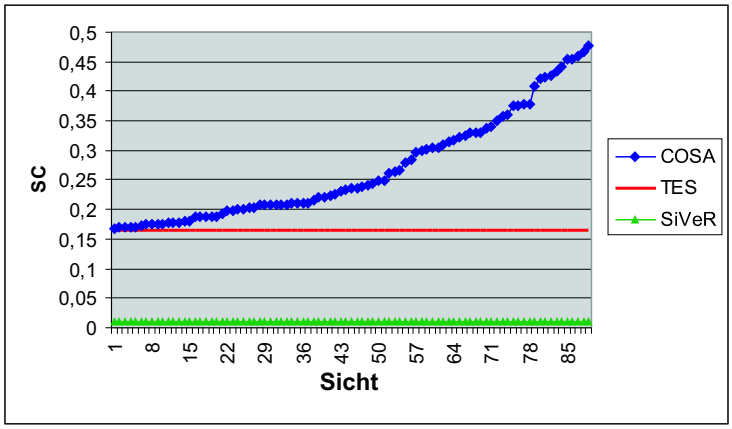

Abbildung 7.1: SiVer und TES im Vergleich zu 89 Sichten von COSA anhand des Silhouetten-Koeffizienten für $|\mathbb{P}| = 10; dim = 15$

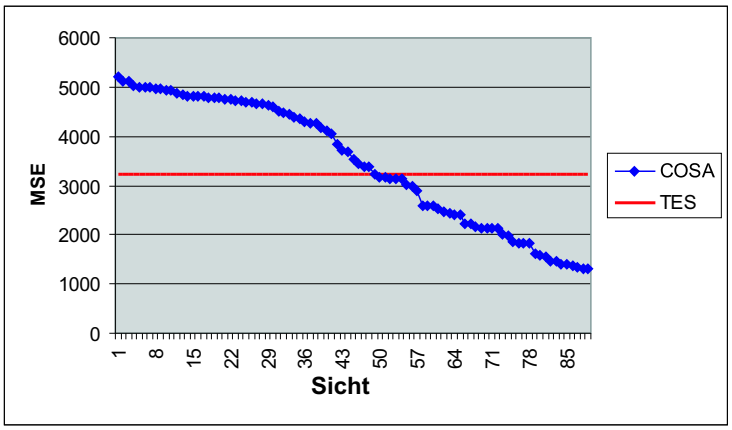

Abbildung 7.2: Vergleich TES mit den 89 Sichten erzeugt von COSA mittels MSE für $|\mathbb{P}| = 10; dim = 15$

Die Ergebnisse mit MSE zeigen zum Teil bessere Ergebnisse für KMeans mit COSA gegenüber TES. 49 der Sichten von COSA sind laut MSE schlechter als TES aber 40 von ihnen zum Teil beträchtlich besser. Das Diagramm in Abbildung 7.2 zeigt die zugehörigen Ergebnisse, wobei TES als Ausgangspunkt bei 3240 und der beste Wert für COSA bei deutlich niedrigeren 1314 liegt.

Nachdem die besten Sichten von COSA deutlich bessere Ergebnisse liefern als TES, soll durch die nächsten Experimente der Einfluss der Dimensionalität des Merkmalsraumes und der Anzahl der Cluster untersucht werden. Bei der Dimensionalität des Merkmalsraumes ist zu erwarten, dass mit steigender Anzahl der Dimensionen die Güte der Clusterung abnimmt (vgl. [25]). Startet man mit zwei Clustern und analysiert den Verlauf der Güte bei steigender Clusteranzahl, so steigt auch die Clustergüte. Ab einer bestimmten Clusteranzahl sollte die Güte wieder fallen.

7.3.3 Variation der Merkmalsanzahl

Für die folgenden Experimente variieren wir im ersten Schritt die Anzahl der Merkmale bzw. Dimensionen dim und wählen $dim = 10, 15, 30, 50, 100$ bei konstanter Anzahl an Clustern $|\mathbb{P}| = 10$. Abbildung 7.3 zeigt die Abhängigkeit zwischen der Merkmalsanzahl und der Güte der Clusterung. Die Linie für COSA gibt den Wert für den Silhouetten-Koeffizienten der besten Sicht, erzeugt durch den Algorithmus `GenerateConceptViews`, wieder. Man sieht in Abbildung 7.3, dass sowohl für TES als auch für COSA die Güte der Ergebnisse sinkt. Für höherdimensionale Datensätze war das auch zu erwarten (vgl. [25]). In jedem Fall liefert COSA im Vergleich zu TES das bessere Ergebnis.

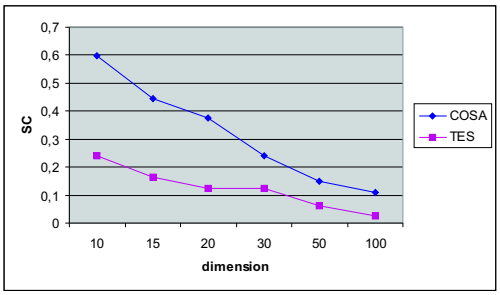

Abbildung 7.3: Vergleich von TES und der besten Sicht von COSA mittels Silhouetten-Koeffizient für $|\mathbb{P}| = 10$ und $dim = 10, 15, 30, 50, 100$

Beim Vergleich der Ergebnisse von TES und COSA verwenden wir für COSA immer das beste Ergebnis. Um die schlechten Ergebnisse von COSA aus der Ergebnismenge zu eliminieren, kann man eine untere Schranke einführen und so den Algorithmus in Bezug auf die schlechten Ergebnisse optimieren. Nach unseren Beobachtungen wäre leicht eine Steigerung der Ergebnisse möglich, indem man z.B. Sichten ausschließt, die zu viele allgemeine Konzepte, wie z.B. THING oder INTANGIBLE (vgl. Abbildung 7.5), enthalten. An dieser Stelle sei noch erwähnt, dass nicht alle Aspekte einer Sicht durch Kennzahlen wie den Silhouetten-Koeffizienten erfasst werden können. Persönliche Präferenzen oder Interessen machen die Clusterung einer bestimmten Sicht für den Anwender wesentlich besser/interessanter als der Silhouetten-Koeffizient vermuten lässt. Auch dies ist ein weiterer Grund, warum wir keine Sicht aus der Evaluierung ausgeschlossen haben.

Abschließend lässt sich aus unserer Erfahrung ableiten, dass der Anwender die Anzahl der Merkmale in Abhängigkeit zum aktuellen Problem angeben sollte. Der Anwender ist ein entscheidender Faktor, da er anschließend die Ergebnisse verstehen und interpretieren muss. Die Dimensionalität spielt hierfür eine wesentliche Rolle. Im Allgemeinen kann man aus den Ergebnissen mit unserem realen Datensatz die folgende obere Schranke für die Anzahl der Merkmale ableiten: Ein Sinken des Silhouetten-Koeffizienten unter 0.25 zeigt eine extrem geringe Strukturierung der Clusterung an. Dies geschieht bei ca. 30 Merkmalen. Die Nutzung von mehr als 30 Merkmalen erscheint aus dieser Perspektive wenig sinnvoll.

7.3.4 Variation der Clusteranzahl

Für das Experiment haben wir die Anzahl der Cluster $|\mathbb{P}|$ zwischen 2 und 100 variiert und die Merkmalsanzahl bei $dim = 15$ fixiert. Abbildung 7.4 zeigt das Ergebnis. Auch hier haben beide Funktionen einen ähnlichen Verlauf. Mit steigender Anzahl der Cluster steigt auch der Silhouetten-Koeffizient leicht an. Dies hat seine Ursache in der steigenden Zahl an Clustern, die genau auf einen Punkt fallen.

Erstaunlicherweise scheint die Anzahl der Cluster nur geringen Einfluss auf das Ergebnis zu haben. Die Güte der Clusterung sinkt bei sehr hoher Clusteranzahl nicht wie erwartet, so dass man die laut Silhouetten-Koeffizient beste Anzahl der Cluster nicht bestimmen kann.

Der extreme Abfall in der Kurve zwischen 2 und 4 Clustern weist auf zwei gut separierte große Cluster hin. Genauere Analysen zeigen, dass es einen Cluster - mit Dokumenten ohne zugehöriges Konzept in der Repräsentation - gibt. Diese Dokumente werden durch den $\vec{0}$-Vektor repräsentiert. Die Dokumente auf dem Nullpunkt lassen sich sehr gut vom Rest der Menge trennen und führen zum Verlauf der Funktion in Abbildung 7.4.

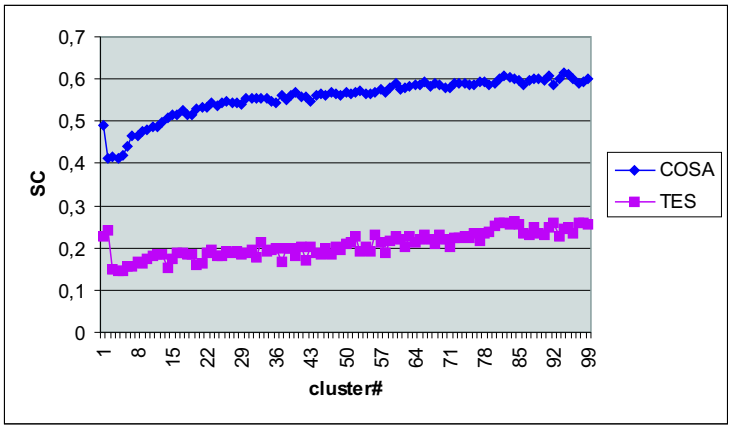

Abbildung 7.4: Vergleich von TES und der besten Sicht von COSA mittels Silhouetten-Koeffizient für $|\mathbb{P}| = 2 \ldots 100$ und $dim = 15$

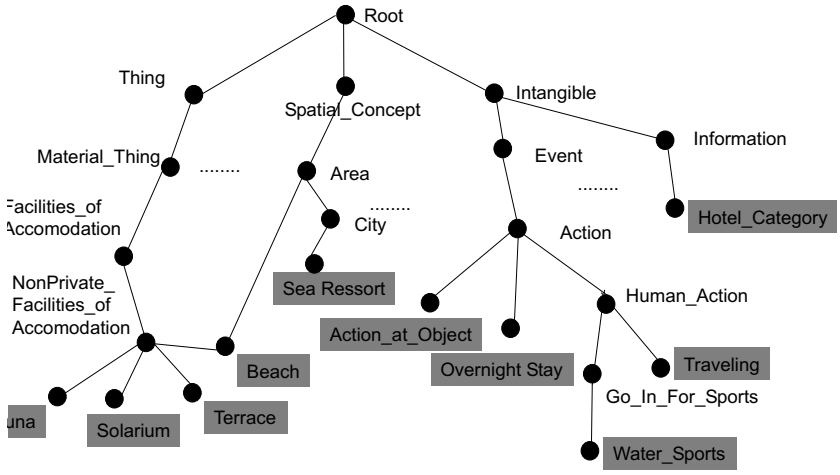

Abbildung 7.5: Eine Beispielsicht erzeugt von COSA

7.3.5 Beispiel einer Sicht

Bevor wir die Ergebnisse kurz zusammenfassen, soll ein konkretes Beispiel einer typischen Sicht helfen, die Art der Ergebnisse besser zu veranschaulichen. Dazu haben wir die beste vom Algorithmus `GenerateConceptView` erzeugte Sicht für $|\mathbb{P}| = 10$ und $dim = 10$ mit einem Silhouetten-Koeffizienten von 0.598 ausgewählt. Die Sicht umfasst die Konzepte:

SAUNA, SOLARIUM, TERRACE, BEACH, SEA_RESSORT, ACTION_AT_OBJECT, OVERNIGHT_STAY, WATER_SPORTS, TRAVELING, HOTEL_CATEGORY

Der Vergleich von Listen mit Konzepten mag dem Anwender eine Einsicht in die unterschiedlichen oder auch nicht sehr unterschiedlichen (wenn sich Sichten sehr ähnlich sind) Clusterergebnisse geben. Viel besser greifbar wird die Vorstellung aber, wenn man sich die relevanten Konzepte der Ontologie hervorhebt, wie das in Abbildung 7.5 anhand eines Teils der Getess-Ontologie zu sehen ist.

In Abbildung 7.5 sind die "wichtigen" Konzepte alle Unterkonzepte von NONPRIVATE_FACILITIES_OF_ACCOMODATION und die Konzepte SEA RESSORT, ACTION_AT_OBJECT, OVERNIGHT STAY, WATER_SPORTS, TRAVELING und HOTEL_CATEGORY d.h. Konzepte die zum Clustern der Dokumente verwendet wurden, grau

unterlegt. Alle ausgewählten Konzepte werden durch einen annähernd gleichen Support gestützt. Das heißt, dass das Konzept HOTEL_CATEGORY (also Hotelkategorien wie drei, vier oder fünf Sterne) ungefähr die gleiche Aussagekraft zum Clustern dieser Dokumente besitzen, wie alle Unterkonzepte von NONPRIVATE_FACILITIES_OF_ACCOMODATION (nichtprivate Ausstattung der Unterkunft). Diese Aussage wird auch durch die guten Clusterergebnisse gestützt.

7.3.6 Vergleich SiVer, TES und COSA

Aus den Experimenten des Kapitel 7.3 konnten wir eine Reihe interessanter Erkenntnisse gewinnen, die sich in folgende Punkte zusammenfassen lassen. Zunächst wurde eine Reihe unserer Erwartungen, die wir z.B. nach der Literaturrecherche hatten, erfüllt:

- Clustern im hochdimensionalen Raum ist wesentlich schwieriger und führt zu schlechteren Ergebnisses als Clustern mit wenigen Dimensionen.
- Clusterergebnisse im hochdimensionalen Raum sind durch die vielen Merkmale extrem schwer zu interpretieren. Versucht man trotzdem die Ergebnisse zu interpretieren, so wendet man dazu Techniken zur Reduktion der Dimensionalität wie z.B. Projektion oder Aggregation an. Das führt unweigerlich zu einem Informationsverlust und zur Beeinträchtigung der Verständlichkeit.
- Neben dem Interpretieren des Inhaltes ist das Bezeichnen (Labeln) von Clustern eine der wichtigen aber sehr schwierigen Aufgaben. Nutzt man die Konzepte einer Ontologie zum Labeln, so erleichtert dies die Interpretation für den Anwender. Obwohl wir keine Untersuchung der Nutzbarkeit (usability study) durchgeführt haben, glauben wir, dass unsere Erfahrung stark genug ist, um diese Aussagen zu rechtfertigen.

Einige der folgenden Ergebnisse waren nicht von Anfang an offensichtlich:

- Aggregate mit wenigen Dimensionen verbessern nicht per se die Clusterergebnisse. Schlechte Ergebnisse bei der Evaluierung stammen oft von Sichten, die zu viele sehr allgemeine Konzepte wie MATERIELLES DING oder IMMATERIELLES enthalten.
- Sichten mit Blattkonzepten oder Konzepten in der Nähe von Blättern führten oft zu guten Ergebnissen. In diesen Aggregaten finden wir einen substantiellen Anteil an Dokumenten die durch den Nullvektor $\vec{0}$ repräsentiert werden. Dieser Umstand lässt sich leicht erklären mit der Tatsache, dass unter Berücksichtigung bestimmter Interessen des Anwenders, repräsentiert durch die Auswahl entsprechender Konzepte, diese Dokumente einfach nicht relevant sind. Man könnte sie auch einfach a priori aus dem Clusterprozess ausschließen.
- Häufig fanden wir bei der Analyse einzelner Cluster von COSA mit dem Silhouetten-Koeffizienten, dass Cluster gut vom Rest der Dokumente getrennt werden konnten und dass diese Trennung auch leicht verständlich ist. Andere Cluster wiederum sind nur schlecht und unverständlich vom Rest getrennt. Wir vermuten, dass für die Interpretation weitere Merkmale bzw. alternative Sichten herangezogen werden müssten. Auch hier sind umfangreiche Studien mit verschiedenen Anwendern nötig, um genauere Einblicke in die Thematik zu erhalten. Der subjektive Charakter dieser Aufgabe lässt eine andere Evaluierung kaum zu. Die vorgestellten Methoden helfen aber, die in den Dokumenten enthaltenen Informationen auf ganz unterschiedliche Art und Weise zugänglich zu machen.

Unsere Ergebnisse unterstützen die allgemeine Aussage, dass im niedrigdimensionalen Raum oft Clusterstrukturen identifiziert werden können, da hier die Nachbarschaftsbeziehungen bedeutungs-

tragend sind (vgl. [25]). Unser Ansatz stellt eine ausgewählte Anzahl von Sichten in unterschiedlichen, aber klar verständlichen Unterräumen zur Verfügung unter Nutzung des Standardclusterverfahrens KMeans. Im Vergleich zu anderen Ansätzen wie dem Clustern auf der Basis von dim ausgewählten Termen, geordnet durch $tfidf$, schneidet unser Verfahren gut ab. Weiterhin erlauben die gewählten Konzepte dem Anwender eine einfachere und umfangreichere Interpretation der Ergebnisse auf der Basis ganz unterschiedlicher Sichten.

7.4 Erweiterung von COSA zum Analysieren von Kommunikationsdaten

In diesem Abschnitt wollen wir die Ideen der erweiterten Version von COSA vorstellen. Motiviert ist die Erweiterung durch die Eigenschaften der gesprächsbeschreibenden Merkmale von Telekommunikationsdaten. Details dazu findet man in Abschnitt 10.1.5.2. Wir folgen mit unseren Ausführungen der Arbeit von [179] und gehen im Folgenden kurz auf die wesentlichen Ideen ein. Dazu definieren wir spezielle Konzeptnotationen, die als Arbeitskonzepte bezeichnet werden und die Verbindung zwischen der Ontologie und COSA darstellen, sowie die Notation von Kreuzkonzepten und der Verfeinerung von Kreuzkonzepten als Erweiterung von COSA.

Im Unterschied zur Anwendung von COSA auf Textdokumenten (der einfachen Version von COSA) bieten die Kommunikationsdaten die Möglichkeit, nicht nur eine sondern mehrere beschreibende Größen zu nutzen (siehe Abschnitt 10.1.5.1). Des Weiteren finden neben der Summe als Aggregationsfunktion z.B. auch das Maximum, Minimum oder der Durchschnitt Anwendung. Wir müssen daher die Berechnung des Supports (Gleichung 7.2) allgemeiner definieren, die Konzepthäufigkeit cf anpassen und auf diesem Wege die Möglichkeit zur Spezifikation der gewünschten Informationen schaffen. Ferner benötigen wir eine neue Version der Funktion $\mathsf{SubConcepts}(C)$.

7.4.1 Notation von Konzepten und Kreuzkonzepten

Um die verschiedenen Aggregationsfunktionen, die beschreibende Größe und den verwendeten Begriff weiterhin durch ein Konzept repräsentieren zu können, haben wir uns entschieden, das Zeichen eines Konzeptes S_C, d.h. die lexikalische Information zu strukturieren. Alternativ hätte man mit mehreren Ontologien und der Verknüpfung dieser arbeiten können. Durch die gewählte Variante konnten wir aber den Algorithmus COSA weitestgehend übernehmen. Wir nennen die Ontologie mit den erweiterten Zeichen auch Arbeitsontologie.

Zur leichteren Unterscheidung führen wir so genannte Arbeitskonzepte ein. Ein Arbeitskonzept ist das Gleiche wie ein Konzept, nur dass die Zeichen C dieser Konzepte einem Muster folgen. Sie bestehen aus dem Aggregationsfunktionsteil Agg und dem Konzeptteil C. Bei der späteren Anwendung wirkt der Konzeptteil wie ein Filter und sucht aus der Datenbank alle Datensätze, die z.B. der Definition von “Nebenzeit” entsprechen. Der Aggregationsfunktionsteil spezifiziert die verwendete Funktion zum Zusammenfassen der ausgewählten beschreibenden Größe.

Definition 16. *Wir definieren den Aufbau des Zeichen S_C eines Arbeitskonzeptes $c \in C$ wie folgt:*

$$S_C := Agg\_\_c$$

wobei Agg die Aggreationsfunktion der beschreibenden Größe (z.B. sum(dauer) oder avg(Verbindungen)) umfasst und c ein Konzept aus der Menge aller Konzepte der Ontologie $\mathcal{O}$ ist. Agg besteht immer aus einer Aggregationsfunktion, die eine Entsprechung in der Datenbank

hat, und einer beschreibenden numerischen Größe cf, die in der Ontologie als Konzept definiert sein muss und ebenfalls eine Entsprechung in der Datenbank hat.

Analog werden die Kreuzkonzepte definiert. Kreuzkonzepte zeichnen sich durch die Nutzung einer Menge von Konzepten aus, die alle unabhängige beschreibende Größen des Kunden darstellen. Um eine Kundenrepräsentation ohne Wiederholungen bzw. teilweiser Mehrfachberechnung (wie in Kapitel 10.1.5.2 ausgeführt) zu erhalten, müssen die Konzepte zu einem Merkmal kombiniert werden. Dies lässt sich in der Ontologie mittels Mehrfachvererbung erreichen. Da nicht jedes Merkmal mit jedem vorab kombiniert werden soll und es einfacher ist, die relevanten Kombinationen auf einem abstrakten Level außerhalb der Ontologie zu definieren, wurde die Definition 16 von S_C wiederum erweitert.

Definition 17. *Wir erweitern das Konzept $c \in C$ aus Definition 16 und ersetzen es durch das Kreuzkonzept kk wie folgt:*

$$S_C := Agg\_\_kk$$

wobei $kk \in KK$ ein Kreuzkonzept aus der Menge aller Kreuzkonzepte einer Ontologie $\mathcal{O}$ ist. Ein Kreuzkonzept besteht aus Konzepten $C \in \mathcal{O}$, die durch das Zeichen _x_ verbunden sind. Ein kk ist daher definiert als:

$$kk := \{c_1\_x\_c_2\_x\_ \ldots c_n\},$$

wobei $c_1, c_2, \ldots, c_n \in \mathcal{O}$ sind. Agg spezifiziert den Aggregationsfunktionsteil des Konzeptes.

Der Aggregationsfunktionsteil Agg wurde bisher implizit durch die Gleichungen 7.2 und 7.4 für den Support definiert. Zur flexibleren Wahl der Aggregationsfunktion und der beschreibenden Größen führen wir diese Funktion hier explizit ein. Man erhält die Berechnung des ursprünglichen Supports nach Gleichung 7.4, wenn man “sum” wieder als Aggregationsfunktion verwendet. Denkbar sind aber auch Funktionen wie z.B. “max” oder “min”.

Die Konzepthäufigkeit cf kann nun auf verschiedene beschreibende Größen wie z.B. “Dauer” oder die “Anzahl der Verbindungen” abgebildet werden. Bisher referenziert cf nur auf die Häufigkeit eines Termes. Das entspricht bei den Kommunikationsdaten der Anzahl der Verbindungen. Die Funktion Support liefert die Ergebnisse nun entsprechend der im Konzept spezifizierten Werte.

Laut Definition entspricht damit das Konzept sum(dauer)__ZEITFENSTER einem einfachen Konzept der Arbeitsontologie und sum(dauer)__ZEITFENSTER_x_LAND einem Kreuzkonzept. Dabei findet man das Konzept ZEITFENSTER_x_LAND und die passenden Einzelkonzepte ZEITFENSTER und LAND in der Ontologie wieder. Der Support würde in diesem Fall durch die Summation der Verbindungsdauern (sum(dauer)) über die einzelnen Konzepte und deren Unterkonzepte ermittelt.

Den Zusammenhang zwischen einfachen Konzepten und Kreuzkonzepten macht Abbildung 7.6 deutlich. Die Konzepte Zeitfenster und Land lassen sich wie in der Abbildung dargestellt in Unterkonzepte zerlegen. Für die Charakterisierung der beschreibenden Merkmale repräsentiert durch die Konzepte stehen bei den einfachen Konzepten nur die letzte Zeile bzw. Spalte zur Verfügung. Alle Konzepte der letzten Zeile verkörpern zusammen 100 % der beschreibenden Größe, wie z.B. Verbindungsdauer. Gleiches gilt für die Konzepte der letzten Spalte. Streichen wir jetzt ein Konzept aus dieser einfachen Repräsentation, wird die Information gegenüber allen anderen Konzepten unterrepräsentiert. Ursache ist die doppelte Zählung der Gespräche (Zeile und Spalte beachten dabei jeweils alle Gespräche). Beim Streichen der Gespräche werden diese nun nicht gänzlich gestrichen, sondern nur noch durch ein Konzept repräsentiert. Eine bessere Repräsentation würde jedes Gespräch eines Kunden genau einmal berücksichtigen. Die Kreuzkonzepte bilden die Grundlage für eine solche Repräsentation. Sie werden von den inneren Zellen der Tabelle in Abbildung 7.6 wiedergegeben. Löschen wir eine Zelle, werden auch alle Gespräche der Kunden in der Repräsentation

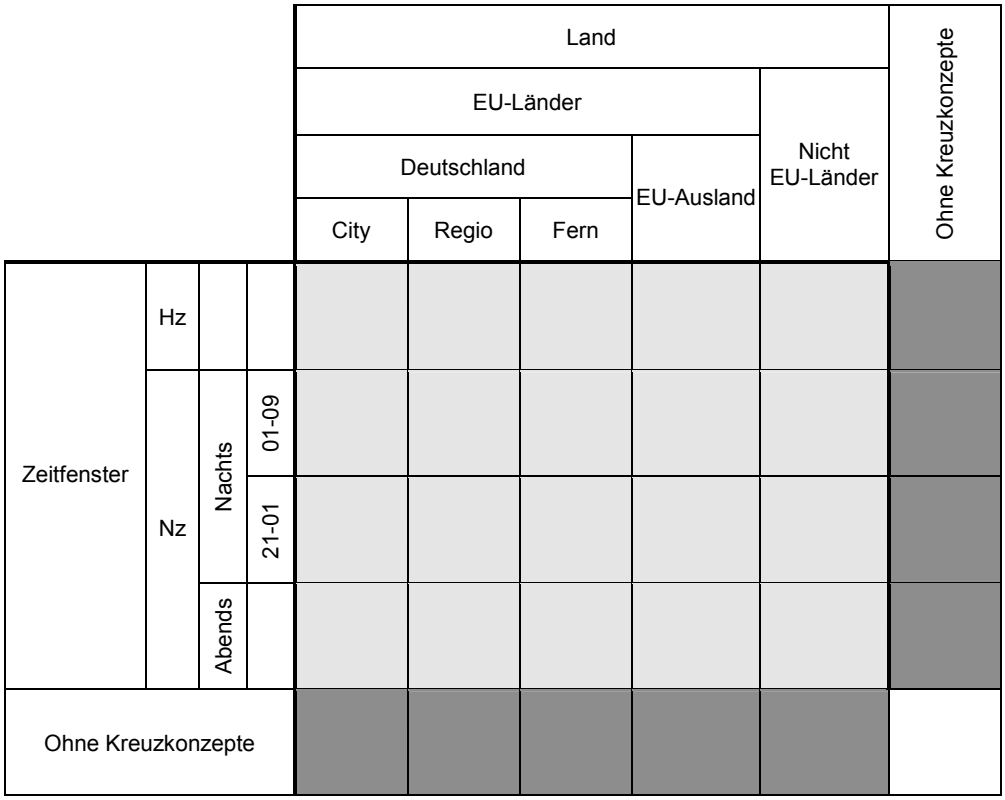

Abbildung 7.6: Vergleich von Kreuzkonzepten mit einfachen Arbeitskonzepten

nicht mehr berücksichtigt. Da alle Gespräche nur einmal berücksichtigt sind, kommt es nicht zu den schon beschriebenen Verzerrungen. Weiterhin können wir aus den Kreuzkonzepten jederzeit die Information der einfachen Konzepte berechnen.

Das Quadrat in der ersten Zeile und ersten Spalte der Abbildung 7.6 repräsentiert das Kreuzkonzept HAUPTZEIT_X_CITY. Es ist ein Unterkonzept des Kreuzkonzeptes ZEITFENSTER_X_LAND.

Beide Repräsentationen haben ihre Vor- und Nachteile. Sollen bestimmmte durch einfache Konzepte repräsentierte Merkmale hervorgehoben werden, ist es sinnvoll diese auch explizit in die Repräsentation zu integrieren (so z.B. bei personalisierten Sichten). Andererseits können sie zu ungewollten Eigenschaften, wie hoch korrelierten Merkmalen führen. Durch die Kombination von einfachen Konzepten zu Kreuzkonzepten stehen uns alle Gespächsinformationen, wie wir dies auch schon aus dem Bereich der Textdokumente kennen, zur Verfügung. Gleichzeitig können wir aber uninteressante Bereiche löschen, ohne auf Verzerrungen in der Repräsentation Rücksicht nehmen zu müssen. Die Auswirkungen auf den COSA-Algorithmus und der damit verbundenen Exploration des Suchraumes widmet sich das nächsten Abschnitt im Detail.

7.4.2 Kreuzkonzepte — die Erweiterung von COSA

Die primäre Erweiterung besteht in der Einführung der Kreuzkonzepte. Sie führt zu einem größeren Suchraum, der sich auf die Methode SubConcepts(C) auswirkt. Sie besteht nun nicht mehr aus einer einfachen Anfrage an die Ontologie, sondern muss alle Kreuzkonzepte durchsuchen. Dazu müssen alle Unterkonzepte aller einfachen Konzepte gebildet und der Support berechnet werden. Die Unterkonzeptkombination mit dem größten Support wird ausgewählt. Die Unterkonzepte der einfachen Konzepte werden erst zu diesem Zeitpunkt gruppiert. Sowohl Ontologie als auch Da-

ten haben Einfluss auf die berechneten Sichten. Schauen wir uns das anhand von Algorithmus 7.2 genauer an:

Algorithmus 7.2 Algorithmus zur Berechnung der Unterkonzeptzerlegung bei Kreuzkonzepten: SubConcepts(C)

INPUT: Kreuzkonzept C
OUTPUT: Konzepte

```
1:  maxsupport = 0;
2:  for all c ∈ SimpleConcept(C) do
3:     RefConcept= refineConcept(c);
4:     for all refConcept ∈ RefConcept do
5:        if Support(refConcept) > maxsupport then
6:           maxsupport = Support(refConcept);
7:           maxRefConcept = RefConcept;
8:        end if
9:     end for
10: end for
11: return maxRefConcept;
```

Input des Algorithmus 7.2 ist ein Kreuzkonzept C. Zeile 2 iteriert über alle im Namen des Konzeptes S_C repräsentierten einfachen Konzepte (SimpleConcept(C)) und verfeinert jedes temporär in Zeile 3. Dazu liefert die Funktion refineConcept alle Unterkonzepte des aktuellen einfachen Konzeptes. Für die Konzepte der aktuellen Verfeinerung RefConcept wird nun geprüft, wie hoch der Support (Zeile 5) der daraus ableitbaren Kreuzkonzepte refConcept ist. Wir merken uns die Verfeinerung RefConcept, die den höchsten Support enthält. Diese stellt gleichzeitig das Ergebnis dar.

Im letzten Schritt wird der Einfluss der Daten auf die Auswahl der Verfeinerung deutlich. Waren die Unterkonzepte beim einfachen COSA nur durch die Ontologie fixiert, wird das zu verfeinernde Konzept bei der erweiterten Version durch die Daten bestimmt. Zentrale Idee ist nach wie vor immer die Konzepte zu splitten, die einen hohen Support haben, da man hierdurch auf eine detailreichere Darstellung und so auf eine genauere Analyse der Kunden durch diese Sicht hofft. Im Gegenzug wählen wir beim Split immer das Konzept mit dem größten Support, um nicht gleich im Graph zu einem (uninteressanten) Blatt mit sehr kleinem Support abzusteigen.

Schauen wir uns noch kurz den Spezialfall der einfachen Konzepte an. Angenommen, wir hätten nur ein einfaches Konzept im Kreuzkonzept. Der Algorithmus liefert in diesem Fall die Unterkonzepte dieses Konzeptes zurück. Damit verhält er sich wie die Funktion SubConcept(C) des einfachen COSA Algorithmus 7.1.

7.5 Verwandte Ansätze

Alle Clusteransätze deren Ähnlichkeits- oder Distanzberechnung auf hochdimensionalen Vektoren beruhen, haben Probleme mit dem gleichen mathematische Phänomen (vgl. [25, 106]). Alle Objekte sind gleich weit von allen anderen Objekten im Raum entfernt. Man kann zwar “gute” Cluster in solchen Räumen auf der Basis von L_p Metriken berechnen, leider spiegeln die Cluster jedoch nicht eine vorhandene Struktur des Datensatzes wider. Die Punkte eines Clusters im Raum sind sich nicht wirklich ähnlicher als zu vielen anderen Punkten im Raum. Für Text-Clustering Ansätze führt dies

zur Forderung, dass in den Prozess mehr Hintergrundwissen einfließen muss. Neue Ähnlichkeitsmaße oder die Konstruktion von entsprechenden Unterräumen als Grundlage der Clusterung stellen zwei mögliche Varianten zur Lösung des Problems dar.

Es existieren Ansätze, für das Clustern in automatisch erzeugten Unterräumen. Auf der einen Seite werden zur Berechnung der "guten" Unterräume statistische Maße benutzt, die dann auch zu guten Clusterergebnissen führen:

- Hinneburg & Keim zeigen in [107] wie man mit Projektionen die Effektivität und Effizienz der Clusteraufgabe steigern kann. In ihrer Arbeit wird der Einfluss von Projektionen auf die Steigerung der Geschwindigkeit von Clusteralgorithmen deutlich herausgearbeitet. Im Gegensatz zu unserer Arbeit geht es bei Ihnen nicht um die Qualität der Cluster in Bezug auf die Offenlegung der internen Struktur der Daten, die eine Clusterung wiedergeben sollte, sondern um die Beschleunigung von Zugriffen auf die Daten in einer Datenbank. Daher steht die Interpretierbarkeit der Ergebnissen nicht im Interesse der Autoren. Die Methode führt im Gegensatz zum Subjektiven Clustern zu schwer verständlichen Ergebnissen.

- Ein Clusterverfahren, das automatisch Unterräume mit maximaler Dimensionalität berechnet, wird von Agrawal et al. in [7] vorgestellt. Der Algorithmus heißt CLIQUE. Clusterbeschreibungen werden in Form von minimalen DNF-Ausdrücken präsentiert. Mehr zur Reduktion der Dimensionalität mittels automatischer statistischer Verfahren findet man in 5.6.6.

- Ein aus der Statistik bekannter Ansatz ist die Hauptkomponentenanalyse, die hier zur Dimensionsreduktion eingesetzt werden kann [59]. Bei der Hauptkomponentenanalyse ersetzt man die vorhandenen Merkmale durch entsprechend viele Hauptkomponenten, die im Wesentlichen durch eine Linearkombination der ursprünglichen Merkmale bestimmt werden. Die Linearkombination macht die Interpretation der anschließend zu berechnenden Cluster wiederum schwer.

- Schuetze and Silverstein stellen in [194] umfangreiche Forschungsergebnisse zu Projektionstechniken für das effiziente Clustern von Textdokumenten vor. So wenden sie verschiedene Projektionstechniken zur Steigerung der Performance von Clusterverfahren an, bei gleichzeitig stabiler Clusterqualität. Sie unterscheiden in ihrer Arbeit zwischen lokaler und globaler Projektion. Bei der lokalen Projektion wird jedes Dokument in einen eigenen Unterraum projiziert, während bei der globalen Projektion die relevanten Terme für alle Dokumente mittels LSI (Latent Semantic Indexing), eingeführt von [48], bestimmt werden.

- McCallum u.a. stellen in [161] einen zweistufigen Ansatz zum Clustern hochdimensionaler Daten vor. In einem ersten Schritt fassen sie auf der Basis eines einfachen und schnell zu berechnenden Abstandsmaßes die Objekte zu überlappenden Gruppen zusammen. Diese Gruppen werden im zweiten Schritt mit einem Standardverfahren wie z.B. KMeans entgültig geclustert. Dieser Ansatz könnte in dieser Arbeit alternativ zu KMeans verwendet werden. Auch eine Kombination mit unserem Ansatz wäre denkbar, indem man die ontologiebasierte Dimensionsreduktion ebenfalls zweistufig gestaltet und jeweils unterschiedliche Merkmale in beiden Schritten verwendet.

Zusammenfassend erlauben die diskutierten Ansätze die Reduktion der Dimensionsanzahl häufig auf Kosten der Verständlichkeit der Clusterergebnisse. Die Clusteransätze stellen die Grundlage für eine Kombination mit dem Subjektiven Clustern dar. Vorteil des Subjektive Clustern ist, das es nicht nur eine Clusterung auf einem niedrigdimensionalen Datensatz bietet, sondern mehrere Sichten und

Clusterungen, deren Merkmale in der Struktur der Ontologie eingebettet sind. Der Anwender hat die Möglichkeit unter diesen Sichten zu wählen und dadurch seine Präferenzen auszudrücken.

Weiterhin kommt es gerade in realen Anwendungen häufig vor, dass die statistisch abgeleitete optimale Projektion, wie eben vorgestellt, sich nicht mit der für Menschen für diese Aufgabe am besten passenden Projektion deckt. Eine solche Aufgabe könnte das Finden einer bestimmten Information in einer großen Menge an Dokumenten sein. Der Anwender präferiert in einem solchen Fall explizit repräsentiertes Hintergrundwissen als Grundlage, um die Clusterung zu steuern und die Ergebnisse der Clusterung zu verstehen.

Hinneburg et al. [108] bezeichnen dieses allgemeine Problem als domänenspezifische Optimierungsaufgabe. Sie schlagen eine interaktive Visualisierungsumgebung vor, um bedeutungsvolle Projektionen zusammen mit dem Anwender abzuleiten. Unser Ansatz kann zur automatischen Lösung von einigen Teilen dieser Aufgabe herangezogen werden. Dem Anwender wird durch die Domänenontologie und der darauf aufbauenden Sichten nicht der gesamte Raum völlig unstrukturiert zur Exploration präsentiert. Vielmehr kann er seinen Zielen entsprechend wesentlich systematischer und verständlicher die gestellte Aufgabe erledigen.

Abschließend sei an dieser Stelle noch auf interessante Ansätze zur Merkmalsauswahl in [50] hingewiesen. Devaney und Ram beschreiben in ihrem Artikel einen Ansatz zur Auswahl von Merkmalen beim unüberwachten Lernen, genauer beim Konzeptuellen Clustern. Sie diskutieren eine sequentielle Merkmalsauswahlstrategie basierend auf dem bekannten Clusterverfahren COBWEB. In ihrer Evaluierung zeigen sie die signifikante Verbesserung der Ergebnisse von COBWEB. Der Nachteil von Devaney und Ram's Methode ist, dass COBWEB ein inkrementelles Verfahren ist. Damit sind die Ergebnisse, im Gegensatz zu KMeans, abhängig von der Reihenfolge der Objekte. Auch müssten alle numerischen Werte für die Verarbeitung mit COBWEB diskretisiert werden.

Das nächste Kapitel wird die Auswirkungen der Übersetzung von Worten und Termen auf Konzepte empirisch untersuchen. Am Beispiel des Clusterns von Textdokumenten werden wir auf der Basis des Kosinus-Ähnlichkeitsmaßes die Parameter für die erfolgreiche Integration von Hintergrundwissen vorstellen und evaluieren.

8 Textclustern mit Hintergrundwissen

Clustern mit Hintergrundwissen beschreibt einen weiteren zentralen Ansatz dieser Arbeit, bei dem wir Hintergrundwissen in den Clusterprozess zur Steigerung der Clustergüte integrieren. Wir beginnen das Kapitel mit einer Diskussion über unterschiedliche Möglichkeiten, Hintergrundwissen in den Clusterprozess einzubringen und grenzen diese Ansätze vom überwachten Lernen bzw. Klassifizieren ab. In Abschnitt 8.2 wird beschrieben, wie Hintergrundwissen in Form einer Ontologie in den Clusterprozess integriert werden kann. Der Ansatz wird anhand von Textdokumenten vergleichend zu bekannten Ansätzen des Textclusterns evaluiert. Die Ansätze und Ergebnisse aus Abschnitt 8.2 folgen den Arbeiten [118, 117]. Abschnitt 8.3 analysiert den veränderten Merkmalsvektor und liefert eine Erklärung für die beobachteten Clusterergebnisse. Ein alternativer Ansatz zur Berechnung neuer Merkmale zum Clustern von Textdokumenten besteht in der Anwendung von LSI (siehe Abschnitt 4.4). Abschnitt 8.4 diskutiert Ergebnisse für ausgewählte Datensätze auf der Basis von LSI-Merkmalen und einer Kombination von LSI und konzeptbasierten Merkmalen. Abschnitt 8.5 verwendet für das Clustern der Textdokumente die Formale Begriffsanalyse und diskutiert in diesem Zusammenhang die Nutzung von Konzepten und KMeans-Clustern als Merkmale für das konzeptuelle Clustern.

8.1 Klassifizieren und Clustern mit Hintergrundwissen

Beim Klassifizieren oder überwachten Lernen besteht die Aufgabe im Ableiten eines möglichst allgemeingültigen Modelles auf der Basis einer gegebenen Menge von kategorisierten Objekten. Das heißt, dass die Objekte in Gruppen, Klassen oder Kategorien eingeteilt sind und entsprechende Klassenbezeichner bekannt sind. Durch die beschränkte Anzahl an kategorisierten Objekten ist es nicht möglich, ein für alle Objekte korrektes, d.h. ein allgemeingültiges-Modell zu berechnen. Je weniger Beispiele zur Verfügung stehen, desto schwieriger wird die Aufgabe das Modell zu schätzen. Da das Erstellen von Beispielklassifikationen meist manuell geschieht und deshalb sehr aufwendig ist, versucht man die Anzahl der Beispiele möglichst klein zu halten. Dies steht im Widerspruch zu der Forderung nach einer möglichst großen Anzahl an kategorisierten Objekten zum Ableiten eines guten Modelles.

Stehen nur sehr wenige Objekte mit Klassenbezeichnern zur Verfügung sinkt die Güte der Modelle. Zur Steigerung der Güte greift man nun auch auf Objekte ohne Klassenbezeichner zurück und kombiniert alle Informationsquellen. Aus Sicht der Klassifikationsaufgabe handelt es sich hierbei um die Nutzung von Hintergrundwissen. Im Bereich des TextMining findet man in der Literatur erfolgreiche Ansätze, die nichtkategorisierte Objekte zur Steigerung der Modellgüte beim Klassifizieren einsetzen. Zum Beispiel nutzen Zelikovitz u.a. [233] Texte oder Webdokumente ohne Klassenbezeichner zur Steigerung der Güte und Nigam u.a. [178] zeigen anhand von realen Datensätzen eine Steigerung von bis zu 30 % bei der Nutzung von nichtkategorisierten Objekten.

Reduziert man die Informationen über die Objekte weiter, so stehen in der nächsten Stufe nicht mehr die Klassenbezeichner zur Verfügung, sondern nur noch Informationen über die paarweise Beziehung zweier Objekte. Denkbar sind hier zum Beispiel die Beziehung “must-link” (steht in enger Beziehung) oder “cannot-link” (steht in keiner Beziehung) (vgl. [225] und [136]). Diese Be-

ziehungsinformationen sind schwächer als die Kategorien eines Objektes. Man kann die paarweisen Beziehungen aus den Kategorien der Objekte ableiten. Stehen nur Informationen über die paarweisen Beziehungen der Objekte zur Verfügung, so wechselt man das Paradigma und man spricht vom Clustern mit Hintergrundwissen und nicht mehr vom Klassifizieren. Das Hintergrundwissen wird hier in Form der Objektbeziehungen bereitgestellt und soll dem Clusterprozess beim Entdecken der Cluster helfen.

Die Informationen über die Zugehörigkeit eines Objektes zu einer bestimmten Kategorie oder die paarweise Beziehung von Objekten untereinander stellen aus Sicht des Clusterns Hintergrundwissen dar. Dabei handelt es sich um *Informationen über die Objekte*, d.h. man besitzt a priori Wissen über die Objekte und stellt es in Form der Beziehungen bereit. Als Alternative können *Informationen über die Merkmale* zur Verfügung gestellt werden, d.h. man weiß a priori etwas über die objektbeschreibenden Merkmale und deren Zusammenhänge untereinander. Beide Alternativen — Informationen über die Objekte oder Merkmale — stellen Ansatzpunkte zur Nutzung von Hintergrundwissen dar. In dieser Arbeit konzentrieren wir uns auf den zweiten Ansatz und stellen in Form einer Ontologie zusätzlich Informationen über die Beziehungen der Merkmale bereit. Ansätze dieser Art findet man z.B. in [89] und [90].

Im folgenden Abschnitt wird die Integration von Hintergrundwissen in Form einer Ontologie in den Clusterprozess vorgestellt. Die Ontologie stellt Informationen über die Merkmale bereit.

8.2 Clustern von Textdokumenten

In diesem Abschnitt werden wir eine neu entwickelte Variante der Integration von Ontologien in die Repräsentation von Objekten, in diesem Fall von Textdokumenten, vorstellen. Abschnitt 8.2.1 wird einen bekannten Ansatz aus der Literatur für das Textclustern als Ausgangs- und Vergleichspunkt einführen. Abschnitt 8.2.2 liefert die Vergleichsergebnisse von geratenen Clusterungen für die PRC-Datensätze (Reuters-Datensätze siehe Abschnitt 2.1). Wir werden für den Ausgangspunkt beim Textclustern eine umfangreiche Evaluierung durchführen und Ergebnisse für verschiedene Parameter präsentieren, die dann in Abschnitt 8.2.5 zum Vergleich und zur Bewertung der neu entwickelten Methoden dienen. Abschnitt 8.2.3 führt die Methoden zur Integration des Hintergrundwissens ein, Abschnitt 8.2.4 beschreibt den Aufbau der Experimente und deren Ergebnisse, die dann in Abschnitt 8.2.5 präsentiert werden. Alle Experimente wurden für den Reuters-Datensatz durchgeführt. Wir beziehen uns in diesem Abschnitt nur auf diesen bzw. auf die aus diesem Datensatz abgeleiteten Teildatensätze PRC, PRC-min15, PRC-min15-max20, PRC-min15-max100, PRC-max20 und PRC-max100, die zum Teil spezielle Eigenschaften aufweisen (siehe Abschnitt 2.1). Ergebnisse für weitere Datensätze, die auf den gleichen Ansatz zur Integration des Hintergrundwissens zurückgreifen, findet man im Anwendungsteil III dieser Arbeit.

8.2.1 Clustern von Textdokumenten ohne Hintergrundwissen

Der Ansatz zur Integration von Hintergrundwissen setzt an der Repräsentation der Dokumente an. Während der Vorverarbeitung wird auf der Basis des Hintergrundwissens die Repräsentation verändert. Typische Vorverarbeitungsschritte für Dokumente sind in Abschnitt 4.2 beschrieben. Als Ergebnis einer solchen Vorverarbeitung liegen die Dokumente als Termvektoren $\vec{t_d}$ vor. Die Terme des Termvektors bestehen im Folgenden immer aus Wortstämmen (siehe Abschnitt 4.2.2), wobei alle Stoppworte entfernt sind. Dazu wird die in Abschnitt 4.2.3 angesprochene Stoppwortliste[1] mit

[1] http://www.aifb.uni-karlsruhe.de/WBS/aho/clustering

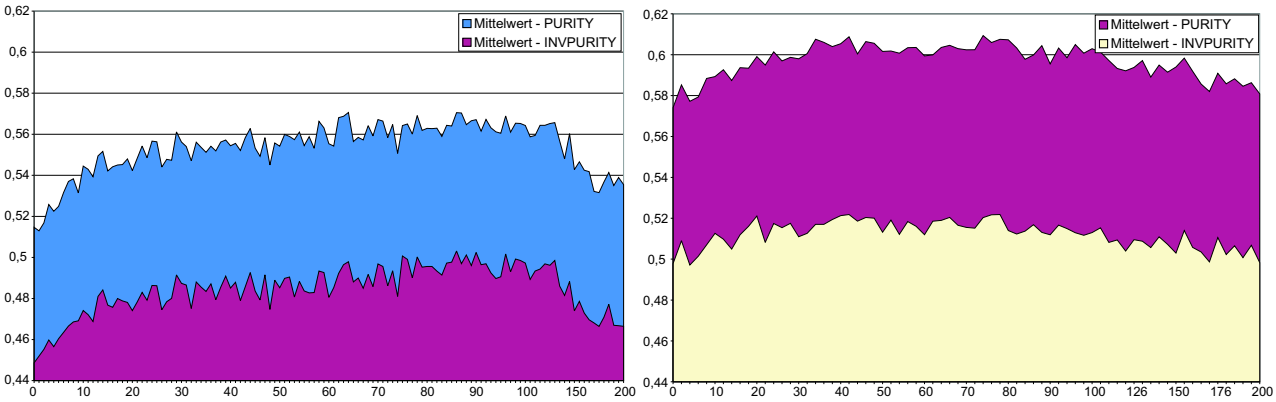

Abbildung 8.1: Analyse des Einflusses von Term-Pruning für Prunethreshold $0 < \delta < 200$ auf Purity/InversePurity beim Clustern von PRC-min15-max100 mit 60 Cluster links ohne Hintergrundwissen und rechts mit Hintergrundwissen (mit tfidf, Stemming, Normalisierung, kein Dokument-Pruning)

571 Stoppworten des SMART-Systems verwendet. Geclustert wird immer mit dem Bi-Sec-KMeans Verfahren.

Verschiedene Einflussgrößen sind aus der Literatur bekannt. In einem ersten Schritt wollen wir systematisch Parameter variieren und so deren Einfluss auf die Clustergüte überprüfen, bevor wir dann die Experimente auf der neuen Repräsentation wiederholen. Folgende Parameter werden in den Experimenten variiert:

tfidf-Gewichtung Während der Vorverarbeitungsphase werden die Termhäufigkeiten $\mathrm{tf}(d,t)$ des Termvektors $\vec{t_d}$ durch die gewichteten Termhäufigkeiten $tfidf(d,t)$ im Vektor ersetzt (siehe Kapitel 4.2.5.1). Alle Versuche sind für beide Vektoren durchgeführt wurden (Auch der Konzeptvektor kann gewichtet oder ungewichtet in den Clusterprozess integriert werden. Wir gehen darauf in Abschnitt 8.2.3 genauer ein.).

Löschen seltener Terme (Prunethreshold δ) Der Prunethreshold δ wird zum Löschen seltener Worte verwendet und stellt eine absolute Schranke für die Häufigkeit von Termen dar (siehe Kapitel 4.2.4). Wir unterscheiden zwei Varianten, das Dokument-Pruning und das Term-Pruning. Um vorab zu bestimmen, ob das Löschen seltener Terme notwendig ist, haben wir für den PRC-min15-max100 Datensatz eine umfangreichere Evaluierung des Term-Prunings durchgeführt. Wir variieren den Prunethreshold δ zwischen 0 und 200.

Das Ergebnis gibt der linke Teil von Abbildung 8.1 wieder. Die Werte der obere Kurve entsprechen den Purity-Werten und die der unteren den InversePurity-Werten. Beide Kurven zeigen einen ähnlichen Verlauf, so dass keine Kannibalisierungseffekte des einen Maßes zu Gunsten des anderen zu beobachten sind. Bei der Analyse des Verlaufs der Purity-Kurve findet man einen steilen Anstieg der Purity am Anfang und einen Abfall am Ende sowie einen ausgedehnten Bereich in der Mitte mit relativ stabilen Ergebnissen. Wir fixierten für die folgenden Experimente den Prunethreshold bei 0, 5 und 30. Den Ergebnissen entnehmen wir, dass man die besten Resultate mit einem Prunethreshold von 30 erhalten wird.

Die Vorverarbeitung der Dokumente erfolgt immer in der Reihenfolge: Stoppworte löschen, Wortstämme berechnen, Löschen der seltenen Worte und Gewichten mit tfidf, falls der Schritt in der Vorverarbeitung enthalten ist.

Tabelle 8.1: Anzahl der Dokumente, Klassen, Wortstämme, Terme der PRC-Datensätze bei unterschiedlichem Prunethreshold

Datensatz	#Dokument	#Klassen	#Wortstämme	$\|T\|$	Prunethreshold
PRC-max20	1035	82	6494	91749	0
			2310	84263	5
			594	64455	30
PRC-min15-max20	899	46	6073	79758	0
			2129	72721	5
			544	54763	30
PRC-max100	2755	82	10177	241005	0
			3847	229733	5
			1239	199606	30
PRC-min15-max100	2619	46	9924	229014	0
			3745	218009	5
			1205	188868	30
PRC	12344	82	20574	863167	0
			7591	840422	5
			2657	784434	30
PRC-min15	12208	46	20432	851176	0
			7536	828574	5
			2629	772865	30

Einige Parameter konnten leider nicht variiert werden, da das Testsetting dann zu groß geworden wäre. So wurde darauf verzichtet, ohne Wortstammbildung zu clustern. Alle Vektoren wurden immer auf die Länge 1 normiert und alle Buchstaben sind immer in Kleinschreibung. Auch das so genannte Dokumentpruning, also das Löschen von Termen, die in weniger als δ Dokumenten vorkommen, zeigte in Vorabstudien wenig Einfluss und wurde daher nicht als Parameter in die Evaluierung einbezogen.

Tabelle 8.1 fast die Eigenschaften der verwendeten Reuters-Datensätze nach der Vorverarbeitung zusammen. Die Datensätze sind in Abschnitt 2.1 im Detail beschrieben. Die Anzahl der unterschiedlichen Wortstämme schwankt zwischen minimal 544 (bei PRC-min15-max20) und maximal 20432 (bei PRC). Durch den unterschiedlichen Prunethreshold schwankt auch die Anzahl der Terme in jedem Datensatz stark. Anzumerken sei an dieser Stelle noch einmal, dass die zwei unterschiedlichen Klassenanzahlen 46 und 82 als Basis der jeweiligen Evaluierung zur Verfügung stehen. Dies ergibt sich bei den Datensätzen mit der Minimalrestriktion von mindestens 15 Dokumenten pro Klasse.

Clusterergebnisse auf Termvektoren Tabelle 8.2 fast die Ergebnisse für die PRC-Datensätze ohne Hintergrundwissen zusammen.[2] Im Folgenden diskutieren wir ein paar Details:

– Der PRC-min15 Datensatz erzielt über alle Messläufe die besten Ergebnisse. Dies gilt für alle Clusterzahlen. Bei 5 Clustern erreicht die Purity schon einen Wert von 54.90 % und bei 100 Clustern steigt sie auf 77.70 %. Keinen großen Unterschied in der Performance erzielt der PRC Datensatz. Die 136 Dokumente verteilt auf 36 Klassen fallen beim Purity-Maß kaum ins Gewicht. Drastischer fällt der Unterschied zum nächstbesten Datensatz PRC-min15-max100 aus, der bei 5 Clustern “nur” einen Purity-Wert von 17.10 % im besten Fall erzielt. Zu beachten ist aber, dass die untere Schranke durch zufälliges Ziehen der Clusterlösung bei PRC-min15-max100 für 5 Cluster mit 5.04 % deutlich niedriger liegt als bei PRC mit 30.46 % für 5 Cluster (siehe Abschnitt 8.2.2).

[2] Tabelle F.1 im Anhang dieser Arbeit bietet alle Werte noch einmal im Überblick sowie zusätzlich die Standardabweichung.

Tabelle 8.2: Purity für Clustering ($k = 5, 10, 20, 30, 50, 60, 70, 100$) ohne Hintergrundwissen, für PRC-Datensätze, Prunethresholds 0, 5, 30, mit und ohne tfidf Gewichtung, Mittelwert über 20 Wiederholungen

PRC	gew.	pr.	5	10	20	30	50	60	70	100
max20	tfidf	0	0.091	0.159	0.249	0.306	0.36	0.385	0.404	0.452
		5	0.092	0.162	0.261	0.325	0.399	0.424	0.446	0.48
		30	0.092	0.169	0.282	0.349	0.447	0.47	0.489	0.531
	ohne	0	0.088	0.148	0.221	0.273	0.341	0.363	0.386	0.435
		5	0.088	0.149	0.229	0.277	0.341	0.367	0.39	0.436
		30	0.088	0.149	0.231	0.281	0.344	0.367	0.393	0.437
min15-max20	tfidf	0	0.104	0.181	0.283	0.343	0.425	0.446	0.448	0.498
		5	0.105	0.188	0.304	0.372	0.464	0.479	0.5	0.54
		30	0.106	0.198	0.335	0.426	0.521	0.543	0.562	0.6
	ohne	0	0.101	0.171	0.26	0.326	0.396	0.419	0.439	0.49
		5	0.1	0.172	0.267	0.321	0.398	0.421	0.45	0.496
		30	0.099	0.173	0.272	0.322	0.401	0.436	0.452	0.503
max100	tfidf	0	0.16	0.265	0.372	0.422	0.482	0.502	0.511	0.536
		5	0.159	0.264	0.375	0.444	0.506	0.515	0.53	0.557
		30	0.162	0.263	0.39	0.452	0.51	0.535	0.548	0.579
	ohne	0	0.147	0.232	0.321	0.363	0.418	0.436	0.45	0.488
		5	0.143	0.224	0.319	0.365	0.428	0.437	0.456	0.49
		30	0.146	0.23	0.316	0.362	0.426	0.447	0.453	0.497
min15-max100	tfidf	0	0.171	0.273	0.401	0.452	0.514	0.526	0.545	0.561
		5	0.173	0.284	0.399	0.463	0.534	0.547	0.556	0.583
		30	0.171	0.287	0.415	0.486	0.54	0.57	0.584	0.608
	ohne	0	0.153	0.245	0.343	0.385	0.446	0.462	0.481	0.515
		5	0.154	0.243	0.34	0.387	0.444	0.461	0.478	0.516
		30	0.154	0.246	0.343	0.395	0.448	0.47	0.482	0.523
PRC	tfidf	0	0.542	0.604	0.696	0.719	0.74	0.748	0.751	0.765
		5	0.539	0.609	0.69	0.721	0.74	0.747	0.753	0.767
		30	0.545	0.604	0.698	0.722	0.743	0.751	0.754	0.77
	ohne	0	0.493	0.555	0.616	0.646	0.677	0.688	0.695	0.71
		5	0.489	0.558	0.616	0.648	0.677	0.685	0.696	0.711
		30	0.491	0.553	0.621	0.651	0.68	0.688	0.696	0.712
min15	tfidf	0	0.544	0.605	0.695	0.722	0.748	0.758	0.761	0.771
		5	0.551	0.613	0.702	0.725	0.752	0.758	0.764	0.772
		30	0.549	0.608	0.701	0.731	0.753	0.76	0.764	0.777
	ohne	0	0.493	0.563	0.621	0.652	0.686	0.695	0.705	0.721
		5	0.494	0.561	0.623	0.655	0.685	0.695	0.705	0.717
		30	0.488	0.562	0.629	0.653	0.687	0.697	0.706	0.721

- Wie zu erwarten war, steigt die Purity mit einer höheren Anzahl an Clustern (siehe Abschnitt 8.2.2). Dies ist bei gleichen Vorverarbeitungsschritten für alle Datensätze und Parameterkombinationen zu beobachten. Zum Beispiel steigt für den PRC-min15-max100 bei tfidf Gewichtung und Prunethreshold 30 der Wert von 17.10 % bei 5 Cluster auf 60.80 % bei 100 Clustern.

- Die Anwendung der tfidf-Gewichtung führt verglichen mit der einfachen Termrepräsentation immer zu besseren Ergebnissen. Dabei liegt die Steigerung im Schnitt um 5 % und führt im besten Fall bei PRC-min15-max100 zu einer 10 %igen Verbesserung.

- Abschließend analysieren wir den Einfluss des Prunethreshold. Er liefert die interessantesten Ergebnisse. Das Löschen von seltenen Worten mit den von uns gewählten Parametern führt immer zu einer Verbesserung oder zu gleich guten Ergebnissen. Der Effekt ist nur zu beobachten, wenn der Termvektor mit tfidf gewichtet wird. Beim Clustern auf der Basis der Termvektoren sind bei gegebener Clusteranzahl die Puritywerte nahezu konstant – der Einfluss des Prunethreshold ist also zu vernachlässigen. Ein gänzlich anderes Verhalten ist bei den gewichteten Termvektoren zu beobachten. Die größte Differenz bei 100 Clustern (alternativ 50) mit ca. 10 % von 49.8 % (42.5 %) auf 60 % (52.1 %) liegt bei PRC-min15-max20 vor. Auch der PRC-max20 Datenatz erfährt durch das Pruning noch einen beachtliche Steigerung in der Qualität der Clusterung.

  Die Steigerungsraten bei den PRC-Datensätzen mit max. 100 Dokumenten fallen schon deutlich geringer aus und bei den PRC-Datensätzen ohne Beschränkung der maximalen Anzahl an Dokumenten ist nur ein sehr geringer Einfluss des Prunethresholds von 0.6 % bei 100 Clustern für PRC-min15 mit Prunethreshold 30 zu beobachten. Experimente mit größeren Prunethreshold-Werten zeigten keine weitere Verbesserung der Clustergüte.

Zusammenfassend lässt sich feststellen, dass die tfidf-Gewichtung immer und Prunethresholds nur bei wenigen Dokumenten zu besseren Ergebnissen führen. Nehmen wir nun die Ergebnisse ohne Hintergrundwissen als Ausgangspunkt für einen Vergleich mit der im Folgenden eingeführten erweiterten Repräsentation mit Hintergrundwissen.

8.2.2 Untergrenzen der Clustergüte für PRC-Datensätze

Bevor die Clusterergebnisse nach der Integration des Hintergrundwissens vorgestellt werden, wollen wir mit Hilfe eines kleinen Experimentes abschätzen, wie gut die Clusterungen ohne Hintergrundwissen gegenüber dem zufälligen Raten sind. Dies wird auch zu einem besseren Verständnis der eingesetzten Maße führen. Die Ergebnisse für Purity und InversePurity (siehe Abschnitt 5.3.3.2) aus Abbildung 8.2 erhält man durch zufälliges Ziehen der Zuordnung von Dokumenten zu Clustern, wobei jedes Dokument genau einem Cluster zugeordnet wird. Die resultierenden Clusterungen haben ungefähr gleich große Cluster, wie das auch beim Bi-Sec-KMeans der Fall ist.

Die Purity des PRC-min15-max100-Datensatzes für einen Cluster liegt bei 3.8 % und entspricht somit dem erwarteten Ausgangspunkt (vgl. Abbildung 8.2 linker Teil). Der Ausgangspunkt berechnet sich als Quotient aus der Anzahl der Objekte im größten Cluster durch die Anzahl aller Objekte des Datensatzes. Bei 60 Clustern liegt der Wert bei ungefähr 9.8 %. Mit 46.1 % (vgl. Tabelle 8.2) ist selbst der Wert für die schlechteste Strategie noch deutlich besser als das Raten der Zuordnung. Die Purity hat wie erwartet ihr Maximum bei eins, was bei einer Clusteranzahl gleich der Objektanzahl der Fall ist.

Die Ergebnisse für die InversePurity sind in der rechten Hälfte der Abbildung 8.2 wiedergegeben. Die InversePurity hat ihr Maximum bei eins und erreicht dies bei einer Clusteranzahl von eins. Bei

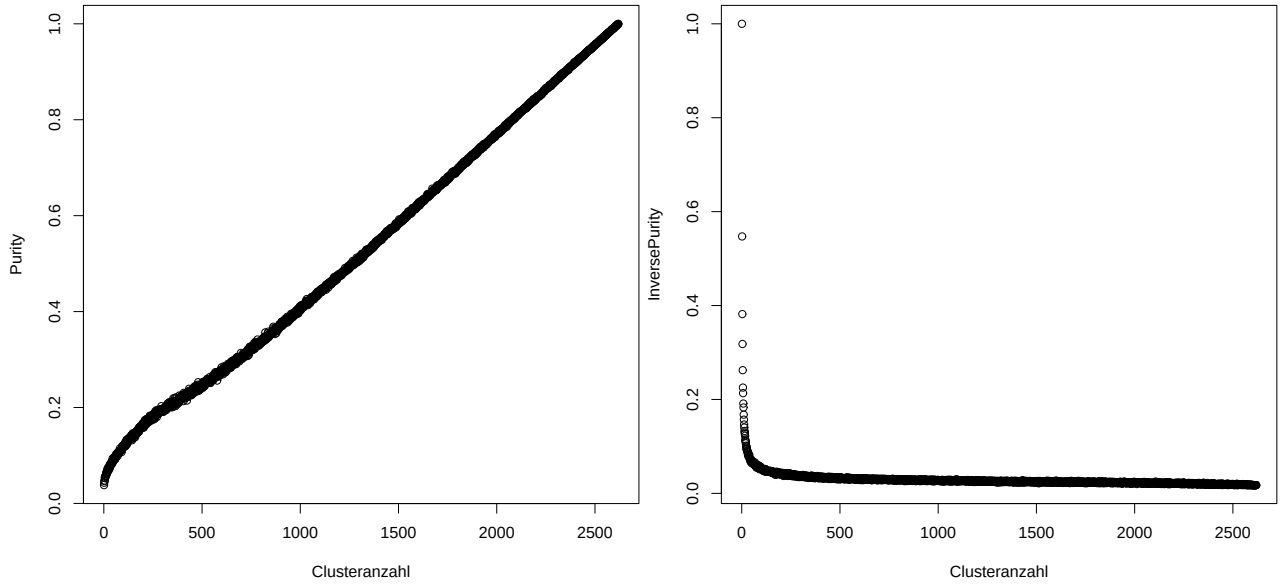

Abbildung 8.2: Purity (links) und InversePurity (rechts) für zufällig gezogene Clusterungen des PRC-min15-max100 Datensatzes mit einer Clusteranzahl von 1 bis $|D| = 2619$

zehn Clustern ist der Wert auf ungefähr 15 % und bei 20 Clustern auf ca. 10 % gesunken. Vergleicht man die InversePurity für 60 Cluster mit den Ergebnissen auf Tabelle 8.5, so stellt man auch hier einen deutlichen Unterschied zwischen geratenen 6 % und den 48 % aus dem Clusterlauf fest.

