Bauhaus-Universität Weimar
Fakultät Medien
Studiengang Mediensysteme

Cluster-Labeling

Paradigmen und Validierung

Masterarbeit

Dennis Hoppe

1. Gutachter: Prof. Dr. Benno Stein

Datum der Abgabe: 21. Juni 2010

Erklärung

Hiermit versichere ich, daß ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Weimar, den 21. Juni 2010

. . . . . .

Dennis Hoppe

Kurzfassung

Im Information Retrieval werden Verfahren eingesetzt, um Textdokumente automatisch in inhaltliche Gruppen einzuteilen. Bei den Verfahren handelt es sich um so genannte Clustering-Verfahren. Die resultierenden Gruppen von Dokumenten werden Cluster genannt. Eines der zentralen Probleme hierbei ist, die Cluster (ebenfalls automatisch) zu beschriften. Für jeden Cluster ist ein Name zu vergeben, der den Inhalt der im jeweiligen Cluster enthaltenen Dokumente prägnant beschreibt und verschiedene Cluster voneinander abgrenzt. Das Problem wird Cluster-Labeling genannt. In dieser Arbeit werden Verfahren zum Cluster-Labeling für Cluster von Textdokumenten untersucht. Hierbei existieren drei verschiedene Herangehensweisen: Datenzentrierte, beschreibungsbeachtende und beschreibungszentrierte Verfahren. Letztgenannte stellen erstmalig die Qualität von Cluster-Labeln in den Vordergrund des Clustering-Prozesses. Folgende Cluster-Labeling-Verfahren werden in dieser Arbeit gegenübergestellt: Frequent and Predictive Words, Weighted Centroid Covering, Suffixbaum-Clustering, Descriptive k-Means und Lingo. Es wird zudem ein neues beschreibungszentriertes Cluster-Labeling-Verfahren namens Topical k-Means vorgestellt. Dieses Verfahren erzielt eine Precision@1 von 0,59 und ist damit das zweitbeste der in dieser Arbeit untersuchten Cluster-Labeling-Verfahren. Das beste Verfahren ist Descriptive k-Means. Es erzielt eine Precision@1 von 0,74. Im Vergleich zum besten Verfahren löst Topical k-Means das Cluster-Labeling-Problem einfacher.

Bislang existiert keine einheitliche Vorstellung darüber, was ein für den Nutzer verständliches Cluster-Label ausmacht. Daher werden in dieser Arbeit wünschenswerte Eigenschaften ermittelt und formalisiert. Diese sind: Verständlichkeit, Überdeckung, Trennschärfe, Minimale Überlappung, Eindeutigkeit und Redundanzfreiheit. Anhand dieser Formalisierung wird ein neues, auf Normalized Discounted Cumulative Gain (NDCG) basierendes, internes Validierungsmaß vorgestellt. Topical k-Means erzielt hier erneut das zweitbeste Ergebnis bei einem NDCG@1 von 0,84. Descriptive k-Means erzielt das beste Ergebnis bei einem NDCG@1 von 0,87.

Inhaltsverzeichnis

1 Einleitung **1**

2 Hintergrund **4**

2.1 Ermittlung wichtiger Worte und Phrasen in Text 5

2.1.1 Verfahren zur Ermittlung von Worten 5

2.1.2 Verfahren zur Ermittlung von Phrasen 11

2.2 Repräsentation von Text 19

2.2.1 Indexierung 21

2.2.2 Dokumentmodelle 21

2.3 Clustering 23

2.3.1 Arten des Clusterings 24

2.3.2 Clustering-Verfahren 25

3 Was sind verständliche Cluster-Label? **29**

4 Problematik des Cluster-Labelings **41**

4.1 Problemstellung 42

4.2 Formalisierung wünschenswerter Label-Eigenschaften 43

4.3 Anforderungen an ein Clustering 47

4.4 Ermittlung von Themen – ein Überblick 48

5 Cluster-Labeling-Verfahren **59**

5.1 Datenzentrierte Verfahren 60

5.1.1 Frequent Predictive Words 60

5.1.2 Weighted Centroid Covering 63

5.1.3 Zusammenfassung 64

5.2 Beschreibungsbeachtende Verfahren 65

5.2.1 Suffixbaum-Clustering 65

5.2.2 Zusammenfassung 67

5.3 Beschreibungszentrierte Verfahren 67
5.3.1 Lingo 69
5.3.2 Descriptive k-Means 73
5.3.3 Zusammenfassung 76

6 Topical k-Means 77
6.1 Clustering einer Dokumentkollektion 78
6.2 Ermittlung von Cluster-Labeln 79
6.3 Erstellung von Clustern 79
6.4 Filterung von Phrasen 80

7 Evaluierung von Cluster-Labeling-Verfahren 82
7.1 Externe Gütekriterien zur Validierung von Cluster-Labeln 83
7.1.1 Precision@R, Match@R und MRR 84
7.1.2 Normalized Discounted Cumulative Gain (NDCG) 85
7.2 Interne Gütekriterien zur Validierung von Cluster-Labeln 87
7.2.1 Quantifizierung wünschenswerter Label-Eigenschaften 87
7.2.2 NDCG-basiertes Validierungsmaß 89
7.3 Externe Gütekriterien zur Validierung eines Clusterings 91

8 Experimente 95
8.1 Experimentbeschreibung 95
8.1.1 Referenzkorpora zur Durchführung von Experimenten 96
8.1.2 Experimentparameter 98
8.2 Experimentergebnisse 101
8.2.1 Ermittlung von Themen 101
8.2.2 Validierung von quantifizierten Label-Eigenschaften 107
8.2.3 Evaluierung von Cluster-Labeling-Verfahren 109

9 Zusammenfassung 117

Literaturverzeichnis 119

A Analyse zur Ermittlung von Themen 132

B Empirische Studie zu quantifizierten Label-Eigenschaften 139

1 Einleitung

Clustering-Verfahren gruppieren Objekte anhand von Ähnlichkeiten in Teilmengen – so genannte Cluster. Verfahren des Cluster-Labelings erzeugen für einen Cluster eine Cluster-Beschriftung (engl. label). Im Bereich des Information-Retrieval werden solche Verfahren häufig eingesetzt. Beispielsweise können ähnliche Suchergebnisse von Suchmaschinen gruppiert werden. Das soll Nutzern dabei helfen, einen schnellen Überblick über die Suchergebnisse zu erlangen. Entscheidend für eine nutzbringende Gruppierung ist dabei eine prägnante, passende Beschriftung der gefundenen Cluster. Fehlen passende Cluster-Label ist ein Nutzer gezwungen, jeden Cluster auf relevante Suchergebnisse hin zu überprüfen. Eine verbesserte Präsentation der Suchergebnisse wird in diesem Fall, unabhängig von der Qualität der gefundenen Cluster, nicht erreicht.

In traditionellen, *datenzentrierten* Clustering-Prozessen nimmt Cluster-Labeling, entgegen der gezeigten Wichtigkeit, eine untergeordnete Rolle ein: Nach passenden Cluster-Labeln wird erst gesucht, nachdem das Clustering-Verfahren eine Dokumentkollektion in Cluster zerlegt hat. Hinzu kommt, dass sich die Menge der potenziellen Cluster-Label aus dem Dokumentmodell des Clustering-Verfahrens ergibt. Dieses basiert in der Regel auf Unigrammen, welche allerdings als Cluster-Label oft ungeeignet sind. Ein Beispiel: Für den Film Eyes Wide Shut sind die Cluster-Label „eyes“, „wide“ und „shut“ alleinstehend nicht verständlich.

Ein neues Paradigma zur Erstellung von Cluster-Labeln führen so genannte *beschreibungszentrierte* Clustering-Verfahren ein. Diese rücken das Erzeugen prägnanter Cluster-Label in den Vordergrund des Clustering-Prozesses. Anstatt potenzielle Cluster-Label aus dem Dokumentmodell des Clustering-Verfahrens zu ermitteln, werden diese in der Dokumentkollektion selbst gesucht.

Dabei kommen Verfahren der Schlüsselwortbestimmung zum Einsatz. Es werden beispielsweise Nominalphrasen ermittelt, die häufig in Dokumenten eines Clusters vorkommen. Sie bilden die Menge potenzieller Cluster-Label. Um aus dieser Menge eine Auswahl für die endgültige Cluster-Label zu treffen, wird mit einem traditionellen, polythetischen Clustering-Verfahren ein Clustering der Dokumentkollektion erzeugt. Die potenziellen

Cluster-Label, die am besten zu den Clustern passen, bilden die Menge der ausgewählten Cluster-Label. Jedes ausgewählte Cluster-Label definiert einen Cluster: die Menge der Dokumente, die das Cluster-Label enthalten. Durch diese monothetische Vorgehensweise entsteht eine für den Nutzer nachvollziehbare Beziehung zwischen Cluster-Label und Cluster-Inhalt.

Diese Arbeit bietet einen Überblick über aktuelle datenzentrierte sowie beschreibungszentrierte Verfahren des Cluster-Labelings. Zusätzlich wird die Gruppe der beschreibungsbeachtenden Cluster-Labeling-Verfahren betrachtet. Diese ermitteln in der Dokumentkollektion zunächst potenzielle Cluster-Label, um mittels dieser den Clustering-Prozess zu beeinflussen. Auch diese Verfahren zeichnen sich durch eine monothetische Vorgehensweise aus.

Die Verfahren im Einzelnen sind: Frequent and Predictive Words (Popescul u. Ungar, 2000), Weighted Centroid Covering (Stein u. Meyer zu Eißen, 2004), Suffixbaum-Clustering (Zamir u. Etzioni, 1998), Descriptive k-Means (Stefanowski u. Weiss, 2007) und Lingo (Osinski u. a., 2004). Zudem wird ein neues, beschreibungszentriertes Verfahren unter dem Namen *Topical k-Means* präsentiert. Für die in einer Untersuchung verwendeten Dokumentkollektionen aus dem Open Directory Project (ODP) wird die Leistungsfähigkeit von Topical k-Means aufgezeigt.

Neben der Erzeugung von Cluster-Labeln stellt die Validierung dieser eine der großen Herausforderungen im Bereich des Cluster-Labelings dar. Während bei einer gegebenen Referenzkategorisierung die Qualität eines Clusterings anhand der Überschneidung der erzeugten Cluster mit den Referenzklassen gemessen werden kann, ist die semantische Nähe zwischen einem gegebenem Referenz-Label und einem erzeugten Cluster-Label nur sehr schwierig formal zu erfassen. Daher gehen gegenwärtige externe Validierungsmaße von einer nach Relevanz sortierten Liste von Cluster-Labeln aus, die ein Cluster-Labeling-Verfahren für einen Cluster erstellt. Die Qualität einer solchen Liste wird auf Grundlage der partiellen sowie exakten Übereinstimmung mit dem Referenz-Label ermittelt. Die Cluster-Labeling Verfahren werden in dieser Arbeit anhand von in der Literatur zu findenden externen Validierungsmaßen bewertet. Diese sind: Precision@R, Match@R und Mean Reciprocal Rank (MRR).

Im Gegensatz zu externen Validierungsmaßen stehen internen Validierungsmaßen keine Referenz-Label zur Bewertung von Cluster-Labeln zur Verfügung. Daher werden Qualitätsmerkmale definiert, die sich aus den intrinsischen Eigenschaften der beschrifteten Cluster ergeben. Im Rahmen dieser Arbeit wird ein, auf Normalized Discounted Cumulative Gain (NDCG) basierendes, internes Validierungsmaß vorgestellt, welches die Posi-

tion eines Cluster-Labels in der Liste der Cluster-Label berücksichtigt. NDCG bewertet die Güte eines erzeugten Clusterings anhand von sechs Qualitätsmerkmalen: Verständlichkeit, Überdeckung, Trennschärfe, Minimale Überlappung, Eindeutigkeit und Redundanzfreiheit. Die einzelnen Merkmale werden dabei zunächst motiviert und anschließend formalisiert. In einer Studie wird gezeigt, dass sich die eingeführten Qualitätsmerkmale für den qualitativen Vergleich von Cluster-Labeling-Verfahren eignen.

Gliederung der Arbeit

Im folgenden Kapitel werden unter anderem Verfahren zur Schlüsselwortbestimmung vorgestellt. Schlüsselworte werden zum Labeling von Clustern verwendet. Kapitel 3 analysiert die von menschlichen Experten erstellten Kategorien im Open Directory Project und Wikipedia. Es werden Anforderungen ermittelt, die gute Cluster-Label erfüllen sollten. Diese Anforderungen werden in Kapitel 4 formalisiert. Kapitel 5 stellt die genannten Cluster-Labeling-Verfahren vor und diskutiert Vor- und Nachteile. Im Anschluss wird das neue beschreibungszentrierte Verfahren Topical k-Means vorgestellt. Verfahren des Cluster-Labelings werden im weiteren Verlauf der Arbeit anhand externer und interner Validierungsmaße bewertet. Kapitel 9 fasst die Ergebnisse zusammen.

2 Hintergrund

Dieses Kapitel diskutiert Aspekte, die zum Erfolg eines Cluster-Labeling-Verfahrens beitragen:

Ermittlung wichtiger Worte und Phrasen in Text Cluster-Label sollen kurz, prägnant und den Inhalt eines Clusters umfassend beschreiben. Solche Forderungen lassen sich leicht informell formulieren, deren Quantifizierung ist jedoch schwierig. Um zunächst Themen im Text zu ermitteln, werden Verfahren zur Schlüsselwortbestimmung eingesetzt. Schlüsselworte sind Kandidaten für Cluster-Label. Es wird argumentiert, dass die Art der Schlüsselwortbestimmung essentiell für den Erfolg eines Cluster-Labeling-Verfahrens ist. Nicht jedes Verfahren der Schlüsselwortbestimmung eignet sich gleichermaßen, um Phrasen aus Text zu ermitteln, die den Anforderungen eines Cluster-Labels gerecht werden. Deshalb werden verschiedene Verfahren diskutiert.

Repräsentation von Text Für das Clustering einer unstrukturierten Textkollektion ist diese zuvor aufzubereiten. Zur maschinellen Verarbeitung der Texte sind hierzu verschiedene Modelle vorstellbar. Ein Text kann durch die Gesamtheit seiner Worte repräsentiert werden. Jedem Wort wird dabei ein Gewicht zugewiesen, welches die Wichtigkeit des Wortes für den Text beschreibt. Texte sind durch die Repräsentation untereinander vergleichbar, so dass diese mittels eines Clusterings gruppiert werden können. Es werden Dokumentrepräsentationen vorgestellt, die in dieser Arbeit verwendet werden.

Clustering Eine Gruppierung von Dokumenten durch ein Clustering kann dem Cluster-Labeling voraus gehen, ist diesem nachgestellt oder mit dem Labeling-Prozess eng verzahnt. Einige Cluster-Labeling-Verfahren nutzen das Clustering, um Themen in der Dokumentkollektion zu ermitteln. Es werden Arten des Clusterings und Clustering-Verfahren vorgestellt, die für diese Arbeit relevant sind.

2.1 Ermittlung wichtiger Worte und Phrasen in Text

Dieses Kapitel gibt einen Überblick über Verfahren zur Schlüsselwortbestimmung.

Eine Möglichkeit zur Ermittlung von Themen in Texten besteht in der automatischen Generierung von Textzusammenfassungen. Für diese werden Sätze ermittelt (Carbonell u. Goldstein, 1998; Goldstein u. a., 1999; Radev u. a., 2002a). Allerdings sind Sätze aufgrund ihrer Länge als Cluster-Label nicht geeignet, da prägnante Cluster-Label bevorzugt werden. Deshalb werden nachfolgend Verfahren zur Bestimmung von Schlüsselphrasen vorgestellt. Insbesondere interessieren Verfahren, die Schlüsselphrasen für eine Menge von Texten ermitteln. Es gilt, möglichst alle in den Texten enthaltenen Themen zu ermitteln. Dies ist ähnlich zum Cluster-Labeling-Problem.

Im weiteren Verlauf wird der Begriff *Dokument* synonym zum Begriff *Text* verwendet. Formal werden nachfolgende Begriffe definiert:

- Ein Vokabular $\mathcal{T}$ über eine Dokumentkollektion besteht aus der Menge der in den Dokumenten enthaltenen Terme $\{t_i|1 \dots L\}$.
- Eine Phrase p wird als Wort N-Gramm der Länge n mit $p = (t_1, \dots, t_n)$ definiert.
- Ein Dokument d ist gegeben durch eine Folge von M Termen $(t_1, \dots, t_M)$.
- Ein Dokument ist Teil einer Dokumentkollektion $\mathcal{D} = \{d_1, \dots, d_N\}$ mit N Dokumenten.
- Eine Dokumentkollektion ist durch eine $L \times N$-Term-Dokument-Matrix mit $A = [d_1, \dots, d_N]$ repräsentierbar. Die Matrix besitzt den Rang $rang(A) \leq min\{L, N\}$.

Um Vor- und Nachteile einzelner Verfahren zur Schlüsselwortbestimmung zu verdeutlichen, wird das Kapitel von einem Beispiel begleitet. Dieses besteht aus sieben Dokumenten, die in in Tabelle 2.1 aufgeführt sind.

2.1.1 Verfahren zur Ermittlung von Worten

Traditionelle Verfahren zur Schlüsselwortbestimmung abstrahieren ein Dokument durch Ermittlung der häufigsten Terme (Luhn, 1958). Die Wichtigkeit eines Terms $t \in d$ wird anhand seiner Häufigkeit im Dokument, der Termhäufigkeit tf_d, bewertet. Dem Term t wird ein Gewicht w zugewiesen:

$$w_{\mathrm{tf}}(t) = \mathrm{tf}_d(t).$$

D1:	Large Scale Singular Value Computations
D2:	Software for the Sparse Singular Value Decomposition
D3:	Introduction to Modern Information-Retrieval
D4:	Linear Algebra for Intelligent Information-Retrieval
D5:	Matrix Computations
D6:	Singular Value Analysis for Cryptograms
D7:	Automatic Information Organization

Tabelle 2.1 *: Dokumente aus den Bereichen „Information-Retrieval“ (IR) und „Singulärwertzerlegung“. D3, D4 und D7 sind dem Bereich IR zugeordnet. Die Dokumente D1, D2, D5 und D6 gehören zum Themenbereich „Singulärwertzerlegung“. Beispiel entnommen aus Osinski u. a. (2004).*

Da die Termhäufigkeit mit der Dokumentlänge zunimmt, sollte diese normiert werden, so dass der Beitrag längerer Dokumente zur Termhäufigkeit vermindert wird.

Ab einer bestimmten Termhäufigkeit erreicht ein Term keine größere Aussagekraft mehr. Deshalb sind zu hohe Termhäufigkeiten durch eine Logarithmusfunktion zu begrenzen (Harman, 1986).

Mit Hilfe der Termhäufigkeit lassen sich die *wichtigsten* Terme für das Beispiel in Tabelle 2.1 ermitteln. Die fünf Terme mit dem höchsten Gewicht sind: *for* (3), *information* (3), *value* (3), *singular* (3) und *computations* (2). Die Termhäufigkeit ist in Klammern angegeben.

In diesem Beispiel wird unter anderem die Präposition *for* als *bester* Term ermittelt. Die Präposition spiegelt allerdings nicht den Inhalt der Dokumente wider. Diese Art der Termgewichtung ist also nicht ausnahmslos zur Ermittlung der wichtigsten Terme in Dokumenten geeignet. Bei Präpositionen wie *for* handelt es sich um sogenannte Stoppworte. Diese transportieren keine Informationen für den Nutzer. Stoppworte können deshalb aus den Texten entfernt werden. Nach der Entfernung dieser ergibt sich für das eingeführte Beispiel eine neue Rangfolge: *information* (3), *singular* (3), *retrieval* (2), *computations* (2) und *decompositions* (1).

Die Entfernung von Stoppworten reicht oftmals nicht aus. Terme, die sehr häufig im Dokument auftreten, besitzen in der Regel nicht genügend Trennschärfe, um das Dokument von anderen zu unterscheiden. Ein Beispiel: Schlüsselworte einer wissenschaftlichen Arbeit aus dem Bereich *Information-Retrieval* sind zu ermitteln. Als häufigste Terme werden *Information-Retrieval*, *Dokument* und *Evaluierung* ermittelt. Diese Begriffe ge-

hören jedoch zum themenspezifischen Vokabular, welches im Information-Retrieval verwendet wird. Die Begriffe besitzen deshalb für das Dokument nur wenig Aussagekraft. Dagegen sind Schlüsselworte wie *Naive-Bayes Klassifikation* oder *Assoziationsanalyse* zu bevorzugen.

Zur Ermittlung von themenspezifischem Vokabular stellen Yang u. Wilbur (1996) entsprechende Methoden vor. Zur Verbesserung der Qualität von Schlüsselphrasen kann dies ebenso wie Stoppworte entfernt werden. Für das dem Kapitel begleitende Beispiel trifft dies auf die Begriffe *information*, *retrieval*, *singular* und *decomposition* zu.

TF-IDF

Wird das Problem der Stoppworte von einem Dokument auf eine Menge von Dokumenten übertragen[1], so existieren ebenfalls Terme, die wenig Informationen transportieren und daher zu vernachlässigen sind. Terme, die in vielen Dokumenten vorkommen sind nicht in der Lage, ein einzelnes Dokument ausreichend gut zu repräsentieren. Wertvoller sind dagegen Terme, die in nur einem einzigen Dokument vorkommen. Spärck Jones (1972) verwenden zur Gewichtung von Termen die inverse Dokumenthäufigkeit *idf*. Dieses Maß bestraft Terme, die sehr häufig in der Dokumentkollektion vorkommen:

$$w_{\text{idf}}(t) = \log\left(\frac{|\mathcal{D}|}{n_t}\right) + 1.$$

n_t ist die Anzahl der Dokumente, die den Term t enthalten. Mittels *idf*-Gewichtung ergibt sich für das Beispiel: *organization* (2,95), *cryptograms* (2,95), *decomposition* (2,95), *analysis* (2,95) und *intelligent* (2,95). Die Werte in den Klammern geben nun die inverse Dokumenthäufigkeit der Terme an. Häufig in der Dokumentkollektion vorkommende Terme wie *information* oder *singular* erzielen einen deutlich niedrigeren Wert (hier: 1,85). Das Maß wird minimal für Terme, die in jedem Dokument auftreten und maximal ($w_{\text{idf}}(t) = 1$) für Terme, die in nur einem einzigen Dokument enthalten sind.

Salton u. Yang (1973) schlagen das Termgewichtungsmaß *tf-idf* vor. Dieses kombiniert die Term- mit der Dokumenthäufigkeit:

$$w_{\text{tf-idf}}(t) = w_{\text{tf}}(t) \cdot w_{\text{idf}}(t).$$

Es existieren viele Varianten der *tf-idf*-Formel. Gemeinsame Eigenschaft der Varianten ist, dass diese die Wichtigkeit eines Terms innerhalb eines Dokuments und in der Dokumentkollektion hervorheben.

[1] Bislang wurden die Dokumente des Beispiels vielmehr als ein Dokument betrachtet.

Für das Beispiel ergibt sich durch Gewichtung mit *tf-idf*: *introduction* (2,54), *singular* (2,54), *intelligent* (2,51), *computations* (2,51) und *algebra* (1,95).

TF-PDF

Termgewichtungsmaße wie *tf*, *idf* und vor allem *tf-idf* sind weit verbreitet und erzielen gute Ergebnisse bei der Ermittlung von Schlüsselworten (Witten u. a., 1999). In Bezug auf die Schlüsselwortbestimmung für eine Menge von Dokumenten und somit auch im Fall des Cluster-Labelings wird argumentiert, dass sowohl *idf* und *tf-idf* für die Ermittlung von wichtigen Termen ungeeignet sind. Ziel des Cluster-Labelings ist es, für ein Cluster Label zu ermitteln, die möglichst alle Dokumente im Cluster repräsentieren. Deshalb ist es kontraproduktiv, die Terme mittels *tf-idf* zu gewichten. Die Multiplikation mit der inversen Dokumenthäufigkeit favorisiert Terme, die nur in wenigen Dokumenten vorkommen. Es besteht das Risiko, die Terme, die in vielen Dokumenten auftreten und daher repräsentativ für das Cluster sind, zu unterschlagen. Deshalb sollte die *idf*-Komponente durch einen Ausdruck ersetzt werden, der Terme, die in vielen Dokumenten der Dokumentkollektion auftreten, bevorzugt. Bun u. Ishizuka (2001) multiplizieren deshalb die Termhäufigkeit mit der proportionalen Dokumenthäufigkeit (*pdf*). Diese Termgewichtung wird mit *tf-pdf* bezeichnet. Die Termhäufigkeit fließt wie bei der ursprünglichen *tf-idf*-Gewichtung weiterhin linear in das Termgewicht ein. Die Dokumenthäufigkeit wird mittels einer Exponentialfunktion stärker gewichtet. Dadurch werden nun Terme bevorzugt, die in vielen Dokumenten vorkommen.

Latent-Semantic-Indexing (LSI)

Die bisher vorgestellten Termgewichtungsmaße behandeln allerdings nicht folgende Sachverhalte natürlicher Sprachen:

Synonyme: Verschiedene Worte besitzen dieselbe Bedeutung. Beispiel: Karotte, Möhre, Mohrrübe. Es handelt sich um sinnverwandte Worte.

Äquivokationen: Dieselben Worte besitzen verschiedene Bedeutungen. Beispiel: Läufer meint unter anderem einen Teppich, eine Spielfigur beim Schach oder eine schnelle Tonfolge in der Musik. Diese lexikalischen Mehrdeutigkeiten lassen Fehlschlüsse zu. Äquivokationen werden in Polyseme und Homonyme unterteilt. Erstere besitzen eine gemeinsame Wortherkunft, letztere unterschiedliche.

	D1	D2	D3	D4	D5	D6	D7
Information	0	0	**1**	**1**	0	0	1
Singular	**1**	**1**	0	0	0	**1**	0
Value	**1**	**1**	0	0	0	**1**	0
Computations	1	0	0	0	1	0	0
Retrieval	0	0	**1**	**1**	0	0	0

Tabelle 2.2 *: Term-Dokument-Matrix des Beispiels aus 2.1 als Inzidenzmatrix. Einem Element* (i, j) *der Matrix wird eine* 1 *zugewiesen, wenn ein Term* i *im Dokument* j *auftritt. Anderenfalls ist* $(i, j) = 0$*. Die Terme „Information“ und „Retrieval“ treten immer in denselben Dokumenten auf (rot). Gleiches gilt für „Singular“ und „Value“ (blau). Jeweils beide Terme korrelieren also miteinander, so dass es sich vermutlich um eine Kollokation handelt. Exakte Schlüsse sind zu ziehen, sobald in der Matrix Termhäufigkeiten der Worte hinterlegt sind.*

Bei der Ermittlung von Termen in Dokumenten sollten Synonyme auf einen einheitlichen Begriff abgebildet werden. Es ist sonst möglich, dass unter anderem drei als Cluster-Label ausgewählte Terme verschieden sind, aber die gleiche Bedeutung besitzen. Synonyme sollten deshalb als Cluster-Label vermieden werden. Angenommen, ein Clustering-Verfahren erzeugt Cluster, indem nur Dokumente einem Cluster zugewiesen werden, wenn diese das Cluster-Label enthalten. Werden Synonyme hier nicht berücksichtigt, resultiert dies in vielen kleinen Clustern, von denen einige eigentlich zusammengehören. Diese sollten zusammengelegt werden.

Werden Äquivokationen nicht aufgelöst, so erkennt eine Schlüsselwortbestimmung vielleicht nicht alle Themen der Dokumente. In Bezug auf ein Clustering gruppiert dieses thematisch unterschiedliche Dokumente. Es ergeben sich wenige, größere Cluster, die besser getrennt werden sollten.

Synonyme und Äquivokationen können mit Hilfe externer Wissensbasen wie Wikipedia oder der lexikalischen Datenbank *WordNet* aufgelöst werden (Mihalcea, 2007; Navigli, 2009). Anstatt die ermittelten Terme zu verwenden, werden diese auf einheitliche Konzepte abgebildet. Ein Beispiel: *Möhre* und *Mohrrübe* werden auf *Karotte* abgebildet. Allerdings steht nicht immer eine externe Wissensbasis zur Verfügung, um diese Abbildung zu leisten. Deerwester u. a. (1990) behaupten, dass eine Singulärwertzerlegung einer Term-Dokument-Matrix das Problem der Synonyme und Äquivokationen bewältigt. Eine Term-Dokument-Matrix für das begleitende Beispiel zeigt Tabelle 2.2.

Synonyme tendieren dazu, gemeinsam mit derselben Menge von Termen in Dokumenten aufzutreten. Dieses nutzen Deerwester u. a. zur Auflösung von Synonymen aus. Eine Dimensionsreduktion projiziert gemeinsam auftretende Terme (Kollokationen) schließlich in dieselbe Dimension eines Vektorraums.

Im Unterschied zu Synonymen besitzen Äquivokationen mehrere Bedeutungen, so dass diese in verschiedenen Kontexten auftreten. Polyseme werden deshalb durch eine Projektion auf verschiedene Dimensionen abgebildet. Die verschiedenen Bedeutungen eines Terms können somit unterschieden werden.

Die Projektion wird mittels einer Singulärwertzerlegung einer $L \times N$-Term-Dokument-Matrix A durchgeführt. Die Zerlegung führt auf eine im Rang reduzierte Matrix A_k mit $k \ll L, N$. A_k ergibt sich aus den k größten Eigenwerten $\{\sigma_i | 1 \dots k\}$ von A:

$$A_k = T_k S_k D_k^T. \tag{2.1}$$

Es gilt $S_k = \mathrm{diag}(\sigma_1, \dots, \sigma_k)$ mit $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_k$. T_k und D_k bestehen jeweils aus den linken und rechten Singulärvektoren. Durch T_k wird das ursprüngliche Dokument im k-dimensionalen Unterraum repräsentiert. Je kleiner k gewählt wird, desto schlechter wird das ursprüngliche Dokument approximiert. Ziel ist es deshalb, die wichtigsten Eigenschaften aus A bei gleichzeitiger Minimierung des Approximationsfehlers zu ermitteln. In Bezug auf die Schlüsselwortbestimmung gibt k zudem vor, wie viele Schlüsselworte zu ermitteln sind.

Es soll der optimale Wert für k bestimmt werden. Sei $q \in [0, 1]$. q gibt an, wie gut A durch A_k approximiert wird. k wird durch den minimalen Wert bestimmt, der der Ungleichung

$$q \leq \frac{\|A_k\|_F}{\|A\|_F} \tag{2.2}$$

genügt. Der Rang k ist mit Hilfe der *Frobenius*-Norm bestimmbar. Diese ist für $X = A - A_k$ definiert durch

$$\|X\|_F = \sqrt{\sum_{i=1}^{L} \sum_{j=1}^{L} x_{ij}^2}.$$

Deerwester u. a. nennen das Verfahren *Latent-Semantic-Indexing* (LSI). Ein wesentlicher Nachteil von LSI ist die aufwändige Berechnung der Singulärwerte für große Datenmengen. Dies gilt bereits, wenn die Term-Dokument-Matrix dünn besetzt ist. Daher werden gegenwärtig effizientere Verfahren eingesetzt: *Principal Component Analysis* (Jolliffe, 2002) oder *Latent Dirichlet Allocation* (Blei u. a., 2003).

2.1.2 Verfahren zur Ermittlung von Phrasen

Bislang wurde ausschließlich die Ermittlung einzelner Terme betrachtet. Alleinstehende Terme besitzen allerdings weniger Aussagekraft als zusammengesetzte. Begriffe wie *Information-Retrieval* und *Singular Value* werden durch die wortweise Indexierung getrennt. Die eigentliche Bedeutung der Begriffe geht gänzlich verloren. Um einen Bedeutungsverlust zu vermeiden, sollten deshalb vorrangig Phrasen ermittelt werden. Diese eignen sich als Cluster-Label, so dass im Folgenden Vor- und Nachteile von Ansätzen zur Ermittlung von Phrasen diskutiert werden.

Wort N-Gramme

Anstelle von einzelnen Worten werden in den Dokumenten Wort N-Gramme ermittelt. Dabei handelt es sich zumeist um Bigramme und Trigramme (Nagao u. Mori, 1994; Zamir u. Etzioni, 1999; Lin u. Hovy, 2003; Weiss, 2006; Kumar u. Srinathan, 2008; Nguyen u. a., 2009). Hierzu wird ein Fenster fester Größe[2] wortweise durch den Text geschoben und jedes erfasste Wort N-Gramm ermittelt. Es ergeben sich für Dokument *D4* aus Tabelle 2.1 folgende Bigramme:

[Linear Algebra], [Algebra for], [for Intelligent],
[Intelligent Information] und [Information-Retrieval].

Das Beispiel zeigt, dass bei einer einfachen Ermittlung aller N-Gramme viele nicht als bedeutungsvoll anzusehen sind. Darunter fallen hier *Algebra for* und *for Intelligent*. Mittels Heuristiken ist die Qualität der Wort N-Gramme allerdings zu steigern (Zhang, 2002). Eine Regel besteht darin, zu verlangen, dass ein Trigramm nicht mit einem Stoppwort anfangen und enden darf. Eine andere ist, dass bei einem Bigramm mindestens eines der beiden Worte ein Nomen ist.

Kollokationen

Um *interessante* N-Gramme zu finden, werden Kollokationen im Text gesucht (Manning u. a., 2002).

Kollokationen sind über die Termhäufigkeit zu ermitteln. Hierbei werden erneut bedeutungslose Ausdrücke wie *for the* ermittelt. Es ergibt sich ein ähnliches Problem wie

[2] Zur Ermittlung von Bigrammen umfasst die Fenstergröße 2 Worte.

mit Stoppworten. Justeson u. Katz (1995) schlagen daher eine Filterung von Wort N-Grammen anhand bestimmter Folgen von Wortarten vor. Für Bigramme sind dies:

(Adjektiv,	Nomen)
(Nomen,	Nomen)

Für Trigramme sind dies:

(Adjektiv,	Adjektiv,	Nomen)
(Adjektiv,	Nomen,	Nomen)
(Nomen,	Adjektiv,	Nomen)
(Nomen,	Nomen,	Nomen)
(Nomen,	Präposition,	Nomen)

Zudem erachten wir aufgrund der Analyse des Open Directory Projects und Wikipedia in Kapitel 3 die Folge (Nomen, Konjunktion, Nomen) ebenfalls als sinnvoll. Siehe hierzu die Tabellen 3.4 und 3.5 auf den Seiten 36 und 37. Für das Dokument *D4* aus dem Beispiel werden unter Berücksichtigung der Heuristiken folgende Bigramme ermittelt:

[Linear Algebra], [Intelligent Information] und [Information-Retrieval].

Hypothesentests Zur Ermittlung von Kollokationen können auch statistische Hypothesentests eingesetzt werden (Manning u. a., 2002). Hierzu zählen der $\mathcal{X}^2$- und der t-Test. Beim t-Test folgt die Testgröße einer Student-t-Verteilung, wenn die Nullhypothese angenommen wird. Die Nullhypothese H_0 besagt, dass zwei Terme t_1 und t_2, beispielsweise $t_1 =$*Information* und $t_2 =$*Retrieval*, unabhängig voneinander in der Dokumentkollektion auftreten. Wird die Hypothese abgelehnt, so wird davon ausgegangen, dass beide Terme in Korrelation zueinander stehen – eine Kollokation bilden.

Nguyen u. a. (2009) setzen den t-Test ein, um bedeutungsvolle Cluster-Label zu ermitteln. Die Autoren verwenden den t-Wert der Teststatistik

$$t = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{N}}}$$

als Maß für die Bedeutsamkeit einer Kollokation. Zur Berechnung werden der Stichprobenmittelwert $\overline{X}$, die empirische Standardabweichung s, der Stichprobenumfang N und

der vermutete Erwartungswert μ_0 der Verteilung benötigt. Diese sind mittels Maximum-Likelihood Methode zu schätzen.

Der Schluss, dass der t-Wert die Bedeutsamkeit einer Kollokation widerspiegelt, trifft nicht zu. Manning u. a. (2002) belegen, dass in einer Dokumentkollektion nahezu jedes Bigramm eine Kollokation bildet. Die natürliche Sprache sei regulärer als die Annahme des t-Tests, die Stichprobe sei normalverteilt. Der t-Wert eignet sich deshalb ausschließlich zur Sortierung von Kollokationen.

Informationstheoretische Maße Neben Hypothesentests existieren informationstheoretische Maße zur Ermittlung von Kollokationen. Eines der bekanntesten Maße ist die *Pointwise Mutual Information* (PMI). Diese und weitere werden in de Winter u. de Rijke (2007) vorgestellt. Die Autoren vergleichen Maße in Bezug auf die Fähigkeit zur Ermittlung von Cluster-Labeln. Dabei erzielt PMI die besten Resultate. Manning u. a. (2002) zeigen, dass die Mutual Information zwar ein gutes Maß für die Unabhängigkeit zweier Terme ist, sich jedoch nicht als Maß für die Abhängigkeit beider eignet. Gerade letzteres ist jedoch für die Ermittlung interessanter Kollokationen Voraussetzung. Dennoch findet Mutual Information beim Cluster-Labeling Zuspruch (Geraci u. a., 2006; Carmel u. a., 2009).

Bislang wurde verlangt, dass Worte einer Kollokation benachbart auftreten müssen. Dies ist jedoch nicht zwingend erforderlich. Insbesondere im Deutschen können Kollokationen durch mehrere Worte getrennt sein. Ein Beispiel: *Der Tag ist hell.* Hier bildet *Tag–hell* eine Kollokation. Ferragina u. Gulli (2005) berücksichtigen beim Cluster-Labeling eben dies. Hierzu ermitteln sie alle Wortpaare (t_i, t_j), die innerhalb einer festen Fenstergröße auftreten.

Suffixbäume

Neben den gerade aufgezeigten Ansätzen zur Ermittlung von Bigrammen und Trigrammen ist die Verwendung eines Suffixbaumes eine weitere Möglichkeit, um gemeinsame Phrasen in Dokumenten zu ermitteln. Hierbei ist das *k-Longest Common Substring*-Problem zu lösen, bei dem die k längsten gemeinsamen Termfolgen in einer Dokumentkollektion gesucht sind. Der Suffixbaum erlaubt eine schnelle Implementation des Problems. Dieser löst das Problem für zwei Dokumente der Länge n und m in $\mathcal{O}(n)$ anstatt

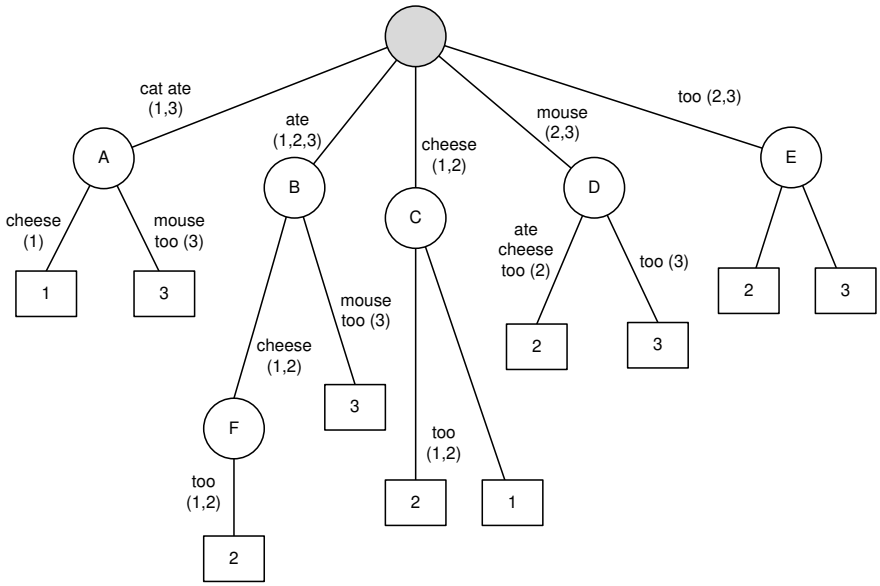

Abbildung 2.1 *: Suffix-Baum für drei Dokumente: (1) „cat ate cheese“ , (2) „mouse ate cheese tooünd (3) „cat ate mouse too“. An den Kanten sind in Klammern die Dokumente angegeben, die das jeweilige Präfix enthalten. Beispiel aus Zamir u. Etzioni (1998).*

in $\mathcal{O}(nm)$[3]. Verallgemeinert für eine Dokumentkollektion $\mathcal{D}$ ergibt sich entsprechend $\mathcal{O}(n_1 + \ldots + n_N)$ zugunsten von $\mathcal{O}(n_1 * \ldots * n_N)$ (Gusfield, 2007).

Ein Suffixbaum für d ist ein Patricia-Trie, der alle Suffixe S von $d = (t_1, \ldots, t_M)$ enthält. Ein Suffix $s \in S$ ist definiert als eine Termfolge in d, die mit dem Wort t_i beginnt und mit t_M endet. Zur Veranschaulichung dient der Suffixbaum in Abbildung 2.1. Um den Suffixbaum für das erste Dokument (*cat ate cheese*) zu erstellen, wird mit dem Einfügen des längsten Suffix $s_1 \in d = \{cat\ ate\ cheese\}$ in den Baum begonnen und mit dem kürzesten, dem leeren Wort, geendet. Jeder Suffix entspricht einem eindeutigen Pfad von der Wurzel des Baumes bis zum Blatt, wobei die Kanten entlang des Pfades die entsprechenden Präfixe von s halten. Beim Einfügen eines neuen Suffix wird ermittelt, ob von der Wurzel ausgehend bereits eine Kante für $t_i \in s_1$ existiert. Ist dies der Fall, wird die Kante durchlaufen und der nächste Knoten besucht. Handelt es sich um einen inneren Knoten, wird erneut geprüft, ob der Präfix der Kante mit t_{i+1} übereinstimmt. Solange es eine Übereinstimmung gibt, wird der Baum weiter traversiert. Existiert dagegen keine Kante, handelt es sich um einen Suffix, der noch nicht im Baum hinterlegt ist. Für den Suffix wird ein neues Blatt unterhalb des zuletzt besuchten Knotens erzeugt. Beide

[3]Gilt durch Lösung mittels dynamischer Programmierung.

Knoten werden durch eine Kante verbunden, welche den Präfix t_{i+h} repräsentiert. h entspricht dabei der Tiefe des Baumes.

Nachdem das erste Dokument eingefügt wurde, wird entsprechend mit den anderen Dokumenten fortgefahren. Ein Suffixbaum, der mehrere Dokumente enthält, wird verallgemeinerter Suffixbaum genannt. Hierbei wird an den Kanten zusätzlich die Anzahl der Dokumente gespeichert, die den Präfix enthalten.

In dem gezeigten Beispiel ist die längste gemeinsame Termfolge *ate*. Der Term kommt als einziges in allen drei Dokumenten vor. Es werden nicht immer aussagekräftige Phrasen ermittelt. Daher ist es sinnvoll, die für die Ermittlung von Kollokationen beschriebenen Heuristiken (siehe Seite 2.1.2) einzusetzen, um bedeutungslose Phrasen zu filtern.

Nominalphrasen

Die Computerlinguistik beschäftigt sich mit Methoden, um mit Hilfe des Computers die menschliche Sprache zu verstehen. Hierbei ist die Ermittlung von Nominalphrasen (engl. *noun phrase*, NP) aus Dokumenten eine wichtige Komponente. Eine Nominalphrase ist eine Phrase, deren Kopf (engl. *head*) ein Nomen oder Pronomen ist. Der Kopf bestimmt immer die Art eines Satzes, hier eine Nominalphrase. Nominalphrasen lassen sich durch folgenden regulären Ausdruck vereinfacht beschreiben:

<DT>?(<JJ>|<NN.?>)*<NN.?>

Eine Nominalphrase beginnt mit einem oder keinem Determinativ (engl. *determiner*, DT), gefolgt von keinem oder beliebig vielen Adjektiven (JJ) oder Nomen (engl. *noun*, NN). Der Kopf der Nominalphrase schließt diese ab. Es werden beliebige Substantive erlaubt (<NN.?>). Darunter fallen unter anderem die Pluralform von Nomen (NNS) und Eigennamen (engl. *proper noun*, NNP). Um Nominalphrasen in Sätzen zu ermitteln, sind zunächst die Wortarten zu ermitteln. Hierzu wird eine *Part-of-speech*-Analyse durchgeführt. Eine Ermittlung der Wortarten für das Dokument *D2* ergibt:

Software/NNP for/IN the/DT Sparse/JJ Singular/JJ Value/NN Decomposition/NN

Worte werden per Konvention durch einen Schrägstrich getrennt. Diesen nachgestellt ist die Wortart. Anhand des regulären Ausdrucks können Nominalphrasen eines Satzes ermittelt werden. Für das Beispiel gilt:

[Software/NNP**]** for/IN **[**the/DT Sparse/JJ Singular/JJ Value/NN Decomposition/NN**]**

Als Nominalphrasen ergeben sich *Software* und *the Sparse Singular Value Decomposition*. Durch die Determinanten erhält man keine zusätzliche Information über die Phrase, so dass in dieser Arbeit der Ausdruck

(<JJ>|<NN.?>)*<NN.?>

bereits ausreichend genau ist, um eine Nominalphrase zu beschreiben. Eine Nominalphrase wird durch das jeweilige Satzende begrenzt.

Der Vorteil von Nominalphrasen zur Ermittlung von Phrasen in Dokumenten ist, dass Phrasen beliebiger Länge ermittelt werden. Die Ermittlung von Wort N-Grammen ist dagegen auf eine feste Länge beschränkt. Nominalphrasen sind grammatikalisch korrekt und gelten als verständlich. Eine Filterung unbedeutender Phrasen, wie sie bei Wort N-Grammen notwendig ist, entfällt.

Allerdings ist der Berechnungsaufwand im Vergleich zur Ermittlung von N-Grammen deutlich höher, da zuerst Wortarten bestimmt werden müssen. Um Nominalphrasen ermitteln zu können, muss ein Verfahren Wissen über die Sprache des Textes besitzen. Ein Training des Verfahrens ist erforderlich. Dagegen ist die Ermittlung von N-Grammen auf beliebigen Texten direkt durchführbar.

Informativeness und Phraseness

Tomokiyo u. Hurst (2003) ermitteln Schlüsselphrasen in einer Dokumentkollektion. Sie fordern, dass eine Schlüsselphrase zwei Eigenschaften zu erfüllen hat: *phraseness* und *informativeness*. *Phraseness* bevorzugt längere Phrasen gegenüber kürzeren, da längere Phrasen als verständlicher gelten. Die zweite Eigenschaft besagt, dass eine Phrase informativ sein sollte. Informativ heißt für Tomokiyo u. Hurst: Je stärker eine Phrase eine Dokumentkollektion repräsentiert, desto größer ist der Informationsgewinn für den Nutzer. Informativeness dient der Abgrenzung gegenüber anderen Dokumentkollektionen.

Tomokiyo u. Hurst stellen Maße vor, um jede Phrase einer Dokumentkollektion auf Erfüllung der Eigenschaften hin zu bewerten. Je höher die Bewertung ausfällt, desto besser eignet sich diese als Schlüsselphrase.

Um Schlüsselphrasen anhand der aufgestellten Eigenschaften zu ermitteln, vergleichen Tomokiyo u. Hurst jeweils zwei Dokumentkollektionen miteinander. Die Dokumentkollektion, für die Schlüsselphrasen zu ermitteln sind, wird als Vordergrund-Korpus (fg)

bezeichnet. Mit Hilfe des Vordergrund-Korpus ist die *Phraseness* verifizierbar. Zur Bewertung der *Informativeness* ist zusätzlich eine zweite Dokumentkollektion erforderlich, die disjunkt zum Vordergrund-Korpus ist. Diese wird Hintergrund-Korpus (bg) genannt.

Für jeden Korpus sind Sprachmodelle (LM) zu definieren. Sprachmodelle im Information-Retrieval gehen auf Ponte u. Croft (1998) zurück. Diese wurden als Alternative beziehungsweise zur Verbesserung traditioneller Termgewichtungsverfahren wie *tf-idf* entwickelt. Ursprünglich definiert jedes Dokument ein Sprachmodell. Bei diesem handelt es sich um eine Wahrscheinlichkeitsverteilung der Terme eines Dokuments, wobei jeder Term mit einer bestimmten Wahrscheinlichkeit erzeugt wird.

In dieser Arbeit ist von Interesse, ob eine Phrase durch ein Dokument besonders gut repräsentiert wird. Die Phrase ist als Anfrage an ein Dokument aufzufassen. Die Relevanz eines Dokuments zu einer Anfrage ist durch die Wahrscheinlichkeit zu bestimmen, mit der die Anfrage vom Sprachmodell des jeweiligen Dokuments erzeugt wird.

Einem Unigramm-Sprachmodell liegt die Annahme zugrunde, dass Terme unabhängig voneinander auftreten. Die Wahrscheinlichkeit einer Termfolge ist deshalb gleich dem Produkt der Einzelwahrscheinlichkeiten:

$$P(t_n|d) = \frac{\mathrm{tf}_{t_{n,d}}}{\sum_{t' \in d} \mathrm{tf}_{t',d}}$$

N-Gramm Modelle höherer Ordnung beziehen dagegen den lokalen Kontext mit ein. Für Bigramm-Modelle wird beispielsweise der vorherige Term bei der Wahrscheinlichkeitsberechnung berücksichtigt. Es gilt allgemein:

$$P(t) = \prod_{i=1}^{n} P(t_i|t_1, t_2, \ldots, t_n)$$

Problematisch ist, dass, wenn ein Term der Phrase nicht durch das Sprachmodell modelliert wird, dieser die Auftrittswahrscheinlichkeit 0 besitzt. Es ist $P(t) = 0$. Zuvor nicht gesehenen Termen sind deshalb ebenso Wahrscheinlichkeiten größer Null zuzuweisen. Hierzu können *Smoothing*-Methoden eingesetzt werden. Eine einfache Methode nimmt von jedem Term an, dass dieser einmal zu viel gesehen wurde.

In dieser Arbeit wird der Vordergrund-Korpus durch das Cluster repräsentiert, für den ein Cluster-Label bestimmt werden soll. Für die Dokumente im Cluster wird ein gemeinsames Sprachmodell mit der Ordnung $N > 1$ erstellt. Jedes Sprachmodell der Ordnung $M < N$ würde zu einer schlechteren Modellierung der Dokumente des Clusters führen. Zur Bewertung der *Phraseness* wird nun der Verlust gemessen, der entstehen

würde, wenn die Dokumente nicht durch LM^N_{fg}, sondern durch ein Unigramm-Modell LM^1_{fg} modelliert würden. In dieser Arbeit ist $N = 3$.

Ein natürliches Maß, um den Verlust zwischen zwei Wahrscheinlichkeitsverteilungen zu messen, ist die Kullback-Leibler Divergenz. Diese ist definiert durch

$$D(p||q) = \sum_x p(x) \cdot \log \frac{p(x)}{q(x)}.$$

Tomokiyo u. Hurst (2003) verwenden den Term innerhalb der Summe, um sowohl Phraseness als auch Informativeness zu berechnen. Es gilt:

$$\delta_t(p||q) := p(t) \cdot \log \frac{p(t)}{q(t)}.$$

Die Formel beschreibt den Beitrag eines Terms t zum Verlust der gesamten Wahrscheinlichkeitsverteilung. Schließlich sind Phraseness und Informativeness zu berechnen:

Phraseness Die Phraseness eines Terms t sagt aus, wie viel Information verloren geht, wenn anstelle eines N-Gramm Modells ein Unigramm-Modell verwendet wird: $\varphi_p := \delta_t(LM^N_{\text{fg}}||LM^1_{\text{fg}})$.

Informativeness Die Informativeness eines Terms t sagt aus, wie viel Information verloren wird, wenn fälschlicherweise angenommen wird, der Term würde durch das Sprachmodell des Hintergrund-Korpus modelliert: $\varphi_i := \delta_t(LM^N_{\text{fg}}||LM^N_{\text{bg}})$. Es ist ebenfalls eine Berechnung mittels Unigramm-Modellen möglich: $\delta_t(LM^N_{\text{fg}}||LM^1_{\text{fg}})$.

Überwachte Verfahren

Bislang wurden Verfahren zur Schlüsselwortbestimmung betrachtet, die ohne Vorwissen über die Dokumente, aus denen Terme ermittelt werden sollen, anwendbar sind. Gegenwärtig existieren überwachte Verfahren, die zunächst auf einer Dokumentkollektion (Trainingsmenge) trainiert werden, um im Anschluss bessere Schlüsselworte bestimmen zu können. Hierbei sollte die Dokumentkollektion aus der gleichen Domäne stammen wie die Dokumente, in denen später Schlüsselworte ermittelt werden. Zu jedem Dokument der Trainingsmenge liegt dem Verfahren eine Liste von menschlich ausgewählten Schlüsselworten vor. Überwachte Verfahren sind gegenwärtig *State-of-the-art* im Bereich der Schlüsselwortbestimmung (Turney, 2000; Hulth, 2003).

KEA KEA ist ein solches überwachtes Verfahren. KEA trainiert mittels eines *Naive-Bayes*-Klassifikators ein Modell auf der Trainingsmenge. Der Klassifikator ordnet jede

ermittelte Phrase der Klasse zu, zu der diese mit der größten Wahrscheinlichkeit gehört. Es werden zwei Klassen unterschieden: *Schlüsselphrase* und *keine Schlüsselphrase*. Für das Training dienen die menschlich ausgewählten Schlüsselphrasen als positive Beispiele. Die nicht von Menschen erzeugten Phrasen sind Negativbeispiele. Eine Phrase wird bei KEA unter anderem durch folgende Merkmale beschrieben:

- TF-IDF,
- die Position des ersten Auftretens einer Phrase im Dokument,
- die Länge der Phrase in Worten.

Bei der Position des ersten Auftretens werden Phrasen bevorzugt, die sehr früh im Dokument auftreten. Diese stehen mit hoher Wahrscheinlichkeit entweder im Titel oder im ersten Absatz des Dokuments. KEA bevorzugt Bigramme, da diese im Unterschied zu einzelnen Termen als verständlicher gelten.

Es kann sowohl eine *freie* Ermittlung der Schlüsselphrasen vorgenommen werden als auch eine *kontrollierte*. Im Vergleich zur freien Ermittlung liegt im letztgenannten Fall ein Thesaurus vor. Dieser definiert das Vokabular, aus dem später Schlüsselphrasen gebildet werden. Liegt eine kontrollierte Situation vor, so werden nur Phrasen berücksichtigt, die im Thesaurus vorkommen. Tabelle 2.2 zeigt einen Ausschnitt aus einem aufbereiteten WordNet-Thesaurus. Dieser enthält Synonyme und eine Kurzbeschreibung des Begriffs *citation*. KEA ist in der Lage, Synonyme mittels des kontrollierten Vokabulars auf ein gemeinsames Konzept abzubilden.

MAUI Das Vokabular von WordNet, welches lexikalisch erfasst wird, liegt gegenwärtig bei 150.000 eindeutigen Worten (Fellbaum u. Miller, 2010). Das ist zu wenig, um als Thesaurus in KEA für beliebige Texte aus den verschiedensten Domänen verwendet zu werden. Basierend auf KEA erlaubt MAUI die Integration von Wikipedia als kontrolliertes Vokabular. Wikipedia existiert in zahlreichen Sprachen und deckt nahezu alle Wissensbereiche ab. Somit ist der Thesaurus von Wikipedia sprach- als auch domänenunabhängig. Er ist einem WordNet-Thesaurus vorzuziehen.

2.2 Repräsentation von Text

Die Ähnlichkeit zweier Dokumente soll bestimmt werden. Dieses ist mittels der vorliegenden Textdokumente nicht ohne Weiteres möglich. Dokumente sind in geeigneter

```
<skos:Concept rdf:about="[..]/synset-citation-noun-3">
  <rdfs:label>citation</rdfs:label>
  <skos:definition>
    (a short note recognizing a source of information or of
     a quoted passage; "the student's essay failed to list
     several important citations"; "the acknowledgments are
     usually printed at the front of a book"; "the article
     includes mention of similar clinical cases")
  </skos:definition>
  <skos:prefLabel>citation</skos:prefLabel>
  <skos:altLabel>acknowledgment</skos:altLabel>
  <skos:altLabel>credit</skos:altLabel>
  <skos:altLabel>mention</skos:altLabel>
  <skos:altLabel>quotation</skos:altLabel>
  <skos:altLabel>reference</skos:altLabel>
</skos:Concept>
```

Abbildung 2.2 *: Ausschnitt aus einem für KEA aufbereitenden Thesaurus. Dieser basiert auf WordNet 2.0. Es wird das Nomen „citation“ durch eine kurze Beschreibung definiert. Als bevorzugte Phrase zur Beschreibung dient ebenfalls „citation“ (prefLabel). Unter „altLabelßind Synonyme von „citationäufgelistet. Wird in KEA als Phrase „quotation“ aus dem Text ermittelt, so wird diese entsprechend auf „citation“ abgebildet.*

Weise durch eine Repräsentation am Computer zu modellieren. Es erfolgt zunächst ein Indexierungsschritt. Dieser abstrahiert eine Dokumentkollektion D durch Indexterme. Anschließend wird ein Dokumentmodell definiert, welches Ähnlichkeitsberechnungen für Dokumentrepräsentationen erlaubt.

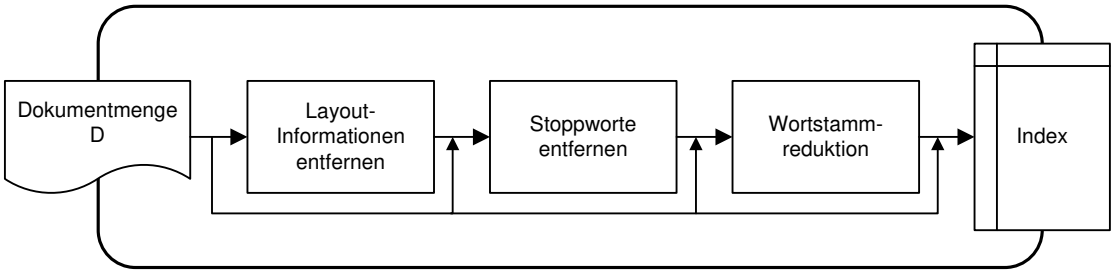

Abbildung 2.3 *: Indexierung einer Dokumentkollektion. Indexierte Terme werden in einem Index gespeichert.*

2.2.1 Indexierung

Die Terme eines Dokuments sind dessen Merkmale. Diese beschreiben die Semantik des Dokuments. Um effizient auf die Merkmale zugreifen zu können, sind die in Abbildung 2.3 aufgeführten Schritte erforderlich.

Strukturinformationen wie Satzzeichen werden aus Texten entfernt, da nur Terme für die Dokumentrepräsentation von Interesse sind. Anschließend wird eine linguistische Vorverarbeitung durchlaufen. Diese beinhaltet sowohl die Entfernung von Stoppworten (siehe Kapitel 2.1.1) als auch die Reduktion der Terme auf ihren Wortstamm (*Stemming*). Durch Stemming können Terme desselben Wortstamms im Index zusammengelegt werden. Dadurch hilft Stemming bei der Verbesserung der Qualität eines Indexierungsverfahrens. Der bekannteste Stemming-Algorithmus für die englische Sprache ist der *Porter*-Algorithmus (Porter, 1980). Zur Wortstammreduktion werden Beugungsformen anhand bestimmter Regeln entfernt. Ein Beispiel: *machines*, *machinery* und *machine* werden auf den gemeinsamen Stamm *machini* reduziert.

Bei der direkten Ableitung von Cluster-Labeln aus der Dokumentrepräsentation ist es schwierig, von dem auf Stammform reduzierten Term wieder auf den ursprünglichen zu schließen. Terme wie *machini* besitzen für den Menschen keine Bedeutung, so dass auf Wortstamm reduzierte Terme sich nicht als Cluster-Label eignen.

2.2.2 Dokumentmodelle

Sei eine Dokumentrepräsentation $\mathbf{d} \in \mathbf{D}$ über eine Abbildung $\alpha : D \rightarrow \mathbf{D}$ definiert. Diese umfasst den Schritt der Indexierung, gefolgt von einer Termgewichtung. Termgewichtungsmaße wurden bereits in Kapitel 2.1.1 in Zusammenhang mit der Schlüsselwortbestimmung diskutiert. Zwischen zwei Dokumentrepräsentationen $\mathbf{d}_i$ und $\mathbf{d}_j$ besteht eine Ähnlichkeitsbeziehung mit $\rho : \mathbf{D} \times \mathbf{D} \rightarrow \mathbb{R}$.

Dokumentrepräsentation und Relevanzfunktion werden vom verwendeten Dokumentmodell festgelegt Es werden zwei Dokumentmodelle vorgestellt, die für diese Arbeit relevant sind: Vektorraummodell und Latent-Semantic-Indexing.

Vektorraummodell

Das Vektorraummodell gehört zu den klassischen Modellen des Information-Retrieval (Salton, 1971). Terme eines Dokuments werden durch eine Dimension in einem L-dimensionalen Vektorraum repräsentiert. Dieser wird vom Vokabular $\mathcal{T}$ aufgespannt. Jedem

Term wird ein Termgewicht zugewiesen, so dass ein reales Dokument d_j durch einen Dokumentvektor $\mathbf{d}_j = \{w_{1,j}, \ldots, w_{L,j}\}$ vollständig beschrieben wird. Ein Dokument wird also durch einen Punkt im Vektorraum repräsentiert.

Da Textdokumente durch Vektoren repräsentiert werden, kann die Ähnlichkeit $\varphi(\mathbf{d}_i, \mathbf{d}_j)$ zweier Dokumentvektoren durch Vektor-Ähnlichkeitsmaße berechnet werden. Hierzu kann mit Hilfe des Skalarprodukts der Kosinus des Winkels zwischen den Vektoren als Maß verwendet werden. Als Ähnlichkeitsfunktion ergibt sich somit für zwei Dokumentrepräsentationen $\mathbf{d}_i, \mathbf{d}_j$:

$$\varphi(\mathbf{d}_i, \mathbf{d}_j) = \frac{\langle \mathbf{d}_i, \mathbf{d}_j \rangle}{|\mathbf{d}_i| \cdot |\mathbf{d}_j|}$$

Je geringer der Winkel zwischen zwei Repräsentationen ausfällt, desto ähnlicher sind sich die Dokumente.

Kritik Das Vektorraummodell trifft keine Annahmen über die Reihenfolge der Terme. Die Wichtigkeit eines Terms liefert somit keine Hinweise über die Wichtigkeit anderer Terme. Der lokale Kontext, indem Terme im Text standen, geht verloren.

Synonyme und Polyseme werden im Vektorraummodell nicht erfasst. Synonyme werden als eigene Dimension im Vektorraum repräsentiert. Dokumente der gleichen Domäne, die Synonyme enthalten, sind sich somit weniger ähnlich, obwohl sie vom gleichen Thema handeln. Im Unterschied dazu werden bei Polysemen die verschiedenen Bedeutungen eines Terms durch eine einzige Dimension im Vektorraum repräsentiert. Dokumente sind zueinander ähnlich, obwohl sie thematisch verschieden sind.

Latent-Semantic-Indexing (LSI)

Latent-Semantic-Indexing kann ebenfalls als Dokumentmodell eingesetzt werden. Kapitel 2.1.1 zeigte bereits, wie eine Dokumentkollektion durch LSI repräsentiert werden kann. Wie zwei Dokumente in diesem Modell auf Ähnlichkeit geprüft werden können, blieb offen. Sei $A = TSD^T$ die Singulärwertzerlegung der $L \times N$ Term-Dokument-Matrix A. Die Vektoren der Matrix D repräsentieren die Dokumentrepräsentationen $\mathbf{d} \in \mathcal{D}$. Durch Multiplikation von A mit ihrer Transponierten ergibt sich

$$AA^T = (TSD^T)\ (TSD^T)^T = TSD^T DST^T = TS^2T^T.$$

AA^T ist eine reguläre $L \times L$-Matrix. Diese wird aus den Termen des Vokabulars $\mathcal{T}$ der Dokumente gebildet. Die Elemente (i, j) der Matrix sind als Maß zu betrachten, welches

das gemeinsame Auftreten der beiden Terme, t_i und t_j, in Dokumenten bewertet. AA^T entspricht also einer Term-Term Ähnlichkeitsmatrix.

Je nach Termgewichtungsmaß variiert die Interpretation. Ist A eine Inzidenzmatrix (siehe Tabelle 2.2 auf Seite 9), so ist a_{ij} gleich der Anzahl der Dokumente, in denen sowohl t_i als auch t_j auftreten. Entsprechend repräsentiert $A^T A$ eine Dokument-Dokument Ähnlichkeitsmatrix.

Dokumente, die bereits im Repräsentationsraum durch einen Vektor repräsentiert sind, sind durch bekannte Vektor-Ähnlichkeitsmaße miteinander vergleichbar. Es sind dieselben Metriken wie beim Vektorraummodell anwendbar. Um dagegen die Ähnlichkeit von Dokumenten zu berechnen, von denen eines nicht durch A repräsentiert wird, muss dieses zunächst in den Repräsentationsraum projiziert werden. Entsprechend des gewählten Termgewichtungsmaßes wird für das nicht repräsentierte Dokument ein Dokumentvektor $\mathbf{q}$ über dem Vokabular $\mathcal{T}$ erzeugt. Dieser wird in den LSI-Raum mittels

$$\mathbf{q}_k = S_k^{-1} T_k^T \mathbf{q}$$

projiziert. $\mathbf{q}_k$ repräsentiert das zuvor nicht erfasste Dokument. Deerwester u. a. sprechen hierbei von einem Pseudodokument. Auf diese Weise können neue Dokumente dem Repräsentationsraum hinzugefügt werden.

Kritik Neben dem bekannten Vorteil der Auflösung von Synonymen und Polysemen besitzt Latent-Semantic-Indexing allerdings den Nachteil, dass durch die einmalige Singulärwertzerlegung der Repräsentationsraum und somit auch dessen Vokabular nicht erweiterbar sind. Es sind Pseudodokumente zu generieren, die durch die stetige Projektion in den Repräsentationsraum diesen verändern. Dessen Qualität nimmt stets ab, bis schließlich eine Neuberechnung erforderlich wird.

2.3 Clustering

Ein Clustering gruppiert Dokumente anhand ihrer Eigenschaften. Es ist eine Form der Klassifikation, bei der Dokumenten Kategorien (Cluster) zugewiesen werden. Cluster werden dabei ausschließlich anhand der vorliegenden Daten erzeugt.

Problemstellung Sei $\mathcal{D} = \{d_1, \ldots, d_N\}$ eine unstrukturierte Dokumentkollektion, k die gewünschte Anzahl an Clustern und ein Zielkriterium gegeben. Dieses bewertet die Qualität eines Clusterings. Als Zielkriterium kann eine Ähnlichkeitsfunktion[4] eingesetzt werden. Es ist eine Zuweisungsfunktion $\gamma : \mathcal{D} \to \{1, \ldots, K\}$ gesucht, die das Zielkriterium maximiert[5]. Sind keine leeren Cluster erwünscht, so ist zusätzlich $\forall c \in \mathcal{C}\ \exists d \in \mathcal{D} : \gamma(d) = c$ zu verlangen. Ein Clustering erzeugt für $\mathcal{D}$ durch γ eine Kategorisierung (Clustering) $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$. Jedes Cluster $c \in \mathcal{C}$ bildet eine Teilmenge von $\mathcal{D}$ mit $\bigcup_{i=1}^{K} c_i = \mathcal{D}$.

Ein Clustering kann durch eine $K \times N$-Matrix H repräsentiert werden. Es gilt $h_{ij} = 1$ gdw. $\gamma(d) = c$. In allen anderen Fällen ist $h_{ij} = 0$.

2.3.1 Arten des Clusterings

Es werden verschiedene Arten des Clusterings unterschieden: partitionierend oder hierarchisch, exklusiv oder überlappend, vollständig oder partiell und monothetisch oder polythetisch (Jain u. a., 1999).

Partitionierend oder hierarchisch Ein partitionierendes Clustering entspricht einer Kategorisierung von $\mathcal{D}$ in disjunkte Teilmengen. Es gilt $\forall_{d \in \mathcal{D}}\ \exists!_{c \in \mathcal{C}} : \gamma(d) = c$. Für die Elemente der Matrix H gilt: $\forall_j : \sum_i h_{ij} = 1$.

Werden Cluster in weitere Cluster unterteilt, so entsteht ein hierarchisches Clustering. Cluster werden in einer Baumstruktur organisiert. Jedes Cluster ist ein innerer Knoten im Baum und setzt sich aus der Vereinigungsmenge seiner Kinder zusammen. Die Wurzel des Baumes entspricht einem Cluster, welches alle Dokumente enthält. Die Blätter des Baumes halten Cluster, die nur aus einem einzigen Dokument bestehen (Singleton-Cluster).

Hierarchische Verfahren werden in agglomerative und divisive Verfahren unterteilt. Agglomerative Verfahren erzeugen ein hierarchisches Clustering *von unten nach oben*. Jedes Dokument ist zu Beginn des Clusterings einem eigenen Cluster zugewiesen. Schrittweise werden jeweils die zwei zueinander ähnlichsten Cluster vereinigt, bis ein Cluster resultiert, welches alle Dokumente enthält – die Wurzel.

Divisive Verfahren erzeugen das Clustering dagegen *von oben nach unten*. Zu Beginn des Clusterings liegt ein Cluster vor, welches alle Dokumente enthält. Das Cluster wird schrittweise geteilt, bis ausschließlich Singleton-Cluster existieren.

[4] Alternativ ist eine Distanzfunktion denkbar.

[5] Im Falle einer Distanzfunktion ist das Zielkriterium entsprechend zu minimieren.

Exklusiv oder überlappend Werden Dokumente eindeutig einem einzigen Cluster zugewiesen, wird von einem exklusiven Clustering gesprochen. Ist dies nicht der Fall, so ist das Clustering überlappend. Es gilt $\forall_{d \in \mathcal{D}} \; \exists_{c \in \mathcal{C}} \; : \; \gamma(d) = c$ und $\forall_j \; : \; \sum_i h_{ij} \geq 1$.

Vollständig oder partiell Ein vollständiges Clustering weist jedem Dokument ein Cluster zu. Im Vergleich dazu werden bei einem partiellen Clustering nicht alle Dokumente einem Cluster zugewiesen, so dass $\gamma : \mathcal{D} \rightarrow_p \{1, \ldots, K\}$ gilt.

Monothetisch oder Polythetisch Findet die Zuweisung eines Dokuments zu einem Cluster anhand eines einzigen Merkmals statt, wird von einem monothetischen Clustering gesprochen. Werden dagegen mehrere Merkmale bei der Zuweisung gleichzeitig berücksichtigt, handelt es sich um ein polythetisches Clustering.

2.3.2 Clustering-Verfahren

Folgende Clustering-Verfahren, die in dieser Arbeit zum Cluster-Labeling verwendet werden, werden vorgestellt: k-Means und Suffixbaum-Clustering.

k-Means

k-Means ist ein partitionierendes, prototyp-basiertes Clustering-Verfahren, welches k Cluster durch ihre jeweiligen Prototypen repräsentiert. Der Prototyp eines Clusters ist dessen Centroid $\mathbf{c}$. Der Centroid des i-ten Clusters c_i ist dessen Mittelwert:

$$\mathbf{c}_i = \frac{1}{|c_i|} \sum_{\mathbf{d} \in c_i} \mathbf{d}.$$

Ein Cluster eines prototyp-basierten Clusterings besteht aus einer Menge von Dokumenten, wobei jedes ähnlicher zum Prototypen des zugewiesenen Clusters ist als zu Prototypen aller anderen Cluster. Alternativ zum Centroiden ist auch der Medoid eines Clusters als Prototyp zu verwenden. Der Medoid eines Clusters ist das Dokument, welches das Cluster am *besten* repräsentiert. Ein Medoid ist immer ein *echtes* Dokument.

Ein ideales prototyp-basiertes Cluster ist kugelförmig – mit dem Centroiden beziehungsweise Medoiden als Schwerpunkt.

k-Means wird initialisiert, indem eine benutzerdefinierte Konstante k bestimmt wird. Diese legt die Anzahl der zu erzeugenden Cluster fest. Nachfolgend wird ein iterativer Prozess durchlaufen, der ein zu wählendes Zielkriterium optimiert:

1. Bestimmung von k Dokumenten aus $\mathcal{D}$, die sich hinreichend voneinander unterscheiden. Diese bilden die Centroiden $\{\mathbf{c}_i | 1 \dots K\}$.
2. Die übrigen Dokumente werden jeweils dem ähnlichsten Centroiden zugewiesen (Zuweisungungsschritt).
3. Die Centroiden der neu zusammengesetzten Cluster werden ermittelt und ersetzen die zuvor verwendeten Centroiden aus Punkt 1 (Aktualisierungsschritt).
4. Wiederholung der Schritte 3 und 4, bis eine definierte Abbruchbedingung erreicht ist.

Der Aktualisierungsschritt optimiert ein Zielkriterium, welches vom gewählten Ähnlichkeitsmaß abhängt. Als Ähnlichkeitsmaß kann die Kosinusähnlichkeit verwendet werden. Hierbei ist die Summe der Kosinusähnlichkeiten der Dokumente zum Cluster-Centroiden zu maximieren. Für das Zielkriterium im Falle der Kosinusähnlichkeit gilt:

$$\sum_{i=1}^{K} \sum_{\mathbf{d} \in c_i} \varphi_{\cos}(\mathbf{c}_i, \mathbf{d}) \rightarrow \max$$

Nach dem Aktualisierungsschritt wird durch eine Abbruchbedingung überprüft, ob die Qualität des Clusterings ausreichend ist. Für prototyp-basierte Clustering-Verfahren sind verschiedene Abbruchbedingungen möglich. Da k-Means immer gegen eine Lösung konvergiert, wird ein Zustand erreicht, indem keine Dokumente mehr ihren Cluster-Centroiden wechseln. Die Centroiden ändern sich nicht mehr. Der Clustering-Prozess kann abgebrochen werden.

Kritik Bei k-Means hat die Auswahl der initialen Centroiden Einfluss auf den Clustering-Prozess. Es ist nicht garantiert, dass das Zielkriterium global maximiert wird. Es ist deshalb eine schlechte Wahl, die initialen Centroiden per Zufall aus der bestehenden Dokumentkollektion zu bestimmen. Besser ist es, die initialen Centroiden sequentiell auszuwählen, so dass jeder möglichst weit von den anderen Centroiden entfernt liegt.

k-Means erzeugt die natürliche Cluster-Anzahl nicht selbstständig. Diese wird durch den Parameter k einmalig festgelegt und ist nicht variabel. Es sind verschiedene Durchläufe des Algorithmus mit unterschiedlichen Werten für k erforderlich, um das Clustering zu bestimmen, welches das Zielkriterium maximiert. Die Laufzeit von k-Means beträgt $\mathcal{O}(KN)$ mit N gleich der Anzahl der Dokumente.

Suffixbaum-Clustering

Suffixbaum-Clustering wird erstmals in Zamir u. Etzioni (1998) vorgestellt. Die Struktur eines Suffixbaumes wird ausgenutzt, um Dokumente, die eine gemeinsame Termfolge enthalten, unter einem Knoten im Baum zu gruppieren. Wird später ein Knoten im Baum ausgewählt, ist sichergestellt, dass alle Dokumente unterhalb des Knotens die Termfolge des ausgewählten Knotens enthalten.

Die Menge der gemeinsamen Termfolgen wird verwendet, um initial Basiscluster $b \in B$ zu erzeugen. Basiscluster werden für Suffixe gebildet, die in mindestens zwei Dokumenten vorkommen. Jedes Basiscluster ist mit dieser Dokumentmenge assoziiert. Hierbei handelt es sich um einen monothetischen Clustering-Schritt, da Basiscluster ausschließlich anhand eines Merkmals, dem Suffix, gebildet werden.

Jedem Basiscluster wird ein Gewicht zugewiesen. Diese ist abhängig von der Länge des Suffixes und von der Anzahl der assoziierten Dokumente. Suffixe, die aus mehreren Termen bestehen und in möglichst vielen Dokumenten vorkommen, sollen bevorzugt werden. Zamir u. Etzioni (1998) gewichten deshalb ein Basiscluster mittels der Funktion $\text{score}(b) = |b| \cdot f(s)$. Die Funktion f wertet einzelne Terme ab, ist linear für Phrasen bestehend aus zwei bis sechs Worten und wird konstant für Phrasen mit mehr als sechs Worten. Eine Realisierung von f geben Zamir u. Etzioni nicht an, so dass folgende Gewichtung von Weiss (2006) übernommen wird:

$$\text{score}(b) = |b| \cdot \exp \frac{-(|s| - m)^2}{2 * d^2} \tag{2.3}$$

mit $m = 4$, $d = 8$. $|b|$ beschreibt die Anzahl der Dokumente, die den Suffix teilen und $|s|$ die Länge des Suffixes. Bei der Exponentialfunktion handelt sich hierbei um eine glockenartige Kurve. Mit m ist die optimale Länge der Suffixe in Worten anzugeben. Je größer d gewählt wird, desto stärker ist die Abwertung längerer Phrasen.

Die k Basiscluster mit den höchsten Gewichten werden für den nächsten Clustering-Schritt ausgewählt. Es kann vorkommen, dass mehrere der ausgewählten Basiscluster dieselbe Dokumentkollektion abdecken. In einem Folgeschritt sind diese Basiscluster zu vereinigen. Hierzu werden alle Basiscluster durch Knoten in einem ungerichteten Graphen repräsentiert. Zwei Basiscluster b_1 und b_2 werden durch eine Kante miteinander verbunden, wenn sie ähnlich zueinander sind. Die Ähnlichkeit zweier Basiscluster wird über ein binäres Ähnlichkeitsmaß $\varphi : B \times B \to \{0, 1\}$ mit

$$\varphi(b_i, b_j) = \begin{cases} 1, & \text{wenn } \frac{|b_i \cap b_j|}{|b_i \cup b_j|} \geq 0,5 \\ 0, & \text{sonst} \end{cases}$$

definiert. Es findet ein polythetisches Clustering der Basiscluster statt. Die resultierenden Cluster werden erneut anhand der Gewichte ihrer Basiscluster und ihrer Dokumentüberlappung gewichtet. Die k besten Cluster bilden das finale Clustering.

Kritik Suffixbaum-Clustering besitzt Vorteile gegenüber k-Means. Die Laufzeit ist linear abhängig von der Anzahl der Dokumente. Aufwändige Ähnlichkeitsberechnungen im Vektorraum entfallen im Vergleich zu k-Means aufgrund des einfachen, binären Ähnlichkeitsmaßes.

Das Clustering ist infolge des Zusammenfassungsschritts der Basiscluster überlappend. Dokumente werden jeweils zu einem Basiscluster vereinigt, wenn diese bereits einen einzigen Suffix gemein haben. Vor allem bei längeren Dokumenten wird die Übereinstimmung mit anderen Suffixen nicht weiter berücksichtigt. In der Folge sind Cluster schlechter Qualität zu erwarten (Meyer zu Eißen, 2007). Suffixbaum-Clustering eignet sich daher eher für kurze Dokumente (Zamir u. Etzioni, 1998; Stefanowski u. Weiss, 2003).