Die Ergebnisse für den PRC-Datensatz sind den Ergebnissen des PRC-min15-max100-Datensatz ähnlich. Abbildung 8.3 fasst sie für Purity und InversePurity für 1 bis 2000 Cluster[3] zusammen. Der Ausgangspunkt für die Purity liegt hier bei ca. 30.5 %. Dieser hohe Wert lässt sich leicht durch die Klasse “earn” mit mehr als 3000 Dokumenten erklären. Sie dominiert alle anderen Klassen. Bei 60 Cluster erhält man noch immer den gleichen Wert für eine geratene Clusterung. Im Vergleich zu einer Bi-Sec-KMeans-Clusterung mit im schlechtesten Fall 68.5% ist die geratene Clusterung deutlich schlechter und man beobachtet die gleichen Unterschiede wie beim PRC-min15-max100 Datensatz.

Die Ergebnisse für die InversePurity sind analog und man kann mit 26 % für eine Bi-Sec-KMeans-Clusterung (60 Cluster) und 3 % für eine geratene Clusterung auch hier einen klaren Unterschied erkennen.

Die Ergebnisse dieser Versuche zeigen, wie deutlich Bi-Sec-KMeans auf der Basis von Termen Zusammenhänge zwischen den Dokumenten finden kann. Unser Ausgangspunkt aus Abschnitt 8.2 scheint daher eine gute Basis für eine vergleichende Evaluierung. Im nächsten Abschnitt werden wir die verschiedenen Strategien zur Integration des Hintergrundwissens diskutieren, bevor wir sie in Abschnitt 8.2.4 empirisch evaluieren.

8.2.3 Integration von Hintergrundwissen in die Textrepräsentation

8.2.3.1 Einleitung

Dieser Abschnitt wird die Integration des Hintergrundwissens in das “Bag of Words” Modell vorstellen. Ausgangspunkt sind die in Kapitel 6.2 eingeführten Ontologien sowie die extrahierten

[3] Die Berechnung aller 12344 verschiedenen Clusterungen war zu aufwendig. Daher wurden nur die ersten 2000 Clusterungen berechnet. Aus den durchgeführten Berechnungen können alle wesentlichen Aussagen abgeleitet werden.

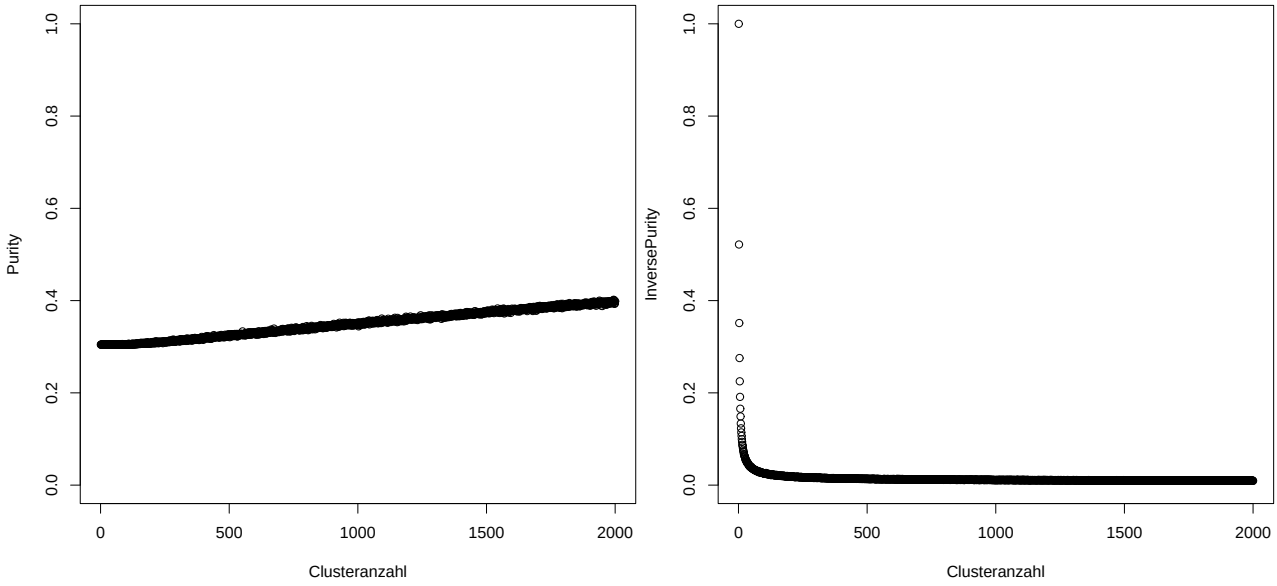

Abbildung 8.3: Purity (links) und InversePurity (rechts) für zufällig gezogene Clusterungen des PRC Datensatzes mit einer Clusteranzahl von 1 bis 2000

Termvektoren $\vec{t_d}$ eines jeden Dokumentes d. Wir folgen mit unseren Ausführungen den Arbeiten [118, 117]. Unser prinzipieller Ansatz für die Integration des Hintergrundwissens basiert auf der Übersetzung der Terme in Konzepte einer Ontologie. Dieser Übersetzungsvorgang ist von zentraler Bedeutung für den Ansatz, da nur bei korrekter Übersetzung des Terms (und hier meinen wir das Problem der Wortsinnerkennung) das Hintergrundwissen in den Clusterprozess eingebracht werden kann. Idee der Übersetzung in Konzepte und der damit verbundenen Änderung der Repräsentation ist es, Synonyme aufzulösen und allgemeine Zusammenhänge der Repräsentation hinzuzufügen. So könnte ein Dokument über Rindfleisch und eines von Schweinefleisch durch den Clusteralgorithmus nicht miteinander in Beziehung gebracht werden, wenn der Termvektor nur die Worte Rind- und Schweinefleisch enthält. Fügen wir aber das generellere Wort “Fleisch” als Term hinzu wird die semantische Beziehung der beiden Worte aufgedeckt und entsprechend repräsentiert.

Die Ontologie stellt in Definition 11 eine Funktion Ref_C bereit. Wir bedienen uns dieser Funktion um für einen Term eine Menge von Konzepten zu erhalten. Wir haben dann verschiedene Optionen zur Verfügung, z.B. können wir alle Konzepte nutzen oder versuchen den richtigen Sinn des Termes, sprich das “am besten passende” Konzept aus der Menge herauszufinden. Wichtig ist an dieser Stelle, dass wir auf die Funktion Ref_C angewiesen sind, die uns zumindest eine Menge von Konzepten passend zu einem Term liefert. Eine Ontologie, wie z.B. WordNet, muss diese Funktion spezifizieren.

Wir zerlegen im Folgenden den Übersetzungsprozess und untersuchen drei zentrale Fragen:

- Wie koexistieren Konzepte und Terme am besten (Abschnitt 8.2.3.2)?
- Ist Wortsinnerkennung für die Integration von Konzepten in den Termvektor wichtig oder nicht (Abschnitt 8.2.3.3)?
- Wie können wir die Generalisierungshierarchie (Abschnitt 8.2.3.4) zur Steigerung der Clustergüte nutzen?

Als Ergebnis des Übersetzungsprozesses werden unsere Dokumente nicht mehr durch einen einfachen Termvektor $\vec{t_d}$, wie wir ihn aus dem letzten Abschnitt 8.2.1 kennen, repräsentiert, sondern durch einen angereicherten Termvektor $\vec{\tau_d}$.

Der Begriff Termvektor ist so allgemein gehalten, dass man sowohl ein Wort als auch ein Konzept darunter subsummieren kann. Wir werden den Übersetzungs- bzw. Anreicherungsprozess im Folgenden entlang des einfachen und angereicherten Termvektors einführen. Im weiteren Verlauf der Arbeit verwenden wir aber nur das Symbol des einfachen Termvektors $\vec{t_d}$, da alle Algorithmen und Operationen für beide Vektoren identisch sind.

8.2.3.2 Strategien: Hinzufügen von Konzepten, Ersetzen von Konzepten durch Terme oder nur Konzeptvektoren

Erinnern wir uns noch einmal an das Beispiel aus der Einleitung dieses Abschnittes. Wir hatten verschiedene Sorten von Fleisch, die wir durch das Konzept FLEISCH in Beziehung zueinander gesetzt haben. Es stellt sich nun die Frage, welche Information nutzt man zum Clustern der Dokumente? Dazu haben wir drei Strategien untersucht:

Hinzufügen von Konzepten (Add Concepts, "add") Wie man dem Namen der Strategie (add[4]) schon entnehmen kann, fügen wir in diesem Fall die Konzepte der Ontologie den Termen hinzu. Wir erweitern den Vektor $\vec{t_d}$ um neue Konzepte c der Dokumentmenge. Der angereicherte Termvektor $\vec{\tau_d}$ ergibt sich durch die Verkettung von einfachem Termvektor $\vec{t_d}$ und Konzeptvektor $\vec{c_d}$:

$$\vec{\tau_d} := (\text{tf}(d, t_1), \ldots, \text{tf}(d, t_z), \text{cf}(d, c_1), \ldots, \text{cf}(d, c_l)) \tag{8.1}$$

Den Konzeptvektor $\vec{c_d} := (\text{cf}(d, c_1), \ldots, \text{cf}(d, c_l))$ erhalten wir unter Anwendung der Referenzfunktion Ref_C auf alle Terme eines Dokumentes d, wobei $z = |T|$ und $l = |C|$ gilt und $\text{cf}(d, c)$ die Häufigkeit des Konzeptes $c \in C$ im Dokument d angibt. Wie der Übersetzungsschritt mit Wortsinnerkennung im Detail definiert ist, findet man in nächsten Abschnitt.

Durch das beschriebene Vorgehen werden Terme, die eine Entsprechung in der Ontologie finden, mindestens zweimal gezählt, einmal als Teil des Termvektors $\vec{t_d}$ und einmal als Teil des Konzeptvektors $\vec{c_d}$. Abhängig von der Wortsinnerkennungsstrategie können Terme wie z.B. "Bank", die mehr als eine Bedeutung in der Ontologie haben, noch häufiger im Termvektor vorkommen.

Ersetzen von Termen durch Konzepte (Replace Terms by Concepts, "repl") Diese Strategie funktioniert wie die Strategie "Hinzufügen von Konzepten", aber sie entfernt anschließend alle Terme aus der Vektorrepräsentation $\vec{t_d}$, für die ein entsprechendes Konzept gefunden wurde. Wir zählen Terme mit einer Entsprechung in der Ontologie nur noch auf der Basis der Konzepte. Terme, die nicht in der Ontologie vorkommen, werden aber nicht gelöscht. Wir reduzieren die Menge der Terme T wie folgt:

$$T_{new} := \{t \in T | Ref_C(t) = \emptyset\} \tag{8.2}$$

und erhalten dadurch den reduzierten Vektor T_{new}. Wir ersetzen $T = T_{new}$ und erhalten so analog zu Gleichung 8.1 den angereicherten Termvektor $\vec{\tau_d}$.

nur Konzeptvektoren (Concept Vector only, "only") Diese Strategie arbeitet wie die Ersetzungsstrategie, mit dem Unterschied, dass wir keinen Term in der Vektorrepräsentation berücksichtigen. Ein Term, der nicht in ein Konzept der Ontologie übersetzt werden kann, wird

[4] Auf die Abkürzungen wird in den Abschnitten 8.2.4 und 8.2.5 zurückgegriffen.

demzufolge im weiteren Verlauf des Clusterprozesses ignoriert. Dazu setzen wir $T := \emptyset$ und nutzen als Vektorrepräsentation den Konzeptvektor $\vec{\tau_d} := \vec{c_d}$.

Wir haben nun verschiedene Varianten zur Verfügung, einfache Terme und Konzepte zu kombinieren. Im nächsten Schritt müssen wir die bei der Übersetzung notwendige Wortsinnerkennung von Termen in Konzepte klären.

8.2.3.3 Strategien zur Wortsinnerkennung

Zentrales Problem bei der Übersetzung von Termen in Konzepte ist die Mehrdeutigkeit der Terme. Fügt man Konzepte zu Termen hinzu oder ersetzt diese, kann so in die Repräsentation Rauschen eingebettet werden oder man verliert Informationen. Es stellt sich daher die Frage, wie man das "am besten passende" Konzept aus einer Menge alternativer Konzepte mit zum Teil sehr unterschiedlichen Bedeutungen für einen Term auswählt bzw. wie diese Auswahl die Clusterergebnisse beeinflusst.

Die Wortsinnerkennung ist ein eigenes großes Forschungsfeld, vgl. [121]. Unsere Intension bei der Integration einer einfachen Wortsinnerkennung in den Prozess liegt schlicht in der Bestimmung, wie viel Wortsinnerkennung wir benötigen. Daher haben wir neben einer Referenzstrategie nur den Einfluss zweier einfacher Wortsinnerkennungsstrategien auf den Clusterprozess untersucht:

Alle Konzepte (All Concepts, "all") Die Referenzstrategie führt keine Wortsinnerkennung durch und nutzt alle Konzepte, um die Termrepräsentation anzureichern. Damit berechnet sich die Konzepthäufigkeit nach der folgenden Formel:

$$\text{cf}(d, c) := \text{tf}(d, \{t \in T \mid c \in Ref_C(t)\}) \tag{8.3}$$

wobei man die Berechnung der Termhäufigkeit $\text{tf}(d, T')$ für Mengen von Termen T' Abschnitt 4.1 entnimmt.

Erstes Konzept (First Concept, "first") Wie in Abschnitt 6.3.3.1 erwähnt, liefert die Ref_C von WordNet eine geordnete Liste der Konzepte. Für Ontologien mit einer solchen Funktion interessiert uns, wie sich diese Ordnung auf den Clusterprozess auswirkt. Unsere Strategie zur Erkennung von Mehrdeutigkeiten stützt sich demzufolge auf die Idee, dass der Schreiber des Textes immer die wahrscheinlichste Bedeutung des Wortes mit der Nutzung des Termes im Text verbunden hat. Dabei ignorieren wir den Kontext des Termes vollkommen.

Für einen Term t, der im Lexikon S_C der Ontologie $\mathcal{O}$ vorkommt, beachtet diese Strategie nur die Konzepthäufigkeit cf für das wichtigste (erste) Konzept der geordneten Menge $Ref_C(t)$. Die Häufigkeiten aller weiteren Konzepte aus $Ref_C(t)$ werden nicht erhöht. Die Konzepthäufigkeit berechnet sich wie folgt:

$$\text{cf}(d, c) := \text{tf}(d, \{t \in T \mid \text{first}(Ref_C(t)) = c\}) \tag{8.4}$$

wobei $\text{first}(Ref_C(t))$ das erste Konzept $c \in Ref_C(t)$ der geordneten Menge liefert. Die Ordnung muss in der Ontologie spezifiziert werden.

Wortsinnerkennung mittels Kontext (Disambiguation by Context, "context") Die letzte und sehr einfache Strategie erkennt den Sinn des Terms t, also die entsprechenden

Konzepte $Ref_C(t) := b, c, \ldots$ mit denen der Term in Beziehung steht, mittels der folgenden einfachen Methode[5]:

1. Wir definieren eine semantische Umgebung eines Konzeptes c als die Menge aller direkten Ober- und Unterkonzepte

$$V(c) := \{b \in C | c \prec b \text{ or } b \prec c\}. \tag{8.5}$$

2. Wir sammeln alle Terme der konzeptuellen Umgebung, die in Beziehung zum Konzept c stehen durch:

$$U(c) := \bigcup_{b \in V(c)} Ref_C^{-1}(b). \tag{8.6}$$

3. Die Funktion dis: $D \times T \rightarrow C$ mit

$$\text{dis}(d,t) := \text{first}\{c \in Ref_C(t) \mid c \text{ maximiert } \text{tf}(d, U(c))\} \tag{8.7}$$

   erkennt den Sinn von Term t anhand des Kontextes, den ein Dokument d darstellt.
4. Damit ergibt sich die Konzepthäufigkeit zu:

$$\text{cf}(d,c) := \text{tf}(d, \{t \in T \mid \text{dis}(d,t) = c\}). \tag{8.8}$$

Intuitiv ausgedrückt, analysiert die Strategie alle Terme im Kontext, also im gleichen Dokument, und wählt dann als konzeptuelle Repräsentation das Konzept, das durch die meisten Termen aus der Nachbarschaft unterstützt wird. Nehmen wir das Beispiel aus Abschnitt 6.3.3.1 (siehe Abbildung 6.3) mit dem Term “Fork” noch einmal auf. Wird der Term “Fork” im Sinne von Besteck (im Folgenden als “Tableware-Konzept” bezeichnet) verwendet, hoffen wir weitere Worte, wie z.B. “tableware”, aus diesem Bereich im Dokument zu finden. Die Termhäufigkeit für dieses Konzept sollte dann in Gleichung 8.7 am höchsten sein. Handelt es sich bei dem Dokument allerdings um Beschreibungen über Computer und die Bedeutung des Wortes liegt eher im Verzweigen von Prozessen, so ist die Termhäufigkeit für das “Tableware-Konzept” eher klein und für “Branch-Konzept” hoch.

Mit Hilfe der vorgestellten Wortsinnerkennungsstrategien ist es uns in einem ersten Schritt möglich, synonyme Terme auf ein Konzept abzubilden. Im nächsten Schritt bietet es sich an, weitere Beziehungen der Ontologie in den Clusterprozess zu integrieren. Der nächste Abschnitt stellt die Integration der Taxonomie vor.

8.2.3.4 Strategien zur Integration von Oberkonzepten

Der dritte Teil unserer Analyse befasst sich mit der Menge an integriertem Hintergrundwissen. Uns stehen durch die Ontologie verschiedene Beziehungstypen zwischen Konzepten zur Verfügung. Die bekannteste Beziehung ist die Taxonomie. Die generelle Idee an dieser Stelle ist die taxonomische Beziehung zwischen den Konzepten auszunutzen und ebenfalls in die Repräsentation zu integrieren. Dazu fügen wir für einen Term nicht nur die Konzepte aus der Menge Ref_C hinzu, sondern auch eine *gewisse Anzahl* an generelleren Konzepten. Greifen wir das laufende Beispiel auf, so fügen wir bis jetzt bei der first-Strategie des letzten Abschnittes für den Term “fork” das passende Konzept (als “Tableware-Konzept”) hinzu und berechnen die entsprechenden Konzepthäufigkeiten. Wir würden

[5] Diese Strategie ist eine vereinfachte Version von [6].

nun auch die Oberkonzepte von “fork”, in diesem Fall “cutlery”, “tableware” usw., hinzufügen und die entsprechenden Konzepthäufigkeiten anpassen.

Die folgende Prozedur realisiert diese Idee und erhöht die Konzepthäufigkeiten der Oberkonzepte für ein Dokument d, indem es die Häufigkeiten der Unterkonzepte (für die nächsten $r \in \mathbb{N}$ Level der Hierarchie) einbezieht:

Wir aktualisieren den Konzeptvektorteil unserer Vektorrepräsentation $\vec{\tau_d}$ (siehe Gleichung 8.1) auf folgende Art und Weise:

Für alle $c \in C$ ersetzen wir $\text{cf}(d, c)$ mit

$$\text{cf}'(d, c) := \sum_{b \in H(c,r)} \text{cf}(d, b) \ , \tag{8.9}$$

wobei $H(c, r)$ in Abschnitt 7.2.2 in Gleichung 7.3 definiert wurde und für ein gegebenes Konzept c die r nächsten Unterkonzepte der Taxonomie liefert. Dies bedeutet für die folgenden Parameter:

$r = 0$**:** Diese Strategie ändert die Konzepthäufigkeiten nicht.
$r = n$**:** Diese Strategie fügt zu jedem Konzept die Häufigkeiten aller Unterkonzepte der n nächsten Ebenen der Ontologie hinzu.
$r = \infty$**:** Diese Strategie fügt zu jedem Konzept die Häufigkeiten aller seiner Unterkonzepte der Ontologie hinzu.

Auf diese Weise sind wir in der Lage, nun auch taxonomische Zusammenhänge der Ontologie in den Clusterprozess zu integrieren.

8.2.4 Aufbau der Experimente

Wir haben im letzten Abschnitt alle zu untersuchenden Fragestellungen sowie eine adequate Referenzclusterung zusammengetragen. Die folgenden Experimente stützen sich auf die Reuters-PRC-Datensätze, die durch ihre Labels/Bezeichner überhaupt erst eine Evaluierung ermöglichen. Ziel der Experimente ist es, mit Hilfe des Clusterprozesses die von den Labels gebildeten Gruppen möglichst gut nachzubilden. Abschnitt 8.2.1 zeigt die Ergebnisse für diese Aufgabe ohne Hintergrundwissen.

Um Hintergrundwissen in den Prozess zu integrieren, benötigen wir eine passende Ressource. Wir haben WordNet gewählt, weil es eine frei verfügbare, sehr umfassende und gut auf den Reuters-Corpus passende Ressource ist. Wir erhoffen uns von der Nutzung der WordNet Informationen eine Steigerung der Clusterergebnisse. Weitere Verbesserungen sollten Ontologien bringen, die speziell auf die analysierten Texte zugeschnitten sind.

Als Clusterverfahren setzen wir Bi-Sec-KMeans aus Abschnitt 5.4.2 ein, welches ein schnelles Cluster-Verfahren ist, das in anderen Studien bessere Ergebnisse als KMeans und vergleichbare Ergebnisse wie hierarchisch-agglomerative Clusterverfahren erzielt hat (vgl. [206]). Zum Vergleich der Ergebnisse werden wir die Maße Purity, InversePurity, F-Measure und Entropy einsetzen, die in der Literatur an unterschiedlichsten Stellen eingesetzt werden (siehe Abschnitt 5.3). Wir werden bei den Experimenten die Gewichtung und das Löschen von seltenen Worten im gleichen Umfang wie in Abschnitt 8.2.1 auch für Hintergrundwissen untersuchen. Zusätzlich werden wir für jede Variante noch alle Kombinationen der Integration von Hintergrundwissen in den Prozess analysieren (siehe letzter Abschnitt). Im Folgenden noch ein paar Punkte zu speziellen Annahmen oder Parametereinstellungen:

WordNet als Ontologie In unseren Experimenten nutzen wir WordNet als Ontologie. Dabei griffen wir nur auf die Substantive zurück, die 68.1 % aller Synsets ausmachen. Diese Synsets

Tabelle 8.3: Liste alle untersuchten Parameterkombinationen

Parameter Name	Werte
Korpus	PRC, PRC-min15, PRC-max100, PRC-min15-max100, PRC-max20, PRC-min15-max20
Stoppworte entfernen	ja
Wortstämme bestimmen	angewendet nur ohne Hintergrundwissen
Seltene Terme löschen	nein, 5 Terme, 30 Terme
Gewichten des Termvektors	tfidf, keine Gewichtung
Integration von Hintergrundwissen	add, replace, only
Anzahl der Oberkonzepte	0 und 5
Wortsinnerkennung	all, first, context
Anzahl Cluster k	5,10,20,30,50,60,70,100

betrachten wir als Konzepte unserer Ontologie. Weiterhin nutzen wir die Hypernymbeziehung als IsA-Taxonomie.

Porter-Stemmer gegenüber WordNet Normalerweise nutzen wir den Porter-Stemmer zum Reduzieren der Worte auf ihre Stammformen. In den Experimenten mit WordNet hat sich aber herausgestellt, dass die morphologische Komponente von WordNet bessere Ergebnisse liefert als der Stemmer. Das Stemmen der Terme basiert bei allen Experimenten mit Hintergrundwissen auf WordNet.

20 Wiederholungen Alle Ergebnisse beruhen auf 20 Wiederholungen mit unterschiedlichen Initialisierungen des Bi-Sec-KMeans-Algorithmusses. Wie in Abschnitt 5.4.2 beschrieben, ist dieser abhängig von der gewählten Startlösung. Wir präsentieren hier immer den Mittelwert dieser 20 Wiederholungen.

Clusteranzahl Wir variierten die Anzahl der Cluster von $k := 5, 10, 20, 30, 50, 60, 70$ bis 100. Unsere Intension war es dabei nicht, genau die gleiche Anzahl an Klassen, die aus dem manuellen Prozess hervorgegangen sind, zu entdecken. Dies erwies sich als nicht sinnvoll, da eine hundertprozentige Übereinstimmung der gefundenen Cluster mit den bekannten Klassen weder zu erwarten ist noch gefunden wurde. Wir führten auch vorab Tests mit sehr vielen Wiederholungen durch und setzten dabei die Anzahl der Cluster k entsprechend der Anzahl der im Originaldatensatz enthaltenen Klassen. Bi-Sec-KMeans zeigte bei dieser Anzahl Ergebnisse, die ähnlich den Ergebnissen mit leicht größerem oder kleinerem k sind. Eine hundertprozentige Übereinstimmung konnte nicht festgestellt werden. Vielmehr sollte ein sinnvolle Anzahl an Clustern bestimmt werden. Grund für diese Überlegung ist die bekannte Tatsache, dass auch beim manuellen Gruppierungsprozess mit mehreren Leuten unterschiedliche Ergebnisse, sprich eine unterschiedlichen Anzahl von Labels entstehen (vgl. [33, 38]). Wir werden daher mit unserem Clusteralgorithmus nur eine von vielen "Meinungen" berechnen können. Außerdem sollte mit der Variation der Clusteranzahl auch untersucht werden, inwieweit die Clusterergebnisse von der Anzahl der Cluster abhängen. Dabei erwarten wir für das Purity Maß ein Steigen der Güte mit steigender Clusteranzahl (bei $k = |D|$ ist Purity = 1) und für die InversePurity im gleichen Fall ein Sinken (bei $k = 1$ ist InversePurity = 1).

Tabelle 8.3 fasst die untersuchten Parametervariationen zusammen. Für die Referenzclusterung aus Abschnitt 8.2.1 untersuchten wir $20 \times 8 \times 6 \times 2 \times 3 = 5760$ Parametervariationen (Anzahl der

Testläufe $\times$ Anzahl der Clusteranzahlen $\times$ Anzahl der Korpora $\times$ Anzahl der Gewichtungsschemata $\times$ Anzahl der Termlöschstrategien), die wir dann mit $20 \times 8 \times 6 \times 2 \times 3 \times 3 \times 3 \times 2 = 103680$ (Anzahl der Testläufe $\times$ Anzahl der Clusteranzahlen $\times$ Anzahl der Korpora $\times$ Anzahl der Gewichtungsschemata $\times$ Anzahl der Termlöschstrategien $\times$ Anzahl der Strategien für die Anwendung von Hintergrundwissen $\times$ Anzahl der Wortsinnerkennungstrategien $\times$ Anzahl der verschiedenen Oberkonzepte) Parametervariationen für die Clusterläufe mit Hintergrundwissen verglichen haben.

8.2.5 Purity-Ergebnisse

Im Allgemeinen stellt man fest, dass Hintergrundwissen beim Clustern zu besseren Ergebnissen führt. Die umfangreichen Untersuchungen zeigen aber auch, dass Hintergrundwissen nicht in jedem Fall zu einer Verbesserung führt. Das Hintergrundwissen muss in geeigneter Form hinzugefügt werden, da sonst zum Teil wesentlich schlechtere Clusterergebnisse entstehen bzw. beobachtet wurden. Die Menge der Fehler während des Integrationprozesses von Hintergrundwissen dürfen nicht so groß werden, dass alle zusätzlichen zu besseren Clusterergebnissen führenden Informationen wieder verloren gehen. Um diese Aussage zu stützen, schauen wir uns gleich im Detail das Balkendiagramm aus Abbildung 8.6 an. Es gibt die Ergebnisse von Clusterungen mit tfidf Gewichtung wieder.

Die Ergebnisse aller Testläufe in Form von Tabellen sind sehr umfangreich und wurden daher auf dem Web unter http://www.aifb.uni-karlsruhe.de/WBS/aho/clustering/ abgelegt. Wir präsentieren im Folgenden eine Auswahl dieser Ergebnisse zum Teil als Grafik und zum Teil als Tabelle. Die Abbildungen 8.4 und 8.5 für die schon Eingangs erwähnten Clusteranzahlen 5, 10, 20, 30, 50, 60, 70 und 100 geben jeweils die durchschnittliche Purity für die Strategien ohne und mit Hintergrundwissen wieder. Man sieht in den Fällen mit 5 und 10 Clustern gleichen sich die Ergebnisse sehr stark für alle Strategien. Die tatsächliche Clusteranzahl und die berechnete liegen hier zu weit auseinander. Im Wesentlichen stellt man für alle überigen Clusteranzahlen (> 10 Cluster) bei genauerer Analyse der Grafiken sehr ähnliche Ergebnisse der verschiedenen Strategien bei den unterschiedlichen Clusteranzahlen fest. Daher konzentrieren wir uns im Folgenden auch auf nur eine Clusteranzahl, nämlich 60.

Auf ein Phänomen sei an dieser Stelle noch hingewiesen. Der schlechteste Wert (bei 5 Clustern) für den PRC-min15-max100 Datensatz liegt bei ca. 17 %. Im Gegensatz dazu findet man den schlechtesten Wert für den PRC-Datensatz bei ca. 50 %. Dies lässt sich leicht anhand der Verteilung der Klassen des PRC- und des PRC-min15-max100-Datensatzes erklären (vgl. hierzu Abschnitt 8.2.2).

Wir werden nun auf das Lesen der Grafiken aus den Abbildungen 8.6, 8.7, 8.9 und 8.10 eingehen, bevor wir dann die Ergebnisse im Detail vorstellen und analysieren.

Der erste Wert der drei Balken ganz links in Abbildung 8.6 (Ontology=false) stellt die Referenz aus den Clusterläufen ohne Hintergrundwissen dar. Alle weiteren Werte in der Grafik sind Clusterläufe mit Hintergrundwissen. Zur besseren Vergleichbarkeit wurde in der Grafik für jeden Referenzwert (Baseline) aus der Clusterung ohne Hintergrundwissen eine waagerechte Linie eingezeichnet. Wir haben drei Referenzwerte für die drei verschiedenen Prunethresholds (0, 5, 30). Werte, die über dem Referenzwert liegen, stellen bessere und Werte unter dem Referenzwert stellen schlechtere Ergebnisse dar. Jeder Balken entspricht dem Mittelwert von 20 wiederholten Clusterungen mit unterschiedlichen Startlösungen initialisiert. Auf der Y-Achse ist der durchschnittliche Purity-Wert abgetragen, wobei die Standardabweichung zwischen 0.6 % und 2.3 % schwankt.

Die erste Spalte der X-Achse in Abbildung 8.6[6] gibt die Ergebnisse für das Clustern ohne Hinter-

[6]Die X-Achse ist in 19 “breite” Spalten eingeteilt, die ihrerseits drei Clusterergebnisse enthalten. Wir verwenden diese

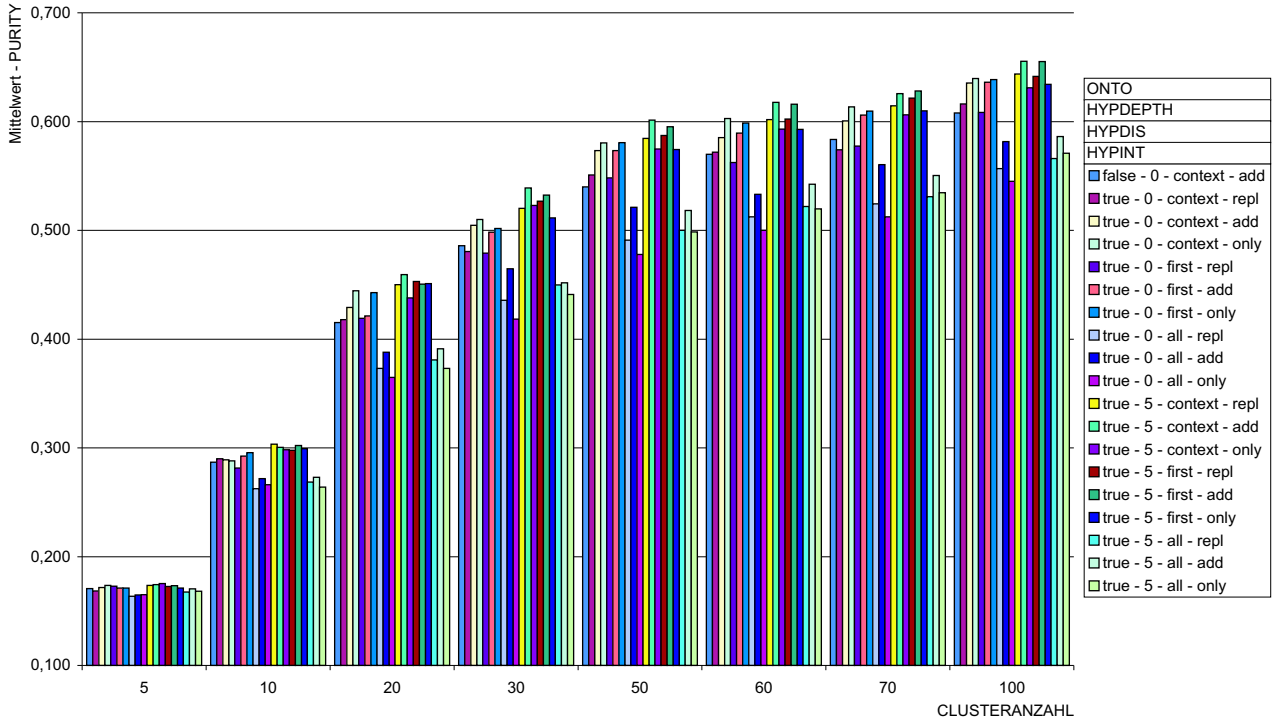

Abbildung 8.4: stellt die Clusterergebnisse für die Anzahl 5, 10, 20, 30, 50, 60, 70, 100 mit Gewichtung, Prunethreshold 30, ohne und mit Hintergrundwissen und hier für alle Strategien für PRC-min15-max100 dar

grundwissen wieder. Den restlichen Spalten entnimmt man die Ergebnisse für “Ontology=true”. Bei den Ergebnissen markiert mit “HYPDEPTH=0” an der X-Achse (Spalten 2-10) wurden keine Oberkonzepte hinzugefügt, bei “HYPDEPTH=5” (Spalten 11-19 der X-Achse) sind es fünf. Die Balken 2-4 zeigen das Ergebnis der Disambiguierungsstrategie “context” bei unterschiedlicher Integration (“repl”, “add”, “only”) der Konzeptvektoren in den Wortvektor. Die Spalten 5-7 und 8-10 sind analog zu den Spalten 2-5 aufgebaut. Wir entnehmen der Abbildung die Gesamtstrategie der Clusterung einer Spalte, indem wir die verschiedenen Strategien ablesen, z.B. entspricht Spalte 2 den Strategien: mit Ontologie, keine Oberkonzepte, Wortsinnerkennung mit “Context”-Strategie und Integration mittels “repl”-Strategie.

Die Ergebnisse aus Abschnitt 8.2.1 lassen sich auch auf das Clustern mit Hintergrundwissen übertragen. Einzige Ausnahme stellt die “all”-Strategie bei der Integration der Konzeptvektoren dar. Sie ist insgesamt deutlich schlechter und für alle Prunethresholds ungefähr gleich. Es konnten keine signifikanten Unterschiede bestimmt werden.

PRC-min15-max100 Bei der Analyse der Abbildungen 8.6 und 8.7 ist zu erkennen, dass die Verbesserung ohne tfidf Gewichtung nur sehr gering ist, 47 % ohne gegenüber 48,6 % mit Hintergrundwissen im besten Fall (first, add, 0). Auffällig ist an der Verbesserung (die sehr gering aber noch signifikant mit einem $\alpha = 0.5\ \%$ ist), dass keine Nutzung der Oberkonzepte erfolgte. Werden diese hinzugefügt, so beobachtet man mindestens 6 % schlechtere Ergebnisse. Auf der anderen Seite findet man die größte Verbesserung der Purity bei den tfidf gewichtete Vektoren unter den um Oberkonzept erweiterte Vektoren (context- und add-Strategie). Der Wert der Baseline wird von 57 % auf 61,8 % gesteigert (vgl. Abbildung 8.6).

Spaltennummer, um die Grafik zu erläutern.

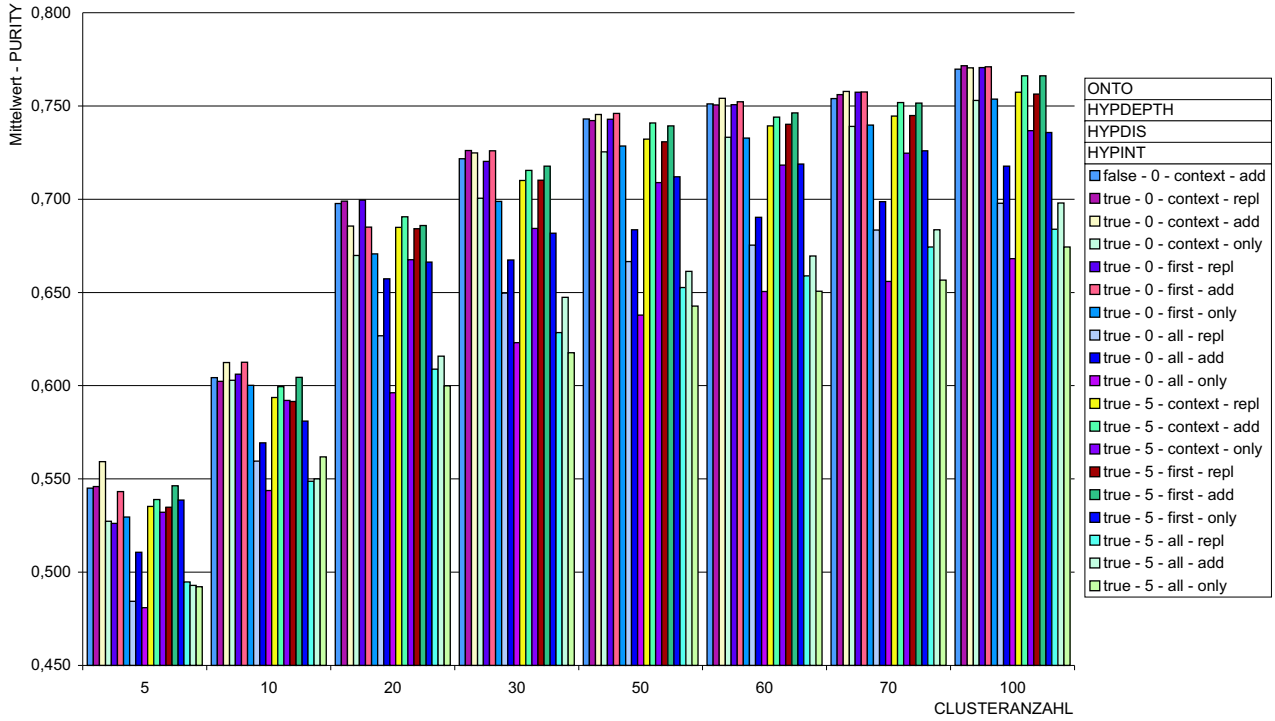

Abbildung 8.5: stellt die Clusterergebnisse für die Anzahl 5, 10, 20, 30, 50, 60, 70, 100 mit Gewichtung, Prunethreshold 30, ohne und mit Hintergrundwissen und hier für alle Strategien für PRC dar

Vergleicht man die drei Wortsinnerkennungsstrategien context, first und all stellt man keine signifikaten Unterschiede zwischen den beiden Strategien context und first fest. Die all Strategie ist in jedem Fall schlechter. Auffällig ist noch, dass bei der Nutzung der ungewichteten Term- bzw. Konzepthäufigkeiten der Abfall der all Strategie nicht so extrem ausfällt (vgl. Abbildung 8.7).

Um fast 15 % sinkt die Purity (vgl. Abbildung 8.7), wenn man die all-Strategie bei ungewichteten Vektoren auch noch mit fünf Oberkonzepten kombiniert. Der Abfall fällt nicht ganz so drastisch bei den anderen beiden Wortsinnerkennungsstrategien aus. Mit ca. 5 % sinkt die Purity aber immer noch beachtlich. Ganz anders sieht die Situation bei den gewichteten Vektoren aus (vgl. Abbildung 8.6). Hier ist die all-Strategie zwar insgesamt immer noch leicht schlechter, aber auf alle Fälle sind die Ergebnisse unter Nutzung der fünf Oberkonzepte besser als ohne die Nutzung dieser. Die beiden Strategien context und first führen unter Nutzung der Oberkonzepte zu den besten Resultaten, die bei der Integration von Hintergrundwissen in Form von Ontologien erzielt wurden. Wir ziehen aus den Beobachtungen den Schluss, dass der Gewichtung der Vektoren eine ganz entscheidende Rolle zukommt. Weiterhin ist es wichtig, eine gewisse Wortsinnerkennung bei der Integration zu berücksichtigen.

Beim Vergleich der drei Integrationsstrategien rep, add und only unabhängig von der Gewichtung schneidet die add-Strategie immer am besten ab. Die only-Strategie ist bei der Nutzung von Oberkonzepten meist etwas schlechter als die repl-Strategie und ohne die Nutzung von Oberkonzepten etwas besser (vgl. Abbildung 8.6 und 8.7 sowohl für die context als auch für die first-Strategie). Die Unterschiede sind aber nur zum Teil signifikant.

Kaum Unterschiede in den Ergebnissen stellt man zum PRC-max100 Datensatz fest.

PRC-max20 Der PRC-max20 Datensatz enthält nur eine sehr geringe Anzahl an Dokumenten pro Klasse. Das Einteilen der Dokumente in Gruppen ist bei einem solchen Datensatz am schwierigsten. Das Hintergrundwissen hilft hier aber auch am meisten. So beträgt die relative Verbesserung

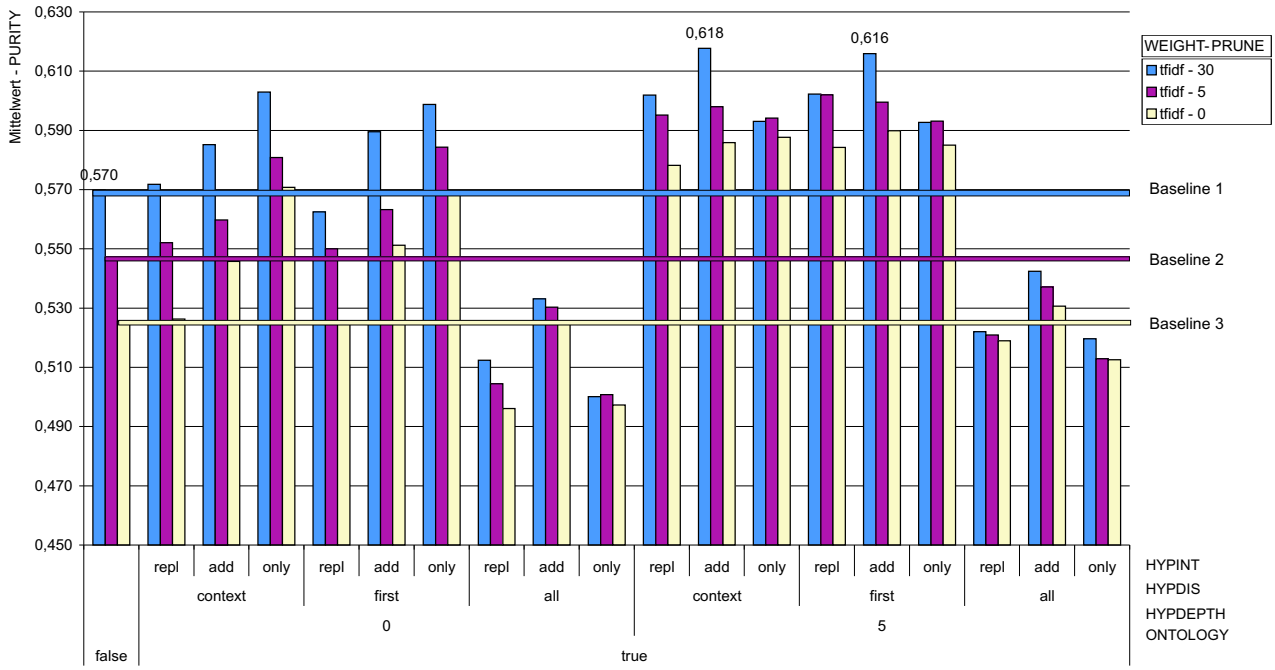

Abbildung 8.6: Vergleicht alle Clusterergebnisse *mit Gewichtung* für Strategien mit Hintergrundwissen mit den Ergebnissen ohne Hintergrundwissen für 60 Cluster für PRC-min15-max100

statt 8.5 % bei PRC-min15-max100 jetzt 11.1 %. Der Link, der durch die Oberkonzepte zwischen den Dokumenten gleicher Klassen in die Vektorrepräsentation integriert wird, erzeugt bei diesem Datensatz den größten Nutzen. Bei Datensätzen mit mehr Dokumenten steigt die Wahrscheinlichkeit, dass zwei Terme, die durch die Ontologie explizit verbunden sind, auch gemeinsam in einem Dokument vorkommen. Anschaulich gesprochen sind daher beide Terme auch in einem gemeinsamen Clusterzentroiden wiederzufinden.

Wie auch schon beim PRC-min15-max100 Datensatz beobachtet, ist die only-Strategie ohne Berücksichtigung von Oberkonzepten besser als die repl- und auch die add-Strategie. Dieses Bild ändert sich, wenn man die Oberkonzepte mit einbezieht. Dann ist die add-Strategie klar besser und only schneidet am schlechtesten ab.

PRC Analysieren wir zum Abschluss dieses Abschnittes den PRC-Datensatz. Wir haben gesehen, wie Hintergrundwissen die Güte von Clusterergebnissen bei Klassen, die kleinere bis mittlere Mengen an Dokumenten enthalten, steigern kann. Wenige Dokumente bedeutet hier weniger als 20 bzw. als 100 Dokumente pro Klasse. In diesen Fällen ist eine Verbesserung des Ergebnisses festzustellen. Keine Verbesserung findet man beim PRC-Datensatz (vgl. Abbildung 8.9 und 8.10).

Mit unserer besten Strategie, in diesem Fall add, context, ohne Oberkonzepte, konnten wir eine kleine aber nicht signifikante Verbesserung erzielen. Auch die für den PRC-min15-max100 Datensatz beobachteten Ergebnisse, dass die all-Strategie schlechter als context- und die first- ähnlich gut wie die context-Strategie ist, findet man beim PRC-Datensatz wieder. Erstaunlich ist, dass durch das Hinzufügen der Oberkonzepte die Ergebnisse nicht besser, sondern schlechter werden. In Abschnitt 8.3 konnten wir zwar eine Erklärung für die schlechten Ergebnisse des PRC-Datensatzes erarbeiten, wobei die Ursache bei den großen Klassen wie “earn” liegt. Den Abfall der Ergebnisse bei der Nutzung der fünf Oberkonzepte konnten wir damit aber nicht erklären.

Abbildung 8.10 gibt die Ergebnisse für die Clusterläufe ohne Gewichtung wieder. Sie decken sich mit den Ergebnissen des PRC-min15-max100 Datensatzes.

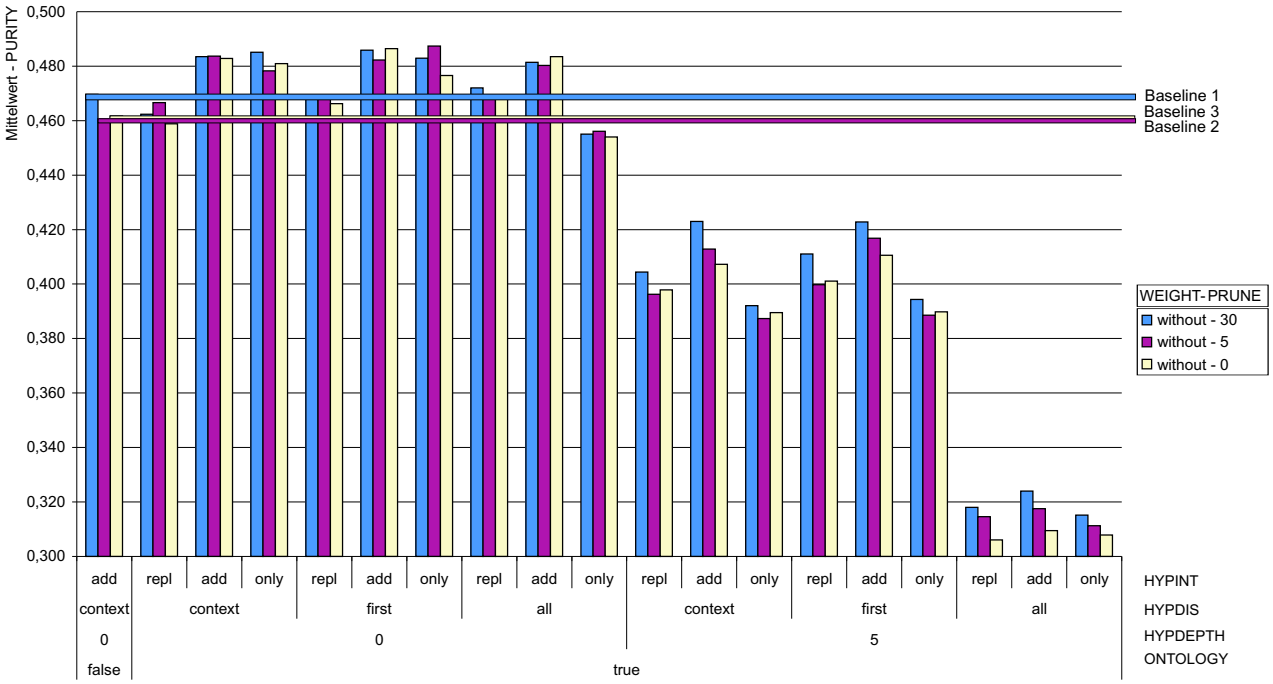

Abbildung 8.7: Vergleicht alle Clusterergebnisse *ohne Gewichtung* für Strategien mit Hintergrundwissen mit den Ergebnissen ohne Hintergrundwissen für 60 Cluster für PRC-min15-max100

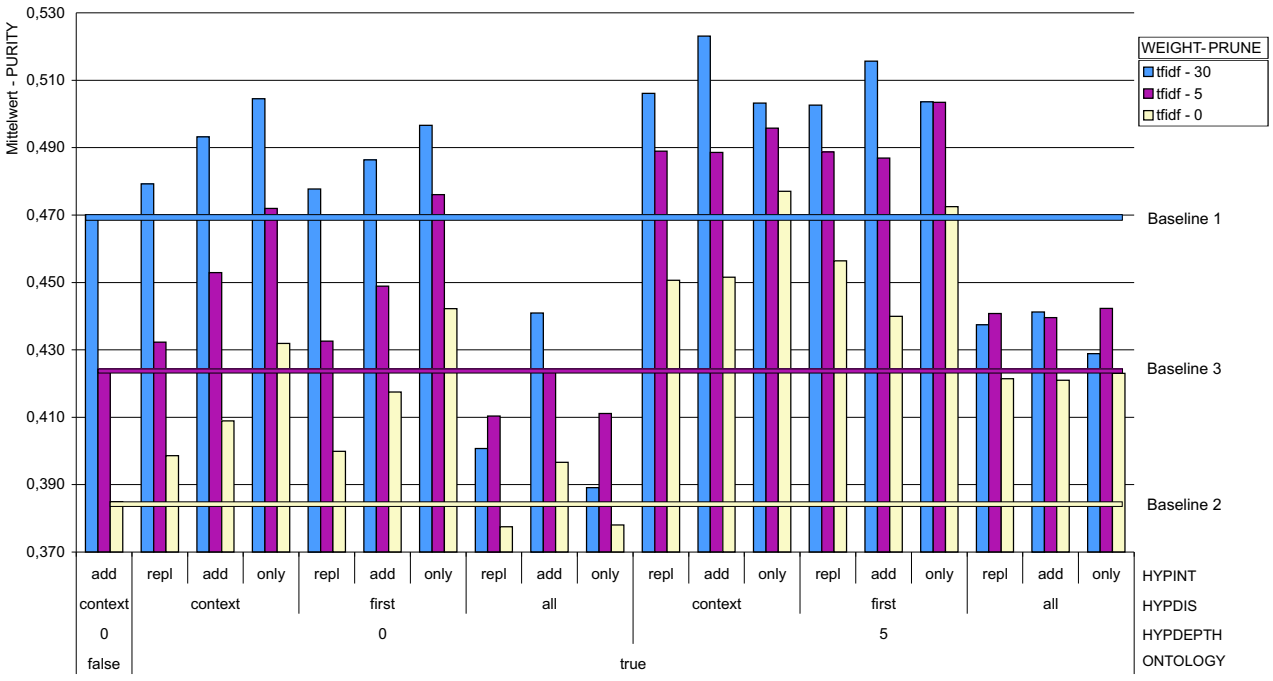

Abbildung 8.8: Vergleicht alle Clusterergebnisse *mit Gewichtung* für Strategien mit Hintergrundwissen mit den Ergebnissen ohne Hintergrundwissen für 60 Cluster für PRC-max20

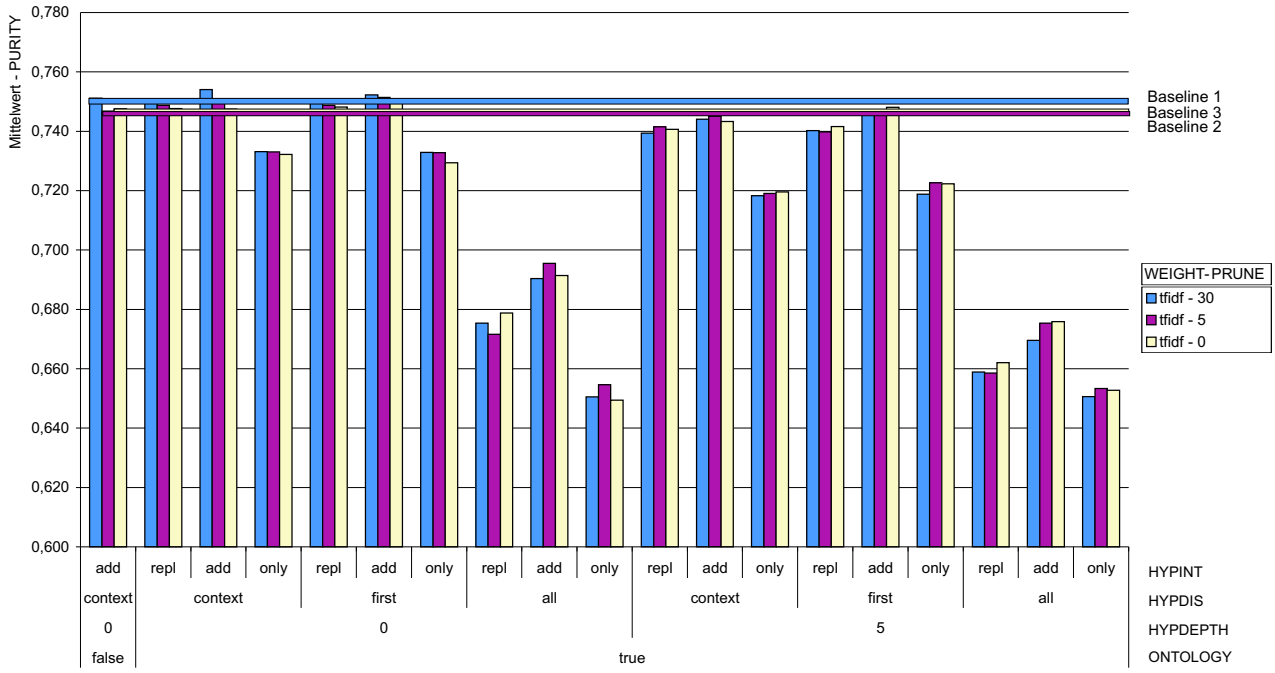

Abbildung 8.9: Vergleicht alle Clusterergebnisse *mit Gewichtung* für Strategien mit Hintergrundwissen mit den Ergebnissen ohne Hintergrundwissen für 60 Cluster für PRC

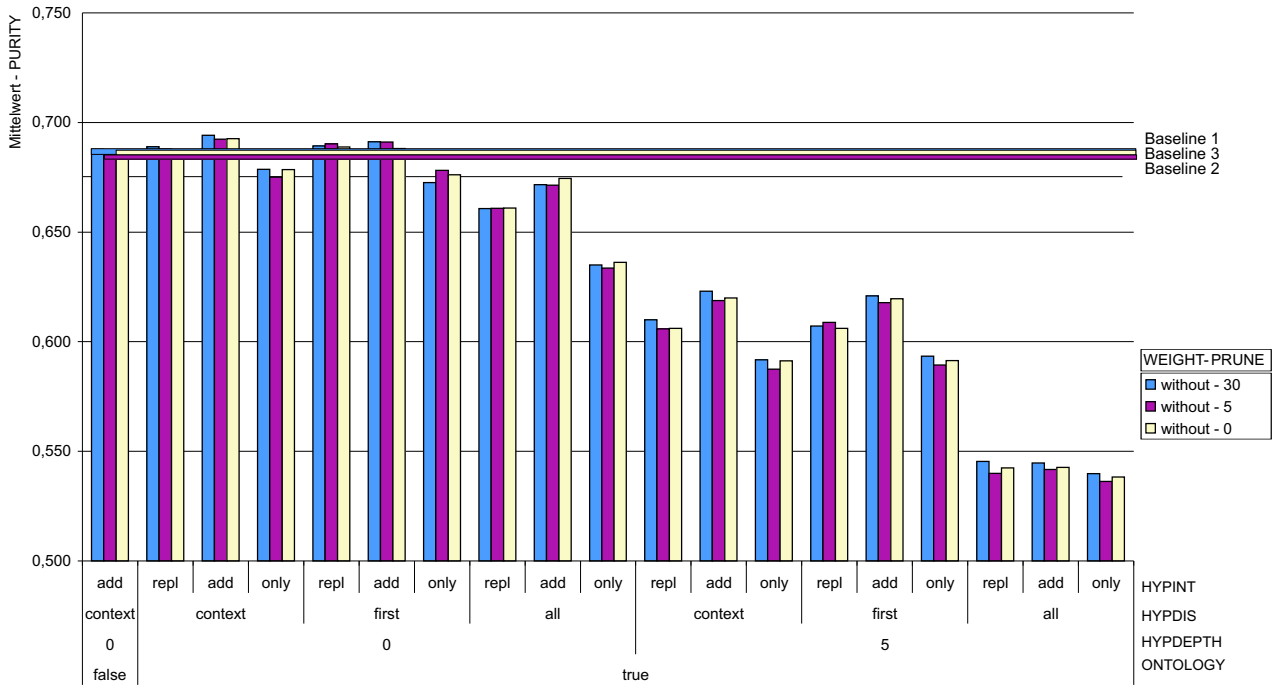

Abbildung 8.10: Vergleicht alle Clusterergebnisse *ohne Gewichtung* für Strategien mit Hintergrundwissen mit den Ergebnissen ohne Hintergrundwissen für 60 Cluster für PRC

Tabelle 8.4: Ergebnisse für den PRC-Datensatz mit $k = 60$, prune = 30 (mit Hintergrundwissen und HYPDIS = context, avg markiert den Mittelwert von 20 Clusterläufen und std die Standardabweichung)

ONTO	HYPDEPTH	HYPINT	Purity avg ± std	InversePurity avg ± std
false			**0,751 ± 0,006**	**0,263 ± 0,007**
true	0	add	**0,755 ± 0,007**	0,269 ± 0,009
		only	0,736 ± 0,008	0,266 ± 0,009
	5	add	0,746 ± 0,006	**0,272 ± 0,007**
		only	0,721 ± 0,007	0,271 ± 0,010

Tabelle 8.5: Ergebnisse für den alternativen PRC-min15-max100-Datensatz (neue Stichprobe) mit $k = 60$, prune=30 (mit Hintergrundwissen und HYPDIS = context, avg makiert den Mittelwert von 20 Clusterläufen und std die Standardabweichung)

Onto	HD	HI	Purity avg ± std	InversePurity avg ± std	F-Measure avg ± std	Entropy avg ± std
false			**0,546 ± 0,015**	**0,435 ± 0,016**	**0,479 ± 0,016**	**1,329 ± 0,038**
true	0	add	0,567 ± 0,020	0,449 ± 0,018	0,492 ± 0,017	1,260 ± 0,052
		only	0,585 ± 0,018	0,460 ± 0,020	0,504 ± 0,021	1,234 ± 0,038
	5	add	**0,602 ± 0,017**	**0,473 ± 0,019**	**0,514 ± 0,019**	**1,178 ± 0,040**
		only	0,589 ± 0,017	0,459 ± 0,017	0,500 ± 0,016	1,230 ± 0,039

8.2.6 InversePurity-Ergebnisse

Neben der Purity als Maß zur Beurteilung der Clusterergebnisse bieten sich Maße wie die InversePurity (vgl. Gleichung 5.15) als entgegengesetztes Maß an. Bevorzugt die Purity eine große Clusteranzahl und bestraft nicht die Aufteilung großer Originalklassen, so bewertet die InversePurity eher wenige große Cluster positiv und reagiert auch sensibel auf das Aufteilen von Clustern. Die InversePurity fragt, welcher Cluster am besten eine gewisse Klasse widerspiegelt. Beide Maße sind Gegenspieler. Werden beide Maße größer, ist die Clusterung auf jeden Fall der zu vergleichenden Klasseneinteilung ähnlicher. Falls die Purity steigt, aber die InversePurity sinkt, wird eine Aussage bzgl. der Clustergüte schwierig. Wir haben daher die beste Referenzclusterung für die beiden Datensätze PRC-min15-max100 und PRC mit den jeweiligen Clusterungen mit Hintergrundwissen anhand der InversePurity verglichen. Die Tabellen 8.4 und 8.5 fassen die Ergebnisse zusammen.

Wie zuvor schon gesehen unterscheiden sich die Purity-Werte des PRC-Datensatzes nicht signifikant für eine typische Strategie mit Hintergrundwissen (Hypdis = context, prune = 30, HYPDEPTH = 5, HYPINT = add) und der Referenzclusterung. Bei der InversePurity beobachten wir eine kleine aber signifikante Verbesserung des Ergebnisses innerhalb des Konfidenzintervalles von 0.5 %. Ganz anders sieht die Situation beim PRC-min15-max100 Datensatz aus. Wir erhalten klare Verbesserungen der Ergebnisse sowohl für die Purity- als auch für die InversePurity-Werte bzgl. der gleichen Hintergrundwissensstrategie (vgl. Tabelle 8.5).

Die Ergebnisse der Tabelle 8.5 basieren auf einer anderen zufällig gezogenen Menge an Dokumenten (Ergebnisse [116] entnommen). Der Datensatz enthält noch immer die PRC-min15-max100-Verteilung der Dokumente in den Klassen. Die Ergebnisse decken sich mit den in der Arbeit bisher vorgestellten Ergebnisses in Bezug auf die Purity. Tabelle 8.5 gibt neben der InversePurity auch noch weitere gängige Maße wie F-Measure und Entropie aus dem Bereich Information Retrieval (siehe Abschnitt 5.3.3.3 und 5.3.3.4) wieder. Die prinzipielle Aussage, dass die Repräsentation mit Hintergrundwissen zur Steigerung der Clustergüte beiträgt, wird durch alle Maße bestätigt.

8.2.7 Zusammenfassung und weitere Schritte

Im Allgemeinen konnten wir bei der Nutzung von Hintergrundwissen Folgendes beobachten:

- Hintergrundwissen steigert die Clustergüte am meisten bei Klassen mit wenigen Dokumenten. Die besten Ergebnisse wurden bei Datensätzen mit maximal 20 Dokumenten pro Klasse erzielt. Aber auch alle anderen Datensätze waren im besten Fall nie schlechter als die Baseline.
- Ohne die Nutzung einer primitiven Wortsinnerkennung (Word Sense Disambiguation) erhält man keine besseren Ergebnisse für die Nutzung von Hintergrundwissen. Auf der anderen Seite reichen schon recht einfache Verfahren aus, um Clusterergebnisse zu verbessern.
- Löscht man seltene Worte nicht, so führt dies immer zu einer Verschlechterung der Ergebnisse. Vergleichen wir die Ergebnisse beim Löschen von 5 und 30 Worten, so sind die Ergebnisse bei 30 immer besser.

Alle Ergebnistabellen sind im Internet unter: http://www.aifb.uni-karlsruhe.de/WBS/aho/clustering/ zu finden.

Fazit ist, dass, wenn Ontologien oder ähnliche Ressourcen passend zu einem Korpus zur Verfügung stehen, diese auf jeden Fall eingesetzt werden sollten.

Bei der Analyse der bisherigen Ergebnisse zeigten sich Punkte, die spannende Aufgaben für die Zukunft darstellen und viel Potential zur weiteren Steigerung der Clustergüte versprechen. So ergab sich, dass z.B. die Klasse “alum” (Aluminium) trotz der Einbeziehung von Oberkonzeptbeziehungen über mehrere Cluster verteilt wurde (Reutersklassen siehe Anhang E). Eine Inspektion der Dokumente führte schnell auf die Ursache. Während einige der Dokumente direkt den Term “Aluminium” enthalten, kommt in anderen Dokumenten kein auch nur aus der Umgebung von Aluminium stammender Term vor. Wir entdeckten aber den Term “Bauxit”, den Namen des Minerals, aus dem Aluminium gewonnen wird. Wir fragten uns, ob WordNet diese beiden Begriffe geeignet in Beziehung setzt und entdeckten die direkte Meronym-Beziehung (Teil von) zwischen beiden. Die Nutzung dieser Informationen zur Steigerung der Clustergüte sollte einer der nächsten Schritte sein. Dabei ist aber zu beachten, dass man nicht jeden Term mit jedem in Beziehung setzt. So würde nur unnötig Rauschen in die Repräsentation der Terme integriert.