Bereits während des Clusterings werden Cluster-Label, repräsentiert durch die Suffixe der Basiscluster, implizit erzeugt. Das Suffixbaum-Clustering besitzt somit einen Vorteil gegenüber anderen Clustering-Verfahren, bei denen dies nicht zutrifft.

3 Was sind verständliche Cluster-Label?

Cluster-Labeling-Verfahren erzeugen automatisch Cluster-Label für einzelne Cluster. Von Menschen ausgewählte Cluster-Label wie Kategorienamen im *Open Directory Project* oder in *Wikipedia* werden als ideal angesehen (Stein u. Meyer zu Eißen, 2004; Treeratpituk u. Callan, 2006; Weiss, 2006). Diese beschreiben eine Menge von Texten und gelten als verständlich und informativ. Ziel bei der Erstellung von Cluster-Labeln muss es demnach sein, von menschlichen Experten gewählte Label möglichst gut nachzubilden. Ein Nutzer soll keinen Unterschied zwischen einem menschlich ausgewählten und einem erzeugten Label feststellen können.

Zu diesem Zweck wird die Struktur der Kategorien vom Open Directory Project und von Wikipedia untersucht. Motivation ist, Rückschlüsse über automatisch erzeugte Label zu ziehen. Zunächst werden beide Projekte sowie die zugrundeliegende Datenbasis vorgestellt, bevor deren Auswertungen gegenübergestellt und diskutiert werden.

Open Directory Project

Das Open Directory Project (ODP) gruppiert Internetseiten durch Unterstützung freiwilliger Editoren. Dem Nutzer wird eine zu durchsuchende Hierarchie von Kategorien geboten. Von der Wurzel ausgehend werden diese mit zunehmender Verzweigung der Hierarchie spezifischer. Eine initiale Anfrage hilft dem Nutzer bei der Suche nach relevanten Internetseiten. Im Anschluss ist die Suche nur durch manuelles Durchsuchen der gefundenen Kategorien zu spezialisieren. Aus diesem Grund versteht sich das Projekt vielmehr als Internetverzeichnis und weniger als Suchmaschine.

Neue Internetseiten zur Aufnahme in das Verzeichnis werden von Nutzern vorgeschlagen. Nach einer Begutachtung durch Editoren werden diese in die entsprechende Kategorie aufgenommen. Bis Ende 2009 wurden mehr als zwei Millionen Internetseiten in über 750.000 Kategorien klassifiziert. Insgesamt sind 82 verschiedene Sprachen vertreten. Englisch und Deutsch nehmen dabei den größten Anteil ein. Siehe hierzu Abbildung 3.1.

Bei der Aufnahme neuer Internetseiten regeln Richtlinien, dass diesen möglichst nur eine einzige Kategorie zugewiesen wird (DMOZ, 2004). Kategorien selbst können nur von

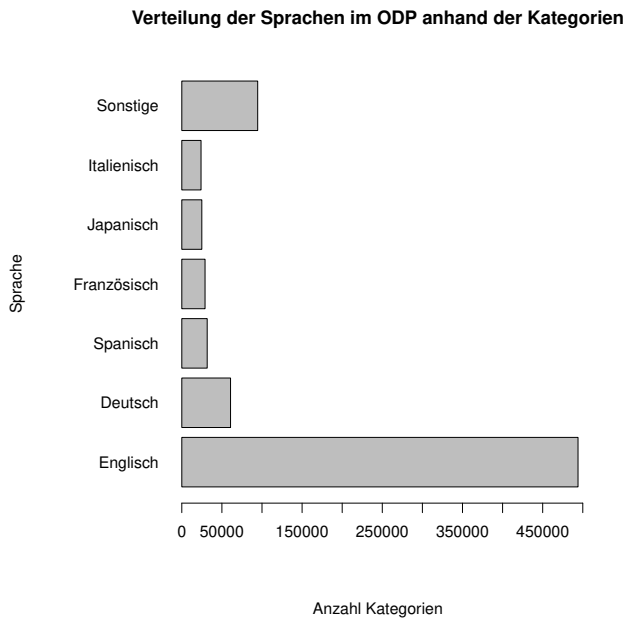

Abbildung 3.1 *: Verteilung der Sprachen im Open Directory Project anhand der Kategorien. „Sonstige“ umfasst 76 Sprachen.*

Editoren angelegt, verändert oder gelöscht werden. Editoren sind angehalten, neue Unterkategorien nur zu erzeugen, wenn in einer Kategorie mindestens 20 Internetseiten gelistet sind. Neue Unterkategorien sollen jeweils mindestens fünf Internetseiten umfassen.

Beinahe jede Internetseite ist nur einer einzigen Kategorie zugewiesen (siehe Abbildung 3.6 auf Seite 37). Somit wird sichergestellt, dass die Hierarchie bis auf wenige Ausnahmen einer Taxonomie entspricht.

Die Auswertung von ODP basiert ausschließlich auf englischen Kategorienamen. Die zugrundeliegende Datenbasis liegt im *RDF*-Format vor und ist vom 30. Dezember 2009[1]. Für diese Arbeit liegen 494.071 englische Kategorien von insgesamt 773.855 Kategorien der ODP zur Auswertung vor.

Wikipedia

Im März 2000 wurde das Vorläuferprojekt von Wikipedia, Nupedia, durch Jimmy D. Wales und Lawrence M. Sanger mit dem Ziel gegründet, die weltweit größte und freie

[1] Ein aktueller Auszug ist unter `http://rdf.dmoz.org/rdf/content.rdf.u8.gz` zu beziehen. ODP stellt im Gegensatz zu Wikipedia jeweils nur den aktuellsten Auszug bereit.

Enzyklopädie im Internet aufzubauen. Nupedia scheiterte allerdings aufgrund des restriktiven, nicht öffentlichen redaktionellen Prozesses, wobei nur registrierte Autoren unter der Aufsicht von Experten Artikel verfassen durften (Nupedia, 2003). Dies ist mit der Aufnahme neuer Internetseiten beim Open Directory Project zu vergleichen, wonach Vorschläge zunächst überprüft werden.

Aufgrund der langsamen Entwicklung von Nupedia forcierte Sanger als Konsequenz im Januar 2001 das Schwesterprojekt Wikipedia. Mit dem gegenteiligen Prinzip – jeder Nutzer darf frei Artikel bearbeiten – sollte der langwierige Zulassungsprozess von Artikeln umgangen und die Entstehung neuer Artikel beschleunigt werden. Anders als Nupedia entwickelte sich Wikipedia in derselben Zeitspanne rasant. Im Januar 2010 existierten Wikipedias in 220 Sprachen mit insgesamt mehr als 21,5 Millionen Artikel. Alleine die vier größten Sprachen – Englisch, Deutsch, Französisch und Polnisch – stellten zusammen mehr als 5,7 Millionen Artikel (Wikipedia, 2010a).

Zur Organisation einzelner Artikel werden Autoren darauf hingewiesen, diese zu mindestens einer Kategorie hinzuzufügen. Zudem sollen Nutzer Kategorienamen ebenso wie beim ODP anhand bestimmter Richtlinien erstellen (Wikipedia, 2010b). Namenskonventionen für Kategorien beinhalten beispielsweise die bevorzugte Bezeichnung von Kategorien, bei denen Institutionen nach Ländern gruppiert sind. Ein Beispiel: Die Kategorie *Fluggesellschaften nach Ländern* soll nur Unterkategorien der Form *Fluggesellschaften in Deutschland* besitzen. Da bei Wikipedia ausschließlich Empfehlungen gegeben werden, häufen sich im Gegensatz zum Open Directory Project Abweichungen. In der Folge sind mehr als die Hälfte aller Artikel in der englischen Wikipedia kategorienlos. Siehe hierzu Abbildung 3.7 auf Seite 38.

Kategorien in Wikipedia bilden im Vergleich zu ODP keine Taxonomie, weil Artikel mehreren Kategorien zugeordnet werden können. Es entsteht ein Kategoriengraph. Um Wikipedia gleichwohl als Taxonomie zu nutzen, existieren Verfahren, um aus dem Graphen einen Baum zu gewinnen (Ponzetto u. Strube, 2007; Kassner u. a., 2008).

Die Auswertung von Wikipedia basiert in dieser Arbeit ausschließlich auf der englischen Version. Wikipedia stellt in unregelmäßigen Abständen die gesamte Datenbasis zur freien Benutzung bereit. Die Kategorien liegen im *SQL*-Format vor[2]. Stand ist der 3. November 2009. Im Vergleich zu ODP sind Kategorien enthalten, die der Organisation von

[2]SQL ist eine Datenbanksprache. Der SQL-Auszug der Wikipedia ist unter `http://download.wikimedia.org/enwiki/20091103/enwiki-20091103-category.sql.gz` zu beziehen. Hierarchische Informationen liegen nicht vor. Letzter Zugriff am 9. April 2010.

Wikipedia selbst dienen. Dazu zählen alle Kategorien, die *wikipedia*, *wikiprojects*, *lists*, *mediawiki*, *templates*, *users*, *portal*, *categories*, *articles* oder *pages* im Namen enthalten. Diese Kategorien beschreiben Strukturen und keine Inhalte. Diese sind in dieser Arbeit nicht von Interesse. Dies entspricht der Vorgehensweise von Ponzetto u. Strube (2007).

Für diese Arbeit liegen 744.339 Kategorien von insgesamt 865.215 Kategorien von Wikipedia zur Auswertung vor.

Analyse von Kategorienamen in ODP und Wikipedia

Im Folgenden wird von Dokumenten gesprochen, wenn sowohl Internetseiten aus ODP und Wikipedia-Artikel gemeint sind. Für die Auswertung sind folgende Fragestellungen von besonderer Bedeutung:

1. Aus wie vielen Worten bestehen Kategorienamen im Durchschnitt?
2. Welche Wortarten werden vorzugsweise als Kategoriename verwendet?
3. Wie vielen Kategorien ist ein Dokument im Durchschnitt zugewiesen?
4. Wie tief ist die Hierarchie? Wie ist die Verteilung der Dokumente pro Hierarchietiefe?

Aus wie vielen Worten bestehen Kategorienamen im Durchschnitt? Hypothese ist, dass ein Label mindestens aus zwei Worten bestehen sollte: Ein Substantiv, welches die Art der Dokumente beziehungsweise deren Inhalt beschreibt und einem zweiten Term, der als Spezialisierung dient. Ein Beispiel: *Blaue Kugeln* ist aussagekräftiger als *Kugeln* und grenzt sich ab von *rote Kugeln*.

Die Verteilung der Worte pro Kategorie im Open Directory Project, siehe Tabelle 3.1, zeigt, dass durchschnittlich ein Kategoriename aus nahezu zwei Worten besteht. Dies stützt die aufgestellte Hypothese. Gleichwohl setzt sich die Hälfte aller Kategorienamen nur aus einem einzigen Wort zusammen. *Essen* ist beispielsweise mehrdeutig. Es eignet sich daher nicht als Label. Warum werden dennoch Kategorienamen aus einzelnen Worten gebildet? Es wird argumentiert, dass in einer strikten Hierarchie alleinstehende Worte bereits aussagekräftig genug sind, wenn diese im Kontext betrachtet werden. Dies ist beim Open Directory Project der Fall. Ein Nutzer springt weniger per Suchanfrage direkt an eine Zielkategorie, sondern navigiert vielmehr von der Wurzel beginnend durch den Hierarchiebaum. Dabei durchläuft er verschiedenste Kategorien, die mit der Tiefe spezieller werden. Die Zielkategorie wird durch einen eindeutigen Pfad beschrieben. Alle

Anzahl Worte	1	2	3	4	5	6	7	≤ 19
Kategorien	232.756	103.566	125.634	11.544	2.888	1.389	589	414
Häufigkeit	48,6%	21,6%	26,3%	2,4%	0,6%	0,3%	0,1%	0,1%

Tabelle 3.1 *: Zusammensetzung von Kategorienamen anhand der Anzahl verwendeter Worte im Open Directory Project. Die maximale Anzahl an Worten pro Kategorie beträgt 19. Die dazugehörige Kategorie ist ein Filmtitel.*

Kategorien entlang des Pfades werden als Kontext aufgefasst. Daher ist der Begriff *Essen* alleinstehend als Kategoriename zwar mehrdeutig, wird jedoch durch den Pfad wieder eindeutig: *Regional: Europa: Deutschland: Nordrhein-Westfalen: Städte und Gemeinden: E: Essen.*

Anders verhält es sich bei Wikipedia. Hier werden Kategorienamen im Durchschnitt aus 3,8 Worten gebildet (siehe Tabelle 3.2). Kategorien von Wikipedia stehen im Vergleich zu ODP nicht im Vordergrund, so dass die Hierarchie nicht so strikt ist. Kategorien stehen bei Wikipedia nicht im Kontext, sondern müssen bereits alleinstehend für den Nutzer aussagekräftig sein. Jeder Kategoriename in Wikipedia ist daher eindeutig. Werden mehrdeutige Begriffe von Nutzern gesucht, helfen Begriffsklärungsseiten bei der Auflösung dieser. Wikipedia enthält zahlreiche Artikel über Personen, Ereignisse, Themen der Wissenschaft und Kultur. Viele Begriffe lassen sich nicht durch zwei Worte fassen, so dass die durchschnittliche Anzahl der Worte pro Kategoriename deutlich höher ausfällt.

Aus den Beobachtungen ist zu schließen, dass Label bestehend aus einem Wort nur eindeutig und aussagekräftig sind, wenn hierarchische Strukturen vorliegen und dem Nutzer das entsprechende Label im Kontext präsentiert wird. Cluster-Label dieser Form eignen sich somit zur Bezeichnung von Clustern in einem hierarchischen Clustering.

Anzahl Worte	1	2	3	4	5	6	7	≤ 61
Kategorien	42.182	149.884	203.873	152.433	95.696	50.317	26.251	25.987
Häufigkeit	5,7%	20,1%	27,3%	20,4%	12,8%	6,7%	3,5%	3,5%

Tabelle 3.2 *: Zusammensetzung von Kategorienamen anhand der Anzahl verwendeter Worte in Wikipedia. Kategorien mit einer Länge von mehr als 20 Worten sind nahezu vollständig von Vandalismus betroffen. Ein Beispiel: „Roman Catholic secondary schools THIS SCHOOOL IS NOT VERY GOOD AND I WOULD NOT ADVISE GOING TO IT in England“.*

Tag	Beschreibung	Tag	Beschreibung
CC	Konjunktion	NNS	Nomen (Plural)
IN	Präposition	NNP	Eigenname (Singular)
JJ	Adjektiv	NNPS	Eigenname (Plural)
NN	Nomen (Singular)	VBG	Verbalsubstantiv

Tabelle 3.3 *: Die im Zusammenhang mit einer Part-of-speech Analyse ermittelten Wortarten, die in Kategorien des Open Directory Projects und Wikipedia vorkommen. Die Abkürzungen der Wortarten – im Englischen als „tag“ bezeichnet – entsprechen der „Penn Treebank“-Notation (Santorini, 1990).*

Im Vergleich dazu sollten bei flachen, partitionierenden Clustering-Verfahren eher längere Label verwendet werden. Eine Beschränkung der Länge eines Labels ist schwierig anzugeben, da unter anderem auch Filmtitel wie *Der Herr der Ringe - Die Schlacht um Mittelerde* vollständig erfasst werden sollten. Es ist sinnvoll, die maximale Länge eines Labels durch das Satzende zu begrenzen. Dieses gewährleistet die grammatikalische Korrektheit.

Welche Wortarten werden vorzugsweise als Kategoriename verwendet? Weiterhin ist von Interesse, aus welchen Wortarten ein Kategoriename zusammengesetzt wird. Label, die aus denselben Wortarten wie im ODP oder in Wikipedia zusammengesetzt werden, sind gegenüber seltenen Kombinationen von Wortarten zu bevorzugen. Annahme ist, dass diese menschlich gewählten Kategorien stärker entsprechen und demnach bedeutsamer sind.

Wortarten werden mit Hilfe einer *Part-of-speech*-Analyse bestimmt. Hierbei handelt es sich um ein Verfahren der Computerlinguistik. Für einzelne Worte im Satz wird versucht, die korrekte Wortart zu ermitteln. Die Erkennung der Wortart für alleinstehende Worte und kurze Phrasen von Kategorien ist Grund, dass Verfahren den Kontext, in dem das jeweilige Wort steht, zur Bestimmung der Wortart benötigen. Dieser liegt bei den hier untersuchten Kategorienamen allerdings nicht vor. Dennoch liefert die Analyse Hinweise über die häufige Verwendung bestimmter Wortarten. Vorkommende Wortarten in Kategorien des Open Directory Projects und Wikipedia sind in Tabelle 3.3 zusammengefasst.

Die Ermittlung von Wortarten für Kategorien aus ODP zeigt, dass vorzugsweise Nomen zur Bildung von Kategorienamen verwendet werden (siehe Tabelle 3.4). Dies verwundert nicht, da diese zumeist aus nur einem einzigen Wort bestehen. Unter anderem lassen sich

Eigennamen wie Personen- oder Ortsnamen unter den Kategorien finden. Interessant ist die Verwendung von Verbalsubstantiven wie *Racing*. Neben einzelnen Worten treten besonders häufig Wort-Trigramme auf. Es werden dabei zwei Nomen mittels einer Konjunktion verknüpft. Dies deutet darauf hin, dass unterhalb einer solchen Kategorie Internetseiten gruppiert sind, die zwei Aspekte eines gemeinsamen Themas behandeln, beispielsweise *Essen und Trinken*.

Wortarten in Wikipedia sind in Tabelle 3.5 auf Seite 37 aufgeführt. Kategorien bestehend aus einzelnen Worten sind dabei nur gering vertreten. Im Unterschied zu ODP werden Kategorien, die aus Wort-Trigrammen gebildet werden, nicht mehr durch eine Konjunktion erzeugt, sondern mit Hilfe von Präpositionen wie *von* und *in*.

Aus den Beobachtungen ist zu schließen, dass für Label, die nur aus einem Wort gebildet werden, sich folgende Wortarten eignen: Nomen, Eigennamen und Verbalsubstantive. Da immer eine Dokumentkollektion beschrieben wird, sind Pluralformen der Wortarten vorzuziehen. Für Label bestehend aus mehreren Worten gelten ähnliche Anforderungen. Konjunktionen tragen dazu bei, zwei miteinander verwandte Themen innerhalb eines Clusters zu beschreiben. Diese Art der Bezeichnung eignet sich insbesondere, wenn ein Clustering überlappende Cluster erzeugt.

Es ist möglich, eine formale Sprache für Kombinationen von Wortarten zu definieren, die zur Bildung von Kategorienamen beim Open Directory Project und Wikipedia verwendet werden. Ist ein Label Teil dieser Sprache, so ist es wahrscheinlicher, dass es für den Menschen verständlich ist.

Wie vielen Kategorien ist ein Dokument im Durchschnitt zugewiesen? Es wird untersucht, ob ein Dokument nur einer einzigen oder mehreren Kategorien zugewiesen wird. Für ein Suchszenario besitzt die Zuweisung mehrerer Kategorien zu einem Dokument Vorteile. Nutzern wird es ermöglicht, ein und dasselbe Dokument über mehrere Pfade in einer Hierarchie zu finden. Sucht ein Nutzer einen Fahrradhändler in Weimar, führen sowohl der Pfad *Sport: Radsport: Händler: Weimar* und *Städte und Gemeinden: Weimar: Händler: Radsport* zu denselben Ergebnissen. Der Händler muss hierzu in beiden Kategorien eingetragen sein. Es wird die Hypothese aufgestellt, dass die Zuweisung eines Dokuments zu mehreren Kategorien vorteilhaft für das Cluster-Labeling ist. Wir sind hier konform mit den Ausführungen von Sacco (2000) aus dem Bereich *Faceted Search* und Krishnapuram u. Kummamuru (2003) bei der Erstellung von Taxonomien. Sacco spricht hier von dynamischen und letztgenannte Autoren von unscharfen Taxonomien.

Wort N-Gramm	absolute Häufigkeit	Beispiel
NNP	87.257	Scotland
NN	62.227	Country
NNS	42.224	Reviews
VBG	16.745	Racing
NNP NNP	55.173	Andy Warhol
NNP NNPS	14.885	Personal Pages
NNP CC NNP	26.439	Society and Culture
NN CC NN	25.652	Import and Export
NN CC NNPS	21.875	Travel and Tourism
NNS CC NNS	14.813	Articles and Interviews

Tabelle 3.4 *: Die häufigsten vorkommenden Abfolgen von Wortarten in Kategorien des Open Directory Projects. Diese sind unterteilt in Wort-Uni-, Wort-Bi- und Wort-Trigramme. Die Auflistung zeigt nur einen Ausschnitt.*

Beide Forschungsbereiche sind mit dem Cluster-Labeling verwandt, so dass sich deren Ergebnisse übertragen lassen.

Beim Open Directory Project ist in nur 7% der Fälle eine Internetseite mehreren Kategorien zugeordnet. Siehe hierzu Tabelle 3.6. Dies erklärt sich durch den geregelten Aufnahmeprozess neuer Internetseiten, bei dem nur eine einzige Kategorie vom Nutzer vorgeschlagen werden kann. Dies garantiert, dass eine Internetseite mindestens einer Kategorie angehört.

Im Gegensatz zum Open Directory Project werden Wikipedia-Artikel in mehr als 57,9% der Fälle keiner einzigen Kategorie zugeordnet (siehe Tabelle 3.7). Wird diese Menge aus der Berechnung herausgenommen, so werden nur 26% der Wikipedia-Artikel einer einzigen Kategorie zugewiesen. Die Mehrheit der Artikel wird zwei bis vier Kategorien zugewiesen. Dieser Umstand wird damit begründet, dass jeder Nutzer die Freiheit besitzt, dem Artikel eine Kategorie zuzuweisen. Da an einem Artikel zahlreiche Nutzer beteiligt sind, werden Artikel häufig mehreren Kategorien zugeordnet. 9% der Artikel sind sogar mehr als 8 Kategorien zugewiesen.

Rückschlüsse auf das Cluster-Labeling sind zu ziehen. Das Clustering kann überlappend sein. Wikipedia zeigt, dass die Zuweisung von Kategorien zu Artikeln eine Aufgabe ist, bei der viele Menschen unterschiedlicher Meinung sind. Deshalb ist es sinnvoll,

Wort N-Gramm	absolute Häufigkeit	Beispiel
NNP	10.892	Chernobyl
NNS	10.024	Mermaids
NNP NNS	41.056	Olympic mascots
NNP NNP	22.789	Maria Sharapova
JJ NNS	20.734	Extraterrestrial supervillians
NNP NN	18.461	Louisiana society
CD NNS	15.032	1983 borns
NNP NNP NNS	27.717	James Bond films
NNS IN NNP	25.772	Airlines of Ireland
NNP IN NNP	18.154	Zoos in Mississippi
JJ NN NNS	16.752	British squash players
NNP NNP NNP NNS	16.795	Andrew Lloyd Webber songs
NNP NNS IN NNP	14.599	Wildlife sanctuaries of India
NNS IN NNP NNP	14.079	Borders of New York

Tabelle 3.5 *: Die häufigsten vorkommenden Abfolgen von Wortarten in Kategorien von Wikipedia. Diese sind unterteilt in Wort-Uni-, Wort-Bi-, Wort-Tri- und Wort-Quad-Gramme. Die Auflistung zeigt nur einen Ausschnitt.*

das Cluster-Labeling an die jeweiligen Nutzerbedürfnisse anzupassen. Dies verlangt semi-überwachte Verfahren. Gegenwärtig existieren nach bestem Wissen keine Verfahren dieser Art für das Cluster-Labeling.

Wie tief ist die Hierarchie? Wie ist die Verteilung der Dokumente pro Hierarchietiefe? Neben den Fragen, die direkt die Zusammensetzung eines Labels betreffen, sind auch Fragen der Kategorienstruktur je Hierarchietiefe insbesondere für hierarchische Cluster-Labeling-Verfahren von Interesse.

Kategorien	1	2	3	4	5	6	7	8	17
Webseiten	2.172.230	154.505	8.088	708	95	22	9	1	1
Häufigkeit	93,0%	6,2%	0,17%	0,01%	0,0%	0,0%	0,0%	0,0%	0,0%

Tabelle 3.6 *: Zuweisung von Internetseiten zu Kategorien im Open Directory-Projekt.*

Kategorien	keine	1	2	3	4	5	6	7	≤ 75
Artikel	4.012.410	761.206	618.330	482.406	321.820	218.471	153.992	106.753	257.547
Häufigkeit	57,9%	11,0%	8,9%	6,9%	4,6%	3,2%	2,2%	1,5%	3,8%
Häufigkeit*	-	26,0%	21,1%	16,5%	11,0%	7,5%	5,3%	3,7%	8,9%

Tabelle 3.7 *: Zuweisung von Artikeln zu Kategorien in Wikipedia. Bei der Berechnung der Häufigkeit* wurden die Artikel, die keiner Kategorie zugewiesen wurden, aus der Datenbasis entfernt. Somit ergibt eine durchschnittliche Zuordnung eines Artikels in 3,5 Kategorien. Der Artikel über Albert Einstein ist insgesamt 75 Kategorien zugewiesen.*

Für Wikipedia liegt keine Hierarchie der Kategorien zur Analyse vor. Stattdessen werden die häufigsten Wort-Bigramme und Wort-Trigramme in Kategorienamen ermittelt, die in Tabelle 3.8 auf Seite 39 aufgeführt sind. Die Verwendung der Präpositionen lässt Rückschlüsse auf hierarchische Zusammenhänge zu. Nastase u. Strube (2008) verwenden Part-of-speech-Muster, um Zusammenhänge aufzudecken. Für Kategorien, die den Mustern [*members of* X], [X *VBG IN* Y] und [X *IN* Y] folgen, wird Y als Unterkategorie von X identifiziert. Derart strukturierte Phrasen sind ebenso in Texten anzutreffen, so dass ein Cluster-Labeling Verfahren auf diese Informationen zurückgreifen könnte.

Für das Open Directory Project liegen hierarchische Informationen über Kategorien bis zur Tiefe 14 vor. Eine Verteilung der Kategorien pro Tiefe zeigt Tabelle 3.9 auf Seite 40. Tiefe 0 und 1 sind für die Analyse nicht von Interesse, da die Wurzelkategorie *Top* für alle Kategorien identisch ist. In Tiefe 1 findet eine Verzweigung der Taxonomie hinsichtlich der Sprache statt. Alle weiteren Kategorien sind aus Nomen und Eigennamen aufgebaut, wobei insbesondere Personennamen erst ab Tiefe 5 auftreten.

Die Auswertung der Hierarchie der Kategorien des ODP bringt keine weiteren Erkenntnisse, die sich auf das Cluster-Labeling übertragen ließen.

Zusammenfassung Erkenntnisse aus ODP sind insbesondere bei der Erstellung von Labeln für ein hierarchisches Clustering hilfreich:

- Label für Element in der Hierarchie sollten aus ein bis drei Worten zusammengesetzt werden.
- Als Wortarten sollten Nomina und Verbalsubstantive verwendet werden.

Bigramm	abs. Häufigkeit	Trigramm	abs. Häufigkeit
of the	23.065	the United States	7.715
People from	20.844	Films directed by	2.997
in the	17.185	and structures in	2.420
established in	8.726	Buildings and structures	2.381
as of	6.347	articles by quality	2.032
university of	4.585	Members of the	1.946
Articles with	4.249	Songs written by	1.550
stations in	4.235		
List of	4.007		
articles by	3.832		
based in	3.781		
by country	3.699		

Tabelle 3.8 *: Die häufigsten Wort-Bigramme und Wort-Trigramme in Kategorienamen von Wikipedia.*

- Label für Kategorien können mehrdeutig sein, da der Pfad von der Wurzel bis zur eigentlichen Kategorie diese eindeutig beschreibt[3].

Die Erkenntnisse der Auswertung der Wikipedia sind dagegen für flache, partitionierende Clustering-Verfahren nützlich:

- Label sollten vorzugsweise aus mindestens zwei Worten bestehen. Damit ein Label grammatikalisch korrekt ist, sollten Satzgrenzen nicht überschritten werden.
- Als Wortarten eignen sich Nominalphrasen und Eigennamen.
- Jedes Label sollte alleinstehend eindeutig sein.

Allgemein ist festzuhalten:

- Die Überlappung der Cluster sollte erlaubt werden.
- Bei Überlappungen der Cluster sind die einzelnen Aspekte eines Themas im Cluster mittels Konjunktion im Label anzuzeigen – beispielsweise *Reise und Tourismus*.

Die Ergebnisse helfen bei der Definition wünschenswerter Label-Eigenschaften.

[3] Bei Hierarchien geringer Tiefe sollten dennoch Label eindeutig gewählt werden, da die Pfadlänge im Zweifelsfall zu wenig Kontext zur Auflösung der Mehrdeutigkeit bietet.

Tiefe	Wort N-Gramm	abs. Häufigkeit	Beispiel
0	JJ	1	Top
1	JJ	82	English
2	NN	8	Science
	NNP	2	Health
	NNS	2	Computers
	VBG	2	Shopping
3	NNS	95	Flowers
	NN	89	Security
	NNP	45	Multimedia
	VBG	34	Cooking
	NNP NNPS	31	Winter Sports
	NNP NNP	29	North America
4	NNS	1.134	Resources
	NN	863	Puppetry
	NNP	599	Software
	NNP NNP	394	Business Development
	VBG	257	Accounting
5	NNP NNP	3.789	John Ritter
	NNS	3.721	Characters
	NN	3.445	Contemporary
	NNP CC NNP	1.245	Film and Video
	NNP NNPS	928	Fan Pages
6	NNP NNP	9.164	Max Ernst
	NNP	8.801	Italy
	NNS	6.839	Cartoons
	NN	6.188	Television

Tabelle 3.9 *: Häufigkeit von Wortarten in Kategorien des Open Directory Projects bis Hierarchietiefe 6. Die Tabelle zeigt einen Ausschnitt häufiger Wort N-Gramme.*

4 Problematik des Cluster-Labelings

Kapitel 3 diskutierte anhand menschlich ausgewählter Kategorienamen von Wikipedia und dem Open Directory Project bereits, was unter einem verständlichen Label zu verstehen ist.

Es existieren verschiedene Arbeiten, die unter anderem informelle Label-Eigenschaften nennen. Kummamuru u. a. (2004) definieren hierzu Anforderungen an eine automatisch zu erzeugende Taxonomie. Diese soll anschließend von einem Nutzer effizient nach relevanten Dokumenten zu durchsuchen sein. Die von Kummamuru u. a. aufgestellten Eigenschaften sind:

Überdeckung Eine Taxonomie soll möglichst alle Dokumente der Dokumentkollektion enthalten. Gleiches ist in dieser Arbeit von ein einem Cluster-Labeling-Verfahren zu verlangen.

Kompaktheit Eine Taxonomie sollte möglichst flach sein und den Großteil der Dokumente enthalten. Dieses erspart dem Nutzer eine langwierige Suche nach relevanten Dokumenten. Es sind in dieser Arbeit also prägnante Cluster-Label gesucht, die den Inhalt eines Clusters beschreiben.

Trennschärfe Jeder Knoten (Cluster) repräsentiert ein Thema beziehungsweise einen Aspekt eines Themas. Daher sollten sich Knoten dergleichen Ebene in der Taxonomie möglichst stark voneinander abgrenzen, um Mehrdeutigkeiten aufzulösen. Ein Nutzer soll effizient die einzelnen Knoten unterscheiden können. Cluster sollten sich also nicht zu stark überlappen.

Vorhersagbarkeit Aussagekräftige Label an jedem Knoten der Taxonomie helfen dem Nutzer bei der Navigation durch diese. Daher sollte jedes Label bestmöglich über den Inhalt der Dokumente am Knoten informieren.

Subsumption Von der Wurzel der Taxonomie ausgehend sollten die Label entlang des Pfades immer spezieller werden. Das Label eines Vaterknotens hat immer allgemeiner zu sein als das seiner Kinder.

Für Weiss (2006) hat ein Cluster-Label folgende drei Eigenschaften zu erfüllen:

Verständlichkeit Cluster-Label sollen für den Nutzer verständlich sein. Anstatt zu definieren, welche Kriterien hierfür zu erfüllen sind, führt Weiss Beispiele für unverständliche Label auf. Grammatikfehler und unvollständige Satzbausteine wie *und Tourismus* im Cluster-Label sind zu vermeiden. Dies gilt auch für Äquivokationen.

Prägnanz Ein Cluster-Label sollte so kurz wie möglich sein, dennoch möglichst viel Informationen über die Dokumente im Cluster bieten. Alle Terme sollten nach Weiss aus einem Cluster-Label entfernt werden, die die Verständlichkeit des Labels nicht verschlechtern.

Transparenz Für einen Nutzer soll nachvollziehbar sein, wieso ein Dokument einem Cluster zugewiesen ist und warum das Cluster genau mit diesem Label assoziiert ist.

Nguyen u. a. (2009) sind an *informativen* und *bedeutungsvollen* Cluster-Labeln interessiert. Sie verwenden dazu den t-Wert der Teststatistik der Studentschen t-Verteilung. In Kapitel 2.1.2 wurde bereits gezeigt, dass dieser keine Aussage über die Bedeutsamkeit eines Labels macht.

Die aufgezeigten Label-Eigenschaften sind ausschließlich informell formuliert. Eine Formalisierung würde dagegen bei der Definition zukünftiger Algorithmen des Cluster-Labelings helfen und eine Evaluierung der Algorithmen hinsichtlich der Erfüllung der Eigenschaften erlauben.

Dieses Kapitel führt in die Problematik des Cluster-Labelings ein. Nach einer kurzen Definition des Problems werden Anforderungen an Cluster-Label formal definiert. Das Kapitel wird abgeschlossen von Anforderungen, die an ein Clustering zu stellen sind.

4.1 Problemstellung

Erzeuge ein Clustering-Algorithmus für eine unstrukturierte Dokumentkollektion $\mathcal{D}$ eine Kategorisierung $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$. Zu einer gegebenen Kategorisierung $\mathcal{C}$ ist ein Cluster-Labeling $\mathcal{L} = \{l_1, l_2, \ldots, l_n\}$ gesucht. Ein Cluster-Label $l \in \mathcal{L}$ besteht aus einer Menge von Phrasen $p \in P$, wobei P die Menge aller Phrasen eines Clusters ist. Eine Cluster-Labeling-Funktion $\tau : \mathcal{C} \to \mathcal{L}$ ordnet jedem Cluster $c \in \mathcal{C}$ ein Cluster-Label zu.

Ein hierarchisches Clustering-Verfahren impliziert zudem eine Cluster-Label-Hierarchie $\mathcal{H}$ auf $\mathcal{C}$. Bei der Hierarchie handelt es sich um eine Taxonomie. Sei $c_i, c_j \in \mathcal{C}$, $c_i \neq c_j$. Ist c_j eine Spezialisierung von c_i, so liegt c_i dichter an der Wurzel der Taxonomie. Es gilt $c_i \succ c_j$.

4.2 Formalisierung wünschenswerter Label-Eigenschaften

Die folgende Formalisierung wünschenswerter Label-Eigenschaften basiert auf den Ausführungen von Stein u. Meyer zu Eißen (2004). Die Autoren betrachten einzelne, unabhängige Terme, aus denen schließlich Cluster-Label gebildet werden. Alleinstehende Terme gelten allerdings nicht als verständlich. In dieser Arbeit wird deshalb die Formalisierung auf Phrasen übertragen und erweitert.

Verständlichkeit

Ein Cluster-Label soll verständlich sein. Hierzu wird angenommen, dass alle Kategorienamen des Open Directory Projects und Wikipedias als verständlich für den Menschen gelten.

In Kapitel 3 wurde motiviert, eine formale Sprache zu definieren, die die Kategorienamen beider Projekte beschreibt. These ist: Ein Cluster-Label ist verständlich, wenn es Teil der formalen Sprache $L(G)$ ist, die durch die Grammatik G definiert wird. Ein Cluster-Label soll Nominalphrase (NP) sein. Für alle Phrasen p eines Cluster-Labels $\tau(c)$ für das Cluster c wird somit verlangt:

$$\forall_{c \in \mathcal{C}} \; \forall_{p \in \tau(c)} \; : \; p \in L(G) \tag{4.1}$$

mit $L(G) = \{w \in \Sigma^* \mid S \rightarrow_G^* w\}$. Sei ferner $G = (V, \Sigma, P, S)$ eine kontextfreie Grammatik definiert durch ein 4-Tupel mit

V – der endlichen Menge der Nichtterminale

Σ – der endlichen Menge der Terminalsymbole, $\Sigma \cap V = \emptyset$

P – der endlichen Menge der Produktionsregeln

S – der Startvariable S, $S \in V$

und den Produktionsregeln

$$
\begin{aligned}
S &\rightarrow NP \\
NP &\rightarrow (\text{<JJ>} \mid N)\ NP \\
NP &\rightarrow N \\
N &\rightarrow \text{<NN>} \mid \text{<NNP>} \mid \text{<NNS>}\ .
\end{aligned}
$$

Nominalphrasen sind im Unterschied zu häufig im Text vorkommenden Phrasen immer grammatikalisch korrekt. Sollte ein Cluster-Label aus nur einem Term bestehen, so garantiert die vorliegende Definition, dass es sich um ein Nomen handelt.

Die *Verständlichkeit* ist eine *intra-cluster* Eigenschaft.

Weiss (2006) motiviert für die Verständlichkeit, dass ein Nutzer nachvollziehen kann, wieso eine Phrase als Cluster-Label ausgewählt wurde. Dieser Aspekt wird bei der Überdeckung berücksichtigt.

Überdeckung

Cluster-Label sollen möglichst alle Dokumente im Cluster repräsentieren. Eine Phrase, die in jedem Dokument auftritt, ist repräsentativer für ein Cluster als eine, die nur in wenigen Dokumenten sehr häufig vorkommt. Daher sollten Phrasen des Cluster-Labels eine höhere Dokumentüberdeckung im Cluster aufweisen als alle anderen Phrasen des Clusters. Es wird verlangt, dass

$$\forall_{c \in \mathcal{C}}\ \exists_{p' \in \tau(c)}\ \forall_{\substack{p \in P_c \\ p \notin \tau(c)}}\ :\ df_c(p) \ll df_c(p'), \tag{4.2}$$

und

$$\forall_{c \in \mathcal{C}}\ :\ \bigcup_{p \in \tau(c)} \{d \mid d \in c\}$$

gilt.

Die letzte Forderung erlaubt sowohl ein überlappendes Clustering als auch eine monothetische Merkmalsauswahl. Es wird nicht ausgeschlossen, dass verschiedene Phrasen $p \in \tau(c)$ dieselben Dokumente überdecken.

Durch eine monothetische Merkmalsauswahl ist es für einen Nutzer zudem nachvollziehbar, warum ein Label das Cluster beschreibt: Es ist in jedem Dokument enthalten.

Die *Überdeckung* ist eine *intra-cluster* Eigenschaft.

Trennschärfe

Neben der Forderung nach hoher Überdeckung einer Phrase im eigenen Cluster sollte zur Unterscheidbarkeit von Clustern dieselbe Phrase eine deutlich geringere Dokumentüberdeckung in allen anderen Clustern besitzen. Es wird gefordert, dass

$$\forall_{\substack{c_i,c_j\in\mathcal{C}\\ c_i\neq c_j}} \exists_{p\in\tau(c_j)} \ : \ \frac{\mathrm{df}_{c_i}(p)}{|c_i|} \ll \frac{\mathrm{df}_{c_j}(p)}{|c_j|} \tag{4.3}$$

zu erfüllen ist.

Die *Trennschärfe* ist eine *inter-cluster* Eigenschaft.