Im folgenden Abschnitt gehen wir auf verwandte Ansätze ein. Abschnitt 8.3 analysiert die Auswirkungen der Repräsentationsänderung und liefert eine Begründung für den Erfolg des Ansatzes.

8.2.8 Verwandte Ansätze zum Textclustern mit Hintergrundwissen

Dieser Abschnitt vergleicht unseren Ansatz mit verwandten Ansätzen aus dem Bereich des Textclusterns. Der Fokus liegt dabei auf der Nutzung von Hintergrundwissen zur Lösung der Clusteraufgabe. Es ist uns zur Zeit keine direkte Nutzung von Hintergrundwissen in der in dieser Arbeit beschriebenen Form während des Clusterns bekannt. Mögliche alternative Clusterverfahren für den Bereich des Textclusterns wurden in Kapitel 5 vorgestellt. Die im Folgenden diskutierten Ansätze sind mit dem Ansatz dieser Arbeit verwandt bzw. basieren auf ähnlichen Ideen aus benachbarten Forschungsgebieten. Wir tragen daher an dieser Stelle die verwandte Literatur zusammen und grenzen unseren Ansatz von den Ideen der anderen Ansätze ab.

WordNet zur Verbesserung von Information Retrieval Im Bereich Information Retrieval haben sowohl Voorhees in [224] als auch Moldovan und Mihalcea in [172] die Möglichkeit

untersucht, WordNet für das wortbasierte Suchen nach Dokumenten nutzbar zu machen. Die Evaluierung erfolgte mittels der im Information Retrieval üblichen Maße Precision und Recall. Die Untersuchungen zeigten schnell, dass die Verbesserung der Ergebnisse nicht so einfach möglich ist. Erfolgreich konnte [86] WordNet zur Wortsinnerkennung nutzen. Gonzalo et.al. erstellten *manuell* einen Synsetvektor. Sie konnten eine Steigerung der Information Retrieval Ergebnisse gegenüber einem Wortvektormodell zeigen.

Die Ergebnisse decken sich mit Teilen unserer Ergebnisse, da, wie gezeigt, einiger Aufwand und die Auswahl der richtigen Strategie notwendig ist, um WordNet erfolgreich zu integrieren. Unser Ansatz hat den Vorteil, dass er nicht auf WordNet beschränkt ist und die Abbildung automatisch und nicht manuell erfolgt.

WordNet zur Text-Dokument-Klassifikation Buenaga Rodríguez u. a. [45] und Ureña Lóez u. a. [150] zeigen die erfolgreiche Integration von WordNet zur Verbesserung der Dokumentklassifikation. Zur Evaluierung nutzten sie den Reuters-Korpus und zeigten für den Rocchio- und den Widrow-Hoff-Algorithmus eine Steigerung von bis zu 20 Prozentpunkten. Ihr Ansatz stellt in gewisser Weise eine obere Schranke bei der Performance dar, da die Auswahl der Synonyme für jede Kategorie manuell erfolgte. Anschließend wurden die im Synset enthaltenen Terme mit ausgewählten Gewichten dem Wortvektor hinzugefügt. Die Menge der Fehler wird bei einem automatischen Verfahren zur Wortsinnerkennung und Auswahl der korrekten Synsets sehr wahrscheinlich höher liegen und dürfte so die Güte der Ergebnisse reduzieren.

Die Ergebnisse sind unseren Ergebnissen sehr ähnlich und zeigen, welche Steigerung der Clustergüte noch möglich ist.

Statistische Konzepte Die Idee “Konzepte” statt Terme für das Clustern zu verwenden ist nicht neu. Es sind verschiedene Ansätze bekannt, denen gemeinsam ist, dass sie nicht die gleiche Art von Konzepten wie in dieser Arbeit verwenden. Deerwester u.a. [48] haben die Methode Latent Semantic Indexing (kurz LSI) erfunden. Sie basiert auf einer Singulärwertzerlegung der Dokument-Term-Matix (vgl. Abschnitt 4.4). Man kann aus dieser Zerlegung einen neuen Raum ableiten, in dem es dann eine Art “statistische Konzepte” gibt. Die Methode geht davon aus, dass nicht alle Singulärwerte gleich viele Informationen enthalten und spannt den neuen Raum nur über die größten Singulärwerte auf. [109] und [32] zeigen die Anwendung einer erweiterten auf Wahrscheinlichkeiten basierenden Version von LSI, nämlich PLSI (probabilistic), auf Textdokumente. Karypis und Han nutzen in [126, 127] Clustering-Methoden zur Berechnung von Wortclustern. Die Wortcluster stellen hier die neuen Konzepte dar. Sie vergleichen ihre Ergebnisse mit LSI und konnten zeigen, dass bei reduziertem Aufwand für die Berechnung der Konzepte ähnlich gute Ergebnisse erzielt werden können.

Entscheidender Nachteil dieser Methoden ist die Art und Weise der Berechnung der Konzepte. Die statistischen Konzepte bestehen im Ergebnis aus einer Linearkombination der Originalterme, d.h. anteilig kommt jeder Term in einem statistischen Konzept vor. Das Auftreten eines statistischen Konzeptes ist daher für einen Anwender nur schwer zu verstehen. Konzepte der Ontologie, wie sie unser Ansatz verwendet, erlauben eine einfache Interpretation und die Ergebnisse sind für den Anwender leicht verständlich. WordNet oder allgemeiner Konzepte einer Ontologie bieten hier einen klaren Vorteil gegenüber den statistischen Konzepten.

Arbeiten zu Ähnlichkeitsmaßen Alexander Strehl schlägt in [209] domänenspezifische Ähnlichkeitsmaße vor. Er zeigt die Notwendigkeit der Anpassung von Ähnlichkeitsmaßen an die gegebene Aufgabe, um gute Ergebnisse zu erzielen. Im Vergleich zu unseren Ergebnissen versucht

er, das gleiche Ziel über die Änderung des Maßes und nicht die Änderung der Repräsentation zu erreichen. Beide Anpassungen sind aber domänenspezifisch, d.h. nicht jedes Maß bzw. nicht jede Ontologie kann für alle Aufgaben eingesetzt werden.

WordNet und Clustern von Textdokumenten Green beschreibt in [89] und [90], wie er lexikalische Synset-Ketten (lexikal chains) aus Dokumenten gewinnt. Die Ketten bestehen aus Synsets von WordNet, auf die die Worte eines Dokumentes abgebildet werden, wobei gleichzeitig der Sinn der Worte anhand verbindbarer Synsets erkannt wird. Die Synsets der Wortketten sowie alle zu diesen Synsets verwandten Synsets bilden die Basis für die Vektorrepräsentation der Dokumente, die dann zum Clustern verwendet werden. Dabei wird ein Dokument durch zwei Vektoren - einer für die direkt gefunden Synsets (Member) und einer für die abgeleiteten Synsets (Linked) - repräsentiert. Lässt man die unterschiedliche Wortsinnerkennung außer Acht und würde Green nur einen Member-Vektor zum Clustern verwenden, so entspricht dies unserem Ansatz mit der Strategie "WordNet only". Der Linked-Vektor von Green folgt der gleichen Idee wie bei uns das Einbeziehen der Hypernyme. Die Ansätze sind aber schwer vergleichbar, da wir nur einen Vektor verwenden. Auch liefert Green keine Aussage, wie gut oder schlecht seine Repräsentation gegenüber der gewöhnlichen "Bag of Words"-Methode abschneidet. Alle anderen von uns diskutierten Strategien sind neu.

Dave u.a. versuchen in [43] ebenfalls Elemente von WordNet als Attribute für das Clustering zu verwenden. Dabei verwenden sie keine Wortsinnerkennung, was zu einer Verschlechterung der Clusterergebnisse führt. Dies deckt sich mit unseren Ergebnissen. Aus der Arbeit geht leider nicht hervor, wie der Wortvektor basierend auf WordNet aufgebaut ist, so dass hier kein Vergleich angestellt werden kann.

Hatzivassiloglou et.al. stellen in [102] einen Vergleich mehrerer Clusterverfahren auf der Basis eines "Bag of Words"-Modell vor. Zwei Verfahren – Single Pass und Groupwise-Average Hierarchial Clustering – werden auf ein um linguistische Features erweitertem "Bag of Words"-Modell angewendet. Dabei konnte anhand des TDT2 Datensatzes gezeigt werden, dass die linguistischen Features allein zu schlechteren Clusterergebnissen führen und zu leicht besseren Ergebnissen in Kombination mit allen anderen Worten. Diese Kombination von Worten mit speziell vorverarbeiteten Merkmalen entspricht unserer "Add" Strategie. Die Ansätze dieser Arbeit verwenden als Merkmale zur Erweiterung des Vektors Synsets von WordNet bzw. Konzepte. Hatzivassiloglou u.a. erkennen mittels "Part of Speech"-Tagger sowie weiterer Heuristiken bestimmte Satzteile oder Namen, die sie als Merkmale in den Vektor integrieren. Denkbar wäre eine Kombination beider Ansätze, um weitere Verbesserungen der Clustergüte zu erzielen.

8.3 Analyse der Repräsentationsänderung

Wie in Abschnitt 8.2 gezeigt, bewirkt die Bag-of-Konzept-Repräsentation häufig eine Verbesserung der Clustergüte. Leider konnte dies nicht durchgängig beobachtet werden. Ziel dieses Abschnittes ist, die Ursache für diese Beobachtung zu finden.

Für die Analyse der Repräsentationsänderung benötigen wir ein geeignetes Mittel, um die Änderung der Repräsentation bewerten zu können. Nehmen wir den Reutersdatensatz, so sollte die Varianz innerhalb einer gegebenen Reutersklasse nach der Änderung der Repräsentation niedriger sein als vorher. Dies wäre eine Möglichkeit zu überprüfen, ob die Nutzung von Hintergrundwissen sich auf die neue Repräsentation auswirkt.

Wir nutzen zu Analysezwecken die Varianz oder auch Streuungsquadratsumme einer Dokumentmenge $X \subset D$. Sie berechnet sich nach folgender Gleichung:

$$\mathrm{var}(X) = \sum_{d \in X} \sum_{t \in T} (t_d - \bar{t_X})^2, \tag{8.10}$$

wobei t_d dem aktuellen Wert für den Term t im Dokument d und $\bar{t_X}$ dem Mittelwert des Terms über alle Dokumente der Menge X entspricht.

Die Varianz über alle Klassen $\mathbb{L}$ einer Clusterung oder auch einer manuellen Klassifikation ergibt sich dann zu:

$$\mathrm{var}(\mathbb{L}) = \sum_{L \in \mathbb{L}} \mathrm{var}(L) \tag{8.11}$$

Bei der Veränderung der Repräsentation wird sich die Varianz des gesamten Datensatz ebenfalls ändern. Gut wäre eine Reduktion der Innerklassenvarianz, da so die Klassen einfacher gefunden werden können. Die Varianzreduktion innerhalb der einzelnen Klassen sollte auch größer sein als die veränderte Gesamtvarianz. Um den Effekt Gesamt- und Innerklassenvarianz herauszurechnen, normieren wir die Varianz wie folgt:

$$\mathrm{var}_{in}(\mathbb{L}) := \frac{\mathrm{var}(\mathbb{L})}{\mathrm{var}(D)}. \tag{8.12}$$

Die Varianz kann für die Vektorrepräsentationen mit und ohne Hintergrundwissen berechnet werden. So erhalten wir zwei Werte $\mathrm{var}_{in}^{with}(L)$ (mit) und $\mathrm{var}_{in}^{without}(L)$ (ohne Hintergrundwissen) für jede Klasse L. Die normalisierte Differenz der Varianzen berechnet man wie folgt:

$$\mathrm{vd}(L) := \frac{\mathrm{var}_{in}^{with}(L) - \mathrm{var}_{in}^{without}(L)}{\mathrm{var}_{in}^{without}(L)}. \tag{8.13}$$

Um zu ermitteln, ob und auf welche Klassen sich die veränderte Varianz auswirkt, berechnen wir mit der *individual inverse purity* (ivp) die Güte, mit der jede Klasse durch eine Clusterung gefunden wurde:

$$\mathrm{ipv}(L, \mathbb{P}) := \max_{P \in \mathbb{P}} \pi(L, P), \tag{8.14}$$

und vergleichen diese wieder für beide Repräsentationen:

$$\mathrm{ipd}(L) := \frac{\mathrm{ipv}^{with}(L, \mathbb{P}) - \mathrm{ipv}^{without}(L, \mathbb{P})}{\mathrm{ipv}^{without}(L, \mathbb{P})}. \tag{8.15}$$

Zum Vergleich nutzen wir die Repräsentationen mit (ipv^{with}) und ohne ($ipv^{without}$) Hintergrundwissen.

Vergleich ausgewählter Datensätze Um den Einfluss des Datensatzes auf das Clusterverfahren etwas besser zu verstehen, schauen wir uns zuerst die Verteilung des Datensatzes PRC-min15-max100 an. Wir erinnern uns, dass der PRC-min15-max100 eine recht homogene Verteilung der Dokumente über die Klassen besitzt (siehe Kapitel 2.1). Der PRC Datensatz ist hingegen sehr ungleichmäßig verteilt. Dies muss man bei der Bewertung der folgenden Abbildungen berücksichtigen.

Abbildung 8.11 gibt den Varianzvergleich gemäß Gleichung 8.13 und passend dazu den Vergleich der Clustergüte entlang der ipd (Gleichung 8.15) wieder. Die absteigende Kurve in Abbildung 8.11 zeigt die normalisierte Differenz der Innerklassenvarianzen zwischen den beiden Repräsentationen mit (Strategie: Hypdepth=5, hypint=add, hypdis=context, prune=30) und ohne Hintergrundwissen.

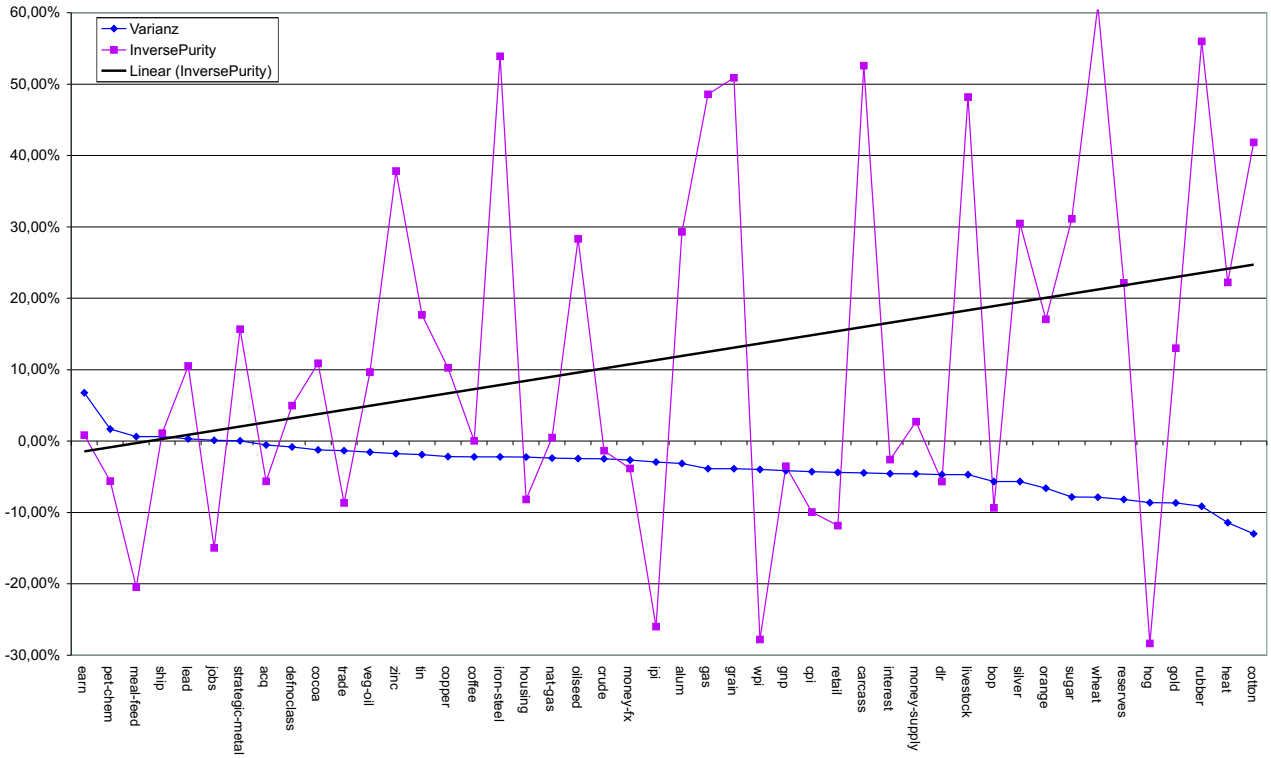

Abbildung 8.11: Vergleicht die Änderung der Varianz für jede Kategorie gegen die Änderung der Clustergüte bzgl. der individual inverse purity (vgl. Gleichung 8.15) am Datensatz PRC-min15-max100, wenn die Vorverarbeitungsstrategie von der besten Referenzclusterung zu einer guten Clusterung mit Hintergrundwissen wechselt (Strategie: Hypdepth=5, hypint=add, hypdis=context, prune=30) für k=60

Wie man erkennen kann, reduziert die Repräsentationsänderung für den größten Teil der Klassen die Varianz. Sieben Klassen erfahren hingegen eine kleine Steigerung der Varianz. Mit deutlichem Abstand steigt die Varianz der Klasse “earn” um 6.76 %. Die größte Reduktion der Varianz erfährt die Klasse “cotton” mit 12.98 %.

Die zweite Kurve zeigt die Steigerung der Clustergüte durch die unüberwachte Reduktion der Varianz für die meisten Klassen. Die schwarze Gerade gibt die lineare Interpolation der ipd-Werte wieder. Man erkennt die deutliche Steigung der Gerade. Größere ipd-Werte gehen einher mit einer größeren Reduktion der Varianz. Die Reduktion der Varianz macht es dem varinanzminimierenden Bi-Sec-KMeans leichter die Originalklassen zu finden. Die Veränderung der Repräsentation bewirkt bei vielen Klassen eine Varianzreduktion, die sich positiv auf die Clusterergebnisse auswirkt. Schauen wir uns das Ganze nun für den PRC-Datensatz an und vergleichen die Ergebnisse.

Abbildung 8.12 stellt wieder die Varianzdifferenzen und ipv-Differenzen (mit linearer Interpolation) diesmal für den PRC-Datensatz dar. Die Varianzreduktion fällt im Durchschnitt über alle Klassen bei beiden Datensätzen ähnlich hoch aus (PRC-min15-max100 = 3.56 % PRC = 3.87 %), was auch zu einer ähnlichen Steigerung der ipd-Werte führt. Trotzdem erhalten wir deutlich schlechtere Ergebnisse für den PRC-Datensatz in Kapitel 8.2.5 und 8.2.6. Berechnen wir den Mittelwert über die ipv-Werte gemäß Gleichung 8.14 so erhalten wir für den Datensatz PRC-min15-max100 ohne Hintergrundwissen $\overline{ipv^{without}} = 47.28\ \%$ und mit $\overline{ipv^{with}} = 52.05\ \%$. Bei PRC-Datensatz ergeben sich folgende Werte: $\overline{ipv^{without}} = 52.9\ \%$ und $\overline{ipv^{with}} = 60.17\ \%$. Vergleichen wir diese Werte mit den InversePurity-Ergebnissen aus den Tabellen 8.5 und 8.4, so stellen wir eine sehr kleine

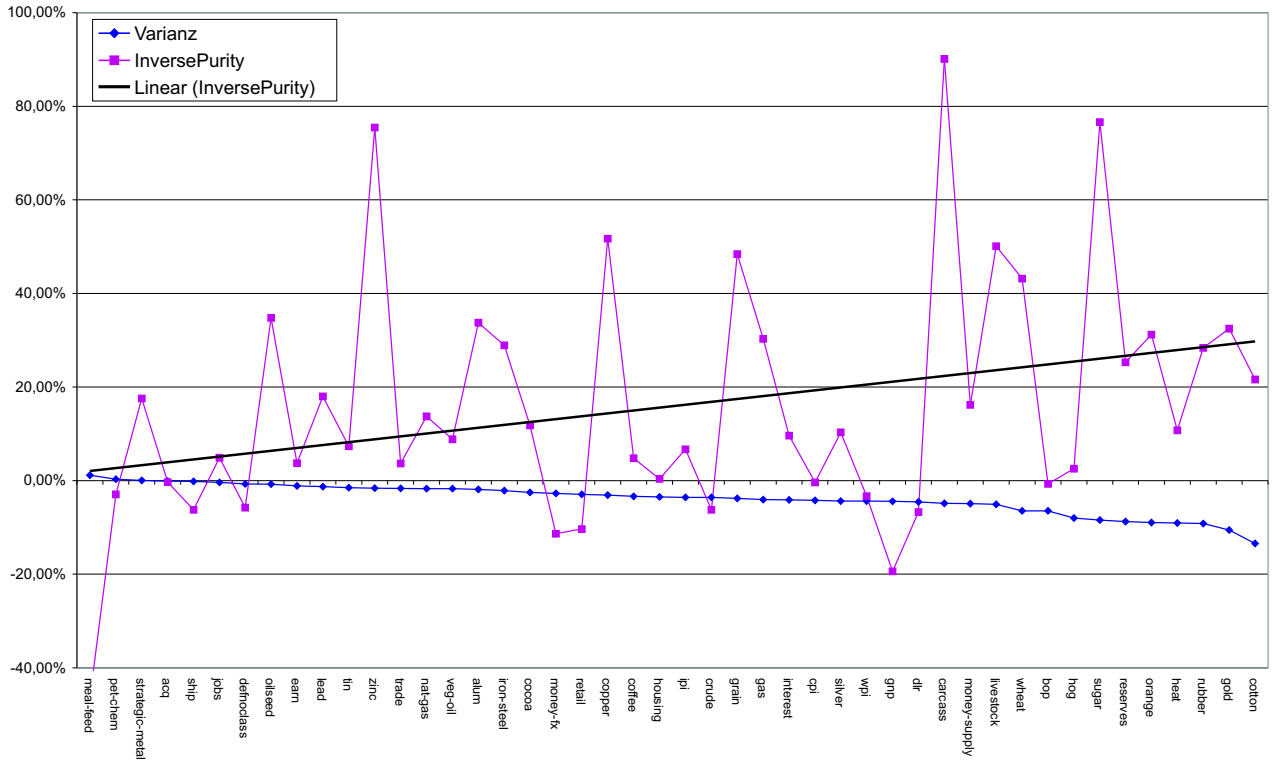

Abbildung 8.12: Vergleicht die Änderung der Varianz für jede Kategorie gegen die Änderung der Clustergüte bzgl. der individual inverse purity (vgl. Gleichung 8.15) am Datensatz PRC, wenn die Vorverarbeitungsstrategie von der besten Referenzclusterung zu einer guten Clusterung mit Hintergrundwissen wechselt (Strategie: Hypdepth=5, hypint=add, hypdis=context, prune=30) für k=60

Abweichung bei den Werten des PRC-min15-max100 Datensatzes fest und eine sehr große beim PRC-Datensatz. Wir erinnern uns an die Definition der InversePurity 5.15, die sich in der Berücksichtigung der Klassengröße vom Mittelwert des ipv aus Gleichung 8.14 unterscheidet. Da aber im PRC-Datensatz die Dokumente der “earn” oder “acq” Klasse sehr häufig vorkommen (sie stellen fast 50 % der Dokumente), diese aber leicht schlechter geclustert werden, wirkt sich die Veränderung der Repräsentation auf das gewichtete Mittel nur unwesentlich aus. Beim ungewichteten Mittel sieht man aber auch für den PRC-Datensatz eine deutliche Steigerung der Clustergüte.

Die Ursache für die schlechtere Clusterung der “earn” Klasse liegt an der Art des Textes. Er enthält in fast jedem Dokument dieser Klasse das Wort “vs.”. Da es nicht in WordNet vorkommt, können wir ihm keine semantische Bedeutung zuordnen. Dies führt per se zu einer sehr geringen Innerklassenvarianz, die durch die neue Repräsentationsänderung nicht weiter verbessert werden kann. Weiterhin enthalten die Texte der Klasse keine weiteren semantisch wichtigen Worte, was den Effekt noch verstärkt. Ein Blick in Anhang D.1 auf zwei Beispieltexte dieser Klasse macht die Problematik noch deutlicher. Im Gegensatz dazu lassen sich z.B. die Texte der Klasse “sugar”, die eine starke Varianzreduktion und eine deutlich bessere Clusterung erfahren, wesentlich besser in die semantische Repräsentation übersetzen (Beispiele siehe Anhang D.2).

Tabelle 8.6: Mittelwert der Purity für Clusterung des PRC-min15-max100 mit $k = 60$ Cluster, prune=30, tfidf-gewichtet, HYPDIS = context, HYPINT = add, HYPDEPTH = 5 (20 Wiederholungen)

Ontologie	ohne LSI	LSI 50	LSI 100	LSI 200
ohne	54,61 %	58,83 %	58,42 %	58,44 %
mit (context, add, 5)	60,16 %	61,54 %	61,98 %	61,28 %

8.4 Clustern mit LSI-Konzepten

Abschnitt 4.4 stellt die statistische Methode Latent Semantic Indexing zur Berechnung von LSI-Konzepten vor. Dieser Abschnitt vergleicht die Textclusterergebnisse basierend auf LSI-Konzepten mit den Ergebnissen auf der in dieser Arbeit entwickelten Ontologie-Repräsentation. Weiterhin werden wir beide Ansätze kombinieren und entsprechende Ergebnisse präsentieren. Vorteil des LSI-Ansatzes ist die geringe Dimensionalität des resultierenden Datensatzes. Leider sind die LSI-Konzepte nicht mehr interpretierbar (siehe Abschnitt 8.2.8). Dies ist der Vorteil des ontologiebasierten Ansatzes.

Die Evaluierung erfolgt für den PRC-min15-max100- und den PRC-Datensatz. Als Maßzahl verwenden wir die Purity. Alle Vergleichswerte beziehen wir aus den Tabellen 8.5 für den PRC-min15-max100 und 8.4 für den PRC-Datensatz.[7] Bei den ontologiebasierten Ansätzen wählen wir immer die beste Strategie als Vergleichsgrundlage. Die Ergebnisse werden für $k = 60$ Cluster, einem Prunethreshold von 30 und tfidf-gewichtet berechnet. LSI wird immer auf die reduzierte und gewichtete Matrix angewendet.

Die erste Zeile der Tabelle 8.6 gibt die Werte für den Vergleich der reinen termbasierten mit der LSI-basierten Bi-Sec-KMeans-Clusterung wieder. Dabei steht LSI 50, 100 und 200 für die Anzahl der berechneten Singulärwerte, die die Näherungsmatrix enthält. Man erhält bei Nutzung von LSI eine ca. vierprozentige Steigerung der Clustergüte unabhängig von der Anzahl der Singulärwerte (signifikant mit $\alpha = 0.5\ \%$).[8] Die Unterschiede zwischen den einzelnen LSI-Clusterungen sind nicht signifikant (Konfidenzintervall von $\alpha = 0.5\ \%$). Die beste ontologiebasierte Clusterung mit der Strategie HYPDIS = context, HYPINT = add, HYPDEPTH = 5 ist mit 60,2 % um weitere signifikante zwei Prozent besser. Berechnet man die LSI-Konzepte für die ontologiebasierte Dokumentrepräsenation, so erfolgt eine weitere knapp zweiprozentige signifikante Steigerung der Clustergüte. Auch hier wurden keine signifikanten Unterschiede zwischen den LSI-Clusterungen entdeckt.

Tabelle 8.7 gibt die Vergleichsergebnisse für den PRC-Datensatz wieder. Auch hier haben wir erst den Vergleich auf der termbasierten Repräsentation durchgeführt. Die Anwendung von LSI führte in diesem Fall zu keiner Verbesserung der Ergebnisse. Bei LSI 50 konnte eine leichte Verschlechterung des Ergebnisses beobachtet werden (der Unterschied ist aber nicht signifikant). Berechnet man für die beste Strategie (in diesem Fall ist das nicht HYPDepth = 5 sondern 0) wieder die LSI-Konzepte, so kann man auch hier keine signifikanten Verbesserungen der Clusterergebnisse beobachten. Vielmehr fällt auf, dass die Purity für LSI 50 wieder leicht schlechter ist. Der Vollständigkeit halber haben wir auch noch einmal die Ergebnisse für die beste Strategie bei den PRC-min15-max100 in der letzten Zeile der Tabelle 8.7 für den PRC-Datensatz wiedergeben. Die Ergebnisse liegen wie in den Fällen davor für LSI auf dem Niveau der Ergebnisse ohne LSI. Das heißt in diesem Fall, dass sie leicht schlechter sind als ohne Ontologierepräsentation.

Im Ergebnis dieses Versuchs zeigen LSI-Konzepte und Ontologie-Konzepte ein ähnliches Verhal-

[7] Auch die Stichprobe der Texte für den PRC-min15-max100 Datensatz ist die selbe.

[8] Test erfolgt mit dem Students t-Test, vgl. [160] oder [169]

Tabelle 8.7: Mittelwert der Purity für Clusterung des PRC mit $k = 60$ Cluster, prune=30, tfidf-gewichtet, HYPDIS = context, HYPINT = add (20 Wiederholungen)

Ontologie	ohne LSI	LSI 50	LSI 100	LSI 200	LSI 300
ohne	75,10 %	74,88 %	75,10 %	75,09 %	75,51 %
mit (HYPDEPTH = 0)	75,50 %	74,85 %	75,33 %	75,26 %	75,29 %
mit (HYPDEPTH = 5)	74,60 %	74,39 %	74,60 %	74,36 %	74,63 %

ten bezüglich der Clustergüte von Textdokumenten. So führen beide Repräsentationen beim PRC-min15-max100-Datensatz zu einer Steigerung der Ergebnisse, wobei der ontologiebasierte Ansatz leicht besser ist. Die Kombination beider Ansätze liefert nochmals bessere Ergebnisse. Für den PRC Datensatz konnte keiner der Ansätze bessere Ergebnisse als die Referenzclusterung liefern. Auch die Kombination war hier nicht besser. Erste Untersuchungen auf der Basis des Java-Datensatzes (vgl. Abschnitt 2.2) führten zu ähnlichen Ergebnissen wie für den PRC-min15-max100 Datensatz.

Die Ergebnisse der Kombination von LSI und ontologiebasiertem Ansatz zur Dokumentrepräsentation sind sehr vielversprechend, da sie auf eine weitere Steigerung der Clustergüte zeigen. Eine umfangreichere Studie liegt außerhalb des Rahmens dieser Arbeit.

Der folgenden Abschnitt beschäftigt sich mit der Einsatzfähigkeit der Formalen Begriffsanalyse zum Clustern von Textdokumenten. Der Vorteil liegt in den vorhandenen Visualisierungstechniken und deren leicht zu verstehenden Ergebnisse.

8.5 Konzeptuelles Clustern von Texten mit Formaler Begriffsanalyse

Die Formale Begriffsanalyse (siehe Abschnitt 5.5) bietet mit ihren Visualisierungstechniken intuitiv verständliche Clusterergebnisse. Daher liegt die Idee nahe, für leicht verständliche Clusterergebnisse aus dem Bereich des Textclusterns, die Formale Begriffsanalyse zu verwenden. Wir werden in diesem Abschnitt zeigen, dass die Formale Begriffsanalyse in der Lage ist, Textcluster zu berechnen. Die Analyse der berechneten Cluster auf der Basis der visualisierten Verbände wird uns die Grenzen dieser Methode zeigen. Gleichzeitig entwickeln wir Wege zur Überwindung der Grenzen durch die Kombination der Formalen Begriffsanalyse mit Ontologien oder durch die Reduktion der Komplexität der Verbände mittels Clusterverfahren wie KMeans. Wir nehmen hier bewusst den Standpunkt der Formalen Begriffsanalyse als Technik zum Clustern von Objekten und im Speziellen von Textdokumenten ein und wollen aus diesem Blickwinkel die Ergebnisse betrachten. Wir sehen in diesem Abschnitt alle zusätzlichen Schritte als Vorverarbeitung zur besseren Berechnung von FBA-Clustern. Die Anwendung der Formalen Begriffsanalyse als Analyse und Visualisierungstechnik von z.B. KMeans-Textclustern steht erst in Abschnitt 9.3 im Vordergrund der Betrachtungen.

Der folgende Abschnitt wird am Beispiel einer wort- bzw. termbasierten Repräsentation das Vorgehen sowie erste Ergebnisse der Anwendung von Formaler Begriffsanalyse auf Textdokumente vorstellen. In Abschnitt 8.5.2 analysieren wir den Einsatz von Ontologien als Basis einer veränderten Repräsentation, bevor wir Textcluster auf einem reduzierten Gegenstandsraum in Abschnitt 8.5.3 für den Einsatz der Formalen Begriffsanalyse diskutieren. Wir beenden diesen Abschnitt mit einem Blick auf verwandte Ansätze.

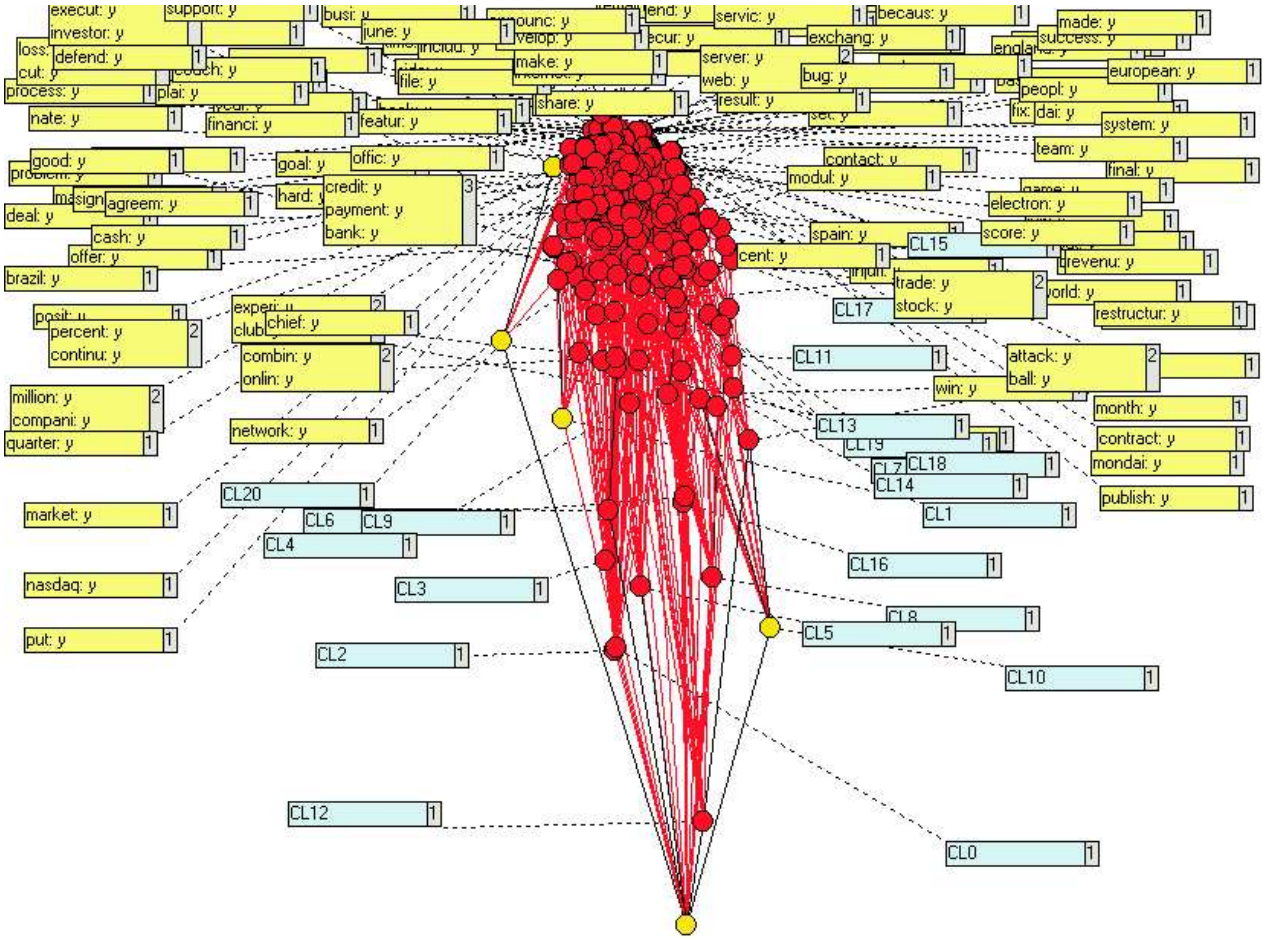

Abbildung 8.13: Begriffsverband für 21 Textdokumenten und 117 Terme (TV1)

8.5.1 FBA-Clustern auf einer Wortrepräsentation

In diesem Abschnitt wollen wir anhand des Datensatzes DS1 aus Abschnitt 5.5 die Einsatzmöglichkeiten der Formalen Begriffsanalyse auf einer termbasierten Repräsentation zum Clustern von Texten diskutieren. Ein sehr übersichtlicher Begriffsverband des Einführungsbeispiels mit zehn Clustern als Gegenstände und acht Termen (Wortstämme) als Merkmale ist in Abbildung 5.4 abgebildet. Die Abbildung ist übersichtlich, leicht verständlich und für einen Experten einfach zu interpretieren.

Abbildung 8.13 visualisiert den Verband für alle 117 Merkmale und 21 Gegenstände, wobei jedes Merkmal ein Term t und jedes Dokument d ein Gegenstand ist. Diesen Begriffsverband nennen wir im Folgenden TV1. Er basiert auf dem Kontext $\mathbb{K}_{TV1} := (G, M, I)$ mit $G := D, M := T$ und $(d, t) \in I$, wenn $(\vec{t_d})_t \geq \theta$ ist[9] (siehe Abbildung C.1).[10] Die Dokumente werden an dieser Stelle nicht zu Cluster zusammengefasst, sondern jedes Dokument wird als Gegenstand betrachtet. Auch die Merkmalsmenge wird nicht weiter eingeschränkt (durch z.B. eine manuelle Auswahl). Der Kontext dieses Verbandes wird aus dem “Bag of Words”-Modell abgeleitet. Die abgeleiteten Termvektoren $\vec{t_d}$ der Dokumente werden mit tfidf gewichtet und auf die Länge eins normiert. Anschließend erfolgt die Umwandlung in den Kontext nach der in Abschnitt 4.5.2 vorgestellten Diskretisierungsmethode. Der Schwellwert θ beträgt für TV1 $10\ \%$. Damit erhält man einen Kontext bestehend aus 21 Textdokumenten (Gegenstände) und 117 Termen (Merkmale).

[9] Abschnitt 4.5.2 beschreibt ausführlich die Reduktion des Termvektors für die Relation I eines Kontextes.
[10] Die weiteren term- und konzeptbasierten Kontexte können analog angegeben werden.

Eine Visualisierung des TV1 findet man in Abbildung 8.13.[11] Ziel der Berechnung und Visualisierung eines Verbandes ist vor allen Dingen die Unterstützung des Anwenders bei der explorativen Analyse der Texte und dem Finden und Verstehen von (konzeptuellen) Clustern — in diesem Fall von Textdokumentclustern. Jeder formale Begriff stellt ein Textdokumentcluster dar. Der Umfang, d.h. die Menge der Gegenstände eines formalen Begriffes sind die Elemente, d.h. die Dokumente eines Clusters, und der Inhalt des formalen Begriffes sind die beschreibenden Merkmale, d.h. die Terme bzw. Wortstämme. Der Verband in Abbildung 8.13 enthält eine große Anzahl an formalen Begriffen mit vielen Beziehungen zwischen diesen. Dies führt zu einer komplexen Struktur des Verbandes, die keineswegs leicht zu verstehen ist. Es sind zu viele Cluster und Beziehungen für eine übersichtliche Visualisierung.

Die Formale Begriffsanalyse erlaubt die Berechnung von Textclustern. Ohne weitere Hilfsmittel gehen die Vorteile der übersichtlichen und verständlichen Visualisierung der Verbände durch die hohe Clusteranzahl und die vielen Beziehungen zwischen den Clustern verloren. Im Folgenden ist daher das Ziel die Berechnung einer *überschaubaren Menge an Clustern* mit der Formalen Begriffsanalyse, deren *Beschreibung mit wenigen und aussagekräftigen Termen* und *leicht nachvollziehbare Beziehungen* zwischen den Clustern. Ein Beispiel für einen solchen Verband liefert Abbildung 5.4. Die folgenden Ansätze erreichen mit unterschiedlichen Methoden dieses Ziel. Sie versuchen auf der einen Seite nur leicht verständliche Teilverbände zu visualisieren oder auf der anderen Seite durch Vorverarbeitung die Gegenstands- oder Merkmalsmenge zu reduzieren. Dieser Abschnitt diskutiert die Visualisierung von Teilverbänden. Weiterhin wird ein Beispiel mit einer manuell reduzierten Merkmalsmenge und zwei Beispiele mit unterschiedlichem Schwellwert (der Schwellwert hat Einfluss auf die Merkmalsmenge) vorgestellt. Weitere Ansätze zur Veränderung der Gegenstands- und Merkmalsmengen werden dann jeweils in den Abschnitten 8.5.2 und 8.5.3 eingeführt.

Teilverbände visualisieren: Die Software stellt direkt Mittel zum Hervorheben und Visualisieren von Teilverbänden zur Verfügung. Aus technischen Gründen erfolgt die Visualisierung der hervorgehobenen Teilverbände in den Abbildungen 8.14 und 8.15 mit einem gedrehten Verband (siehe Abschnitt 5.5.3).

Auf der Suche nach Textclustern, die mehrere Dokumente umfassen, untersucht man im ersten Schritt die allgemeinen Begriffe und lässt sich diese und alle Unterbegriffe in der Visualisierung hervorheben. Allgemeine formale Begriffe findet man in der Visualisierung unten. Sie sind direkt mit dem Top-Begriff verbunden (Der Top-Begriff ist der Begriff, der in dieser Visualisierung am weitesten unten liegt). Man sieht an der Anzahl der vom Top-Begriff abgehenden Kanten, dass sehr viele solche Begriffe existieren. Der Begriff mit der Bezeichnung “cup” hat sechs Dokumente im Umfang und den Term “cup” im Inhalt. Abbildung 8.14 zeigt den gesamten Verband und den hervorgehobenen Teilverband, der durch “cup” erzeugt wird. Fasst dieser Begriff mehrere Dokumente zum gleichen Thema zusammen, so stellt er einen guten Cluster da. Die Themen der Dokumente sind durch eine manuelle Analyse bekannt. Die sechs gewählten Dokumente stammen alle aus dem Bereich Fußball (CL9 fehlt). Das erste Ziel, einen Cluster mit Dokumenten vom gleichen Thema zu finden, haben wir erreicht. Bei der weiteren Analyse finden wir die Begriffe erzeugt durch “player” und “game”. Auch sie haben sechs der sieben Fußballdokumente im Umfang (CL7 fehlt bei player und CL11 bei game).

Es ist also möglich, formale Begriffe, d.h. Cluster zu finden, die Dokumente zum gleichen Thema im Umfang haben. Durch die große Anzahl an solchen formalen Begriffen, finden wir auch eine

[11] Die Berechnung der Visualisierung erfolgte mit der Software Cernato der NaviCon AG. Die Texte haben die Namen CL0-CL20. Texte CL0-CL6 sind über Finanzen, Texte CL7-CL13 über Fußball und Texte CL14-CL20 über Software.

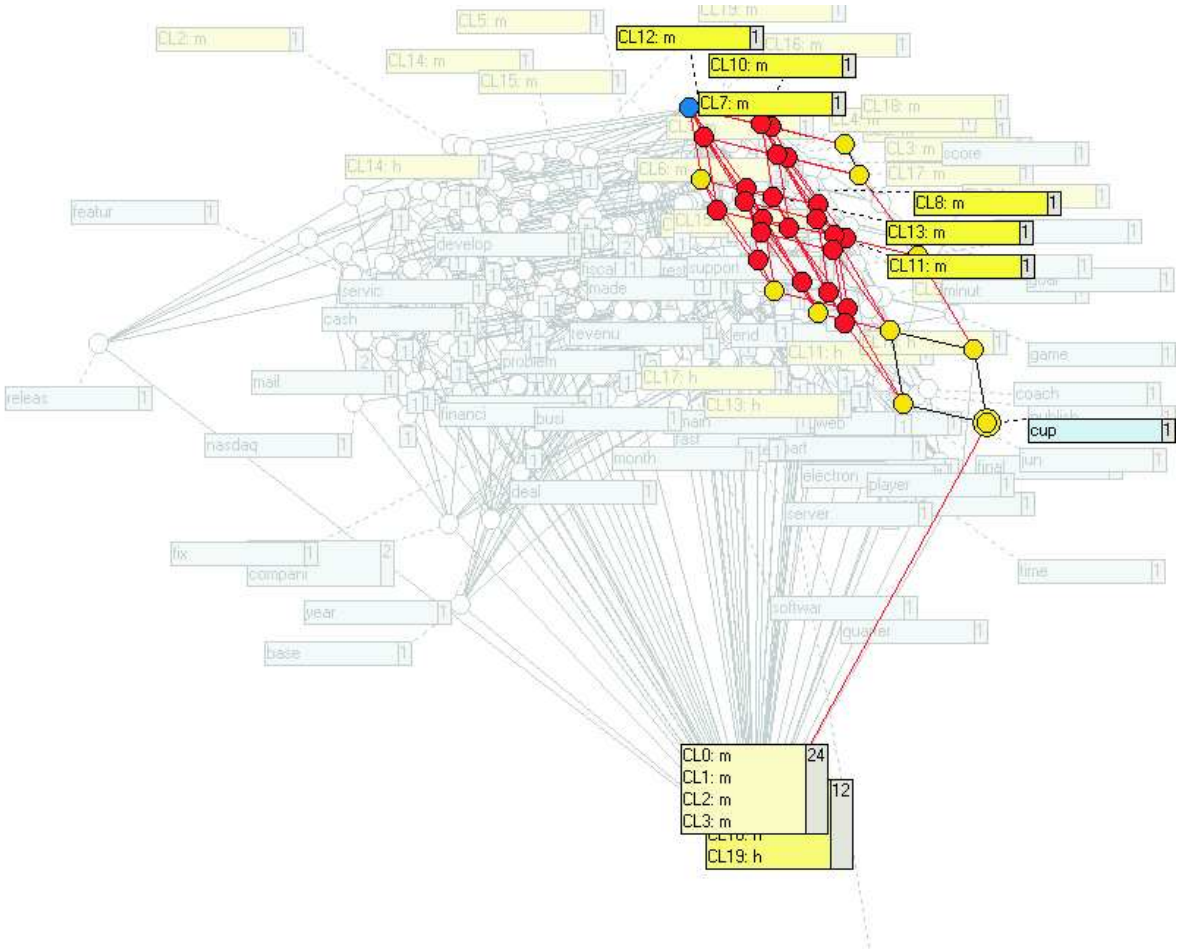

Abbildung 8.14: stellt den Begriffsverband TV1 mit dem hervorgehobenen Teilverband, erzeugt von “cup”, dar

große Anzahl an formalen Begriffen, die Dokumente zu unterschiedlichen Themen enthalten. Zur Bestimmung des Clusterthemas brauchen wir eine gute und leicht verständliche Beschreibung. Nur dann sind wir in der Lage, die formalen Begriffe zu einem Thema zu bestimmen. Dazu sollten im Idealfall alle Dokumente zu einem Thema unter einem formalen Begriff mit einem allgemeinen Term wie z.B. “Fußball” für unser Beispiel zusammengefasst werden. Anschließend sollten Unterbegriffe die Dokumente in weitere Cluster zu unterschiedlichen Themen z.B. betreffend Fußball teilen. Auch hier sind wieder aussagekräftige Terme für das Verständnis wichtig. Die Anzahl der Untercluster sollte nicht zu groß sein, d.h. der Verband sollte für eine einfache Exploration wenige formale Begriffe und nicht zu viele Beziehungen enthalten. Wir überprüfen unsere Forderungen anhand des Verbandes, der nur die Fußballdokumente als Gegenstände enthält, für die wir schon allgemeine Begriffe im Gesamtverband entdeckt haben (siehe Abbildung 8.15). Keine der Forderungen ist erfüllt, so dass es sowohl schwierig ist, das Thema der Dokumente zu bestimmen, als auch die Untercluster und deren Themen. Vielmehr sind viele formale Begriffe abgebildet. Damit wird es auch schwierig, formale Begriffe mit Hilfe der Visualisierung von Teilverbänden für das Textclustern zu verwenden, da die relevanten Cluster in der Menge aller Cluster im Gesamtverband nur schwer identifiziert werden können.

Eine Ursache für die vielen Begriffe und Beziehungen sind synonyme Terme. Synonyme Terme wie z.B. “Ball”, “Fußball” oder “Leder” im Kontext eines Dokumentes über Fußball transportieren den gleichen Inhalt. Im Begriffsverband führen sie zu einem eigenen Begriff oder Teilverband. Dies

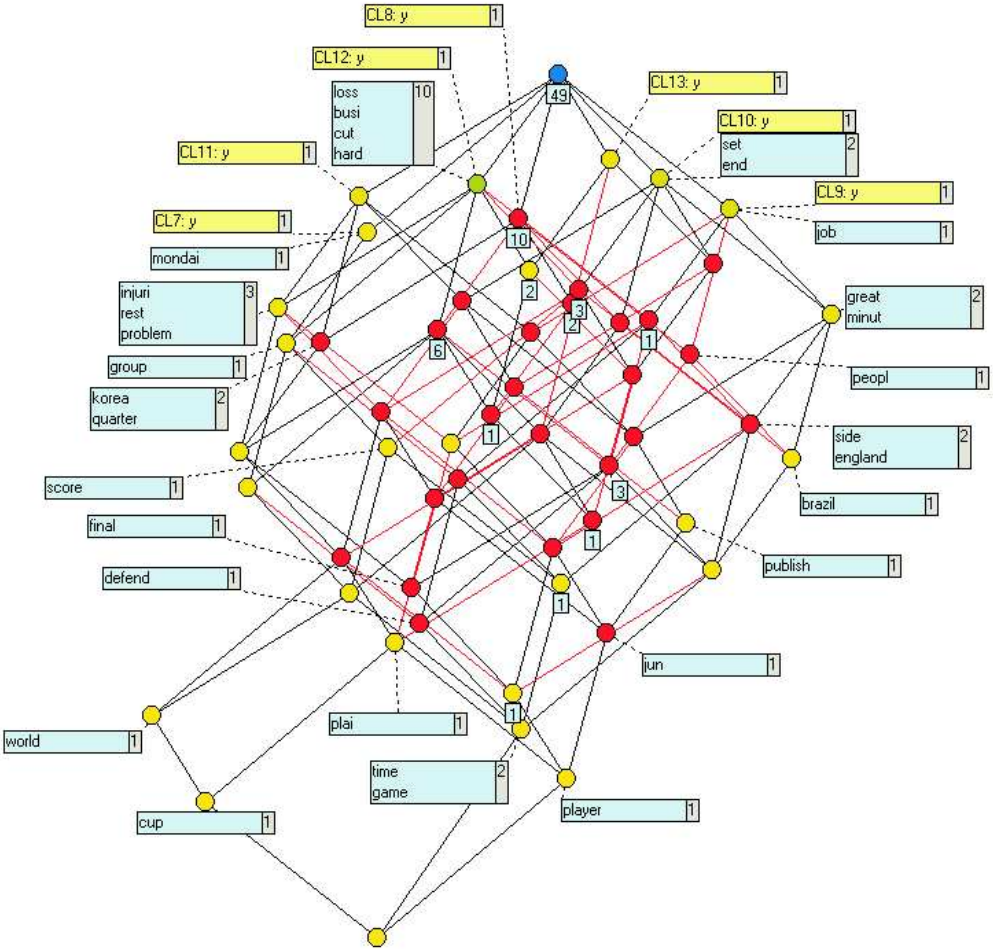

Abbildung 8.15: gibt den von den Dokumenten (über Fußball) CL6-CL13 erzeugte Teilverband von TV1 wieder

ist zwar korrekt aber nicht hilfreich beim Bestimmen der Beziehungen zwischen den einzelnen Dokumenten auf der Basis der verwendeten Terme/Worte. Ein einzelner aussagekräftiger Term wäre hier von Vorteil.

Eine weitere Ursache sind fehlende allgemeine Terme bzw. Terme, die in jedem Dokument einer Klasse vorkommen. Zum Beispiel kommt der Term “cup” nur in sechs Dokumenten vor. In keinem der sieben Dokumente CL6 bis CL13 kommt das Wort Fußball vor. Aus diesem Grund ist es auch nicht möglich einen formalen Begriff mit dieser Bezeichnung abzuleiten. Dafür finden wir Terme wie “job”, “side” oder “score”, die durchaus mit Fußball in Verbindung gebracht werden können. Sie haben aber nicht offensichtlich etwas damit zu tun. Diese detailreiche Betrachtung der Zusammenhänge (im Sinne der vielen verwendeten Terme) durch die Begriffsanalyse ist für das Verständnis des Verbandes nachteilig. Eine “unscharfe” Betrachtung, sprich auf einem abstrakteren bzw. allgemeineren Niveau, wäre an dieser Stelle von Vorteil. Hintergrundwissen kann solche Informationen liefern. Einen solchen Ansatz betrachten wir in Abschnitt 8.5.2.

Im Folgenden wollen wir untersuchen, ob wir mit den Termen nicht doch einen verständlichen Begriffsverband ableiten können. Dazu reduzieren wir in einem ersten Schritt die Merkmalsanzahl manuell. Für das laufende Beispiel lesen wir dazu die Texte des DS1-Datensatz und versuchen mar-

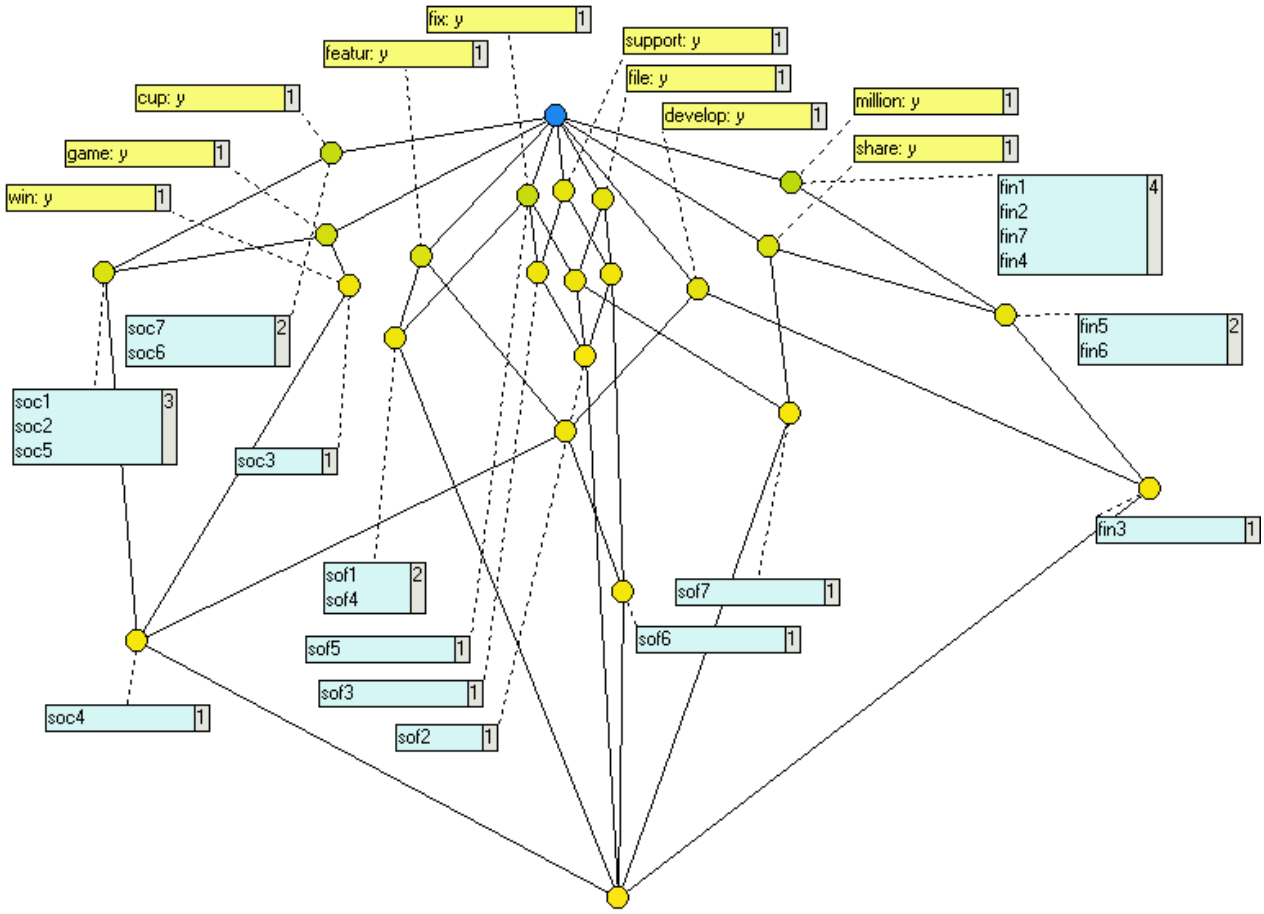

Abbildung 8.16: Begriffsverband mit manuell ausgewählten Termen, so dass sich die gegebenen Klassen in den konzeptuellen Clustern wiederfinden lassen (TV2)

kante Terme auszuwählen. In einem zweiten Schritt erfolgt die Auswahl auf der Basis des Schwellwertes, wobei wir hoffen, dass auch bei dieser Methode die aussagekräftigen Terme zum Erzeugen des Verbandes über dem Schwellwert liegen.

Manuelle Merkmalsauswahl: Die manuelle Auswahl von Merkmalen ist sehr zeitaufwendig, erlaubt aber eine sehr effektive Auswahl der Merkmale. Für einen verständlichen Verband müssen die Merkmale idealerweise die Eigenschaft besitzen, dass sie in allen Dokumenten einer gegebenen Klasse vorkommen und in keinem oder möglichst wenigen Dokumenten einer anderen Klasse. Auch kann man bei der manuellen Auswahl ein paar wenige für den Anwender leicht verständliche Terme auswählen und muss nicht alle vorhandenen Terme nutzen. Für die Abbildung 8.16 wurde das beschriebene Vorgehen bei der Wahl der Merkmale berücksichtigt. Da für die allgemeine Anwendung der Formale Begriffsanalyse zum Textclustern die manuelle Merkmalsauswahl wegen des hohen Aufwandes nicht sinnvoll ist, stellt sie an dieser Stelle nur eine Referenz für einen guten und verständlichen Begriffsverband dar. Alle weiteren Methoden versuchen ähnlich verständliche und übersichtliche Verbände abzuleiten.

Das Ergebnis dieser gezielten Auswahl spiegelt sich in einem deutlich übersichtlicheren und verständlicheren Verband wider. Auf der linken Seite gruppieren sich die Fußballtexte, auf der rechten Seite die finanzwirtschaftlichen Texte und in der Mitte die Texte, in denen es um Software geht. Auch der Inhalt der Dokumente lässt sich anhand der verwendeten Worte gut erfassen. Querbezie-

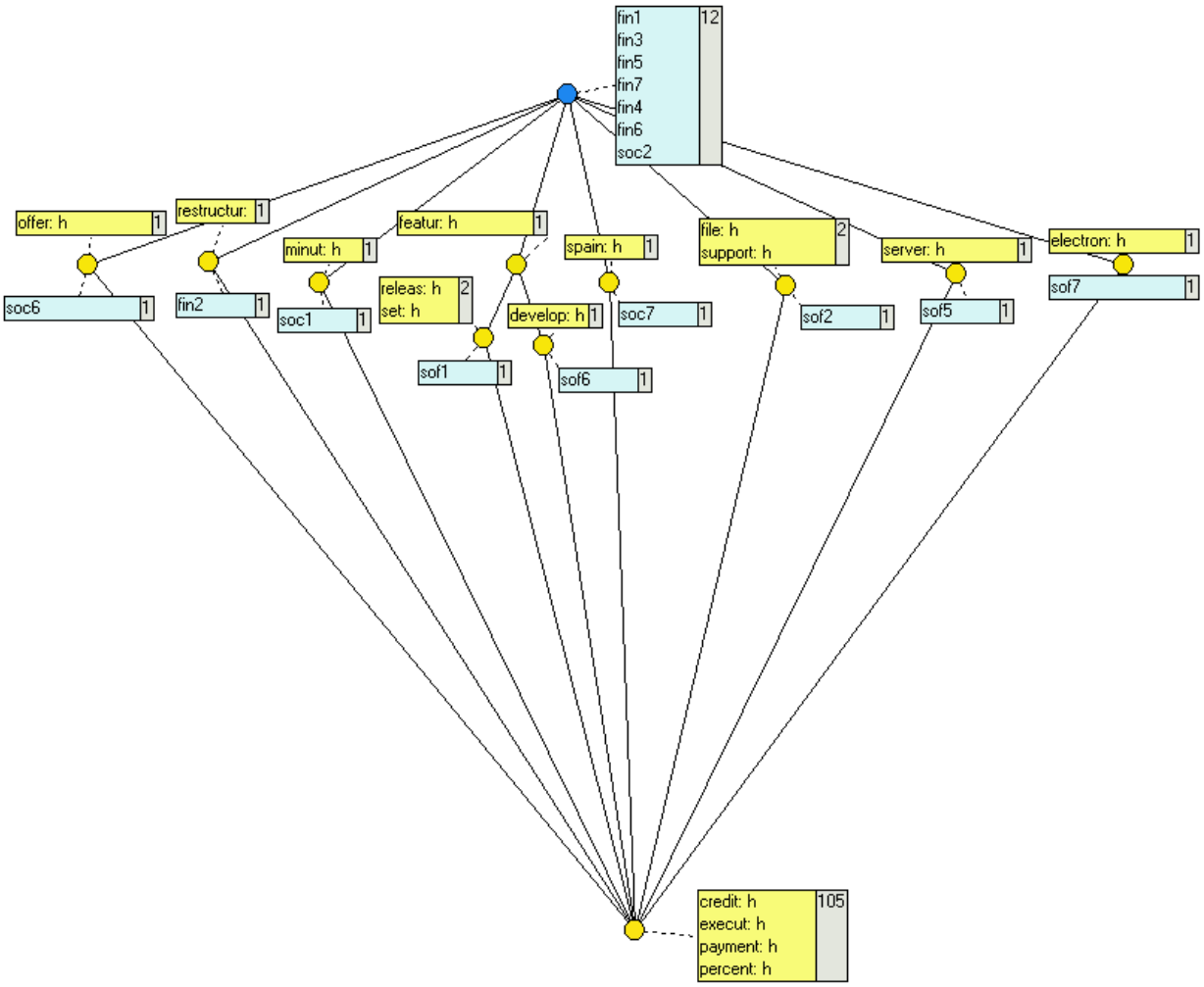

Abbildung 8.17: Begriffsverband mit $\theta = 80\ \%$ (TV3)

hungen zwischen den Klassen wie z.B. bei Dokument “Soc4” und “Sof6” mit “develop” (man kann sowohl Software entwickeln als auch Strategien im Fußball) sind durchaus nachvollziehbar.

Die manuell ausgewählten Merkmale zeigen, dass das konzeptuelle Clustern von Textdokumenten zu guten Ergebnissen führen kann. Man benötigt nun eine effektive Strategie für die Merkmalsauswahl. Im Folgenden untersuchen wir die Möglichkeit, die manuelle Merkmalsauswahl durch die Wahl unterschiedlicher Schwellwerte, die eine unterschiedliche Merkmalsanzahl bewirken, zu ersetzen.

Merkmalsauswahl per Schwellwert: Der Schwellwert θ bei der Berechnung der Merkmale für den Kontext hat Einfluss auf die Anzahl der Merkmale. Ein hoher Schwellwert bewirkt eine geringere Merkmalsanzahl, ein niedriger eine hohe Anzahl. Ein naheliegender Ansatz zur Merkmalsauswahl (Reduktion) besteht in der Steigerung des Schwellwertes. Abbildung 8.17 zeigt den Verband mit einem Schwellwert von $\theta = 80\ \%$. Zwölf der 21 Dokumente sind im Umfang des Top-Begriffes und haben keinen Term, dessen Gewicht über dem Schwellwert liegt. Die Aussagekraft des Verbandes wird nicht nur durch die nicht beschriebenen zwölf Dokumente, sondern auch durch die sehr wenigen Beziehungen zwischen den restlichen neun Dokumenten stark reduziert. Auch die Terme der neun beschriebenen Dokumente sind nicht eindeutig. “offer” oder “restructur” muss man nicht unbedingt mit den Themen Fußball bzw. Finanzwirtschaft verbinden. “feature”, “releas” oder

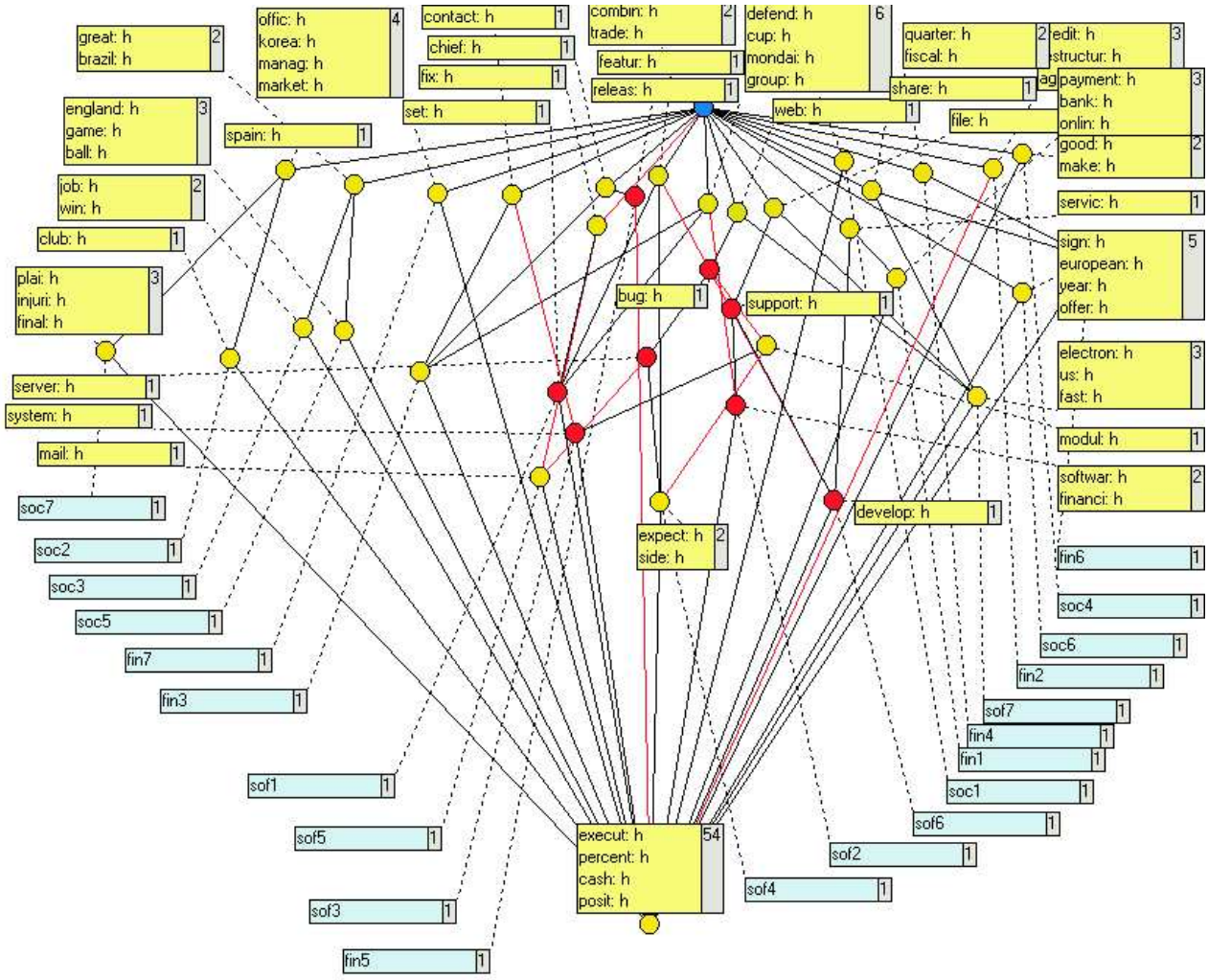

Abbildung 8.18: Begriffsverband mit $\theta = 45\ \%$ (TV4)

"server" dagegen schon eher mit Software. Wir haben zwar das eine Ziel, die Anzahl der formalen Begriffe des Verbandes zu reduzieren erreicht, dies aber auf Kosten der Verständlichkeit. Auch fehlen die themenbezogenen Cluster, die z.B. alle Fußballdokumente zusammenfassen. Der Zielverband sollte mehr formale Begriffe enthalten, d.h. wir benötigen eine größere Merkmalsmenge.

Betrachten wir abschließend den Fall eines mittleren Schwellwertes. Bei der Reduktion des Schwellwertes können die Dokumente, die im letzten Kontext kein Merkmal hatten, wieder durch Merkmale im Kontext beschrieben werden. Abbildung 8.18 gibt den Verband für den Fall wieder, wenn jedes der Dokumente mindestens einen Term im Inhalt hat, d.h. der Top-Begriff keinen Gegenstand enthält. Dies ist bei einem Schwellwert von ca. 45 % für θ der Fall.

Bei der Analyse des Verbandes der Abbildung 8.18, aber auch der Abbildungen 8.17 und 8.16, erkennt man eine deutlich höhere Anzahl an Beziehungen zwischen den formalen Begriffen, die die Software-Dokumente repräsentieren, gegenüber den Begriffen der anderen beiden Originalklassen. Fünf der neun Dokumente, die durch mindestens einen Term in Abbildung 8.16 beschrieben werden, sind aus der Klasse Software. Die Terme der Software-Dokumente erhalten höhere Gewichte als die meisten Terme der anderen Klassen. Das führt auch bei höherem Schwellwert zu mehr Termen und so zu mehr formalen Begriffen und Beziehungen zwischen diesen, also einem komplexeren Verband. Damit stehen weder für die Softwareklasse, hier gibt es zu viele Beziehungen, noch für die beiden anderen Klassen, hier gibt es zu wenige Beziehungen, aussagekräftige Begriffe zur Verfügung, die alle Dokumente einer Klasse im Umfang haben und klar beschreiben.

Beim Vergleich von Abbildung 8.18 mit 8.17 fällt außerdem die stark gestiegene Anzahl an formalen Begriffen auf. Es wurden nicht nur wichtige Terme einer Klasse hinzugefügt, sondern auch viele Terme, die Verbindungen zu Dokumenten aus anderen Klassen schaffen. Dies führt aber zu vielen Begriffen und zu einem schwer verständlichen Verband. Das Ziel, die Dokumente anhand automatisch ausgewählter Merkmale zu beschreiben, wurde nicht erreicht. Die Textcluster und deren Abhängigkeiten, die dem Verband der Abbildung 8.18 entnommen werden können, entsprechen nicht den Erwartungen eines einfachen und leicht zu interpretierenden Verbandes, die sich aus dem Verband aus Abbildung 8.16 und dessen Herleitung ergeben. Weder kann man die drei gegebenen Klassen entdecken, noch bekommt man eine adäquate Beschreibung. Auch sind nur wenige formale Begriffe vorhanden, die Dokumente gleicher Originalklassen enthalten.

Das Ziel, mit diesem einfachen schwellwertbasierten Ansatz die Worte auszuwählen, die die Dokumente einer Originalklasse beschreiben und in Beziehung zueinander setzen, wie dies in Abbildung 8.16 bei der manuellen Auswahl gezeigt werden konnte, kann der Ansatz nicht erfüllen. Wir werden daher im Folgenden aus Sicht der Formalen Begriffsanalyse weitere Vorverarbeitungsschritte durchführen, um die zur Verfügung stehenden Merkmale so zu verändern, dass ein Clustern mittels der Formalen Begriffsanalyse ermöglicht wird und so auch leicht verständliche Verbände entstehen.

Zusammenfassung: Die direkte Anwendung der Formalen Begriffsanalyse auf Textdokumente führt zu einem unübersichtlichen und schwer verständlichen Verband und unterstützt so den Anwender nicht sehr gut bei der explorativen Analyse von Textdokumenten. Folgende Gründe wurden herausgearbeitet: Die Verbindung von jedem Dokument zu jedem anderen anhand von Worten/Wortstämmen als dokumentbeschreibende Terme erzeugt einen sehr detaillierten Verband mit vielen Querbeziehungen zwischen Dokumenten auch unterschiedlicher Originalgruppen. Außerdem ist die Gewichtung und Auswahl der Merkmale sehr entscheidend. Merkmale, die eine Gruppe gut beschreiben, aber auch Dokumente anderer Gruppen, helfen bei der Diskriminierung nicht und führen nicht zu den gewünschten Textclustern. Eine Balance von beschreibenden und diskriminierenden Merkmalen wäre wünschenswert. Auch fehlen allgemeine Merkmale, die den Inhalt einer Klasse zusammenfassen, da die Terme nicht im Text vorkommen. Mit solchen Merkmalen wäre die Abstraktion von zu vielen Details möglich. Das würde die Lesbarkeit und Verständlichkeit des Verbandes steigern. Die einfache Methode der Adaption des Schwellwertes zur Begrenzung der Merkmalsanzahl war nicht erfolgreich.

Im folgenden Abschnitt 8.5.2 fügen wir mit Hilfe einer Ontologie allgemeinere Terme in die Repräsentation ein. Wir wollen untersuchen, ob diese Terme die Verständlichkeit des Verbandes steigern. Auch möchten wir wissen, ob formale Begriffe vorkommen, die alle Dokumente einer Originalklasse enthalten.

8.5.2 FBA auf einer Konzeptrepräsentation

Abschnitt 8.5.1 hat gezeigt, dass die termbasierte Dokumentrepräsentation Schwierigkeiten bei der Abstraktion und Generalisierung hat. Die Konzepte einer Ontologie nach Definition 8 sowie die die taxonomischen Beziehungen bieten sich für eine abstraktere Repräsentation an. Im Folgenden wurde mit Methoden des Ontology Learnings (siehe [153]) eine Ontologie Names SO1 (siehe Abbildung 8.19) für die Texte des Datensatzes DS1 modelliert. Im Folgenden werden wir erst die Ontologie einführen und dann untersuchen, ob eine konzeptbasierte Repräsentation der Dokumente zu einer verbesserten Clusterung durch die Formale Begriffsanalyse führt.

Die Ontologie SO1 in Abbildung 8.19 besteht nur aus einem Lexikon, den Konzepten und ei-

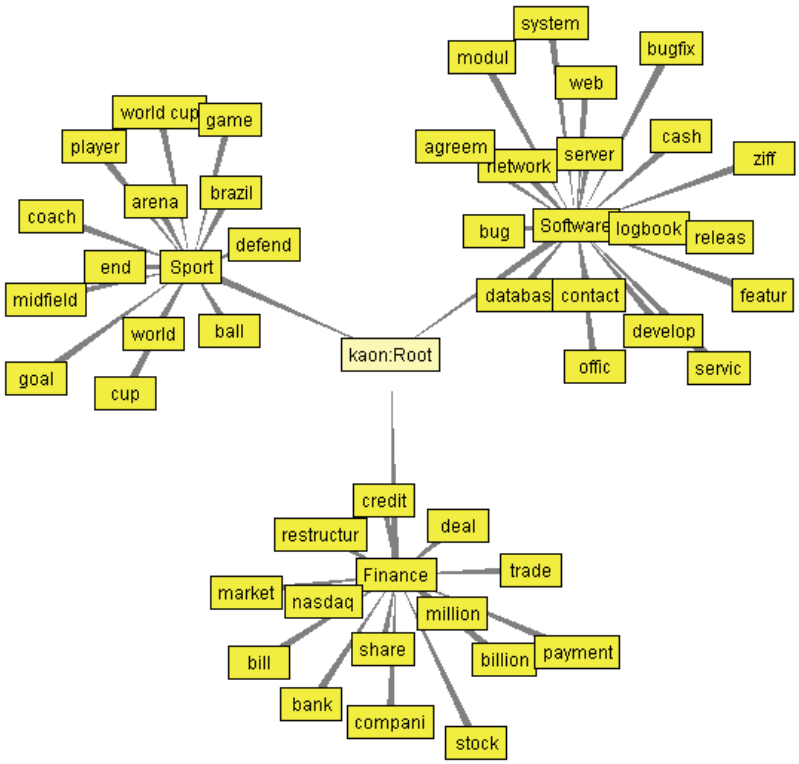

Abbildung 8.19: Beispielontologie passend zum Datensatz DS1 in Kapitel 5.5.1

ner Taxonomie. Die taxonomische Beziehung zwischen den Blattkonzepten und den dazugehörigen Oberkonzepten ist nicht immer eine “isa” Beziehung. Das allgemeinste Konzept ist das KAON:ROOT-Konzept. Jede Kante von diesem Konzept führt zu einem Unterkonzept. Ausgehende Kanten von diesen Unterkonzepten führen wieder zu deren Unterkonzepten usw., d.h. SERVER ist ein Unterkonzept von SOFTWARE und SOFTWARE ist ein Unterkonzept von KAON:ROOT.

Bei der Modellierung der taxonomischen Beziehung wurden Konzepte, die vorrangig in Dokumenten einer Klasse vorkommen, in der Taxonomie auch dem passenden Oberkonzept zugeordnet, z.B. CUP dem Konzept SPORT. Wir erhoffen uns von diesem Vorgehen bei der Modellierung eine veränderte Verbandsstruktur, die zu formalen Begriffen führt, die Dokumente einer Klasse zusammenfassen und mit dem entsprechenden Bezeichner, wie z.B. “Sport” eine leichtere Interpretation durch den Anwender zulassen. Diese einfache Ontologie werden wir im weiteren Verlauf dieses Abschnittes zur Erläuterung der Idee einer konzeptbasierten Repräsentation verwenden. Durch die einfache Struktur kann man die Einbettung der Konzepte in die Repräsentation und deren Wirkung im resultierenden Verband leicht nachvollziehen.

Die Ontologie SO1 wurde während des Vorverarbeitungsprozesses, wie er in Abschnitt 8.2.3 beschrieben ist, integriert. Die Merkmale im Kontext und Verband bestehen nun nur noch aus Konzepten der Ontologie SO1 (siehe Abbildung 8.19), wobei keine Wortsinnerkennung erfolgte. Für Terme, die auf mehr als ein Konzept abgebildet werden können, wurde zufällig eines gewählt. Jeder Bezeichner in der Abbildung enthält nun die lexikalischen Einträge der Ontologie SO1. Bei der Vorverarbeitung wurden auch die verschiedenen Schreibweisen eines Wortstammes der Ontologie automatisch hinzugefügt und später in die Merkmalsnamen der FBA übernommen. Das führt dazu, dass Worte in unterschiedlichen Schreibweisen oder bei kurzen Worten, das Wort zweimal im Inhalt eines formalen Begriffes auftaucht (siehe Abbildung 8.20).

Zwei Fragen werden anhand der Ontologie im Folgenden untersucht:

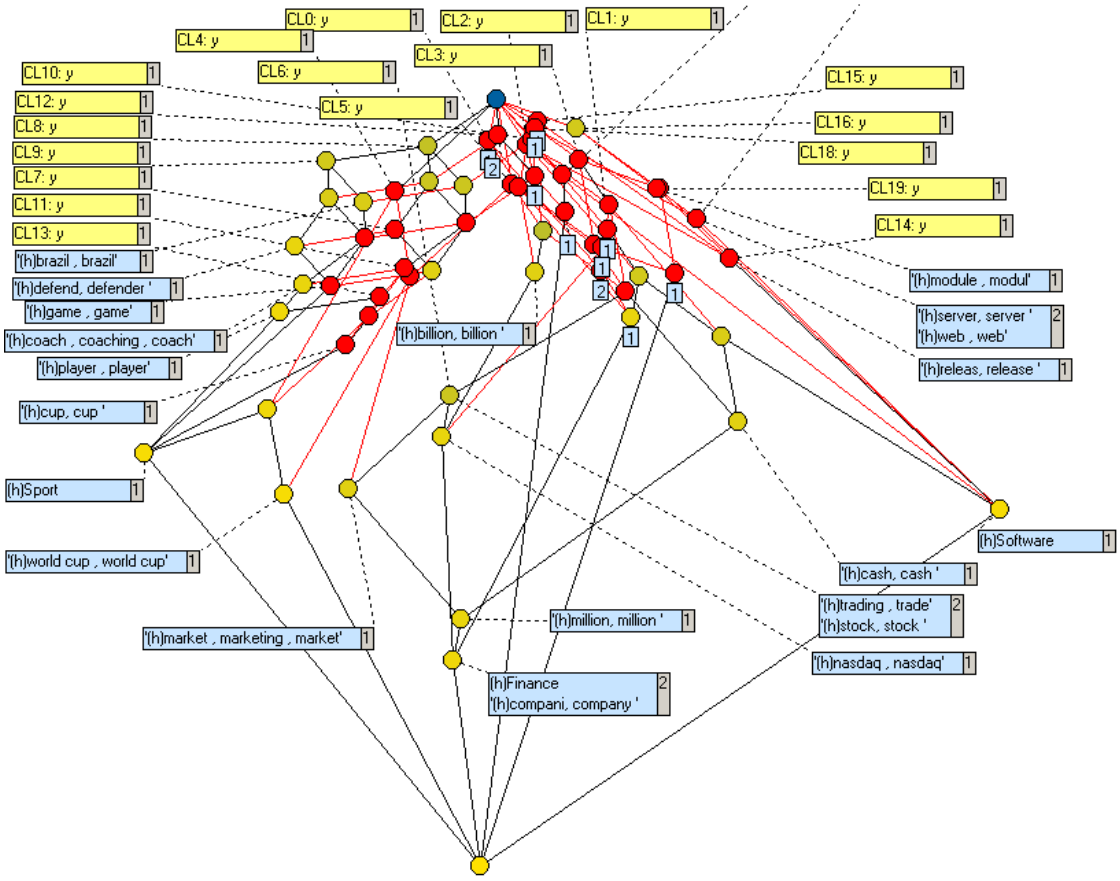

Abbildung 8.20: Verband CV1 des Datensatzes DS1 auf Basis der Ontologie OS1 ($\theta = 10\ \%$)

– Findet man die Hierarchie der Ontologie auch in ähnlicher Form im Begriffsverband wieder?
– Gewinnt der Verband durch die Ontologie an Struktur, Übersichtlichkeit und damit Verständlichkeit?

Dazu wurde der Verband auf der Basis der Ontologie berechnet. Die Visualisierung des Verbandes gibt Abbildung 8.20 wieder, den wir im Folgenden CV1 nennen werden. Alle Dokumente aus dem Bereich Fußball findet man unter dem von SPORT erzeugten Begriff. Der Gegenstandsbegriff Finanzdokument 4 (CL4) ist ebenfalls unter dem Merkmalsbegriff SPORT zu finden. Dies liegt an der fehlerhaften Abbildung des Wortes “world” auf das Konzept WORLD CUP. Die Phrase “in the world” des Dokumentes CL4 hat nichts mit dem “World Cup” zu tun. Diesen Fehler bei der Zuordnung der Konzepte kann man nur vermeiden, wenn man entsprechende Wortsinnerkennung beim Abbilden der Terme auf die Konzepte einsetzt, die hier nicht zum Einsatz gekommen ist. Völlig unabhängig davon ist das Wort “World” an sich kritisch zu betrachten, kann es doch in allen drei Bereichen vorkommen.

Bei der Analyse der beiden Klassen “Finanzen” und “Software” im Verband erkennt man, dass ausschließlich Dokumente des Finanzbereiches unter dem Merkmalsbegriff FINANCE zu finden sind. Bei den Softwaredokumenten ist das Ergebnis leider nicht so eindeutig. Drei Dokumente aus dem Finanzbereich erzeugen ebenfalls Unterbegriffe vom SOFTWARE erzeugten Begriff. Hier gibt es zwei Gründe: Einerseits ist in einigen Dokumenten tatsächlich die Rede von Finanzsoftware. Diese Verbindung ist damit korrekt, wobei die Frage, welches Thema vordergründig im jeweiligen Do-

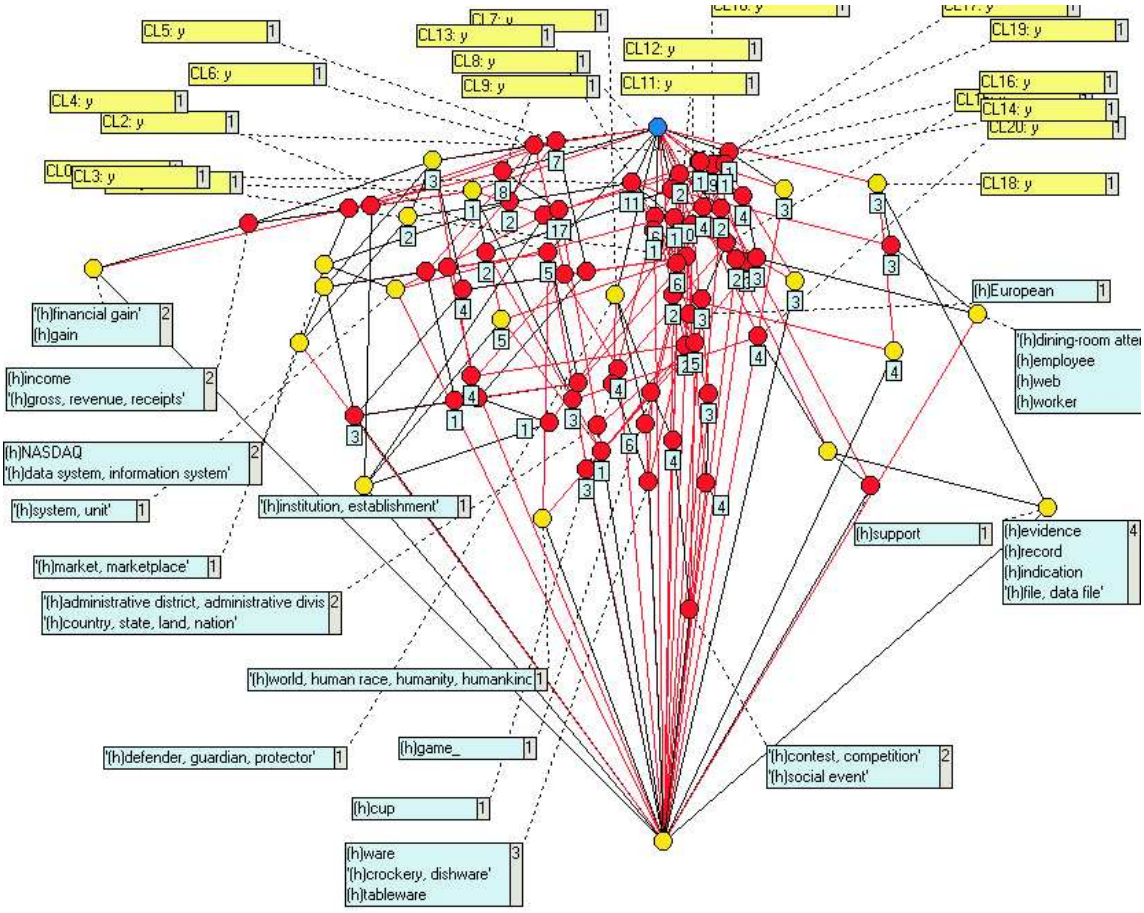

Abbildung 8.21: Verband WV1 des Datensatzes DS1 auf der Basis von WordNet ($\theta = 20\ \%$)

kument behandelt wird und ob die Verbindung zum Merkmalsbegriff SOFTWARE gewünscht ist, offen bleibt. Auf der anderen Seite ist die Zuordnung von DEVELOP als Unterkonzept von SOFTWARE in der Ontologie eher kritisch zu betrachten. Diese Beziehung ist die Ursache für die Verbindung zum Merkmalsbegriff SOFTWARE des Dokumentes zwei. Da die Entwicklung von Software etwas sehr Zentrales in diesem Bereich ist, das Wort selber aber eher unabhängig von der Domäne, ist die Aufnahme DEVELOP als Konzept in der Ontologie nicht immer zu empfehlen.

Die bisher verwendete Ontologie wurde für die Beispieltexte mit dem Ziel, die Bildung von Gruppen zu unterstützen, modelliert. Abbildung 8.21 stellt den Verband auf der Basis von WordNet-Konzepten (es wurden nur Konzepte berücksichtigt) unter Nutzung der “Kontext”-Strategie mit drei zusätzlichen generellen Konzepten dar. Wir nennen diesen WV1. WordNet als sehr allgemeine Ressource hat den Vorteil, dass sie sehr umfangreich ist, aber auch den Nachteil, dass sie sehr spezielle Themen nicht abdeckt. Wichtig zum konzeptuellen Clustern sind aber die in der Ontologie enthaltenen Worte und deren entsprechende Bedeutung. Eine domänenspezifische Ontologie ist in einem solchen Fall unter Umständen WordNet vorzuziehen.

Vergleichen wir die Visualisierung der Verbände CV1 und WV1 in Abbildung 8.20 und 8.21, dann fällt sofort die gestiegene Komplexität des Verbandes WV1 gegenüber CV1 auf, obwohl der Schwellwert für WV1 mit 20 % höher gewählt wurde als der von CV1 mit 10 %. Die gestiegene Anzahl von formalen Begriffen lässt sich mit der großen Zahl von referenzierten Konzepten erklären. Durch die Größe von WordNet konnten viele Worte erfolgreich auf Konzepte abgebildet werden.

Viele der generellen Begriffe im Verband WV1 enthalten zur Verständlichkeit beitragende Labels, wie z.B. FINANCIAL GAIN. Es sind aber auch viele Konzepte zu finden, die bei der Erklärung des Inhaltes nicht oder nur wenig helfen. So sind z.B. “evidence” oder “indication” nur schwer ohne den genauen Kontext verständlich und helfen nicht, die Dokumente dieser drei Themen gut zu trennen.

Durch die große Zahl an Konzepten wird der Verband unübersichtlicher und bei Fehlern der Wortsinnerkennung, wie sie im vorherigen Absatz angesprochen werden, bekommt man zusätzlich eine Reihe falscher Oberkonzepte hinzu. Diese Oberkonzepte stellen ihrerseits wieder Beziehungen zu Konzepten her, die nichts mit dem ursprünglichen Konzept zu tun haben. Diese Beziehungen findet man auch im Verband wieder. Der Effekt führt dann zu einer Verschlechterung der Clusterung (weitere Beispiele zu Problemen mit der Wortsinnerkennung findet man in Abschnitt 9.3.2). Auf der anderen Seite kann man auch beim Hinzufügen von WordNet-Konzepten die Oberkonzepte im Verband als allgemeine Begriffe wiederfinden. Damit erfüllt die WordNet-Integration eines der beiden Ziele. Durch die Fehler beim Abbilden der Terme auf die Konzepte, konnte das zweite Ziel nicht erreicht werden. Techniken aus dem NLP-Bereich (siehe Abschnitt 3.1.3) versprechen Verbesserungen im Bereich der Wortsinnerkennung und führen folglich auch zu einer übersichtlicheren Verbandsstruktur.

Zusammenfassung: Trotz der beschriebenen Schwierigkeiten durch zu viele “nichts sagende” Konzepte aus WordNet, fehlerhafter Abbildungen von Worten auf die “richtigen” Konzepte bzw. Probleme mit der modellierten Ontologie spiegelt sich die in der Ontologie modellierte Hierarchie klar im Verband wider. Generelle Konzepte der Ontologien entsprechen in unserem kleinen Beispiel allgemeinen Begriffen im Verband. Die Übersichtlichkeit wurde durch die Strukturierung der manuell modellierten Ontologie und die Reduktion der Anzahl der Begriffe erhöht. Bei der Nutzung von WordNet konnte dieser Effekt nicht erzielt werden. Um dieses Problem zu lösen, werden wir uns im nächsten Abschnitt 8.5.3 ansehen, wie man mit KMeans die Komplexität des Verbandes durch die Reduktion der Gegenstandsmenge weiter senken kann.

8.5.3 Reduktion der Gegenstandsmenge durch KMeans

Um Einfluss auf die Anzahl der dargestellten formalen Begriffen zu nehmen, wurde bisher die Anzahl der Merkmale verändert. Dieser Abschnitt analysiert Veränderungen der Gegenstandsmenge, wobei indirekt auch die Merkmalsmenge beeinflusst wird. Dazu wenden wir in einem ersten Vorverarbeitungsschritt KMeans zum Clustern der Textdokumente an. Wir möchten so die Anzahl der Dokumente auf eine überschaubare Clusteranzahl reduzieren. Die Formale Begriffsanalyse wird die Menge der Cluster als Gegenstandsmenge verarbeiten. Die Anzahl der Cluster ist kleiner als die Menge der Dokumente. Jeder Cluster fast im Allgemeinen mehrere Dokumente zusammen, die dann als ganzes durch Terme beschrieben werden.

Durch das Clustern von Dokumenten werden diese nicht nur in Gruppen eingeteilt, sondern die Terme, die zur Beschreibung der Cluster verwendet werden, sind die zentralen Terme aus allen Dokumenten einer Gruppe und nicht nur aus einem einzelnen Dokument. In gewisser Weise verändern wir so auch die Anzahl der Merkmale. Nicht jeder bisher wichtige bzw. beschreibende Term eines Dokumentes wird später als beschreibendes Merkmal auch wichtig zur Beschreibung eines Clusters sein. Weiterhin müssen die wichtigen Terme nicht unbedingt in jedem Dokument vorkommen. Die Zusammenfassung des Inhaltes der Dokumente eines Clusters durch wenige wichtige Terme aus allen Dokumenten führt zu einem übersichtlichen und abstrakten Verband. Ähnlich wie beim Auflösen von Synonymen und dem Hinzufügen von Oberkonzepten hilft das gemeinsame Auftreten von Termen in einem KMeans-Cluster bei der Abstraktion vom Detail eines Dokumentes. Im

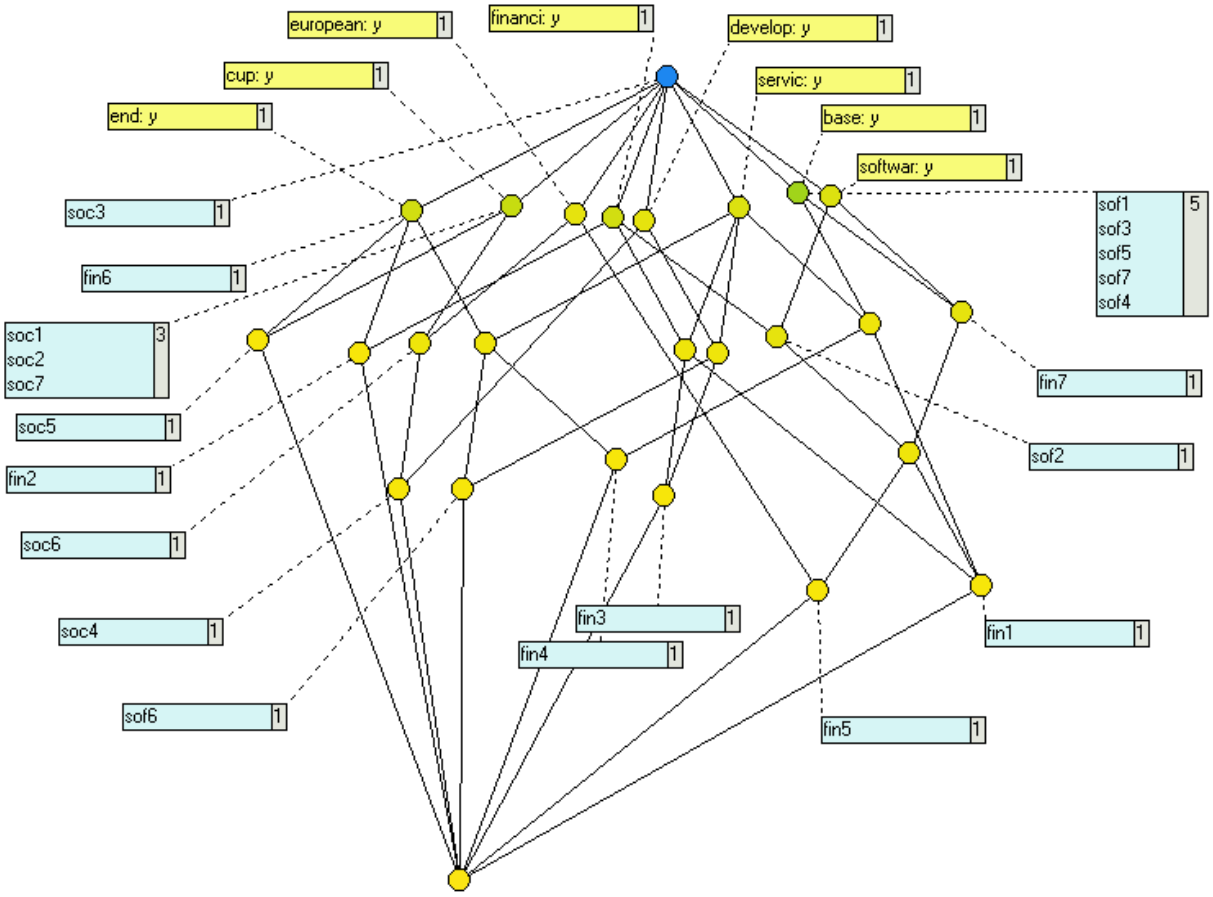

Abbildung 8.22: Begriffsverband TV5 erzeugt mit den gleichen Merkmalen wie Verband KV1

Folgenden werden wir die unterschiedlichen Beschreibungen – dokumentbasiert und clusterbasiert – anhand zweier Verbände diskutieren.

Abbildung 5.4 zeigt den Verband einer KMeans Clusterung $\mathbb{P}$ für den Beispieldatensatz DS1 mit der beschränkten Termmenge T, den wir im Folgenden KV1 nennen. Der Verband KV1 basiert auf dem Kontext $\mathbb{K}_{KV1} := (G, M, I)$ mit $G := \mathbb{P}, M := T$ und $(P, t) \in I$, wenn $(\vec{t_P})_t \geq \theta$ ist (siehe Abbildung 5.3). Den Kontext $\mathbb{K}_{TV1}$ ohne Einschränkung der Termmenge gibt Abbildung C.1 wieder. Der visualisierte Verband basiert auf zehn Clustern. Wie gewünscht erhält man einen einfachen und überschaubaren Verband. Abbildung 8.22 gibt den Verband unter Verwendung der gleichen Terme T als Merkmale, aber nicht für Cluster, sondern für jedes einzelne Dokument wieder. Man erhält den zugrundeliegenden Kontext $\mathbb{K}_{TV5} := (G', M', I')$ aus den Termvektoren der Dokumente mit $G' := D, M' := T$ und $(d, t) \in I'$, wenn $(\vec{t_d})_t \geq \theta$ ist. Wir nennen den aus $\mathbb{K}_{TV5}$ berechneten Verband TV5. Der Verband enthält, wie nicht anders zu erwarten, deutlich mehr Begriffe als der Verband KV1. Dies spiegelt sich auch in den vielen Abhängigkeiten zwischen den Begriffen, zum Teil erzeugt durch Dokumente unterschiedlicher Klassen, wider. Dies führt zu einem unübersichtlichen Verband, aus dem man schlecht die Clusterstruktur ablesen kann.

Analysieren wir dies anhand zweier Begriffe etwas detaillierter. Der Begriff mit dem Label “end” führt in beiden Verbänden KV1 (siehe Abbildung 5.4) und TV5 (siehe Abbildung 8.22) zur Verbindung von Clustern bzw. Dokumenten mit unterschiedlichen Themen; z.B. werden bei KV1 jeweils drei Dokumente aus allen Bereichen (Cluster 3, 7, 8) miteinander verbunden. Genauer sieht man

die Verbindung im Verband TV5, wo die Dokumente “fin2” und “fin3” mit “soc5” und “sof6” verbunden sind. Es sind also wirklich Dokumente aus allen drei Bereichen beteiligt, wobei durch die Abstraktion beim Clustern auch Dokumente aus Cluster mit “end” in Verbindung gebracht werden, die das Wort an sich nicht enthalten. Ein ähnliches Phänomen beobachtet man beim Begriff mit dem Label “cup”. Enthalten in TV5 nur zwei der sieben Sporttexte den Term “cup” und werden auf diesem Weg mit dem “World Cup” in Verbindung gebracht, umfasst der Merkmalsbegriff “cup” im Verband KV1 alle sieben Sporttexte. Dieser Effekt wird auf die so genannte co-occurrence, dem gemeinsamen Auftreten von Termen in Texten gleichen Inhaltes, zurückgeführt.

Es sei noch darauf hingewiesen, dass der Verband KV1 eine “Näherung” des Verbandes TV5 ist, d.h. der Verband KV1 kann als Begriffshierarchie der Dokumente (die in den entsprechenden Clustern sind) interpretiert werden. Ein Dokument d wird nicht durch den eigenen Termvektor $\vec{t_d}$, sondern durch den Termvektor $\vec{t_P}$ seines Clusters P beschrieben, d.h. wenn $d \in P$ dann gilt für die Relation I': $(d,t) \in I'$ wenn $(\vec{t_P})_t \geq \theta$ und nicht $(\vec{t_d})_t \geq \theta$. Alle Dokumente eines Clusters haben dann exakt den gleichen Termvektor und fallen so auf den gleichen Begriff.

Eine offene Frage ist, wie weit inhaltlich gesehen die genäherte Clusterrepräsentation $(\vec{t_P})_t$ eines Dokumentes von der jeweiligen Dokumentrepräsentation $(\vec{t_d})_t$ abweicht. D.h. welche Terme mehr und welche Terme in beiden Repräsentationen weniger Gewicht haben und dann entsprechend im Verband erscheinen.

In Abschnitt 8.5.4 diskutieren wir Arbeiten, die analog zu den Ansätzen des Abschnittes 8.5 die Formale Begriffsanalyse zum Verarbeiten und speziell zum Clustern von Textdokumenten eingesetzt haben.

8.5.4 Verwandte Ansätze

Es sind keine Arbeiten bekannt, in denen die Formale Begriffsanalyse zum Clustern von Textdokumenten eingesetzt wurde. Der ähnlichste bekannte Ansatz, bei dem mit Hilfe der Formalen Begriffsanalyse Textdokumente verarbeitet wurden, stammt aus dem Bereich des Information Retrieval.

Kim u.a. [132], [133] verbessern das Information Retrieval, indem sie das Browsing, d.h. die Präsentation von Suchergebnissen mittels Formaler Begriffsanalyse unterstützen. Dabei greifen sie auf Schlagworte und Thesauri zurück, um die Dokumente kompakt zu beschreiben. Dies kommt unserem ontologiebasierten Ansatz am nächsten, wobei wir durch die automatische Übersetzung der Terme eines Dokumentes mit wesentlich mehr Termen umgehen müssen als das in den Arbeiten von [132] und [133] der Fall ist. Die manuell zugeordneten Schlagworte stellen eine bessere Grundlage zum Clustern durch die Formale Begriffsanalyse dar. Da dieser Ansatz nicht skaliert, arbeiten wir mit einem automatischen, skalierenden Abbildungsmechanismus.

Zusammenfassung: In diesem Abschnitt haben wir den Einfluss unterschiedlicher Methoden auf die Clusterung von Textdokumenten mit Hilfe der Formalen Begriffsanalyse zur Steigerung der Verständlichkeit untersucht. Neben Techniken zur Exploration des Verbandes durch die Darstellung von Teilverbänden wurden auch Methoden zur Reduktion der Merkmals- und Gegenstandsanzahl analysiert. Auch die Visualisierung von Teilverbänden stellte sich als hilfreich heraus. Für eine leicht verständliche Visualisierung fehlten aber u.a. Zusammenhänge zwischen den formalen Begriffen. Mit Hilfe der Ontologie konnten diese in den Verband integriert werden. Trotz dieser Zusammenfassung und Strukturierung der Merkmale lässt sich mit Hilfe der Formalen Begriffsanalyse nur eine begrenzte Anzahl an Dokumenten verarbeiten. Daher wurde die Reduktion der Gegenstandsanzahl mit Hilfe von KMeans ohne den Einsatz einer Ontologie analysiert. Die berechneten KMeans-Textcluster ließen sich übersichtlich visualisieren. Auch wurde nicht mehr jedes Detail

eines Dokumentes in den Vordergrund gestellt, sondern mehr das Thema eines Clusters. Es fehlte aber eine übersichtliche und leicht verständliche Strukturierung zwischen den Clustern verschiedener Themen.

Im nächsten Kapitel untersuchen wir daher die Kombination dieser Ansätze. Wir berechnen Textcluster mit KMeans auf der Basis einer ontologiebasierten Repräsentation und nutzen zum Finden von interessanten Clustern oder Clustergruppen die Möglichkeit Teilverbände zu explorieren. Auf diese Weise kombinieren wir die Vorteile der verschiedenen Ansätze. Das Vorgehen wird im nächsten Kapitel im Detail anhand der Visualisierung von Textclustern erläutert.

9 Beschreibung von Textclustern mit Hintergrundwissen

Neben dem Berechnen von Clustern und der Steigerung der Clustergüte ist ein zentrales Thema des Clusterprozesses die Präsentation der Clusterergebnisse. Der Anwender muss die berechneten Cluster leicht verstehen und deren Zustandekommen nachvollziehen können. Dieses Kapitel vergleicht zwei Ansätze, einen einfachen Ansatz, der die Ergebnisse in Tabellen- oder Listenform präsentiert, und einen Ansatz auf der Grundlage der Formalen Begriffsanalyse anhand des Reuters-Datensatzes. Beide Ansätze werden zur Präsentation von Textclusterergebnissen verwendet. Neu an diesen Ansätzen sind die zu Grunde liegenden Merkmale sowie deren Auswahl für die Präsentation der Cluster. Als Merkmale werden Konzepte einer Ontologie und nicht wie in Kapitel 8 einfache Terme herangezogen.

Ziel der in diesem Kapitel vorgestellten Methoden ist *die Beschreibung von Textclustern*, die mit KMeans oder verwandten Algorithmen berechnet werden. Der Anwender soll in die Lage versetzt werden, den Inhalt der Dokumente eines Cluster erfassen bzw. abschätzen zu können. Wir benötigen daher eine kurze, leicht verständliche und prägnante Zusammenfassung jedes Clusters, die gleichzeitig diesen Cluster von den anderen Clustern einer Clusterung abgrenzt.

Abschnitt 9.1 stellt die Parameter für den zur empirischen Untersuchung verwendeten Reuters-Datensatz vor. Die Ergebnisse eines Clusterlaufes werden mit Hilfe von Tabellen auf der Basis von Konzepten in Abschnitt 9.2 diskutiert. Dabei werden auch Probleme dieses Ansatzes herausgearbeitet. Abschnitt 9.3 nutzt zur Präsentation der Ergebnisse die Formale Begriffsanalyse und diskutiert verschiedene Wege der explorativen Analyse der visualisierten Begriffsverbände. Auf den Einsatz von alternativen und verwandten Ansätzen gehen wir in Abschnitt 9.4 ein.

9.1 Der PRC_{30}-Datensatz

Ein speziell vorverarbeiteter PRC-Datensatz, den wir im Folgenden PRC_{30}-Datensatz nennen, wird im Rest dieses Kapitels als Beispiel-Datensatz genutzt. Der PRC-Datensatz, bei dem es sich um alle 12344 Reuters-Dokumente handelt, wurde in Abschnitt 2.1 ausführlich vorgestellt. Die zur Vorverarbeitung verwendeten Parameter werden im Folgenden eingeführt. Dabei geht es nicht nur um die Schritte zum Aufbau des Konzeptvektors, sondern auch um die angewendeten Strategien, mit denen das Hintergrundwissen in die Repräsentation integriert wurde.

Vorverarbeitung von PRC_{30}: Die Vorverarbeitung des Datensatzes PRC_{30} erfolgt mit der eigens entwickelten TextMining-Umgebung im KAON-Framework (siehe Anhang A), um das "Bag of Words"-Modell (siehe Abschnitt 4.2.1) abzuleiten. Die folgenden Schritte werden auf den Datensatz PRC_{30} angewendet: Als erstes werden die Großbuchstaben aller Worte in Kleinbuchstaben umgewandelt. Alle Stoppworte (siehe Kapitel 4.2.3) werden gelöscht. Dabei kommt eine Stoppwortliste mit 571 Einträgen zum Einsatz; es werden 416 Stoppworte aus dem Datensatz entfernt. Weiterhin werden nur Worte berücksichtigt die häufiger als 30 mal (Prunethreshold $\delta = 30$) und in mindestens zwei Dokumenten im Datensatz vorkommen (siehe Abschnitt 4.2.4). 17917 Terme

werden durch diesen Schritt gelöscht, so dass 2657 verschiedene Terme als Merkmale im Datensatz PRC_{30} verbleiben. Insgesamt werden noch 784434 Terme berücksichtigt. Einen Überblick über die Eigenschaften aller PRC-Datensätze nach der Vorverarbeitung findet man in Abschnitt 8.2.1 in Tabelle 8.1.

WordNet als Hintergrundwissen: Weiterhin werden wir für unser Beispiel den Termvektor durch Konzepte einer Ontologie ersetzen. Als Ontologie verwenden wir WordNet (siehe Abschnitt 6.3.3.1). Die Konzepte nennt man bei WordNet auch Synset. Beim Abbilden der Terme auf die Konzepte berücksichtigen wir nur Konzepte, die in WordNet als Substantive gekennzeichnet sind und ignorieren den Rest. Die verschiedenen Strategien für die Abbildung von Termen auf die Konzepte sind in Abschnitt 8.2.3 beschrieben. Wir wenden die folgende Strategie an, um den vorverarbeiteten PRC_{30}-Datensatz abzuleiten:

- Termauswahl: only,
- Wortsinnerkennung: first,
- zusätzliche Oberkonzepte: 5.

Die Menge der Terme T unseres Beispieldatensatzes PRC_{30} enthält nun nur noch Konzepte, die auch als Substantive in WordNet enthalten sind. Zur Wortsinnerkennung nutzen wir die WordNet-interne Ordnung der Wortsinne aus. Die Ordnung spiegelt die Häufigkeitsverteilung der Sinne in der englischen Sprache wider. Zusätzlich fügen wir mit der letzten Option fünf weitere Oberkonzepte (sofern vorhanden) hinzu. Diese Oberkonzepte entsprechen den Hypernymen von WordNet. Wir erhalten durch diesen Schritt eine Termmenge T bestehend aus 1935 Konzepten.

Gewichtung des Termvektors: Im letzten Schritt berechnen wir für jedes Dokument $d \in PRC_{30}$ die Gewichte für den Termvektor $\vec{t_d}$ nach dem tfidf-Maß (siehe Abschnitt 4.2.5.1).

Nach diesen Vorverarbeitungsschritten erhalten wir eine spezielle, um Hintergrundwissen angereicherte und tfidf-gewichtete Version des PRC-Datensatzes, den PRC_{30}-Datensatz. Wenn wir im Folgenden auf den Datensatz PRC_{30} bezugnehmen, setzen wir die genannten Vorverarbeitungsschritte voraus und gehen vom abgeleiteten Datensatz aus.

9.2 Tabellarische Ergebnispräsentation von Textclustern

Wie in der Einleitung dieses Kapitels ausgeführt, wird eine kurze, prägnante und leicht verständliche Zusammenfassung zur Präsentation eines Clusters benötigt. Die zum Clustern verwendeten Merkmale sind bei Textdokumenten Terme. Terme können z.B. Worte, Wortstämme oder Konzepte einer Ontologie sein. Im Allgemeinen sind die Terme für den Anwender leicht verständlich. Die Dokumente können aber sehr viele Terme enthalten. Für eine kurze Zusammenfassung müssen die wichtigen Terme eines Clusters extrahiert werden. Anschließend kann man diese Terme mittels unterschiedlicher Methoden dem Anwender präsentieren.

Im Folgenden werden wir die tabellarische Ergebnispräsentation von wichtigen Termen eines Clusterzentroiden diskutieren und die Schwächen dieser Methode herausarbeiten. Alternative Ansätze zur Extraktion wichtiger Merkmale für die Präsentation von Textclustern findet man im Abschnitt 4.5.3.

Tabelle 9.1: Anzahl der Dokumente, größte Reutersklasse, Precision pro Cluster, geordnet nach Clusternummer

Cluster	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Anzahl Dokumente	156	128	2	100	136	120	54	160	146	70	95	149	72	183	140	173	78	129	186	201	160	174	201	207	166
Reutersklasse	trade	money	earn	ship	defnoc	acq	earn	acq	acq	vegoil	acq	acq	earn	sugar	earn	earn	defnoc	acq	earn	earn	defnoc	grain	earn	crude	coffee
Precision	46%	44%	100%	50%	74%	82%	93%	87%	95%	57%	98%	64%	100%	77%	100%	100%	65%	17%	99%	99%	63%	81%	100%	81%	54%

Cluster	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49
Anzahl Dokumente	74	111	99	126	129	176	13	137	103	145	126	67	89	107	115	147	186	125	163	132	47	79	156	117	52
Reutersklasse	earn	defnoc	money	trade	acq	grain	earn	earn	earn	defnoc	earn	defnoc	earn	defnoc	cpi	acq	defnoc	acq	jobs	earn	earn	earn	earn	coffee	earn
Precision	100%	82%	40%	36%	87%	82%	46%	100%	100%	53%	100%	88%	100%	97%	37%	65%	98%	82%	26%	98%	100%	97%	75%	17%	96%

Cluster	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74
Anzahl Dokumente	147	165	154	132	74	108	6	252	179	165	112	165	137	53	113	130	75	26	125	119	204	154	115	107	151
Reutersklasse	defnoc	money	grain	earn	defnoc	ship	acq	gold	acq	money	trade	oilsee	money	earn	earn	earn	earn	earn	acq	acq	money	earn	ship	earn	earn
Precision	90%	39%	67%	100%	96%	56%	50%	37%	85%	62%	44%	15%	85%	100%	100%	100%	100%	100%	70%	71%	52%	70%	16%	100%	19%

Cluster	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99
Anzahl Dokumente	126	201	35	137	186	151	158	68	22	113	152	80	146	144	65	132	101	84	119	190	106	68	84	95	206
Reutersklasse	trade	earn	acq	trade	crude	acq	defnoc	defnoc	earn	defnoc	crude	earn	interes	earn	earn	acq	earn	money	earn	acq	crude	defnoc	earn	acq	defnoc
Precision	48%	94%	37%	57%	45%	96%	82%	99%	95%	90%	26%	100%	81%	70%	98%	59%	90%	61%	99%	98%	92%	94%	100%	63%	55%

Zur Auswahl der wichtigsten Terme greifen wir auf die Merkmalsextraktion aus Clusterzentroiden (siehe Abschnitt 4.5.2) zurück. Wir berechnen für den PRC_{30}-Datensatz mit dem Bi-Sec-KMeans-Algorithmus (siehe Abschnitt 5.4.2) 100 Cluster. Tabelle 9.1 gibt die Verteilung der Dokumente auf die Cluster wieder. Ebenfalls wird die zahlenmäßig größte Reutersklasse und deren Anteil am gesamten Cluster (Precision) wiedergegeben. Bei einer Clusterung können diese Informationen im Normalfall nicht zur Verfügung gestellt werden, da keine Klasseneinteilung bekannt ist. Für den PRC_{30}-Datensatz steht eine manuelle Klasseneinteilung zur Verfügung. Diese apriori Information über den Datensatz werden weder für die Clusterung noch für Beschreibung der Textcluster genutzt. Um die folgenden Aussagen besser nachvollziehen zu können, werden die Informationen über die Zusammenhänge zwischen Cluster und manueller Klassifikation in Tabelle 9.1 bereitgestellt.

Für alle 100 Cluster werden die zehn wichtigsten Merkmale, d.h. die Merkmale mit dem größten Gewicht extrahiert. Aus Platzgründen haben wir für die folgende Diskussion nur die Cluster 0 bis 9 aus allen 100 Clustern ausgewählt. Tabelle 9.2 gibt zu jedem dieser Cluster die wichtigsten zehn Terme und deren Wert im entsprechenden Zentroiden wieder. Alle aufgeführten Werte liegen über dem unteren Schwellwert $\theta_1 = 7\,\%$. Im Allgemeinen steigt die Anzahl der Terme, die den Schwellwert θ_1 pro Cluster überschreiten, auf bis zu 50 Terme. Nutzt man statt zehn Termen alle Terme über dem Schwellwert, führt diese große Menge an Termen zu einer unübersichtlichen Tabelle. Die Erfassung des Clusterinhaltes anhand einer solchen Tabelle ist sehr schwierig und unterstützt die explorative Analyse der Dokumente nicht.

Aber auch Tabelle 9.2 zeigt schon deutlich, dass eine Interpretation der Clusterergebnisse nicht trivial ist. Einige der Schwierigkeiten stammen von der einfachen Präsentation der Ergebnisse in Tabellenform, andere wiederum sind mehr substantieller Natur, da dem Anwender kaum strukturelle Zusammenhänge präsentiert werden. Im Folgenden werden wir anhand der gegebenen Tabelle die Erfassung des Inhaltes einiger Cluster diskutieren und Unzulänglichkeiten der Tabellenform herausarbeiten.

Zum Beispiel entnimmt man Tabelle 9.2 die Ähnlichkeit der Cluster 2 und 6, da beide von “loss” (Verlust), “failure” (Scheitern) und “non-accomplishment” (Unfähigkeit) handeln. Der Nutzer könnte beim Betrachten die Liste der Terme “depository financial institution’, ‘financial institution’, ‘rate’, ‘charge’, ‘institution’, ‘loss’, ‘monetary unit’, ‘financial loss’ und ‘expenditure’ für den Cluster 1 als verlustreiche Finanztransaktionen interpretieren (was sich bei der Betrachtung der entsprechenden Reuters-Dokumente als richtig herausstellt). Man kann sich leicht vorstellen, dass – entsprechende Benutzeroberflächen vorausgesetzt – diese Beobachtungen auch über alle Cluster gemacht werden können. Eine Tabelle stellt dafür aber kein adequates Mittel dar und führt zu erhöhtem Aufwand. Man ist aber prinzipiell in der Lage, den Inhalt eines Clusters zu erfassen.

Tabelle 9.2: Die wichtigsten zehn Terme (Synsets) der ersten zehn von 100 Clustern für den Reuters-Datensatz PRC_{30} sortiert nach Werten im Zentroid

Cluster 0		Cluster 1		Cluster 2		Cluster 3		Cluster 4	
amount	0,12	depository financial instit	0,09	loss	0,34	Irani, Iranian, Persian'	0,14	indebtedness, liability, fin	0,12
billion, one million million,	0,11	financial institution, financ	0,09	failure	0,33	Iran, Islamic Republic of	0,13	obligation	0,12
large integer'	0,11	rate, charge per unit'	0,09	nonaccomplishment, non	0,32	gulf	0,13	debt	0,12
integer, whole number'	0,11	charge	0,09	Connecticut, Nutmeg Sta	0,28	vessel, watercraft'	0,12	written agreement'	0,1
insufficiency, inadequacy	0,1	institution, establishment	0,09	ten, 10, X, tenner, decad	0,24	ship	0,12	agreement, understandin	0,08
deficit, shortage, shortfall	0,1	loss	0,08	American state'	0,23	craft	0,12	creditor	0,08
number	0,09	monetary unit'	0,07	state, province'	0,22	Asian, Asiatic'	0,11	lender, loaner'	0,08
excess, surplus, surplusa	0,09	central, telephone exchar	0,07	system, unit'	0,19	person of color, person o	0,10	statement	0,07
overabundance, overmuc	0,09	financial loss'	0,06	network, net, mesh, mest	0,19	Asian country, Asian nati	0,10	billion, one million million,	0,06
abundance, copiousness	0,09	outgo, expenditure, outla	0,06	September, Sep, Sept'	0,18	oil tanker, oiler, tanker, ta	0,10	large integer'	0,05

Cluster 5		Cluster 6		Cluster 7		Cluster 8		Cluster 9	
text, textual matter'	0,15	loss	0,34	gross sales, gross revent	0,11	tender, legal tender'	0,15	metric weight unit, weight	0,15
matter	0,15	failure	0,33	sum, sum of money, amo	0,09	offer, offering'	0,14	metric ton, MT, tonne, t'	0,15
letter, missive'	0,15	nonaccomplishment, non	0,32	income	0,09	medium of exchange, mo	0,11	mass unit'	0,14
sign, mark'	0,13	common fraction, simple	0,22	financial gain'	0,09	speech act'	0,1	palm, thenar'	0,14
clue, clew, cue'	0,13	fraction	0,22	gain	0,09	indicator	0,1	area, region'	0,12
purpose, intent, intention,	0,11	rational number'	0,22	enterprise	0,05	standard, criterion, meas	0,1	unit of measurement, unit	0,10
evidence	0,11	real number, real'	0,22	business, concern, busin	0,05	reference point, point of r	0,09	organic compound'	0,10
indication, indicant'	0,11	complex number, comple	0,22	assets	0,05	signal, signaling, sign'	0,08	oil	0,10
goal, end'	0,1	one-half, half'	0,22	division	0,05	acquisition	0,06	lipid, lipide, lipoid'	0,10
writing, written material, p	0,07	revolutions per minute, rp	0,22	army unit'	0,05	giant	0,06	compound, chemical com	0,08

Weiterhin existieren wichtige Strukturen, die eine Tabelle nur unzureichend präsentiert. Analysieren wir z.B. das Vorkommen des Terms “oil” (Öl). “oil” findet man nur in Cluster 9. Cluster 3 enthält zwar auch “oil”, leider ist der Term aber nicht unter den zehn wichtigsten Termen, sondern an Position 13. Der Term setzt die beiden Cluster (und noch ein paar weitere Cluster, die nicht in der Tabelle wiedergegeben werden und auf die wir an dieser Stelle nicht weiter eingehen wollen) in Beziehung zueinander. Ein allgemeinerer Term wie “chemical compound” würde wahrscheinlich wesentlich mehr Cluster als der spezielle Term “oil” umfassen. Informationen über die Veränderungen der Clustermengen beim Wechsel von “chemical compound” zu “oil” würde wesentlich zum Verständnis beitragen. Da Öl auch in Zusammenhang mit Ölfarben im Englischen auftreten kann, würde man einen Cluster, der z.B. den Term “covering” (welcher ein Hypernym von Öl in diesem Sinne ist) enthält, sehr gut von den anderen unterscheiden können. Bei der Suche nach weiteren Öl-Sorten wäre es wichtig zu wissen, ob z.B. der Term “palm” aus Cluster 9 nur in diesem und keinem weiteren Cluster auftaucht, um herauszufinden, wie wichtig und alleinstellend dieser Term für den Cluster ist.

Die Probleme bei der Extraktion von beschreibenden Termen für Cluster entstehen außerdem durch die Annahme, dass die Ordnung der Terme die Bedeutung des Terms für die Beschreibung adäquat widerspiegelt. Dies ist häufig nicht der Fall, wie das Beispiel von Cluster 6 zeigt, wo die Art des Verlustes (“loss”), die in den Dokumenten angesprochen wird, unklar bleibt. Die restlichen Terme zur Zusammenfassung des Inhaltes von Cluster 6 haben es nicht unter die besten zehn Terme geschafft. Gleiches fällt bei Cluster 3 für “oil” auf. Tatsächlich hängt die Wichtigkeit eines Terms zur Zusammenfassung des Inhaltes häufig von der Fähigkeit der Terme ab, Gemeinsamkeiten und Unterschiede zwischen Clustern herauszuarbeiten. Die Gewichte der Terme sind meist in der richtigen Größenordnung, aber nicht immer erhalten die Terme durch die Gewichte der Zentroide die richtige Reihenfolge bzw. befinden sich unter den ersten zehn Termen. Auch kann es vorkommen, dass man unterschiedlich viele Terme zur Beschreibung des Clusterinhaltes benötigt.

Zusammenfassung Die Beispiele zeigen, dass man anhand der Terme in Tabellenform den Inhalt der Cluster verstehen kann. Leider wird die explorative Analyse von Zusammenhängen und Beziehungen zwischen den Clustern in der Tabelle nur unzureichend unterstützt. Weiterhin findet man bedingt durch Probleme bei der Bestimmung der Termreihenfolge nicht immer die zur Inhalts-

erfassung benötigten Terme in der Tabelle wieder. Für die Extraktion der Beziehungen benötigt man weitere Analyseschritte, wie z.B. die Formale Begriffsanalyse. Zusätzlich erlaubt dieser nachgeschaltete Schritt die Steigerung der Termanzahl für die explorative Analyse. Die Bedeutung der Termreihenfolge sinkt, da durch die erhöhte Anzahl auch schlechter gewichtete Terme noch präsentiert werden können.

Wir analysieren im nächsten Abschnitt anhand des Beispieldatensatzes PRC_{30} die Clusterergebnisse auf der Basis der Formalen Begriffsanalyse (Einführung siehe Abschnitt 5.5) und gehen in diesem Zusammenhang auf verschiedene explorative Vorgehensweisen zur Analyse von Clustererergebnissen ein. Ein Überblick über alternative Ansätze, wie z.B. Regellerner, ist in Abschnitt 9.4.1 zu finden.

9.3 Konzeptuelles Clustern zur Beschreibung von KMeans-Clustern

In Kapitel 8 (siehe auch [119, 115]) haben wir für das konzeptuelle Clustern von Textdokumenten die Formale Begriffsanalyse (FBA) eingesetzt und in Kapitel 5.5 soweit in dieser Arbeit benötigt eingeführt. Die folgenden Abschnitte geben die Ansätze aus [119, 115] wieder. Abschnitt 9.3.1 zeigt am Beispiel eines Begriffsverbandes für den PRC_{30}-Datensatz die berechneten Beziehungen zwischen den Clustern, während Abschnitt 9.3.2 die Vorteile eines visualisierten Begriffsverbandes herausstellt. Abschnitt 9.3.3 diskutiert zwei Methoden zur explorativen Analyse von Verbänden zur Beschreibung von Textclustern.

Stand in Abschnitt 8.5 das Berechnen von Textclustern auf der Basis unterschiedlicher Repräsentationen im Mittelpunkt der Betrachtungen, stellen wir uns in diesem Abschnitt auf den Standpunkt, dem Anwender Informationen über geclusterte Textdokumente vermitteln zu wollen. Der Anwender hat eine Menge von Dokumenten mit einem statistischen bzw. maschinellen Lernverfahren geclustert und versucht nun diese Cluster zu interpretieren bzw. den Inhalt zu erfassen. Wir unterstützen ihn, indem wir Verbände zur Bestimmung von Gemeinsamkeiten und Unterschieden zwischen den Clustern mit der Formalen Begriffsanalyse berechnen und visualisieren.

9.3.1 Beschreibung von Textclustern durch formale Begriffe

Die Term-Selektion zum Ableiten eines Kontextes $\mathbb{K} := (G, M, W, I)$ erfolgt für unser laufendes Beispiel PRC_{30} analog zum Abschnitt 9.2 für die dort berechneten 100 Cluster. Wir nutzen aber im Gegensatz zum letzten Abschnitt nun nicht nur einen Schwellwert $\theta_1 = 7\ \%$ sondern auch einen zweiten höheren Schwellwert von $\theta_2 = 20\ \%$, wobei wir dies im Folgenden durch (m) für θ_1 und (h) für θ_2 kennzeichnen (siehe Abschnitt 4.5). Dadurch erhält man einen mehrwertigen Kontext mit der Gegenstandsmenge $G := \mathbb{P}$, der Merkmalsmenge $M := T$, der Wertemenge $W := \mathbb{R}$ und mit $x \in W$ erhält man die Relation I wie folgt: $(P, t, x) \in I :\Leftrightarrow (\vec{t}_P)_t = x$. Den mehrwertigen Kontext überführt man mit Hilfe des begrifflichen Skalierens (siehe Abschnitt 5.5.2) unter Nutzung einer Ordinalskala in einen einwertigen Kontext $\mathbb{K}' := (G', M, J)$ mit der Gegenstandsmenge $G' := \mathbb{P} \times m, h$, der Merkmalsmenge $M := T$ und der Relation $((P, m), t) \in J :\Leftrightarrow (\vec{t}_P)_t \geq \theta_1$ und $((P, h), t) \in J :\Leftrightarrow (\vec{t}_P)_t \geq \theta_2$. Da wir in diesem Abschnitt einen "gedrehten" Verband visualisieren, wird nicht die Menge der Terme, sondern die Menge der Cluster (Gegenstände) skaliert. Dadurch werden die Namen der Gegenstände um (m) oder (h), z.B. für Cluster CL3 zu CL3(m) oder CL3(h), erweitert. Aus dem Kontext $\mathbb{K}'$ berechnen wir den Begriffsverband $\mathfrak{B}(\mathbb{K})$ (siehe Kapitel 5.5).

Der Verband enthält einige hundert formale Begriffe. Jeder Begriff fasst Cluster des KMeans-Schrittes zusammen. Ein Begriff eines Verbandes spiegelt die konzeptuelle Ähnlichkeit der enthaltenen KMeans-Cluster wider. Folgendes Beispiel macht dies deutlich: Ein formaler Begriff, den wir im Folgenden mit (*) referenzieren wollen, hat {CL3(m), CL9(m), CL23(m), CL79(m), CL85(m), CL95(m)} als Umfang und {organic compound, oil,'lipid, lipide, lipoid','compound, chemical compound'} als Inhalt. Der formale Begriff gibt die Gemeinsamkeiten der genannten KMeans-Cluster wieder. Die Mehrheit der Dokumente in dem konzeptuellen Cluster (Begriff) handelt von Öl.

Der formale Begriff (*) hat drei Unterbegriffe: Der erste hat {CL3(m)} im Umfang sowie die Merkmale von oben und zusätzlich die Merkmale 'oil tanker' und 'Iranian' im Inhalt. Der zweite hat {CL9(m)} im Umfang sowie die Merkmale von oben und als weitere 'area', 'palm' und 'metric ton' im Inhalt. Der dritte Unterbegriff hat {CL23(m), CL79(m), CL85(m), CL95(m)} im Umfang und die Merkmale von oben und zusätzlich 'substance, matter' im Inhalt. Diese drei Unterbegriffe von (*) zeigen die Unterschiede der Cluster, die im formalem Begriff (*) zusammengefasst wurden. Wir wissen also, dass die meisten Dokumente der Cluster in (*) von 'oil' handeln und Cluster 3 speziell vom Transport von Öl (vom/zum Iran). Cluster 9 handelt eher von Palmöl und die verbleibenden Cluster von Rohöl (crude oil).