Minimale Überlappung

Eine minimale Überlappung der Cluster führt dazu, dass ein Cluster-Label möglichst eindeutig die Dokumente eines Clusters beschreibt. Somit fällt es einem Nutzer leichter, zwei Cluster voneinander zu unterscheiden. Diese Eigenschaft ist verwandt mit der *Trennschärfe*. Es wird gefordert, dass

$$\forall_{\substack{c_i,c_j\in\mathcal{C}\\ c_i\neq c_j}} \exists_{p\in\tau(c_j)} : \ \frac{|c_i(p)\cap c_j(p)|}{|c_i(p)\cup c_j(p)|} \ll 1 \tag{4.4}$$

mit $c(p) := \{d \mid d \in c\}$ zu erfüllen ist.

Die *minimale Überlappung* ist eine *inter-cluster* Eigenschaft.

Eindeutigkeit

Cluster-Label sollen eindeutig sein. Das heißt, dass zwei Cluster nicht dasselbe Cluster-Label zugewiesen bekommen sollten. Es gilt nach Stein u. Meyer zu Eißen (2004):

$$\forall_{\substack{c_i,c_j\in\mathcal{C}\\ c_i\notin c_j}} \ : \ \tau(c_i)\cap\tau(c_j)=\emptyset \tag{4.5}$$

Die *Eindeutigkeit* ist eine *inter-cluster* Eigenschaft.

Die Forderung nach Eindeutigkeit kann durch die Auflösung von Äquivokationen erweitert werden, indem ein mehrdeutiges Wort durch die Zunahme eines erklärenden Terms eindeutig wird. Ein Beispiel: Bank (Sitzgelegenheit). Ein Cluster-Label sollte daher keine Äquivokation sein. Es gilt:

$$\forall_{c\in\mathcal{C}} \ \forall_{p\in\tau(c)} \ : \ p \text{ ist keine Äquivokation}$$

Es wird externes Wissen benötigt, um Mehrdeutigkeiten zu erkennen.

Die *Eindeutigkeit* ist eine *inter-cluster* und *intra-cluster* Eigenschaft. Die Auflösung von Äquivokationen wird in dieser Arbeit nicht behandelt.

Redundanzfreiheit

Die Redundanzfreiheit ergänzt nach Stein u. Meyer zu Eißen (2004) die Forderung nach Eindeutigkeit. Es sind Synonyme im Cluster-Label eines Clusters

$$\forall_{c \in \mathcal{C}} \; \forall_{\substack{p,p' \in \tau(c) \\ p \neq p'}} \; : \; p \text{ und } p' \text{ sind nicht synonym}$$

und zwischen Clustern

$$\forall_{\substack{c_i,c_j \in \mathcal{C} \\ c_i \neq c_j}} \; \forall_{\substack{p,p' \in \tau(c) \\ p \neq p'}} \; : \; p \text{ und } p' \text{ sind nicht synonym}$$

zu vermeiden.

Die *Redundanzfreiheit* ist eine *inter-cluster* und eine *intra-cluster* Eigenschaft. Es wird externes Wissen benötigt, um Synonyme zu erkennen.

Hierarchische Konsistenz

Stellt ein Cluster c_j eine Spezialisierung eines Cluster c_i dar, $c_j \sqsubseteq c_i$, so spiegelt sich dies auch in den zugehörigen Cluster-Labeln wieder (Stein u. Meyer zu Eißen, 2004):

$$\forall_{\substack{c_i,c_j \in \mathcal{C} \\ c_i \neq c_j}} : c_j \sqsubseteq c_i \Rightarrow P(p_i|p_j) = 1 \wedge P(p_j|p_i) < 1,$$

mit $p_i \in \tau(c_i)$ und $p_j \in \tau(c_j)$. Dies entspricht der Formulierung der Subsumption: w_i verallgemeinert w_j (Sanderson u. Croft, 1999). Sanderson u. Croft (1999) argumentieren, die Forderung sei zu streng, so dass $P(p_i|p_j) = 0.8 \wedge P(p_j|p_i) < 1$ gilt.

Subsumption wird durch ein Beispiel verständlich (siehe auch Abbildung 4.1):

Italienisches Restaurant $\sqsubseteq$ Restaurant

Chinesisches Restaurant $\sqsubseteq$ Restaurant

Chinesisches Restaurant $\sqsubseteq$ **not** Italienisches Restaurant

Die *hierarchische Konsistenz* ist eine *inter-cluster* Eigenschaft. Sie wird nur bei hierarchischen Cluster-Labeling-Verfahren vorausgesetzt. Ein weitere hierarchische Label-Eigenschaft ist die Cluster-Label-Kohärenz von Schwesterknoten.

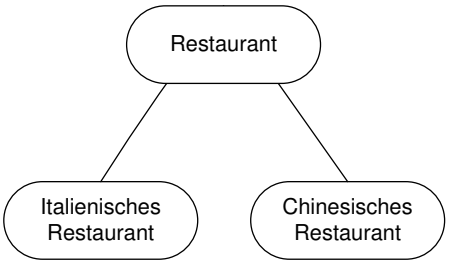

Abbildung 4.1 *: Hierarchische Darstellung der Beziehung zwischen italienischem und chinesischem Restaurant. Beide werden durch den Begriff „Restaurant“ verallgemeinert.*

Cluster-Label-Kohärenz von Schwesterknoten

Schwesterknoten in einer Hierarchie sollten bezüglich ihrer Cluster-Label semantisch enger verbunden sein als Schwesterknoten anderer Teilbäume. Es wird gefordert, dass

$$\forall_{\substack{c \in \mathcal{C} \\ d^+(c)>0}} \; \forall_{\substack{c_i, c_j \in N^+(\mathcal{C}) \\ c_i \neq c_j}} \; : \; \varphi(\tau(c_i), \tau(c_j)) \approx 1$$

gilt. Für φ ist ein geeignetes thesaurus-basiertes Ähnlichkeitsmaß einzusetzen.

Die Kohärenz von Schwesterknoten ist eine *inter-cluster* Eigenschaft. Es wird externes Wissen benötigt, um diese zu erfüllen.

4.3 Anforderungen an ein Clustering

Neben formalen Forderungen an ein Cluster-Label können ebenso Forderungen an ein Clustering gestellt werden, so dass die erzeugte Kategorisierung für den Nutzer nachvollziehbar ist. Diese sind:

Monothetisch Ein Clustering sollte monothetisch sein. Jedes Dokument eines Clusters enthält ein gewähltes Merkmal. Wird dieses Merkmal als Cluster-Label verwendet, so ist für einen Nutzer direkt nachvollziehbar, warum die Dokumente genau diesem Cluster zugewiesen wurden (Weiss, 2006).

Überlappend Ein Dokument kann mehrere Aspekte eines Themas abdecken. Ein Artikel über *Parlamentswahlen in der Slowakei 2010* kann sowohl einem Cluster *Parlamentswahlen*, einem Cluster *Slowakei* als auch einem Cluster *2010* zugewiesen werden. Durch überlappende Cluster wird dem Nutzer erlaubt, ein relevantes Dokument unter Berücksichtigung verschiedenster Aspekte zu finden.

Partiell Traditionelle Clustering-Verfahren erzeugen in der Regel ein vollständiges Clustering. Für das Cluster-Labeling wird dagegen ein partielles Clustering bevorzugt, weil Cluster, die nicht prägnant beschrieben werden können, für den Nutzer keinen Wert besitzen. Somit können das Cluster sowie die zugewiesenen Dokumente verworfen werden.

Hierarchisch Die Analyse des Open Directory Projects in Kapitel 3 zeigt, dass durch den Kontext einer Kategorie Cluster bereits durch alleinstehende Nomen eindeutig beschrieben werden können. Der Kontext erhöht also die Verständlichkeit eines Cluster-Labels.

4.4 Ermittlung von Themen – ein Überblick

Die Ermittlung von Themen und Konzepten innerhalb von Dokumenten oder für Dokumentkollektionen, sowie die Zuweisung von Dokumenten zu Kategorien sind ein wichtiger Bestandteil, den viele Systeme der Informationsverarbeitung teilen. Im Folgenden soll ein Überblick über die Forschungsbereiche gegeben werden, die eine Themenfindung untersuchen. Es werden jeweils relevante wissenschaftliche Beiträge vorgestellt. Gemeinsamkeiten und Unterschiede zum Cluster-Labeling werden diskutiert.

Automatische Ermittlung von Schlüsselwörtern

Verfahren zur Schlüsselwortbestimmung wurden bereits in Kapitel 2.1 vorgestellt. Einen zusätzlichen Überblick über die aktuelle Forschung bietet Klüger (2006). Frühe Arbeiten im Bereich der Schlüsselwortbestimmung verwenden unter anderem Wort- und Phrasenhäufigkeiten (Luhn, 1958), die Position im Text (Baxendale, 1958) oder Schlüsselphrasen (Edmundson, 1969) als Merkmale, um relevante Terme zu ermitteln.

Automatische Zusammenfassung von Texten

Es werden Verfahren entwickelt, die Zusammenfassungen für gegebene Texte automatisch erzeugen. Radev u. a. (2002a) heben drei wichtige Merkmale von Zusammenfassungen hervor. Zusammenfassungen

- können für ein oder mehrere Dokumente erstellt werden,
- sollten die wichtigsten Aspekte des Originaltexts beinhalten und

- sollten kurz sein.

Texte diskutieren in der Regel mehrere Aspekte eines Themas oder handeln von mehreren verschiedenen Themen. In der Forschung werden themenspezifische Zusammenfassungen erstellt, um diesem gerecht zu werden. Hovy u. Lin (1999) entwickeln hierzu ein System namens SUMMARIST. Das System identifiziert zunächst die unterschiedlichen Themen eines Textes. Hierbei wird sich auf frühe Arbeiten der Schlüsselwortbestimmung gestützt, um relevante Terme oder Sätze zu ermitteln. Sind alle Themen identifiziert, werden diese auf allgemeinere Konzepte abgebildet. *Rad*, *Lenker* und *Luftpumpe* bilden zum Beispiel auf das Konzept *Fahrrad* ab. Eine solche Verallgemeinerung ist nicht möglich, wenn ausschließlich die Worthäufigkeiten herangezogen werden. Stattdessen zählen Hovy u. Lin Konzepte. Jedes Wort wird auf ein Konzept in *WordNet* abgebildet. Existiert kein entsprechendes Konzept, wird die Worthäufigkeit verwendet.

Neben Methoden der Schlüsselwortbestimmung werden ebenso Clustering-Verfahren eingesetzt, um Themen in Dokumentkollektion zu ermitteln (Radev u. a., 2000). Insbesondere werden dabei centroid-basierte Verfahren eingesetzt. Zunächst wird ein Clustering der Dokumentkollektion durchgeführt. Jeder Cluster entspricht dabei einem Thema. Zusammenfassungen für jedes Thema werden anschließend erzeugt, in dem für die häufigsten Terme des jeweiligen Centroiden die dazugehörigen Sätze ermittelt werden.

Der Beitrag der Forschung über die automatische Erstellung von Zusammenfassungen für das Cluster-Labeling ist jedoch gering, da Zusammenfassungen sich aufgrund ihrer Länge nicht als Cluster-Label eignen.

Einen Überblick über die aktuelle Forschung bietet Das (2007).

Clustering von Suchergebnissen

Kommerzielle Suchmaschinen wie *Google* oder *Yahoo!* präsentieren dem Nutzer eine nach Relevanz sortierte Ergebnisliste zu einer gestellten Suchanfrage. Für den Nutzer ist es jedoch oft schwierig, die Vielzahl an Ergebnissen zu überblicken und zu entscheiden, welche tatsächlich für ihn relevant sind. Somit müssen häufig mehrere Anfragen durch Hinzunahme weiterer Terme an die Suchmaschine gestellt werden, um den Suchraum sukzessiv einzuschränken. Diese Art der Suche wird *direkte Suche* genannt.

Um einen Nutzer bei der Suche zu unterstützen, wird ein Clustering der Suchergebnisse durchgeführt (engl. *web search result clustering*). Dabei werden ähnliche Resultate gruppiert. Diese werden zusätzlich zu der nach Relevanz sortierten Ergebnisliste präsentiert. Für ungerichtete Anfragen wie *Jaguar* sind somit die verschiedenen Bedeutungen

Anwendung	Cluster-Label	Berechnung	Eingabedaten	Anzahl Cluster	Zuweisung	GUI
Dokumenten-Clustering	Centroid	offline	Dokumente	fest	disjunkt	nein
Suchergebnis-Clustering	Natürliche Sprache	online	Snippets	variabel	überlappend	ja

Tabelle 4.1 *: Unterschiede zwischen Dokumenten- und Suchergebnis-Clustering nach Carpineto u. a. (2009). Beim Clustering von Suchergebnissen werden Cluster-Label, im Gegensatz zum Dokumenten-Clustering, nicht aus dem Centroiden gewonnen. Angenommen wird, dass solche hochfrequente Terme nicht repräsentativ für alle im Cluster enthaltenen Dokumente sind. Zudem besitzen einzelne Terme zu wenig Aussagekraft. Da das Clustering für jede Suchanfrage neu erstellt wird (online), ist die optimale Cluster-Anzahl im voraus nicht bekannt. Im Gegensatz zum Dokumenten-Clustering können Dokumente jedoch mehr als nur einem Cluster angehören. Begründet wird dies damit, dass auf diese Weise in einer Hierarchie ein Nutzer über mehrere Pfade das gewünschte Dokument finden kann. Eine zusätzliche Herausforderung beim Suchresultate-Clustering ist es, die ermittelten Kategorien dem Nutzer anschaulich zu präsentieren.*

des Wortes für einen Nutzer direkt sichtbar: *Tier*, *Automarke* und *Betriebssystem* – um nur einige zu nennen. Eine ausschließlich nach Relevanz sortierte Ergebnisliste leistet dies nicht. Als gegenwärtig einzige kommerzielle Suchmaschine führt *Yippy!*[1] ein solches Clustering durch.

Gemeinsamkeiten zum Cluster-Labeling finden sich bei der Bedeutung, die der Bezeichnung der einzelnen Kategorien zugesprochen wird. Weiss (2006) prägt in diesem Zusammenhang die Aussage „*description comes first*“. Die aussagekräftige Bezeichnung der Cluster steht im Vordergrund. Cluster seien für den Nutzer nicht von Wert, wenn diese bedeutungslose oder mehrdeutige Cluster-Label besitzen.

Gegenüber Dokumenten-Clustering stellt das Clustering von Suchergebnissen besondere Anforderungen an ein Clustering. Tabelle 4.1 zeigt diese auf. Diese Anforderungen sind auch für das Cluster-Labeling sinnvoll und sollten bei der Wahl des Clustering-Verfahrens berücksichtigt werden.

Das Clustering von Suchergebnissen grenzt sich in zwei Punkten vom Cluster-Labeling ab. Es dienen ausschließlich sehr kurze Texte[2] als Eingabe für das Clustering. Es werden hohe Anforderungen an die Reaktionszeit des Systems gestellt, welches sich mit traditionellen Suchmaschinen messen lassen muss.

[1] Die Suchmaschine *Yippy!* ist früher unter dem Namen *Vivisimo* bekannt geworden.

[2] Es wird hierbei von *Snippets* gesprochen. Ein Snippet besteht aus der Internet-Adresse des Suchergebnisses, dem Titel und einer Kurzbeschreibung.

Einen Überblick über die aktuelle Forschung bieten Carpineto u. a. (2009).

Facetten-Suche

Alternativ zur direkten Suche ist die sogenannte *navigierende Suche* einsetzbar, um Nutzer bei der Suche nach relevanten Informationen zu unterstützen. Ein Nutzer ist hierbei in der Lage, den Suchraum durch fest vorgegebene Kategorien schrittweise solange einzuschränken, bis er die Informationen findet, die für ihn relevant sind. Je tiefer er die Hierarchie der Kategorien hinabsteigt, desto spezieller werden diese.

Forscher, die sich mit der *Facetten-Suche* beschäftigen, argumentieren, dass beide Paradigmen – sowohl *direkte* als auch *navigierende Suche* – den Nutzer nicht ausreichend genug bei der Suche unterstützen. Daher wird eine Kombination beider vorgeschlagen. Der Nutzer stellt zunächst eine Suchanfrage. Um den Suchraum einzuschränken, werden ihm anschließend sogenannte *Facetten* (Merkmalskategorien) des gesuchten Begriffs präsentiert. Vor allem in E-Commerce Systemen wie *Ebay* oder *Amazon*, bei denen Produkte sich nach Farben, Preisspanne oder Gerätegröße sortieren lassen, ist dieses neue Paradigma populär. Die Forschung beschäftigt sich unter anderem mit folgenden Themen (SIGIR, 2006):

- Ermittlung von Facetten aus Metadaten oder direkt aus Dokumenten,
- Gestaltung von Nutzerschnittstellen bei der Facetten-Suche,
- Facetten-Suche für Dokumentkollektion mit verrauschten Daten und
- Facetten-Suche in E-Commerce Anwendungen.

Für viele E-Commerce Systeme ist der Produktbereich geschlossen. Facetten werden hierbei durch Metadaten a priori bestimmt. Für offene Systeme, wie beispielsweise eine Internet-Suchmaschine, liegen jedoch keine Metadaten vor. Hier ist es notwendig, Facetten direkt aus dem Inhalt der Dokumente zu generieren. Fujimura u. a. (2006) erstellen beispielsweise Facetten für Weblogs, indem sie *Named Entities* wie Personen, Organisationen oder Orte ermitteln. Diese Aufgabe ist ähnlich der des Cluster-Labelings.

Facetten werden üblicherweise hierarchisch organisiert. Es bilden sich sogenannte dynamische Taxonomien heraus (Sacco, 2000). Diese unterscheiden sich primär von traditionellen Taxonomien, indem erlaubt wird, dass ein Element mehrere Vaterknoten besitzen darf. Ein Fahrrad lässt sich anhand zahlreicher Merkmale beschreiben, so dass es sinnvoll erscheint, ein bestimmtes Modell unter anderem zu den Facetten *28 Zoll* und *Rennrad*

zuzuordnen. Diederich u. a. (2007) verwenden hierzu Metadaten aus elektronischen Dokumenten, um Facetten dynamisch zu erzeugen. Da ein Dokument unter mehrere Konzepte eingeordnet werden kann, entsteht ein direkter azyklischer Graph. Eine traditionelle Taxonomie ist dagegen ein Baum.

Die Berücksichtigung des Konzepts dynamischer Taxonomien für ein hierarchisches Cluster-Labeling ist sinnvoll, da sich überlappende Cluster ergeben. Somit ist ein Nutzer in der Lage, über mehrere Pfade ein gewünschtes Dokument zu finden, ohne den ganzen Pfad bis zur Wurzel wieder hinauf wandern zu müssen.

Einen Überblick über die aktuelle Forschung bietet Tunkelang (2009).

Ermittlung und Verfolgung von Themen

Es wird versucht, automatisch themenspezifische Informationen aus Datenströmen von Nachrichtenagenturen oder -sendungen zu filtern und miteinander in Beziehung zu setzen. Ziel ist es unter anderem, online neue Ereignisse (engl. *event detection*) und das erste Auftreten einer Nachricht zu einem neuen Thema (engl. *first story detection*) in einem Nachrichtenstrom zu entdecken. Im Gegensatz zu anderen Forschungsgebieten wird bei der Ermittlung und Verfolgung von Themen (engl. *topic detection and tracking*) stets die chronologische Reihenfolge der Dokumente berücksichtigt.

Verfahren zur Erkennung neuer Ereignisse basieren auf Clustering-Verfahren (Allan u. a., 1998, 2000; Liu u. Croft, 2004). Angenommen wird, dass alle Dokumente innerhalb eines Clusters über dasselbe Thema handeln. Ähnlich der themenspezifischen Zusammenfassung von Texten repräsentiert jedes Cluster somit ein bestimmtes Thema. Neue Nachrichten werden entweder einem bereits bestehenden Cluster hinzugefügt oder dienen als Ausgangspunkt für ein neues Cluster. Im letzteren Fall ist dies äquivalent mit dem Auftreten einer ersten Nachricht zu einem neuen Thema. Die Herausforderung besteht darin, die einzelnen Cluster entsprechend bedeutsam zu benennen.

Muthukrishnan u. a. (2008) ermitteln diskriminierende Cluster-Label für verschiedene Themen, die innerhalb einer Dokumentkollektion auftreten. Zur Ermittlung von Cluster-Labeln setzen Muthukrishnan u. a. das bei der Schlüsselwortbestimmung vorgestellte Verfahren von Tomokiyo u. Hurst (2003) ein. Cluster-Label sollen jeweils eine möglichst große Dokumentkollektion repräsentieren. Die einzelnen Dokumentkollektionen sollen sich dabei jedoch nur geringfügig überschneiden. Diese Aufgabe ist ähnlich dem Mengenüberdeckungsproblem. Dieses ist NP-vollständig (Karp, 1972), so dass Muthukrishnan u. a. ein Greedy-Verfahren zur Lösung einsetzen.

Im Gegensatz zum Cluster-Labeling wird hier ein Thema weniger als eine Kategorie, sondern vielmehr als ein Ereignis oder eine Aktivität verstanden. Damit eingeschlossen sind ebenso verwandte Ereignisse und Aktivitäten (Wayne, 2000). Des weiteren ist die chronologische Reihenfolge von Dokumenten beziehungsweise Themen beim Cluster-Labeling nicht relevant.

Einen Überblick über die aktuelle Forschung bietet Allan (2002).

Automatische Erstellung von Taxonomien

Verfahren zur automatischen Erstellung von Taxonomien versuchen, eine Hierarchie ausschließlich mittels der zugrundeliegenden Dokumentkollektion zu gewinnen. Krishnapuram u. Kummamuru (2003) geben einen Überblick über die Schwerpunkte des Forschungsgebiets, die im Einklang mit denen des Cluster-Labelings stehen:

- Ermittlung von Dokumenten mit ähnlichem Inhalt beziehungsweise Thema,
- Ermittlung einer hierarchischen Struktur für Themen und
- Ermittlung passender Cluster-Label für Themen.

Die ersten beiden Schwerpunkte sind mit Hilfe von Clustering-Verfahren zu bewältigen. Krishnapuram u. Kummamuru unterscheiden hierbei monothetische und polythetische Verfahren. Hierarchische Verfahren lassen sich weiterhin in hierarchie-synthetisierend und -analysierend unterteilen. Bei letzteren werden Cluster-Label für Themen nicht direkt aus Termen der Dokumentkollektion gewonnen, sondern aus externen Wissensbasen wie *WordNet* oder *Wikipedia*. Das Clustering wird entweder auf Dokumenten, auf Worten oder auf beiden gleichzeitig (*co-clustering*) durchgeführt (Dhillon, 2001).

Sanderson u. Croft (1999) verwenden ein hierarchie-synthetisierendes Verfahren, um eine Hierarchie zu erstellen. Dabei ermitteln diese in der Dokumentkollektion wichtige Worte und Phrasen. Anschließend wird durch eine Subsumptionsanalyse eine Term-Hierarchie erzeugt.

Bei der automatischen Erstellung von Taxonomien wird die Verwendung *unscharfer Taxonomien* motiviert (Krishnapuram u. Kummamuru, 2003). Diese sind mit dynamischen Taxonomien, wie sie bei der Facetten-Suche eingesetzt werden, vergleichbar. Ein Dokument ist wiederum unterhalb mehrerer Konzepte einsortierbar. Jedoch besitzt es zu jedem Vaterkonzept einen unterschiedlichen Grad der Mitgliedschaft, der sich zu Eins summiert. Krishnapuram u. Kummamuru empfehlen daher, bei der Erstellung von Taxonomien unscharfe Clustering-Verfahren, beispielsweise *Fuzzy k-Means*, einzusetzen.

Die Forschungsergebnisse im Bereich der automatischen Erstellung von Taxonomien sind für das Cluster-Labeling relevant, wenn dieses hierarchisch ist. Unscharfe Taxonomien motivieren ein überlappendes Clustering. Bezüglich des Clusterings werden von Kummamuru u. a. (2004) monothetische Verfahren präferiert.

Einen Überblick über die aktuelle Forschung bieten Krishnapuram u. Kummamuru (2003).

Text-Klassifikation

Bei einer Klassifikation ist eine feststehende Menge an Kategorien gegeben. Aufgabe ist es, jedem Dokument aufgrund seines Inhalts einer oder mehreren Kategorien zuzuweisen. Kategorien dienen als Cluster-Label für Dokumente. Im Gegensatz zum Cluster-Labeling werden ausschließlich einzelne Dokumente betrachtet.

Früher wiesen Experten jedem Dokument manuell Kategorien zu. Bei großen Datenbeständen ist dieses jedoch zeitlich eine nicht mehr zu leistende Aufgabe. Beispielsweise ist es sinnvoll, bei einem Unternehmen zentral per E-Mail eingesandte Support-Anfragen an die entsprechende Fachabteilung automatisch weiterzuleiten. An welche Fachabteilung die E-Mail weitergeleitet werden soll, entscheidet eine Klassifikation. Zur Lösung des Klassifikationsproblems existieren derzeit drei verschiedene Verfahren: überwacht, semi-überwacht und unüberwacht.

Überwachte Verfahren (engl. *supervised classification*) basieren auf statistischen Klassifikationsverfahren und Techniken des Maschinellen Lernens: Naive-Bayes Klassifikatoren und Entscheidungsbäume (Lewis u. Ringuette, 1994), Support Vector Machines (Joachims, 1997) und Neuronale Netze (Yang, 1999) – um nur einige zu nennen. Eine von menschlichen Experten gruppierte Dokumentkollektion wird als Trainingsmenge verwendet, um einen Klassifikator zu trainieren. Dieser ist anschließend in der Lage, zuvor nicht gesehene Dokumente zu kategorisieren.

Bei der Klassifikation werden oftmals nicht alle Terme der Dokumente berücksichtigt, sondern nur eine Untermenge. Dies hat zwei Vorteile. Training und Klassifikation neuer Dokumente sind aufgrund des deutlich reduzierten Vokabulars effizienter zu realisieren. Die Klassifikationsgenauigkeit kann sich erhöhen, da vermeintlich schlechte Merkmale entfernt werden. Merkmale sind schlecht, wenn diese den Klassifikationsfehler erhöhen. Zur Merkmalsauwahl existieren verschiedene Methoden: *Mutual Information*, *Information Gain*, $\mathcal{X}^2$-Statistik und häufigkeitsbasierende Techniken. Eine Evaluierung der genannten Methoden liefern Yang u. Pedersen (1997).

Um gute Klassifikationsergebnisse zu erzielen, ist jedoch eine große Trainingsmenge notwendig. Aufgrund der manuellen Erhebung der Daten ist dies eine zeitaufwändige Arbeit. Daher werden vermehrt semi-überwachte (engl. *semi-supervised classification*) eingesetzt (Blum u. Mitchell, 1998; Nigam, 2001). Verfahren dieser Art reichern die Trainingsmenge mit weiteren Dokumenten an.

Unüberwachte Verfahren (engl. *unsupervised classification*) verzichten bei der Klassifikation von Dokumenten auf eine Trainingsmenge. Rao u. a. (2002) setzen Clustering ein, um Kategorien zu bestimmen. Diese führen zunächst ein Wort-Clustering durch, um semantisch ähnliche Worte zu gruppieren. Jedes Cluster wird durch die fünf häufigsten Worte repräsentiert, die am dichtesten zum jeweiligen Centroiden liegen. Anschließend wird jedes Dokument mit jedem Cluster verglichen. Dabei wird für jedes Dokument-Cluster-Paar die Summe der Häufigkeiten der fünf Worte, die den Cluster repräsentieren, im Dokument ermittelt. Die Worte mit der größten Summe werden dem Dokument zugewiesen. Es ist möglich, dass ein Dokument mehreren Clustern zugewiesen wird. Somit ergeben sich die endgültigen Kategorien für jedes Dokument aus der Vereinigungsmenge der Worte, die die Cluster repräsentieren.

Neueste Verfahren setzen Wikipedia ein, um Kategorien für einzelne Dokumente zu ermitteln (Schonhofen, 2006; Syed u. a., 2008; Coursey u. Mihalcea, 2009). Schonhofen ermittelt und sortiert alle Kategorien aus Wikipedia, die mit einem Dokument verwandt sind. Hierzu werden die Titel der Wikipedia-Artikel mit den Worten des Dokuments auf Übereinstimmung hin überprüft. Die auf diese Weise ermittelten Kategorien werden anhand von vier Faktoren gewichtet. Die Kategorie mit dem höchsten Gewicht wird als Cluster-Label für das Dokument verwendet. Wikipedia dient hier als externe Ontologie.

Für das Cluster-Labeling sind die Verfahren zur Merkmalsauswahl von Interesse. Diese werden ebenso beim Cluster-Labeling eingesetzt, um Cluster-Label für Kategorien zu ermitteln (Popescul u. Ungar, 2000; de Winter u. de Rijke, 2007). Des weiteren zeigt die Arbeit von Rao u. a. (2002), dass unüberwachte Verfahren verwandt mit denen des Cluster-Labelings sind. Ebenso ist der Einsatz von Wikipedia als externe Ontologie eine interessante Idee, um die Qualität von Cluster-Labeln zu verbessern.

Einen Überblick über die Forschung bieten Aas u. Eikvil (1999).

Automatische Bezeichnung themenzentrierter Modelle

Texte werden durch Dokumentmodelle repräsentiert. Klassische Modelle wie das Vektorraummodell (Salton u. a., 1975) sind termbasiert und bauen direkt auf dem Vokabular

eines Textes auf. Für jedes Wort wird dabei eine eigene Dimension aufgespannt. Auf diese Weise entsteht eine Term-Dokument-Matrix. Obwohl termbasierte Modelle intuitiv sind und gute Ergebnisse liefern, modellieren sie wichtige Aspekte der natürlichen Sprache nicht: Synonyme und Äquivokationen (Manning u. Schütze, 1999). Dieses wurde bereits in Kapitel 2.1.1 diskutiert. Neben Latent Semantic Indexing (LSI) lösen konzeptbasierte Dokumentmodelle wie *probabilistisches* LSI (Hofmann, 1999) und *Latent-Dirichlet-Allocation* (Blei u. a., 2003) (LDA) ebenfalls Synonyme und Äquivokationen auf. Diese basieren auf einem erzeugenden Wahrscheinlichkeitsmodell. Es wird angenommen, dass jedes Wort in einem Dokument durch ein bestimmtes Thema erzeugt wird. Unterschiedliche Worte eines Dokuments werden somit von verschiedenen Themen erzeugt. Mit anderen Worten generiert ein sogenanntes themen-zentriertes Modell für das Thema *Katze* die Worte *Milch* und *Maus* mit einer höheren Wahrscheinlichkeit als für ein Modell über Hunde. Mit Hilfe des LDA-Modells ist es also möglich, anhand von Wahrscheinlichkeitsverteilungen verschiedene Themen von Dokumenten oder Dokumentmengen zu ermitteln. Für einen Nutzer haben sie keinen Wert, solange nicht jeder Verteilung ein passendes Cluster-Label zugewiesen wird, welches die Interpretation erleichtert. Hierzu existieren derzeit drei verschiedene Verfahren:

- Subjektive Auswahl geeigneter Cluster-Label (Hofmann, 1999; Blei u. a., 2003; Griffiths u. Steyvers, 2004),
- Auswahl der k besten Terme einer Wahrscheinlichkeitsverteilung (Mei u. Zhai, 2006; Mei u. a., 2006) und
- die automatische Erzeugung von Cluster-Labeln (Mei u. a., 2007).

Mei u. a. sind nach bestem Wissen die einzigen, die sich der letztgenannten Problematik widmen. Sie fordern, dass ein automatisch ermitteltes Cluster-Label für ein themen-zentriertes Modell bestimmte Eigenschaften erfüllen sollte. Ein Cluster-Label sollte verständlich und semantisch relevant sein, diskriminierend gegenüber anderen Cluster-Labeln sein und eine hohe Überdeckung bieten. Die Autoren argumentieren, dass einzelne Terme nicht aussagekräftig genug sind, um als Thema für eine Wahrscheinlichkeitsverteilung herangezogen zu werden. Mei u. a. motivieren somit ebenfalls die Verwendung von Phrasen.

Die Verwendung von themen-zentrierten Modellen in Verbindung mit Cluster-Labeling ist spannend. Gegenwärtig setzen Nguyen u. a. (2009) solche Modelle ein, um die Themenfindung beim Clustering zu verbessern.

Bezeichnung selbstorganisierender Karten (Kohonennetze)

Bei diesen Verfahren werden hochdimensionale Daten mittels eines neuronalen Netzes auf eine zweidimensionale Karte abgebildet (Kohonen u. a., 1999). Es wird versucht, die ursprüngliche Topologie der Daten soweit wie möglich zu erhalten. Selbstorganisierende Karten (engl. *self-organizing maps*), auch Kohonennetze genannt, sind unüberwachte neuronale Netze, die nicht trainiert werden müssen. Aufgrund der Projektion in den zweidimensionalen Raum sind die Originaldaten besser zu visualisieren und zu interpretieren. Die Projektion kann als ein Clustering der Daten verstanden werden, da ähnliche Dokumente auf eine gemeinsame Karte abgebildet werden. In diesem Fall ist eine Dimension ausreichend.

Die Interpretation der projizierten Daten kann zudem erleichtert werden, indem ermittelten Karten passende Cluster-Label zugewiesen werden. Rauber u. Merkl (1999) weisen hierbei jeder Karte das Cluster-Label zu, welches mit der Karte assoziierte Dokumente möglichst gut beschreibt. Rauber u. Merkl argumentieren, dass der mittlere Quantisierungsfehler eines Merkmals der Dokumentvektoren minimal wird, wenn viele Dokumente dasselbe Merkmal teilen. Dieses Merkmal eignet sich, um die Dokumentkollektion zu beschreiben.

Generell unterscheiden sich selbstorganisierende Karten von anderen Clustering-Verfahren dadurch, dass diese eine Reduktion des hochdimensionalen Raumes vornehmen. Eine Gemeinsamkeit mit dem Cluster-Labeling ist, dass durch Auffinden passender Cluster-Label die Interpretation von Daten erleichtert werden soll.

Einen Überblick über die Forschung bietet Kohonen (1997).

Erstellung von Schlagwortwolken

Tagging beschreibt die Aktivität, eine Ressource mit einem oder mehreren Schlüsselworten oder Schlüsselphrasen zu assoziieren. Ein Tag ist als Etikett oder Notiz zu verstehen, um zu einem späteren Zeitpunkt Dinge leichter wieder finden zu können. Tags dienen somit der Organisation von Ressourcen.

Großer Beliebtheit erfreuen sich unter anderem kollaborative Tagging-Dienste wie *Flickr* (Bilder), *del.icio.us* (Internetseiten) oder *Facebook* (soziales Tagging). Beim kollaborativen Tagging annotieren Nutzer selbst die entsprechenden Ressourcen. Der auf diese Weise implizit entstehende sogenannte *tagspace* soll anschließend effizient durchsuchbar sein. Jedoch führt das manuelle Annotieren von Ressourcen zu Problemen, wie Golder

u. Huberman (2006) aufzeigen. Synonyme und Polyseme verursachen dabei Probleme, die bereits in Kapitel 2.1.1 angesprochen wurden. So schließt eine Bildersuche mittels des Schlüsselworts *Auto* keine Bilder mit ein, die nur mit dem Synonym *Pkw* assoziiert sind. Ein weiteres Problem ist, das dieselben Ressourcen von verschiedenen Nutzern unterschiedlich annotiert werden. Fügt ein Nutzer als Schlüsselwort einem Bild *Auto* als Schlüsselwort hinzu, so wird ein anderer durch Zuweisung von *Volkswagen Golf Mk5 (A5/Typ 1K, 2003-2009)* deutlich konkreter.

Um Inkonsistenzen bei der Zuweisung von Tags zu vermeiden und dadurch die Retrievalqualität von Suchmaschinen zu steigern, wird automatisches Tagging eingesetzt. Dieses dient vor allem beim Annotieren von neuen Weblog-Nachrichten dazu, Empfehlungen für passende Schlüsselworte einem Nutzer vorzuschlagen.

Brooks u. Montanez (2006) annotieren automatisch neue Nachrichten in Weblogs. Sie verwenden dazu die drei häufigsten Terme einer Nachricht, gewichtet nach dem *tf-idf*-Schema. Nachteil ist, dass Schlüsselworte somit auf Wort-Unigrammen und dem der Nachricht zugrundeliegenden Vokabular beschränkt sind.

Mishne (2006) ermitteln zu neuen Nachrichten in Weblogs zunächst ähnliche Nachrichten, die von anderen Nutzern zuvor verfasst wurden. Anschließend werden einem Nutzer die dort verwendeten Schlüsselworte als Tags für eine neue Nachricht vorgeschlagen.

Lee u. Chun (2007) setzen dagegen ein überwachtes Lernverfahren ein – ein neuronales Netz. Dieses wird auf einer Menge von Blog-Nachrichten trainiert, denen Tags zugewiesen sind. Anschließend können Empfehlungen für Tags für neue Nachrichten gegeben werden. Grundlage zur Erfassung von Tags sind Verfahren der Schlüsselwortbestimmung. Dabei werden häufige, durch *tf-idf* gewichtete Wort-Bigramme in den Nachrichten ermittelt. Ob die ermittelten Bigramme auch allgemein häufig verwendet werden, wird anschließend mittels WordNet evaluiert.

Neben der automatischen Empfehlung von Schlüsselworten wird ebenso versucht, aus dem *tagspace* eine Hierarchie abzuleiten (Brooks u. Montanez, 2006; Wu u. a., 2006). Solche Hierarchien dienen der Suchoptimierung. Da im Allgemeinen bei der Erstellung von Schlagwortwolken auf vorhandene Schlüsselworte zurückgegriffen wird, finden sich bis auf die Erstellung von Taxonomien aus Schlagworten keine Gemeinsamkeiten mit Ansätzen des Cluster-Labelings.

5 Cluster-Labeling-Verfahren

Die Klassifikation von Cluster-Labeling-Verfahren ist in der Literatur nicht einheitlich. Dies überrascht nicht, da die Forschung in diesem Bereich im Vergleich zum Clustering selbst noch relativ jung ist. Einleitend werden existierende Klassifikationsschemata vorgestellt und bewertet, bevor im Anschluss Cluster-Labeling Verfahren sowie dazugehörige Verfahren im Detail diskutiert werden.

Stein u. Meyer zu Eißen (2004) unterscheiden strukturbasierte und inhaltsbasierte Verfahren. Erstere nutzen bei der Ermittlung von Cluster-Labeln vorhandene Metadaten wie Titel, Autor oder Schlüsselworte eines Dokuments aus (Cutting u. a., 1992). Eine weitere Unterteilung strukturbasierter Verfahren bleibt offen. Inhaltsbasierte Verfahren werden dagegen weiter in polythetisch und monothetisch unterteilt. Der Vorteil monothetischer Verfahren für das Verständnis eines Labelings wurde bereits in Kapitel 4 motiviert. Dennoch zeigt die Recherche von Stein u. Meyer zu Eißen, dass 2004 polythetische Verfahren weitaus stärker verbreitet waren als andere.

Bei der Klassifikation von Stein u. Meyer zu Eißen wird das Labeling als Funktion betrachtet, die im Anschluss an ein Clustering jedem Cluster Terme als Label zuweist. Dagegen berücksichtigen Nguyen u. a. (2009) in ihrem Klassifikationsschema, dass einem Clustering das Labeling auch vorausgehen kann. Sie unterscheiden demnach nur zwei Fälle: *finding clusters first* und *finding labels first*. Synonym werden ebenso *clustering comes first* und *labeling comes first* verwendet. Das Cluster-Labeling ist entweder einem Clustering vorangestellt oder wird erst im Anschluss durchgeführt.