Wie eben am Beispiel beschrieben, hilft der Begriffsverband tatsächlich, Gemeinsamkeiten und Unterschiede verschiedener KMeans-Cluster aufzudecken und herauszuarbeiten. Dabei nutzen wir als Basis die gleichen Informationen wie in Abschnitt 9.2. Anhand dieser Informationen werden die Cluster während der Berechnung des Verbandes in Beziehung zueinander gesetzt und das Ergebnis wird leicht verständlich visualisiert. Der Aufwand zur manuellen Exploration (d.h. zur manuellen Bestimmung des Verbandes), der sich in der Berechnung des Verbandes widerspiegelt, macht den Vorteil des Einsatzes der FBA deutlich.

Formale Begriffe können Textcluster in der gewünschten Form zusammenfassen. Unklar bleibt aber, wie wir aus den vielen hundert Begriffen schnell und einfach die "interessanten" Begriffe herausfinden. Dabei betrachten wir solche Begriffe als interessant, die uns helfen, schnell und einfach den Inhalt einer größeren Clusteranzahl zu erfassen. Im Forschungsbereich der Formalen Begriffsanalyse wurden dazu Visualisierungstechniken entwickelt, die wir im Folgenden auf den Verband des PRC_{30}-Datensatz anwenden werden.

9.3.2 Visualisierung von Textclustern

In Kapitel 5.5.3 wird die Visualisierung des Begriffsverbandes durch Hasse-Diagramme vorgestellt sowie das Lesen und Interpretieren erläutert. Abbildung 9.1 gibt das Hasse-Diagramm des durch die Cluster 3, 9, 23, 39, 79, 85, 95 erzeugten Teilverbandes für unser laufendes Beispiel wieder. Der Begriffsverband ist der selbe wie in Abschnitt 9.3.1. Alle dargestellten Cluster besitzen einen Wert θ_1 für "chemical compound", der größer als 7 $\%$ ist. Aus technischen Gründen wurde dieses Diagramm gedreht (siehe Kapitel 5.5.3). Im Folgenden analysieren wir den Begriffsverband, den das Diagramm wiedergibt, im Detail.

Der Knoten unten in der Mitte von Abbildung 9.1 mit der Bezeichnung "oil" stellt den formalen Begriff (*) aus dem letzten Abschnitt dar. Wir erkennen weiterhin eine Kette von formalen Begriffen mit steigender Spezifizität. Der generellste Begriff von dieser Kette, markiert mit (**) in Abbildung 9.1, enthält im Umfang Dokumentcluster, die etwas mit chemischen Verbindungen zu tun haben. Es handelt sich um die Cluster: 3,9,23,39,79,85,95, die mit mittlerer (m) Wichtigkeit vorkommen (Schwellwert θ_1). Der nächste Begriff ist (*). Sein Umfang ist auf Cluster beschränkt, die mit "oil" in Beziehung stehen. Dies sind alle Cluster des vorherigen Begriffes außer 39. Begriff (*) haben wir schon in Abschnitt 9.3.1 diskutiert. Betrachten wir das Diagramm, so finden wir tatsächlich seine drei Unterbegriffe wieder. Der Begriff (***) (ebenfalls markiert in Abbildung 9.3.1) mit den Cluster

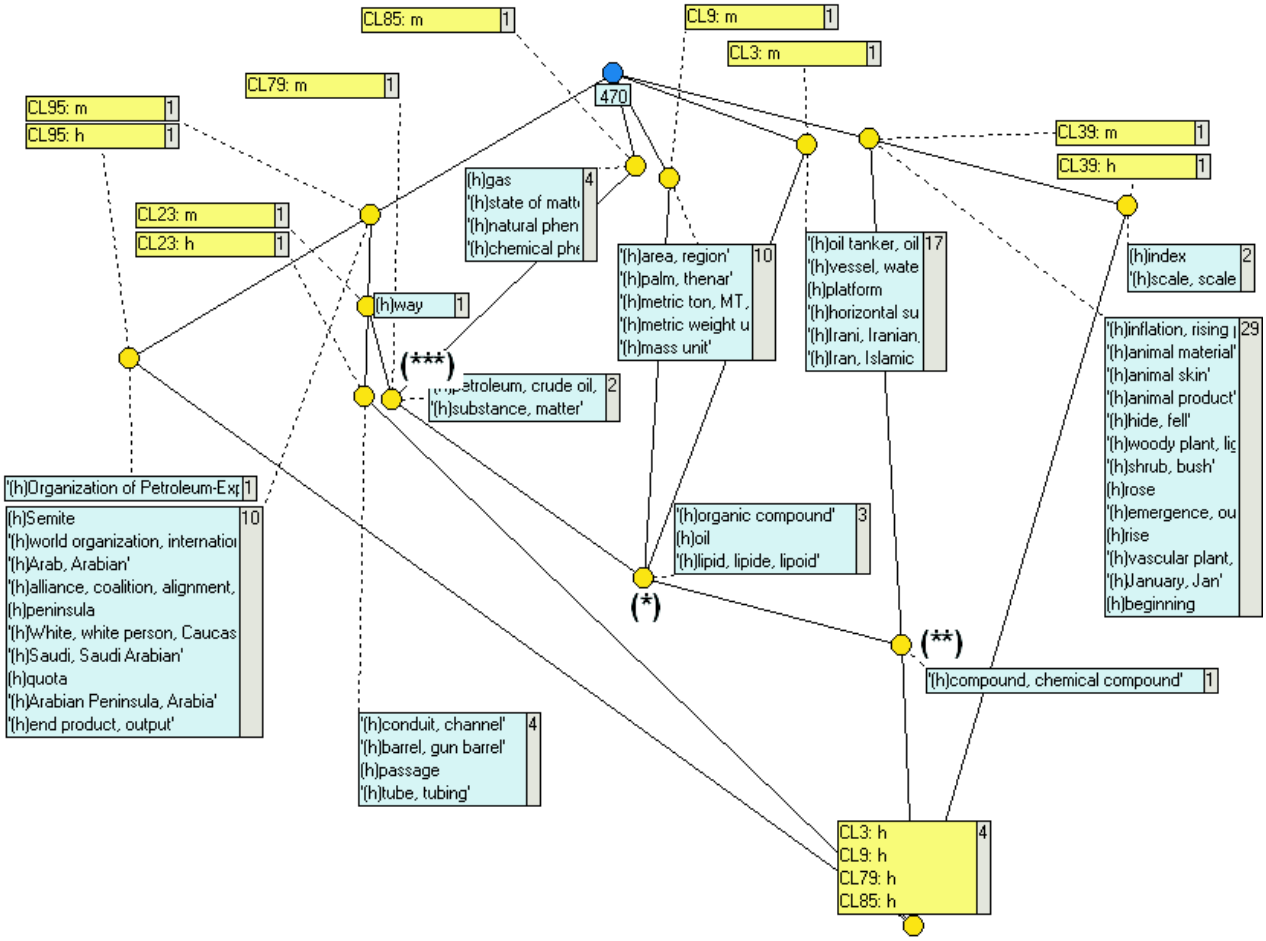

Abbildung 9.1: Das resultierende konzeptuelle Clusterergebnis der KMeans-Text-Cluster (visualisiert für die Cluster, die mit “chemical compounds” in Beziehung stehen)

23, 79, 85, 95 im Umfang, in dem es um Rohöl “crude oil” geht, wird noch einmal aufgespalten. Während keine weiteren Informationen zum Cluster 79 im Diagramm enthalten sind, steht bei den Dokumenten der Cluster 23 und 95 der Transport und bei Cluster 95 zusätzlich “oil quotas” der OPEC im Vordergrund, wie sich im Diagramm erkennen lässt. Damit steht uns eine Methode zur Verfügung, um leicht die interessanten Begriffe im Verband zu entdecken.

Interessant ist der Ursprung der beobachteten Begriffshierarchie, vom z.B. Begriff (**) zu Begriff (*). Er beruht auf der Ontologiehierarchie. Eine solche Beziehung wird durch das Hinzufügen von Oberkonzepten der Ontologie in die Textrepräsentation eingebracht, d.h. kein Reuters-Text enthält das Wort “chemical compound”. Der Term wird aber den Dokumenten hinzugefügt, die Unterkonzepte von “chemical compound” enthalten, wie dies z.B. bei “oil” der Fall ist. Auf diesem Wege werden Cluster durch allgemeine Terme in Beziehung zueinander gebracht und man erhält die Ketten mit steigender Spezifizität der beschreibenden Terme.

Mit Hilfe der Visualisierung ist eine einfache Analyse der Textcluster möglich. Bei einer umfangreichen Untersuchung der Textcluster stößt man auf Fehler, die auf den automatischen Übersetzungsprozess der syntaxtischen Terme in Konzepte zurückzuführen sind. Diese Fehler können zu Problemen bei der Interpretation der Textcluster führen. Folgendes Beispiel verdeutlicht das Phä-

nomen: Wie man Abbildung 9.1 entnimmt, handeln die Dokumente in Cluster 85 von Öl, wobei der Begriff auch “gas” im Inhalt hat. Neben “gas” kommt auch “state of matter” als wichtiges beschreibendes Konzept vor. Es wird durch die WordNet-Ontologie hinzugefügt, welches “gas” als wahrscheinlichsten Wortsinn “state of matter” zuordnet. Die entsprechenden Konzepte werden dann als Hypernyme dem Vektor hinzugefügt. Eine manuelle Untersuchung der Dokumente zeigt, dass der Fehler bei der Wortsinnerkennung liegt. In den Dokumenten wird “gas” als Synonym zu “gasoline” (Benzin) und nicht zu “state of matter” verwendet, was wiederum besser zur inhaltlichen Beschreibung der Cluster passt.

Außerdem fehlen einige wichtige Konzepte in unserer Clusterbeschreibung, die zu einer besseren Erklärung des Inhaltes führen würden. Das wichtigste Konzept in diesem Fall ist “refinement”. Es hat ein Gewicht leicht unter dem Schwellwert θ_1. Unser Ansatz lieferte leider nicht die korrekte Erklärung für den Cluster 85: Die Dokumente im Cluster sind hauptsächlich über die Raffinierung von Rohöl zu Benzin.

Die Konzeptliste des Clusters 39 birgt einen ähnlichen Fehler. So findet man in der Liste das Wort “rose”. Im Text wird es als Verb verwendet (von “to rise”). Beim Aufbau des “Bag of Words”-Modell verliert man die Information über die Stellung des Wortes im Satz. Wie in Abschnitt 9.1 beschrieben, interpretieren wir alle Worte als Substantive, was hier zur Verwendung “rose” als Rose im Sinne einer Blume führt. Durch die fehlerhafte Wortsinnerkennung werden wiederum die falschen Oberkonzepte der Repräsentation hinzugefügt und so die Interpretation des Clusterinhaltes erschwert.

Erst mit Hilfe der Formale Begriffsanalyse konnten die Fehler bei der Wortsinnerkennung entdeckt werden. Fehler dieser Art lassen sich in Tabellenform schwerer entdecken, da keine Unterstützung durch die Visualisierung von Beziehungen zwischen Clustern und deren Themen exisitiert. Auch die geringe Anzahl der Terme in der Tabellenform erschwert die Entdeckung. Zur Lösung der beschriebenen Probleme kann man sowohl eine bessere Wortsinnerkennung als auch Techniken zur Wortarterkennung einsetzen. Beide Aufgaben werden im Forschungsbereich NLP (siehe Abschnitt 3.1.3) untersucht. Eine Kombination der Ansätze dieser Arbeit mit den NLP-Techniken erscheinen daher vielversprechend.

Zusammenfassend kann man feststellen, dass die Visualisierung des Begriffsverbandes eine Navigation der Strukturen zur Erklärung von Gemeinsamkeiten und Unterschieden der einzelnen Cluster erlaubt. Der Verband erweitert die schon extrahierte Information zur Beschreibung der Cluster, indem er sie in Beziehung zueinander setzt. Gleichzeitig finden sich Strukturen der Ontologiehierarchie im Verband wieder. Das wiederum steigert die Verständlichkeit der Erklärungskomponente.

Durch die Clusterung mit KMeans auf einer ontologiebasierten Basis erfolgt eine Komprimierung bzw. Zusammenfassung der Informationen einzelner Dokumente in einer Form, die eine verständliche Visualisierung durch einen Begriffsverband überhaupt erst möglich macht. Durch die Komprimierung der Information erfolgt eine Abstraktion vom Detail. Dadurch wird der Verband kleiner und übersichtlicher und die Visualisierung verständlicher.

9.3.3 Methoden zur explorativen Analyse der visualisierten Verbände

Nachdem wir in Abschnitt 9.3.2 den Nutzen der visualisierten Verbandsstruktur untersucht haben, beschreiben wir hier zwei Methoden zur Analyse des Verbandes, die nach der Berechnung und Visualisierung des Verbandes Anwendung finden können. Ziel der Methoden ist die einfache Bestimmung von interessanten Teilverbänden, die ihrerseits übersichtlich visualisiert werden können. Die erste Methode greift dabei auf die Struktur des Verbandes zurück, siehe Abschnitt 9.3.3.1, während die zweite Methode auf der Analyse der KMeans Cluster und deren Ähnlichkeit basiert, siehe Abschnitt 9.3.3.2.

9.3.3.1 Nutzung der Verbandsstruktur bei der explorativen Analyse

Abbildung 9.2 gibt den vollständigen Begriffsverband unseres Beispieldatensatzes PRC_{30} für 100 Cluster wieder. Die Komplexität und Größe des Verbandes erlauben keine übersichtliche Darstellung. Trotzdem kann man wichtige und interessante Teilverbände lokalisieren, die anschließend unabhängig vom restlichen Verband visualisiert werden. Bevor wir das Vorgehen erläutern, benötigen wir eine genauere Vorstellung, was wir unter interessanten Teilverbänden verstehen. In Abschnitt 9.3.2 haben wir uns Ketten von formalen Begriffen angesehen, die das Verständnis des Clusterinhaltes erleichterten. Diese Ketten bestehen aus Begriffen von unterschiedlichem Generalisierungsgrad. Es gibt daher in jeder Kette sehr allgemeine und sehr spezifische Begriffe. Wir betrachten diese Ketten als interessant, da die allgemeinen Begriffe häufig eine Menge an KMeans-Clustern strukturiert zusammenfasst und so den Ausgangspunkt für interessante Teilverbände bilden können. Details liefert die anschließende Visualisierung dieser Teilverbände. Wir verlieren zwar durch die Beschränkung auf Teile des Verbandes auch einen gewissen Grad der Beziehungen zwischen den Begriffen, erhalten aber dadurch einen übersichtlichen Verband in der Visualisierung.

Wie schon erwähnt, kann der Verband aus Abbildung 9.2 auf so beschränktem Platz und mit so vielen Beziehungen nicht übersichtlich und leicht verständlich visualisiert werden. Trotzdem findet man unterstützt durch die Visualisierungssoftware Cernato nach sehr kurzer Zeit erste Clusterketten. Einige Bezeichner der in Abbildung 9.2 dargestellten formalen Begriffe wurden aufgeklappt. Die formalen Begriffe, die “loss” oder “rate” im Inhalt haben, gruppieren Cluster zum Thema “earn”. Zum gleichen Thema finden wir eine weitere Kette, die in der Mitte der Abbildung zu erkennen ist. Es handelt sich um die Begriffe mit “income”, “financial gain” usw. im Inhalt. Auf der linken Seite ganz unten sind die schon bekannten Begriffe aus dem Bereich “oil” und allgemeiner “chemical compound” abgebildet. Diese Begriffe entsprechen denen aus Abbildung 9.1. Unterstützt durch die Software Cernato kann man sich sehr leicht den Teilverband visualisieren lassen.

Wie man an den Beispielen erkennen kann, sind die inhaltlich mit allgemeinen Worten beschriebenen Begriffe auch die, die am dichtesten zum Top-Begriff positioniert sind. Diese Tatsache macht es leicht, einen Ausgangspunkt einer Clusterkette zu finden. Wir müssen uns dazu nur die Begriffe mit direkter Verbindung zum Top-Begriff ansehen und nicht den gesamten Verband explorieren. Auf diese Weise wurde nicht nur der interessante Teilverband aus Abbildung 9.1 entdeckt, sondern auch Clusterketten zu anderen Themen wie Zucker, Getreide, Kaffee oder Geld.

9.3.3.2 Nutzung der Ähnlichkeitsbeziehungen zwischen Textclustern bei der explorativen Analyse

Die Nutzung von sehr allgemeinen Begriffen und strukturellen Informationen des Verbandes zur explorativen Analyse von Verbänden haben wir im letzten Abschnitt diskutiert. Wir werden im Folgenden eine weitere Methode zur Visualisierung von Teilverbänden vorstellen. Dazu berechnen wir die Kosinus-Ähnlichkeit zwischen allen Textclustern und bestimmen in einem ersten Schritt ähnliche Textcluster. Im zweiten Schritt erfolgt die Visualisierung der Teilverbände nur noch für diese ähnlichen Textcluster und nicht mehr alle Textcluster. Abbildung 9.3 gibt eine Visualisierung der Ähnlichkeitsbeziehungen zwischen den Textclustern unseres laufenden Beispiel für den PRC_{30}-Datensatz mit 100 Clustern wieder. Im Folgenden erläutern wir die Entstehung dieser Grafik und geben dann eine inhaltliche Interpretation.

Um die Grafik in Abbildung 9.3 zu erzeugen, wird die Ähnlichkeit zwischen den Zentroiden $\vec{t}_P$ der Textcluster P mit dem Kosinus-Maß (siehe Abschnitt 5.2.2) berechnet und ins Verhältnis zur maximalen Ähnlichkeit gesetzt. Zur Visualisierung des Grafen nutzen wir den “magnetic spring”

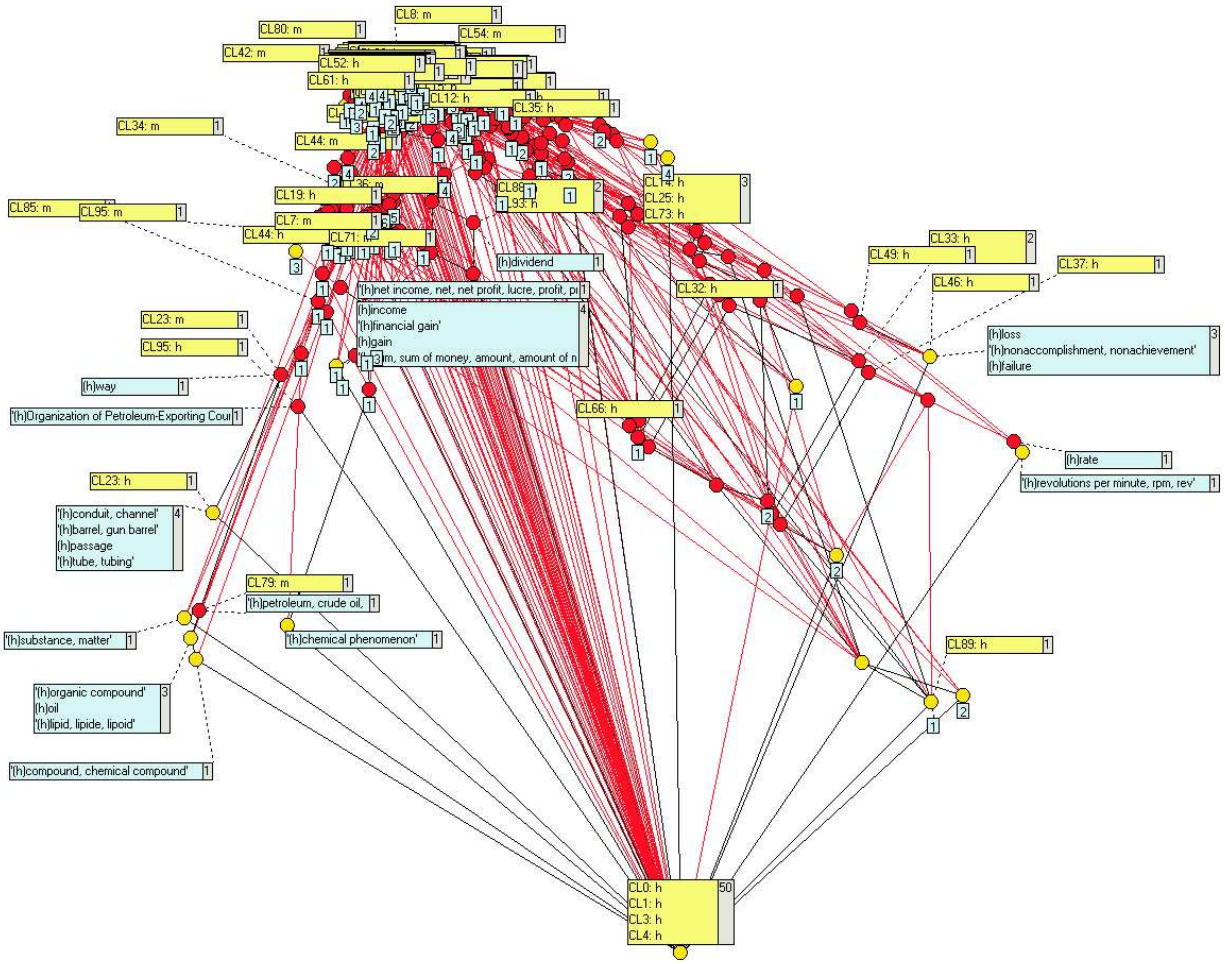

Abbildung 9.2: Vollständiger Begriffsverband der 100 Cluster des Datensatzes PRC_{30}; 3 Ketten sind zu erkennen

Referenzalgorithmus, der als Demo mit dem Java SDK geliefert wird.[1] Ziel des Algorithmus ist es, alle Cluster, die sich sehr ähnlich sind, räumlich nah beieinander zu platzieren. Ab einer gewissen Unähnlichkeit sollten die Cluster möglichst weit voneinander entfernt liegen. Um dieses Ziel zu erreichen, lassen wir uns nicht alle Ähnlichkeitsbeziehungen darstellen. Wir führen einen Schwellwert ein und lassen so nur Kanten mit hoher Ähnlichkeit zwischen den Clustern darstellen.

In Abbildung 9.3 wird eine Kante nur dann dargestellt, wenn sie über dem Schwellwert von 70 % der maximalen Ähnlichkeit liegt. Jedes Rechteck symbolisiert in der Abbildung einen Cluster. Im Rechteck ist die Clusternummer und die Reuters-Klasse[2] dargestellt. Geht von einem Rechteck keine Kante zu einem anderen Rechteck, so ist das Rechteck zu allen anderen Rechtecken relativ unähnlich. Diese Art der Cluster stehen in einem gewissen Sinne allein da, d.h. es gibt keinen anderen Cluster, der bzgl. des Schwellwertes diesem Cluster ähnlich ist. Existieren Kanten zwischen den Clustern, dann gibt die Länge der Kante die Unähnlichkeit zwischen den beiden Clustern an, d.h. bei einer Kantenlänge von Null sind die Zentroide identisch. Der Algorithmus versucht nun die

[1] http://java.sun.com/applets/jdk/1.0/demo/GraphLayout/

[2] Die Reuters-Klasse wurde nach dem Clustern zu Evaluierungszwecken vom Programm mit dargestellt. Dabei wird immer die Klasse angegeben, deren Dokumente im Cluster am häufigsten vorkommen. Es wäre auch denkbar z.B. beschreibende Terme hier darzustellen.

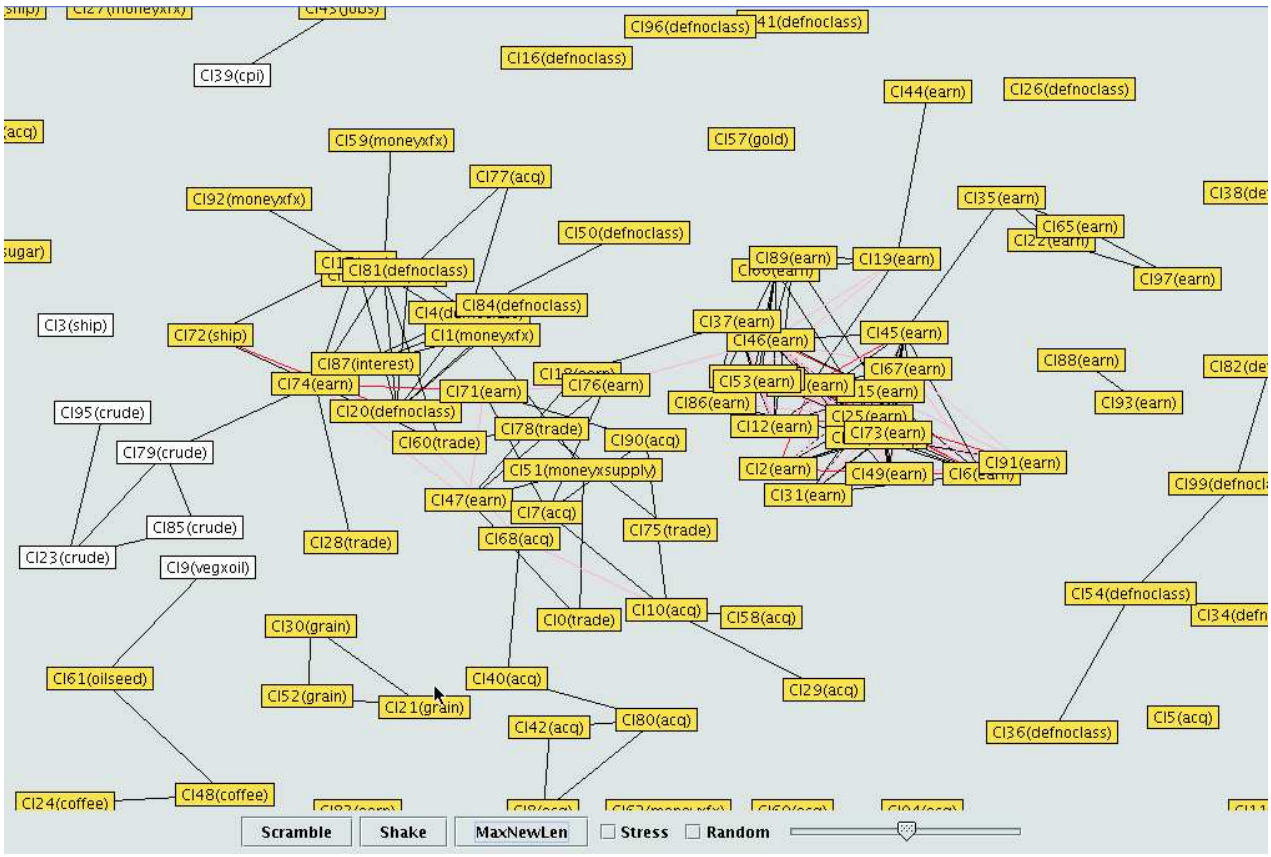

Abbildung 9.3: Die Abbildung zeigt die Ähnlichkeiten zwischen den Textclustern auf der Basis der Ähnlichkeit der Zentroide

Cluster in der Ebene so zu platzieren, dass der Abstand der Cluster ihrer Kantenlänge entspricht. Dies klappt unterschiedlich gut. Die schwarzen Linien symbolisieren eine korrekte Kantenlängen und die roten (hellen) Linien eine inkorrekte Kantenlängen. Wie weit die dargestellte Kantenlänge von der vorgegebenen abweicht, kann mit einer Zahl neben der Kante angegeben werden.

Kommen wir zur inhaltlichen Analyse von Abbildung 9.3. Rechts vom Zentrum erkennen wir eine große Gruppe von Clustern, die alle mit der Bezeichnung “earn” versehen sind. Die kurzen Strecken zwischen den Clustern zeigen eine hohe Ähnlichkeit an. Außerdem findet man nicht nur bilaterale Beziehungen zwischen den Clustern vor, sondern auch größere Gruppen von Clustern, die alle mit Kanten untereinander verbunden sind. Neben diesem Ballungszentrum findet man Ketten von Clustern, wobei zum Teil die Cluster nicht unmittelbar in inhaltlicher Beziehung zueinander stehen. Beispielsweise findet man in der linken unteren Ecke die Cluster CL24, CL48, CL61 und CL9, wobei Cluster 9, der von pflanzlichem Öl handelt, wenig mit Cluster 24 mit dem Thema Kaffee zu tun hat. Eine Visualisierung des Begriffsverbandes für diese Cluster macht dies schnell deutlich. Die Beziehungen der Cluster beruhen laut visualisiertem Verband auf den Termen “metric ton, MT, tonne, t” und “metric weight unit, weight unit”. Da dies sehr allgemeine Terme sind, ist eine weitere Analyse wahrscheinlich auf Basis der Dokumente notwendig.

Eine weitere interessante Clustergruppe stellen die Cluster CL23, CL79, CL85 und CL95 dar. Die Cluster sind durch kurze Strecken und im Kreis miteinander verbunden. Man erkennt in der Abbildung 9.3 (auf der linken Seite) die beschriebenen Eigenschaften dieser Clustergruppe. Die Clustergruppe entspricht dem Umfang des formalen Begriffs (***) im linken Teil der Abbildung 9.1. Anhand des Verbandes kann man als Thema dieser Cluster “crude oil” (Rohöl) identifizieren. Der Vergleich mit den Reuters-Labels der Clustergruppe bestätigt die Korrektheit der Verbandsanalyse.

Aus Abbildung 9.1 entnehmen wir weiterhin einen Zusammenhang der Clustergruppe mit den Clustern 3 und 9 über das Konzept “oil” und über “chemical compound” auch zu Cluster 39. Diese Cluster sind in Abbildung 9.3 (weiß) hervorgehoben. Die Unähnlichkeit der Cluster von der Clustergruppe ist so stark, dass bei der verwendeten Schranke von 70 % keine Kanten zwischen den Clustern existieren. Wir erkennen in Abbildung 9.1 eine kleine Überlappung an wichtigen Worten zwischen der Clustergruppe und den Clustern 3, 9 und 39, die aber nicht zu einer starken Ähnlichkeit zwischen den Zentroiden führt. Bei genauer Analyse der Texte stellt sich heraus, dass z.B. die Dokumente des Clusters 3 zur einen Hälfte der Klasse “crude” und zur anderen Hälfte der Klasse “ship” angehören. Die “crude”-Dokumente des Clusters 3 liefern die Begriffe, die zur Verbindung mit den anderen “crude”-Clustern betragen würden. Die “ship”-Dokumente des Clusters 3 führen hingegen bei der Ähnlichkeitsberechnung zu einer größeren Unähnlichkeit des Clusters 3 zu anderen “crude”-Clustern. Das erklärt die fehlende Kante zwischen Cluster 3 und den anderen Clustern über Rohöl in Abbildung 9.3 und die beobachtete Beziehung in Abbildung 9.1.

Zusammenfassung: Der vorgestellte Ähnlichkeitsgraf präsentiert die Beziehungen zwischen den Clustern übersichtlich und erlaubt es einfach Gruppen von Clustern zu identifizieren. Diese Gruppen können die Basis für eine begriffliche Analyse durch die Formale Begriffsanalyse bilden. Im Artikel [119] beschreiben wir, wie wir mit dieser Methode weitere interessante Clustergruppen identifizieren konnten.

Im folgenden Abschnitt gehen wir auf alternative Ansätze zur Berechnung von Beziehungen zwischen den Textclustern ein.

9.4 Alternative und verwandte Ansätze

Der folgende Abschnitt geht auf die Anwendung eines alternativen Ansatzes aus dem Bereich der Regellerner ein und stellt dafür erste Ergebnisse vor. Abschnitt 9.4.2 diskutiert die verwandten Ansätze zu den Methoden dieses Kapitels.

9.4.1 Alternative Ansätze

In diesem Abschnitt wollen wir auf alternative Ansätze zur Beschreibung und Präsentation von Textclustern eingehen. Ziel ist der Einsatz von Klassifikationsalgorithmen, wie Regellerner oder Entscheidungsbäume anstelle der Formalen Begriffsanalyse zur Beschreibung der berechneten KMeans-Textcluster. Für diese Aufgabe kommen nur Klassifikationsalgorithmen in Frage, die ihr Modell leicht interpretierbar ausgeben. Dies trifft z.B. auf den Regellerner C4.5 (vgl. [187]), Ripper (vgl. [39]) oder PART (vgl. [76]) zu. Alternativ bieten sich Entscheidungsbäume wie z.B. der C4.5 (vgl. [187]) an. Im Folgenden wollen wir nur kurz auf erste Ergebnisse eingehen. Weitere vertiefenden Analysen sind zur Überprüfung dieser Ergebnisse notwendig.

Wir nutzen den in Abschnitt 9.3.1 abgeleiteten mehrwertigen Kontext des PRC_{30}-Datensatzes mit 100 Clustern. D.h., der Datensatz enthält nur 100 Elemente. Jede Klasse enthält nur ein Beispiel. Zur Berechnung der Regeln setzen wir das Weka-System[3] und hier den PART-Algorithmus ein (vgl. [76]). Da wir an einer möglichst genauen Beschreibung des Datensatzes interessiert sind, verwenden wir die komplette Trainingsmenge zur Berechnung des Modells. Wir erhalten 98 Regeln (Cluster 17 und 74 enthalten keinen Term und Cluster 25 hat die gleichen Merkmale wie Cluster 73).

Für Cluster 52 lieferte das Verfahren folgende Regel:

```
wheat__corn = m AND
```

[3] http://www.cs.waikato.ac.nz/ml/weka/

```
crop__harvest = 0: CL52 (1.0)
```

Die angegebenen Terme sind Konzepte. “m” und “h” entsprechen den Werten aus Abschnitt 9.3.1 und “0” bedeutet, dass der Wert für dieses Merkmal im Termvektor kleiner als der Schwellwert θ_1 war. Die Regel besagt, dass es sich um den Cluster 52 handelt, wenn “wheat__corn” mit “m” im Datensatz auftritt und “crop__harvest” nicht vorkommt. Bei diesem Cluster geht es also um Weizen. Die Dokumente des Cluster sind laut Reuters-Label in der Mehrzahl über Korn und nicht über Weizen. Schaut man sich alle Terme des Clusters an, so findet man neben Weizen auch Korn.

Analysieren wir ein zweites Beispiel. Folgende Regel wurde für Cluster 79 berechnet:

```
metric_ton__MT__tonne__t = 0 AND
rate = 0 AND
living_quarters__quarters = 0 AND
petroleum__crude_oil__crude__coal_oil__rock_oil__fossil_oil = m AND
conduit__channel = 0: CL79 (1.0)
```

Offensichtlich geht es in diesem Cluster um “crude oil”. Dies Ergebnis deckt sich mit den Ergebnissen aus Abschnitt 9.3. Abbildung 9.1 entnehmen wir, dass die Dokumente der Cluster 23, 85, 95 auch über “crude oil” sind. Keine Regel der anderen Cluster enthält den Term “oil” oder “crude”. Wir sind daher nicht in der Lage, eine Verbindung der verschiedenen Cluster zueinander aus diesen Regeln abzuleiten.

Zusammenfassung: Mit Hilfe dieses einfachen Experimentes konnten wir zeigen, dass man beschreibende Regeln mit Regellernern ableiten kann. Der Regellerner liefert nach unseren Beobachtungen Regeln, die möglichst wenige alleinstellende Merkmale eines Clusters enthalten. Die Beziehungen zwischen den Clustern, wie sie die Formale Begriffsanalyse liefert, konnten daher mit Hilfe der abgeleiteten Regeln nicht entdeckt werden. Für die Bestätigung dieser ersten Ergebnisse sind weitere umfangreiche Experimente notwendig.

Uns ist bewusst, dass der verwendete Datensatz sehr klein und sehr wenige Beispiele enthält. Es handelt sich um einen eher untypischen Datensatz. Alternativ zum untersuchten Datensatz kann man den kompletten nicht diskretisierten PRC_{30}-Datensatz mit 12344 Elementen nutzen. In diesem Fall wird jedem Dokument sein Cluster zugeordnet und man erhält auf diesem Wege eine Klasseneinteilung. Auch auf diesen Datensatz könnte man einen Regellerner anwenden. Wir haben dies aber nicht weiter untersucht. Im nächsten Abschnitt gehen wir auf die verwandten Ansätze dieses Kapitels ein.

9.4.2 Verwandte Ansätze

Die in diesem Kapitel vorgestellte Visualisierung von ontologiebasierten Textclustern mittels der Formalen Begriffsanalyse besteht aus einer Menge von Teilschritten. Diese Schritte und ihre verwandten Ansätze werden in anderen Kapitel schon ausführlich diskutiert und sollen an dieser Stelle nicht noch einmal wiederholt werden. Man findet verwandte Ansätze zur Merkmalsextraktion in Abschnitt 4.5.3, zum ontologiebasierten Textclustern in Abschnitt 8.2.8 und zur Nutzung der Formale Begriffsanalyse zum Clustern von Textdokumenten in Abschnitt 8.5.4.

Verwandte Arbeiten, die wie die Formale Begriffsanalyse Mengen von Objekten anhand von symbolischen Merkmalen gruppieren, kommen aus dem Bereich des konzeptuellen Clusterns und werden in Abschnitt 5.6.8 diskutiert. Die Anwendung dieser Verfahren zur Beschreibung von Textclustern ist eine offene Forschungsfrage.

Arbeiten zum Thema Visualisieren von Textclustern, die ebenfalls den in diesem Kapitel vorgestellten Methoden ähnlich sind, findet man im Bereich der Self Organizing Maps (vgl. [143]), die in Abschnitt 5.6.3 eingeführt werden. Eines der bekannten Projekte ist das WEBSOM-Projekt.[4] Denkbar wäre eine SOM-basierte Visualisierung der in diesem Abschnitt verwendeten ontologiebasierten Dokumentrepräsentation. In wieweit die Ergebnisse dieser Visualisierung mit den FBA-Visualisierungen vergleichbar sind, müssen zukünftige Experimente noch zeigen.

[4] http://websom.hut.fi/websom/

Teil III

Anwendung

10 Anwendungen des Subjektiven Clusterns

In diesem Teil der Arbeit beschäftigen wir uns mit den Anwendungsgebieten der in der Arbeit entwickelten Methoden. Wir gliedern die Ausführungen in zwei Teile: Die Anwendung des "*Subjektiven Clusterns*" und des "*Clusterns und Visualisierens mit Hintergrundwissen*" und beziehen die Ergebnisse auf die in Kapitel 2 eingeführten Fragestellungen und Datensätze. Die Ergebnisse für den Reuters-Korpus wurden schon ausführlich während der Vorstellung der Methoden im Teil II diskutiert und werden an dieser Stelle nicht wiederholt.

In diesem Kapitel gehen wir in Abschnitt 10.1 auf die Ergebnisse bei der Anwendung des Subjektiven Clusterns auf den Telekomdatensatz ein. Abschnitt 10.2 beschreibt die Architektur eines Wissensportals, wobei eine Komponente dieses Portals das Subjektive Clustern zur Strukturierung von Webseiten einsetzt.

10.1 Subjektives Clustern von Kommunikationsdaten

Im ersten Teil dieses Kapitels wollen wir die Anwendung des Subjektiven Clusterns auf Kommunikationsdatensätze vorstellen. Ziel ist es dabei, Kundencluster anhand des Kommunikationsverhaltens der Kunden zu berechnen. Die Kommunikationsdatensätze stellt die Deutschen Telekom AG zur Verfügung. Die abgeleiteten Datensätze sind Thema von Abschnitt 2.5. Der folgende Abschnitt gliedert sich wie folgt:

Nach Einführung der Aufgabenstellung in Abschnitt 10.1.1 und der Diskussion praktischer Probleme bei der Vorverarbeitung von großen Datenmengen untersuchen wir vorverarbeitete Kommunikationsdaten nach typischen Phänomenen des hochdimensionalen Raumes in Abschnitt 10.1.3. Wir leiten aus den Ergebnissen dieser Analyse die Forderung nach einfachen, verständlichen und benutzerbezogenen Clusterlösungen ab, die in Abschnitt 10.1.4 beschrieben werden. Sie motivieren den Einsatz des Subjektiven Clusterns aus Kapitel 7 in einer erweiterten Version auf den Kommunikationsdaten während der Vorverarbeitungsphase. Die Ergebnisse werden in Abschnitt 10.1.5 präsentiert. Wir folgen bei den Betrachtungen den Arbeiten [111, 112, 158].

10.1.1 Einleitung

Die Deutsche Telekom AG besitzt als größter Telekommunikationsanbieter in Deutschland auch das größte deutsche Festnetz. Bei 35 Millionen Kunden findet man unterschiedlichste Kunden mit unterschiedlichsten Kommunikationsbedürfnissen. Um die Kunden zufrieden zu stellen und an das Unternehmen zu binden, muss die Deutsche Telekom AG ihre Kunden und deren Kommunikationsverhalten analysieren und verstehen. Die Einblicke in das Kommunikationsverhalten der Kunden erlauben es der Deutschen Telekom AG, die Tarife bedarfsgerecht zu erstellen und Kunden an sich zu binden. Auch der Unternehmenserfolg kann so gesichert werden. Um dieses Ziel zu erreichen und effizient zu arbeiten, kann die Deutsche Telekom AG nicht jeden Kunden direkt ansprechen,

sondern versucht die Kunden in Gruppen (Cluster oder Segmente) mit gleichem Kommunikationsverhalten einzuteilen. Diesen Gruppen kann sie gezielt neue maßgeschneiderte Tarife anbieten. Um optionale oder Spezialtarife, die gezielt auf ausgewählte Kundengruppen zugeschnitten sind, anbieten zu können, muss das Kommunikationsverhalten dieser Gruppe für den Anwender bei der Telekom verständlich bzw. *interpretierbar* sein.

Um das Kommunikationsverhalten der Kunden beurteilen zu können, muss man dieses geeignet repräsentieren. Man versteht unter dem Kommunikationsverhalten eines Kunden die Menge an Gesprächen dieses Kunden, die er in einem bestimmten Zeitraum geführt hat. Sehr schnell wird verständlich, dass der Vergleich zweier Kunden anhand der geführten Gespräche sehr schwierig wird. Typischerweise wird man nur sehr selten zwei Kunden finden, die auch nur ein Gespräch genau zur gleichen Zeit geführt haben. Noch schwieriger wird es, wenn das Gespräch auch noch gleich lang gewesen sein soll. Weitere Merkmale verschärfen das Problem weiter und machen das Zusammenfassen von einzelnen Gesprächen zur Analyse des Kundenverhaltens notwendig. Eine zentrale Frage ist die Bestimmung des richtigen Aggregationsniveaus bzw. des richtigen Maßes für die Ähnlichkeit von Gesprächen. In einem Vorverarbeitungsschritt, der in Abschnitt 4.3 beschrieben ist, überführt man die Kommunikationsdaten in kundenbeschreibende Merkmale. Dieser Schritt ist sehr aufwendig und erfordert für die von der Telekom zur Verfügung gestellten Daten relativ viele Ressourcen.

Um die kundenbezogenen Kommunikationseigenschaften zu berechnen, müssen pro Monat ca. 130 GB Rohdaten ausgewertet werden. Bei der Telekom standen zehn handelsübliche PC's mit 500 MHz, 384 MB Hauptspeicher und 60 GB Festplattenplatz zur Auswertung und Analyse der Daten zur Verfügung. Dieses verteilte System besteht aus preisgünstigen Einheiten, die gemeinsam eingesetzt sehr leistungsfähig sind. Man nennt solche Systeme auch "shared nothing" Systeme. Sie sind in aller Regel wesentlich günstiger als äquivalente Hochleistungssysteme, verlangen aber mehr Aufwand bei der Administration. Um ein verteiltes System nutzen zu können, muss die zu lösende Aufgabe auch verteilbar (parallelisierbar) sein, d.h. es muss kleine Teilaufgaben geben, die ein Rechner unabhängig von jedem anderem Rechner lösen kann und die einen Teil zum Gesamtergebnis beitragen. Außerdem steht man vor der Aufgabe, dass der Kommunikationsaufwand zwischen den Rechnern möglichst gering sein muss.

Zwei Berechnungen müssen auf diesem System erfolgen. Als erstes sind die kundenbeschreibenden Merkmale abzuleiten und anschließend müssen die Kunden anhand dieser Merkmale geclustert werden. Der erste Schritt lässt sich leicht mit Hilfe einer verteilten Datenbank auf einem solchen System durchführen. Die Berechnung aller Vorverarbeitungsschritte erfolgen unter Nutzung einer verteilten DB2-Datenbank. Um einen Eindruck vom Aufwand dieser Datenverarbeitung zu vermitteln, geben wir im nächsten Abschnitt eine kurze Beschreibung des Aufwands für den Reverse-Pivoting-Schritt.

10.1.2 Merkmalsberechnung in der Praxis

Für die folgenden Ergebnisse wurden alle Gespräche der Kunden der 10 % Stichprobe (vgl. Abschnitt 2.5) für den Januar 2000 betrachtet. Es standen ca. 130 GB Rohdaten des Monats Januar zur Verfügung, die ca. 500 Mill. Datensätze für ca. 3,5 Mill. Kunden umfassen. Die Tabelle mit den Kommunikationsdaten enthält neben einer Kundennummer auch Informationen über die Tarifzone, Tagart usw. Weiterhin findet man zu jedem Gespräch die Gesprächsdauer. Das prinzipielle Vorgehen zum Ableiten eines 76-dimensionalen kundenbeschreibenden Datensatzes ist in Abschnitt 4.3 beschrieben. Für die folgenden Betrachtungen wollen wir für die Daten des genannten Zeitraums genau einen solchen 76-dimensionalen Datensatz aus den Kommunikationsdaten ableiten. Die Rohdaten werden dazu in die verteilte DB2 geladen.

Mit Hilfe von SQL-Befehlen kann man den Reverse-Pivoting-Schritt auf unterschiedliche Art und Weise implementieren. Leider sind nicht alle Varianten gleich performant und bedingen unterschiedlichen viel temporären Plattenplatz. Die schnellste Variante bildet in einem ersten Schritt die Summe für jedes Merkmal mit Hilfe der "group by" Klausel. Diese resultierende Tabelle enthält dann sechs Spalten – Kundennummer, die vier Dimensionen (Tarifzone, Tagart, Uhrzeit und Verbindungsnetzbetreiber) und die Summe der Verbindungsminuten (das Aggregat) – und muss in einem zweiten Schritt noch auf die gewünschte Form (jedes Merkmal eine Spalte) transferiert werden. Mit Hilfe einer case Anweisung lässt sich dies realisieren. Das Auslesen der Daten ohne Reverse-Pivoting-Schritt aus der Datenbank direkt in das Clusterprogramm wäre ebenfalls denkbar. Das Anwendungsprogramm würde diesen Schritt dann automatisch während des Lesens der Daten vollziehen. Alles in allem benötigen zehn handelsübliche PC's für den Reverse-Pivoting-Schritt ca. 5-6 Stunden. Dabei reduziert sich die Datenmenge von 130 GB auf ca. 2,5 GB. Eine weitere Variante, die ca. 41 Stunden für die gleiche Datenmenge benötigt, berechnet in einem ersten Schritt 76 Tabellen und fügt diese zum Schluss zu einer zusammen. Im Ergebnis liefern beide Varianten aus den Kommunikationsdaten die beschreibenden Kundenmerkmale, die zum Clustern der Kunden verwendet werden können.

Die Auswahl von 76 Dimensionen für unseren Datensatz war bisher "willkürlich" (vgl. Abschnitt 4.3.2). Dieser 76-dimensionale Datensatz weist für das Clustern ungünstige Eigenschaften auf. Bevor wir uns im Folgenden mit der Auswahl von "sinnvollen" Merkmalen zum Clustern und Beschreiben von Kunden beschäftigen, analysieren wir im folgenden Abschnitt das Problem der gewählten "hochdimensionalen" Repräsentation des Datensatzes. Aus der Forderung nach Interpretierbarkeit der Clusterergebnisse, die bei niedrigdimensionalen Datensätzen gegeben ist, sowie wegen der besseren Eigenschaften der Repräsentation von niedrigdimensionalen Datensätzen für das Clustern leiten wir die Notwendigkeit zur Reduktion der Merkmalsanzahl auf eine verständliche und anwendbare Anzahl ab. Lösungen präsentieren wir in Abschnitt 10.1.4.

10.1.3 Hohe Dimensionalität bei Kommunikationsdaten

Der folgende Abschnitt führt in die Probleme eines hochdimensionalen Datensatzes ein. Abschnitt 10.1.3.2 prüft darüber hinaus, ob der aus den Kommunikationsmerkmalen abgeleitete Datensatz die Eigenschaften eines hochdimenionalen Datensatzes aufweist.

10.1.3.1 Phänomene des hochdimensionalen Raumes

In [158] werden Clusterergebnisse präsentiert, basierend auf der gleichen Repräsentation wie sie in Abschnitt 4.3 eingeführt und in Abschnitt 10.1.2 praktisch berechnet werden. Obwohl die Ergebnisse sehr vielversprechend sind, bleiben einige Fragen ungeklärt, zum Beispiel die Frage nach der automatischen Bestimmung der Clusteranzahl eines Datensatz. Kein bekanntes Maß [226] bestätigte zuverlässig bei den unterschiedlichen Clusterläufen die Anzahl der errechneten Cluster. Eine Erklärung war, dass die Anzahl der Cluster deutlich höher liegen musste als bei den Clusterläufen getestet. Da aber schon 100 Cluster eine zu große Anzahl für die Referenten bei der Telekom darstellt, wurde dazu übergegangen, dass die Referenten der Telekom die Anzahl der Cluster vorgeben. Diese Lösung ist zwar praktikabel, liefert aber nicht den Grund, warum die Anzahl der Cluster nicht bestimmt werden konnte. Die folgenden Betrachtungen werden zeigen, dass der erzeugte "hochdimensionale" Raum zur Beschreibung der Telekomkunden (siehe Abschnitt 10.1.2) die Ursache dieses Problems darstellt.

Das Phänomen des hochdimensionalen Raumes ([25] und [106]) lässt sich anhand der Abbildung 10.1 anschaulich erläutern. Dabei geht es um den minimalen und maximalen Abstand eines

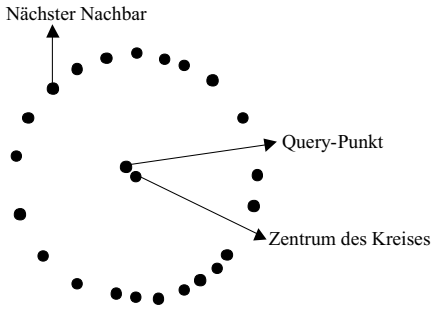

Abbildung 10.1: Anfragepunkt (Query Punkt) und sein nächster Nachbar

Punktes im Raum zu den anderen Punkten. Die zur Berechnung des Abstandes verwendete Metrik spielt dafür eine wichtige Rolle. Wir erläutern die Problematik des nächsten und weitesten Nachbarn unter Nutzung der Euklid-Metrik. In Abbildung 10.1 ist eine Punktewolke so angeordnet, dass der Ausgangspunkt oder besser Query-Punkt Q nahe dem Zentrum der Wolke liegt und der nächste Nachbar ein Punkt des dargestellten Kreises ist. Es ist nun leicht, den nächsten Nachbarn zu bestimmen (markierter Punkt in der Abbildung 10.1). Der Unterschied in der Entfernung zwischen dem nächsten Nachbarn und dem weitesten Nachbarn ist aber sehr gering. Versetzt man den Anfragepunkt nur ein klein wenig, so wird sich ein neuer nächster Nachbar ergeben. Die Aussagekraft des nächsten Nachbarn für eine Aufgabe wie das Clustern ist an dieser Stelle eher gering, da im Schnitt alle Punkte ungefähr gleich weit entfernt sind. Auch die Bedeutung des nächsten Nachbarn sinkt, da es sehr viele Punkte gibt, die in unmittelbarer Umgebung zum nächsten Nachbarn liegen. Man kann vom Standpunkt Q aus gesehen, die Punkte kaum unterscheiden, geschweige denn einen sinnvollen nächsten Nachbarn bestimmen.

Im hochdimensionalen Raum kann man ein analoges Phänomen nachweisen. Bei steigender Dimensionsanzahl m wird das Verhältnis zwischen der Distanz zum entferntesten Nachbarn ($dist_{max}$) und die Distanz zum nächsten Nachbarn ($dist_{min}$) immer kleiner und geht bei unendlich großer Dimensionsanzahl m gegen 1.

[25] zeigen, dass für $m \to \infty$ unter der Voraussetzung:

$$\lim_{m\to\infty} var\left(\frac{(dist_m(P_m, Q_m))^r}{E[(dist_m(P_m, Q_m))^r]}\right) = 0 \tag{10.1}$$

folgendes erfüllt ist:

$$\frac{dist_{max_m} - dist_{min_m}}{dist_{min_m}} \to_r 0 \tag{10.2}$$

Dabei ist $0 < r < \infty$ ein Konstante und P_m und Q_m sind zwei Punkte im m-dimensionalen Raum. Wählt man zur Berechnung der Distanz $dist$ die Minkowski-Metrik L_r (siehe Abschnitt 5.2.1), so wird r der Parameter von L_r. Weitere Details findet man in [25].

Bei der Betrachtung von Gleichung 10.1 fällt auf, dass der Abstand $dist_m$ bezüglich einer Metrik zwischen den Punkten P und Q berechnet wird. Auf die Rolle der Metrik wird auch in [106] eingegangen. [106] weisen nach, dass ab der L3-Metrik die Aussage von Gleichung 10.2 noch verschärft werden kann. Dann konvergiert $dist_{max} - dist_{min}$ gegen 0. Für die L1- und L2- Metrik konvergiert die Differenz gegen $C_1 * \sqrt{m}$ bzw. eine Konstante C_2.

Wir entnehmen den Ausführungen, dass es Fälle gibt, in denen auch im hochdimensionalen Raum Clusterstrukturen entdeckt werden können. Vor jedem Clusterlauf ist für einen gegebenen Datensatz eine Prüfung der Voraussetzung notwendig. Erst wenn diese nicht erfüllt wird, besteht überhaupt

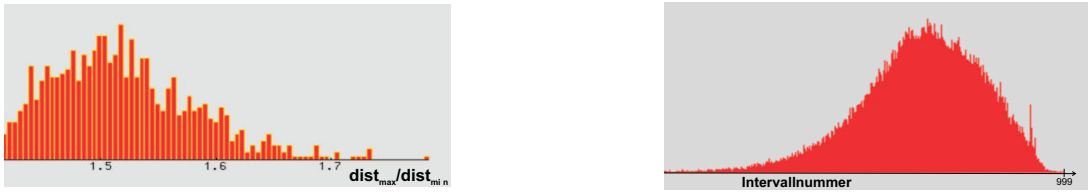

Abbildung 10.2: a)Häufigkeitsverteilung des Quotienten zwischen $dist_{max}$ und $dist_{min}$ für 76-dim. Datensatz, b) Häufigkeitsverteilung mit 1000 Intervallen, Entfernung zwischen einem beliebigen Punkt und allen Punkten des 76-dimensionalen Datensatzes

eine Chance, mit distanzbasierten Clusterverfahren Cluster zu bestimmen. Ansonsten sind bei hoher Anzahl von Dimensionen alle Punkte ungefähr gleich weit entfernt.

Für praktische Anwendungen muss noch geklärt werden, was eine hohe Anzahl an Dimensionen ist, da Gleichung 10.2 nur Aussagen für das Unendliche macht. Dazu haben [25] ebenfalls Untersuchungen durchgeführt. Sie stellten fest, dass schon bei 15 bis 20 Dimensionen das Phänomen beobachtet werden kann und das Verhältnis zwischen $dist_{max}$ und $dist_{min}$ gegen 1 strebt.

Um Cluster anhand von Kommunikationsdaten berechnen zu können, müssen wir in einem ersten Schritt den abgeleiteten Datensatz auf das Vorliegen des Phänomens des hochdimensionalen Raumes prüfen. Die Ergebnisse gibt der nächsten Abschnitt wieder.

10.1.3.2 Ergebnis für den 76-dimensionalen Telekomdatensatz

Nach den ersten Clusterergebnissen für die Telekomdatensätze stand die Vermutung im Raum, dass bei dem 76-dimensionalen Datensatz (vgl. 4.3) das beschriebene Phänomen des hochdimensionalen Raumes beobachtet werden kann. In Anlehnung an [25] wurden zwei verschiedene Tests auf dem 76-dimensionalen Datensatz durchgeführt. Da die Dimensionalität auf realen Daten nicht ohne weiteres variiert werden kann, wurden folgende Hilfsgrößen bestimmt, die Indikatoren für das Phänomen darstellen. Als erstes wurde für einen beliebig gewählten Anfragepunkt Q der Quotient zwischen dem entferntesten und dem nächsten Nachbarn auf der Basis der L_2-Metrik bestimmt:

$$\frac{dist_{max}}{dist_{min}} \tag{10.3}$$

Je näher dieser Quotient an 1 heranreicht, um so geringer ist der Unterschied der beiden Werte $dist_{max}$ und $dist_{min}$.

Abbildung 10.2a zeigt die Häufigkeitsverteilung für $\frac{dist_{max}}{dist_{min}}$ des 76 dimensionalen Datensatzes bei der Telekom für 1000 zufällig ausgewählte Punkte im Raum. Wie zu erkennen ist, wird dieser Wert nicht größer als 2. Dies entspricht den Ergebnissen, die Beyer in [25] als Problem beschreibt. Intuitiv sollte der Wert des Quotienten deutlich größer sein, z.B. 10000 und mehr.

Ein weiterer interessanter Test für einen realen Datensatz ist die Frage nach der Entfernung eines jeden Punktes von einem beliebig gewählten Anfragepunkt. Häufen sich die Punkte in einer bestimmten Entfernung, so ist kaum ein Unterschied zwischen $dist_{max}$ und $dist_{min}$ auszumachen. Ist der Unterschied des nächsten und entferntesten Punktes aber sehr gering, so ist auch die Bedeutung des nächsten Nachbarn, als Punkt zum Cluster gehörig, fraglich. Existiert dagegen in unmittelbarer Umgebung eine Menge von Punkten, die dicht beieinander liegen, und etwas entfernt wieder eine solche Menge, dann gewinnt das nächste Nachbar-Konzept an Bedeutung und man kann auf diesem Datensatz Cluster entdecken.

Abbildung 10.2b zeigt die Häufigkeitsverteilung des 76-dimensionalen Datensatzes der Telekom.

Dafür wurde die Entfernung zwischen einem beliebig gewählten Anfragepunkt und den Punkten des Datensatzes berechnet. Man sieht nur *einen* Häufungspunkt, was darauf schließen lässt, dass alle Punkte ungefähr im gleichen Abstand zum Anfragepunkt liegen. Wählt man einen anderen Anfragepunkt und wiederholt das Experiment auf diesem Datensatz, so ergibt sich das gleiche Bild, obwohl die Abbildung sehr stark vom Anfragepunkt abhängt (jeder Punkt sollte von seiner Position aus alle übrigen Punkte in unterschiedlicher Entfernung "sehen"). Ein deutlich besser geeigneter Datensatz liegt Abbildung 10.3b zu Grunde. Hier erkennt man deutlich die Häufung von Punkten in verschiedenen Abständen. Wir gehen auf diesen Datensatz in Kapitel 10.1.4 genauer ein.

10.1.4 Lösungen für Clustern im hochdimensionalen Raum

Wir entnehmen dem letzten Abschnitt, dass die von uns gewählte Repräsentation des Kommunikationsverhaltens der Telekomkunden zum Clustern in der vorliegenden Form nur bedingt geeignet ist. Der folgende Abschnitt analysiert diese Ergebnisse. In den Abschnitten 10.1.4.2 und 10.1.4.3 gehen wir auf verschiedene Formen der Dimensionsreduktion eines hochdimensionalen Datensatzes ein.

10.1.4.1 Analyse der Ergebnisse aus Abschnitt 10.1.3.2

Es stellt sich nun die Frage, wie die Clusterergebnisse aus [158] vor diesem neuen Hintergrund zu bewerten sind. Was bedeutet die Erkenntnis, dass alle Kunden bei der von uns gewählten Repräsentation sich annähernd gleich ähnlich/unähnlich sind, für eine Clusterung mit einem K-Means-Algorithmus?

Ziel der Clusterung bei der Telekom ist es, die Kunden in Cluster einzuteilen, die jeweils möglichst ein einheitliches Verhalten aufweisen. Wenn ein großer Teil der Kunden sich ungefähr gleich verhält und dieser Teil sich kaum von allen anderen Kunden unterscheidet, wird es schwer, Kunden überhaupt nach solchen Kriterien in Gruppen einzuteilen. K-Means wird immer eine Clusterung liefern, auch für den 76-dimensionalen Telekomdatensatz. Bei K-Means wird jede Clusterung auf dem gleichen Datensatz durch die Abhängigkeit des Verfahrens vom Startwert anders aussehen, d.h. es werden jedes Mal neue Gruppen gebildet. Die Kunden einer Gruppe sind sich durchaus ähnlich und unterscheiden sich von den Kunden der anderen Gruppen, aber nicht so gravierend, dass man eine inhärente Struktur entdecken könnte, die bei jedem Clusterlauf wiedergefunden werden würde. Damit erklären sich auch die durchaus plausiblen Clusterergebnisse. Die Clusterung ist nicht falsch. Die Repräsentation liefert aber keine Kriterien, nach denen man die Kunden gut in Gruppen einteilen kann.

Betrachtet man noch einmal die Kunden der Deutschen Telekom AG und stellt sich die Aufgabe, anhand des Kommunikationsverhaltens die Kunden in Gruppen einzuteilen, so kommt man als Mensch zu dem Schluss, dass die Kunden durchaus nicht alle gleich sind und eigentlich gravierende Unterschiede zwischen verschiedenen Kunden existieren müssten. Zum Beispiel enthält der verwendete Referenzdatensatz neben Privatkunden auch Geschäftskunden. Beide Kundengruppen unterscheiden sich zum Teil extrem. Nimmt man sich nur die Tageszeit, so wird der Teil der Privatkunden, der arbeitet, tagsüber nicht bzw. wenig telefonieren, die Geschäftskunden werden tagsüber sehr viel telefonieren. Diese Information ist im Datensatz enthalten und sollte eigentlich zum Unterscheiden der beiden Gruppen nutzbar sein. Für einen Menschen ist dies auf jeden Fall möglich. Durch die ungeschickte Repräsentation der Daten scheinen diese Informationen aber verloren zu gehen. Findet man einen Weg, die versteckten Informationen wieder nutzbar zu machen oder hervorzuheben, sollte auch eine Clusterung möglich sein. Mit den Wegen, mittels derer sich in den vorhandene Kommunikationsdaten doch noch Clusterstrukturen finden lassen, befassen sich die nächsten beiden Abschnitte.

Abbildung 10.3: a)Häufigkeitsverteilung des Quotienten zwischen $dist_{max}$ und $dist_{min}$ für 7-dim. Datensatz, b) Häufigkeitsverteilung mit 1000 Intervallen, Entferung zwischen einem beliebigen Punkt und allen Punkten des 7-dimensionalen Datensatzes

10.1.4.2 Reduktion der Dimensionsanzahl

Eine einfache Idee ist, die Anzahl der Dimensionen zu reduzieren. Verfahren aus der klassischen Statistik, wie die Hauptkomponentenanalyse, können dafür leider nicht eingesetzt werden. Wie in [112] gezeigt, würde man so sehr viel Informationen verschenken. Eine wesentlich einfachere Methode ergibt sich aus der Merkmalsgenerierung. Abbildung 4.1 zeigt alle Dimensionen, die in die Generierung der Merkmale eingeflossen sind. Nutzt man nicht alle Dimensionen gleichzeitig, sondern z.B. nur die Tarifzone, so ergeben sich sieben Merkmale.

Vergleichen wir die Häufigkeitsverteilungen aus Abbildung 10.2 mit den Häufigkeitsverteilungen für diese sieben Merkmale in Abbildung 10.3. Man findet unter a) wieder den Quotienten zwischen $dist_{max}$ und $dist_{min}$ für den siebendimensionalen Datensatz (Tarifzone). Man erkennt den deutlichen Unterschied zur Abbildung 10.2. Sowohl der mittlere Quotient als auch der maximale Quotient sind deutlich größer als für den 76-dimensionalen Datensatz. Auch die Verteilung der Entfernung zwischen einem Anfragepunkt und dem Rest des Datensatzes zeigt ein anderes Bild. Man erkennt die Gebiete mit größerer Dichte in unterschiedlicher Entfernung. Diese erlauben die Annahme, dass in unterschiedlicher Entfernung zum Anfragepunkt mehrere Gebiete mit höherer Konzentration der Punkte vorhanden sind. Die Bedeutung des nächsten und weitesten Nachbarn ist aus diesem Grund gegenüber der 76-dimensionalen Repräsentation deutlich gestiegen.

Alternativ könnte man statt eines siebendimensionalen Datensatzes auch die restlichen 12 Merkmale, also Tageszeit und Tagart und Verbindungsnetzbetreiber zur Dimensionsreduktion wählen. Beide Datensätze wurden in [112] erfolgreich mittels OPTICS [11] (Kapitel 5.6.7) geclustert. Man erkennt neben den deutlichen Strukturen, die auf Cluster schließen lassen, auch Unterschiede zwischen den Strukturen der beiden Clusterungen. Durch die unterschiedliche Vorverarbeitung wurden verschiedene wesentliche Eigenschaften aus den Daten hervorgehoben. Die Clusterung gruppiert die Kunden nach diesen Eigenschaften unterschiedlich und liefert so auch ein anderes Bild bzw. andere Gruppen, die unabhängig voneinander sind.

Die verschiedenen Sichten auf die Kunden der Telekom (vgl. Abschnitt 7.1.2 zum Begriff Sicht), die im Prinzip nur unterschiedliche Aggregate darstellen, führen zu der Frage, welches die “richtige” Sicht oder die “beste” Sicht auf die Kunden ist. Die Frage sei vor dem Hintergrund gestellt, neue Preisangebote zu machen und die Kunden vorher besser verstehen zu wollen. Das bedeutet aber eigentlich, dass alle Facetten des Kunden beleuchtet werden müssten, also jede Sicht auf den Kunden wichtig ist.

Um die Frage besser verstehen zu können und auch das vorhandene Wissen bei der Telekom zu nutzen (bzw. von den Referenten und Analysten der Telekom zu akquirieren), wurden in einem ersten Schritt zwei Experten der Telekom befragt. Mit Hilfe der Experten sollte auf der einen Seite die Frage geklärt werden, ob und welche Sichten existieren bzw. neue und interessante Sichten akquiriert werden. Auf der anderen Seite stand die Frage nach einer Gewichtung der unterschiedlichen

Sichten auf die Kunden zur Diskussion.

10.1.4.3 Nutzung von Expertenwissen zur Auswahl von Dimensionen

Für die Auswahl der richtigen Sicht auf die Kunden wurden zwei Personen (wir nennen sie im Folgenden Person A und Person B) der Deutschen Telekom AG befragt. Die unterschiedlichen Erfahrungen beider ergab ein differenziertes Bild zur Priorisierung von Merkmalen und der Beschreibung von Kunden der Deutschen Telekom. Im Folgenden werden die Antworten zusammengetragen und dann bewertet.

Ergebnisse der Expertenbefragung Person A nimmt als erstes eine Trennung der Kunden nach (Privatkunden) PK und (Geschäftskunden) GK vor und sieht sich dann folgende Merkmale an:

- monatlicher Gesamtumsatz in DM für Anschluss und Verbindungen
- monatlicher Gesamtumsatz in Minuten
- Tarifart optional, Standard oder Preselection
- Verbindungsnetzbetreiber (Anteile)
- Nutzungszeit (Mo-Fr, 9-18 Uhr. . .)
- Tarifzone

Person B wählt eine andere Herangehensweise. Ihre erste Frage galt der zu untersuchenden Größe. Dabei steht neben der Summe der Verbindungsminuten, die in dieser Arbeit immer Zielgröße ist, die Anzahl Verbindungen, Umsatz, Preiselastizität zur Auswahl. Alle Größen können für jeden Kunden berechnet und zur Clusterung herangezogen werden. Beispielhaft und um die Vergleichbarkeit zu waren, wurde das weitere Vorgehen anhand der Summe der Verbindungsminuten besprochen. Wichtige Merkmale für Person B sind Merkmale mit einer hohen Varianz. Folgende Merkmale zählen dazu:

- Trennung PK/GK
- Tarifzone
- Tageszeit (“Tagesverkehrskurve”)
- Ortsnetzgröße

Person B wies darauf hin, dass eine Trennung nach PK und GK sehr wahrscheinlich notwendig ist. Sowohl die Tarifzone als auch die Tagesverkehrskurve sind wichtige Merkmale. Ihre Bedeutung und Nutzung hängt aber von der durchzuführenden Analyse ab und kann unabhängig davon nur schwer angegeben werden.

Bei der Befragung der beiden Personen kristallisierten sich einige wenige Merkmale wie PK/GK heraus, die unbedingt zu beachten sind. Eine Trennung der Kunden a priori in diese beiden Gruppen scheint notwendig. Die anschließende Clusterung muss für jede Gruppe separat durchgeführt werden und führt zu unterschiedlichen Ergebnissen. Bei der Wahl weiterer Dimensionen bzw. auch der Zielgröße konnte keine eindeutige Priorisierung angegeben werden. Person A fasst dies folgendermaßen zusammen: “Allerdings würde ich die Reihenfolge der Merkmale nicht starr festhalten, sondern von der Aufgabe abhängig machen.“ Diese Aussage macht deutlich, dass das dynamische Erzeugen von Datensätzen mit unterschiedlichem Blickwinkel auf die Kunden einen interessanten Ansatz zum Clustern der Kunden darstellt. Die Auswahl der Sichten muss in Zusammenarbeit mit dem Referenten bei der Telekom erfolgen. Notwendig dafür sind Methoden und Werkzeuge, die eine schnelle und effiziente Generierung der gewünschten Kommunikationsmerkmale erlauben. Dazu sollte das gesammelte Domänenwissen formalisiert und so genutzt werden.

Der folgende Abschnitt beschäftigt sich mit der Akquisition einer Domänen-Ontologie zur Beschreibung von Kommunikationsdaten bei der Telekom. Weiterhin berechnen wir eine Menge von Clusterungen auf der Basis von automatisch durch COSA (siehe Abschnitt 7.2) generierten Merkmalen. COSA nutzt die vorher erstellte Ontologie.

10.1.5 Ergebnisse von COSA auf Kommunikationsdaten

Die diskutierten Phänomene eines hochdimensionalen Datensatzen stellen die Motivation für unseren Ansatz des "Subjektiven Clusterns" dar. Wir erreichen damit neben einer verbesserten Verständlichkeit auch eine gesteigerte Clustergüte. In einem ersten Schritt müssen wir für die Gesprächsdaten eine Ontologie akquirieren (Abschnitt 10.1.5.1). Wir gehen dann spezielle Eigenschaften unseres Kommunikationsdatensatzes in Bezug auf die Anwendung des COSA-Algorithmus in Abschnitt 10.1.5.2 ein und leiten daraus die Erweiterungen für COSA ab. Anschließend präsentieren wir Ergebnisse für einen ausgewählten Kommunikationsdatensatz in Abschnitt 10.1.5.3.

10.1.5.1 Akquisition einer Telekom-Ontologie

Um die erweiterte Version von COSA einsetzen zu können, benötigen wir eine Domänenontologie, die zu den Kommunikationsdatensätzen bei der Telekom passt. In der Literatur werden unterschiedliche Modelle zur Akquisition von Ontologien vorgeschlagen (vgl. [214]). Wir benötigen neben den verwendeten Begrifflichkeiten/Konzepten und deren taxonomischer Beziehung auch eine Abbildung auf die Feldbezeichner der Datenbank. Zusätzlich müssen wir sinnvolle Größen zur Bewertung des Kundenverhaltens in der Ontologie speichern. Durch unsere speziellen Anforderungen an die Ontologie – wir benötigen ausschließlich Konzepte, die Taxonomie und eine Abbildung vom Konzept auf die Daten der Datenbank – vollzogen wir nicht den kompletten Akquisitionsprozess, sondern nur den relevanten Teil (vgl. [214]).

Wir entschieden uns für die Nutzung eines Fragebogens, den wir zusammen mit den Mitarbeitern der Telekom entwarfen. Dieser bietet den Vorteil nicht nur Konzepte und Beziehungen erfassen zu können, sondern er spiegelt auch die persönliche Perspektive der befragten Mitarbeiter (Domänenexperten und spätere Nutzer unserer Clusterung) auf die Daten wider. Auf diese Weise sind wir auch in der Lage, "anwenderbezogene" Sichten auf den Datensatz zu generieren.

Die letzte Version unseres Fragebogens ist in Anhang G wiedergeben. Er wurde mehrfach verändert und optimiert. Gleichzeitig wurde die Ontologie angepasst und erweitert. Neben der einleitenden Motivation ist der Fragebogen so aufgebaut, dass er möglichst verschiedene Bereiche der Tarifgestaltung anspricht. Auf diese Weise sollten alle Facetten der Tarifierung und damit auch alle typischen Verbindungstypen für Telefongespräche erfasst werden. Die befragten Mitarbeiter wurden angehalten, möglichst ausführlich zu antworten.

Die erste Version des Fragebogens wurde von Mitarbeitern, die sich vorrangig mit der Datenanalyse beschäftigen, ausgefüllt. Die Ergebnisse der Auswertung dieser Fragebögen flossen nun in die Optimierung selbiger und in eine erste Version der Domänenontologie ein. Nach einer zweiten Runde bei diesen Mitarbeitern wurde der Fragebogen an Referenten, die ihrerseits Tarife bei der Telekom gestalten, gesendet. Hiervon erhofften wir uns spezifischere Einblicke und Details. Aus diesen Information erstellten wir die Domänenontologie, die hauptsächlich aus Begriffen besteht. Ein Ausschnitt ist in Abbildung 10.4 und eine etwas ausführlichere Version in Abbildung G.1 im Anhang G zu finden.

Die aus den Fragebögen extrahierten Konzepte und deren Beziehung untereinander bilden die Domänenontologie. Wie schon angedeutet müssen diese Konzepte die Kommunikationsverbindungen der Kunden eindeutig beschreiben. Die Kommunikationsverbindungen stehen in einer Daten-

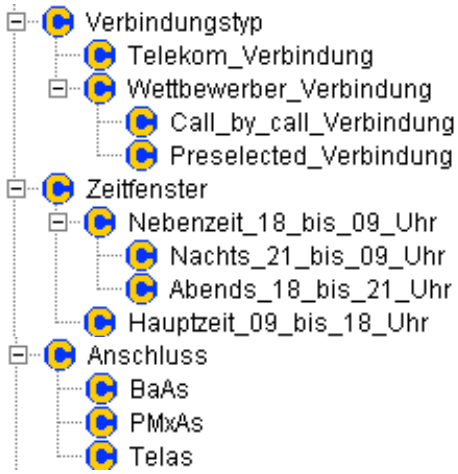

Abbildung 10.4: Ausschnitt aus der Domänenontologie

```
<ConceptMap>
        <ConceptLabel>Telekom_Verbindung</ConceptLabel>
        <SQLCondition>(verbtyp = 'Telekom Direct Access')</SQLCondition>
</ConceptMap>
<ConceptMap>
        <ConceptLabel>CallByCall_Verbindung</ConceptLabel>
        <SQLCondition>(verbtyp = 'Wettbewerber CbC')</SQLCondition>
</ConceptMap>
<ConceptMap>
        <ConceptLabel>Preselect_Verbindung</ConceptLabel>
        <SQLCondition>(verbtyp = 'Wettbewerber Preselect')</SQLCondition>
</ConceptMap>
```

Abbildung 10.5: Abbildung der Konzepte auf SQL-Bedingungen

bank, die jede Verbindung anhand von Merkmalen eindeutig charakterisiert. Um Sichten auf den Kommunikationsdaten unter Berücksichtigung der Domänenontologie berechnen zu können, benötigen wir eine Abbildung der Konzepte auf die Merkmale in der Datenbank. Das Konzept “Telekom_Verbindung” beschreibt alle Verbindungen, die über das Netz der Telekom geführt werden und muss nun auf das Merkmal “Telekom Direct Access” der Datenbank abgebildet werden. Gleiches gilt für das Konzept “Wettbewerber_Verbindung”.

Wir benötigen eine formale Spezifikation dieser Abbildungen für jedes Konzept. Zur Beschreibung der durch ein Konzept betroffenen Datensätze wird jedem Konzept eine SQL-Bedingung zugeordnet. Beispiele sind in Abbildung 10.5 zu sehen.

Die erste Abbildungsvorschrift bildet das Konzept “Telekom_Verbindung” auf die Spalte “verbtyp” ab und schränkt die Menge der Datensätze auf alle die ein, die gleich “Telekom Direct Access” enthalten. Sowohl zur Bestimmung des Supports als auch zur Erzeugung der Sichten wird die Abbildung eingesetzt. Im Endeffekt bildet diese Abbildungsvorschrift eine erweiterte Version der Ref_C-Funktion, die Konzepte auf lexikalischen Einträge abbildet (vgl. Abschnitt 6.2).

Betrachten wir nun noch einige Besonderheiten der Mappingdatei. Um den Spezifikationsaufwand zu reduzieren, müssen nur die Blattkonzepte ein Mapping in der Mappingdatei enthalten. Alle übrigen Konzepte können das Mapping anhand der Unterkonzepte bestimmen. Zum Beispiel ist dem Konzept “Wettbewerber_Verbindung” kein Mapping zugeordnet. Das Programm stellt bei der Ausführung das fehlende Mapping fest und versucht, dies anhand der Unterkonzepte “CallBy-

Abbildung 10.6: Ausschnitt aus der Arbeitsontologie

Call_Verbindung” und “Preselect_Verbindung” zu berechnen. Sollte diese wiederum kein Mapping enthalten, wird rekursiv ein weiterer Verfeinerungsschritt durchgeführt. Erst wenn ein Blattkonzept kein Mapping enthält, terminiert das Programm mit einer Fehlermeldung.

Das Mapping und die Domänenontologie erlauben die gezielte Auswahl von Kommunikationsdatensätzen durch die Ontologie aus der Datenbank. Wir müssen nun noch festlegen, welche Größe wir analysieren wollen. Die verschiedenen Größen können wir dem Fragebogen entnehmen. Direkt in der Datenbank enthalten sind:

- dauer,
- umsatz,
- verbindung.

Das Feld “dauer” enthält die Länge des Gespräches, “umsatz” die Kosten tarifiert nach Standardtarif der Telekom und “verbindung” den Anteil der Verbindung, der zu dieser Stunde stattfand. Geht eine Verbindung über die volle Stunde hinaus, so wird mehr als ein Datensatz in der Datenbank abgelegt.

Nachdem wir nun die Zielgröße der Analyse beschreiben können, fehlt uns noch die entsprechende Aggregationsfunktion. Sie fasst die Menge der Datensätze pro Kunde zusammen und führt so zu einem kundenbeschreibenden Merkmal. Wir können alle von der Datenbank zur Verfügung gestellten Funktionen einsetzen. Den Ergebnissen des Fragebogens entnehmen wir, dass neben der Summe der Minuten (also Dauer) auch der Marktanteil eine wichtige Größe darstellt. Mehr Details dazu findet man in [179]. Die Notation der Merkmale entnehmen wir Kapitel 7.4. So ergibt sich der Ausschnitt der Arbeitsontologie in Abbildung 10.6 passend zur Domänenontologie.

Personalisierte Sichten Wie eingangs schon erwähnt, erlauben die Fragebögen die leichte Erstellung von personalisierten Sichten. Frage 5 in Anhang G enthält die passenden Fragen. Schon durch die Nutzung nur der wichtigsten Merkmale reduziert sich die Dimensionalität der Daten drastisch und die Clusterergebnisse werden verständlicher. Weiterhin kann man die erfragten Merkmale in eine Arbeitsontologie übersetzen und mittels COSA weitere personalisierte Sichten erzeugen, oder man nutzt nur die wichtigen Merkmale (falls es nicht zu viele sind) zum Clustern. Ist z.B. das Merkmal “Summe des Umsatzvolumens” als wichtig im Fragebogen markiert, so erscheint in

der Arbeitsontologie das Merkmal “sum(umsatz)__AlleVerbindungen”. Dies ist möglich, da alle Merkmale des Fragebogens ein entsprechendes Konzept in der Ontologie haben. Auch die Aggregatsfunktion kann entsprechend der Definition in Abschnitt 7.4 im Konzept einer Arbeitsontologie spezifiziert werden. Es ist also möglich, alle Informationen des Fragebogens direkt in Konzepte der Arbeitsontologie zu übersetzten.

10.1.5.2 Analyse der Telekomontologie und -daten in Verbindung mit COSA

Zentrale Idee von COSA ist die Nutzung von formalisiertem Hintergrundwissen zur Generierung von aussagekräftigen Merkmalen. Die Merkmale stellen die Grundlage einer späteren Datenanalyse dar. Um sowohl die Auswahl der Merkmale im COSA als auch die Datenanalyse sinnvoll durchführen zu können, müssen die ursprünglichen Merkmale alle Informationen der Objekte umfassen. Ein Beispiel illustriert diese Problemstellung.

Jedes Telefongespräch eines Kunden der Deutschen Telekom AG wird zu einer bestimmten Zeit, mit einer bestimmten Dauer und zu einem bestimmten Ort bzw. mit einer bestimmten Entfernung geführt. Zusätzlich kann das Gespräch über verschiedene Verbindungsnetzbetreiber und zu unterschiedlichen tariflichen Konditionen abgewickelt werden. Diese Merkmale charakterisieren ein Gespräch und machen es eindeutig. Übernimmt man bei der Transformation (siehe Abschnitt 10.1.2) alle Informationen, ergibt sich nicht wie in Abschnitt 10.1.2 ein 76-dimensionaler Raum, sondern die Kombination aller möglichen Ausprägungen der angeführten Merkmale. Ermittelt man allein den Startzeitpunkt eines Gespräches auf Sekundenbasis, ergibt sich eine sehr hohe Anzahl an Merkmalen. Diese Merkmale müssen dann wiederum mit allen anderen Merkmalen kombiniert werden.

Die beschriebene Kombination aus Merkmalen kommt in der Ontologie nicht vor. Wir finden dort nur jedes Merkmal einzeln, namentlich z.B. Tageszeit oder Ort. Wählen wir zum Beispiel den Ort als beschreibendes Merkmal und schauen uns die Summe der Verbindungsminuten eines Kunden an, so erhalten wir im Ergebnis alle Verbindungsminuten des gewählten Zeitraumes (z.B. 1 Monat). Das gleiche Ergebnis erhalten wir bei der Wahl von Tageszeit, Tagart oder Verbindungsnetzbetreiber, sprich bei allen Konzepten unter ROOT bekommen wir 100 % der Minuten. COSA geht aber davon aus, dass dies erst beim ROOT geschieht. Bei Textdokumenten in Kapitel 7 ist dies auch der Fall. Die Anzahl der Worte eines Dokumentes besteht aus der Summe der Vorkommenshäufigkeiten aller Worte. Auf diesem Weg erhalten wir die Länge des Dokumentes bei ROOT.

Analysieren wir den folgenden Spezialfall: Starten wir COSA nicht von ROOT, sondern vom Konzept LAND (vgl. Abbildung G.2), können wir COSA ohne Modifikationen zum Generieren von Auslandssichten einsetzen. Wurde ein Gespräch zum Beispiel nach Frankreich geführt, dann fallen diese Minuten nur unter Frankreich bzw. unter die Oberkonzepte Europa und Ausland. Gleiches gilt, wenn man z.B. nur die Tageszeit untersucht. Ein Gespräch wird auf diesem Wege also immer auch nur einmal gezählt. Wir benötigen eine Lösung, wenn wir mit allen Merkmalen wie Ort und Tageszeit gemeinsam arbeiten wollen. Zur Lösung kombinieren wir die Merkmale zu so genannten *Kreuzkonzepten*.

In einem ersten Schritt müssen wir alle Konzepte der Domänenontologie kennzeichnen, die bei gemeinsamer Verwendung Kreuzkonzepte bilden können. Die benötigten Modifikationen am COSA-Algorithmus zur korrekten Auswertung der Kreuzkonzepte sowie deren Notation sind in Kapitel 7.4 spezifiziert. Wir präsentieren im nächsten Abschnitt einige Ergebnisse für Kommunikationsdatensätze der Telekom.

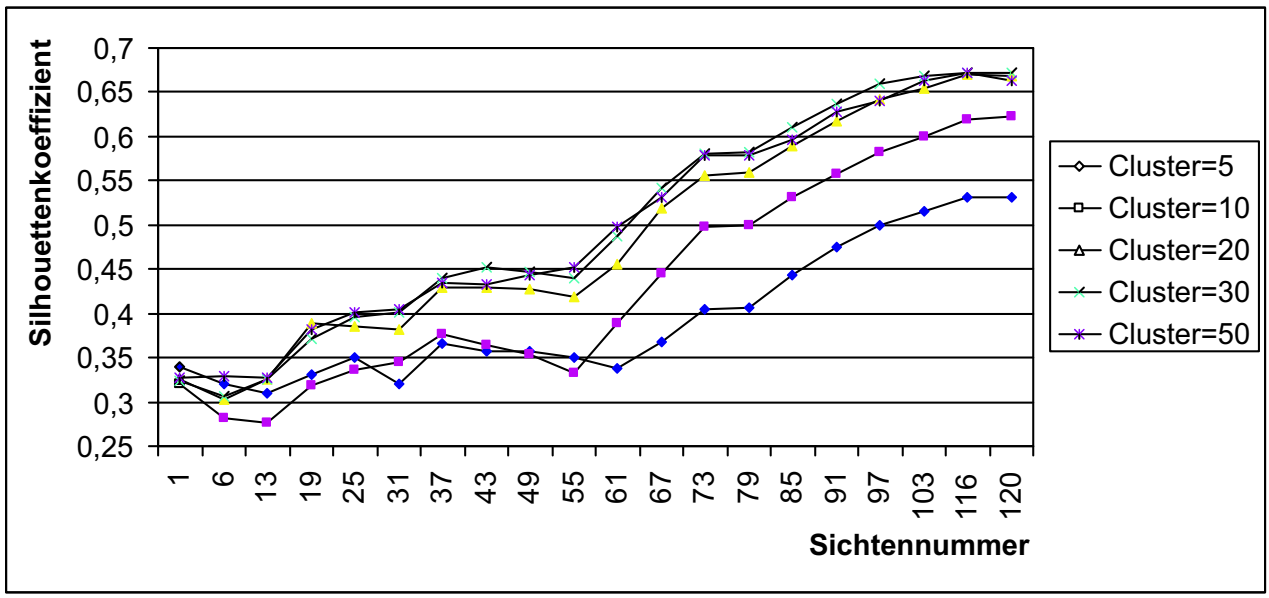

Abbildung 10.7: Silhoutten-Koeffizient für verschiedene Sichten mit unterschiedlicher Anzahl von Clustern für die Auslandsontologie

10.1.5.3 Ergebnisse des (erweiterten) COSA-Algorithmus auf Telekomdaten

Abschließend stellen wir in diesem Abschnitt Ergebnisse mit dem COSA-Algorithmus auf den Telekomdaten vor. Dabei schauen wir uns zwei Aspekte an. Auf der einen Seite wollen wir sehen, inwieweit die Reduktion der Dimensionalität zur Steigerung der Clusterergebnisse beiträgt. Wir messen dies mit dem Silhouetten-Koeffizienten. Weiterhin interessiert uns die Verteilung der Sichten über den Merkmalsraum. Hier sollten unterschiedliche Sichten zu unterschiedlichen Ergebnissen führen. Aus Gründen der Geheimhaltung können wir an dieser Stelle keine konkreten Zahlen nennen, sondern schauen uns die Verteilung der Kunden über zwei verschiedene Clusterungen an.

Güte der Sichten bei Auslandsgesprächen Für das folgende Experiment wählen wir aus der 10 % Stichprobe (siehe Abschnitt 2.5) alle großen Unternehmen (insgesamt 24156) aus. Wir betrachten alle Gespräche dieser Kunden für den Oktober 2000 und interessieren uns nur für den Auslandsverkehr, wobei wir hoffen, dass viele der großen Unternehmen auch ins Ausland telefonieren, ihr Verhalten aber sehr heterogen sei. Der Auslandsbereich der Domänenontologie umfasst ca. 230 Länder, die ihrerseits wieder in Regionen und Tarifzonen unterteilt sind. Die Länder bieten die Chance, das Verhalten von COSA genauer zu analysieren und ganz unterschiedliche Sichten auf die Daten generieren zu lassen. COSA wurde mit einer maximalen Dimensionalität von 10 Dimensionen gestartet.

Abbildung 10.7 gibt für jede Sicht den Silhouetten-Koeffizienten für die Clusteranzahl 5, 10, 20, 30 und 50 wieder. Die Sichtennummer der X-Achse spiegelt den Zeitpunkt der Berechnung der Sicht in COSA wider. Eine Sicht umfasst immer eine Menge von Konzepten, die dann mit Hilfe der Abbildungsfunktion auf die Datenbank abgebildet wird. Die Sichten mit den kleinen Nummern werden zuerst generiert. Sie enthalten sehr generelle Konzepte, wie z.B. die Kontinente und dort als Ziel Mobilfunk- oder Festnetzanschlüsse. Sichten mit den höheren Nummern enthalten immer häufiger Blattkonzepte der Ontologie. Das sind zum Teil Länder, wie USA oder Kanada, oder zum Teil Zonen, die wiederum Länder zusammenfassen. Wir erkennen in der Abbildung 10.7 deutlich die Steigerung der Güte bzgl. des Silhouetten-Koeffizienten. Erstaunlich ist, dass bei den Sichten bis 25 die Clusterergebnisse mit 5 Clustern besser bzw. ähnlich gut sind als die mit 10 Clustern.

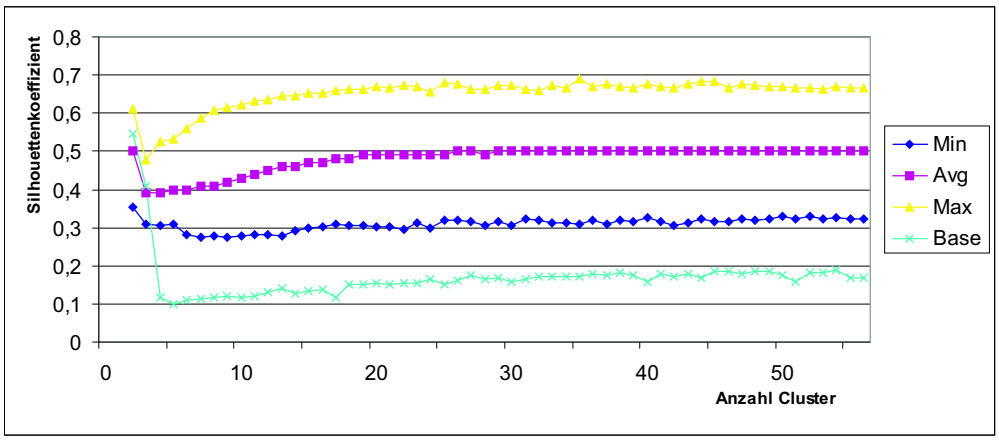

Abbildung 10.8: minimaler, mittlerer und maximaler Silhoutten-Koeffizient über alle Sichten der Auslandsontologie für 2 bis 100 Cluster, sowie Referenzclusterung mit allen Merkmalen

Erst ab Sicht 61 steigt die Güte für 10 Cluster deutlich über die von 5 Clustern. Der Anstieg ist so stark, dass man auf klare Strukturen innerhalb der Daten schließen kann (vgl. Abschnitt 5.3.4.2). Die Clusterungen mit einer Clusteranzahl größer 10 sind immer besser als mit einer Anzahl von 5 oder 10. Einzige Ausnahme bilden die Sichten 1, 6 und 13. Hier kann man kaum Unterschiede zwischen den unterschiedlichen Clusterungen feststellen.

Um den Einfluss der Clusteranzahl besser beurteilen zu können, wurden für alle Sichten die Clusteranzahlen zwischen 2 und 60 berechnet. Abbildung 10.8 gibt den Silhouetten-Koeffizienten der Referenzclusterung sowie den Durchschnitt, das Minimum und das Maximum über alle Sichten an. Die Referenzclusterung (Base) wurde mit 230 Attributen berechnet. Alle Sichten basieren auf 10 Dimensionen. Bis auf die Clusterungen mit 2 und 3 Clustern sind die Sichten immer besser als die Referenzclusterung. Die besten Sichten weisen klare Strukturen auf. Die Ausnahme mit zwei Clustern, bei der die Referenzclusterung ähnlich gute Ergebnisse erzielt wie die Clusterung auf Basis der Sichten, lässt sich leicht erklären. Der Datensatz enthält viele Kunden, die sehr wenig telefonieren. Die werden im ersten Schritt, also bei zwei Clustern, vom Rest getrennt. Warum auch die Referenzclusterung mit drei Clustern so gute Ergebnisse erzielt, wurde nicht herausgefunden. Festzustellen bleibt, dass die Referenzclusterung hier nur noch im Schnitt der Sichtenclusterungen liegt.

Analysieren wir zum Abschluss die Anzahl der Cluster für eine ausgewählte Sicht. Abbildung 10.9 gibt für die Sicht 91 den Silhouetten-Koeffizienten für Clusterungen mit der Anzahl 2 bis 100 wieder. Wir erkennen, wie auch schon in Abbildung 10.8 beobachtet, die sehr gute Bewertung der Clusterung mit zwei Clustern und den deutlichen Abfall bei drei Clustern. Ab drei Clustern steigt das Ergebnis bis zum Maximum bei 42 Clustern an und bleibt ab dort fast konstant. Das gute Ergebnis bei zwei Clustern lässt sich auch hier sehr gut erklären. Durch die begrenzte Anzahl an Merkmalen, die nicht von allen Kunden genutzt werden, gibt es eine große Anzahl, die für fast alle verwendeten Merkmale kein Gespräch geführt hat bzw. sehr wenige Gespräche insgesamt. Diese werden in einem ersten Schritt von den Vieltelefonierern getrennt, was sehr gut funktioniert. Mit drei Clustern wird eine weitere Gruppe mit großem Gesprächsaufkommen von den Wenigtelefonierern getrennt. Diese Trennung ist laut Maß für die gesamte Clusterung nicht von Vorteil. Die Anzahl der schlecht geclusterten Kunden steigt so stark, dass durch den dritten (schlechten) Cluster das Ergebnis drastisch sinkt. Die Clusterung entdeckt nur unzureichend die Struktur der Daten.

Die berechneten Sichten führen also zu ganz unterschiedlichen Perspektiven auf den Datensatz (allgemein vs. speziell) und steigern die Güte der Clusterergebnisse ganz erheblich. Auf diese Weise kann dem Preismanagement der Telekom ein effizientes Mittel zur Analyse der Kunden in die Hand

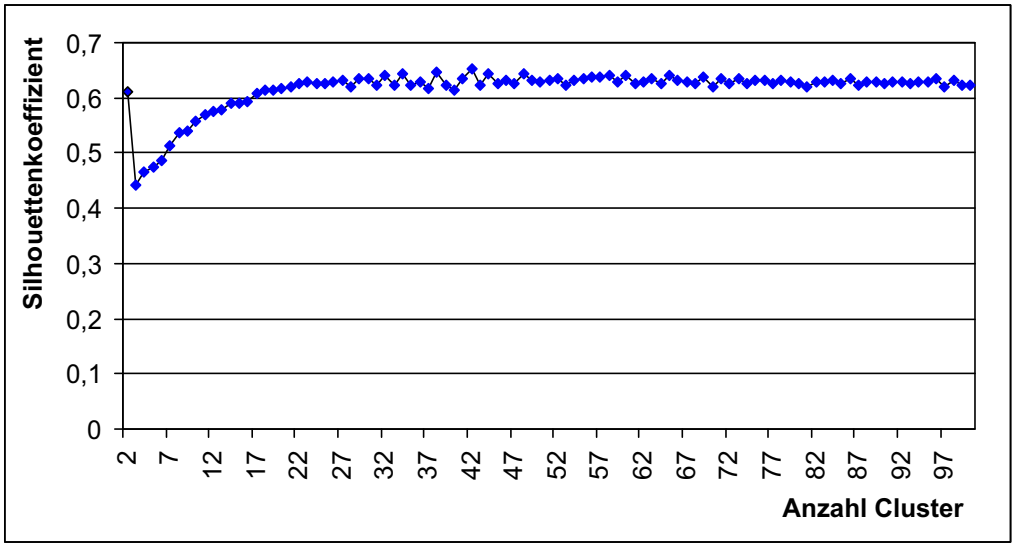

Abbildung 10.9: Silhoutten-Koeffizient für Sicht Nummer 91 der Auslandsontologie für 2 bis 100 Cluster, sowie Referenzclusterung mit allen Merkmalen

gegeben werden. Um die durch die Sichten behandelten unterschiedlichen Aspekte der Kunden ein wenig genauer zu untersuchen, diskutieren wir eine Clusterung personalisierter Sichten und vergleichen die Zuordnung der einzelnen Kunden zu den jeweiligen Clustern.

Vergleich von Clusterergebnissen verschiedener Sichten Wir führen zwei Experimente durch. Einmal nutzen wir einen Datensatz mit 77997 Kunden. Dies war eine zufällige Stichprobe der 10 % Stichprobe und enthielt nur Verbindungsdaten des Monats Juni 2002. Die Ergebnisse auf diesem Datensatz sind vergleichbar mit den Ergebnissen von 2000 zufällig gezogenen Kunden, wobei der Zeitraum der betrachteten Verbindungen drei Monate von Juni bis August 2002 betrug. Wir berechneten mit dem Bi-Sec-KMeans jeweils 10 Cluster auf Basis des Kosinus-Maßes und logarithmierten Werten der abgeleiteten Merkmale eines jeden Kunden.

Eine wichtige Eigenschaft der Clusterergebisse sind die verschiedenen Aspekte, die durch einzelne Sichten repräsentiert werden. Dabei findet man jedes Blatt-Merkmal[1] in mindestens einer Sicht. Die Kombination der Merkmale und ob ein nicht Blatt-Merkmal in einer Sicht vorkommt, ergibt sich im Falle der Kommunikationsdaten anhand der geführten Gespräche der Kunden. Aggregate, die “zu viele” Informationen (z.B. Gesprächsminuten) zusammenfassen[2], können schon vor der Ausgabe der ersten Sicht durch deren Unterkonzepte ersetzt werden. Schauen wir uns nun an, inwieweit die verschiedenen Sichten Kundengruppen unterschiedlich betrachten. Wir initialisieren Bi-Sec-KMeans mit dem gleichen Seed für den Zufallszahlengenerator, so dass die gleiche Startlösung für die verschiedenen Sichten verwendet wird. Die unterschiedlichen Ergebnisse lassen sich dann auf die unterschiedlichen Merkmale zurückführen. Alles in allem erhielten wir für eine kleine personalisierte Ontologie 11 Sichten.

Tabelle 10.1 stellt die Clusterung der Sicht 1 und Sicht 11 gegenüber. Jeder Zelle der Tabelle enthält die Anzahl der Kunden, die bei der jeweiligen Clusterung in diesen Cluster gefallen ist. Die erste Zelle ist die Zelle mit Cluster 0 der Sicht 1 und Cluster 0 der Sicht 11. Sie enthält 34 Kunden, die in beiden Fällen dem Cluster 0 zugeordnet worden. Cluster 0 bleibt in beiden Sichten fast vollständig erhalten. Nur 6 bzw. 8 Kunden werden anderen Clustern zugeordnet. Auch die Eigen-

[1] Ein Blatt-Merkmal wird durch ein Konzept ohne Unterkonzept repräsentiert.
[2] Informationen werden anhand der spezifizierten Größe beschrieben, siehe Abschnitt 10.1.5.1.

Tabelle 10.1: Sicht 1 (Zeilen) vs. Sicht 11 (Spalten), 10 Cluster mit Bi-Sec-KMeans

	0	1	2	3	4	5	6	7	8	9	$\sum$
0	34	1	7	0	0	0	0	0	0	0	42
1	6	42	26	0	1	0	0	0	0	0	75
2	0	68	65	0	0	1	0	0	0	0	134
3	0	0	61	0	0	0	0	0	0	0	61
4	0	0	0	0	0	11	106	2	8	0	127
5	0	0	0	18	0	341	0	40	171	62	632
6	0	0	0	2	43	88	0	0	5	46	184
7	0	0	0	0	9	63	18	1	5	10	106
8	0	0	0	4	29	141	0	7	19	30	230
9	0	0	0	5	0	283	0	13	71	37	409
$\sum$	40	111	159	29	82	928	124	63	279	185	2000

schaften der Cluster sind nahezu identisch. Auffällig ist außerdem der ausschließliche Austausch an Kunden zwischen den Clustern 1 und 2 der Sicht 11 mit den Clustern 1, 2 und 3. Nur zwei Kunden der Cluster 1 und 2 aus Sicht 1 finden sich in anderen Clustern wieder. Eine wesentlich stärkere Durchmischung ist zwischen den übrigen Clustern der Sichten zu beobachten. Hierbei kommt es zum Teil zu völlig neuen Gruppen. Zum Beispiel findet man 928 Kunden in Cluster 5 der Sicht 11. Cluster 5 aus Sicht 11 verteilt sich hauptsächlich auf die Cluster 5, 8 und 9 der Sicht 1.

Die unterschiedlichen Merkmale der Sichten führen tatsächlich zu Clusterungen mit ganz unterschiedlichen Eigenschaften. Dabei konnten wir beobachten, dass auf der einen Seite einige Cluster erhalten bleiben, auf der anderen Seite auch völlig neue Cluster berechnet werden.

Wir konnten in diesem Abschnitt anhand der Kommunikationsdaten der Deutschen Telekom AG zeigen, dass Sichten zur Verbesserung der Clustergüte führen und man sehr unterschiedliche Clusterergebnisse basierend auf den jeweiligen Sichten beobachtet. Die Betonung unterschiedlicher Merkmale einzelner Sichten ist der Schlüssel für dieses Ergebnis.

Im folgenden Abschnitt stellen wir eine Architektur eines Wissensportals vor, die die Methode des Subjektiven Clusterns zum Strukturieren von Webseiten einsetzt.

10.2 Weitere Anwendungen des Subjektiven Clusterns

Text Mining und speziell das Clustern von Textdokumenten kann für unterschiedlichste Aufgaben eingesetzt werden. Das Clustern von Dokumenten für eine strukturiertere Präsentation ist ein spannendes Anwendungsszenario. Die in dieser Arbeit entwickelte Methode des Subjektiven Clusterns erlaubt es dem Anwender, mittels einer Ontologie die wesentlichen Elemente zur Strukturierung der Dokumente vorzugeben. Auch die Präsentation der Ergebnisse ist wesentlich mehr auf den Anwender fixiert, da er mit Hilfe der Ontologie die beschreibenden Merkmale, Terme und Begrifflichkeiten in strukturierter Form vorgeben und einschränken kann.

Im Folgenden werden zwei Anwendungsgebiete für Subjektives Clustern vorgestellt. Abschnitt 10.2.1 geht auf die SEAL-II-Architektur ein. Insbesondere wird die Clustering-Komponente der Architektur vorgestellt, die den Dokumentenbestand strukturiert und die Navigation durch ihn wesentlich erleichtert (vgl. [114]). In der zweiten Anwendung in Abschnitt 10.2.2 steht das Navigieren und Browsing von Lernmaterialien im Vordergrund. Auch hier kann das Subjektive Clustern zur

Strukturierung der Lernmaterialien eingesetzt werden.

10.2.1 Wissensportale

Bei SEAL-II handelt es sich um eine Architektur für Semantische Portale, wobei das ontologiebasierte Clustern bei der zielgerichteten Strukturierung von unstrukturierten Informationen hilft (vgl. [114]). Es baut auf dem SEAL-Framework (SEmantic portALs) auf (siehe [201, 205, 156, 157]). Mit SEAL wurde eine umfassende Architektur mit einer Reihe von Tools zur Verbesserung des Verhältnisses zwischen Aufwand und Nutzen bei der Erstellung, Pflege und Wartung von Portalen eingeführt. Die Technologie zur Präsentation und zum einfachen Austausch von Informationen basiert auf Ontologien. Um den Einstieg in ein solches Portal zu erleichtern, erweitert SEAL-II das SEAL-Framework um die Möglichkeit, auch unstrukturierte Informationen verarbeiten und präsentieren zu können.

SEAL-II vermittelt dazu zwischen völlig unstrukturiertem und reichhaltig strukturiertem Wissen. Wir nutzen die Ontologie sowohl für Wissenselemente, die Metadaten enthalten, als auch zur Steuerung weiterer Techniken, die unstrukturierte Daten sammeln und für die explorative Analyse durch den Menschen aufbereiten. Um beim Aufbau eines semantischen Portals mit wenig Aufwand schon großen Nutzen zu erzeugen, wird unstrukturiertes Wissen z.B. in Form von Textdokumenten laufend dem Portal hinzugefügt. Techniken wie Information Retrieval, Textclustern oder auch einfaches Keyword Matching helfen bei den ersten Zugriffen. Findet der Nutzer relevante Informationen, kann er diese leicht markieren und fügt so automatisch Metadaten ins System ein. Diese Metadaten, aber auch die im System enthaltene Ontologie, erlauben es, Verfahren wie das Crawlen und Clustern von Texten zu modifizieren. Die Ontologie steuert Verfahren und kann so auf die Bedürfnisse des Anwenders eingehen. Der Nutzer findet immer einfacher, schneller und wesentlich mehr der gesuchten Informationen. Gleichzeitig bietet das Portal immer die Möglichkeit, zu jeder gefundenen Webseite bzw. zu jedem gefundenen Textdokument auch beschreibende Metadaten des Nutzers abzulegen, die dann wieder in den Prozess einfließen. Wurden genügend Metadaten dem System hinzugefügt, so wird nicht nur die Keyword-basierte Suchanfrage erfolgreich sein, sondern die wirklich relevanten Antworten werden immer häufiger aus der Wissensbasis mit den strukturierten Daten stammen.

Abbildung 10.10 macht die Idee der gleichzeitigen Nutzung von unstrukturierten (links in der Abbildung) über angereicherte bis zu den strukturierten Informationen (rechts in der Abbildung) deutlich. Abbildung 10.11 bettet diese Elemente in die eine Architektur für Wissensportale ein. Die Architektur enthält neben dem Knowledge Warehouse zum Speichern der strukturierten Information auch die Komponenten wie Clustering und Crawling zum Sammeln und Verarbeiten von unstrukturierten Informationen wie Webseiten oder Textdokumente. Wir wollen an dieser Stelle nicht auf alle Komponenten im Detail eingehen, sondern nur die für das Subjektive Clustern relevanten beschreiben. Mehr findet man in [114].

Die Komponenten ontologiefokussiertes Crawling und ontologiebasiertes Clustern werden im Folgenden detailliert beschrieben.

Ontologiefokussierter Crawler: Eine wichtige Komponente in der Architektur ist der ontologiebasierte fokussierte Crawler für Dokumente und (Meta-) Daten. Mit Hilfe der Ontologie des Knowledge Warehouse bewertet der Crawler die gesammelten Daten und steuert so den gesamten Suchprozess. Daten können sowohl Textdokumente im Intranet, Web-Seiten als auch Metadaten sein, die z.B. durch Annotierung in die Webseite eingebettet wurden [99, 100]. Die gesammelten Web-Seiten werden zur Bewertung und Verarbeitung in einen Konzeptvektor überführt (vgl. 4.2.1). Relevante Metadaten werden direkt im Knowledge

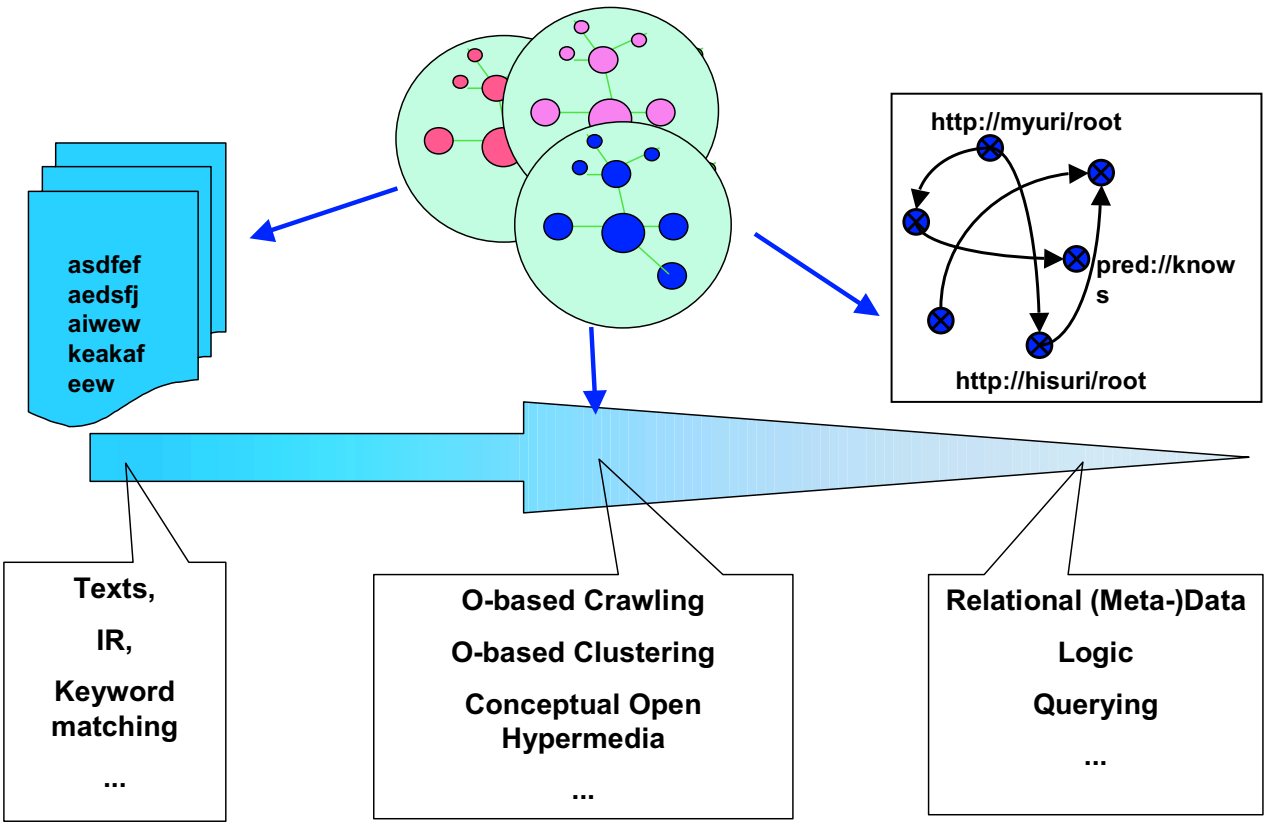

Abbildung 10.10: Bewältigung verschiedener Anforderungen: Wissensmanagementtechniken für strukturierte und unstrukturierte Informationen

Warehouse abgespeichert. Details zum fokussierten Crawler findet man in [114, 154]. Die so gesammelten Textdokumente bilden neben den manuell ins Portal eingestellten Dokumenten die Grundlage für die Anwendung der Techniken zur Strukturierung der Informationen im Portal.

Ontologiebasiertes Clustern: Ein spezielles Feature zur Analyse von Dokumenten in SEAL-II stellt das ontologiebasierte Clustern dar, das auf der in dieser Arbeit entwickelten Methode des Subjektiven Clusterns basiert. Durch das Clustern wird eine erste Struktur für die unstrukturierte Dokumentensammlung berechnet. Ähnliche Dokumente werden dabei auf Grundlage der im Knowledge Warehouse gespeicherten Ontologie in Gruppen zusammengefasst.

Im Gegensatz zu herkömmlichen Clustermethoden wird nicht nur eine, sondern es werden mehrere niedrigdimensionale Clusterungen berechnet. Die Merkmale, die zur Berechnung der Clusterung verwendet werden, sind die Konzepte der Ontologie. Sie spiegeln die Interessen der Anwender und die relevanten Themen des Portals wider. Im Allgemeinen sind die Anwender mit den in der Ontologie enthaltenen Konzepten vertraut. Da die Konzepte auch die Merkmale einer Clusterung darstellen und die Cluster auf der Basis der Merkmale präsentiert werden, werden die Cluster in für den Anwender leicht verständlicher Form wiedergegeben. Auch die Clusterung an sich konzentriert sich durch die Nutzung der Konzepte beim Gruppieren der Dokumente auf die wesentlichen und relevanten Informationen.

Zwar sollte jeder Anwender eines Wissensportals die zugrunde liegende Ontologie kennen. Er ist aber nicht immer an jedem Detail interessiert. Um den unterschiedlichen Interessen der Anwender gerecht zu werden, wird nicht eine Clusterung auf der Basis aller Konzepte der

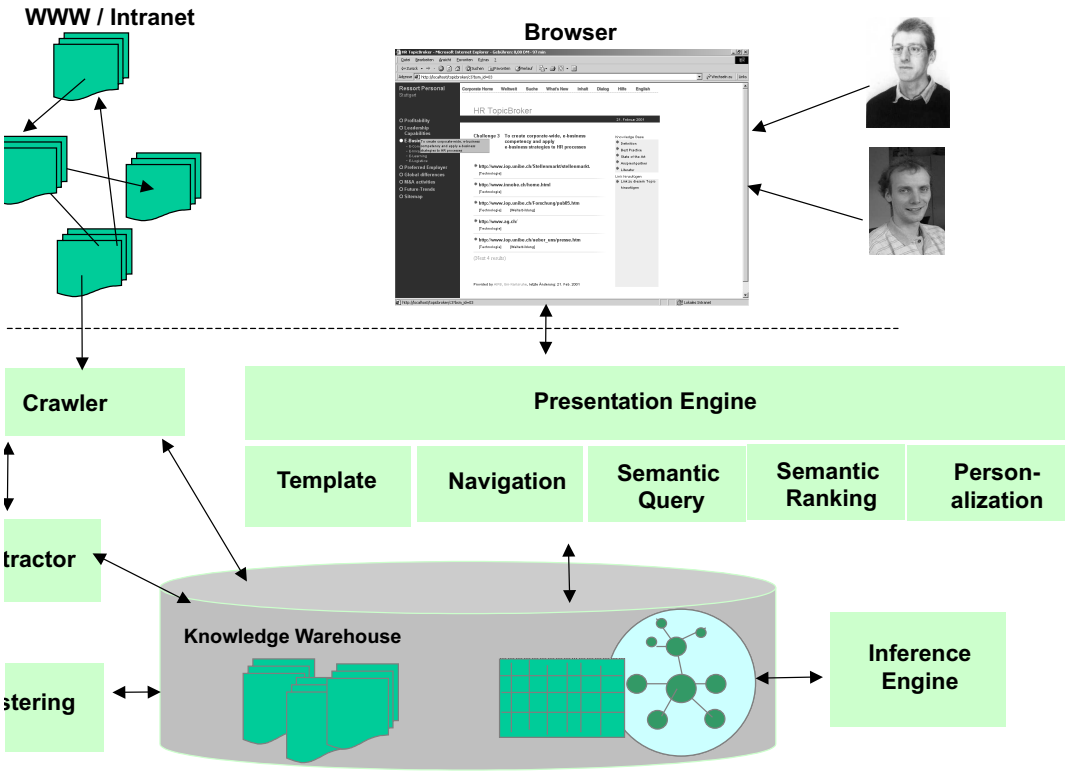

Abbildung 10.11: Architektur SEAL-II

Ontologie durchgeführt, sondern es werden verschiedene Clusterungen auf der Basis ganz unterschiedlicher Merkmalsmengen – genannt Sichten – berechnet (vgl. Kapitel 7). In einem ersten Schritt wählt der Anwender die für ihn relevante Sicht aus. Die Dokumente werden anhand dieser Merkmale geclustert und dann präsentiert. Neben den vorberechneten Sichten kann der Anwender auch selber Merkmale zum Clustern vorgeben, um eine personalisierte Strukturierung der Dokumente zu erhalten.

Für die Strukturierung von unstrukturierten Informationen in Wissensportalen bietet sich die Verwendung des Subjektiven Clusterns an. Auf diesem Wege werden unstrukturierte Informationen leichter nutzbar gemacht und stehen zusammen mit strukturierten Informationen zur Verfügung.

Im nächsten Abschnitt gehen wir auf die Nutzung des Subjektiven Clusterns zur Struktuierung von Lernmaterialien ein.

10.2.2 Subjektives Clustern von Lernmaterialien

Ein vielversprechendes Anwendungsgebiet für den Einsatz des Subjektiven Clusterns ist der “Courseware Watchdog” (siehe [217] und [216]). Dabei handelt es sich um ein ontologiebasiertes Werkzeug zum Suchen, Finden und Organisieren von Lernmaterialien. Die Lernmaterialien stehen dabei in elektronischer Form zur Verfügung und sind zum Teil mit Metadaten annotiert. Der Austausch der Metadaten erfolgt dezentral, d.h. über ein Peer-to-Peer-System. Das Werkzeug besteht aus einer Browsing-, Crawling-, Such-, Anfrage-, Cluster- und Evolutionskomponente. Mit Hilfe dieser Elemente wird das Suchen und Erstellen von Kursunterlagen unterstützt.

Wie schon im letzten Abschnitt bei Wissensportalen dargestellt, nutzt das System einen “fokussierten Crawler”, der Webseiten, die relevant für den Anwender sind, aus dem Internet einsammelt.

Die Relevanz wird mit Hilfe der Ontologie bestimmt. In einem nächsten Schritt werden die eingesammelten Seiten durch das Subjektive Clustern strukturiert. Dabei erlaubt das Subjektive Clustern multiple Sichten durch die Nutzung von mehreren niedrigdimensionalen Merkmalsmengen auf ein und dem selben Dokumentenbestand. Durch die im System integrierte Browsing-Komponente kann der Anwender leicht eine Merkmalsmenge fixieren und bekommt dann eine entsprechende Clusterung in leicht verständlicher Form präsentiert.

Im nächsten Kapitel wenden wir das Clustern und Visualisieren mit Hintergrundwissen auf drei weiteren realen Datensätzen an. Wir werden für zwei Datensätze eine Verbesserung der Clustergüte zeigen und präsentieren für alle drei Datensätze FBA-basierte Visualisierungen der Clusterungen.

11 Clustern und Visualisieren mit Hintergrundwissen

In diesem Kapitel werden wir die Anwendung der Methoden aus Kapitel 8 und 9 anhand dreier weiterer Datensätze diskutieren. Für den Java-eLearning-Datensatz des Abschnitts 11.1 und für den AGROVOC-Datensatz des Abschnitts 11.2 können wir sowohl qualitative als auch quantitative Ergebnisse präsentieren. Für den Tourismusdatensatz des Abschnitts 11.3 fehlt uns eine manuelle Klasseneinteilung, so dass wir nur die visualisierte Verbandstruktur der Textcluster vorstellen können. Wir folgen [116] bei der Präsentation einiger Ergebnisse dieses Kapitels.

11.1 Lernmaterialien

11.1.1 Ergebnisse des Textclusterns auf dem Java-eLearning-Datensatz

Wir präsentieren im Folgenden die Ergebnisse der Clusterung des eLearning-Datensatzes, den wir in Abschnitt 2.2 eingeführt haben. Für die Berechnung der Textcluster mit Hintergrundwissen verwenden wir auf der einen Seite die domänenspezifische Ontologie (siehe Abschnitt 6.3.2.3) und auf der anderen Seite WordNet (siehe Abschnitt 6.3.3.1). Neben der Validierung unserer Ergebnisse aus Kapitel 8, dass Hintergrundwissen beim Clustern von Textdokumenten zu besseren Ergebnissen führt, sind wir auch am unterschiedlichen Einfluss von domänenspezifischer und domänenunabhängiger Ontologie interessiert.

Für alle Clusterungen des Java-Datensatzes werden zehn Cluster berechnet. Der Prunethreshold liegt bei 17. Bei der Nutzung der Ontologien wird der wahrscheinlichste Sinn bei der Wortsinnerkennung verwendet (HYPDIS = first) und es wird ein Oberkonzept hinzugefügt (HYPDEPTH = 1), da die Java-Ontologie nur eine flache Hierarchie von durchschnittlich fünf Konzepten besitzt (WordNet liegt bei durchschnittlich 13). Bei der Nutzung von Wort- bzw. Konzeptvektoren wird zwischen den Strategien “add” und “only” variiert. Details zu den Strategien findet man in Abschnitt 8.2.3.

Tabelle 11.1: Ergebnisse für den Java-Datensatz mit $k = 10$ Cluster, prune = 17; bei Nutzung von Hintergrundwissen: HYPDIS = first, HYPDEPTH = 1, (avg. gibt den durchschnittlichen Wert für 20 Clusterläufe und std. die Standardabweichung an)

Ontologie	H.INT	Purity avg ± std	InversePurity avg ± std	F-Measure avg ± std	Entropy avg ± std
ohne		**0,61 ± 0,051**	**0,662 ± 0,062**	**0,602 ± 0,047**	**0,845 ± 0,102**
Wordnet	add	0,634 ± 0,070	0,665 ± 0,051	0,626 ± 0,062	0,803 ± 0,125
Java	add	0,651 ± 0,076	**0,685 ± 0,064**	**0,646 ± 0,061**	**0,745 ± 0,122**
Wordnet	only	0,630 ± 0,052	0,635 ± 0,051	0,610 ± 0,051	0,825 ± 0,093
Java	only	**0,669 ± 0,041**	0,646 ± 0,026	0,637 ± 0,036	0,751 ± 0,085

Im Ergebnis entnimmt man Tabelle 11.1 eine Steigerung der Clustergüte bei der Nutzung von

Hintergrundwissen. Dies trifft sowohl für die Nutzung von WordNet als auch für die Nutzung der domänenspezifischen Java-Ontologie zu. Damit werden unsere Ergebnisse aus Kapitel 8 bestätigt. Sie scheinen nicht vom Reuters-Datensatz abhängig zu sein.

Die Analyse der Ergebnisse ergab, dass die Steigerung der Clustergüte mit der Java-Ontologie für das F-Measure signifikant mit $\alpha = 2\ \%$ ist. Die Unterschiede für WordNet sind für $\alpha = 2\ \%$ nicht signifikant. Man entnimmt der Tabelle weiterhin, dass mit Hilfe der domänenspezifischen Ontologie die Steigerung der Clustergüte größer ist als unter Verwendung von WordNet. Bei der "add" Strategie beträgt die Differenz 1,7 % (Unterschied nicht signifikant) und bei der "only" Strategie sogar 3,9 % (Unterschied nur $\alpha = 2\ \%$ signifikant) für die Purity-Werte. Die InversePurity-Werte verhalten sich analog.

Weiterhin ist zu beobachten, dass die "add"-Strategie bei WordNet leicht besser ist als die "only"-Strategie. Bei der Java-Ontologie beobachtet man dies nur mit Hilfe der F-Measure- und Entropie-Werte. Interessant ist der Wert für die InversePurity von 64,4 % bei der Verwendung der Java-Ontologie und der only-Strategie. Dieser ist deutlich schlechter als der Wert für die Referenzclusterung, wobei der gleiche Wert bei der Purity deutlich besser ist. Zieht man die beiden Werte für F-Measure und Entropy mit in Betracht, kommt man zu dem Schluss, dass die Purity auf Kosten der InversePurity gesteigert wird. Warum dies bei dieser Strategie passiert ist, konnte nicht festgestellt werden.

Weiterhin können wir beobachten, dass die Nutzung von mehr als einem Oberkonzept bei der Java-Ontologie zu einer Verschlechterung der Ergebnisse geführt hat. Dies scheint mit der Größe der Ontologie und der Anzahl der Oberkonzepte zusammenzuhängen. Die Java-Ontologie ist eine sehr kleine und flache Ontologie (durchschnittliche Tiefe beträgt ca. 5 Konzepte).

Der folgende Abschnitt präsentiert für eine Clusterung des Java-Datensatzes auf der Basis der Java-Ontologie den visualisierten Begriffsverband und diskutiert daran die Extraktion der Clusterthemen.

11.1.2 Visualisierung der Java-eLearning-Textcluster

Für die Visualisierung der Textcluster für den Java-eLearning-Datensatz berechnen wir eine Clusterung mit der Strategie HYPINT=only, HYPDIS=first und HYPDEPTH=1 unter Verwendung der Java-Ontologie und bestimmen dafür einen formalen Kontext. Dazu verwenden wir die Schwellwerte $\theta_1 = 10\ \%$ und $\theta_2 = 35\ \%$ (siehe Abschnitt 9.3.1 für Details zur Verwendung von zwei Schwellwerten). Das Clusterergebnis entspricht den Clusterergebnissen aus der letzten Zeile von Tabelle 11.1.

In Abbildung 11.1 wird der gedrehte Begriffsverband für die zehn Textcluster als Gegenstände und 22 Konzepte der Java-Ontologie als Merkmale visualisiert. Wir nennen diesen Verband KV2. Die 22 Merkmale ergeben sich durch die Nutzung des Schwellwertes θ_2. Im Folgenden gehen wir auf die Bestimmung der Themen ausgewählter Cluster anhand ihrer Merkmale ein.

Der Verband ist übersichtlich und enthält wenige aussagekräftige formale Begriffe. Dies liegt an der ausschließlichen Nutzung des Schwellwertes θ_2. Dem sehr übersichtlichen Verband entnimmt man z.B., dass die Dokumente des Clusters 0 von ARRAYS handeln oder die des Clusters 6 von OPERATOR, also Operatoren. Beim Vergleich dieser Ergebnisse mit den Bezeichnern der gegebenen Klasseneinteilung der Dokumente stellt man eine Übereinstimmung der Konzeptbezeichner mit den Klassenbezeichnern fest.

Interessant ist Cluster 7. Cluster 7 ist sowohl im Umfang des mit (*) markierten formalen Begriffes und des mit (**) markierten Begriffes, wobei die Merkmale im Inhalt dieser Begriffe sehr unterschiedliche Themen ansprechen. Die Merkmale von (*) handeln von Klassen in Java (CLASS) und die von (**) beschäftigen sich mit Applets, womit wir zwei heterogene Themen für Cluster 7

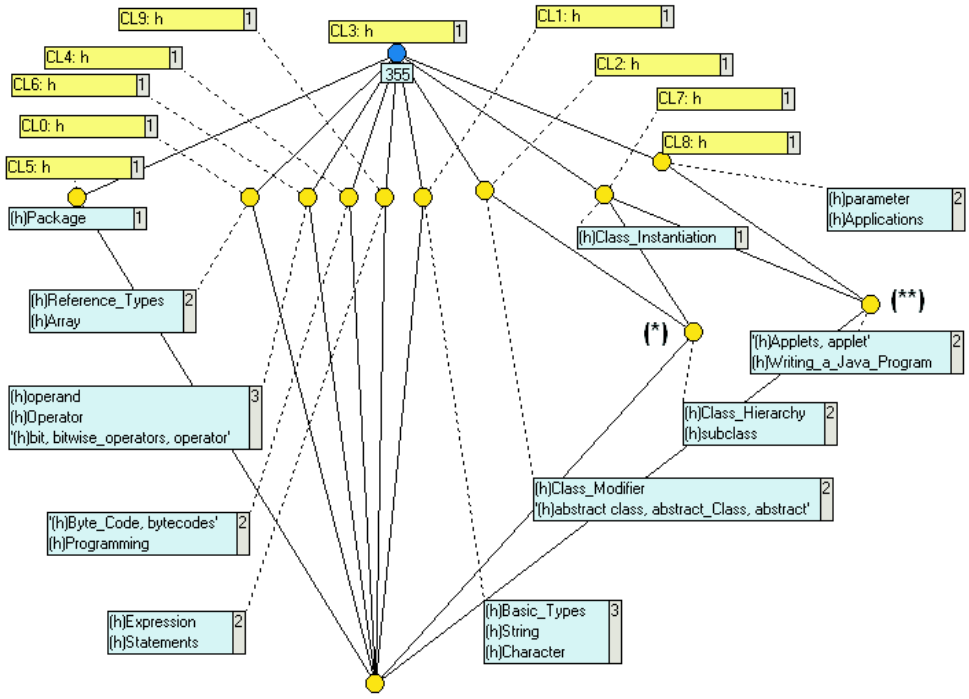

Abbildung 11.1: Begriffsverband KV2 (gedreht) des Java-eLearning-Datensatzes mit zehn Clustern für den Schwellwert $\theta_2 = 35\ \%$

identifizieren können. Cluster 7 ist über den Begriff (*) mit Cluster 2 und über den Begriff (**) mit Cluster 8 verbunden, wobei wir als Themen der Cluster jeweils “Class” bzw. “Applet” identifizieren. Bei der anschließenden Analyse der Dokumente des Clusters 7 stellt man fest, dass in der Tat die Hälfte der Dokumente dieses Clusters der Klasse “Classes” und die andere Hälfte der Klasse “Applet” angehören. Die Themen der Cluster 2 und 8 stimmen ebenfalls mit den identifizierten Themen überein.

Wir konnten zeigen, dass die Visualisierung des Verbandes eine einfache Analyse der Textclusterthemen erlaubt. Dabei können sowohl die Themen einzelner Cluster klar und einfach zugeordnet werden. Auch ist es möglich, Cluster, die mehr als ein Thema umfassen, zu identifizieren und mit Clustern, die gleiche Themen ansprechen, in Beziehung zu bringen.

Im nächsten Abschnitt präsentieren wir Clusterergebnisse für den AGROVOC-Datensatz.

11.2 Landwirtschaftliche Texte

11.2.1 Textcluster der landwirtschaftlichen Texte

Entlang des in Abschnitt 2.3 eingeführten AGROVOC-Datensatz AGeD diskutieren wir im Folgenden Ergebnisse anhand eines weiteren Datensatzes für das Textclustern mit Hintergrundwissen sowie dessen Visualisierung. Auch zu diesem Datensatz steht eine domänenspezifische Ontologie (siehe Abschnitt 6.3.2.1) zur Verfügung, die neben WordNet zum Einsatz kommt.

Tabelle 11.2 fasst die Ergebnisse für den AGROVOC-Datensatz zusammen. Die Clusterung erfolgt mit zehn Clustern, einem Prunethreshold von 30 und den Strategien HYPDIS = first und

HYPINT = only. Bei der Anzahl der zusätzlich integrierten Oberkonzepte werden bei WordNet fünf und bei AGROVOC-Thesaurus eins gewählt, da der AGROVOC-Thesaurus nur eine flache Hierarchie von durchschnittlich drei Konzepten besitzt (WordNet liegt bei durchschnittlich 13). Details zu den Strategien findet man in Abschnitt 8.2.3.