Anhand der Klassifikation von Nguyen u. a. fällt die Einordnung von Ansätzen, die Clustering und Labeling eng miteinander verzahnen, schwer. Wünschenswert wäre daher eine dritte Möglichkeit zur Einordnung dieser Cluster-Labeling-Verfahren. Carpineto u. a. (2009) kommen diesem nach, indem die Autoren Verfahren in drei Herangehensweisen gliedern: datenzentrierte, beschreibungsbeachtende und beschreibungszentrierte Verfahren.

Dieses Klassifikationsschema ermöglicht gegenwärtig die Kategorisierung aller Cluster-Labeling-Verfahren. Im Folgenden wird jede der drei Herangehensweisen erläutert und zugehörige Verfahren hinsichtlich ihrer Labeling-Qualität miteinander verglichen.

5.1 Datenzentrierte Verfahren

Datenzentrierte Verfahren sind die ersten Verfahren, die zur Bezeichnung von Clustern eingesetzt wurden. Hierbei werden zunächst die Cluster einer Dokumentkollektion mittels traditioneller Clustering-Verfahren erzeugt. Im Anschluss werden mit Hilfe der gewonnenen Cluster-Centroiden[1] Terme ermittelt, die als Label für das jeweilige Cluster dienen. Centroiden sind als Metadokument eines Clusters zu verstehen, welches sich aus allen zugehörigen Dokumentrepräsentationen zusammensetzt. Ein Centroid ist somit eine Ansammlung verschiedener, nicht in Zusammenhang stehender Terme aus allen Dokumenten. Annahme datenzentrierter Verfahren ist nun, dass die *besten* Terme im Centroiden repräsentativ für das Cluster sind und sich somit als Label eignen.

Zur Bestimmung der k besten Terme werden neben bekannten Termgewichtungsverfahren wie *tf-idf* (siehe Kapitel 2.1.1) insbesondere traditionelle Merkmalsextraktionsverfahren aus dem Bereich der Textklassifikation eingesetzt. Dazu zählen informationstheoretische Maße wie *Mutual Information*. Diese wurden bereits in Kapitel 2.1.2 im Zusammenhang mit Schlüsselwortbestimmungsverfahren diskutiert.

Abbildung 5.1 verdeutlicht nochmals die sequentielle Abfolge bei der Erzeugung eines Labelings durch datenzentrierte Verfahren. Nachfolgend werden zwei Verfahren näher vorgestellt.

5.1.1 Frequent Predictive Words

Es ist trivial, die k am häufigsten vorkommenden Terme im Cluster als Label zu verwenden. Jedoch finden sich darunter viele unbedeutende Begriffe, ungeachtet dessen, dass zuvor gegebenenfalls Stoppworte entfernt wurden. Zudem ist nicht gewährleistet, dass Label eines Clusters tatsächlich repräsentativ für dieses sind. Bei der Formulierung wünschenswerter Label-Eigenschaften wurden in diesem Zusammenhang von der Trennschärfe eines Labels gesprochen.

Popescul u. Ungar (2000) argumentieren daher, dass bei der Merkmalsauswahl insbesondere Worte zu berücksichtigten sind, die sowohl häufig im Cluster vorkommen und

[1] Neben Centroiden sind ebenfalls Medoiden denkbar.

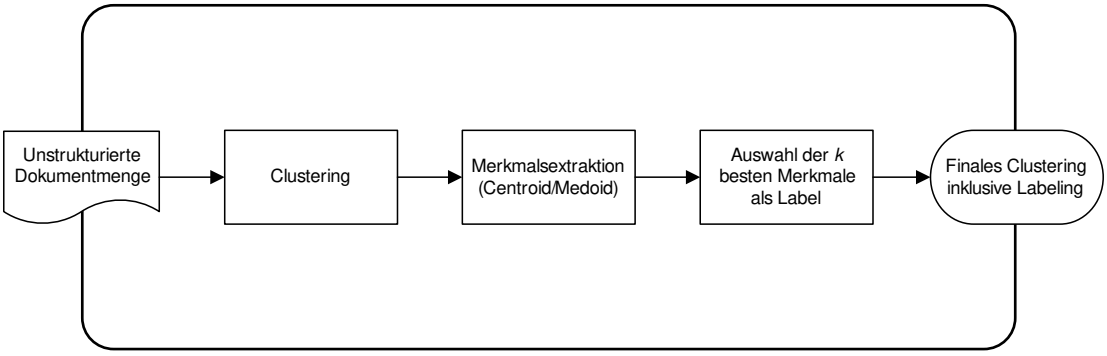

Abbildung 5.1 *: Zu durchlaufende Schritte zum Erzeugen eines Labelings anhand eines datenzentrierten Ansatzes.*

diskriminierend gegenüber anderen Clustern sind. Hierzu verwenden die Autoren zur Erzeugung des Labelings ein statistisches Modell, welches auf den Ausführungen von Yarowsky (1992) basiert. Yarowsky's Interesse gilt vorrangig der Auflösung sprachlicher Mehrdeutigkeiten (engl. *word sense disambiguation*). Zur Abgrenzung der verschiedenen Bedeutungen eines Terms ist Yarowsky an Termen interessiert, die in einer Domäne (hier ein einzelnes Cluster) signifikant häufiger vorkommen als in der gesamten Dokumentkollektion. Popescul u. Ungar übertragen die Idee der Merkmalsextraktion auf das Clustering, so dass sich ein Maß f_c zur Bewertung eines Terms t im Cluster c wie folgt formulieren lässt:

$$f_c(t) = P(t|c) \cdot \frac{P(t|c)}{P(t)},$$

wobei $P(t|c)$ die bedingte Wahrscheinlichkeit angibt, mit der ein Term t aus dem Cluster c stammt. Wie häufig t in der gesamten Dokumentkollektion vorkommt, zeigt $P(t)$ an. Die Wahrscheinlichkeiten sind mittels der Maximum-Likelihood-Methode zu schätzen. Sei $\mathrm{ctf}(t)$ die Termhäufigkeit eines Terms t in der Dokumentkollektion. Es gilt:

$$f_{c,\mathrm{MLE}}(t) = \mathrm{tf}_c(t) \cdot \frac{\mathrm{tf}_c(t)}{\mathrm{ctf}(t)}.$$

Wie gefordert präferiert der erste Faktor Terme, die häufig im Cluster auftreten. Der zweite Faktor favorisiert dagegen Terme, die im Schnitt häufiger im eigenen Cluster als in der gesamten Dokumentkollektion vorkommen. Damit grenzen sich Label verschiedener Cluster voneinander ab – die Trennschärfe der Cluster-Label ist sichergestellt.

Seien durch k-Means für das in Kapitel 2.1 eingeführte Beispiel zwei Cluster erzeugt worden. Es ergibt sich mittels des von Popescul u. Ungar vorgestellten Verfahrens folgendes Cluster-Labeling[2]:

Cluster 1: [computations, scale, large, singular]

D1: Large Scale Singular Value Computations

D2: Software for the Sparse Singular Value Decomposition

D5: Matrix Computations

D6: Singular Value Analysis for Cryptograms

Cluster 2: [information, retrieval, modern, introduction]

D3: Introduction to Modern Information Retrieval

D4: Linear Algebra for Intelligent Information Retrieval

D7: Automatic Information Organization

Stärken des Verfahrens

- Durch die Trennung von Clustering und Cluster-Labeling ist ein beliebiges Clustering-Verfahren einsetzbar.

Schwächen des Verfahrens

- Die Güte von Centroiden beeinflusst die Qualität von Cluster-Labeln.
- Das Verfahren generiert Label direkt aus der gegebenen Dokumentrepräsentation, was dazu führt, dass Terme, die auf ihren Wortstamm reduziert sind, für den Nutzer schwer verständlich sind.

Erfüllung von geforderten Label-Eigenschaften

- Das Verfahren von Popescul u. Ungar erfüllt durch das Maß f_c die Forderungen nach *Überdeckung* und *Trennschärfe* eines Labels.
- Es ist möglich, dass Label verschiedener Cluster identisch sind. Somit ist die *Eindeutigkeit* der Label nicht gewährleistet.

[2] Es werden vier Terme pro Cluster-Label betrachtet. Auf Stemming wird verzichtet.

5.1.2 Weighted Centroid Covering

Stein u. Meyer zu Eißen (2004) definieren Label-Eigenschaften formal. Der von Stein u. Meyer zu Eißen vorgestellte Algorithmus namens *Weighted Centroid Covering* wird anhand dieser Label-Eigenschaften erstellt und maximiert eine Auswahl dieser. Hierbei wird das Cluster-Label $\tau(c)$ im Gegensatz zur Verwendung in dieser Arbeit nicht durch eine Menge von Phrasen repräsentiert, sondern durch eine Termmenge. Es ergibt sich:

Eindeutigkeit Keine zwei Label von zwei verschiedenen Clustern verwenden einen Term gemeinsam:

$$\forall_{\substack{c_i,c_j \in \mathcal{C} \\ c_i \notin c_j}} \; : \; \tau(c_i) \cap \tau(c_j) = \emptyset$$

Überdeckung Das Label eines Clusters c enthält von jedem Dokument $d \in c$ mindestens einen Term $t \in d$. Es gilt

$$\forall_{c \in \mathcal{C}} \; \forall_{d \in c} \; : \; \tau(c) \cap T_d \neq \emptyset,$$

wobei T_d die Menge der Terme in Dokument d repräsentiert.

Ausdrucksstärke Die Terme im Label eines Clusters c gehören bezüglich der Dokumente in c zu den häufigsten:

$$\forall_{c \in \mathcal{C}} \; \exists_{t' \in \tau(c)} \; \forall_{d \in c} \; \forall_{\substack{t \in T_d \\ t \notin \tau(c)}} \; : \; tf_c(t) \leq tf_c(t')$$

Diskriminanz In dem Label eines Clusters c existiert ein Term t, dessen relative Häufigkeit in Bezug auf die Dokumentkollektion in c signifikant größer ist, als bei allen anderen Clustern:

$$\forall_{\substack{c_i,c_j \in \mathcal{C} \\ c_i \neq c_j}} \; \exists_{t \in \tau(c_j)} \; : \; \frac{\mathrm{tf}_{c_i}(t)}{|c_i|} \ll \frac{\mathrm{tf}_{c_i}(t)}{|c_j|}$$

Der Algorithmus erzeugt Cluster-Label in drei Schritten:

1. Es werden zunächst die k häufigsten Terme in jedem Cluster c identifiziert. Für diese Termmenge T wird ein Vektor $\mathcal{T}$ bestehend aus $k \times |T|$-Tupel $\langle t, \mathrm{tf}_{\kappa(w,i)}(w) \rangle$ mit $i \in \{1, \ldots, k\}$ erstellt. Sei $\kappa : T \times \{1, \ldots, |\mathcal{C}|\} \to \mathcal{C}$ eine Funktion mit $\kappa(t, i) = c$ gdw. der Term im Cluster c am i-t häufigsten vorkommt.

2. Die Elemente in $\mathcal{T}$ werden absteigend nach ihrer Termhäufigkeit sortiert.

3. Im letzten Schritt werden l verschiedene Terme jedem Cluster im Round-Robin-Verfahren zugewiesen. Da $\mathcal{T}$ nach der Termhäufigkeit sortiert ist, wird sichergestellt, dass die Terme, die die Dokumente im Cluster am stärksten überdecken, auch als Cluster-Label verwendet werden. Diese Überdeckungsannahme (engl. *coverage*) verleiht dem Verfahren seinen Namen.

Die vom *Weighted Centroid Covering*-Verfahren erzeugten Cluster-Label für das in Kapitel 2 eingeführte Beispiel unterscheiden sich geringfügig von denen des *Frequent Predictive Words*-Verfahrens:

Cluster 1: [singular, scale, large, decomposition]

D1: Large Scale Singular Value Computations

D2: Software for the Sparse Singular Value Decomposition

D5: Matrix Computations

D6: Singular Value Analysis for Cryptograms

Cluster 2: [information, retrieval, modern, introduction]

D3: Introduction to Modern Information Retrieval

D4: Linear Algebra for Intelligent Information Retrieval

D7: Automatic Information Organization

Stärken und Schwächen von Weighted Centroid Covering sind identisch mit Frequent Predictive Words.

5.1.3 Zusammenfassung

- Datenzentrierte Verfahren sind unabhängig vom Clustering, so dass jedes Clustering-Verfahren a posteriori durch ein Labeling zu ergänzen ist.
- Eine wortbasierte Repräsentation des Clusters in Form des Centroiden ist aus Sicht des Nutzers unzureichend, um daraus Cluster-Label zu erzeugen. Diese genügen nur schwer den Ansprüchen der *Verständlichkeit* an ein Label.
- Bezeichnend für datenzentrierte Verfahren ist, dass Cluster-Label aus Termmengen bestehen. Da ein Centroid nicht die Reihenfolge von Termen erhält und diese unabhängig voneinander aus dem Centroiden gewonnen werden, stehen Terme eines Cluster-Labels in den meisten Fällen semantisch nicht miteinander in Beziehung.

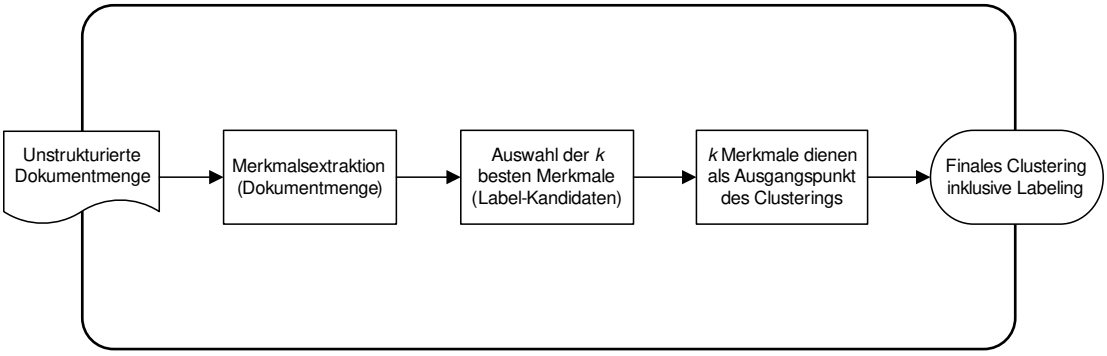

Abbildung 5.2 *: Zu durchlaufende Schritte zum Erzeugen eines Labelings anhand eines beschreibungsbeachtenden Ansatzes.*

5.2 Beschreibungsbeachtende Verfahren

Verfahren dieser Kategorie lösen das Labeling-Problem während des Clusterings. Ein für den Nutzer nachvollziehbares Clustering wird durch eine monothetische Merkmalsauswahl erreicht.

In der Literatur findet sich gegenwärtig ausschließlich ein Verfahren, welches diesem Cluster-Labeling Paradigma folgt: Suffixbaum-Clustering.

5.2.1 Suffixbaum-Clustering

Suffixbaum-Clustering wurde bereits im Kontext der Clustering-Verfahren in Kapitel 2.3.2 vorgestellt. Unter Verwendung eines Suffixbaumes wird das Labeling-Problem implizit gelöst, da häufige Suffixe direkt als Label verwendet werden.

Beim Suffixbaum-Clustering handelt es sich allerdings nicht um ein reines monothetisches Clustering. Der abschließende Schritt, bei dem gegenseitig sich stark überlappende Basiscluster iterativ zusammengeführt werden, ist polythetisch. Maslowska (2003) zeigen, dass das Clustering derart modifiziert werden kann, so dass es vollständig monothetisch ist.

Für das eingeführte Beispiel ergibt sich nun folgendes Cluster-Labeling:

Cluster 1: [Singular Value, Singular, Value]

D1: Large Scale Singular Value Computations

D2: Software for the Sparse Singular Value Decomposition

D5: Matrix Computations

D6: Singular Value Analysis for Cryptograms

Cluster 2: [Information, Information Retrieval, Retrieval]

D3: Introduction to Modern Information Retrieval

D4: Linear Algebra for Intelligent Information Retrieval

D7: Automatic Information Organization

Stärken des Verfahrens

- Suffixbaum-Clustering erzeugt Cluster-Label direkt aus dem Text und umgeht auf diese Weise das Problem datenzentrierter Verfahren, Cluster-Label aus mathematischen Modellen zu gewinnen.
- Das Verfahren weist Dokumente nur einem Cluster zu, wenn diese einen gemeinsamen Suffix aufweisen. Das Clustering ist somit partiell. Dokumente werden nicht berücksichtigt, die nicht aussagekräftig genug beschrieben werden können[3].
- Die Anzahl resultierender Cluster wird implizit festgelegt.
- Suffixbaum-Clustering ist bezüglich des Zeitbedarfs mit $\mathcal{O}(n)$ sehr effizient.
- Das Verfahren erzeugt ein überlappendes Clustering.

Schwächen des Verfahrens

- Das abschließende polythetisches Clustering führt dazu, dass Basiscluster zusammengeführt werden, die keine gemeinsamen Dokumente teilen. Ein exemplarisches Beispiel hierzu führt Meyer zu Eißen (2007) auf: Seien vier Basiscluster S_1 bis S_4 mit $S_1 = \{d_1, d_2, d_3\}$, $S_2 = \{d_2, d_3, d_4\}$, $S_3 = \{d_3, d_4, d_5\}$, $S_4 = \{d_4, d_5, d_6\}$ gegeben. Alle Dokumente $d_1, \ldots, d_6$ werden schrittweise zu einem einzigen Cluster vereinigt. Jedoch überlappen sich die Dokumente von S_1 und S_4 nicht, so dass nicht sichergestellt werden kann, dass diese ähnlich zueinander sind.
- Die Verwendung häufiger Phrasen als Cluster-Label stellt nicht sicher, dass diese verständlich sind.

[3] Ein unvollständiges Clustering wird traditionell als Schwäche eines Verfahrens angesehen. Im Zusammenhang mit Cluster-Labeling-Verfahren ist dies jedoch ein Vorteil.

- Die Qualität eines Clusterings und der Label ist abhängig von verschiedenen Faktoren: Wie viele Basiscluster werden für das Clustering verwendet? Nach welchem Kriterium werden die Basiscluster zusammengeführt? Wieviele Suffixe werden später als Cluster-Label dem Nutzer präsentiert?

Erfüllung von geforderten Label-Eigenschaften

- Die Bewertung eines Basisclusters bezieht sowohl die Anzahl der Dokumente als auch die Länge des Suffixes mit ein. Letzteres gewährleistet, dass längere Suffixe bevorzugt werden, die in vielen Dokumenten vorkommen. Beides optimiert sowohl die Eigenschaft *Überdeckung* und die Forderung nach *Verständlichkeit*.

- Jedes Basiscluster besitzt einen exklusiven Suffix, so dass zwei Cluster, die denselben Suffix als Cluster-Label besitzen, nicht vorkommen. Die Forderung nach *Eindeutigkeit* wird erfüllt.

5.2.2 Zusammenfassung

- Beschreibungsbeachtende Verfahren verzahnen Clustering und Labeling eng miteinander. In Folge dessen beeinflusst das Labeling im Gegensatz zu datenzentrierten Ansätzen das Clustering von Beginn an. Es ist somit nicht möglich, das Labeling mit einem beliebigen Clustering-Verfahren zu kombinieren.

5.3 Beschreibungszentrierte Verfahren

Verfahren dieser Kategorie verfolgen einen gegensätzlichen Ansatz zu datenzentrierten Verfahren. Durch das Paradigma *description comes first* rückt erstmals die Qualität der Cluster-Label in den Fokus. Gefragt ist nicht mehr zwingend ein optimales Clustering der Dokumente, sondern vielmehr aussagekräftige Cluster-Label. Dieses hat zur Folge, dass ein Cluster verworfen wird, falls kein passendes Label gefunden wird. Begründet wird dieser restriktive Schritt damit, dass ein Cluster ohne ausdrucksstarkes Label keinen Wert für den Nutzer besitzt.

Beschreibungszentrierte Verfahren werden aktuell ausschließlich beim Clustering von Suchergebnissen eingesetzt und arbeiten effizient auf kleinen Dokumentkollektionen (Carpineto u. a., 2009).

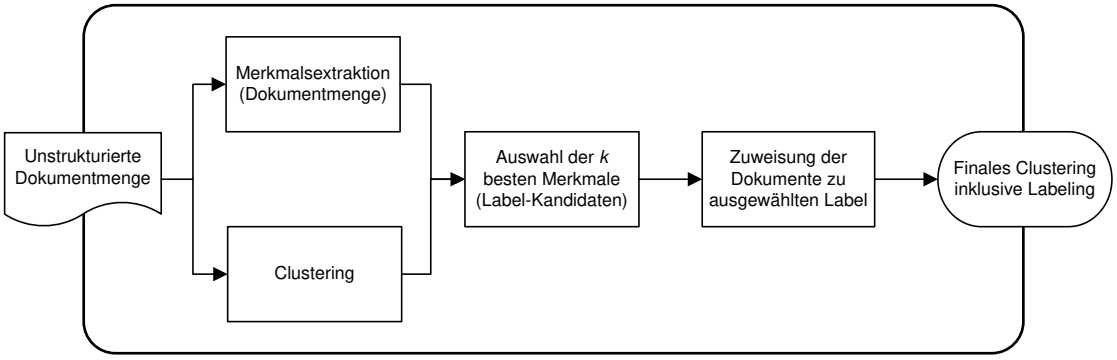

Abbildung 5.3 *: Zu durchlaufende Schritte zum Erzeugen eines Labelings anhand eines beschreibungszentrierten Ansatzes.*

Motivation für die Entwicklung beschreibungszentrierter Verfahren ist, dass das traditionelle Clustering zwar eine Kategorisierung von Dokumenten erzeugt, diese jedoch aufgrund des zugrundeliegenden mathematischen Modells von Nutzern nicht nachvollzogen werden kann. Charakteristisch für beschreibungszentrierte Verfahren ist daher, dass Clustering und Labeling zunächst getrennt voneinander betrachtet werden. Dadurch wird das Problem bisheriger Verfahren vermieden, Cluster-Label aus den Dokumentrepräsentationen zu ermitteln, die für das Clustering verwendet wurden. Das Labeling-Problem verschiebt sich somit von der Ermittlung von Termen aus einem mathematischen Modell, der Dokumentrepräsentation, hin zur Selektion geeigneter Phrasen direkt aus der Dokumentkollektion.

Bei beschreibungszentrierten Ansätzen dient das Clustering ausschließlich der Ermittlung von Themen in der Dokumentkollektion. Die resultierende Kategorisierung ist dabei nicht von Interesse und wird verworfen. Jedem der ermittelten Themen, beispielsweise repräsentiert durch einen Cluster-Centroiden, werden geeignete Label-Kandidaten zugewiesen. Diese bilden den Ausgangspunkt eines abschließenden monothetischen Clusterings der Dokumentkollektion. Jedes Cluster-Label repräsentiert also ein Cluster. Monothetisches Clustering gewährleistet, dass das Merkmal (hier Cluster-Label) in den Dokumenten des Clusters auftritt. Das Clustering ist überlappend.

Im Folgenden werden zwei Verfahren vorgestellt, die dem aufgezeigten Schema folgen. Die Vorgehensweise beschreibungszentrierter Verfahren ist zusammenfassend in Abbildung 5.3 dargestellt.

5.3.1 Lingo

Idee bei Lingo ist, mittels Latent-Semantic-Indexing (siehe Kapitel 2.1.1) die in der Dokumentkollektion verborgenen Konzepte aufzudecken. Von diesen abstrakten Konzepten wird angenommen, dass diese die Themen von Dokumenten beschreiben. Folglich repräsentiert jedes Konzept ein Cluster.

Eine linguistische Vorverarbeitung umfasst bei Lingo sowohl die Entfernung von Stoppworten als auch Stemming. Weiterhin werden nur Terme berücksichtigt, die eine benutzerdefinierte Häufigkeit in der Dokumentkollektion überschreiten. Die Autoren argumentieren, auf diese Weise im weiteren Verlauf bei der Ermittlung abstrakter Konzepte bedeutungslose Phrasen zu vermeiden. Für den Clustering-Schritt wird eine Term-Dokument-Matrix A benötigt.

Bei der Erstellung von A wird jeder Term anhand der *tf-idf*-Formel gewichtet. Osinski u. a. gewichten Terme zudem stärker, die in Titeln[4] vorkommen, da diese das Thema eines Dokuments stärker widerspiegeln. Da in dieser Arbeit Informationen über die Dokumentstruktur nicht berücksichtigt werden, wird auf diese zusätzliche Gewichtung verzichtet.

Die Dimensionsreduktion der Term-Dokument-Matrix auf einen neuen Rang k wird durch eine Singulärwertzerlegung erreicht (siehe Formel 2.1). k ist mit Bedacht zu wählen, da der neue, reduzierte Rang die Anzahl resultierender Cluster vorgibt. Osinski u. a. bestimmen k mit Hilfe der Frobenius-Norm (siehe Formel 2.2). Diesbezüglich ist die Approximationsqualität q anzugeben – ein weiterer benutzerdefinierter Parameter bei Lingo.

Parallel zur Singulärwertzerlegung sind Label-Kandidaten in der Dokumentkollektion zu ermitteln. Es werden dabei Phrasen ermittelt, die folgenden Bedingungen genügen:

- Eine Phrase soll dieselbe benutzerdefinierte Häufigkeit in der Dokumentkollektion überschreiten, die bereits bei der Erstellung von A für einzelne Terme gilt.
- Eine Phrase ist durch die Satzgrenze in ihrer Länge limitiert.
- Eine Phrase darf weder mit einem Stoppwort beginnen noch enden.

Ziel ist es nun, die durch die Singulärwertzerlegung entdeckten abstrakten Konzepte durch geeignete Phrasen zu beschreiben. Jedem Konzept ist die ähnlichste Phrase zuzuweisen. Diese entspricht dem endgültigen Cluster-Label. Hierzu sind alle ermittelten

[4] Für Internetseiten liegen unter anderem Informationen über den Titel vor.

Phrasen, die den zuvor aufgestellten Bedingungen genügen, in den LSI-Repräsentationsraum zu projizieren. Jede Phrase entspricht einem Pseudodokument (siehe Kapitel 2.1.1). Alle Phrasen werden gemeinsam durch eine Term-Dokument-Matrix P repräsentiert, wenngleich nun Phrasen die Dokumente beschreiben. Für die Projektion gilt

$$M = U_k^T P,$$

wobei die Terme in P ebenfalls mittels *tf-idf* gewichtet werden. Die Matrix U_k^T stellt die abstrakten Konzepte dar. Ein Spaltenvektor m_i der Ergebnismatrix M hält die Kosinusse der Winkel zwischen dem i-ten abstrakten Konzept und den Phrasen-Vektoren. Dabei ist die größte Komponente in m_i die ähnlichste Phrase zum Konzept i. Diese wird Cluster-Label von Cluster i.

Bislang ist offen, welche Dokumente den Clustern zuzuweisen sind. Hier führt Lingo einen monothetischen Clustering-Schritt durch. Es wird eine weitere Term-Dokument-Matrix Q benötigt, bei der die Dokumente durch die zuvor bestimmten Cluster-Label repräsentiert werden. Somit repräsentieren die Spaltenvektoren Cluster-Label. Q wird anschließend mit A_k multipliziert. A_k ist die auf den Rang k reduzierte ursprüngliche Term-Dokument-Matrix der Eingabedokumente. Es gilt

$$C = Q^T A_k.$$

Jede Zeile in C repräsentiert ein Cluster, die Spalten repräsentieren die ursprünglichen Eingabedokumente von A. Liegt eine Komponente des Zeilenvektors über einen benutzerdefinierten Schwellwert, wird das Dokument, welches zur Komponente gehört, dem Cluster zugewiesen. Ein Dokument kann dabei mehreren Clustern zugewiesen werden. Dokumente, die keinem Cluster zugeordnet werden konnten, werden einem separaten Cluster hinzugefügt.

Ein Beispiel (Osinski u. a., 2004): Für die Dokumente des in Kapitel 2.1 eingeführten Beispiels ergibt die Singulärwertzerlegung:

$$A = \begin{vmatrix} 0 & 0 & 0,56 & 0,56 & 0 & 0 & 1 \\ 0,49 & 0,71 & 0 & 0 & 0 & 0,71 & 0 \\ 0,49 & 0,71 & 0 & 0 & 0 & 0,71 & 0 \\ 0,72 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0,83 & 0,83 & 0 & 0 & 0 \end{vmatrix} \qquad U = \begin{vmatrix} 0 & 0,75 & 0 & -0,66 & 0 \\ 0,65 & 0 & -0,28 & 0 & -0,71 \\ 0,65 & 0 & -0,28 & 0 & 0,71 \\ 0,39 & 0 & 0,92 & 0 & 0 \\ 0 & 0,66 & 0 & 0,75 & 0 \end{vmatrix}$$

Es werden mittels einer Schlüsselwortbestimmung zwei Phrasen in der Dokumentkollektion ermittelt: *Singular Value* und *Information Retrieval*. Sei $q = 0,9$. Es folgt für die Abschätzung von $k = 0 \rightarrow q = 0,62$, $k = 1 \rightarrow q = 0,856$ und schließlich $k = 2 \rightarrow q = 0,956$. Nun kann $M = U_2^T P$ bestimmt werden:

$$P = \begin{vmatrix} 0 & 0,56 & 1 & 0 & 0 & 0 & 0 \\ 0,71 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0,71 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0,83 & 0 & 0 & 0 & 0 & 1 \end{vmatrix} \qquad M = \begin{vmatrix} 0,92 & 0 & 0 & 0,65 & 0,65 & 0,39 & 0 \\ 0 & 0,97 & 0,75 & 0 & 0 & 0 & 0,66 \end{vmatrix}$$

Da die Zeilen in M die Cluster repräsentieren und die größte Komponente des Zeilenvektors das korrespondierende Cluster-Label, ergeben sich zwei Cluster mit den Labeln *Singular Value* sowie *Information Retrieval*.

Es wird $C = Q^T A_k$ bestimmt:

$$Q = \begin{vmatrix} 0 & 0,56 \\ 0,71 & 0 \\ 0,71 & 0 \\ 0 & 0 \\ 0 & 0,83 \end{vmatrix} \qquad C = \begin{vmatrix} 0,69 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0,56 \end{vmatrix}$$

Sei der Schwellwert, bei dem ein Dokument einem Cluster zugewiesen wird, 0.5. Es ergibt sich folgendes Clustering:

Cluster 1: [Information Retrieval]

D3: Introduction to Modern Information Retrieval

D4: Linear Algebra for Intelligent Information Retrieval

D7: Automatic Information Organization

Cluster 2: [Singular Value]

D2: Software for the Sparse Singular Value Decomposition

D6: Singular Value Analysis for Cryptograms

D1: Large Scale Singular Value Computations

Cluster 3: [Sonstige Dokumente]

D5: Matrix Computations

Das Cluster *Sonstige Dokumente* nimmt Dokumente auf, die keinem Cluster zugewiesen wurden.

Stärken des Verfahrens

- Durch die Dimensionsreduktion der Term-Dokument-Matrix wird implizit durch den Parameter q die resultierende Cluster-Anzahl festgelegt.
- Es wird davon ausgegangen, dass durch die Anwendung von Latent-Semantic-Indexing alle verborgenen Konzepte (Themen) der Dokumente entdeckt werden.
- Lingo erzeugt ein überlappendes Clustering.

Schwächen des Verfahrens

- Der hohe Rechenaufwand von Latent-Semantic-Indexing schließt die Verwendung von Lingo für große Dokumentkollektionen mit langen Texten aus. Das Einsatzgebiet von Lingo beschränkt sich somit auf das Clustering von Suchergebnissen. Hier ist die Anzahl der zu gruppierenden Dokumente und deren Textlänge limitiert.
- Lingo verlangt benutzerdefinierte Parameter, die entscheidend die Qualität der Cluster-Label beeinflussen: (1) Wahl der Qualität q der Approximation der ursprünglichen Term-Dokument-Matrix, (2) Wahl einer Häufigkeitsschwelle zur Auswahl von Phrasen aus Texten, (3) Wahl eines Schwellwerts zur Auswahl von finalen Cluster-Labeln, (4) Wahl einer Ähnlichkeitsschwelle, die Dokumente bei der Zuweisung zu einem Cluster überschreiten müssen. Die Vielzahl der Parameter erschwert deren Konfiguration.
- Die Zuweisung der Dokumente zu Clustern erfolgt nicht monothetisch. Es werden auch Dokumente zu einem Cluster zugewiesen, die das Cluster-Label nur zu einem Teil enthalten. In Hinblick auf die Forderung nach Verständlichkeit und Transparenz eines Cluster-Labelings ist dies als Nachteil anzusehen.
- In Osinski u. a. (2004) werden häufige Phrasen als endgültige Cluster-Label verwendet. Somit ist nicht sichergestellt, dass diese verständlich sind.

Erfüllung von geforderten Label-Eigenschaften

- Werden bei der Ermittlung wichtiger Phrasen in der Dokumentkollektion initial Nominalphrasen anstatt häufiger Phrasen ermittelt, wird die *Verständlichkeit* der Cluster-Label garantiert.
- Sowohl *Trennschärfe* eines Clusters als auch die *Dokumentüberdeckung* innerhalb eines Clusters werden durch die Verwendung von Latent-Semantic-Indexing sichergestellt.

5.3.2 Descriptive k-Means

Descriptive k-Means ist die erste Realisierung des beschreibungszentrierten Ansatzes auf Basis eines traditionellen Clustering-Algorithmus (Stefanowski u. Weiss, 2007). Mittels k-Means sollen die Themen der Dokumentkollektion im Gegensatz zu Lingo effizient bestimmt werden. Annahme ist, dass jeder Cluster-Centroid ein Thema repräsentiert. Stefanowski u. Weiss sprechen von Konzept-Vektoren.

Descriptive k-Means folgt streng dem *description comes first*-Paradigma. Es werden zuerst sowohl häufige Phrasen mit Hilfe eines Suffixbaumes und Nominalphrasen in der gesamten Dokumentkollektion identifiziert. Phrasen mit einer Termlänge von vier werden als Cluster-Label bevorzugt. Hierbei wird angenommen, dass Label, die nur aus einem Term oder zu vielen Termen bestehen, nicht aussagekräftig genug sind. Zur Abwertung von Phrasen aufgrund ihrer Länge verwenden Stefanowski u. Weiss die folgende Funktion:

$$\text{penalty}(p) = \exp \frac{-(|p| - m)^2}{2 * d^2} \tag{5.1}$$

mit $m = 4$, $d = 8$, p die Phrase und $|p|$ dessen Länge in Worten.

Das Clustering wird mittels k-Means durchgeführt. k-Means wird hierzu unverändert eingesetzt. Als Termgewichtungsmodell für die Dokumentrepräsentationen wird *tf-idf* verwendet.

Wie bereits angesprochen dient das Clustering ausschließlich der Ermittlung von Themen, die hier durch Cluster-Centroiden formal repräsentiert werden. Beschreibungszentrierten Ansätzen entsprechend werden nun die zuvor ermittelten Phrasen bestimmt, die am ähnlichsten zu den Cluster-Centroiden sind.

Für die Ähnlichkeitsbestimmung sind alle Phrasen und Cluster-Centroiden in einem gemeinsamen Vektorraum zu repräsentieren. Jedem Centroiden wird eine Menge von

Label-Kandidaten T_c zugewiesen. Es gilt: $\forall c \in \mathcal{C}: \; T_c := \{p \mid p \in P \wedge \varphi(p, \mathbf{c}) \geq \theta\}$ mit θ als benutzerdefinierten Schwellwert. Die Menge T_c beschreibt das finale Cluster-Label.

Durch ein anschließendes monothetisches Clustering werden jedem Cluster-Label die Dokumente zugewiesen, die das Cluster-Label exakt enthalten. Hierbei kann es vorkommen, dass einem Cluster-Label wie *Information Retrieval* kein Dokument zugewiesen wird, da potenzielle Dokumente jeweils nur einen der beiden Terme enthalten. In der Folge beschreibt das Label keine Dokumentgruppe aussagekräftig genug, so dass das Cluster für dieses Label verworfen wird. Ebenso werden Cluster nicht berücksichtigt, denen weniger als fünf Dokumente zugewiesen werden konnten. Dokumente, die mit keinem Cluster assoziiert wurden, werden in einem gesonderten Cluster gruppiert. Somit erzeugt Descriptive k-Means im Resultat ein monothetisches, überlappendes Clustering, obwohl zuvor der polythetische, exklusive Clustering-Algorithmus k-Means verwendet wird.

Für das in dieser Arbeit eingeführte Beispiel ergibt sich nun folgendes Cluster-Labeling:

Cluster 1: [Sparse Singular Value Decomposition]

D2: Software for the Sparse Singular Value Decomposition

Cluster 2: [Cryptograms]

D6: Singular Value Analysis for Cryptograms

Cluster 3: [Modern Information Retrieval]

D3: Introduction to Modern Information Retrieval

Cluster 4: [Intelligent Information Retrieval]

D4: Linear Algebra for Intelligent Information Retrieval

Cluster 5: [Sonstige Dokumente]

D1: Large Scale Singular Value Computations

D5: Matrix Computations

D7: Automatic Information Organization

Stärken des Verfahrens

- Die Laufzeit von Descriptive k-Means ist abhängig von der Laufzeit von k-Means und beträgt somit $\mathcal{O}(KN)$. N sei die Anzahl der Dokumente. Damit ist dieses

Verfahren im Gegensatz zu Lingo effizient für große Dokumentkollektionen einsetzbar. Dazu trägt ebenso die Möglichkeit bei, sowohl das Clustering basierend auf k-Means und die Identifikation wichtiger Phrasen im Text parallel auszuführen.

- Descriptive k-Means erzeugt ein überlappendes Clustering.

Schwächen des Verfahrens

- Themen innerhalb einer Dokumentkollektion werden initial durch Centroiden repräsentiert. In der Folge ist die optimale Anzahl der Themen durch Variation von k bei k-Means zunächst zu ermitteln. Dieses leistet Descriptive k-Means nicht.
- Die Festlegung auf ein festes k garantiert nicht, dass am Ende auch k Cluster erzeugt werden.
- Da Cluster-Label durch räumliche Nähe zum Centroiden bestimmt werden, ist die Qualität der Label von der Güte der Centroiden abhängig. Treeratpituk u. Callan (2006) zeigten dies.
- Descriptive k-Means verlangt die Festlegung von benutzerdefinierten Parametern: (1) Anzahl zu identifizierender Cluster bei k-Means, (2) Auswahl von Kandidaten als Cluster-Label durch Überschreitung eines Ähnlichkeitswertes zu Termen im Centroiden, (3) Wahl der optimalen Länge eines Cluster-Labels. Dies erschwert, ähnlich wie bei Lingo, die optimale Konfiguration des Algorithmus.

Erfüllung von geforderten Label-Eigenschaften

- Descriptive k-Means bewertet jedes Cluster anhand der Länge der zugewiesenen Phrase und der Anzahl der Dokumente im Cluster. Dies ist ähnlich zur Bewertung der Basiscluster beim Suffixbaum-Clustering und optimiert somit die Eigenschaften *Überdeckung* und *Verständlichkeit*. Die *Verständlichkeit* wird zudem durch die Festlegung auf Nominalphrasen als Cluster-Label sichergestellt.
- Die Trennschärfe wird erfüllt, da Centroiden als Ausgangspunkt zur Ermittlung von Cluster-Labeln dienen. Centroiden repräsentieren dabei die Themen der Dokumentkollektion.

5.3.3 Zusammenfassung

- Beschreibungszentrierte Verfahren führen ein monothetisches Clustering durch.
- Cluster-Label bestehen im Vergleich zu bisher betrachteten Verfahren aus nur einer einzigen Phrase.
- Es ist ein beliebiges Clustering-Verfahren einsetzbar.

6 Topical k-Means

Die Diskussion bisheriger Cluster-Labeling-Verfahren zeigt, dass kein Verfahren alle geforderten Label-Eigenschaften erfüllt. Daher wird ein neues Verfahren vorgestellt. Dieses optimiert folgende Eigenschaften:

- Eine hohe Überdeckung eines Cluster-Labels wird durch die Auswahl der häufigsten Nominalphrasen in der Dokumentkollektion erfüllt.
- Die Trennschärfe eines Cluster-Labels wird durch die Gewichtung von Phrasen mittels Informativeness erfüllt.
- Die Verständlichkeit eines Cluster-Labels wird durch die ausschließliche Berücksichtigung von Nominalphrasen sichergestellt.
- Die Redundanz in der Auswahl der Cluster-Label wird reduziert, indem nur die allgemeinsten und spezifischsten Phrasen angezeigt werden.

Im Folgenden wird der Clustering-Prozess und die Ermittlung passender Cluster-Label erläutert. Topical k-Means unterteilt sich in folgende Schritte:

1. Ermittlung von häufigen Nominalphrasen in der Dokumentkollektion.
2. Repräsentation von Textdokumenten durch Nominalphrasen.
3. Clustering der repräsentierten Textdokumente.
4. Ermittlung der k_c besten Phrasen für jedes Cluster.
5. Gewichtung der ausgewählten Phrasen mittels Informativeness.
6. Die k_l besten ausgewählten Phrasen werden Cluster-Label.
7. Zuweisung von Dokumenten zu den Cluster-Labeln.

6.1 Clustering einer Dokumentkollektion

Den beschreibungszentrierten Ansätzen folgend werden zunächst Nominalphrasen aus der Dokumentkollektion ermittelt. Ein Clustering wird allerdings nicht parallel durchgeführt. Stattdessen werden die k besten Nominalphrasen zur Repräsentation der Dokumente verwendet. Folgende Ziele werden dabei verfolgt:

- Die Integration der Nominalphrasen in den Clustering-Prozess führt dazu, dass in den Centroiden bereits die Nominalphrasen enthalten sind, die das Cluster am stärksten repräsentieren. Diese Vorgehensweise ist von datenzentrierten Ansätzen übernommen, wobei diese das Problem besitzen, das Terme des Centroiden unabhängig voneinander sind. Durch die Verwendung von Phrasen bleibt hier die ursprüngliche Reihenfolge der Terme erhalten.

- Ein Problem von Descriptive k-Means und Lingo ist, nachdem Clustering-Schritt die besten Schlüsselphrasen jedem Cluster zuzuweisen. Dieser Schritt entfällt für Topical k-Means aufgrund der Repräsentation von Textdokumenten durch Nominalphrasen.

- Die Verwendung von Nominalphrasen anstatt Termen führt zu einer Reduktion des Vektorraumes. Dies resultiert in einer effizienteren Berechnung von Dokumentähnlichkeiten beim Clustering.

Das Clustering wird mittels k-Means durchgeführt. k entspricht der Anzahl an Themen in einer Dokumentkollektion. Es wird das Problem mit Descriptive k-Means geteilt, k zuvor optimal zu bestimmen.

Die durch Nominalphrasen repräsentierten Textdokumente sind mittels eines Termgewichtungsmodells zu gewichten. Neben dem *tf-idf*-Modell bietet sich das in Kapitel 2.1.1 motivierte *tf-pdf*-Modell an, um Phrasen zu bevorzugen, die in vielen Dokumenten vorkommen. Experimente zeigen allerdings, dass die Qualität des Clusterings durch letztgenanntes Modell abnimmt. Daher wird eine Gewichtung mit *tf-idf* vorgezogen. Es werden Stoppworte aus den Nominalphrasen entfernt und Stemming durchgeführt[1].

[1] Damit die ursprünglichen Nominalphrasen zum Labeling von Clustern verwendet werden können, müssen diese gespeichert werden.

6.2 Ermittlung von Cluster-Labeln

Eine Zuweisung von Nominalphrasen zu den in der Dokumentkollektion entdeckten Themen entfällt, da diese bereits in den Centroiden hinterlegt sind. Jeder Centroid repräsentiert ein Thema. Die k_c besten Nominalphrasen eines Centroiden werden erneut anhand folgender Maße gewichtet:

1. Bestrafung von zu kurzen und zu langen Nominalphrasen durch die in Formel 5.1 definierte Funktion (siehe Descriptive k-Means).

2. Berechnung der Informativeness einer Nominalphrase.

Informativeness einer Nominalphrase

Zur Ermittlung repräsentativer Phrasen für ein Cluster sind diese gegenüber anderen Clustern abzugrenzen. Dies motivieren die Label-Eigenschaften Trennschärfe und minimale Überlappung. Die Berechnung der Informativeness wurde bereits in Kapitel 2.1.2 eingeführt. Zur Erinnerung:

Informativeness Die Informativeness eines Terms t sagt aus, wie viel Information verloren wird, wenn fälschlicherweise angenommen wird, der Term würde durch das Sprachmodell des Hintergrund-Korpus modelliert: $\varphi_i := \delta_t(LM^N_{\text{fg}} || LM^N_{\text{bg}})$.

Das Cluster, für welches potenzielle Cluster-Label anhand der Informativeness bewertet werden sollen, wird als Vordergrund-Korpus *fg* definiert. Dokumente, die nicht im Cluster vorkommen, bilden den Hintergrund-Korpus *bg*. Der Informationsverlust bei Phrasen ist besonders hoch, die schlecht oder gar nicht durch den Hintergrund-Korpus modelliert werden. Diese eignen sich als Cluster-Label.

Zur Berechnung der Informativeness werden Sprachmodelle der Ordnung 3 eingesetzt.

6.3 Erstellung von Clustern

Die jeweils k_l besten Nominalphrasen mit der höchsten Informativeness sind potenzielle Cluster-Label für ein Cluster. Für jede Nominalphrase werden die Dokumente bestimmt, die diese enthalten. Werden weniger als fünf Dokumente gefunden, wird die Nominalphrase verworfen. In allen anderen Fällen wird ein Cluster gebildet mit der Nominalphrase als Cluster-Label.

Die Anzahl resultierender Cluster wird durch einen Schwellwert begrenzt. Dieser gibt in Prozent an, wie viele Dokumente eines Clusters abgedeckt werden sollen. Ist der Schwellwert zu niedrig festgelegt, ist es möglich, dass bestimmte Themen in Dokumentkollektionen nicht erkannt werden. Ist der Schwellwert dagegen zu hoch, besteht die Gefahr, zu viele Cluster zu erzeugen.

6.4 Filterung von Phrasen

Die Qualität eines Cluster-Labels ist mittels folgender Heuristiken zu verbessern.

Einem Nutzer sollten nur die *allgemeinsten* und *spezifischsten* Phrasen als Cluster-Label angezeigt werden. Die allgemeinste Phrase ist eine, für die keine andere Phrase existiert, die eine Teilzeichenfolge (engl. substring) dieser ist. Die spezifischste Phrase ist eine, die in keiner anderen Phrase als Teilzeichenfolge auftritt.

Seien *Information Retrieval*, *Information* und *Retrieval* potenzielle Cluster-Label. *Information Retrieval* ist die spezifischste Phrase, die anderen beiden die allgemeinsten. Die spezifischste Phrase ist bereits aussagekräftig genug, da sowohl *Information* und *Retrieval* in dieser enthalten sind. Allgemeinste Phrasen sollten deshalb nur angezeigt werden, wenn keine spezifischere Phrase mit ähnlicher Dokumentüberdeckung existiert.

Die genannten Verbesserungsvorschläge für die Aufbereitung der Label sind für alle Cluster-Labeling-Verfahren gültig. Diese sind in einem Nachbearbeitungsschritt am Ende des Cluster-Labelings ausführbar.

Da im Englischen der Plural für viele Worte durch Anfügen des Buchstabens „s“ gebildet wird, bildet die Singularform anhand obiger Betrachtungen die allgemeinste Phrase. Der Plural ist in diesem Fall spezifischer als der Singular. Bei ähnlicher Dokumentüberdeckung wird somit die Singularform verworfen und der Plural bevorzugt. Somit werden implizit Pluralformen der Terme als Cluster-Label verwendet.

Topical k-Means erzeugt für das der Arbeit begleitende Beispiel folgendes Cluster-Labeling:

Cluster 1: [Sparse Singular Value Decomposition]

D2: Software for the Sparse Singular Value Decomposition

Cluster 2: [Intelligent Information Retrieval]

D4: Linear Algebra for Intelligent Information Retrieval

Cluster 3: [Cryptograms]

D6: Singular Value Analysis for Cryptograms

Cluster 4: [Modern Information Retrieval]

D3: Introduction to Modern Information Retrieval

Cluster 5: [Sonstige Dokumente]

D1: Large Scale Singular Value Computations

D5: Matrix Computations

D7: Automatic Information Organization

Stärken des Verfahrens

- Topical k-Means erzeugt ein monothetisches, überlappendes Clustering.
- Die Integration von Nominalphrasen in den Clustering-Prozess löst implizit das Problem, passende Cluster-Label für die ermittelten Themen zu finden.
- Das nachfolgende Ranking potenzieller Cluster-Label durch die Informativeness bevorzugt Label, die ein Cluster stärker repräsentieren als andere.
- Für das Clustering von Textdokumenten ist ein beliebiges Verfahren einsetzbar.

Schwächen des Verfahrens

- Die resultierende Anzahl an Clustern hängt vom Schwellwert ab, der die maximale Überdeckung von Dokumenten eines Clusters festlegt.

7 Evaluierung von Cluster-Labeling-Verfahren

Neben der Erzeugung von Cluster-Labeln stellt die Validierung dieser eine der großen Herausforderungen im Bereich des Cluster-Labelings dar.

Ein Cluster-Labeling ist auf verschiedene Arten zu bewerten. Dazu gehören die Güte der Cluster-Label, die Effizienz des Verfahrens als auch die Qualität des Clusterings. Diese Arbeit beschränkt sich auf die Validierung der Güte von Cluster-Labeln, da diese die wichtigste Repräsentation für den Nutzer darstellen.

Für die Mehrheit der Cluster-Labeling-Verfahren ist das zugrundeliegende Clustering-Verfahren frei wählbar, so dass von einer Bewertung der Effizienz eines Verfahrens und der Clustering-Qualität abgesehen wird. Ein Vergleich zwischen dem Cluster-Labeling-Verfahren Descriptive k-Means und dem traditionellen Clustering-Algorithmus k-Means findet sich in Weiss (2006). Es wird gezeigt, dass die Qualität des Clusterings, welches durch Descriptive k-Means erzeugt wird, schlechter ist.

Dieses Kapitel gibt einen Überblick über bestehende und neue Validierungsmaße. Diese gehen von einer nach Relevanz sortierten Liste von Cluster-Labeln aus, die ein Cluster-Labeling-Verfahren für einen Cluster erstellt.

Empirische Nutzerstudien

Empirische Nutzerstudien finden im Bereich des Cluster-Labelings breite Unterstützung (Sanderson u. Croft, 1999; Zamir u. Etzioni, 1999; Popescul u. Ungar, 2000; Lawrie u. Croft, 2003; Maslowska, 2003; Kummamuru u. a., 2004; Ferragina u. Gullì, 2007). In dieser Arbeit wird auf Nutzerstudien verzichtet, da diese folgende Nachteile besitzen:

- Zeitaufwändige Vorbereitung von Umfragen und Durchführung von Experimenten.
- Nicht-Reproduzierbarkeit der Experimente, so dass die Ergebnisse nicht durch Dritte nachvollziehbar sind.

- Bewertungen von Testpersonen sind subjektiv. Diese werden zudem von vielen Umgebungsfaktoren wie dem Vorwissen der Testperson bezüglich des Experiments beeinflusst.
- Geringe Anzahl an Testpersonen führt dazu, dass die Ergebnisse nicht verallgemeinerbar sind.

Anstatt einer empirischen Bewertung eines Cluster-Labelings wird in dieser Arbeit eine qualitative Bewertung durchgeführt.

Überblick

Am Ende eines Cluster-Labelings interessieren folgende Fragestellungen.

1. Wie lässt sich die Güte eines Cluster-Labels unter Zuhilfenahme einer externen Referenz messen?
2. Wie lässt sich die Güte eines Cluster-Labels allein auf Grundlage der vorliegenden Daten messen?
3. Wie lässt sich entscheiden, welches von zwei erzeugten Cluster-Labelings besser ist?

Um Fragestellung 1 zu beantworten, werden in Kapitel 7.1 externe Gütekriterien für das Cluster-Labeling vorgestellt. Ist keine externe Referenz zur Hand, sind interne Validierungsmaße zu verwenden. Um Fragestellung 2 zu beantworten, werden in Kapitel 7.2 interne Gütekriterien für das Cluster-Labeling vorgestellt. Zur Beantwortung der Fragestellung 3 sind interne und externe Gütekriterien anwendbar.

7.1 Externe Gütekriterien zur Validierung von Cluster-Labeln

Das Cluster-Labeling Problem wird als Ranking-Problem aufgefasst. Liefere ein Retrieval-System, beispielsweise eine Suchmaschine, für eine Suchanfrage eine Menge relevanter Dokumente zurück. Die Herausforderung besteht darin, eine optimale Reihenfolge der Dokumente zurückzugeben, bei der die Dokumente der Relevanz nach absteigend sortiert sind. Das Dokument mit der höchsten Relevanz zur Anfrage steht an erster Stelle.

Übertragen auf das Cluster-Labeling wird die Anfrage durch das Referenz-Label vorgegeben. Die Menge relevanter Dokumente für ein Cluster wird durch die Menge der

Phrasen definiert, die das tatsächliche Cluster-Label bilden. Idealerweise ist die erste Phrase des Cluster-Labels identisch mit dem Referenz-Label. Durch entsprechende Maße kann nun die Güte gemessen werden, mit der ein erzeugtes Cluster-Label das Referenz-Label wiedergibt.

Maße, welches dieses berücksichtigen, sind: *Precision@R*, *Match@R* und *Mean Reciprocal Rank* (MMR). Annahme hierbei ist, dass ein Nutzer sich ausschließlich die ersten R Resultate anschaut.

7.1.1 Precision@R, Match@R und MRR

Treeratpituk u. Callan (2006) präsentieren erstmals ein Framework zur Evaluierung von Cluster-Labeln. Als Referenzkorpus dient ein Auszug des Open Directory Projects. Das Cluster-Labeling Problem wird ebenfalls als Ranking-Problem aufgefasst. Dieses spiegelt sich in den vorgestellten externen Maßen wider. Treeratpituk u. Callan vergleichen die automatisch erzeugten Cluster-Label mit dem korrekten Kategorienamen aus dem Open Directory Project. Diese dienen als Referenz-Label für das jeweilige Cluster.

Treeratpituk u. Callan verwenden folgende Validierungsmaße[1] zur Bewertung der Qualität von Cluster-Labeln:

Precision@R: Das Maß berechnet die *Precision* (Genauigkeit), also den Anteil der korrekten Label unter den ersten R Resultaten. Die Anzahl der als korrekt eingestuften Label wird durch R geteilt.

Match@R: Das Maß gibt an, ob innerhalb der besten R Resultate ein korrektes Label enthalten ist. Mit zunehmenden R wird dieser Wert größer. Es handelt sich um ein binäres Maß.

Mean Reciprocal Rank (MRR): Vom Rang i des ersten korrekten Labels wird der Kehrwert gebildet. Das Maß ist das Mittel der Kehrwerte für alle Cluster. Ist kein korrektes Label vorhanden, so wird der Kehrwert mit Null definiert.

$$\text{MMR} := \frac{1}{|\mathcal{C}|} \sum_{i=1}^{|\mathcal{C}|} \frac{1}{\text{Rang}_i}$$

Da der Algorithmus eine Label-Hierarchie erzeugt, werden die Referenz-Label des aktuellen Clusters S und des Vater-Clusters P berücksichtigt, wenn die Übereinstimmung

[1] Die Maße wurden ursprünglich in Radev u. a. (2002b) vorgestellt.

mit dem Referenz-Label überprüft wird. Ein Label L gilt als korrekt, wenn es entweder eine exakte oder partielle Übereinstimmung mit S und P gibt. Eine exakte Übereinstimmung liegt vor, wenn L gleich der Zeichenfolge S, SP oder PS ist. Dagegen liegt eine partielle Übereinstimmung bereits vor, wenn L mindestens einen Term mit S, SP oder PS gemein hat. Ein Beispiel:

Exakte Übereinstimmung Die Label *Health*, *Pharmacy*, *Drugs and Medications* sowie *Antibiotics* stimmen exakt mit der Referenzkategorie *Health/Pharmacy/Drugs and Medications/Antibiotics* überein.

Partielle Übereinstimmung Die Label *Medications*, *Drugs* und *Antibiotics Health* stimmen partiell mit der Referenzkategorie *Health/Pharmacy/Drugs and Medications-/Antibiotics* überein.

Anstelle von L sind ebenfalls Synoyme von L zugelassen. Diese werden in der Arbeit von Treeratpituk u. Callan mittels *WordNet* ermittelt.

Da in dieser Arbeit keine hierarchischen Informationen vorliegen, kann ausschließlich das aktuelle Cluster S zur Überprüfung der Übereinstimmung berücksichtigt werden. Zudem erscheint es nicht sinnvoll, beispielsweise für die Kategorie Vertigo auch die Vaterkonzepte „Film“ oder „Kunst“ zu berücksichtigen. Diese sagen nichts über die eigentliche Kategorie aus, das sie zu allgemein gehalten sind. In dieser Arbeit wird ausschließlich die speziellste Kategorie S als Referenz herangezogen.

7.1.2 Normalized Discounted Cumulative Gain (NDCG)

Nachteil der bisher vorgestellten Maße ist, dass diese nicht sensitiv für die Reihenfolge relevanter Dokumente in der Ergebnismenge sind. Die Reihenfolge von Phrasen eines Cluster-Labels ist jedoch entscheidend, da ein Nutzer möglichst durch Lesen der ersten Phrase eine Ahnung davon bekommen soll, welche Dokumente sich im Cluster befinden. Es ist für die Qualität eines Cluster-Labelings schlechter, wenn eine relevante Phrase eines Cluster-Labels erst auf einem der hinteren Ränge aufgeführt wird.

Ein Maß, welches sensitiv gegenüber der Reihenfolge ist, ist *Normalized Discounted Cumulative Gain* (Järvelin u. Kekäläinen, 2002). Grundlegende Ideen sind:

- Je höher ein relevantes Dokument im Ranking positioniert ist, desto besser ist das Ranking.
- Relevante Dokumente sollten höher positioniert werden als nicht relevante Dokumente.

Beim *Cumulative Gain* wird davon ausgegangen, dass die binäre Relevanzbewertung eines Dokuments zu streng ist, so dass verschiedene Relevanzbewertungen rel_i für ein Dokument an Position i im Ranking unterstützt werden. Im einfachen binären Fall ist $rel_i \in \{0, 1\}$ mit 0 gleich *nicht relevant* und 1 gleich *relevant*. Für die Berechnung des Cumulative Gain bis zur Position R ergibt sich:

$$\text{CG@}R = \sum_{i=1}^{R} rel_i$$

Der Name des Maßes leitet sich wie folgt ab: Jedes relevante Dokument besitzt einen bestimmten Beitrag aufgrund seiner Relevanz zur Anfrage. Je höher der Beitrag, desto höher ist der Informationsgewinn (engl. *gain*) für den Nutzer. Das Maß misst dabei den gesamten Informationsgewinn aller Dokumente bis zur Position R durch Bildung der Summe der Einzelbeiträge (engl. *cumulative*).

Die Grundannahme, dass ein sehr relevantes Dokument möglichst zu Beginn des Rankings aufgeführt wird, erfüllt Cumulative Gain nicht. Der Beitrag eines jeden Dokuments fließt unabhängig von dessen Position in die Summe ein. Um dies dennoch zu gewährleisten, ist zusätzlich der Beitrag eines schlechter positionierten Dokuments zu mindern. Der Beitrag ist durch eine entsprechende Funktion, beispielsweise durch eine logarithmische Funktion, zu reduzieren. *Discounted* Cumulative Gain (DCG) wird wie folgt definiert:

$$\text{DCG@}R = \sum_{i=1}^{R} \frac{2^{rel_i} - 1}{\log_2(1 + i)}$$

Neben der Logarithmusfunktion sind ebenso weitere Funktionen wie eine Linearfunktion oder eine Wurzelfunktion denkbar, um den Beitrag schlechter positionierter Dokumente zu mindern.

Da verschiedene Anfragen eine unterschiedliche Anzahl an Resultaten zurückliefern, eignet sich Discounted Cumulative Gain nicht, um verschiedene Auswertungen untereinander zu vergleichen. Das Maß ist in der Folge zu normieren. Eine obere Schranke zum DCG bildet die ideale Reihenfolge der Dokumente, bei der diese anhand der Relevanz absteigend sortiert sind (iDCG). Es ergibt sich somit für die Berechnung des *Normalized* Discounted Cumulative Gain:

$$\text{NDCG@}R = \frac{\text{DCG@}R}{\text{iDCG@}R}$$

Das Maß wird 1, wenn das Cluster-Label relevant ist und alle Phrasen des Cluster-Labels anhand ihrer Relevanz ideal sortiert sind. Das Maß wird 0, falls im Cluster-Label keine einzige relevante Phrase existiert.

Relevanzstufe	Definition
0	keine Übereinstimmung mit dem Referenz-Label
1	partielle Übereinstimmung mit dem Referenz-Label
2	exakte Übereinstimmung mit dem Referenz-Label

Tabelle 7.1 *: Relevanzstufen eines Cluster-Labels für die Verwendung bei Normalized Discounted Cumulative Gain.*

Es stellt sich die Frage, welche Relevanzstufe einer Phrase eines Cluster-Labels zugesprochen werden sollte. Im binären Fall ist ein Cluster-Label relevant, wenn es mit dem Referenz-Label übereinstimmt. In allen anderen Fällen ist es nicht relevant. Diese Unterscheidung ist allerdings zu strikt und wird dem Cluster-Labeling nicht gerecht.

Andererseits ließe sich die von Treeratpituk u. Callan aufgestellte Unterscheidung zwischen exakter und partieller Übereinstimmung einer Phrase mit dem Referenz-Label zu einem Maß zusammenfassen. Die Relevanzstufen für NDCG sind hierfür in Tabelle 7.1 aufgeführt. Auf diese Weise ist die Bewertung eines Cluster-Labels ebenfalls von der Reihenfolge der Phrasen abhängig.

7.2 Interne Gütekriterien zur Validierung von Cluster-Labeln

Die in Kapitel 4.2 geforderten Label-Eigenschaften sind quantifizierbar. Ein Cluster-Labeling lässt sich somit anhand dieser Maße bewerten. Kapitel 7.2.1 stellt diese vor.

Im zweiten Abschnitt dieses Kapitels wird ein neues internes Validierungsmaß basierend auf Normalized Discounted Cumulative Gain (NDCG) vorgestellt.

7.2.1 Quantifizierung wünschenswerter Label-Eigenschaften

Die in Kapitel 4.2 geforderten Label-Eigenschaften werden quantifiziert:

Verständlichkeit: Die Funktion $f_1(\tau)$ nimmt den Wert 1 an, wenn jede Phrase eines Cluster-Labels eine Nominalphrase ist und zudem nah an der gewünschten Termlänge $|p|_{\text{opt}} = 4$ liegt. Die erste Forderung prüft die Funktion $\text{NP}(p)$, die zweite Forderung wird durch die Funktion $\text{penalty}(p)$ aus Kapitel 2.3.2 überprüft.

Die Auswertung von Wikipedia in Kapitel 3 zeigte, dass alleinstehende Terme aufgrund des fehlenden Kontextes weniger verständlich sind als längere Phrasen. Phrasen, die nur aus einem Term bestehen, werden deshalb bei flachen, partitionierenden

Clustering-Verfahren stärker abgewertet. Es gilt:

$$f_1(\tau) = \frac{1}{k} \sum_{c \in \mathcal{C}} \frac{1}{|\tau(c)|} \sum_{p \in \tau(c)} \text{penalty}(p) \cdot \text{NP}(p) \tag{7.1}$$

mit

$$\text{penalty}(p) := \begin{cases} \exp \frac{-(|p| - |p|_{\text{opt}})^2}{2 \cdot d^2} & \text{, wenn } |p| = 1 \\ 0,5 & \text{, sonst} \end{cases}$$

$$\text{NP}(p) := \begin{cases} 1 & \text{, wenn } p \in L(G) \\ 0 & \text{, sonst} \end{cases}$$

Überdeckung: Je näher die Funktion $f_2(\tau)$ bei 1 liegt, desto besser überdeckt das Cluster-Label die Dokumente des dazugehörigen Clusters. Ein Wert nahe 0 zeigt an, dass die Phrasen eines Cluster-Labels in keinem oder nur in wenigen Dokumenten eines Clusters vorkommen. Sei P_c die Menge der Phrasen in Cluster c. Es gilt:

$$f_2(\tau) = 1 - \frac{1}{k} \sum_{c \in \mathcal{C}} \operatorname*{argmin}_{p' \in \tau(c_j)} \frac{1}{|P_c \setminus \tau(c)|} \sum_{\substack{p \in P_c \\ p \notin \tau(c)}} \frac{\text{df}_c(p)}{\text{df}_c(p')} \tag{7.2}$$

Trennschärfe: Je näher die Funktion $f_3(\tau)$ bei 1 liegt, desto stärker ist die Diskriminanzkraft der Cluster-Label. Der Funktionswert kann negativ werden, wenn die Phrasen eines Cluster-Labels schlecht gewählt sind. Es gilt:

$$f_3(\tau) = 1 - \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \operatorname*{argmin}_{p \in \tau(c_j)} \frac{|c_j| \cdot \text{df}_{c_i}(p)}{|c_i| \cdot \text{df}_{c_j}(p)} \tag{7.3}$$

Minimale Überlappung: Je näher die Funktion $f_4(\tau)$ bei 1 liegt, desto eindeutiger werden die Dokumente eines Clusters durch das Cluster-Label beschrieben. Es gilt:

$$f_4(\tau) = 1 - \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \operatorname*{argmin}_{p \in \tau(c_j)} \frac{|c_i(p) \cap c_j(p)|}{|c_i(p) \cup c_j(p)|} \tag{7.4}$$

Eindeutigkeit: Die Funktion $f_5(\tau)$ nimmt den Wert 1 an, wenn alle Phrasen aller Cluster-Label unterschiedlich sind. Je näher der Wert bei 0 liegt, desto mehr überschneiden sich die Cluster-Label verschiedener Cluster. Es gilt:

$$f_5(\tau) = 1 - \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{|\tau(c_i) \cap \tau(c_j)|}{|\tau(c_i) \cup \tau(c_j)|} \tag{7.5}$$

Redundanzfreiheit: Die Funktion $f_6(\tau)$ nimmt den Wert 1 an, wenn alle Phrasen eines Cluster-Labels nicht synonym zueinander sind. Je mehr Synonyme auftreten, desto näher liegt der Wert bei 0. Es gilt:

$$f(\tau) = 1 - \frac{1}{k} \sum_{c \in \mathcal{C}} \frac{2}{|\tau(c)| \left(|\tau(c)| - 1\right)} \sum_{i=1}^{|\tau(c)|-1} \sum_{j=i+1}^{|\tau(c)|} \mathrm{syn}(p_i, p_j)$$

mit syn: $p \times p \mapsto \{0, 1\}$.

Die Maße sind voneinander unabhängig für ein Cluster-Labeling bestimmbar. Bislang ist unklar, welche Maße besonders wichtig für die Bewertung eines Cluster-Labelings sind. Die Maße f_1 bis f_6 sind beispielsweise zu summieren, um nur einen Zahlenwert als Güte für ein Cluster-Labeling zu erhalten. Ebenfalls kann ein Produkt der Maße gebildet werden, wobei der Einfluss eines Maßes durch ein entsprechendes Gewicht beeinflusst werden kann.

Eine Quantifizierung hierarchischer Label-Eigenschaften entfällt, da in dieser Arbeit keine hierarchischen Cluster-Labeling-Verfahren betrachtet werden.

7.2.2 NDCG-basiertes Validierungsmaß

Die in Kapitel 7.2.1 aufgestellte Quantifizierung von Label-Eigenschaften wird zur Bewertung der Relevanz einer Phrase herangezogen. Die folgenden Maße werden angepasst, so dass diese anstatt eines Cluster-Labelings jeweils nur eine einzige Phrase p eines Clusters c bewerten. Da NDCG die Reihenfolge der Phrasen in einem Cluster-Label berücksichtigt, beziehen einige Maße den Rang R mit in die Berechnung ein. Es gilt:

Verständlichkeit

$$f_{1,\mathrm{NDCG}}(c, p) = \mathrm{penalty}(p) * \mathrm{NP}(p)$$

mit

$$\mathrm{penalty}(p) := \begin{cases} \exp \frac{-(|p| - |p|_{\mathrm{opt}})^2}{2 \cdot d^2} & \text{, wenn } |p| = 1 \\ 0,5 & \text{, sonst} \end{cases}$$

$$\mathrm{NP}(p) := \begin{cases} 1 & \text{, wenn } p \in L(G) \\ 0 & \text{, sonst} \end{cases}$$

Sei $L(G)$ ist die in Formel 4.1 definierte Sprache.

Überdeckung

$$f_{2,\mathrm{NDCG}}(c,p,R) = 1 - \frac{1}{|P_c \setminus \tau(c)_R|} \sum_{\substack{p' \in P_c \\ p' \notin \tau(c)_R}} \frac{\mathrm{df}(p')}{\mathrm{df}(p)}$$

mit $\tau(c)_R$, die Menge der Phrasen in $\tau(c)$ bis zur Position r.

Trennschärfe

$$f_{3,\mathrm{NDCG}}(c,p) = 1 - \frac{1}{k-1} \sum_{\substack{c' \in \mathcal{C} \\ c' \neq c}} \frac{|c| \ \mathrm{df}_{c'}(p)}{|c'| \ \mathrm{df}_c(p)}$$

Minimale Überlappung

$$f_{4,\mathrm{NDCG}}(c,p) = 1 - \frac{1}{k-1} \sum_{\substack{c' \in \mathcal{C} \\ c' \neq c}} \frac{|c(p) \cap c'(p)|}{|c(p) \cup c'(p)|}$$

Eindeutigkeit

$$f_{5,\mathrm{NDCG}}(c,p,R) = 1 - \frac{1}{k-1} \sum_{\substack{c' \in \mathcal{C} \\ c' \neq c}} \frac{|p \cap \tau(c')_R|}{|p \cup \tau(c')_R|}$$

Redundanzfreiheit

$$f_{6,\mathrm{NDCG}}(c,p) = 1 - \frac{1}{|\tau(c)_r| - 1} \sum_{\substack{p' \in \tau(c)_r \\ p' \neq p}} \mathrm{syn}(p,p')$$

Es wird erneut auf die Quantifizierung hierarchischer Label-Qualitätsmaße verzichet, da kein in dieser Arbeit aufgeführtes Verfahren ein hierarchisches Labeling erzeugt.

Die Relevanz rel_i der i-ten Phrase eines Cluster-Labels berechnet sich durch

$$rel_i = \sum_{i=1}^{|Y(\cdot)|} y_i(\cdot),$$

mit $Y(\cdot) = \{y_1(\cdot), \ldots, y_6(\cdot)\}$ und der Abbildung $y_i \ : \ p \times c \times r \mapsto \{0,1\}$ mit

$$y_i(\cdot) = \begin{cases} 1 & \text{, wenn } f_i(\cdot) \geq 0,5 \\ 0 & \text{, sonst} \end{cases}$$

Je mehr Label-Eigenschaften eine Phrase erfüllt, desto höher ist deren Relevanz.

Eine Gewichtung der einzelnen Maße ist möglich. Es ist anzumerken, dass für die definierten Maße es nicht gleich schwer ist, die Schwelle von 0,5 zu überschreiten.

7.3 Externe Gütekriterien zur Validierung eines Clusterings

Im Folgenden werden externe Validierungsmaße vorgestellt, die zur Bewertung der Clustering-Qualität herangezogen werden können. Für ein traditionelles Clustering bedeutet dies, dass zur Validierung eine menschliche Kategorisierung einer Dokumentkollektion (Referenzkategorisierung) vorliegt[2].

Ziel eines Clusterings ist es, Dokumente Clustern derart zuzuweisen, dass diese alle und nur diese Dokumente enthalten, die Mitglied derselben Kategorie sind. Ist das Clustering mit der Referenzkategorisierung identisch, ist es *perfekt*. Sind Kategorien a priori gegeben, ist dieses trivial zu überprüfen. Jedoch wird selten ein perfektes Clustering erzielt. Validierungsmaße sind deshalb erforderlich, um zu bewerten, wie stark sich das tatsächliche Clustering vom Ideal unterscheidet.

Problemstellung

Wie bereits in Kapitel 2.3 eingeführt, ist ein Clustering $\mathcal{C}$ eine Partitionierung einer Dokumentkollektion $\mathcal{D}$ in Teilmengen $c_1, c_2, \ldots, c_n$, so dass $\bigcup_{k=1}^{n} c_k = \mathcal{D}$. Je nach gewähltem Clustering-Verfahren ist die Zuordnung zu einem Cluster exklusiv oder nicht. Im ersten Fall sind die Teilmengen disjunkt und es gilt:

$$\forall_{d \in \mathcal{D}} \; \exists!_{c \in \mathcal{C}} \; : \; d \in c \; .$$

Die Zugehörigkeit der Dokumente $d \in \mathcal{D}$ zu den Clustern $c \in \mathcal{C}$ ist durch eine $|\mathcal{C}| \times |\mathcal{D}|$-Matrix A gegeben. Es gilt $a_{ij} = 1$ genau dann, wenn $d \in c$. In allen anderen Fällen ist $a_{ij} = 0$. Da die Teilmengen disjunkt sind, gilt zudem $\forall_i \; : \; \sum_j a_{ji} = 1$.

Clustering-Verfahren, die beim Cluster-Labeling eingesetzt werden, sind in der Regel überlappend. Sie bilden nicht zwingend disjunkte Teilmengen aus, so dass nun

$$\forall_{d \in \mathcal{D}} \; \exists_{c \in \mathcal{C}} \; : \; d \in c \qquad \text{und} \qquad \forall_i \; : \; \sum_j a_{ji} \geq 1$$

gilt. Sei $\mathcal{C}^* = \{c_1^*, c_2^*, \ldots, c_{n'}^*\}$ ein zweites Clustering für $\mathcal{D}$. Die Anzahl der Cluster kann unterschiedlich sein.

Sei $\mathcal{C}^*$ die menschlich gewählte Referenzkategorisierung. $\mathcal{C}$ beschreibt das tatsächlich erzielte Clustering. Mit Hilfe einer Kreuztabelle lässt sich bewerten, wie stark $\mathcal{C}$ von $\mathcal{C}^*$ abweicht. Hierfür wird eine $|\mathcal{C}^*| \times |\mathcal{C}|$-Matrix H mit $h_{ij} = |c_i^* \cap c_j|$ erstellt. Sie beschreibt

[2] Die Kategorisierung wird in der Regel durch Experten vorgenommen und unterliegt einer Mehrheitsentscheidung.

das gemeinsame Auftreten von Dokumenten, die gleichzeitig einem Cluster $c \in \mathcal{C}$ zugewiesen wurden und einer Klasse $c^* \in \mathcal{C}^*$ angehören. Bei einer Referenzkategorisierung wird von Klassen anstatt Clustern gesprochen. Ein perfektes Clustering erzeugt für H eine Diagonalmatrix.

Im Folgenden wird zur Bewertung eines traditionellen Clusterings das bekannteste Maß, *F-Measure*, vorgestellt. Zur Bewertung des Clusterings eines Cluster-Labeling-Verfahrens, welches unter anderem überlappend ist, wird das Maß *Cluster Contamination* diskutiert.

F-Measure

Ein externes Validierungsmaß zur Bewertung eines Clusterings sollte zwei Kriterien erfüllen (Rosenberg u. Hirschberg, 2007): Homogenität und Vollständigkeit. Ein Cluster ist homogen, wenn es ausschließlich Dokumente einer einzigen Klasse enthält. Das ist einfach zu erreichen, in dem jedes Dokument ein eigenes Cluster bildet. Das Clustering ist somit perfekt homogen. Im Unterschied dazu ist ein Cluster vollkommen inhomogen, wenn alle Dokumente aller Klassen einem einzigen Cluster zugewiesen werden. In diesem Fall ist das Cluster jedoch perfekt vollständig, da es alle Dokumente aller Klassen enthält. Vollständigkeit bedeutet also, dass in einem Cluster alle Dokumente einer Klasse zu finden sind. Beide Kriterien sind gegensetztlich.

F-Measure ist ein Qualitätsmaß, welches ein Clustering positiv bewertet, wenn beide Kriterien maximal werden. Hierbei wird das Maß der Precision verwendet, um die Homogenität eines Clusters zu messen. Die Precision setzt die Übereinstimmung der Dokumente im Cluster mit Dokumenten der Referenzklasse (h_{ij}) ins Verhältnis zur Anzahl der im Cluster c_i enthaltenen Dokumente. Es gilt:

$$\text{Precision}(c_i^*, c_j) = \frac{h_{ij}}{|c_j|}.$$

Zur Bewertung der Vollständigkeit wird der Recall verwendet. Dieser setzt nun die Schnittmenge $h_{ij} = |c_i^* \cap c_j|$ ins Verhältnis zur Anzahl der Dokumente, die zur Klasse c_i^* gehören. Es gilt:

$$\text{Recall}(c_i^*, c_j) = \frac{h_{ij}}{|c_i^*|}.$$

Motivation für das F-Measure ist, dass Precision und Recall eines guten Clusterings hoch sein sollten. Hierzu wird das harmonische Mittel beider Maße gebildet:

$$\text{F-Measure}(c_i^*, c_j) = \frac{2 \cdot \text{Recall}(c_i^*, c_j) \cdot \text{Precision}(c_i^*, c_j)}{\text{Recall}(c_i^*, c_j) + \text{Precision}(c_i^*, c_j)}. \tag{7.6}$$

Cluster Contamination

Die Cluster Contamination ist ein Maß, welches speziell für die Validierung von Cluster-Labeling-Verfahren entwickelt wurde. Hintergrund ist, dass durch das überlappende Clustering Dokumente verschiedener Klassen einem Cluster zugewiesen sind. Anhand der bisherigen Definition von Homogenität wäre dieses Cluster nicht mehr als homogen anzusehen. Dennoch gilt es für das Cluster-Labeling als homogen, solange jeweils Dokumentpaare im Cluster enthalten sind, die eine gemeinsame Klasse besitzen. Ein Beispiel: Ein Cluster *Film* ist homogen, wenn es ausschließlich Dokumente der beiden Filme *Psycho* und *Vertigo* enthält. Befindet sich darunter allerdings ein einziges Dokument, welches beispielsweise der Klasse *Antibiotika* angehört, ist das Cluster nicht mehr perfekt homogen.

Die Precision wird dieser Forderung nicht gerecht, da die Homogenität hier anhand der stärksten Klasse im Cluster bewertet wird. Die Zusammensetzung des restlichen Clusters bleibt unberücksichtigt.