Tabelle 11.2: Ergebnisse für den AGROVOC-Datensatz mit $k = 10$ Cluster, prune = 30; bei Nutzung von Hintergrundwissen: HYPDIS = first, HYPINT = only, bei WordNet HYPDEPTH = 5 und bei AGROVOC-Thesaurus HYPDEPTH = 1 (avg. gibt den durchschnittlichen Wert für 20 Clusterläufe und std. die Standardabweichung an)

Ontologie	Purity avg $\pm$ std	InversePurity avg $\pm$ std	F-Measure avg $\pm$ std	Entropy avg $\pm$ std
ohne	**0,552 $\pm$ 0,026**	**0,455 $\pm$ 0,046**	**0,489 $\pm$ 0,035**	**1,050 $\pm$ 0,046**
WordNet	0,558 $\pm$ 0,023	0,467 $\pm$ 0,037	0,501 $\pm$ 0,031	1,047 $\pm$ 0,039
AGROVOC	**0,576 $\pm$0,023**	**0,468 $\pm$ 0,041**	**0,512 $\pm$ 0,026**	**0,998 $\pm$ 0,041**

Bei der Nutzung von WordNet als Hintergrundwissen konnte keine signifikante Verbesserung der Ergebnisse beobachtet werden. Die Steigerung der Clustergüte unter Verwendung des AGROVOC-Thesaurus beträgt bei den Purity-Werten 2,4 % (bei $\alpha = 0,5\ \%$). Bei allen übrigen Maßen ist die Steigerung nur noch für ein $\alpha = 4\ \%$ signifikant.

Die beobachtete Verbesserung der Clustergüte fällt damit nicht so hoch aus wie bei den anderen Datensätzen. Dies liegt wahrscheinlich an der fehlenden Nutzung von Worten, die aus mehreren Termen bestehen. Solche Worte kommen häufig im AGROVOC-Thesaurus vor, werden aber bei der Berechnung der "Bag of Words"-Repräsentation nicht berücksichtigt. Dadurch können diese Terme auch nur teilweise korrekt auf die Konzepte abgebildet werden.

Der folgende Abschnitt präsentiert eine Visualisierung mit Hilfe der Formalen Begriffsanalyse für die Clusterergebnisse des AGROVOC-Datensatzes basierend auf einer Clusterung mit dem AGROVOC-Thesaurus.

11.2.2 Anwendung der FBA auf landwirtschaftliche Texte

Für die visualisierte Clusterung der Abbildungen 11.2, 11.3, 11.4 und 11.5 werden die Strategien aus dem letzten Abschnitt zur Vorverarbeitung des Datensatzes angewendet. Es wurden zehn Cluster berechnet. Als Ontologie kam der AGROVOC-Thesaurus zum Einsatz. Die Schwellwerte sind $\theta_1 = 15\ \%$ und $\theta_2 = 25\ \%$. Abbildung 11.2 gibt den vollständigen und gedrehten Begriffsverband mit den zehn Clustern als Gegenstände und den Konzepten des AGROVOC-Thesaurus als Merkmale wieder. Wir nennen diesen Verband KV3.

Den Ausführungen in Kapitel 9 entnimmt man, dass die Exploration von Teilverbänden als eine Vorgehensweise zur Analyse von Begriffsverbänden dienen kann. Man untersucht dabei in einem ersten Schritt möglichst allgemeine Begriffe und visualisiert entsprechende Teilverbände. Wir nutzen im Folgenden diese Vorgehensweise zur Analyse des Verbandes KV3. Bei der Anwendung auf den Verband KV3 der Abbildung 11.2 erhält man unter anderem die in den Abbildungen 11.3, 11.4 und 11.5 dargestellten Teilverbände.

Die Analyse der drei Teilverbände von KV3 macht die Themen der in den Abbildungen hervorgehobenen Cluster deutlich. Cluster 0, 2, 3 und 7 stehen in Zusammenhang mit Wald, was wir den Konzepten WOOD INDUSTRIE und FORREST RANGE entnehmen. Die Cluster 5 und 9 haben etwas mit "Clover" (Klee) zu tun und die Cluster 1, 4, 8 mit "Professional Services". Um zu diesen Ergebnissen zu kommen, werden nur die sehr allgemeinen Begriffe, die in den Abbildungen auch

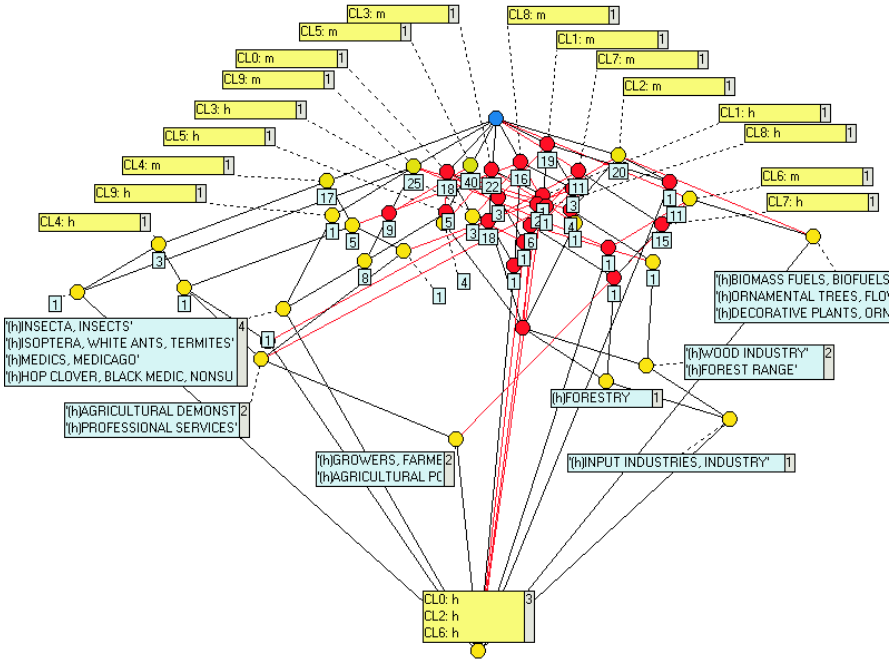

Abbildung 11.2: Vollständiger Begriffsverband KV3 für den AGROVOC-Datensatz mit 10 Clustern, $\theta_1 = 15\ \%$ und $\theta_2 = 25\ \%$

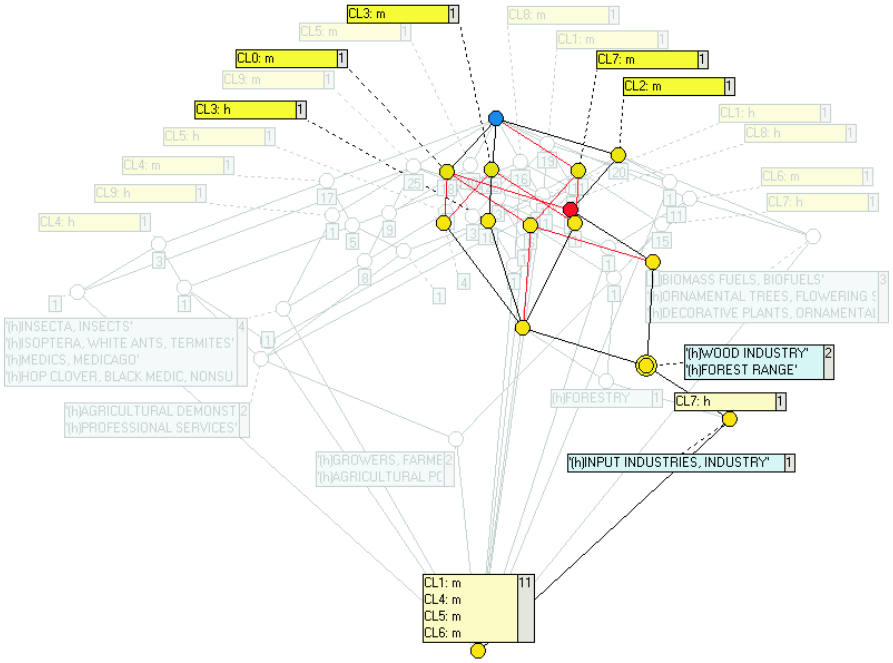

Abbildung 11.3: hervorgehobener Teilverband von KV3 mit den Clustern zum Thema “Forest”

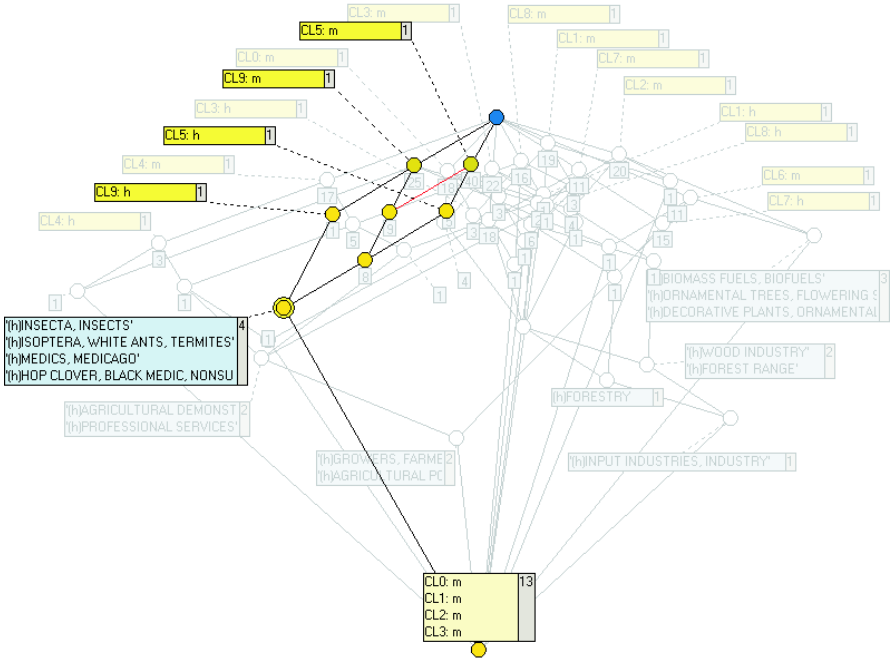

Abbildung 11.4: hervorgehobener Teilverband von KV3 mit den Clustern zum Thema “Clover”

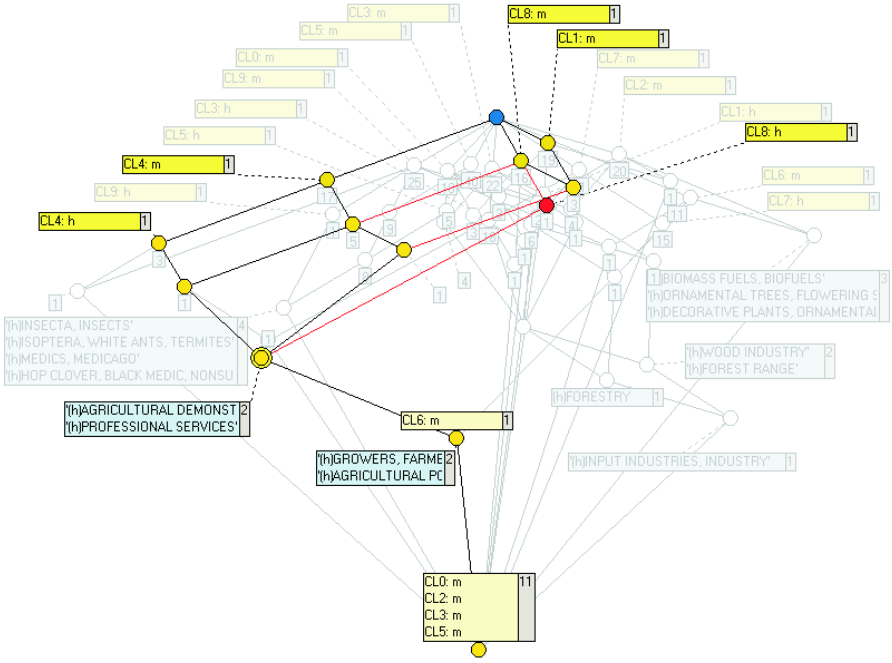

Abbildung 11.5: b) hervorgehobener Teilverband von KV3 mit den Clustern zum Thema “Activities”

hervorgehoben sind, herangezogen. Vergleicht man die Ergebnisse mit den Bezeichnern der Klassen, so stimmen diese weitestgehend überein. Einzig bei der letzten Clustergruppe mit den Clustern 1, 4, 8, deren Klasse mit “Extension Activities” überschrieben ist, wird die Verbindung nicht sofort deutlich.

Da der AGROVOC-Thesaurus sehr viele Fachbegriffe enthält, kann nur ein Experte aus diesem Gebiet eine fundierte und detaillierte Analyse der Textcluster durchführen. Dieser stand leider nicht zur Verfügung. Daher wird an dieser Stelle von einer Detailanalyse der Verbände abgesehen.

Im folgenden Abschnitt stellen wir Visualisierungen von 25 Textclustern des Getess-Datensatzes vor.

11.3 Tourismus-Web-Seiten

In diesem Abschnitt präsentieren wir für den Getess-Datensatz (vgl. Abschnitt 2.4) die Clusterergebnisse einer Clusterung mit 25 Clustern. Wir möchten dabei noch einmal die Auswirkungen auf den visualisierten Verband und die Analyseergebnisse mittels einer term- und konzeptbasierten KMeans-Clusterung untersuchen (vgl. Abschnitt 8.5 und Kapitel 9). Dazu erfolgte eine Clusterung auf einer “Bag of Words”-Repräsentation und auf einer ontologiebasierten Repräsentation. Als Ontologie wählen wir die allgemeine Ressource GermaNet. Sie bildet das deutsche Äquivalent zu WordNet und wurde in Abschnitt 6.3.3.2 eingeführt.

Wir berechnen zwei Begriffsverbände. Ein Verband, den wir TV6 nennen, nutzt die Worte der Dokumente als Merkmale und die 25 Cluster als Gegenstände. Der andere Verband, genannt KV4, basiert auf den Konzepten aus GermaNet als Merkmale und den Clustern als Gegenstände. Die Merkmalsmengen des Kontextes für die Verbandsberechnung ergeben sich mit den Schwellwerten $\theta_1 = 20\ \%$ und $\theta_2 = 35\ \%$ für TV6 und mit den Schwellwerten $\theta_1 = 20\ \%$ und $\theta_2 = 35\ \%$ für KV4. Die Visualisierungen der Verbände findet man jeweils in Abbildung 11.6 und 11.7.

Der Visualisierung in Abbildung 11.6 entnimmt man, dass Cluster 13 etwas mit Orten bzw. Inseln an der Ostsee zu tun haben muss, da sowohl die Insel Usedom als auch die Orte Bansin, Heringsdorf und Ahlbeck im äussersten Nordosten von Deutschland direkt an der Ostsee liegen. Dabei handelt es sich um Orte mit viel Tourismus. In den Dokumenten des Cluster 13 scheint es um die Orte an der Ostsee zu gehen.

Unter Berücksichtigung der präsentierten Ergebnisse analysierten wir anschließend den Verband KV4 und waren an den Orten und Inseln im konzeptbasierten Verband interessiert. Der Verband KV4 enthält zwar das Konzept INSEL,EILAND, aber keiner der Ortsnamen taucht im Verband auf. Grund sind die fehlenden Ortsnamen in GermaNet. Damit fällt ein großer Teil an Informationen der Dokumente bei der Übersetzung weg. Bei der manuellen Analyse können die Ortsnamen bei einer spätere Interpretation der Cluster von Vorteil sein, wenn man sie kennt. Bei Unkenntnis der Namen wäre eine Verbindung zu einem Oberkonzept wie z.B. OSTSEEBAD oder INSELN AN DER OSTSEE sehr hilfreich. Diese könnten sich dann im Verband widerspiegeln, wie man das auch für andere Konzepte im Verband KV4 beobachten kann. Folgendes Beispiel illustriert dies.

Bei der Analyse von Abbildung 11.7 erkennt man sehr schön die in Kapitel 9 schon beobachteten Ketten von Konzepten mit steigenden Spezifizität. Handelt es sich bei AUFENTHALTSORT um ein sehr allgemeines Konzept, so ist PENSION oder LOKAL,GASTSTÄTTE ein Unterkonzept des Konzeptes AUFENTHALTSORT. Diese durch die Ontologie bereitgestellte Information wird an dieser Stelle wieder in den Verband übernommen.

Um die Vorteile beider Ansätze (wort- und konzeptbasiert) auszunutzen, wird ein Verband, den wir KTV1 nennen, basierend auf einem Wort-Konzept-Vektor (dies entspricht der add Strategie) berechnet. Ziel ist es, sowohl Ortsnamen als auch allgemeine Konzepte der Ontologie, in einem

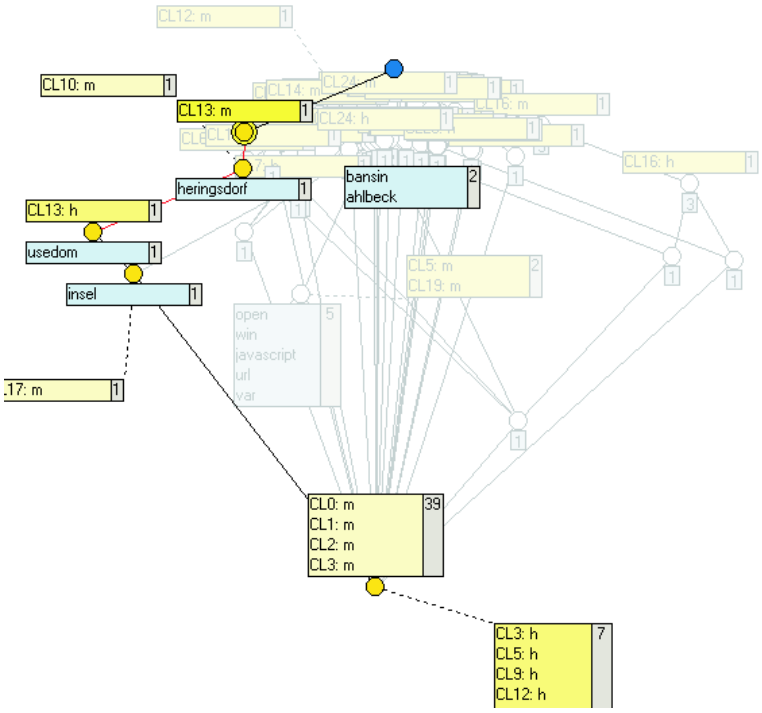

Abbildung 11.6: Begriffsverband TV6 mit hervorgehobenem Cluster 3 der Getess-Clusterung mit 25 Clustern ohne Hintergrundwissen

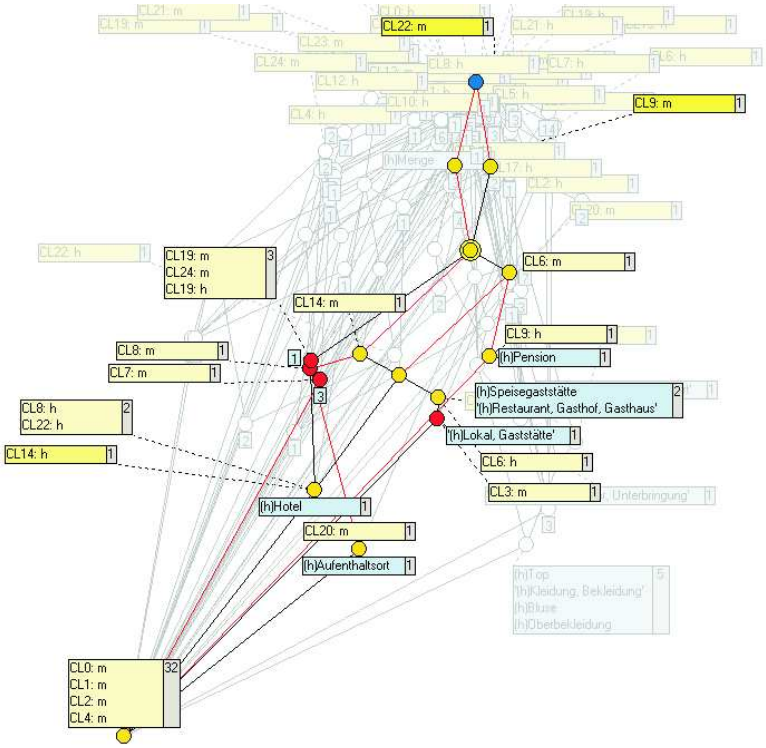

Abbildung 11.7: Begriffsverband KV4 mit hervorgehobenen Begriff erzeugt durch die Gegenstände “CL22: m”,“CL9: m” (Aufenthaltsort als Oberkonzept von Pension)

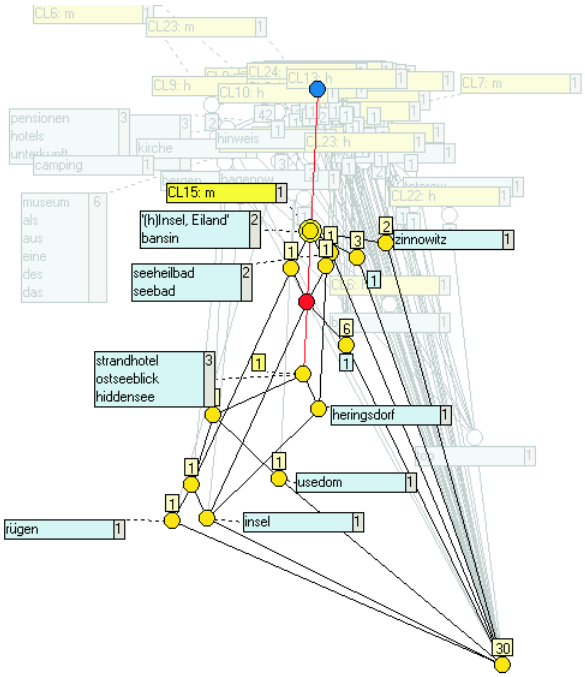

Abbildung 11.8: Begriffsverband KTV1 mit hervorgehobenem Begriff erzeugt durch den Gegenstand “CL15: m” (Term “insel” im Inhalt eines allgemeineren Begriffes als Konzept INSEL, EILAND)

formalen Begriff zu lokalisieren. Der berechnete Verband ist in Abbildung 11.8 visualisiert. Die Schwellwerte sind $\theta_1 = 10\ \%$ und $\theta_2 = 20\ \%$. Es wurden wieder 25 Cluster berechnet.

Wir entnehmen der Abbildung die aus TV6 bekannte Beziehung zwischen dem Term “Inseln” und den Ortsnamen wie “Heringsdorf”. Allerdings befindet sich das Konzept INSEL, EILAND passend zum Term “Insel” nicht im Inhalt des gleichen Begriffes, sondern im Inhalt eines viel spezielleren Begriffes. Dieses Ergebnis wurde nicht erwartet. Die Analyse des Wortes und Konzeptes zeigt, dass “Insel” 1995-mal im Korpus vorkommt und einen idf-Wert von 1,527 hat. Das Konzept INSEL, EILAND kommt 2714-mal vor und hat einen idf-Wert von 1,195. Wir vermuten, dass dieses kleine idf-Gewicht des Konzeptes dazu geführt hat, dass die beiden Terme nicht zusammen auftauchen. Auch konnten die Oberbegriffe von INSEL, EILAND wie z.B. GEOGRAPHISCHES GEBIET, GEGEND nicht im Verband identifiziert werden. Wir vermuten, dass auch hier die Gewichte für die Übername der Konzepte in die Menge der Merkmale zu klein sind. Die Nutzung größerer Gewichte würde aber die Übersichtlichkeit des Verbandes erheblich reduzieren. Dies sollte bei der Nutzung weiterer Merkmale berücksichtigt werden.

Wie wir gesehen haben, bieten wortbasierten Visualisierungen andere Informationen als konzeptbasierte. Dies tritt immer dann auf, wenn die Ontologie nur unzureichend auf die Dokumente abgestimmt ist. In einem solchen Fall ist eine zusätzliche Clusterung und Visualisierung auf der Basis von Worten ein sinnvoller Weg zur Analyse der Dokumente. Die ausschließliche Clusterung auf Wortbasis kann nicht empfohlen werden, da man ansonsten wertvolle Informationen und Beziehungen zwischen Clustern auf der Basis von generellen Konzepten der Ontologie verliert. Vielmehr wird die Anreicherung der Ontologie um die fehlenden Informationen empfohlen. Die gleichzeitige Nutzung von Worten und Konzepten bleibt eine spannende Forschungsfrage.

Zusammenfassung:

Wir konnten in diesem Kapitel zeigen, dass die Integration von Hintergrundwissen in die Vek-

torrepräsentation zum Clustern von Textdokumenten zu einer Verbesserung der Clustergüte führt. In einigen Fällen konnte keine signifikante Verbesserung der Ergebnisse beobachtet werden. Die Ergebnisse waren aber in keinem Fall signifikant schlechter. Die Anwendung der Clusterung mit Hintergrundwissen kann anhand der vorliegenden Ergebnisse für die beste Strategie nur empfohlen werden.

Die präsentierten Visualisierungen erlaubten eine einfache und intuitive Analyse der Clusterergebnisse, wobei häufig der Inhalt der Cluster schnell erfasst werden kann. Auch die Beziehungen der einzelnen Cluster untereinander helfen beim Verständnis des Clusterinhaltes.

12 Zusammenfassung und Ausblick

In dieser Arbeit haben wir die drei Methoden *Subjektives Clustern*, *Clustern mit Hintergrundwissen* und die *Beschreibung von Textclustern mit Hintergrundwissen auf der Basis der Formalen Begriffsanalyse* eingeführt. Dabei konnten wir zeigen, dass die Integration von formal repräsentiertem Hintergrund in Form einer Ontologie die Güte der Clusterergebnisse steigerte. Außerdem konnten leicht verständliche Visualisierungen der Textcluster erzeugt werden.

Subjektives Clustern: Subjektives Clustern berechnet benutzerbezogene Cluster bei gleichzeitiger Dimensionsreduktion. Ziel ist nicht wie bisher die Berechnung einer Clusterung, sondern mehrerer Clusterungen auf der Basis subjektiver benutzerbezogener Sichten. Die Sichten werden mit Hilfe der Ontologie und der Daten abgeleitet. Sie spiegeln die verschiedenen Präferenzen einzelner Benutzer wider. Der Benutzer hat die Möglichkeit, aus mehreren niedrigdimensionalen Clusterungen mit unterschiedlichen Merkmalen auszuwählen, wobei die Merkmale die Konzepte einer Ontologie sind. Die geringe Dimensionsanzahl erleichtert dem Benutzer auch die spätere Interpretation der Clusterergebnisse. Wir konnten zeigen, dass die Clusterungen basierend auf Sichten zu besseren und leichter verständlichen Ergebnissen führen. Wir wendeten das Subjektive Clustern erfolgreich auf Textdokumente aus der Praxis und zum Berechnen von Kundensegmenten anhand von Verbindungsdaten bei der Deutschen Telekom AG an.

Textclustern mit Hintergrundwissen: Bei der zweiten Methode wird das Hintergrundwissen in Form einer Ontologie während der Vorverarbeitung der Dokumente in den Clusterprozess integriert. Es konnte gezeigt werden, dass diese neue ontologiebasierte Repräsentation für Textdokumente gegenüber der herkömmlichen “Bag of Words”-Repräsentation zu einer signifikanten Steigerung der Clustergüte führt. Dazu wurde neben den verschiedenen Strategien zur Abbildung der Worte eines Textes auf die Konzepte einer Ontologie auch die Nutzung taxonomischer Beziehungen zur Steigerung der Clustergüte anhand dreier Datensätze aus der Praxis untersucht. Einer der Textkorpora besteht aus Nachrichtentexten der Agentur Reuters, einer aus Lernmaterialen der Programmiersprache Java und einer aus Texten landwirtschaftlicher Fachzeitschriften. Die Anwendung der Clusterung mit Hintergrundwissen kann anhand der vorliegenden empirischen Ergebnisse auf alle Fälle empfohlen werden, da die Ergebnisse immer gleich gut und meistens sogar besser als die Referenzclusterung basierend auf der “Bag of Words”-Repräsentation waren.

Beschreibung von Textclustern mit Hintergrundwissen: Erstmals wurden Verfahren der Formalen Begriffsanalyse zur Präsentation von Textclustern verwendet. Die visualisierten Verbände liefern eine für Menschen leicht verständliche Beschreibung der berechneten Textcluster. Grund dafür sind die berechneten Beziehungen zwischen den Textclustern, die Gemeinsamkeiten und Unterschiede zwischen den Clustern hervorheben. Die in die Textrepräsentation integrierte Ontologie führt zu einer weiteren Verbesserung der Verständlichkeit. Sie strukturiert den Verband durch die bereitgestellten Oberkonzepte und ermöglicht so die einfache Exploration des Verbandes ausgehend von allgemeinen Begriffen bis hin zu speziellen.

Wir konnten dies vor allen Dingen anhand von Textclustern auf dem Reuters-Korpus zeigen. Experimente auf anderen praxisnahen Textkopora bestätigten diese Ergebnisse.

Auf weitere Anwendungen der entwickelten Methoden wurde ebenfalls in der Arbeit eingegangen. So erlauben die Methoden z.B. die Strukturierung von Textdokumenten und stellen damit einen ersten Schritt von unstrukturierten zu strukturierten Informationen in einem Wissensportal oder einer eLearning-Umgebung dar. Im Folgenden werden wir auf offene Forschungsfragen für die Zukunft eingehen, die in Zusammenhang mit dieser Arbeit stehen.

Die Berechnung der Sichten beim Subjektiven Clustern erfolgt zur Zeit mittels eines Top-Down-Ansatzes, d.h. die allgemeineren Konzepte werden schrittweise verfeinert. Alternativ könnte man auch mit den Blattknoten der Taxonomie einer Ontologie starten. Dies führt zu einem so genannten Bottom-Up-Ansatz. Ein Vergleich der dann berechneten Sichten und Clusterungen mit dem Top-Down-Ansatz scheint vielversprechend. Sowohl für das Subjektive Clustern als auch für die Visualisierung der Begriffsverbände stellt die Durchführung von Studien auf der Basis von speziell designten Benutzerschnittstellen zur Untersuchung der Anwendbarkeit der berechneten Visualisierung eine interessante Aufgabe dar.

Um eine weitere Steigerung der Clustergüte durch den Einsatz von Hintergrundwissen zu erreichen, stellt der Einsatz verbesserter Strategien zur Erkennung von Wortsinnen, aber auch die Nutzung weiterer Beziehungen zwischen den Konzepten einen sinnvollen Schritt dar. Auch eine Wortarterkennung, wie sie im Bereich der Sprachverarbeitung entwickelt wird, sollte in den Prozess integriert werden. Die gezielte Auswahl von Oberkonzepten verspricht ebenfalls Verbesserungen der Clustergüte. Wie in der Arbeit gezeigt, ist die Gewichtung der Termvektoren ein weiterer wichtiger Faktor für gute Textcluster. Der Einsatz von alternativen Maßen zur Gewichtung der Termvektoren sollte eruiert werden. Entsprechende Maße wurden in den vergangen Jahren im Bereich des Information Retrieval entwickelt. Die Anwendung dieser Maße auf die ontologiebasierte Textdokument-Repräsentation wäre eine weitere interessante Aufgabe für die Zukunft.

Erste positive Ergebnisse für die Nutzung von Latent Semantic Indexing (LSI) in Kombination mit Hintergrundwissen konnten in der Arbeit empirisch gezeigt werden. Der Einsatz von Probabilistic Latent Semantic Indexing (PLSI) auf der ontologiebasierten Repräsentation bietet viel Potential für die weitere Verbesserung der Ergebnisse.

Beim Einsatz der Formalen Begriffsanalyse zur Beschreibung berechneter Textcluster wurde beobachtet, dass teilweise die hierarchischen Beziehungen der einzelnen Konzepte in der Verbandsstruktur wiedergefunden wurden. Dies trifft aber nicht für alle Beziehungen zu. Wünschenswert wäre ein Verband, der für alle Konzepte der Ontologie deren hierarchische Beziehung enthält. Um dies zu erreichen, müsste man diese Beziehungen explizit in die Verbandsstruktur übernehmen.

Der schon angesprochene Einsatz von verbesserter Wortsinn- und Wortarterkennung kann nicht nur zur Steigerung der Clustergüte, sondern auch zur Verbesserung der visualisierten Ergebnisse beitragen, da die Anzahl der Fehler beim Abbilden der Worte auf die Konzepte reduziert wird. Die Kenntnis der Wortart einzelner Worte und Konzepte bei der Präsentation im Verband würde außerdem eine leichtere Interpretation des Inhaltes erlauben. Eine gezieltere Auswahl von Termen zur Visualisierung der Textcluster kann sowohl durch eine geänderte Gewichtung der Terme im Vektor als auch durch Methoden zur Merkmalsextraktion erreicht werden. Vielversprechend wäre auch die Nutzung von Self Organizing Maps zum Clustern und Visualisieren der Clusterung auf der Basis einer ontologiebasierten Repräsentation.

Die Arbeit stellt einen wichtigen Schritt zur Nutzung von formal repräsentiertem Hintergrundwissen in Form von Ontologien im Knowledge Discovery oder Data, Text und Web Mining dar. Weitere neue Anwendungsfelder sind das so genannte Semantic Web Mining (vgl. [212], [19]). Dabei geht es auf der einen Seite um die Nutzung von Data Mining Verfahren zur Unterstützung des Aufbaus

des Semantic Web, genannt Ontology Learning (vgl. [153]). Auf der anderen Seite steht die Analyse von strukturierten Daten und Informationen durch die Verfahren und Methoden des Data, Text und Web Minings im Vordergrund, wobei wir mit dieser Arbeit einen Beitrag zur Erreichung des zweiten Zieles liefern.

Teil IV

Anhang

A Text Mining Environment

Die Text Mining Environment (TME) ist ein Tool zum Clustern und Klassifizieren von Textdokumenten. Es wurde innerhalb des KAON-Frameworks entwickelt.[1] Das Tool bildet die Grundlage für die in der Arbeit durchgeführten Evaluierungen der Kapitel 8 und 11. Sämtliche in der Arbeit referenzierten Schritte, angefangen von der Vorverarbeitung der Dokumente über die Anreicherung der Termvektoren mit Hintergrundwissen bis hin zum Clustern und Klassifizieren, können mit dem Tool durchgeführt werden. Alle Schritte lassen sich parametrisieren. Die Parameter werden in einer XML-basierten Konfigurationsdatei abgelegt. Im Folgenden beschreiben wir anhand des prinzipiellen Ablaufs einer Clusterung die Elemente der TME, wobei wir auch auf ausgewählte Elemente der grafischen Oberfläche eingehen.

Die Steuerung der TME erfolgt über die Konfigurationsdatei bzw. den entsprechenden grafischen Dialog, der im linken unteren Teil der Abbildung A.1 zu sehen ist. Über den Dialog kann man neben den Pfaden zu den Textdokumenten, der Stopwortliste oder den Ergebnisfiles auch die Nutzung des Hintergrundwissens aktivieren und steuern oder die Anzahl der Cluster festlegen.

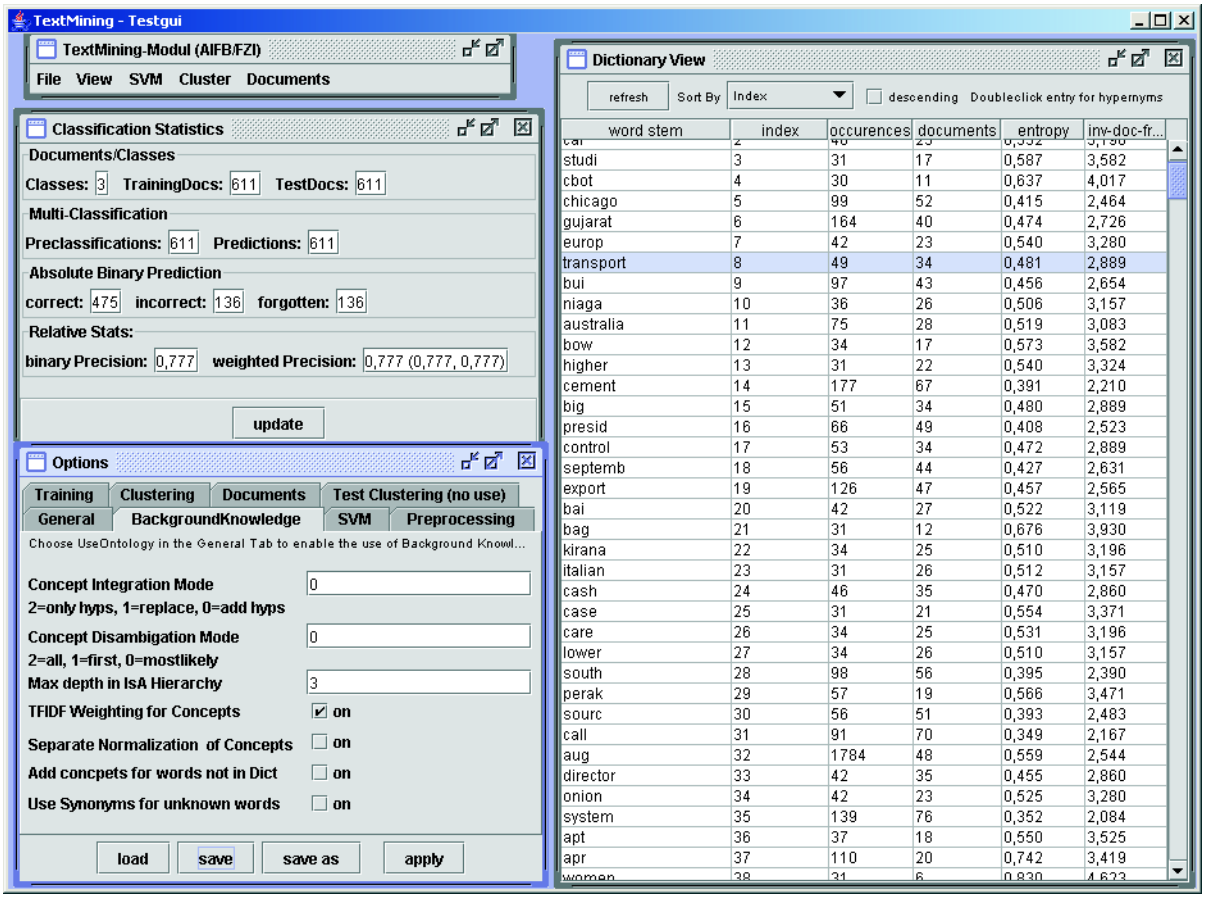

Abbildung A.1: Screenshot der Text-Mining-Umgebung mit dem Optionsdialog, dem Wörterbuch und dem Ergebnisfenster

[1] Mehr zu den KAON-Tools findet man unter: http://kaon.semanticweb.org

Abbildung A.2: Screenshot der Text Mining Umgebung mit der Hypernym-Ausgabe für das Wort “Transport”

Ausgehend von einer Menge von Dokumenten erfolgt die Vorverarbeitung in der TME in mehreren Schritten. Das Ergebnis der Vorverarbeitung eines Dokumentes ist ein so genannter Sparsevektor. In einem ersten Schritt wird ein internes Wörterbuch aufgebaut, auf das über einen Dialog zugegriffen werden kann. Einen Ausschnitt gibt Abbildung A.1 rechts wieder. Man entnimmt der Tabelle für jeden Wortstamm die absolute Häufigkeit im Korpus, die Anzahl der Dokumente, in denen der Wortstamm vorkommt und die Gewichte wie z.B. die “inverted document frequency” des tfidf-Maßes (vgl. Abschnitt 4.2.5.1).

Nach dem Aufbau des Wörterbuches wird jedes Dokument noch einmal verarbeitet und der entsprechende “Bag of terms” bestehend aus Worten, Wortstämmen oder Konzepten wird abgeleitet. Die Abbildung der Worte auf die Konzepte kann durch einen Klick auf den Wortstamm in der Wörterbuchtabelle nachvollzogen werden. Ein Beispiel für den Term “transport” unter Verwendung von WordNet gibt Abbildung A.2 wieder.

Die Clusteralgorithmen sind in Weka[2] implementiert, so dass das Ergebnis der Dokumentvorverarbeitung, der Sparsevektor, ins Weka-interne Sparsevektorformat überführt wird. Damit stehen für den vorverarbeiteten Vektor auch sämtliche in Weka implementierten Algorithmen zur Verfügung. Der einfache KMeans-Algorithmus ist bereits in Weka enthalten und kann entsprechend auf die vorverarbeiteten Daten angewendet werden. Die Bi-Sec-KMeans Variante wurde neu implementiert. Die Ergebnisse des Clusterlaufes werden anschließend wieder in die TME übernommen und stehen nun zu Analysezwecken zur Verfügung. Eine Zusammenfassung des Clusterergebnisses liefert der Dialog “Statistics” links oben in Abbildung A.1.

Für die Auswertung der Ergebnisse steht eine Liste aller gegeben Dokumentklassen zur Verfügung. Die TME geht davon aus, dass die Dokumente über eine gegebene Klassifikation verfügen. Diese wird beim Einlesen der Dokumente im System abgelegt und auf sie kann, wie in Abbildung A.3 oben zu sehen, über den Dialog “Document Classes” zugegriffen werden. In unserem

[2] http://www.cs.waikato.ac.nz/ml/weka/

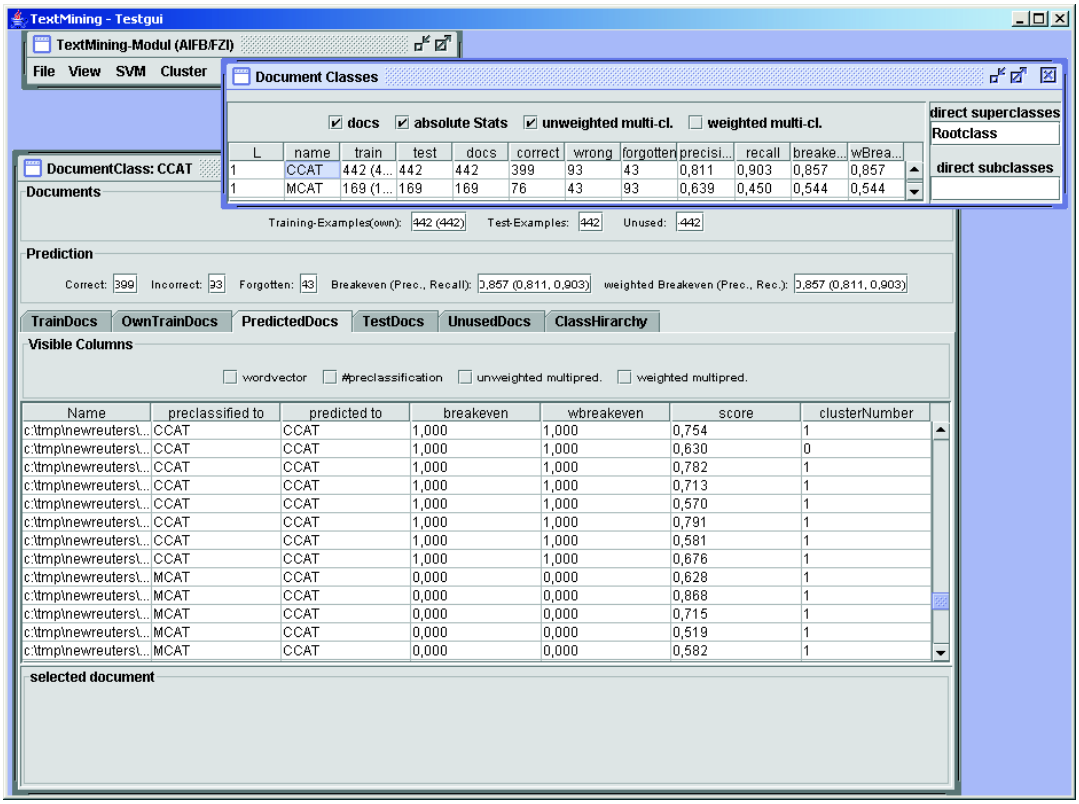

Abbildung A.3: Screenshot der Text-Mining-Umgebung mit der Liste der Dokumentklassen und der Liste der Dokumente einer Klasse

Beispiel war der Datensatz in zwei Klassen aufgeteilt. Diese Informationen werden ausschließlich für die Evaluierung der Clusterung verwendet. Man erhält im gleichen Dialog für jede Klasse nach einem erfolgreichen Clusterlauf Zugriff auf Precision, Recall usw. Werte der Klasse. Mit einem Doppelklick auf einen Klassennamen öffnet sich eine weiteres Fenster. Es gibt für diese Klasse die Menge der Dokumente wieder (siehe Abbildung A.3 unten). Unter dem Reiter “predictedDocs” findet man neben den Dokumenten, die in diese Klasse vorhergesagt wurden, auch den passenden Cluster dieses Dokumentes und den Abstand zum Zentroiden (Score).

Ein Klick auf eines der Dokumente öffnet ein weiteres Fenster. Ein Reiter (siehe Abbildung A.4 links unten) gibt den Text des Dokumentes, ein anderer Reiter (siehe Abbildung A.4 rechts unten) gibt den aktuellen Wortvektor des Systems für dieses Dokument wieder. Beim gegebenen Beispielvektor handelt es sich um einen gemischten Vektor aus Wortstämmen und Konzepten. Dies erkennt man am führenden “(h)” bei einigen Termen des Vektors, die die Konzepte markieren. Mit Hilfe der Dialoge kann man neben der Güte der Clusterung auch falsch geclusterte Dokumente identifizieren und explorieren. Bei der Suche nach der Ursache von fehlerhaft zugeordneten Dokumenten wird man durch die leicht zu erreichenden Termvektoren effektiv unterstützt.

Neben der Liste aller gegebenen Klassen, kann man die Analyse der Ergebnisse auch clusterbezogen durchführen. Der Dialog “Clusterliste” (siehe Abbildung A.5 links unten) gibt für jeden Cluster die Liste der verwendeten Terme und deren Gewichte im Cluster wieder (“(h)” kennzeichnet auch in dieser Tabelle Konzepte). Auch die Clustergröße, die größte Klasse der gegebenen Klassen im Cluster und das Label dieser Klasse kann dem Dialog entnommen werden. Abbildung A.5 rechts gibt den Clustergrafen wieder. Er wird in Abschnitt 9.3.3.2 ausführlich beschrieben.

Die TME entstand in Zusammenarbeit mit den Studenten Gert Pache, Henning Blum und Boris Lauser sowie in den Arbeiten [181, 146].

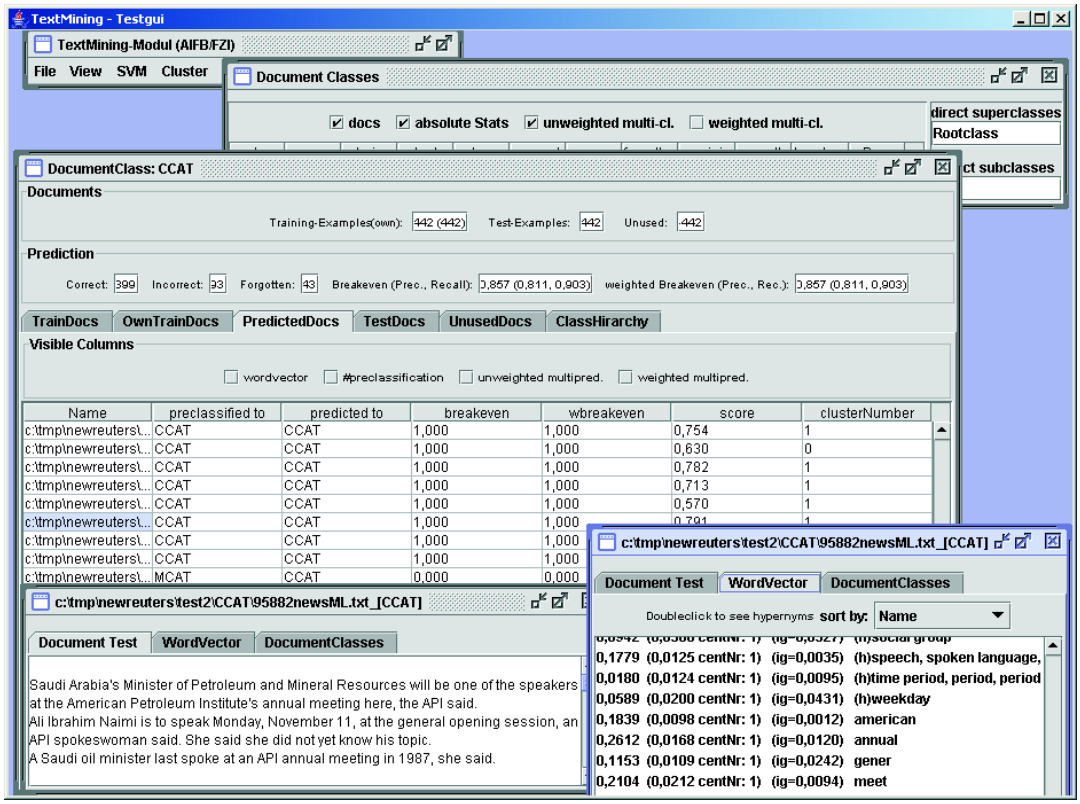

Abbildung A.4: Screenshot der Text-Mining-Umgebung mit der Liste der Dokumente einer Klasse und für ein Dokument dieser Klasse der Text und der zugehörige “Bag of Terms”

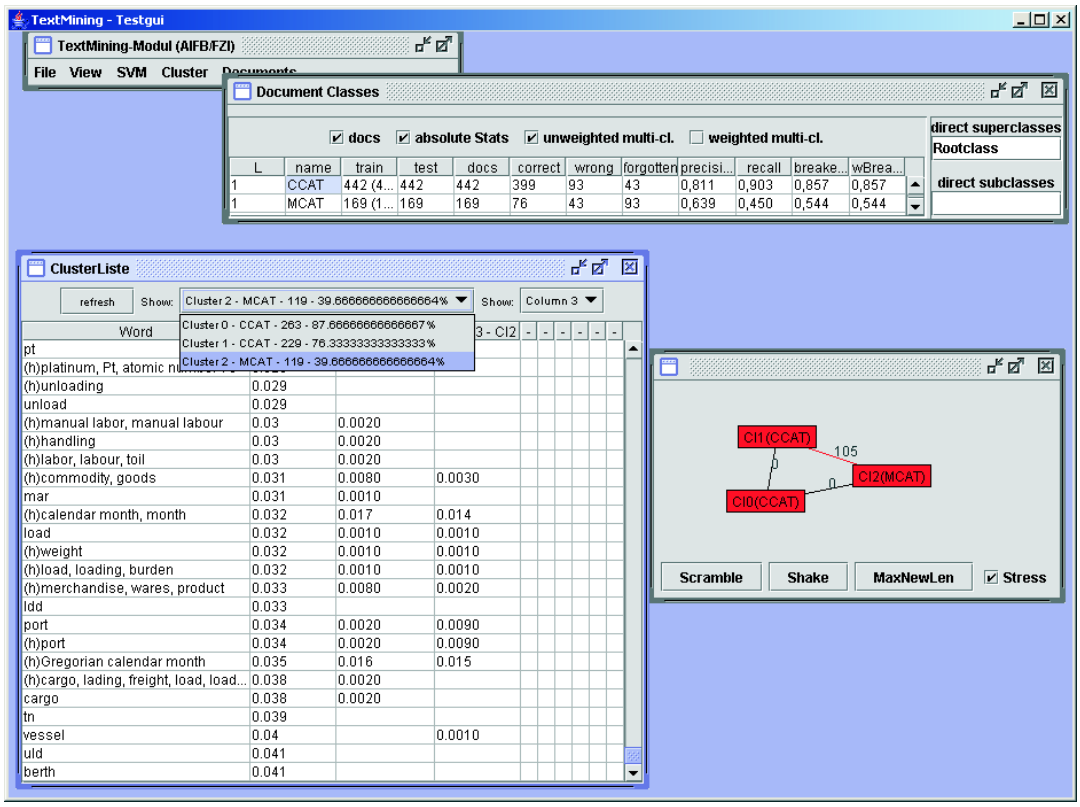

Abbildung A.5: Screenshot der Text-Mining-Umgebung mit der Clusterliste, dem Clustergrafen und der Liste der Dokumentklassen

B Ontologien

Im Folgenden findet man einen Teil einer in RDF serialisierte Version der Beispielontologie aus Abbildung 8.19 des Abschnitts 8.5.2.

```
<?xml version='1.0' encoding='UTF-8'?> <!DOCTYPE rdf:RDF [
    <!ENTITY rdfs 'http://www.w3.org/2000/01/rdf-schema#'>
    <!ENTITY kaon 'http://kaon.semanticweb.org/2001/11/kaon-lexical#'>
    <!ENTITY rdf 'http://www.w3.org/1999/02/22-rdf-syntax-ns#'>
]>

<?include-rdf logicalURI="http://kaon.semanticweb.org/2001/11/kaon-root"
  physicalURI="jar:file:/C:/users/build/kaon_build_root/kaon/release/lib/kaonapi.jar!/edu/unika/aifb/kaon/api/res/kaon-root.xml"?>
<?include-rdf logicalURI="http://kaon.semanticweb.org/2001/11/kaon-lexical"
  physicalURI="jar:file:/C:/users/build/kaon_build_root/kaon/release/lib/kaonapi.jar!/edu/unika/aifb/kaon/api/res/kaon-lexical.xml"?>

<?model-attribute key="OIModel.version" value="165"?>

<rdf:RDF xml:base="http://www.aifb.uni-karlsruhe.de/beispiel"
    xmlns:rdfs="&rdfs;"
    xmlns:kaon="&kaon;"
    xmlns:rdf="&rdf;">

<rdfs:Class rdf:ID="1052723569763-1045505043">
    <rdfs:subClassOf rdf:resource="#software"/>
</rdfs:Class>
<kaon:Label rdf:ID="1052723569763-1967682672" kaon:value="ziff">
    <kaon:inLanguage rdf:resource="&kaon;en"/>
    <kaon:references rdf:resource="#1052723569763-1045505043"/>
</kaon:Label>
<kaon:Stem rdf:ID="1052723569763-894313619" kaon:value="ziff">
    <kaon:inLanguage rdf:resource="&kaon;en"/>
    <kaon:references rdf:resource="#1052723569763-1045505043"/>
</kaon:Stem>
<kaon:Label rdf:ID="1052723569773-1115815941" kaon:value="world cup">
    <kaon:inLanguage rdf:resource="&kaon;en"/>
    <kaon:references rdf:resource="#1052723569773-53037770"/>
</kaon:Label>
<kaon:Label rdf:ID="1052723569773-1132729702" kaon:value="server">
    <kaon:inLanguage rdf:resource="&kaon;en"/>
    <kaon:references rdf:resource="#1052723569773-265089400"/>
</kaon:Label>

...

<kaon:Synonym rdf:ID="1052725499638-41283456" kaon:value="database ">
    <kaon:inLanguage rdf:resource="&kaon;en"/>
    <kaon:references rdf:resource="#1052725499638-234471161"/>
</kaon:Synonym>
<kaon:Stem rdf:ID="1052725502121-134211956" kaon:value="develop">
    <kaon:inLanguage rdf:resource="&kaon;en"/>
    <kaon:references rdf:resource="#1052725502121-1717306782"/>
</kaon:Stem>
<rdfs:Class rdf:ID="1052725502121-1717306782">
    <rdfs:subClassOf rdf:resource="#software"/>
</rdfs:Class>
<kaon:Synonym rdf:ID="1052725502121-541479" kaon:value="development ">
    <kaon:inLanguage rdf:resource="&kaon;en"/>
    <kaon:references rdf:resource="#1052725502121-1717306782"/>
</kaon:Synonym>
<kaon:Label rdf:ID="1052725502121-735531866" kaon:value="develop">
    <kaon:inLanguage rdf:resource="&kaon;en"/>
    <kaon:references rdf:resource="#1052725502121-1717306782"/>
</kaon:Label>
<kaon:Label rdf:ID="1052725511695-1637199877" kaon:value="featur">
    <kaon:inLanguage rdf:resource="&kaon;en"/>
    <kaon:references rdf:resource="#1052725511695-178584536"/>
</kaon:Label>
<rdfs:Class rdf:ID="1052725511695-178584536">
    <rdfs:subClassOf rdf:resource="#software"/>
</rdfs:Class>
<kaon:Synonym rdf:ID="1052725511695-71642800" kaon:value="feature ">
    <kaon:inLanguage rdf:resource="&kaon;en"/>
    <kaon:references rdf:resource="#1052725511695-178584536"/>
</kaon:Synonym>
<kaon:Stem rdf:ID="1052725511695-958739121" kaon:value="featur">
    <kaon:inLanguage rdf:resource="&kaon;en"/>
    <kaon:references rdf:resource="#1052725511695-178584536"/>
</kaon:Stem>
<rdfs:Class rdf:ID="Sport">
    <rdfs:subClassOf rdf:resource="&kaon;Root"/>
</rdfs:Class>
<rdfs:Class rdf:ID="finance">
    <rdfs:subClassOf rdf:resource="&kaon;Root"/>
```

```
</rdfs:Class>
<rdfs:Class rdf:ID="software">
    <rdfs:subClassOf rdf:resource="&kaon;Root"/>
</rdfs:Class>
</rdf:RDF>
```

C Beispielkontext

Im Folgenden ist der Kontext zum Datensatz DS1 aus Abschnitt 5.5 dargestellt.

	CL0	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CL9
agreem				X	X		X			X
fiscal				X			X			X
quarter		X	X	X			X			X
billion				X	X		X			X
make		X			X		X			X
injuri		X	X				X			X
hard	X	X					X			X
network	X				X		X			X
continu	X			X			X			X
result				X			X			X
contract		X			X		X			X
loss				X			X			X
combin	X				X		X			X
cent	X			X			X			X
offic	X						X			X
manag	X	X				X	X			X
execut	X						X			X
restructur				X			X			X
posit				X	X		X			X
club		X					X			X
korea	X	X	X				X			X
month		X		X	X		X			X
million	X			X	X		X			X
june	X			X			X			X
spain		X	X				X			X
softwar	X				X		X	X		X
includ				X	X		X	X		X
financi	X			X	X		X	X		X
base	X			X	X	X	X	X		X
secur	X				X		X	X		X
servic					X		X	X		X
support							X	X		X
file						X	X	X		X
featur						X	X	X		X
develop	X						X	X		X
fix						X	X	X		X
releas	X					X	X	X		X
experi		X					X			X
compani	X			X	X		X			X
dai	X				X		X			X
cut				X			X			X
investor	X						X			X
onlin					X		X			X
market	X			X			X			X
us					X	X	X	X		X
exchang	X				X		X	X		X
busi				X	X		X			X
internet	X			X	X		X			X
offer		X			X		X			X
announc		X		X			X			X
cash	X			X			X			X
process					X	X	X			X
payment				X	X		X			X
success		X					X			X
bank				X	X		X			X
contact	X					X	X			X
system	X					X	X			X
share	X			X		X	X			X

	CL0	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CL9
problem			X			X	X			X
improv				X		X	X			X
fast		X				X	X			X
made		X					X			X
stock	X			X			X			X
chief	X						X			X
put	X			X			X			X
percent	X			X			X			X
revenu				X	X		X			X
nasdaq	X			X	X		X			X
european		X					X			X
remain		X		X			X			X
becaus		X		X			X			X
trade	X			X			X			X
move	X	X		X			X			X
bug						X	X			X
credit				X	X		X			X
part	X	X					X			X
modul						X	X			X
mail					X	X	X			X
rest			X				X			X
sign		X			X		X			X
deal	X	X		X	X		X			X
web						X	X			X
electron	X				X	X	X			X
server						X	X			X
expect				X		X	X			X
year	X	X			X		X			X
plai		X	X				X			X
set	X				X		X	X		X
nate		X			X		X		X	X
end				X			X	X	X	X
back	X						X		X	X
team		X				X	X		X	X
good		X					X		X	X
peopl	X						X		X	X
attack		X					X		X	X
final		X	X				X		X	X
publish		X	X				X		X	X
player		X	X				X		X	X
jun		X	X				X		X	X
group	X						X		X	X
mondai	X			X			X		X	X
ball		X					X		X	X
world	X	X	X				X		X	X
coach		X	X				X		X	X
win		X					X		X	X
job						X	X		X	X
cup		X	X				X		X	X
england							X		X	X
time		X					X		X	X
side		X				X	X		X	X
defend		X					X		X	X
great							X		X	X
score							X		X	X
goal		X					X		X	X
brazil							X		X	X
game		X					X		X	X
minut							X		X	X

Abbildung C.1: Kontext zu Datensatz DS1 (Gegenstände und Merkmale sind vertauscht)

Die Abbildung gibt zehn Gegenstände und 117 Merkmale wieder. Die Gegenstände sind Cluster

eines KMeans-Clusterlaufs für den DS1-Datensatz. Die Merkmale sind Wortstämme. Die Cluster sind folgendermaßen gelabelt: CL0 = Finance(3); CL1 = Soccer(3); CL2 = Soccer(1); CL3 = Finance(3); CL4 = Finance(1); CL5 = Software(4); CL6 (0); CL7 = Software(3); CL8 = Soccer(3); CL9 (0). Die Zahl in Klammern gibt die Anzahl der Dokumente im Cluster wieder.

D Texte des Reuters-Datensatzes

In dem folgenden Kapitel werden beispielhaft Texte aus dem Reuters-Korpus zu Illustrationszwecken wiedergegeben. Dabei handelt es sich im ersten Fall um Dokumente der Klasse “earn” und im zweiten Fall um Dokumente der Klasse “sugar”.

D.1 Texte der Klasse “earn”

Text 1 (NEWID=21014)

```
Shr 96 cts vs 87 cts
    Net 10.8 mln vs 9,671,000
    Revs 103.9 mln vs 97.5 mln
    Nine mths
    Shr 2.73 dlrs vs 2.62 dlrs
    Net 30.7 mln vs 29.3 mln
    Revs 325.7 mln vs 302.8 mln
 Reuter
```

Text 2 (NEWID=15002)

```
Shr 10 cts vs 32 cts
    Net 975,000 vs 3,145,000
    Sales 159.1 mln vs 147.3 mln
 Reuter
```

Text 3 (NEWID=5012)

```
Qtrly 15 cts vs 15 cts prior
    Pay May Eight
    Record April 24
 Reuter
```

D.2 Texte der Klasse “sugar”

Text 1 (NEWID=5175)

```
Taiwan is not expected to export sugar this year because of
falling production and growing domestic consumption, state-owned
Taiwan Sugar Corp said.
    A company spokesman told Reuters this will be the first
time in more than 40 years Taiwan has not exported sugar. Last
year, sugar exports totalled 149,755 tonnes.
    He said the actual production during the 1986/87 season
(November/May) is about 480,000 tonnes, barely enough to meet
local consumption. This compares with actual 1985/86 output of
```

570,000. He said the production fall was due to typhoon damage to more than 6,000 hectares of canefields last year.
 REUTER

Text 2 (NEWID=10014)

The New York Coffee, Sugar and Cocoa Exchange (CSCE) elected former first vice chairman Gerald Clancy to a two-year term as chairman of the board of managers, replacing previous chairman Howard Katz.

Katz, chairman since 1985, will remain a board member.

Clancy currently serves on the Exchange board of managers as chairman of its appeals, executive, pension and political action committees.

The CSCE also elected Charles Nastro, executive vice president of Shearson Lehman Bros, as first vice chairman. Anthony Maccia, vice president of Woodhouse, Drake and Carey, was named second vice chairman, and Clifford Evans, president of Demico Futures, was elected treasurer.
 Reuter

E Reuters-Klassen

Die folgende Tabelle gibt alle 82 verwendeten Reuters-Klassen, die Anzahl der Dokumente pro Klassen sowie den Anteil der Klasse an allen 12344 Dokumenten wieder.

Name	**Anzahl Dok/Kl**	Anteil	Anzahl Dok/Kl kumuliert	Anteil kumuliert
austdlr	1	0,01%	1	0,01%
saudriyal	1	0,01%	2	0,02%
hk	1	0,01%	3	0,02%
wool	1	0,01%	4	0,03%
naphtha	1	0,01%	5	0,04%
rand	1	0,01%	6	0,05%
soy-meal	1	0,01%	7	0,06%
tapioca	1	0,01%	8	0,06%
fishmeal	1	0,01%	9	0,07%
barley	1	0,01%	10	0,08%
nzdlr	1	0,01%	11	0,09%
plywood	2	0,02%	13	0,11%
inventories	2	0,02%	15	0,12%
rapeseed	2	0,02%	17	0,14%
f-cattle	2	0,02%	19	0,15%
coconut	2	0,02%	21	0,17%
cpu	2	0,02%	23	0,19%
l-cattle	2	0,02%	25	0,20%
rice	3	0,02%	28	0,23%
propane	3	0,02%	31	0,25%
groundnut	3	0,02%	34	0,28%
palm-oil	3	0,02%	37	0,30%
stg	4	0,03%	41	0,33%
platinum	4	0,03%	45	0,36%
soybean	4	0,03%	49	0,40%
jet	4	0,03%	53	0,43%
potato	5	0,04%	58	0,47%
nickel	5	0,04%	63	0,51%
instal-debt	5	0,04%	68	0,55%
yen	6	0,05%	74	0,60%
income	7	0,06%	81	0,66%
corn	8	0,06%	89	0,72%
tea	9	0,07%	98	0,79%
lei	12	0,10%	110	0,89%
fuel	13	0,11%	123	1,00%

Name	**Anzahl**	Anteil	# kumuliert	Anteil kumuliert
lumber	13	0,11%	136	1,10%
housing	16	0,13%	152	1,23%
hog	16	0,13%	168	1,36%
silver	16	0,13%	184	1,49%
heat	16	0,13%	200	1,62%
orange	18	0,15%	218	1,77%
retail	19	0,15%	237	1,92%
lead	19	0,15%	256	2,07%
strategic-metal	19	0,15%	275	2,23%
zinc	20	0,16%	295	2,39%
meal-feed	21	0,17%	316	2,56%
wheat	21	0,17%	337	2,73%
wpi	24	0,19%	361	2,92%
cotton	27	0,22%	388	3,14%
carcass	29	0,23%	417	3,38%
pet-chem	29	0,23%	446	3,61%
tin	32	0,26%	478	3,87%
gas	32	0,26%	510	4,13%
dlr	37	0,30%	547	4,43%
rubber	40	0,32%	587	4,76%
bop	47	0,38%	634	5,14%
nat-gas	48	0,39%	682	5,52%
alum	48	0,39%	730	5,91%
ipi	49	0,40%	779	6,31%
jobs	50	0,41%	829	6,72%
iron-steel	51	0,41%	880	7,13%
reserves	51	0,41%	931	7,54%
livestock	57	0,46%	988	8,00%
cocoa	59	0,48%	1047	8,48%
copper	62	0,50%	1109	8,98%
cpi	75	0,61%	1184	9,59%
oilseed	78	0,63%	1262	10,22%
veg-oil	93	0,75%	1355	10,98%
money-supply	113	0,92%	1468	11,89%
gnp	117	0,95%	1585	12,84%
gold	121	0,98%	1706	13,82%
coffee	124	1,00%	1830	14,83%
sugar	145	1,17%	1975	16,00%
ship	203	1,64%	2178	17,64%
interest	262	2,12%	2440	19,77%
trade	441	3,57%	2881	23,34%
crude	482	3,90%	3363	27,24%
grain	488	3,95%	3851	31,20%
money-fx	572	4,63%	4423	35,83%
defnoclass	1975	16,00%	6398	51,83%
acq	2186	17,71%	8584	69,54%
earn	3760	30,46%	12344	100,00%
Summe	12344			

F Ausgewählte Ergebnistabellen

Dieses Kapitel gibt die Ergebnisse für den Clusterlauf auf dem Reuters-Datensatz ohne Hintergrundwissen wieder. Tabelle F.1 fasst die Purity für alle Clusterergebisse ohne Hintergrundwissen zusammen.