Weiss (2006) formuliert aus diesen Gründen das Maß der Cluster Contamination. Dieses bezieht alle Dokumentpaare eines Clusters in die Bewertung mit ein. Mittels der Matrix H sind zunächst alle Dokumentpaare anzugeben, die einem gemeinsamen Cluster c_j, aber verschiedenen Klassen c_i^* angehören[3]. Die Funktion f_{10} bewertet die Inhomogenität im Cluster:

$$f_{10}(j) = \sum_{i=2}^{|\mathcal{C}|} \sum_{t=1}^{i-1} h_{ij} \cdot h_{tj}$$

Es handelt sich also um alle Paare im Cluster, die beim Cluster-Labeling nicht erwünscht sind. Um die schlechteste Anzahl aller Dokumentpaare im Cluster zu ermitteln, die verschiedenen Klassen angehören würden, wird $f_{\max}$ für ein Cluster c_j berechnet:

$$f_{\max}(j) = \sum_{i=2}^{|\mathcal{C}|} \sum_{t=1}^{i-1} \hat{h}_{ij} \cdot \hat{h}_{tj}$$

mit

$$\hat{h}(ij) = \begin{cases} \left\lfloor \frac{\sum_{t=1}^{|\mathcal{C}|} h_{tj}}{|\mathcal{C}|} \right\rfloor & \text{wenn } i < \sum_{t=1}^{|\mathcal{C}|} h_{tj} \pmod{|\mathcal{C}|} \\ \left\lfloor \frac{\sum_{t=1}^{|\mathcal{C}|} h_{tj}}{|\mathcal{C}|} \right\rfloor + 1 & \text{sonst} \end{cases}$$

[3]Die Definition der Funktion f_{10} orientiert sich an der Notation des Rand-Index. Die erste Ziffer im Index gibt an, ob die Dokumentpaare einem gemeinsamen Cluster angehören (=1) oder verschiedenen (=0). Ebenso verhält es sich mit der zweiten Ziffer. Hierbei wird sich allerdings auf eine gemeinsame Klasse (=1) oder verschiedene (=0) bezogen.

Die Cluster Contamination wird schließlich als Verhältnis zwischen der tatsächlichen Inhomogenität f_{10} und der maximalen Inhomogenität $f_{\max}$ bestimmt. Es gilt:

$$\text{Cluster Contamination}(j) = \frac{f_{10}(j)}{f_{\max(j)}}. \tag{7.7}$$

Im Gegensatz zum F-Measure bewertet die Cluster Contamination ausschließlich die Homogenität eines Clusters. Sie macht allerdings keine Aussagen über die Güte von Cluster-Labeln, so dass dieses Maß im weiteren Verlauf der Arbeit nicht berücksichtigt wird.

8 Experimente

Die in dieser Arbeit diskutierten Cluster-Labeling-Verfahren werden mittels der in Kapitel 7 vorgestellten Validierungsmaße ausgewertet. Es sind folgende Fragestellungen von Interesse:

1. Können Cluster-Labeling-Verfahren verschiedene Themen in einer Dokumentkollektion unterscheiden?
2. Bevorzugen die quantifizierten Label-Eigenschaften tatsächlich prägnante und verständliche Cluster-Label?
3. Führt ein perfektes Clustering zu einer verbesserten Qualität von Cluster-Labeln?
4. Welchen Einfluss hat die Wahl eines Schlüsselwortbestimmungsverfahrens auf die Qualität von Cluster-Labeln?
5. Welches Cluster-Labeling-Verfahren ist das beste?
6. Korreliert das NDCG-basierte interne Validierungsmaß mit externen Validierungsmaßen?

Zur Beantwortung einiger Fragestellungen wird ein Referenzkorpus benötigt. Dieser wird zunächst vorgestellt. Im Anschluss wird die Parametrisierung der zu untersuchenden Cluster-Labeling-Verfahren für die Experimente festgelegt. Kapitel 8.2.1 diskutiert die Fragestellung 1. Anschließend beantwortet Kapitel 8.2.2 die Fragestellung 2, bevor in Kapitel 8.2.3 die Fragestellungen 3 bis 6 diskutiert werden.

8.1 Experimentbeschreibung

In diesem Kapitel wird ein Referenzkorpus für eine externe Validierung vorgestellt. Im weiteren Verlauf des Kapitels wird die Parametrisierung der Cluster-Labeling-Verfahren beschrieben.

ID	Kategorie	Dokumente
Film/Kubrick	Arts/Movies/Titles/2/2001 - A Space Odyssey	33 (43)
	Arts/Movies/Titles/E/Eyes Wide Shut	17 (20)
	Arts/Movies/Titles/K/Killer's Kiss	15 (15)
Film/Hitchcock	Arts/Movies/Titles/P/Psycho - 1960	13 (15)
	Arts/Movies/Titles/R/Rear Window	9 (9)
	Arts/Movies/Titles/V/Vertigo	15 (17)
Datenbanken	Computers/Software/Databases/Data Warehousing	17 (39)
	Computers/Software/Databases/IBM DB2	28 (30)
	Computers/Software/Databases/MySQL	42 (47)
	Computers/Software/Databases/PostgreSQL	35 (40)
Gesundheit	Health/Pharmacy/Drugs and Medications/Antibiotics	16 (22)
Freizeit	Recreation/Birding/Backyard Birding/Bluebirds	43 (46)

Tabelle 8.1 *: Kategorien aus dem Open Directory Project, die für die Experimente ausgewählt wurden. Bei der Kategorie* `Computers/Software/Databases/Data Warehousing` *wurde die Unterkategorie* `Data Integrity and Cleansing Tools` *gewählt. Jeweils in Klammern ist die maximal verfügbare Anzahl hinterlegter Internetseiten pro Kategorie angegeben. Die tatsächliche Anzahl fällt geringer aus, da einige beim Open Directory Project hinterlegte Internetadressen nicht erreichbar waren.*

8.1.1 Referenzkorpora zur Durchführung von Experimenten

In Kapitel 3 wurde das Open Directory Project hinsichtlich der Zusammensetzung von menschlich ausgewählten Kategorienamen untersucht. Diese werden für die Experimente als Referenz verwendet. Der in dieser Arbeit verwendete Korpus setzt sich aus den in Tabelle 8.1 aufgeführten Kategorien zusammen. Es werden vier thematisch nicht-verwandte Kategorien ausgewählt: Filme, Datenbanken, Gesundheit und Freizeit. Zu jeder dieser Kategorien werden zusätzlich ein bis sechs Unterkategorien bestimmt.

Zu jeder Kategorie werden die mit der Kategorie assoziierten Internetseiten heruntergeladen. Beim Herunterladen werden auch die ausgehenden Verweise einer Internetseite bis zur Tiefe 1 berücksichtigt. HTML-Elemente und JavaScript-Quelltexte werden entfernt. Ist die resultierende Textlänge für eine Internetseite geringer als 1.000 Zeichen, wird diese nicht berücksichtigt.

Verschiedene Kategorien werden für die Experimente jeweils zu einem von fünf Testkorpora zusammengestellt. Die Motivation für die jeweiligen Experimente einschließlich

ID	Kategorien	Fragestellung
E1	2001 - A Space Odyssey, MySQL, Antibiotics, Bluebirds	Ist das Verfahren in der Lage, die nicht-verwandten Kategorien zu unterscheiden und korrekt zu benennen?
E2	Film/Kubrick, Film/Hitchcock	Ist das Verfahren in der Lage, die sechs thematisch verwandten, aber nicht identischen Kategorien zu unterscheiden und korrekt zu benennen? Erkennt ein hierarchisches Verfahren, dass es sich um Filme handelt, die von zwei unterschiedlichen Regisseuren stammen? (Verallgemeinerung)
E3	Datenbanken, Gesundheit	Ist das Verfahren in der Lage, sowohl die vier thematisch verwandten Kategorien (Datenbanken), als auch die thematisch nicht verwandte Kategorie (Gesundheit) zu unterscheiden? Erkennt ein hierarchisches Verfahren, dass vier Kategorien aus dem Bereich *Computer/Datenbanken* stammen und eine aus dem Bereich *Gesundheit/Pharmazie*?
E4	Film/Kubrick, Film/Hitchcock, Datenbanken	Ist ein Verfahren in der Lage, die beiden thematischen Gruppen Filme und Datenbanken zu trennen?
E5	Film/Kubrick, Film/Hitchcock, Datenbanken, Gesundheit, Freizeit	Ist ein Verfahren in der Lage, beiden Gruppen Filme und Datenbanken, sowie Gesundheit und Freizeit zu trennen?

Tabelle 8.2 *: Zusammenstellung der Datensätze für die Experimente. Jedem Experiment wird eine Identifikationsnummer zugewiesen. Es wird jeweils eine bestimmte Fragestellung verfolgt. Auf hierarchische Fragestellungen wird in dieser Arbeit nicht eingegangen.*

der Zusammensetzung der Kategorien ist in Tabelle 8.2 aufgeführt.

Stand der Daten ist der 30. Dezember 2009. Die Internetseiten wurden im Mai 2010 heruntergeladen. Das Testkorpus wird veröffentlicht.

Wikipedia könnte ebenfalls als Referenzkorpus verwendet werden. Allerdings zeigt sich, das Wikipedia für die in dieser Arbeit durchzuführenden Experimente als Korpus nicht geeignet ist. Ein Referenz-Label, also eine Wikipedia-Kategorie, muss in den Wikipedia-Artikeln vorkommen, um von einem Schlüsselwortverfahren ermittelt zu werden. Eine

Auswertung von 5.500 Kategorien[1] zeigt aber, dass dies nur auf 7,9% der Kategorien zutrifft. Das eigentliche Thema eines Wikipedia-Artikels kommt also selten im Text vor. Eine exakte Übereinstimmung der erzeugten Cluster-Label mit der menschlich ausgewählten Referenz ist in diesen Fällen selten gegeben.

Im Unterschied dazu wird beim zuvor vorgestellten Open Directory Project-Korpus eine exakte Übereinstimmung mit dem menschlich ausgewählten Kategorienamen für 65,5% der Kategorien erreicht. Eine Übereinstimmung ist somit wahrscheinlicher.

Die Analyse des Wikipedia-Korpus zeigt, dass zur Ermittlung eines Cluster-Labels, welches mit dem Kategorienamen Wikipedias übereinstimmt, eine Schlüsselwortbestimmung nicht zum Erfolg führt. Daher wird auf einen Wikipedia-basierten Referenzkorpus verzichtet.

8.1.2 Experimentparameter

In diesem Kapitel wird auf die für die Experimente vorzunehmende Parametrisierung der Cluster-Labeling-Verfahren eingegangen.

Ermittlung von Schlüsselphrasen

Als Schlüsselphrasen werden für alle Cluster-Labeling-Verfahren Nominalphrasen aus der jeweiligen Dokumentkollektion ermittelt. Hiervon ausgenommen sind datenzentrierte Verfahren und Suffixbaum-Clustering. Erstere ermitteln keine Schlüsselphrasen, letztgenannte Verfahren ermitteln dagegen die häufigsten Suffixe in einer Dokumentkollektion.

Zur Ermittlung von Nominalphrasen wird ein Verfahren von Barker u. Cornacchia (2000) verwendet. Barker u. Cornacchia bestimmen die häufigsten Substantive in Dokumenten. Anhand dieser werden Nominalphrasen ermittelt. Nominalphrasen werden gewichtet, indem die Auftrittshäufigkeit mit der Anzahl der Worte dieser Phrase multipliziert wird. Das Verfahren wird in dieser Arbeit in Tabellen und Abbildungen mit NPE abgekürzt. Cluster-Labeling-Verfahren, die zur Schlüsselwortbestimmung Nominalphrasen ermitteln, werden *ohne* Zusatz von NPE gekennzeichnet.

Um in den Experimenten die Frage zu beantworten, ob die Wahl eines Schlüsselwortbestimmungsverfahrens die Qualität von Cluster-Labeln beeinflusst, werden neben der Ermittlung von Nominalphrasen auch häufige Phrasen in Dokumenten ermittelt.

[1] Für die Clustering-Aufgabe sind Wikipedia-Artikel zu berücksichtigen, die nur einer einzigen Kategorie zugewiesen sind. Dies trifft auf 11% der Artikel zu. Siehe hierzu Tabelle 3.7 auf Seite 38.

Das von Tseng (1998) präsentierte Verfahren ermittelt zunächst in Dokumenten häufige Wort-Bigramme. Überlappen sich Paare von Bigrammen mit ähnlicher Häufigkeit, werden diese konsekutiv zu immer längeren Phrasen zusammengefasst. Auf Basis häufiger Wort-Bigramme können Phrasen beliebiger Länge ermittelt werden. Die resultierenden Phrasen werden anhand der Häufigkeit aggregierter Wort-Bigramme gewichtet. Das Verfahren wird in dieser Arbeit in Tabellen und Abbildungen mit RPE abgekürzt.

Parametrisierung von Cluster-Labeling-Verfahren

Für die in dieser Arbeit vorgestellten Cluster-Labeling-Verfahren sind benutzerdefinierte Parameter für die Experimente festzulegen. Diese werden für die Dauer eines Experiments nicht verändert. Wenn nicht anders angegeben, werden Terme in Dokumenten mittels *tf-idf* gewichtet. Es werden Stoppworte aus den Texten entfernt und Stemming durchgeführt. Als Clustering-Verfahren wird k-Means verwendet. Für die Experimente wird k auf die Anzahl der im Experiment enthaltenen Kategorien gesetzt. Wird anstatt eines tatsächlichen Clusterings die Referenzkategorisierung von einem Cluster-Labeling-Verfahren verwendet, wird das Verfahren in dieser Arbeit in Tabellen und Abbildungen mit dem Zusatz REF versehen.

Frequent Predictive Words Für die Bewertung der Qualität von Cluster-Labeln anhand externer Maße wird die zu erzeugende Anzahl an Phrasen pro Label auf fünf festgelegt. Keine weiteren Parameter sind festzusetzen. Das Verfahren wird in dieser Arbeit in Tabellen und Abbildungen mit FPW abgekürzt.

Weighted Centroid Covering Einstellungen sind dieselben wie für Frequent Predictive Words. Das Verfahren wird in dieser Arbeit in Tabellen und Abbildungen mit WCC abgekürzt.

Suffixbaum-Clustering Um die Verständlichkeit von Cluster-Labeln sicherzustellen, werden aus den Eingabedokumenten keine Stoppworte entfernt. Es wird auch kein Stemming durchgeführt. Das Verfahren wird in Tabellen und Abbildungen mit STC abgekürzt.

Zur Bildung des finalen Clusterings werden 15 Cluster mit der höchsten Bewertung ausgewählt. Die Bewertung eines Clusters erfolgt nach Formel 2.3 von Seite 27. Ein zusätzlicher Cluster enthält Dokumente, die zuvor keinem Cluster zugewiesen wurden.

Im Cluster-Label stehen nur Phrasen, die weder am allgemeinsten noch am spezifischsten sind. Siehe hierzu die Ausführungen in Kapitel 6.4.

Descriptive k-Means Zur Abwertung zu kurzer oder zu langer Phrasen wird die *penalty*-Funktion aus Formel 5.1 verwendet:

$$\text{penalty}(p) = \exp \frac{-(|p| - m)^2}{2 * d^2},$$

mit $m = 4$, $d = 8$. Sei p die Phrase und $|p|$ die Anzahl der Worte, aus denen p besteht. Experimente zeigten, dass die Wahl der Parameter für Descriptive k-Means schwierig ist. In Osinski u. a. (2004) wird eine optimale Parametrisierung nicht angegeben. Experimente dieser Arbeit zeigen, dass eine Zuweisung von Schlüsselphrasen zu Centroiden ohne Überschreitung eines benutzerdefinierten Schwellwertes subjektiv zu einer besseren Qualität der Cluster-Label führt. Allerdings wird dadurch eine große Anzahl Cluster erzeugt. Somit werden für die Auswertungen nur die 20 besten Cluster, die von Descriptive k-Means erzeugt werden, berücksichtigt. Eine Ausnahme bildet das Experiment zur Ermittlung von Themen in einer Dokumentkollektion. Hier werden so viele Cluster angezeigt, bis möglichst alle Kategorien der Dokumentkollektion repräsentiert sind.

Das Verfahren wird in dieser Arbeit in Tabellen und Abbildungen mit DK abgekürzt.

Lingo Im Unterschied zu Ausführungen in Osinski u. a. (2004) werden als Schlüsselphrasen für Dokumente nicht die häufigsten Phrasen, sondern Nominalphrasen ermittelt. Dieses erlaubt in dieser Arbeit die Vergleichbarkeit mit anderen Cluster-Labeling-Verfahren.

Aus Effizienzgründen werden die 2.000 häufigsten Nominalphrasen in einer Dokumentkollektion für die weitere Verarbeitung in Lingo ausgewählt. Zur Repräsentation der Dokumente werden die 10.000 Terme mit der höchsten Termhäufigkeit in der Dokumentkollektion verwendet.

Die Term-Dokument-Matrix A der Eingabedokumente wird auf einen neuen Rang k reduziert. k soll abhängig von der Anzahl der Eingabedokumente sein, so dass k anhand folgender Heuristik festgelegt wird: $k = \min(3 \cdot |\mathcal{D}|, |\mathcal{D}|)$. Der Parameter q zur Bestimmung der Qualität einer Approximation von A wird nicht benötigt.

Entscheidend für die Erzeugung der finalen Cluster ist, ab welchem Schwellwert ein Dokument einem Label zugewiesen wird. In dieser Arbeit wurden experimentell für die Testkorpora E1 bis E5 folgende Schwellwerte ermittelt: 0.15, 0.04, 0.1, 0.15, 0.15. Diese Schwellwerte ergeben jeweils das Clustering mit dem besten F-Measure Wert.

Das Verfahren wird in Tabellen und Abbildungen mit LINGO abgekürzt.

Topical k-Means Experimentell wurde in dieser Arbeit ermittelt, dass zur Repräsentation von Textdokumenten das Vokabular auf die 2.500 häufigsten in der Dokumentkol-

lektion vorkommenden Nominalphrasen zu beschränken ist. Aus jedem Centroiden sind die 250 besten Terme als potenzielle Cluster-Label zu ermitteln.

Zur Maximierung der Überdeckung von Dokumenten eines Clusters werden so viele Cluster für einen Centroiden erzeugt, bis 80% der Dokumente eines Clusters abgedeckt sind. In dieser Arbeit wird das Verfahren in Tabellen und Abbildungen mit TK abgekürzt.

Zusätzlich gilt für Suffixbaum-Clustering, Descriptive k-Means, Lingo und Topical k-Means: Ein Cluster wird erzeugt, wenn mindestens fünf Dokumente diesem zugewiesen werden können. Ein zusätzlicher Cluster enthält Dokumente, die zuvor keinem Cluster zugewiesen werden konnten.

Zur Validierung von Cluster-Labeling-Verfahren durch externe Maße, die von einer nach Relevanz sortierten Liste von Phrasen in einem Cluster-Label ausgehen, werden Descriptive k-Means und Topical k-Means modifiziert. Diese Verfahren erzeugen nur eine Phrase pro Cluster-Label. Die Cluster-Label ergeben sich nun wie folgt: Die fünf besten Cluster-Label, die für ein Cluster ermittelt werden, werden zu einem einzigen Cluster-Label für das Cluster konkateniert. Es werden nun k Cluster mit Cluster-Labeln, die aus fünf Phrasen bestehen, erzeugt.

Die Änderung ist für Lingo und Suffixbaum-Clustering nicht durchführbar, so dass diese Verfahren in Experimenten, die durch externe Validierungsmaße evaluiert werden, nicht berücksichtigt werden.

Es werden ebenfalls mit Hilfe der in Kapitel 8.1.2 vorgestellten Schlüsselwortbestimmungsverfahren Nominalphrasen und häufige Phrasen ermittelt. Dabei bilden jeweils die fünf besten Phrasen, die für ein Cluster ermittelt werden, das Cluster-Label.

8.2 Experimentergebnisse

In diesem Kapitel werden die Ergebnisse der Experimente vorgestellt.

8.2.1 Ermittlung von Themen

Dieses Experiment beantwortet Fragestellung 1. Es ist zu zeigen, dass Cluster-Labeling-Verfahren Themen in Dokumentkollektionen ermitteln können. Hierfür werden die Testkorpora E1 bis E4 ausgewertet. Für jeden Korpus zeigt ein Balkendiagramm die vom Cluster-Labeling-Verfahren erzeugten Cluster und die Verteilung der Dokumente in diesen an. Für alle Cluster-Labeling-Verfahren gilt, dass die Relevanz der Cluster, die entlang der x-Achse angeordnet sind, vom Ursprung aus abnimmt. Eine Ausnahme bildet

Weighted Centroid Covering (WCC), bei dem keine Rangfolge der Cluster erzeugt wird. Eine Auswertung des Verfahrens Frequent Predictive Words entfällt, da das Clustering identisch zu WCC ist. Die resultierenden Label unterscheiden sich nur geringfügig voneinander.

Im Folgenden werden die Ergebnisse für Topical k-Means und Suffixbaum-Clustering diskutiert. Eine Auswertung weiterer Verfahren ist im Anhang A auf Seite 132 angeführt.

Topical k-Means

Topical k-Means erzeugt für die Testkorpora E1 bis E4 zwischen 14 und 27 Cluster. Die Ergebnisse sind in Abbildung 8.1 auf Seite 105 zusammengefasst.

E1 Topical k-Means erzeugt Cluster, die alle vier Themenbereiche durch mindestens drei Cluster abdecken. Dabei werden aussagekräftige Cluster-Label wie *Stanley Kubrick film review* gegenüber allgemeineren wie *Film Odyssey* bevorzugt. Dies liegt daran, dass die Anzahl der Dokumente nicht in die Bewertung eines Clusters mit einfließt. Die Bewertung eines Clusters basiert ausschließlich auf der Bewertung des assoziierten Cluster-Labels.

E2 Von den zu erkennenden Film-Kategorien wird ausschließlich *Psycho* eindeutig erkannt. Im Gegensatz zu Clustern in E1 werden nun vermehrt Dokumente mehrerer Kategorien in einem Cluster zusammengefasst. Topical k-Means erkennt beispielsweise alle drei Filme von Stanley Kubrick und gruppiert diese in einem Cluster namens *Stanley Kubrick Stanley Kubrick*. Phrasen mit sich wiederholenden Termen kommen hierbei zu Stande, da durch Entfernung der HTML-Elemente im Open Directory Project-Korpus unter anderem Überschrift und Satzanfang zusammenfallen.

Alle Kategorien von E2 finden sich im Cluster *Review* wieder. Überraschend ist, dass die größte Kategorie, A Space Odyssey, nicht als eigenständiges Cluster erkannt wird. Die Dokumente der Kategorie verteilen sich vielmehr über mehrere Cluster. Ein Fehlen der Kategorie wird durch die Begrenzung der Dokumentüberdeckung eines Clusters auf 80% begründet. Wird der Wert auf 100% erhöht, wird auch A Space Odyssey gefunden.

E3 Im Unterschied zu E2 werden hier die thematisch verwandten Datenbank-Kategorien klar voneinander unterschieden. Dies wird damit begründet, dass die Kategorien jeweils ein eigenes Vokabular verwenden und nicht so große Überschneidungen wie bei der Film-Kategorie auftreten.

Neben der Erkennung von Datenbank-Kategorien wird auch die Ausreißer-Kategorie Antibiotika erkannt und prägnant beschriftet. Auch hier werden erneut aussagekräftige Cluster-Label gegenüber allgemeineren bevorzugt.

Das Cluster *Database management system* enthält Dokumente aller vier Datenbank-Kategorien. Hier findet wie bereits in E1 eine Generalisierung von Themen statt.

E4 Sowohl Film-Kategorien als auch Datenbank-Kategorien werden durch die Cluster *Director Stanley Kubrick* und *Database development* zusammengefasst. In jedem Cluster befinden sich ausschließlich Film-Kategorien oder Datenbank-Kategorien. Topical k-Means ist hier in der Lage, beide thematisch nicht verwandten Hauptthemen voneinander zu unterscheiden.

Suffixbaum-Clustering

Das verwendete Suffixbaum-Clustering erzeugt für jeden Testkorpus 17 Cluster. Die Ergebnisse für die Experimente sind in Abbildung 8.2 auf Seite 106 zusammengefasst.

E1 Im Vergleich zu Topical k-Means erkennt das Suffixbaum-Clustering die Kategorie Antibiotika nicht. Suffixbaum-Clustering erzeugt Cluster-Label anhand deren Auftrittshäufigkeit in einer Dokumentkollektion. Antibiotika wird hier nicht berücksichtigt, da diese Kategorie mit 16 Dokumenten im Testkorpus am kleinsten ist. Terme aus dem Bereich Datenbanken überwiegen.

Im Unterschied zu Topical k-Means werden für E1 ebenso Cluster erzeugt, die Dokumente aller Kategorien enthalten, obwohl eine Generalisierung inhaltlich hier nicht sinnvoll ist. Kennzeichnend für diese Cluster ist, dass deren Cluster-Label aus kollektionsspezifischem Vokabular gebildet werden: *Search*, *Contact*, *Navigation, Skip*, *Page*, *Web*, *News* und *Links*. Dieses ist ebenfalls damit zu begründen, dass besonders häufige Terme als Cluster-Label verwendet werden. Dieses ist nicht nur auf den Testkorpus E1 beschränkt, sondern auch in E2, E3 und E4 zu beobachten.

E2 Es werden ausschließlich die Filme Eyes Wide Shut und Space Odyssey eindeutig ermittelt. Alle anderen Filme werden nicht identifiziert. Beide Kategorien stellen jeweils die meisten Dokumente im Cluster. Auch hier handelt es sich neben kollektionsspezifischem Vokabular um die häufigsten Phrasen in der Dokumentkollektion.

E3 Aus bereits für E1 und E2 genannten Gründen wird die Ausreißer-Kategorie Antibiotika nicht erkannt.

E4 Die erzeugten Cluster sind im Vergleich zu bisher betrachteten Verfahren inhomogener. Cluster, die bereits in E1, E2 und E3 erkannt wurden, werden ebenfalls wieder erkannt. Dies ist auf deren hohe Dokumentanzahl zurückzuführen.

Schlussfolgerungen

- Cluster-Labeling-Verfahren sind in der Lage, Themen in Dokumentmengen eindeutig zu erkennen. Probleme ergeben sich, wenn eine Dokumentkollektion, hier E2, viele inhaltlich sehr ähnliche Kategorien enthält. Hier fällt eine klare Trennung der Kategorien schwerer.
- Descriptive k-Means zeigte die beste Erkennungsleistung. Das Verfahren erkannte unter anderem die meisten Film-Kategorien in E2.
- Suffixbaum-Clustering zeigte die schlechteste Leistung bei der Erkennung von Themen in Dokumentmengen. Dieses ist auf die Ermittlung häufiger Suffixe in einer Dokumentkollektion zurückzuführen. Anstatt der eigentlichen Themen wird somit fälschlicherweise kollektionspezifisches Vokabular hervorgehoben.

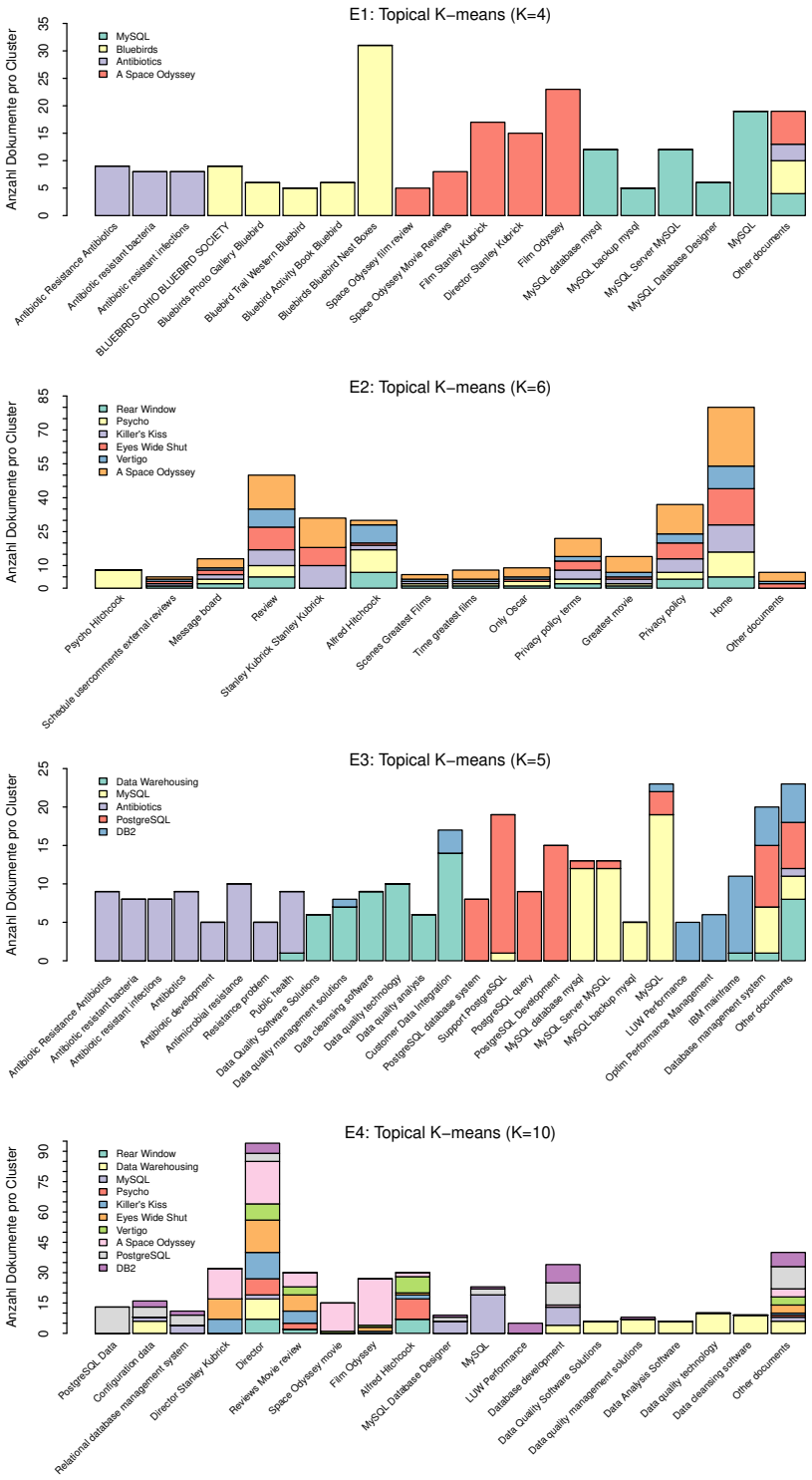

Abbildung 8.1 *: Cluster-Labeling für die Korpora E1 bis E4 mittels Topical k-Means.*

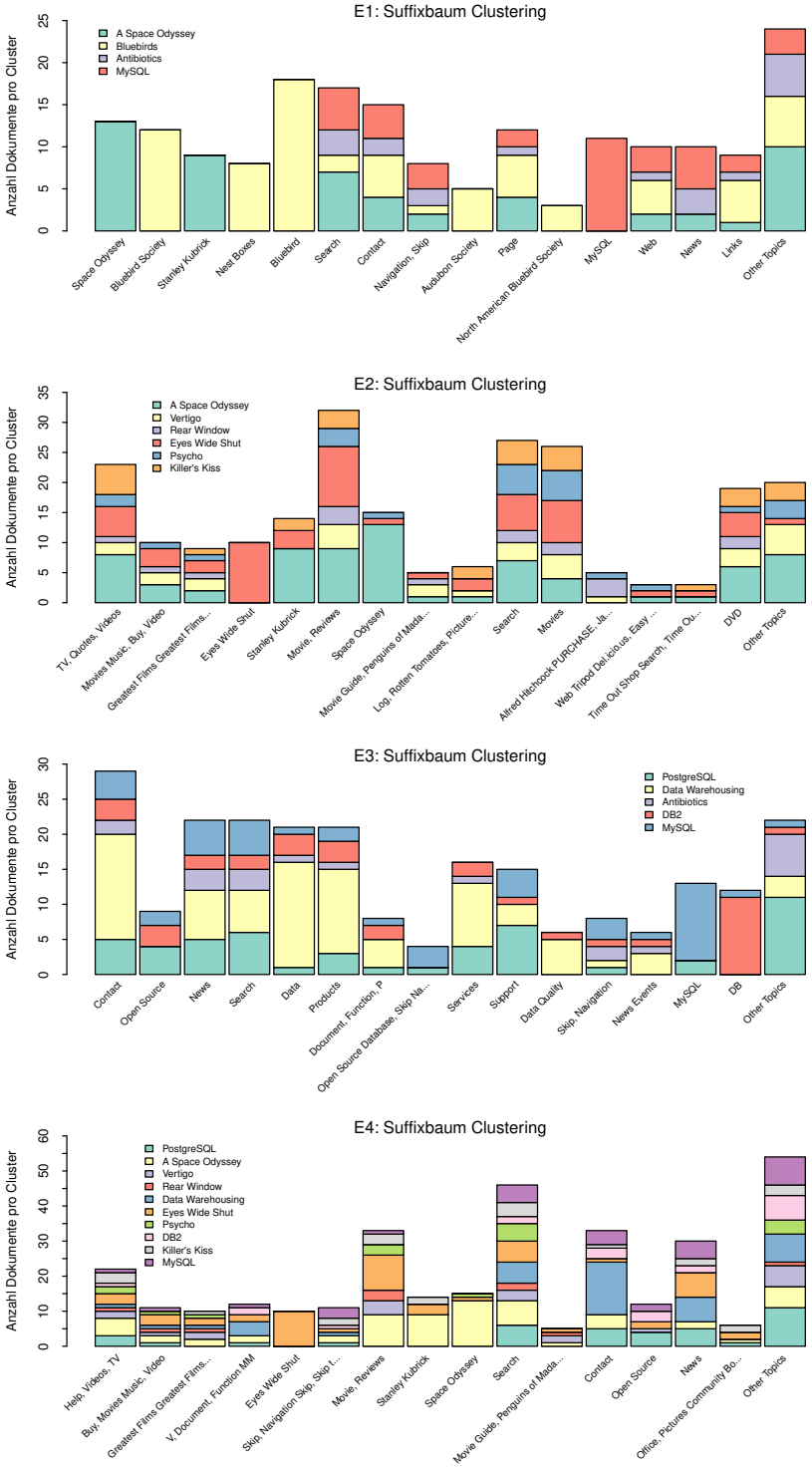

Abbildung 8.2 *: Cluster-Labeling für die Korpora E1 bis E4 mittels Suffixbaum-Clustering.*

ODP-Kategorie	Besten 5 Phrasen	Schlechteste 5 Phrasen
Antibiotics	used antibiotics (5,909)	Technology (-3,773)
	other antibiotics (5,889)	queries (-0,591)
	Antibiotics Health (5,889)	project (1,045)
	antibiotics Antibiotics (5,875)	Print (1,045)
	Antibiotics Work (5,875)	time (1,389)
Psycho	Psycho psycho (5,879)	User (-0,955)
	Bates Motel Norman (5,75)	TOPIC (-0,955)
	Marion Crane Janet Leigh (5,75)	mail (-0,909)
	shower scene Hitchcock (5,75)	list (0,482)
	Martin Balsam (5,745)	release (0,5)
Vertigo	Scottie Stewart (5,818)	team (-2,682)
	film Scottie (5,8)	June (-2,045)
	San Francisco detective (5,8)	interface (-0,864)
	Novak Barbara (5,8)	Company (-0,773)
	Alfred Hitchcocks Vertigo (5,758)	database (-0,636)

Tabelle 8.3 *: Auszug der besten und schlechtesten Phrasen, die durch intrinsische Label-Eigenschaften für Kategorien des ODP bestimmt wurden.*

8.2.2 Validierung von quantifizierten Label-Eigenschaften

Es wird Fragestellung 2 beantwortet. Es ist zu zeigen, dass die in Kapitel 7.2.1 definierten Maße tatsächlich Phrasen als potenzielle Cluster-Label auswählen, die die geforderten Label-Eigenschaften erfüllen.

Für das Experiment werden alle Kategorien, die in Tabelle 8.1 aufgeführt sind, betrachtet. Für jedes Dokument einer Kategorie werden bis zu 100 Nominalphrasen[2] bestimmt. Nominalphrasen aller Dokumente einer Kategorie sind potenzielle Cluster-Label für diese. Jede Phrase wird anhand der in Kapitel 7.2.2 vorgeschlagenen Maße bewertet. Ziel ist, die besten fünf sowie die schlechtesten fünf Phrasen zu ermitteln. Folgende Maße werden berücksichtigt: Verständlichkeit, Überdeckung, Trennschärfe, Minimale Überlappung, Eindeutigkeit und Redundanzfreiheit. Im Anschluss werden die einzelnen Maße summiert, so dass ein potenzielles Cluster-Label einen maximalen Relevanzwert von 6 erreichen kann. Da einige Maße, wie beispielsweise die Trennschärfe, negativ werden können, sind ebenso negative Werte für schlechte Cluster-Label zu erwarten.

[2] Abhängig von der Länge eines Dokumentes werden auch weniger als 100 Nominalphrasen ermittelt.

Phrase	Überdeckung	Trennschärfe	Überlappung
web site	0.9	-0.033	0
use	0.956	-0.115	0
community	0.888	-0.407	0.333
time	0.944	-0.296	0
information	0.941	-0.372	0
click	0.909	-0.454	0

Tabelle 8.4 *: Kollektionsspezifisches Vokabular der Kategorie MySQL.*

Ein Auszug der Ergebnisse des Experiments ist in Tabelle 8.3 aufgeführt. Die vollständige Auswertung findet sich im Anhang B auf Seite 139.

Prägnante Phrasen werden durch die definierten Maße als potenzielle Cluster-Label ausgewählt. Beispielsweise ist eines der besten Cluster-Label für die Kategorie *Antibiotics Health*. Für die Kategorie Vertigo ist es *Alfred Hitchcocks Vertigo*. In vielen Fällen enthalten die Cluster-Label Terme der Referenzkategorie. Ist dies nicht der Fall, sind die Cluster-Label zumindest thematisch verwandt mit der Kategorie. Dieses ist zum Beispiel für die Kategorie Psycho der Fall. Bei den ermittelten Phrasen handelt es sich um Schauspieler oder Filmfiguren. Es wird zudem die Duschszene in Psycho gelistet. Das Schauspieler für Filmkategorien ausgewählt werden, trifft häufiger zu. Dieses ist durch Verwendung von Named Entities zu vermeiden.

Für die Kategorie DB2 wird die Phrase *DB2* nicht ermittelt. Hierbei handelt es sich nicht um eine Nominalphrase.

Der Auswertung ist zu entnehmen, dass die fünf besten Phrasen die geforderten Label-Eigenschaften vollständig erfüllen. Ausschließlich die Überdeckung einer Phrase im Cluster ist nicht maximal, so dass die resultierenden Relevanzwerte nahezu bei 6 liegen.

Die Auswertung der schlechtesten Phrasen in einer Kategorie zeigt, dass sich in jeder Kategorie bestimmte Terme wiederholen: *mail*, *download*, *Click*. Solche Terme sind repräsentativ für Internetseiten. Die quantifizierten Label-Eigenschaften filtern also kollektionsspezifisches Vokabular. Dieses wird deutlicher, wenn nicht nur die schlechtesten fünf Phrasen betrachtet werden. Repräsentativ für alle Kategorien wurde dazu MySQL näher untersucht. Kollektionsspezifisches Vokabular folgt dabei einem bestimmten Muster. Siehe hierzu Tabelle 8.4. Kennzeichnend ist eine hohe Überdeckung in der Kategorie, eine negative Trennschärfe und eine minimale Überlappung nahe 0.

Schlussfolgerungen

- Die internen Validierungsmaße wählen prägnante Phrasen aus einer Dokumentkollektion aus.
- Die internen Validierungsmaße bestrafen potenzielle Cluster-Label, die zwar sehr häufig in einer Kategorie auftreten, aber nicht repräsentativ für die Kategorie sind. Dieses wurde am Beispiel von kollektionsspezifischem Vokabular gezeigt.