Alle übrigen Tabellen zum Clustern mit Hintergrundwissen für den Reuters-Datensatz findet man unter: http://www.aifb.uni-karlsruhe.de/WBS/aho/clustering/

Tabelle F.1: Purity für Clustering ohne Hintergrundwissen, passend zu Tabelle 8.2, Durchschnitt ±Standardabweichung von 20 Wiederholungen

PRC	#Klassen	# Wortstämme	#Worte	Gewichtung	Prunethreshold	5	10	20	30	50	60	70	100
max20	82	6494	91749	tfidf	0	9,10% ± 0,003	15,90% ± 0,009	24,90% ± 0,011	30,60% ± 0,016	36,00% ± 0,017	38,50% ± 0,023	40,40% ± 0,017	45,20% ± 0,013
		2310	84263	tfidf	5	9,20% ± 0,002	16,20% ± 0,008	26,10% ± 0,019	32,50% ± 0,018	39,90% ± 0,021	42,40% ± 0,014	44,60% ± 0,017	48,00% ± 0,023
		594	64455	tfidf	30	9,20% ± 0,004	16,90% ± 0,005	28,20% ± 0,014	34,90% ± 0,019	44,70% ± 0,016	47,00% ± 0,017	48,90% ± 0,019	53,10% ± 0,013
		6494	91749	ohne	0	8,80% ± 0,002	14,80% ± 0,009	22,10% ± 0,014	27,30% ± 0,011	34,10% ± 0,015	36,30% ± 0,018	38,60% ± 0,016	43,50% ± 0,017
		2310	84263	ohne	5	8,80% ± 0,002	14,90% ± 0,006	22,90% ± 0,012	27,70% ± 0,013	34,10% ± 0,015	36,70% ± 0,019	39,00% ± 0,015	43,60% ± 0,016
		594	64455	ohne	30	8,80% ± 0,002	14,90% ± 0,006	23,10% ± 0,012	28,10% ± 0,013	34,40% ± 0,013	36,70% ± 0,02	39,30% ± 0,013	43,70% ± 0,015
min15-max20	46	6073	79758	tfidf	0	10,40% ± 0,003	18,10% ± 0,01	28,30% ± 0,02	34,30% ± 0,016	42,50% ± 0,019	44,60% ± 0,022	44,80% ± 0,012	49,80% ± 0,015
		2129	72721	tfidf	5	10,50% ± 0,003	18,80% ± 0,01	30,40% ± 0,016	37,20% ± 0,02	46,40% ± 0,024	47,90% ± 0,014	50,00% ± 0,018	54,00% ± 0,014
		544	54763	tfidf	30	10,60% ± 0,005	19,80% ± 0,008	33,50% ± 0,017	42,60% ± 0,025	52,10% ± 0,023	54,30% ± 0,023	56,20% ± 0,021	60,00% ± 0,019
		6073	79758	ohne	0	10,10% ± 0,003	17,10% ± 0,009	26,00% ± 0,011	32,60% ± 0,019	39,60% ± 0,014	41,90% ± 0,023	43,90% ± 0,015	49,00% ± 0,015
		2129	72721	ohne	5	10,00% ± 0,003	17,20% ± 0,009	26,70% ± 0,014	32,10% ± 0,013	39,80% ± 0,019	42,10% ± 0,015	45,00% ± 0,013	49,60% ± 0,018
		544	54763	ohne	30	9,90% ± 0,005	17,30% ± 0,01	27,20% ± 0,015	32,20% ± 0,011	40,10% ± 0,018	43,60% ± 0,015	45,20% ± 0,014	50,30% ± 0,011
max100	82	10177	241005	tfidf	0	16,00% ± 0,006	26,50% ± 0,013	37,20% ± 0,018	42,20% ± 0,022	48,20% ± 0,018	50,20% ± 0,015	51,10% ± 0,018	53,60% ± 0,015
		3847	229733	tfidf	5	15,90% ± 0,008	26,40% ± 0,009	37,50% ± 0,016	44,40% ± 0,021	50,60% ± 0,013	51,50% ± 0,015	53,00% ± 0,015	55,70% ± 0,015
		1239	199606	tfidf	30	16,20% ± 0,005	26,30% ± 0,017	39,00% ± 0,019	45,20% ± 0,017	51,00% ± 0,015	53,50% ± 0,016	54,80% ± 0,016	57,90% ± 0,008
		10177	241005	ohne	0	14,70% ± 0,004	23,20% ± 0,012	32,10% ± 0,015	36,30% ± 0,019	41,80% ± 0,011	43,60% ± 0,012	45,00% ± 0,01	48,80% ± 0,013
		3847	229733	ohne	5	14,30% ± 0,005	22,40% ± 0,012	31,90% ± 0,013	36,50% ± 0,02	42,80% ± 0,014	43,70% ± 0,016	45,60% ± 0,015	49,00% ± 0,009
		1239	199606	ohne	30	14,60% ± 0,004	23,00% ± 0,01	31,60% ± 0,016	36,20% ± 0,015	42,60% ± 0,018	44,70% ± 0,015	45,30% ± 0,013	49,70% ± 0,01
min15-max100	46	9924	229014	tfidf	0	17,10% ± 0,007	27,30% ± 0,02	40,10% ± 0,025	45,20% ± 0,028	51,40% ± 0,02	52,60% ± 0,023	54,50% ± 0,022	56,10% ± 0,016
		3745	218009	tfidf	5	17,30% ± 0,006	28,40% ± 0,017	39,90% ± 0,022	46,30% ± 0,026	53,40% ± 0,021	54,70% ± 0,018	55,60% ± 0,021	58,30% ± 0,017
		1205	188868	tfidf	30	17,10% ± 0,007	28,70% ± 0,013	41,50% ± 0,018	48,60% ± 0,02	54,00% ± 0,017	57,00% ± 0,02	58,40% ± 0,013	60,80% ± 0,012
		9924	229014	ohne	0	15,30% ± 0,004	24,50% ± 0,012	34,30% ± 0,015	38,50% ± 0,016	44,60% ± 0,013	46,20% ± 0,015	48,10% ± 0,017	51,50% ± 0,016
		3745	218009	ohne	5	15,40% ± 0,005	24,30% ± 0,012	34,00% ± 0,012	38,70% ± 0,019	44,40% ± 0,019	46,10% ± 0,015	47,80% ± 0,011	51,60% ± 0,013
		1205	188868	ohne	30	15,40% ± 0,005	24,60% ± 0,01	34,30% ± 0,014	39,50% ± 0,02	44,80% ± 0,014	47,00% ± 0,012	48,20% ± 0,015	52,30% ± 0,012
	82	20574	863167	tfidf	0	54,20% ± 0,027	60,40% ± 0,019	69,60% ± 0,015	71,90% ± 0,014	74,00% ± 0,009	74,80% ± 0,01	75,10% ± 0,01	76,50% ± 0,006
		7591	840422	tfidf	5	53,90% ± 0,027	60,90% ± 0,018	69,00% ± 0,015	72,10% ± 0,012	74,00% ± 0,008	74,70% ± 0,008	75,30% ± 0,008	76,70% ± 0,005
		2657	784434	tfidf	30	54,50% ± 0,027	60,40% ± 0,021	69,80% ± 0,013	72,20% ± 0,011	74,30% ± 0,008	75,10% ± 0,007	75,40% ± 0,008	77,00% ± 0,006
		20574	863167	ohne	0	49,30% ± 0,011	55,50% ± 0,008	61,60% ± 0,013	64,60% ± 0,008	67,70% ± 0,006	68,80% ± 0,005	69,50% ± 0,006	71,00% ± 0,005
		7591	840422	ohne	5	48,90% ± 0,013	55,80% ± 0,003	61,60% ± 0,014	64,80% ± 0,006	67,70% ± 0,007	68,50% ± 0,005	69,60% ± 0,004	71,10% ± 0,007
		2657	784434	ohne	30	49,10% ± 0,012	55,30% ± 0,008	62,10% ± 0,012	65,10% ± 0,008	68,00% ± 0,007	68,80% ± 0,006	69,60% ± 0,005	71,20% ± 0,006
min15	46	20432	851176	tfidf	0	54,40% ± 0,026	60,50% ± 0,019	69,50% ± 0,012	72,20% ± 0,015	74,80% ± 0,007	75,80% ± 0,008	76,10% ± 0,009	77,10% ± 0,007
		7536	828574	tfidf	5	55,10% ± 0,024	61,30% ± 0,022	70,20% ± 0,012	72,50% ± 0,012	75,20% ± 0,007	75,80% ± 0,007	76,40% ± 0,006	77,20% ± 0,008
		2629	772865	tfidf	30	54,90% ± 0,031	60,80% ± 0,02	70,10% ± 0,013	73,10% ± 0,013	75,30% ± 0,007	76,00% ± 0,009	76,40% ± 0,007	77,70% ± 0,006
		20432	851176	ohne	0	49,30% ± 0,012	56,30% ± 0,005	62,10% ± 0,012	65,20% ± 0,01	68,60% ± 0,008	69,50% ± 0,006	70,50% ± 0,005	72,10% ± 0,005
		7536	828574	ohne	5	49,40% ± 0,012	56,10% ± 0,007	62,30% ± 0,014	65,50% ± 0,008	68,50% ± 0,007	69,50% ± 0,007	70,50% ± 0,006	71,70% ± 0,007
		2629	772865	ohne	30	48,80% ± 0,013	56,20% ± 0,007	62,90% ± 0,011	65,30% ± 0,008	68,70% ± 0,006	69,70% ± 0,006	70,60% ± 0,006	72,10% ± 0,006

G Telekom-Fragebogen und Ontologie

Kapitel 10.1.5 beschäftigt sich mit der Anwendung des Subjektiven Clusterns auf die Kommunikationsdaten der Deutschen Telekom AG. Dazu ist es notwendig, eine geeignete Ontologie zu akquirieren. Den verwendeten Fragebogen findet man auf den folgenden Seiten. Er wurde in Zusammenarbeit mit Michael Nuhn erarbeitet (siehe [179]). Die daraus resultierende Ontologie ist ausschnittsweise in Abbildung G.1 zu sehen.

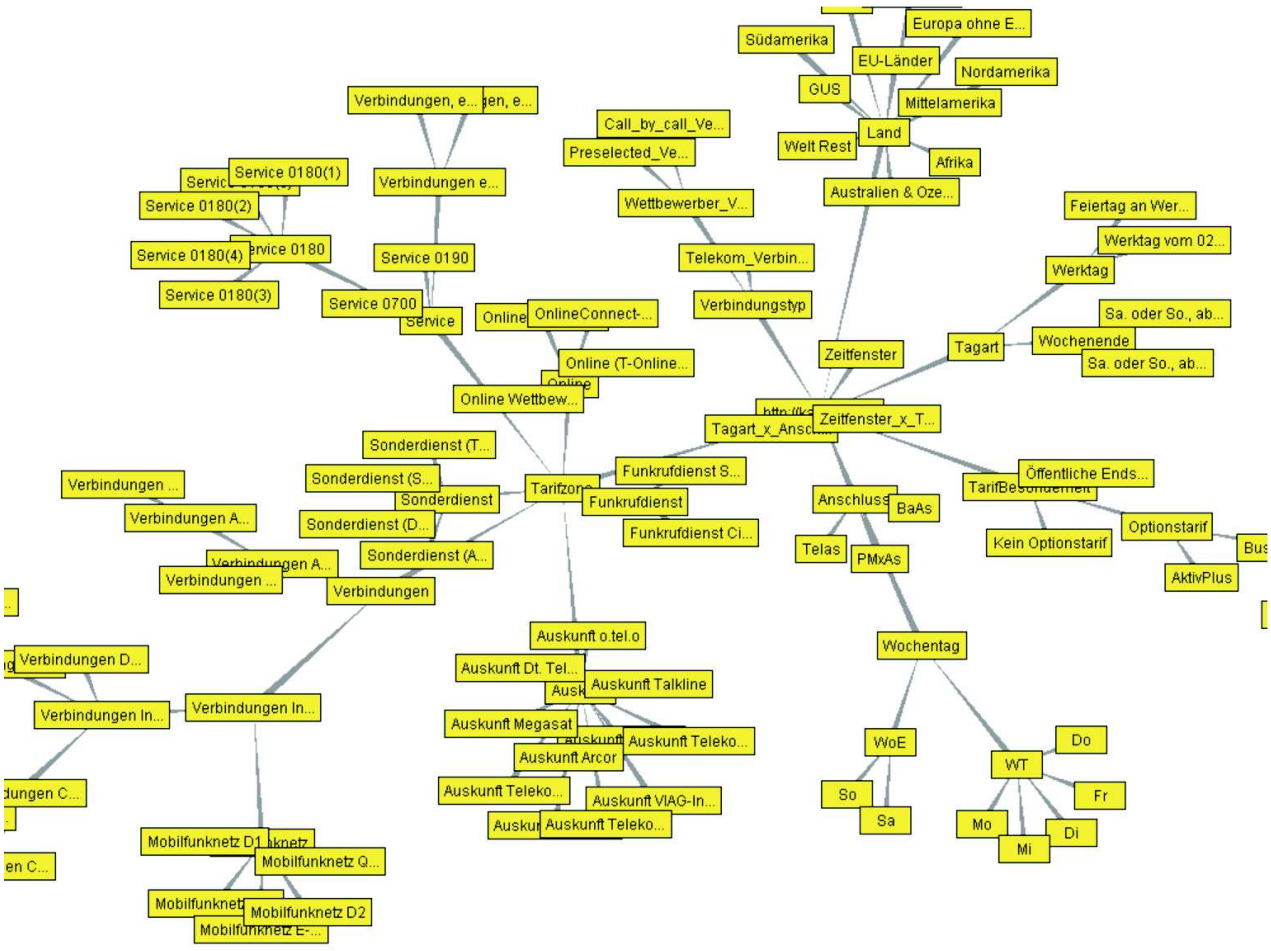

Abbildung G.1: Ausschnitt aus der mittels Fragebogen akquirierten Telekom-Ontologie

Fragebogen zur Kundensegmentierung

Im Rahmen meiner Diplomarbeit wende ich einen Clusteralgorithmus auf Gesprächsdaten bei der Telekom an, um algorithmisch Kundengruppen zu erstellen.

Die Daten, die ich durch diese Umfrage ermittele, werden in Form einer Ontologie dem Clusteralgorithmus als Hintergrundwissen zur Verfügung gestellt. Dieser wird dann aus den in dieser Umfrage ermittelten Blickwinkeln selbständig aus Verbindungsdaten repräsentative Kategorien erstellen, in die sich die Kunden der Telekom einteilen lassen. Auf diese Weise soll der Telekom eine neue Methode zur Verfügung gestellt werden, Kundenkategorien durch eine automatische Analyse ihrer Kommunikationsdaten zu gewinnen.

Den Fragebogen können Sie auch im Internet unter http://www.michael-nuhn.de/Fragebogen.doc herunterladen und ausgefüllt per Email an folgende Adresse schicken: an@michael-nuhn.de

1. Für welche Kunden entwerfen Sie Tarife?

a. Privatkunden (MVC)
b. Geschäftskunden (MVB)
c. Sonstige:

---

2. Auf welche Datenquellen stützen Sie sich, wenn Sie Tarife entwerfen?

a. Statistiken über Kommunikationsdaten (wenn ja, welche?)

---

b. Umfragen
c. Andere:

---

3. Mit welchen Tarifen beschäftigen Sie sich oder haben Sie sich früher beschäftigt und sind diese noch am Markt?

Bitte kreuzen Sie entsprechend an oder ergänzen Sie die Liste!

	Aktuell beschäftigt	Früher beschäftigt	Noch am Markt?	
			Ja	Nein
AktivMobil				
AktivMobil				
AktivPlus				
AktivPlus Basis				
Bonus 8				
BusinessCall 300				
BusinessCall 500				
BusinessCall 700				
Dial & Benefit CN				
Select 5/30				
T DSL				
T ISDN				
T Net				
T Online				
XXL				
XXL				
Sonstige:				

4. In welche Gruppen teilen Sie die Kunden der Telekom ein?

Ich möchte Sie bitten, hier anzugeben, wie Sie persönlich aufgrund Ihrer Erfahrung die Kunden einteilen würden. Falls die übliche Terminologie des Marketings nicht ausreicht, um ein bestimmtes Kundensegment zu beschreiben, so bitte ich Sie, dieses Segment hier kurz mit Ihren eigenen Worten zu beschreiben.

Lfdnr	Abwanderungs-Gefährdet?	Beschreibung
0.		z.B.: Geschäftskunden mit hohem Gesprächsaufkommen (über 4.000 min/Monat)

5.a Welche der folgenden Merkmale würden Sie zum Charakterisieren Ihrer Kunden nutzen? Welche Merkmale eignen sich Ihrer Meinung nach besonders gut, um Kundengruppen zu beschreiben oder sie voneinander zu unterscheiden?

1 – unwichtig, 5 – wichtig
Bitte ergänzen Sie die Tabelle ggf. um Merkmale, die Ihnen wichtig sind!

	1	2	3	4	5
Summe des Umsatzvolumens					
Summe der Verbindungsminuten					
Summe der Minuten zur Hauptzeit					
Summe der Minuten zur Nebenzeit					
Summe der Minuten über AktivPlus					
Summe der Minuten über 1. Telekom, 2. Call by call und 3. Preselection					
Marktanteil, den die Telekom am Umsatz des Kunden hat					
Marktanteil, den die Telekom an den City-Verbindungen hat					
Marktanteil, den die Telekom an den Regio-und Fernverbindungen hat					

Marktanteil, den die Telekom an den Welt-Verbindungen hat					
Marktanteil, den die Telekom an den Draht-Funk-Verbindungen hat					
Anteil der Draht-Funk-Verbindungen ins Ausland am gesamten Minutenvolumen					
Anteil der Drahtverbindungen ins Ausland am gesamten Minutenvolumen					
Anteil der Onlinezeit am Minutenvolumen					
Anteil der Verbindungen über einen Optionstarif an den gesamten Verbindungen					

5.b Würden Sie einige der oben genannten Mekmale gerne abändern? Sie können die Merkmale variieren:

- Zum Beispiel die Dimension ändern:
  Aus:
  „Summe der Minuten über AktivPlus“ können Sie
  „Anzahl der Gespräche über AktivPlus“ machen.

- oder verfeinern:
  Aus:
  „Summe der Minuten über AktivPlus“ können Sie
  „Summe der Minuten über AktivPlus zur Hauptzeit“ oder
  „Summe der Minuten über AktivPlus zur Hauptzeit in ein Mobilfunknetz“ machen.
  oder aus
  „Summe der Onlineminuten“
  „Summe der Onlineminuten Nachts“ machen.

  Sie brauchen sich nicht an die üblichen Telekomkategorien zu halten. Wenn
  „Summe der Minuten zur Hauptzeit“
  nicht das ist, was Sie haben wollen, da Sie vielleicht die Gespräche interessieren, die zu Bürozeiten geführt werden, können Sie auch
  „Summe der Minuten von 8:00 – 16:00“
  wählen. Meinetwegen auch:
  „Summe der Minuten von 8:00 – 16:00 aber nicht von 12:00 –12:30 und 9:00-9:15“

- Oder Marktanteile und sonstige Anteile hinzufügen:
  Aus:
  “Summe der Verbindungsminuten“ können Sie
  “Marktanteil der Telekom an der Summe der Verbindungsminuten“ machen
  oder aus
  „Summe der Onlineminuten“
  „Anteil der Onlineminuten an der Summe des Gesamtminutenvolumens“ machen.

- Sowie alle diese Möglichkeiten kombinieren:
  “Anteil der Telekom am Minutenvolumen Abends an Verbindungen in Mobilfunknetze nach Ghana“

Tragen Sie Ihre Ideen hier ein und kreuzen Sie an, welche Priorität die Merkmale beim Unterscheiden von Kundengruppen für Sie haben!

	1	2	3	4	5

5.c Sind Ihnen ganz andere Merkmale wichtig? Hier würde ich gerne erfahren, ob Sie etwas aus der obigen Liste vollkommen vermissen. Sie können hier ganz frei Merkmale notieren, die für Sie wichtig sind, aber oben nicht angeboten wurden.

Die Ergebnisse einer Befragung ergeben nicht nur die Domänenontologie, sondern auch eine für den erweiterten COSA-Algorithmus notwendige Arbeitsontologie. Ein Auschnitt einer Arbeitsontologie ist in Abbildung G.2 zu sehen.

Abbildung G.2: Ausschnitt aus der mittels Fragebogen akquirierten Telekom-Arbeitsontologie

Literaturverzeichnis

[1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison Wesley, 1995.

[2] Steven P. Abney. Parsing by chunks. In Robert C. Berwick, Steven P. Abney, and Carol Tenny, editors, *Principle-Based Parsing: Computation and Psycholinguistics*, pages 257–278. Kluwer Academic Publishers, Boston, 1991.

[3] Sudhir Agarwal, Jorge Gonzalez, Jens Hartmann, Solivia Hollfelder, Anthony Jameson, Stefan Klink, Michael Ley, Emma Rabbidge, Eric Schwarzkopf, Nitesh Shrestha, Nenad Stojanovic, Rudi Studer, Gerd Stumme, Bernd Walter, Alexander Weber, Patrick Lehti, and Peter Fankhauser. Semantic methods and tools for information portals. In *Informatik03 - Jahrestagung der Gesellschaft für Informatik*, pages 116–131. GI, SEP 2003.

[4] Sudhir Agarwal, Siegfried Handschuh, and Steffen Staab. Surfing the service web. In *ISWC2003 2nd International Semantic Web Conference*, volume 2870, pages 211–226, Sanibal Island, Florida, USA, 2003. Springer.

[5] Charu C. Aggarwal and Philip S. Yu. Finding generalized projected clusters in high dimensional spaces. In Weidong Chen, Jeffrey F. Naughton, and Philip A. Bernstein, editors, *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA*, pages 70–81. ACM, 2000.

[6] E. Agirre and G. Rigau. Word sense disambiguation using conceptual density. In *Proc. of 15th International Conference on Computational Linguistics, COLING'96. Copenhagen, Denmark, 1996*, 1996.

[7] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the ACM SIGMOD Int'l Conference on Management of Data, Seattle, Washington*, pages 94–105. ACM Press, June 1998.

[8] B. Amann and I. Fundulaki. Integrating ontologies and thesauri to build RDF schemas. In S. Abiteboul and A.-M. Vercoustre, editors, *Proceedings of the Third European Conference on Digital Libraries (ECDL-99): Research and Advanced Technology for Digital Libraries*, volume 1696 of *Lecture Notes in Computer Science (LNCS)*, pages 234–253, Paris, France, September 1999. Springer.

[9] G. Amati, C. Carpineto, and G. Romano. Fub at trec-10 web track: A probabilistic framework for topic relevance term weighting. In *Proceedings of the Tenth Text Retrieval Conference (TREC-10)*, pages 182–191. National Institute of Standards and Technology (NIST), 2001.

[10] S. Amit, G. Salton, M. Mitra, and C. Buckley. Document length normalization. Technical report, Technical Report TR95-1529, Cornell University, Computer Science, 1995.

[11] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. In *Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'99)*, pages 49–60, Philadelphia, PA, 1999.

[12] Anupriya Ankolekar, Mark Burstein, Jerry R. Hobbs, Ora Lassila, Drew McDermott, David Martin, Sheila A. McIlraith, Srini Narayanan, Massimo Paolucci, Terry Payne, and Katia Sycara. Daml-s: Web service description for the semantic web. In *1st Int'l Semantic Web Conf. (ISWC 02)*, pages 348–363, 2002.

[13] A. P. Azcarraga and Teddy N. Yap Jr. Extracting meaningful labels for websom text archives. In *Proc of the 10th ACM International Conference on Information and Knowledge Management (CIKM 2001)*, pages 41–48, Atlanta, Georgia, USA, 2001.

[14] F. Baader, D. Calvanese, D. McGuiness, D. Nardi, and P. Patel-Schneider, editors. *The Description Logic Handbook*. Cambridge University Press, 2003.

[15] J. Bacher. *Clusteranalyse: Anwendungsorientierte Einführung*. R. Oldenbourg Verlag Wien München GmbH, 1994.

[16] G. Ball and D. Hall. Isodata: A novel method of data analysis and pattern classification. Technical report, Stanford Research Institute, Menlo Park, 1965.

[17] A. Bauer and H. Günzel. *Data Warehouse Systeme.* dpunkt.verlag, Heidelberg, 2001.

[18] Florian Beil, Martin Ester, and Xiaowei Xu. Frequent term-based text clustering. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 436–442. ACM Press, 2002.

[19] B. Berendt, A. Hotho, and G. Stumme. Towards semantic web mining. In I. Horrocks and J. A. Hendler, editors, *Proceedings of the First International Semantic Web Conference: The Semantic Web (ISWC 2002)*, volume 2342 of *Lecture Notes in Computer Science (LNCS)*, pages 264–278, Sardinia, Italy, 2002. Springer.

[20] Pavel Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.

[21] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 2001(5), 2001. available at http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.

[22] Tim Berners-Lee. Semantic web road map. http://www.w3.org/DesignIssues/Semantic.html, 1998.

[23] Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien. Using linear algebra for intelligent information retrieval. Technical Report UT-CS-94-270, Computer Science Department, University of Tennessee, Knoxville, 1994.

[24] Michael Berthold and David J. Hand (eds.). *Intelligent data analysis*. Springer-Verlag New York, Inc., 1999.

[25] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is 'nearest neighbor' meaningful. In *Proc. of ICDT-1999*, pages 217–235, 1999.

[26] Avrim Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997.

[27] E. Bozsak, M. Ehrig, S. Handschuh, A. Hotho, A. Maedche, B. Motik, D. Oberle, C. Schmitz, S. Staab, L. Stojanovic, N. Stojanovic, R. Studer, G. Stumme, Y. Sure, J. Tane, R. Volz, and V. Zacharias. Kaon - towards a large scale semantic web. In *Proceedings of EC-Web*, pages 304–313, Aix-en-Provence, France, 2002. LNCS 2455 Springer.

[28] R. Brachman and T. Anand. The process of knowledge discovery in databases: A humancentered approach. In *Advances in Knowledge Discovery & Data Mining*, pages 37–57. AAAI Press & The MIT Press, 1996.

[29] P. Bradley, U. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In *Proc. of KDD-1998*, pages 9–15. AAAI Press, August 1998.

[30] Paul S. Bradley and Usama M. Fayyad. Refining initial points for K-Means clustering. In *Proc. 15th International Conf. on Machine Learning*, pages 91–99. Morgan Kaufmann, San Francisco, CA, 1998.

[31] C. Brewster, F. Ciravegna, and Y. Wilks. Background and foreground knowledge in dynamic ontology construction. In *Proceedings of the SIGIR Semantic Web Workshop*, 2003.

[32] Wray Buntine and Henry Tirri. Multi-faceted learning of web taxonomies. In G. Stumme B. Berendt, A. Hotho, editor, *Proc. of the Semantic Web Mining Workshop of the 13th European Conference on Machine Learning (ECML'02)/ 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02)*, 2002.

[33] J. Carletta. Asessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[34] Claudio Carpineto, Renato de Mori, Giovanni Romano, and Brigitte Bigi. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems (TOIS)*, 19(1):1–27, 2001.

[35] Surajit Chaudhuri and Umeshwar Dayal. An overview of data warehousing and olap technology. *SIGMOD Record*, 26(1):65–74, 1997.

[36] Ming-Syan Chen, Jiawei Han, and Philip S. Yu. Data mining: an overview from a database perspective. *IEEE Transaction on Knowledge and Data Engineering*, 8(6):866–883, 1996.

[37] Yizong Cheng and George M. Church. Biclustering of expression data. In *Proc. of the 8th ISMB*, pages 93–103. AAAI Press, 2000.

[38] Yi-Ming Chung, William M. Pottenger, and Bruce R. Schatz. Automatic subject indexing using an associative neural network. In *Proceedings of the 3 rd ACM International Conference on Digital Libraries (DL'98)*, pages 59–68, 1998.

[39] William W. Cohen. Fast effective rule induction. In Armand Prieditis and Stuart Russell, editors, *Proc. of the 12th International Conference on Machine Learning*, pages 115–123, Tahoe City, CA, July 9–12, 1995. Morgan Kaufmann.

[40] Cross industry standard process for data mining. http://www.crisp-dm.org/.

[41] K. Dahlgren. A linguistic ontology. *International Journal of Human–Computer Studies*, 43(5/6):809–818, 1995.

[42] Stephen D'Alessio, Keitha Murray, Robert Schiaffino, and Aaron Kershenbaum. A. category levels in hierarchical text categorization. In *Proceedings of EMNLP-3, 3rd Conference on Empirical Methods in Natural Language Processing*, 1998.

[43] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the Twelfth International World Wide Web Conference, WWW2003*. ACM, 2003.

[44] T. H. Davenport and L. Prusak. *Working Knowledge – How organisations manage what they know*. Havard Business School Press, Boston, Massachusetts, 1998.

[45] M. de Buenaga Rodrıguez, J. M. Gomez Hidalgo, and B. Díaz-Agudo. Using WordNet to complement training information in text categorization. In *Recent Advances in Natural Language Processing II*, volume 189. John Benjamins, 2000.

[46] S. Decker, M. Daniel, M. Erdmann, and R. Studer. An enterprise reference scheme for integrating model based knowledge engineering and enterprise modeling. In E. Plaza and V. R. Benjamins, editors, *Proceedings of the 10th European Workshop on Knowledge Acquisistion, Modeling, and Management (EKAW'97)*, volume 1319 of *Lecture Notes in Artificial Intelligence (LNAI)*. Springer, 1997.

[47] S. Decker, M. Erdmann, D. Fensel, and R. Studer. Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information. In R. Meersman et al., editors, *Database Semantics: Semantic Issues in Multimedia Systems*, pages 351–369. Kluwer Academic Publisher, 1999.

[48] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[49] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[50] M. Devaney and A. Ram. Efficient feature selection in conceptual clustering. In *Proceedings 14th International Conference on Machine Learning, Nashville, TN*, pages 92–97. Morgan Kaufmann, 1998.

[51] Inderjit S. Dhillon, Yuqiang Guan, and J. Kogan. Iterative clustering of high dimensional text data augmented by local search. In *2nd SIAM International Conference on Data Mining (Workshop on Clustering High-Dimensional Data and its Applications)*, 2002.

[52] Inderjit S. Dhillon, Subramanyam Mallela, and Dharmendra S. Modha. Information-theoretic co-clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 89–98. ACM Press, 2003.

[53] Inderjit S. Dhillon and Dharmendra S. Modha. A data-clustering algorithm on distributed memory multiprocessors. In *Large-Scale Parallel Data Mining, Workshop on Large-Scale Parallel KDD Systems, SIGKDD, August 15, 1999, San Diego, CA, USA, revised papers*, volume 1759 of *Lecture Notes in Computer Science*, pages 245–260. Springer, 2000.

[54] DIN 2330. *Begriffe und Benennungen – Allgemeine Grundsätze*. DIN-Deutsches Institut für Normung e.V. (Normenausschuß Terminologie), 1993.

[55] DIN 2331. *Begriffssysteme und ihre Darstellung*. DIN-Deutsches Institut für Normung e.V. (Normenausschuß Terminologie), April 1980.

[56] James Dougherty, Ron Kohavi, and Mehran Sahami. Supervised and unsupervised discretization of continuous features. In *International Conference on Machine Learning*, pages 194–202, 1995.

[57] Jason Dowling. Information retrieval using latent semantic indexing (lsi) and a semi-discrete matrix decomposition (sdd). Bcomp(hons) thesis, Monash University, 2002.

[58] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2001.

[59] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. Wiley-Interscience Publication, New York, 1973.

[60] A. J. Duineveld, R. Stoter, M. R. Weiden, B. Kenepa, and V. R. Benjamins. Wondertools? a comparative study of ontological engineering tools. *International Journal of Human-Computer Studies*, 6(52):1111–1133, 2000.

[61] Andreas Eberhart. OntoAgent: A platform for the declarative specification of agents. In M. Schroeder and G. Wagner, editors, *Proceedings of the international Workshop on Rule Markup Languages for Business Rules on the Semantic Web. In conjunction with the first International Semantic Web Conference (ISWC 2002)*, pages 58–71, Chia, Sardinia, Italy, July 2002.

[62] R. Engels. *Component-Based User Guidance in Knowledge Discovery and Data Mining*. PhD thesis, Universität Karlsruhe, 1999.

[63] Michael Erdmann. *Ontologien zur konzeptuellen Modellierung der Semantik von XML*. Isbn: 3831126356, Universität Karlsruhe, 10 2001.

[64] M. Ester and J. Sander. *Knowledge Discovery in Databases: Techniken und Anwendungen*. Springer Verlag, Berlin, September 2000.

[65] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Evangelos Simoudis, Jiawei Han, and Usama M. Fayyad, editors, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. AAAI Press, 1996.

[66] John O. Everett, Daniel G. Bobrow, Reinhard Stolle, Richard Crouch, Valeria de Paiva, Cleo Condoravdi, Martin van den Berg, and Livia Polanyi. Making ontologies work for resolving redundancies across documents. *Communications of the ACM*, 45(2):55–60, 2002.

[67] U. Fayyad, C. Reina, and P. Bradley. Initialization of iterative refinement clustering algorithms. In *Proc. of KDD-1998*, pages 194–198. AAAI Press, August 1998.

[68] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. Knowledge discovery and data mining: Towards a unifying framework. In *Knowledge Discovery and Data Mining*, pages 82–88, 1996.

[69] Ronen Feldman and Ido Dagan. Kdt - knowledge discovery in texts. In *Proceedings of the First International Conference on Knowledge Discovery (KDD)*, pages 112–117, 1995.

[70] D. Fensel. *Problem Solving Methods: Understanding, Description, Development, and Reuse*, volume 1791 of *Lecture Notes in Computer Science (LNCS)*. Springer, 2000.

[71] D. Fensel. *Ontologies: Silver bullet for knowledge management and electronic commerce*. Springer-Verlag, Berlin, 2001.

[72] Reginald Ferber. *Information Retrieval. Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web*. dpunkt.verlag, 2003.

[73] N. Fickel. Clusteranalyse mit gemischt-skalierten merkmalen: Abstrahierung vom skalenniveau. *Allgemeines Statistisches Archiv, Vandenhoeck & Ruprecht in Göttingen*, 81(3):249–265, 1997.

[74] Douglas H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2(2):139–172, September 1987.

[75] E. Forgy. Cluster analysis of multivariate data: Efficiency versus interpretability of classification. *Biometrics*, 21(3):768–769, 1965.

[76] Eibe Frank and Ian H. Witten. Generating accurate rule sets without global optimization. In *Proc. 15th International Conf. on Machine Learning*, pages 144–151. Morgan Kaufmann, San Francisco, CA, 1998.

[77] J. Fuernkranz, T. Mitchell, and E. Riloff. A Case Study in Using Linguistic Phrases for Text Categorization on the WWW. In *Proc. of AAAI/ICML Workshop Learning for Text Categorization*, pages 5–12. AAAI Press, 1998.

[78] Robert Gaizauskas. An information extraction perspective on text mining: Tasks, technologies and prototype applications. http://www.itri.bton.ac.uk/projects/euromap/TextMiningEvent/Rob_Gaizauskas.pdf, 2003.

[79] B. Ganter and R. Wille. *Formale Begriffsanalyse: Mathematische Grundlagen*. Springer Verlag, Berlin, 1996.

[80] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, 1999.

[81] W. Gaul and M. Schader. A new algorithm for two-mode clustering. In H. H. Bock and W. Polasek, editors, *Data Analysis and Information Systems*, pages 15–23, Berlin, 1995. Springer.

[82] A. Gersho and R. M. Gray. *Vector quantization and signal compression*. Kluwer Academic Publishers, 1992.

[83] R. Gnanadesikan. *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York, 1977.

[84] A. Gómez-Pérez, J. Angele, M. Fernandéz-López, V. Christophides, A. Stutt, Y. Sure, et al. A survey on ontology tools. OntoWeb deliverable 1.3, Universidad Politecnia de Madrid, 2002.

[85] A. Gómez-Pérez, David Manzano-Macho, Enrique Alfonseca, Rafael Nú nez, Ian Blacoe, Steffen Staab, Oscar Corcho, Ying Ding, Jan Paralic, and Raphael Troncy. A survey of ontology learning methods and techniques. OntoWeb deliverable 1.5, Universidad Politecnia de Madrid, 2002.

[86] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarrán. Indexing with WordNet synsets can improve text retrieval. In *Proceedings ACL/COLING Workshop on Usage of WordNet for Natural Language Processing*, 1998.

[87] E. Gose, R. Johnsonbaugh, and S. Jost. *Pattern Recognition & Image Analysis*. Prentice-Hall, 1996.

[88] Robert M. Gray, Keren Perlmutter, and Richard A. Olshen. Quantization, classification, and density estimation for kohonen's gaussian mixture. In *Data Compression Conference*, pages 63–72, 1998.

[89] Stephen J. Green. Building hypertext links in newspaper articles using semantic similarity. In *Proc. of third Workshop on Applications of Natural Language to Information Systems (NLDB '97)*, 1997.

[90] Stephen J. Green. Building hypertext links by computing semantic similarity. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 11(5):713–730, 1999.

[91] T. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5:199–220, 1993.

[92] T. Gruber. Towards principles for the design of ontologies used for knowledge sharing. *Intl. J. of Human and Computer Studies*, 43(5/6):907–928, 1994.

[93] N. Guarino. Understanding, building and using ontologies. *Intl. J. of Human and Computer Studies*, 46(2/3):293–310, 1997.

[94] N. Guarino. Formal ontology and information systems. In N. Guarino, editor, *Proceedings of the First International Conference on Formal Ontologies in Information Systems (FOIS)*, volume 46 of *Frontiers in Artificial Intelligence and Applications*, Trento, Italy, 1998. IOS-Press.

[95] N. Guarino, C. Masolo, and G. Vetere. OntoSeek: Content-based access to the web. *IEEE Intelligent Systems*, 14(3), 1999.

[96] Martin Gutschke. Kategorisierung von Textuellen Lernobjekten mit Methoden des Maschinellen Lernens. Studienarbeit, Universität Hannover, Hannover, 2003.

[97] Udo Hahn and Klemens Schnattinger. Towards text knowledge engineering. In *AAAI/IAAI*, pages 524–531, 1998.

[98] Birgit Hamp and Helmut Feldwig. GermaNet — A lexical-semantic net for German. In Piek Vossen, Geert Adriaens, Nicoletta Calzolari, Antonio Sanfilippo, and Yorick Wilks, editors, *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15. Association for Computational Linguistics, New Brunswick, New Jersey, 1997.

[99] S. Handschuh, S. Staab, and A. Maedche. CREAM – creating relational metadata with a component-based, ontology-driven annotation framework. In *Proceedings of the First International Conference on Knowledge Capture (K-Cap 2001)*, Victoria, B.C., Canada, October 2001.

[100] Siegfried Handschuh and Steffen Staab. Authoring and annotation of web pages in CREAM. In *Proc. of the 11th International World Wide Web Conference, WWW*, Honolulu, Hawaii, 2002. ACM Press.

[101] J. Hartigan. *Clustering Algorithms*. John Wiley and Sons, New York, 1975.

[102] Vasileios Hatzivassiloglou, Luis Gravano, and Ankineedu Maganti. An investigation of linguistic features and clustering algorithms for topical document clustering. In *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 24-28, 2000, Athens, Greece*. ACM, 2000.

[103] M. Hearst. Untangling text data mining. In *Proceedings of ACL'99 the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.

[104] N. Henze. Towards open adaptive hypermedia. In *9. ABIS-Workshop 2001, im Rahmen der Workshopwoche "Lernen - Lehren - Wissen - Adaptivität" (LLWA 01)*, Dortmund, 2001.

[105] José María Gómez Hidalgo. Tutorial on text mining and internet content filtering. Tutorial Notes Online: http://ecmlpkdd.cs.helsinki.fi/pdf/hidalgo.pdf, 2002.

[106] A. Hinneburg, C. Aggarwal, and D. A. Keim. What is the nearest neighbor in high dimensional spaces? In *Proc. of VLDB-2000*, pages 506–515. Morgan Kaufmann, September 2000.

[107] A. Hinneburg and D. A. Keim. Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. In *Proceedings of the 25th International Conference on Very Large Databases (VLDB-99), Edinburgh, Scotland*, pages 506–517. Morgan Kaufmann, 1999.

[108] A. Hinneburg, M. Wawryniuk, and D. A. Keim. Hd-eye: visual mining of high-dimensional data. *IEEE Computer Graphics and Applications*, 19(5):22–31, September 1999.

[109] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pages 50–57, 1999.

[110] I. Horrocks and P. F. Patel-Schneider. DL systems comparison. In E. Franconi, G. De Giacomo, R. M. MacGregor, W. Nutt, C. A. Welty, and F. Sebastiani, editors, *Collected Papers from the International Description Logics Workshop (DL'98)*, pages 55–57. CEUR (http://ceur-ws.org/), 1998.

[111] A. Hotho. Analyse von Wettbewerbsverlusten im Telekommunikationsmarkt und mögliche Gegenmaßnahmen. Projektbericht 1999 für die Deutsche Telekom AG, AIFB, 2000.

[112] A. Hotho. Analyse von Wettbewerbsverlusten im Telekommunikationsmarkt und mögliche Gegenmaßnahmen. Projektbericht 2000 für die Deutsche Telekom AG, AIFB, 2001.

[113] A. Hotho, A. Maedche, and S. Staab. Ontology-based text clustering. In *Proceedings of the IJCAI-2001 Workshop "Text Learning: Beyond Supervision", August, Seattle, USA*, 2001.

[114] A. Hotho, A. Maedche, S. Staab, and R. Studer. SEAL-II — the soft spot between richly structured and unstructured knowledge. *Journal of Universal Computer Science (J.UCS)*, 7(7):566–590, 2001.

[115] A. Hotho, S. Staab, and G. Stumme. Explaining text clustering results using semantic structures. In *Proc. of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD*, pages 217–228, 2003.

[116] A. Hotho, S. Staab, and G. Stumme. Ontologies improve text document clustering. In *Proc. of the ICDM 03, The 2003 IEEE International Conference on Data Mining*, pages 541–544, 2003.

[117] A. Hotho, S. Staab, and G. Stumme. Text clustering based on background knowledge. Technical report, Institute AIFB, Universität Karlsruhe, 2003. 36 pages.

[118] A. Hotho, S. Staab, and G. Stumme. Wordnet improves text document clustering. In *Proc. of the SIGIR 2003 Semantic Web Workshop*, Toronto, Canada, 2003.

[119] A. Hotho and G. Stumme. Conceptual clustering of text clusters. In *Proceedings of FGML Workshop*, pages 37–45. Special Interest Group of German Informatics Society (FGML — Fachgruppe Maschinelles Lernen der GI e.V.), 2002. http://www.aifb.uni-karlsruhe.de/WBS/aho/pub/tc_fca_2002_submit.pdf.

[120] F. Höppner, F. Klawon, R. Kruse, and T. Runkler. *Fuzzy Cluster Analysis: Methods for classification, data analysis and image recognition*. John Wiley and Sons Ltd, 1999.

[121] N. Ide and J. Véronis. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40, 1998.

[122] ISO 704. *Terminology Work — Principles and Methods*. International Organization of Standardization, 2000.

[123] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys,*, 31(3):264–323, 1999.

[124] R. Jäger. Zusammenhang zwischen Gebühren und Einnahmen im Telekommunikationsbereich. *Der Fernmelde-Ingenieur*, 1990.

[125] Kyo Kageura and Bin Umino. Methods of automatic term recognition: a review. *Terminology*, 3(2):259–289, 1996.

[126] G. Karypis and E. Han. Fast supervised dimensionality reduction algorithm with applications to document categorization and retrieval. In *Proc. of 9th ACM International Conference on Information and Knowledge Management, CIKM-00*, pages 12–19, New York, US, 2000. ACM Press.

[127] George Karypis and Eui-Hong Han. Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval and categorization. Technical report tr-00-0016, University of Minnesota, 2000.

[128] V. Kashyap. Design and creation of ontologies for environmental information retrieval. In *Proceedings of the 12th Workshop on Knowledge Acquisition, Modeling and Management (KAW-99)*, Banff, Canada, 1999. available at http://sern.ucalgary.ca/KSI/KAW/KAW99/papers/Kashyap1/kashyap.pdf.

[129] Leonard Kaufman and Peter J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Wiley, New York, 1990.

[130] M. Kesseler. A schema based approach to HTML authoring. *World Wide Web Journal*, 96(1), 1996.

[131] M. Kifer, G. Lausen, and J. Wu. Logical foundations of object-oriented and frame-based languages. *Journal of the ACM*, 42:741–843, 1995.

[132] M. Kim and P. Compton. Formal concept analysis for domain-specific document retrieval systems. In Markus Stumptner, Dan Corbett, and Mike Brooks, editors, *AI 2001: Advances in Artificial Intelligence: 14th Australian Joint Conference on Artificial Intelligence*, pages 237–248, Adelaide Australia, 2001. Springer.

[133] M. Kim and P. Compton. Evolutionary document management and retrieval for specialised domains. *International Journal of Human Computer Studies (IJHCI)*, page to appear, 2004.

[134] Mathias Kirsten and Stefan Wrobel. Relational distance-based clustering. In D. Page, editor, *Proc. Eighth Int. Conference on Inductive Logic Programming*, pages 261–270. Springer, LNAI 1446, 1998.

[135] Mathias Kirsten and Stefan Wrobel. Extending k-means clustering to first-order representations. In James Cussens and Alan M. Frisch, editors, *Inductive Logic Programming, 10th International Conference, ILP 2000, London, UK, July 24-27, 2000, Proceedings*, volume 1866 of *Lecture Notes in Computer Science*, pages 112–129. Springer, 2000.

[136] Dan Klein, Sepandar D. Kamvar, and Christopher D. Manning. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *Procedings of the Nineteenth International Conference on Machine Learning, ICML*, 2002.

[137] M. Klettke, M. Bietz, I. Bruder, A. Heuer, D. Priebe, G. Neumann, M. Becker, J. Bedersdorfer, H. Uszkoreit, A. Maedche, S. Staab, and R. Studer. GETESS — Ontologien, objektrelationale Datenbanken und Textanalyse als Bausteine einer Semantischen Suchmaschine. *Datenbank-Spektrum*, 1(1), 2001.

[138] Y. Kodratoff. Knowledge discovery in texts: A definition and applications. *Lecture Notes in Computer Science*, 1609:16–29, 1999.

[139] T. Kohonen. *Self-organizing Maps*. Springer Verlag, 1997.

[140] Tamara G. Kolda. *Limited-Memory Matrix Methods with Applications*. PhD thesis, University of Maryland Applied Mathematics, 1997.

[141] V. Kumar and M. Joshi. What is data mining? http://www-users.cs.umn.edu/~mjoshi/hpdmtut/sld004.htm, 2003.

[142] Y. Labrou and T. W. Finin. Yahoo! as an ontology: Using Yahoo! categories to describe documents. In *Proceedings of the 1999 ACM CIKM International Conference on Information and Knowledge Management*, pages 180–187, Kansas City, Missouri, November 1999. ACM Press.

[143] K. Lagus. *Text Mining with the WEBSOM*. PhD thesis, Acta Polytechnica Scandinavica, Mathematics and Computing Series no. 110, Helsinki University of Technology, Finland., 2000.

[144] B. Larsen and Ch. Aone. Fast and effective text mining using linear-time document clustering. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, KDD 1999*, pages 16–22. ACM, 1999.

[145] B. Lauser and A. Hotho. Automatic multi-label subject indexing in a multilingual environment. In *Proc. of the 7th European Conference in Research and Advanced Technology for Digital Libraries, ECDL*, pages 140–151, 2003.

[146] Boris Lauser. Semi-automatic ontology engineering and ontology supported document indexing in a multilingual environment. Diplomarbeit, Universität Karlsruhe, 2003.

[147] Edda Leopold. Das zipfsche gesetz. *Schwerpunkt Textmining; Künstliche Intelligenz*, 16(2):34, 2002.

[148] Kristina Lerman. Document clustering in reduced dimension vector space. http://www.isi.edu/ lerman/papers/Lerman99.pdf, 1999.

[149] D. D. Lewis. Reuters-21578 text categorization test collection, 1997.

[150] L. A. Ureña Lóez, M. de Buenaga Rodríguez, and J. M. Gómez Hidalgo. Integrating linguistic resources in tc through wsd. *Computers and the Humanities*, 35(2):215–230, 2001.

[151] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[152] S. A. Macskassy, A. Banerjee, B. D. Davison, and H. Hirsh. Human performance on clustering web pages: a preliminary study. In *Proc. of KDD-1998*, pages 264–268. AAAI Press, August 1998.

[153] A. Maedche. *Ontology Learning for the Semantic Web*. Kluwer, 2002.

[154] A. Maedche, M. Ehrig, S. Handschuh, L. Stojanovic, and R. Volz. Ontology-focused crawling of documents and relational metadata. In *Proceedings of the Eleventh International World Wide Web Conference WWW-2002*, Hawaii, 2002.

[155] A. Maedche and S. Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2):72 –79, 2001.

[156] A. Maedche, S. Staab, N. Stojanovic, and R. Studer. SEAL - A Framework for Developing SEmantic portALs. In *Proceedings of the 18th British National Conference on Databases, July, Oxford, UK*, LNCS. Springer, 2001.

[157] A. Maedche, S. Staab, R. Studer, Y. Sure, and R. Volz. SEAL – tying up information integration and web site management by ontologies. *IEEE Computer Society Data Engineering Bulletin, Special Issue on Organizing and Discovering the Semantic Web*, 25(1):10–17, 2002.

[158] Alexander Maedche, Andreas Hotho, and Markus Wiese. Enhancing preprocessing in data-intensive domains using online-analytical processing. In *Data Warehousing and Knowledge Discovery, Second International Conference, DaWaK 2000, London, UK*, LNCS, pages 258–264. Springer, 2000.

[159] Ranjan Maitra. A statistical perspective on data mining. *J. Ind. Soc. Prob. Statist.*, 2002.

[160] C. D. Manning and H. Schuetze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts, 1999.

[161] A. McCallum, K. Nigam, and L. H. Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. In *Proc. of KDD-2000*, pages 169–178, 2000.

[162] Marina Meilă and David Heckerman. An experimental comparison of several clustering and initialization methods. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 386–395. Morgan Kaufmann, Inc., San Francisco, CA, 1998.

[163] E. Mena, V. Kashyap, A Illarramendi, and A. Sheth. Domain specific ontologies for semantic information brokering on the global information infrastructure. In N. Guarino, editor, *Proceedings of the First International Conference on Formal Ontologies in Information Systems (FOIS)*, volume 46 of *Frontiers in Artificial Intelligence and Applications*, Trento, Italy, 1998. IOS-Press.

[164] R. S. Michalski. Knowledge acquisition through conceptual clustering: A theoretical framework and an algorithm for partitioning data into conjunctive concepts. *Journal of Policy Analysis and Information Systems*, 4(3):219–244, September 1980.

[165] R. S. Michalski and R. E. Stepp. Learning from observation: Conceptual clustering. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors, *Machine Learning: An Artificial Intelligence Approach*, pages 331–363. TIOGA Publishing Co., Palo Alto, 1983.

[166] G. W. Miligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–179, 1985.

[167] G. Miller. WordNet: A lexical database for english. *CACM*, 38(11):39–41, 1995.

[168] G. W. Milligan and L. M. Sokol. A two stage clustering algorithm with robust recovery characteristics. *Educational and Psychological Measurement*, 40:755–759, 1980.

[169] T. M. Mitchell. *Maschine Learning*. McGraw-Hill, 1997.

[170] D. Mladenic. Text-learning and related intelligent agents: a survey. *IEEE Intelligent Systems*, 14(4):44–54, 1999.

[171] D. Mladenic and M. Grobelnik. Feature selection for classification based on text hierarchy. In *Working Notes of Learning from Text and the Web, Conf. Automated Learning and Discovery (CONALD-98)*. Carnegie Mellon Univ., Pittsburgh,, 1998.

[172] D. I. Moldovan and R. Mihalcea. Using WordNet and lexical operators to improve internet searches. *IEEE Internet Computing*, 4(1):34–43, 2000.

[173] Luis Carlos Molina, Lluís Belanche, and Àngela Nebot. Feature selection algorithms: A survey and experimental evaluation. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), 9-12 December 2002, Maebashi City, Japan*, pages 306–313. IEEE Computer Society, 2002.

[174] B. Motik, A. Maedche, and R. Volz. A conceptual modeling approach for semantics–driven enterprise applications. In R. Meersman, Z. Tari, et al., editors, *Proceedings of the Confederated International Conferences: On the Move to Meaningful Internet Systems (CoopIS, DOA, and ODBASE 2002)*, volume 2519 of *Lecture Notes in Computer Science (LNCS)*, pages 1082–1099, University of California, Irvine, USA, 2002. Springer.

[175] Fionn Murtagh, Jean-Luc Starck, and Michael W. Berry. Overcoming the curse of dimensionality in clustering by means of the wavelet transform. *The Computer Journal*, 43(2):107–120, 2000.

[176] U. Nahm and R. Mooney. Text mining with information extraction. In *Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, 2002.

[177] G. Neumann, R. Backofen, J. Baur, M. Becker, and C. Braun. An information extraction core system for real world german text processing. In *Proc. of ANLP-1997*, pages 208–215, 1997.

[178] Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

[179] Michael Nuhn. Clustern mit Hintergrundwissen — Implementierung eines Data Mining Tools zur Detektion von Kundengruppen bei der Deutschen Telekom AG. Mastersthesis, Institute AIFB, Universität Karlsruhe, 2003.

[180] C. K. Ogden and I. A. Richards. *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. Routledge & Kegan Paul Ltd., London, 10 edition, 1923.

[181] G. Pache. Textklassifikation mit support-vektor-maschinen unter zuhilfenahme von hintergrundwissen. Studienarbeit, Universität Karlsruhe, Germany, April 2002.

[182] Patrick Pantel and Dekang Lin. Document clustering with committees. In *Proceedings of SIGIR'02, Tampere, Finland*, 2002.

[183] Dan Pelleg and Andrew Moore. Accelerating exact k -means algorithms with geometric reasoning. In *Knowledge Discovery and Data Mining*, pages 277–281, 1999.

[184] Dan Pelleg and Andrew Moore. X-means: Extending K-means with efficient estimation of the number of clusters. In *Proc. 17th International Conf. on Machine Learning*, pages 727–734. Morgan Kaufmann, San Francisco, CA, 2000.

[185] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[186] D. Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann Publishers, San Francisco, California, 1999.

[187] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, California, 1993.

[188] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, UK., 2nd ed. edition, 1979.

[189] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In D. K. Harman, editor, *Third Text Retrieval Conference (TREC-3)*, 1995.

[190] G. Salton, editor. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice-Hall, 1971.

[191] G. Salton and M. J. McGill. *Introduction to Modern Retrieval*. McGraw-Hill Book Company, 1983.

[192] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[193] Sergio M. Savaresi, Daniel Boley, Sergio Bittanti, and Giovanna Gazzaniga. Cluster selection in divisive clustering algorithms. In Robert L. Grossman, Jiawei Han, Vipin Kumar, Heikki Mannila, and Rajeev Motwani, editors, *Proceedings of the Second SIAM International Conference on Data Mining, Arlington, VA, USA, April 11-13, 2002*. SIAM, 2002.

[194] H. Schuetze and C. Silverstein. Projections for efficient document clustering. In *Proc. of SIGIR-1997*, pages 74–81. Morgan Kaufmann, July 1997.

[195] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[196] Gholamhosein Sheikholeslami, Surojit Chatterjee, and Aidong Zhang. Wavecluster: A multi-resolution clustering approach for very large spatial databases. In Ashish Gupta, Oded Shmueli, and Jennifer Widom, editors, *VLDB'98, Proceedings of 24rd International Conference on Very Large Data Bases, August 24-27, 1998, New York City, New York, USA*, pages 428–439. Morgan Kaufmann, 1998.

[197] Amit Singhal, Gerard Salton, Mandar Mitra, and Chris Buckley. Document length normalization. *Information Processing and Management*, 32(5):619–633, 1996.

[198] N. Slonim, N. Friedman, and N. Tishby. Unsupervised document classification using sequential information maximization. In *SIGIR 02*. ACM, 2002.

[199] N. Slonim and N. Tishby. The power of word clusters for text classification. In *23rd European Colloquium on Information Retrieval Research*, 2001.

[200] K. Sparck-Jones and P. Willett, editors. *Readings in Information Retrieval*. Morgan Kaufmann, 1997.

[201] S. Staab, J. Angele, S. Decker, M. Erdmann, A. Hotho, A. Maedche, H.-P. Schnurr, R. Studer, and Y. Sure. Semantic community web portals. In *WWW9 — Proceedings of the 9th International World Wide Web Conference, Amsterdam, The Netherlands*, pages 473–491. Elsevier, 2000.

[202] S. Staab, C. Braun, I. Bruder, A. Duesterhoeft, A. Heuer, M. Klettke, G. Neumann, B. Prager, J. Pretzel, H.-P. Schnurr, R. Studer, H. Uszkoreit, and B. Wrenger. A system for facilitating and enhancing web search. In *Proceedings of International Working Conference on Artificial and Natural Neural Networks: Engineering Applications of Bio-Inspired Artificial Neural Networks (IWANN'99)*, volume 1607 of *LNCS*, pages 706–714, Berlin, 1999. Springer Verlag.

[203] S. Staab, C. Braun, A. Düsterhöft, A. Heuer, M. Klettke, S. Melzig, G. Neumann, B. Prager, J. Pretzel, H.-P. Schnurr, R. Studer, H. Uszkoreit, and B. Wrenger. GETESS — searching the web exploiting german texts. In *Proceedings of the 3rd Workshop on Cooperative Information Agents, Uppsala, Sweden*, LNCS, pages 113–124. Springer, 1999.

[204] S. Staab and A. Maedche. Ontology engineering beyond the modeling of concepts and relations. In *ECAI-2000 - European Conference on Artificial Intelligence. Proceedings of the 13th European Conference on Artificial Intelligence*. IOS Press, Amsterdam, 2000.

[205] Steffen Staab and Alexander Maedche. Knowledge portals — ontologies at work. *AI Magazine*, 21(2), Summer 2001.

[206] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*, 2000.

[207] Michael Steinbach, Levent Ertoz, and Vipin Kumar. Challenges of clustering high dimensional data. In L. T. Wille, editor, *New Vistas in Statistical Physics – Applications in Econophysics, Bioinformatics, and Pattern Recognition*. Springer-Verlag, 2003.

[208] D. Steinhausen and K. Langer. *Clusteranalyse Einführung in Methoden und Verfahren der automatischen Klassifikation*. Walter de Gruyter & Co., 1977.

[209] A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In *Proc. of Workshop of Artificial Intelligence for Web Search*, pages 58–64. AAAI, 2000.

[210] R. Studer, V. R. Benjamins, and D. Fensel. Knowledge engineering principles and methods. *Data and Knowledge Engineering*, 25(1–2):161–197, 1998.

[211] R. Studer, Y. Sure, and R. Volz. Managing focused access to distributed knowledge. *Journal of Universal Computer Science (J.UCS)*, 8(6):662–672, 2002.

[212] G. Stumme, A. Hotho, and B. Berendt, editors. *Semantic Web Mining*, Freiburg, September 3rd 2001. 12th Europ. Conf. on Machine Learning (ECML'01) / 5th Europ. Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'01).

[213] Gerd Stumme. Using ontologies and formal concept analysis for organizing business knowledge. In *Proc. Wissensmanagement mit Referenzmodellen – Konzepte für die Anwendungssystem- und Organisationsgestaltung*, pages 163–174. Physica, Heidelberg, 2002.

[214] Y. Sure. *Methodology, Tools & Case Studies for Ontology based Knowledge Management*. PhD thesis, Universität Karlsruhe (TH), Institut für Angewandte Informatik und Formale Beschreibungsverfahren, 2003.

[215] Y. Sure and J. Angele, editors. *Proceedings of the First International Workshop on Evaluation of Ontology based Tools (EON 2002)*, volume 62 of *CEUR Workshop Proceedings*, Sundial Resort, Sanibel Island, Florida, USA, 2002. 2nd International Semantic Web Conference. available at http://CEUR-WS.org/Vol-62/.

[216] Julien Tane, Christoph Schmitz, and Gerd Stumme. Semantic resource management for the web: An elearning application. In *Submitted to the Thirteenth International World Wide Web Conference (WWW 2004)*, New York, May 2004.

[217] Julien Tane, Christoph Schmitz, Gerd Stumme, Steffen Staab, and Rudi Studer. The courseware watchdog: an ontology-based tool for finding and organizing learning material. In *Fachtagung "Mobiles Lernen und Forschen"*, Kassel, Germany, Nov 2003. Uni Kassel.

[218] Alexandre Termier, Michèle Sebag, and Marie-Christine Rousset. Combining statistics and semantics for word and document clustering. In *Ontology Learning Workshop*, pages 49–54, Seattle, August 4 2001. IJCAI'01.

[219] B. Thiesson, C. Meek, D. Chickering, and D. Heckerman. Learning mixtures of bayesian networks, 1997.

[220] TOVE: Manual of the Toronto Virtual Enterprise, 1995. available at http://www.eil.utoronto.ca/enterprise-modelling/.

[221] M. Uschold and M. Gruninger. Ontologies: Principles, methods and applications. *Knowledge Sharing and Review*, 11(2):93–155, June 1996.

[222] M. Uschold, M. King, S. Moralee, and Y. Zorgios. The enterprise ontology. *Knowledge Engineering Review*, 13(1):31–89, 1998.

[223] G. van Heijst, A. Th. Schreiber, and B. J. Wielinga. Using explicit ontologies for kbs development. *International Journal of Human-Computer Studies*, 46(2/3):183–292, 1997.

[224] Ellen M. Voorhees. Query expansion using lexical-semantic relations. In W. Bruce Croft and C. J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*, pages 61–69. ACM/Springer, 1994.

[225] Kiri Wagsta, Claire Cardie, Seth Rogers, and Stefan Schroedl. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 577–584, 2001.

[226] A. Weingessel, E. Dimitriadou, and S. Dolnicar. An examination of indexes for determining the number of clusters in binary data sets. Technical Report Working Paper 29, SFB "Adaptive Information Systems and Modeling in Economics and Management Science", 1999.

[227] G. Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, 25(3):38–49, 1992.

[228] Y. Wilks. Information extraction as a core language technology. In M-T. Pazienza, editor, *Information Extraction*. Springer, Berlin, 1997.

[229] R. Wille. Restructuring lattice theory: an approach based on hierarchies of concepts. In I.Rival, editor, *Ordered sets*, pages 445–470, Dordrecht-Boston, 1982. Reidel.

[230] Karsten Winkler and Myra Spiliopoulou. Structuring domain-specific text archives by deriving a probabilistic xml dtd. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *Principles of Data Mining and Knowledge Discovery, 6th European Conference, PKDD 2002, Helsinki, Finland, August 19-23, 2002, Proceedings*, volume 2431 of *Lecture Notes in Computer Science*, pages 461–474. Springer, 2002.

[231] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.

[232] Yiming Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1/2):69–90, 1999.

[233] Sarah Zelikovitz and Haym Hirsh. Improving text classification with lsi using background knowledge. In *IJCAI01 Workshop Notes on Text Learning: Beyond Supervision*, 2001.

[234] G. K. Zipf. *Selected Studies of the Principle of Relative Frequency in Language.* Harvard University Press, Cambridge, Massachusetts, 1932.

[235] Youyong Zou, Tim Finin, Li Ding, Harry Chen, and Rong Pan. Using semantic web technology in multi-agent systems: a case study in the taga trading agent environment. In *Proceedings of the 5th international conference on Electronic commerce*, pages 95–101. ACM Press, 2003.

Die oben angegebenen URLs wurden zuletzt am 20.12.2003 überprüft.

Sachwortregister