8.2.3 Evaluierung von Cluster-Labeling-Verfahren

In diesem Abschnitt sollen die Fragestellungen 3 bis 6 beantwortet werden. Diese sind:

- Führt ein perfektes Clustering zu einer verbesserten Qualität von Cluster-Labeln?
- Welchen Einfluss hat die Wahl eines Schlüsselwortbestimmungsverfahrens auf die Qualität von Cluster-Labeln?
- Welches Cluster-Labeling-Verfahren ist das beste?
- Korreliert das NDCG-basierte interne Validierungsmaß mit externen Validierungsmaßen?

Die Auswertung basiert auf den Korpora E1 bis E5. Als externe Validierungsmaße werden Precision@R, Match@R und MRR mit $R = 1, \ldots, 5$ verwendet. Als internes Maß wird das auf NDCG-basierende Validierungsmaß eingesetzt, welches in Kapitel 7.2.2 vorgestellt wurde.

Bewertung von Cluster-Labeln durch externe Validierungsmaße

Die Auswertung der Cluster-Labeling-Verfahren anhand der externen Validierungsmaße Precision@R, Match@R und MMR ist in den Tabellen 8.5, 8.6 und 8.7 aufgeführt. Die Ergebnisse der drei Maße korrelieren miteinander, so dass ausschließlich auf die Auswertung durch Precision@R eingegangen wird.

Für Precision@1 erzielt DKREF mit 0,74 (partiell) und 0,59 (exakt) die besten Ergebnisse. Die Precision nimmt jeweils bis $R = 3$ auf 0,82 (partiell) und 0,65 (exakt) zu, um anschließend wieder leicht abzufallen. Die Konstanz der Precision@R-Werte ist kennzeichnend für monothetische Cluster-Labeling-Verfahren. Dieses lässt sich ebenfalls für Topical k-Means beobachten. Monothetische Verfahren sind in diesem Experiment

	Precision@R (partiell)					Precision@R (exakt)				
Verfahren	R=1	R=2	R=3	R=4	R=5	R=1	R=2	R=3	R=4	R=5
NPE	0,40	0,30	0,22	0,19	0,17	0,29	0,20	0,13	0,12	0,11
NPEREF	0,46	0,33	0,22	0,21	0,21	0,31	0,22	0,14	0,13	0,12
RPE	0,13	0,18	0,16	0,16	0,16	0,11	0,17	0,15	0,13	0,12
RPEREF	0,08	0,19	0,19	0,17	0,18	0,08	0,19	0,19	0,15	0,15
FPW	0,20	0,11	0,10	0,09	0,07	0,12	0,06	0,04	0,03	0,02
FPWREF	0,50	0,32	0,25	0,21	0,16	0,21	0,12	0,08	0,06	0,05
WCC	0,21	0,14	0,11	0,09	0,09	0,12	0,06	0,04	0,03	0,02
WCCREF	0,21	0,11	0,10	0,08	0,06	0,17	0,09	0,06	0,04	0,03
DK	0,61	0,58	0,59	0,56	0,55	0,57	0,55	0,53	0,51	0,49
DKREF	**0,74**	**0,80**	**0,82**	**0,80**	**0,78**	**0,59**	**0,65**	**0,65**	**0,63**	**0,61**
DKRPE	0,51	0,46	0,50	0,48	0,47	0,39	0,39	0,43	0,41	0,40
DKRPEREF	0,69	0,60	0,56	0,56	0,52	0,53	0,49	0,45	0,45	0,41
TK	0,55	0,52	0,52	0,51	0,51	0,46	0,43	0,41	0,41	0,42
TKREF	0,59	0,57	0,56	0,55	0,53	0,49	0,46	0,44	0,43	0,36
TKRPE	0,48	0,48	0,46	0,46	0,45	0,36	0,38	0,37	0,37	0,40
TKRPEREF	0,42	0,43	0,46	0,50	0,48	0,34	0,38	0,39	0,41	0,40

Tabelle 8.5 *: Auswertung von Cluster-Labeling-Verfahren durch Precision@R.*

die besten. Die hohe partielle Precision@R für diese Verfahren resultiert durch die längeren Phrasen in den Cluster-Labeln. Zur Erinnerung: Bei partieller Übereinstimmung mit einer Referenzkategorie reicht ein gemeinsames Wort bereits aus.

Gefolgt werden monothetische Verfahren von der Nominalphrasen-Extraktion. NPEREF erzielt eine Precision@1 von 0,46 (partiell). Dagegen schneidet RPEREF deutlich schlechter ab. Hier wird eine Precision@1 von nur 0,08 (partiell) erreicht. Dieses ist der schlechteste Wert unter allen Verfahren. Verfahren, die RPE zur Extraktion von Schlüsselworten einsetzen, zeigen einheitlich eine Verschlechterung der Precision-Werte gegenüber der Verwendung von Nominalphrasen als Cluster-Label. TKREF erzielt hier eine Precision@1 von 0,59 (partiell). Werden anstatt Nominalphrasen häufige Phrasen mittels RPE für TKREF extrahiert, fällt dieser Wert auf 0,42 (TKRPEREF, partiell). Dieses ist ebenfalls für Descriptive k-Means zu beobachten. Eine Nominalphrasen-Extraktion ist deshalb vorzuziehen.

FPWREF erzielt für ein datenzentriertes Verfahren eine überraschend hohe Precision@1 von 0,5 (partiell). Allerdings fällt diese für Precision@2 auf 0,32 (partiell) ab. Datenzentrierte Verfahren und schlüsselwortbasierte Verfahren wie NPE und RPE zeigen allgemein bereits für R=2 einen deutlichen Abfall der Precision.

Datenzentrierte Verfahren verwenden die besten fünf Terme aus einem Centroiden als

	Match@R (partiell)					Match@R (exakt)				
Verfahren	R=1	R=2	R=3	R=4	R=5	R=1	R=2	R=3	R=4	R=5
NPE	0,40	0,52	0,52	0,52	0,59	0,29	0,39	0,39	0,39	0,46
NPEREF	0,46	0,58	0,58	0,58	0,72	0,31	0,43	0,43	0,43	0,50
RPE	0,13	0,30	0,37	0,47	0,49	0,11	0,28	0,35	0,37	0,39
RPEREF	0,08	0,38	0,38	0,46	0,46	0,08	0,38	0,38	0,38	0,38
FPW	0,20	0,22	0,31	0,31	0,31	0,12	0,12	0,12	0,12	0,12
FPWREF	0,50	0,60	0,60	0,65	0,65	0,21	0,24	0,24	0,24	0,24
WCC	0,21	0,25	0,27	0,30	0,39	0,12	0,12	0,12	0,12	0,12
WCCREF	0,21	0,21	0,29	0,32	0,32	0,17	0,17	0,17	0,17	0,17
DK	0,61	0,63	0,77	0,82	0,82	0,57	0,59	0,65	0,70	0,70
DKREF	**0,74**	**0,86**	**0,98**	**1,00**	**1,00**	**0,59**	**0,71**	**0,78**	**0,78**	**0,78**
DKRPE	0,51	0,56	0,67	0,67	0,67	0,39	0,44	0,53	0,53	0,53
DKRPEREF	0,69	0,76	0,76	0,79	0,81	0,53	0,60	0,60	0,64	0,66
TK	0,55	0,55	0,65	0,65	0,65	0,46	0,46	0,49	0,49	0,49
TKREF	0,59	0,64	0,73	0,73	0,75	0,49	0,55	0,58	0,58	0,60
TKRPE	0,48	0,62	0,64	0,73	0,75	0,36	0,48	0,50	0,56	0,58
TKRPEREF	0,42	0,58	0,63	0,77	0,81	0,34	0,48	0,52	0,62	0,66

Tabelle 8.6 *: Auswertung von Cluster-Labeling-Verfahren durch Match@R.*

Cluster-Label. Jeder Term existiert nur einmal im Centroiden, so dass beispielsweise für die Referenzkategorie Antibiotika maximal eine relevante Phrase im Cluster-Label enthalten sein kann. Dieses begründet die schlechte Performanz datenzentrierter Verfahren im Experiment und erklärt gleichzeitig, warum eine einfache Nominalphrasen-Extraktion bessere Ergebnisse erzielt. Hier besteht die Möglichkeit, dass alle fünf Phrasen eines Cluster-Labels das Wort Antibiotika enthalten. Gleiches gilt für die untersuchten monothetischen Verfahren.

Alle Verfahren zeigen einen Anstieg der Precision@1 um bis zu 0,3 (FPWREF, partiell), wenn das Cluster-Labeling auf Grundlage der Referenzkategorisierung durchgeführt wird. Dies belegt, dass ein qualitativ hochwertiges Clustering die Qualität von Cluster-Labeln erhöht.

Schlussfolgerungen

- Monothetische Cluster-Labeling Verfahren schneiden am besten ab. In mehr als der Hälfte der Fälle (partielle Übereinstimmung) wird das korrekte Referenz-Label ermittelt.

Verfahren	MMR	
	(partiell)	(exakt)
NPE	0,47	0,35
NPEREF	0,55	0,38
RPE	0,26	0,23
RPEREF	0,25	0,23
FPW	0,24	0,12
FPWREF	0,56	0,23
WCC	0,26	0,12
WCC	0,24	0,17
DK	0,68	0,62
DKREF	**0,85**	**0,67**
DKRPE	0,57	0,45
DKRPEREF	0,73	0,58
TK	0,58	0,47
TKREF	0,65	0,54
TKRPE	0,58	0,44
TKRPEREF	0,56	0,46

Tabelle 8.7 *: Auswertung von Cluster-Labeling-Verfahren durch MMR.*

- Descriptive k-Means erzielt von allen untersuchten Cluster-Labeling-Verfahren für Precision@R, Match@R und MRR die besten Ergebnisse.
- Die Extraktion von Nominalphrasen erzielt bei allen Validierungsmaßen für $R = 1$ gute Ergebnisse und schneidet besser ab als datenzentrierte Verfahren. Hier zeigt sich die Überlegenheit von längeren Phrasen gegenüber einzelnen Termen in einem Cluster-Label.
- Datenzentrierte Verfahren besitzen den Nachteil, dass maximal eine relevante Phrase in einem Cluster-Label enthalten sein kann. Hier ist es besonders wichtig, einen hohen Precision@1-Wert zu erzielen. Die Experimente zeigen allerdings, dass dies nicht gelingt.
- Die Qualität von Cluster-Labeln steigt, wenn eine Referenzkategorisierung verwendet wird.
- Die Extraktion von häufigen Phrasen in Dokumenten eignet sich nicht, um Cluster-Label zu erzeugen. Alle Verfahren, die auf RPE basieren, schneiden schlechter ab als

Verfahren, die Nominalphrasen zur Bestimmung von potenziellen Cluster-Labeln einsetzen.

Bewertung von Cluster-Labeln durch interne Validierungsmaße

Nachdem gezeigt wurde, dass die in dieser Arbeit aufgestellten Label-Eigenschften für ein Cluster repräsentative Phrasen auswählen, werden die vorgestellten Cluster-Labeling-Verfahren anhand dieser evaluiert. Es ist zur Beantwortung von Fragestellung 6 zu zeigen, dass diese mit einem externen Maß korrelieren.

Die Experimente basieren auf den Korpora E1 bis E5.

NDCG-basiertes Validierungsmaß Das NDCG-basierte Validierungsmaß bewertet das Cluster-Labeling für Cluster eines Testkorpus. Tabelle 8.8 führt die gemittelten Werte über alle Testkorpora auf. Zur Bewertung der Relevanz einer Phrase in einem Cluster-Label werden die in Kapitel 7.2.2 aufgeführten Maße eingesetzt.

Die Ergebnisse sind in Tabelle 8.8 aufgeführt. Für die Evaluierung wurde auf Verfahren verzichtet, die zur Bestimmung von potenziellen Cluster-Labeln häufige Phrasen verwenden. Diese erzielten in zuvor durchgeführten Experimenten die schlechtesten Ergebnisse.

Die Auswertung zeigt, dass auch hier die untersuchten monothetischen Verfahren am besten abschneiden. Es erzielt erneut DKREF mit einem NDCG@1 von 0,87 das beste Ergebnis. Hier fällt allerdings der Unterschied zu Topical k-Means deutlich geringer aus. Die Performanz von TK ist sogar marginal besser als die von DK.

Im Vergleich zur Auswertung mittels Precision@R und Match@R zeigt sich allerdings, dass RPE nur marginal schlechtere Ergebnisse liefert als NPE. NPE erzielt für NDCG@1 einen Wert von 0,74. Die Differenz zu RPE liegt nur bei 0,02. Dieses ist darauf zurückzuführen, dass häufige Phrasen im Cluster viele Label-Eigenschaften erfüllen, die hier zur Relevanzbewertung herangezogen werden. Dieses gilt vor allem für die Überdeckung. Längere Phrasen kommen zudem weniger häufig in anderen Clustern vor als kurze, so dass auch die Trennschärfe und minimale Überlappung gegeben ist. Die Verständlichkeit kann auch erfüllt sein, falls es sich bei der häufigen Phrase um eine Nominalphrase handelt. Also liefert auch RPE passende Cluster-Label, die allerdings nur wenig mit Referenz-Labeln überlappen. Dies zeigte die Auswertung externer Validierungsmaße im vorigen Kapitel.

Im Unterschied zu den Ergebnissen von Precision@R fällt hier die Konstanz der Werte auf. Diese steigen von NDCG@1 bis NDCG@5 für die meisten Verfahren sogar leicht

	NDCG@R				
Verfahren	R=1	R=2	R=3	R=4	R=5
NPE	0,74	0,77	0,78	0,78	0,78
NPEREF	0,76	0,77	0,77	0,78	0,77
RPE	0,72	0,77	0,79	0,79	0,78
RPEREF	0,74	0,78	0,79	0,78	0,77
FPW	0,64	0,69	0,71	0,73	0,74
FPWREF	0,65	0,70	0,73	0,74	0,74
WCC	0,66	0,70	0,72	0,73	0,74
WCCREF	0,69	0,72	0,73	0,74	0,75
DK	0,82	0,86	0,88	0,88	0,89
DKREF	**0,87**	**0,91**	**0,92**	**0,93**	**0,93**
TK	0,84	0,87	0,89	0,89	0,90
TKREF	0,83	0,87	0,88	0,89	0,89

Tabelle 8.8 *: Auswertung von Cluster-Labeling-Verfahren durch NDCG@R.*

an. Alle Cluster-Labeling-Verfahren erzeugen somit Cluster-Label, deren fünf Phrasen jeweils ungefähr die gleiche Relevanz besitzen.

Schlussfolgerungen

- Das interne NDCG-basierte Validierungsmaß korreliert mit den externen Maßen Precision@R, Match@R und MRR. Es bewertet die monothetischen Cluster-Labeling-Verfahren am besten, gefolgt von Schlüsselwortbestimmungsverfahren und datenzentrierten Ansätzen. Liegt keine externe Referenz vor, eignet sich das interne Maß somit zur Bewertung eines Cluster-Labelings.
- Im Unterschied zu Precision@R, Match@R und MRR bewertet NDCG das Verfahren RPE beziehungsweise RPEREF gleichauf mit der Nominalphrasen-Extraktion.

Bewertung von Cluster-Labeln anhand der quantifizierten Label-Eigenschaften Im Gegensatz zu den bisherigen Experimenten werden die Verfahren Descriptive k-Means und Topical k-Means nicht modifiziert. Es wird durch das Validierungsmaß das erzeugte Cluster-Labeling mittels sechs Eigenschaften bewertet. Diese sind: Verständlichkeit (f_1), Überdeckung (f_2), Trennschärfe (f_3), Minimale Überlappung (f_4) und Eindeutigkeit ($f_{5,\text{intra}}$, $f_{5,\text{inter}}$). Es werden alle in dieser Arbeit vorgestellten Cluster-Labeling-Verfahren untersucht.

	Validierungsmaß						
Verfahren	f_1	f_2	f_3	f_4	$f_{5,\text{intra}}$	$f_{5,\text{inter}}$	f_6
NPE	0,76	0,67	0,38	**1,00**	**1,00**	0,98	0,99
NPEREF	0,78	**0,74**	0,53	**1,00**	**1,00**	0,92	0,98
RPE	0,80	0,55	-0,01	**1,00**	0,99	0,99	0,99
RPEREF	0,84	0,69	0,59	**1,00**	0,99	0,89	0,99
FPW	0,38	0,66	0,73	**1,00**	**1,00**	**1,00**	**1,00**
FPWREF	0,40	0,37	-4,56	**1,00**	**1,00**	0,96	0,99
WCC	0,39	0,68	0,69	**1,00**	**1,00**	0,98	**1,00**
WCCREF	0,40	0,64	0,48	**1,00**	**1,00**	0,95	**1,00**
STC	0,73	0,70	0,89	0,68	**1,00**	**1,00**	0,99
LINGO	0,75	0,42	0,89	0,69	**1,00**	**1,00**	**1,00**
DK	0,98	0,72	**0,93**	0,36	**1,00**	**1,00**	**1,00**
DKREF	**0,99**	0,72	**0,93**	0,32	**1,00**	**1,00**	**1,00**
TK	0,89	0,65	0,91	0,42	**1,00**	**1,00**	**1,00**
TKREF	0,96	0,69	0,90	0,39	**1,00**	0,99	**1,00**

Tabelle 8.9 *: Auswertung von Cluster-Labeling-Verfahren durch interne Validierungsmaße.*

Auch hier erzielt DKREF die besten Ergebnisse. Es wird eine sehr hohe Verständlichkeit von 0,99 erzielt, so dass fast keine einzelnen Terme als Cluster-Label verwendet werden. An zweiter Stelle erzielt DK eine Verständlichkeit von 0,98 gefolgt von TKREF mit 0,96. Die Verständlichkeit datenzentrierter Verfahren ist dagegen am schlechtesten. Hier erzielt FPW nur einen Wert von 0,38. Ein Wert unterhalb von 0,5 unterstreicht, dass ausschließlich einzelne Terme als Cluster-Label verwendet werden (Verständlichkeit = 0,5). Durch eine Wortstammreduktion sind diese oftmals nicht verständlich, so dass es sich um keine Nominalphrasen handelt (Verständlichkeit = 0). STC und Lingo erzielen eine Verständlichkeit von 0,7. STC erzeugt somit tendenziell längere Nominalphrasen.

Bei der Überdeckung der Dokumente in einem Cluster erzielt NPEREF mit 0,74 das beste Ergebnis und FPWREF mit 0,37 das schlechteste. Auch Lingo schneidet mit 0,42 schlecht ab. Lingo zeigte bereits bei der Ermittlung von Themen in Dokumentkollektionen die Tendenz zur Erzeugung kleiner Cluster.

Bei der Trennschärfe unterscheiden sich die untersuchten Cluster-Labeling-Verfahren stark voneinander. Hier erreichen DK und DKREF den besten Wert mit 0,93, gefolgt von TK mit 0,91. Auch STC und LINGO erzielen mit 0,89 eine hohe Trennschärfe. Dieses spricht dafür, dass beschreibungsbeachtende als auch beschreibungszentrierte Verfahren Cluster-Label derart auswählen, dass diese eindeutig die Dokumente eines Cluster be-

schreiben. Dagegen schneidet FPWREF erneut am schlechtesten ab. Die Trennschärfe liegt bei unter -4. Die schlechte Trennschärfe ist darauf zurückzuführen, dass datenzentrierte Verfahren in den Experimenten auf Wortstamm reduzierte, einzelne Terme als Cluster-Label erzeugen. Terme wie *db* oder *data* kommen natürlich auch in vielen Dokumenten anderer Cluster vor. Im Unterschied dazu würden die beschreibungszentrierte Verfahren eher Phrasen wie *database* oder *data quality* wählen, deren Trennschärfe allein aufgrund der Länge der Phrase höher ist.

Das Maß der minimalen Überlappung zeigt deutlich, welche Art des Clusterings einem Verfahren zu Grunde liegt. Datenzentrierte Verfahren, die ein polythetisches, exklusives Clustering durchführen, erzielen hier die höchste Bewertung mit 1,0. Dagegen überlappen sich die Dokumente von Clustern bei beschreibungsbeachtenden und beschreibungszentrierten Verfahren deutlich. Hier erzielt DKREF mit 0,32 den schlechtesten Wert. Lingo und STC liegen mit 0,69 bzw. 0,68 im Mittelfeld.

Die restlichen drei Validierungsmaße, die zwei Varianten der Eindeutigkeit und die Redundanzfreiheit, zeigen keine weiteren Tendenzen an. Die Werte der Maße für die Verfahren liegen zwischen 0,96 und 1,0. Dabei erzielen beschreibungsbeachtende und beschreibungszentrierte Verfahren leicht bessere Ergebnisse. Die Maße selbst erscheinen für die Bewertung eines Cluster-Labelings sinnvoll. Es zeigt sich aber, dass alle untersuchten Verfahren diese Eigenschaften nahezu vollständig erfüllen. Somit ist diesen weniger Gewicht zuzusprechen.

Werden die Validierungsmaße f_1 bis f_3 stärker gewichtet als die restlichen Maße, so ergibt sich erneut eine Präferenz für beschreibungsbeachtende und beschreibungszentrierte Verfahren. An zweiter Stelle liegen die Schlüsselwortbestimmungsverfahren, gefolgt von den datenzentrierten Ansätzen.

9 Zusammenfassung

In dieser Arbeit werden Verfahren zur Beschriftung von Clustern vorgestellt und validiert. Die Einleitung motiviert den Nutzen passender und verständlicher Cluster-Label im Kontext von Suchmaschinen. Cluster-Label sollen Textdokumente eines Clusters beschreiben und sich von Labeln anderer Cluster abgrenzen. Kapitel 2 führt in die theoretischen Grundlagen der Schlüsselwortbestimmung und des Clusterings ein. Für die Schlüsselwortbestimmung werden Verfahren wie die Nominalphrasen-Extraktion herausgestellt. Diese eignet sich insbesondere, um Cluster-Label zu erzeugen. Nominalphrasen gelten für den Menschen als verständlich und grammatikalisch korrekt. Das folgende Kapitel untersucht das Open Directory Project und Wikipedia empirisch. Es wird die Frage beantwortet, was verständliche Cluster-Label sind. Es können Rückschlüsse für das Cluster-Labeling gezogen werden. Auf diesen basieren die in Kapitel 4 vorgestellten Eigenschaften, die ein Cluster-Label erfüllen sollte. Diese sind: Verständlichkeit, Überdeckung, Trennschärfe, minimale Überlappung, Eindeutigkeit und Redundanzfreiheit. Für ein hierarchisches Cluster-Labeling sind dies zusätzlich: Hierarchische Konsistenz und Kohärenz von Schwesterknoten. Kapitel 5 ordnet Cluster-Labeling-Verfahren in drei Kategorien ein: datenzentriert, beschreibungsbeachtend und beschreibungszentriert. Hierzu werden jeweils entsprechende Verfahren vorgestellt und deren Vor- und Nachteile diskutiert. Im folgenden Kapitel wird ein neues beschreibungszentriertes Cluster-Labeling-Verfahren namens Topical k-Means vorgestellt. Die Cluster-Labeling-Verfahren werden anschließend in Kapitel 8 mit Hilfe der in Kapitel 7 beschriebenen externen und internen Validierungsmaße gegenübergestellt. Hierzu wird ein neues internes, NDCG-basiertes Validierungsmaß vorgeschlagen.

Die Experimente zeigen, dass beschreibungszentrierte als auch beschreibungsbeachtende Cluster-Labeling-Verfahren denen datenzentrierter Verfahren deutlich überlegen sind. Hierbei erzielte Descriptive k-Means die besten Ergebnisse mit einer Precision@1 von 0,74 (partiell) und 0,59 (exakt). Überraschend gute Ergebnisse liefert die Nominalphrasen-Extraktion mit einer Precision@1 von 0,46 (partiell) und 0,31 (exakt). Es werden einfach die besten (häufigsten) Nominalphrasen als Cluster-Label verwendet. Da hier keine weite-

ren Heuristiken angenommen werden, erzielen reine Schlüsselwortbestimmungsverfahren aber eine schlechtere Trennschärfe der Cluster-Label. Hier liegt die Stärke beschreibungszentrierter Verfahren, die ihre Cluster-Label explizit anhand solcher Anforderungen auswählen. Allerdings zeigte die Vorstellung der Verfahren in Kapitel 5 auch, dass vor allem beschreibungsbeachtende und beschreibungszentrierte Verfahren benutzerdefinierte Parameter besitzen, die teilweise schwer einzustellen sind. Dieses gilt für Lingo als auch Descriptive k-Means.

Für die in Kapitel 4 vorgestellten intrinsischen Eigenschaften von Cluster-Labeln konnte gezeigt werden, dass deren Quantifizierung mit externen Maßen wie Precision@R, Match@R und MRR korreliert. Es zeigte sich, dass die Eigenschaften Eindeutigkeit und Redundanzfreiheit nicht stark genug sind, um Cluster-Labeling-Verfahren voneinander abzugrenzen. Eine reine Bewertung von Cluster-Labeling-Verfahren anhand Precision@R, Match@R und MRR zeigt nicht die eigentliche Qualität eines Cluster-Labelings auf. Es ist erforderlich, die in dieser Arbeit vorgestellten Validierungsmaße weiterzuentwickeln. Die Maße könnten um eine Eigenschaft erweitert werden, die alle Phrasen als Cluster-Label bestraft, die in der Schnittmenge von Termen der Textdokumente eines Clusters liegen. Für ein Cluster über Filme wären dies beispielsweise *Film* oder *Schauspieler*.

Die Experimente belegen den starken Einfluss der Schlüsselwortbestimmung auf die Qualität des Cluster-Labelings. Es ist anzuraten, weitere Schlüsselwortbestimmungsverfahren zu untersuchen. Hier bieten sich auf Graphen basierte Verfahren wie TextRank an (Mihalcea u. Tarau, 2004). Allerdings zeigt die Analyse von Wikipedia, dass die eigentlichen Themen von Textdokumenten nicht immer im Text selbst vorkommen müssen. Es ist daher zu überlegen, ob zusätzlich zur Schlüsselwortbestimmung externes Wissen herangezogen wird. Hier bietet sich Wikipedia als Wissensquelle an. Zu diesem Thema existieren bereits Arbeiten von Stein u. Meyer zu Eißen (2007), Coursey u. Mihalcea (2009) und Carmel u. a. (2009).

Die Aufmerksamkeit, die das Cluster-Labeling in den letzten Jahren seitens der Forschung bekommt, eröffnet die Möglichkeit für viele weitere Verfahren. Dieses verlangt allerdings auch nach einem einheitlichen Framework zur Validierung. Hierzu trägt diese Arbeit ihren Teil bei.

Literaturverzeichnis

[Aas u. Eikvil 1999] Aas, Kjersti ; Eikvil, Line: Text Categorization: A Survey. (1999), November, S. 1–38

[Allan 2002] Allan, James: Introduction to topic detection and tracking. (2002), S. 1–16. ISBN 0–7923–7664–1

[Allan u. a. 2000] Allan, James ; Lavrenko, Victor ; Jin, Hubert: First story detection in TDT is hard. In: *CIKM '00: Proceedings of the ninth international conference on Information and knowledge management*. New York, NY, USA : ACM, 2000. – ISBN 1–58113–320–0, S. 374–381

[Allan u. a. 1998] Allan, James ; Papka, Ron ; Lavrenko, Victor: On-line new event detection and tracking. In: *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA : ACM, 1998. – ISBN 1–58113–015–5, S. 37–45

[Barker u. Cornacchia 2000] Barker, Ken ; Cornacchia, Nadia: Using Noun Phrase Heads to Extract Document Keyphrases. In: *English* (2000)

[Baxendale 1958] Baxendale, P B.: Machine-Made Index for Technical Literature - an Experiment. In: *IBM Journal of Research and Development* 2 (1958), S. 354–361

[Blei u. a. 2003] Blei, D M. ; Ng, A Y. ; Jordan, M I.: Latent dirichlet allocation. In: *The Journal of Machine Learning Research* 3 (2003), S. 993–1022

[Blum u. Mitchell 1998] Blum, A ; Mitchell, T: Combining labeled and unlabeled data with co-training. In: *Proceedings of the eleventh annual conference on Computational learning theory*, ACM, 1998, 100

[Brooks u. Montanez 2006] Brooks, C.H. ; Montanez, Nancy: Improved annotation of the blogosphere via autotagging and hierarchical clustering. In: *Proceedings of the 15th international conference on World Wide Web*, ACM, 2006, 632

[Bun u. Ishizuka 2001] Bun, K ; Ishizuka, M: Emerging topic tracking system in WWW. In: *Knowledge-Based Systems* 19 (2001), Juli, Nr. 3, 164–171. http://dx.doi.org/10.1016/j.knosys.2005.11.008. – DOI 10.1016/j.knosys.2005.11.008. – ISSN 09507051

[Carbonell u. Goldstein 1998] Carbonell, J ; Goldstein, J: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: *Proceedings of the 21st annual international ACM SIGIR {...}* (1998), Januar. http://portal.acm.org/citation.cfm?id=291025&dl=GUIDE,

[Carmel u. a. 2009] Carmel, D ; Roitman, H ; Zwerdling, N: Enhancing cluster labeling using wikipedia. In: *Proceedings of the 32nd international ACM SIGIR {...}* (2009), Januar. http://portal.acm.org/citation.cfm?id=1571941.1571967

[Carpineto u. a. 2009] Carpineto, Claudio ; Osiński, Stanislaw ; Romano, Giovanni ; Weiss, Dawid: A survey of Web clustering engines. In: *ACM Computing Surveys* 41 (2009), Nr. 3, 1–38. http://dx.doi.org/10.1145/1541880.1541884. – DOI 10.1145/1541880.1541884. – ISSN 03600300

[Coursey u. Mihalcea 2009] Coursey, Kino ; Mihalcea, R.: Topic identification using Wikipedia graph centrality. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, Association for Computational Linguistics, 2009 (Juni), 117–120

[Cutting u. a. 1992] Cutting, Douglass R. ; Karger, David R. ; Pedersen, Jan O. ; Tukey, John W.: Scatter/Gather: a cluster-based approach to browsing large document collections. In: *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '92* (1992), 318–329. http://dx.doi.org/10.1145/133160.133214. – DOI 10.1145/133160.133214. ISBN 0897915232

[Das 2007] Das, Dipanjan: A Survey on Automatic Text Summarization Single-Document Summarization. In: *Language* (2007), S. 1–31

[Deerwester u. a. 1990] Deerwester, Scott ; Dumais, S.T. ; Furnas, G.W. ; Landauer, T.K. ; Harshman, Richard: Indexing by latent semantic analysis. In: *Journal of the American society for information science* 41 (1990), Nr. 6, 391–407. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.1152&rep=rep1&type=pdf

[Dhillon 2001] Dhillon, I S.: Co-clustering documents and words using bipartite spectral graph partitioning. In: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (2001), S. 269–274

[Diederich u. a. 2007] Diederich, J. ; Balke, W.T. ; Thaden, Uwe: Demonstrating the semantic growbag: automatically creating topic facets for faceteddblp. In: *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, ACM, 2007, 505

[DMOZ 2004] DMOZ: *How to suggest a site to the Open Directory*. `http://www.dmoz.org/add.html`. Version: 2004

[Edmundson 1969] Edmundson, H P.: New Methods in Automatic Extracting. In: *Computing* 16 (1969), Nr. 2, S. 264–285

[Fellbaum u. Miller 2010] Fellbaum, Christiane ; Miller, George A.: *WordNet 3.0 Statistics*. `http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html#toc2`. Version: 2010

[Ferragina u. Gulli 2005] Ferragina, P ; Gulli, A: A personalized search engine based on web-snippet hierarchical clustering. In: *Special interest tracks and posters of the 14th {...}* (2005), Januar. `http://portal.acm.org/citation.cfm?id=1062745.1062760`

[Ferragina u. Gullì 2007] Ferragina, Paolo ; Gullì, Antonio: Anatomy of a Hierarchical Clustering Engine for Web-page, News and Book. In: *Citeseer* (2007). `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1.8536&rep=rep1&type=pdf`

[Fujimura u. a. 2006] Fujimura, Ko ; Toda, H. ; Inoue, T. ; Hiroshima, N. ; Kataoka, R. ; Sugizaki, M.: Blograngera multi-faceted blog search engine. In: *Proceedings of the WWW 2006 3nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, Citeseer, 2006

[Geraci u. a. 2006] Geraci, F ; Pellegrini, M ; Maggini, M: Cluster generation and labeling for web snippets: a fast, accurate hierarchical solution. In: *Internet Mathematics* (2006), Januar. `http://akpeters.metapress.com/index/L304G3K3KV66X709.pdf`

[Golder u. Huberman 2006] Golder, Scott A. ; Huberman, Bernardo A.: Usage patterns of collaborative tagging systems. In: *Journal of Information Science* 32 (2006), Nr. 2, 198–208. `http://dx.doi.org/10.1177/0165551506062337`. – DOI 10.1177/0165551506062337. – ISSN 0165–5515

[Goldstein u. a. 1999] Goldstein, J ; Kantrowitz, M ; Mittal, V ; Carbonell, J: Summarizing text documents: sentence selection and evaluation metrics. In: *Proceedings of the 22nd annual international ACM SIGIR {...}* (1999), Januar. http://portal.acm.org/citation.cfm?id=312665

[Griffiths u. Steyvers 2004] Griffiths, T ; Steyvers, M: Finding scientific topics. In: *Proceedings of the National Academy of Sciences* (2004), Januar. http://www.pnas.org/cgi/content/abstract/101/suppl_1/5228

[Gusfield 2007] Gusfield, Dan: *Algorithms on strings, trees, and sequences : computer science and computational biology.* ambridge Univ. Press, 2007. – 94–207 S. http://www.worldcat.org/isbn/0521585198. – ISBN 0521585198

[Harman 1986] Harman, Donna W.: An experimental study of factors important in document ranking. In: *SIGIR '86: Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval.* New York, NY, USA : ACM, 1986. – ISBN 0–89791–187–3, S. 186–193

[Hofmann 1999] Hofmann, T: Probabilistic latent semantic indexing. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (1999), S. 50–57

[Hovy u. Lin 1999] Hovy, Eduard ; Lin, Chin-yew: Automated Text Summarization in SUMMARIST. In: *Evaluation* (1999)

[Hulth 2003] Hulth, Anette: Improved automatic keyword extraction given more linguistic knowledge. In: *Proceedings of the 2003 conference on Empirical methods in natural language processing* Bd. 10, 2003, 216–223

[Jain u. a. 1999] Jain, A K. ; Murty, M N. ; Flynn, P J.: Data clustering: a review. In: *ACM computing surveys (CSUR)* 31 (1999), Nr. 3, S. 264–323

[Järvelin u. Kekäläinen 2002] Järvelin, Kalervo ; Kekäläinen, Jaana: Cumulated gain-based evaluation of IR techniques. In: *ACM Transactions on Information Systems* 20 (2002), Oktober, Nr. 4, 446. http://dx.doi.org/10.1145/582415.582418. – DOI 10.1145/582415.582418. – ISSN 10468188

[Joachims 1997] Joachims, Thorsten: Text Categorization and Support Vector Machines: Learning with Many Relevant Features. (1997). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.6124

[Jolliffe 2002] Jolliffe, I T.: *Principal Component Analysis.* Second. Springer, 2002 http://www.worldcat.org/isbn/0387954422. – ISBN 0387954422

[Justeson u. Katz 1995] Justeson, J ; Katz, S: Technical terminology: some linguistic properties and an algorithm for identification in text. In: *Natural Language Engineering* (1995), S. 9–27

[Karp 1972] Karp, Richard M.: Reducibility Among Combinatorial Problems. In: Miller, R E. (Hrsg.) ; Thatcher, J W. (Hrsg.): *Complexity of Computer Computations.* Plenum Press, 1972, S. 85–103

[Kassner u. a. 2008] Kassner, L ; Nastase, V ; Strube, M: Acquiring a taxonomy from the German Wikipedia. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation* (2008)

[Klüger 2006] Klüger, Karsten: *Automatische Extraktion von Schlüsselwörtern aus Text.* Juni 2006

[Kohonen u. a. 1999] Kohonen, T ; Kaski, S ; Lagus, K ; Salojärvi, J: Self organization of a massive text document collection. In: *Kohonen maps* (1999), Januar

[Kohonen 1997] Kohonen, Teuvo: *Self-Organizing Maps (2nd ed)(Springer Series in Information Sciences, 30).* Springer, 1997. – ISBN 3540620176

[Krishnapuram u. Kummamuru 2003] Krishnapuram, Raghu ; Kummamuru, Krishna: Automatic taxonomy generation: Issues and possibilities. In: *Lecture notes in computer science* (2003), 52–63. http://cat.inist.fr/?aModele=afficheN&cpsidt=15509131. – ISBN 3–540–40383–3

[Kumar u. Srinathan 2008] Kumar, Niraj ; Srinathan, Kannan: Automatic keyphrase extraction from scientific documents using N-gram filtration technique. In: *DocEng '08: Proceeding of the eighth ACM symposium on Document engineering.* New York, NY, USA : ACM, 2008. – ISBN 978–1–60558–081–4, S. 199–208

[Kummamuru u. a. 2004] Kummamuru, K ; Lotlikar, R ; Roy, S ; Singal, K: A hierarchical monothetic document clustering algorithm for summarization and browsing {...}. In: *Proceedings of the 13th international conference on {...}* (2004), Januar. http://portal.acm.org/citation.cfm?id=988762&dl=

[Lawrie u. Croft 2003] Lawrie, D ; Croft, W: Generating hierarchical summaries for web searches. In: *Proceedings of the 26th annual international ACM {...}* (2003), Januar. http://portal.acm.org/citation.cfm?id=860435.860549&type=series

[Lee u. Chun 2007] Lee, SOK ; Chun, AHW: Automatic Tag Recommendation for the Web 2.0 Blogosphere Using Collaborative Tagging and Hybrid ANN Semantic Structures. In: *Proceedings of the 6th WSEAS Internation Conference on Applied Computer Science*, City University of Hong Kong, 2007, 88–93

[Lewis u. Ringuette 1994] Lewis, D.D. ; Ringuette, M.: A comparison of two learning algorithms for text categorization. In: *Third annual symposium on document analysis and information retrieval* Bd. 33, Citeseer, 1994, 81–93

[Lin u. Hovy 2003] Lin, C Y. ; Hovy, E: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (2003), S. 78

[Liu u. Croft 2004] Liu, X ; Croft, W B.: Cluster-based retrieval using language models. In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (2004), S. 186–193

[Luhn 1958] Luhn, H.P.: The Automatic Creation of Literature Abstracts. (1958), Nr. April, S. 159–165

[Manning u. a. 2002] Manning, C ; Schütze, H ; 2002: Foundations of statistical natural language processing. In: *MIT Press* (2002). http://www.mitpressjournals.org/doi/pdfplus/10.1162/coli.2000.26.2.277

[Manning u. Schütze 1999] Manning, Christopher D. ; Schütze, Hinrich: *Foundations of Statistical Natural Language Processing.* Cambridge, Massachusetts : The {MIT} Press, 1999

[Maslowska 2003] Maslowska, Irmina: Phrase-based hierarchical clustering of Web search results. In: *Lecture notes in computer science* (2003), 555–562. http://cat.inist.fr/?aModele=afficheN&cpsidt=15283050. – ISBN 3–540–01274–5

[Mei u. a. 2006] Mei, Q ; Liu, C ; Su, H ; Zhai, C: A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In: *Proceedings of the 15th internatio-*

nal conference on {...} (2006), Januar. http://portal.acm.org/citation.cfm?id=1135777.1135857

[Mei u. a. 2007] Mei, Q ; Shen, X ; Zhai, C: Automatic labeling of multinomial topic models. In: *Proceedings of the 13th ACM SIGKDD international {...}* (2007), Januar. http://portal.acm.org/citation.cfm?id=1281246

[Mei u. Zhai 2006] Mei, Q ; Zhai, C: A mixture model for contextual text mining. In: *{...} conference on Knowledge discovery and data mining* (2006), Januar. http://portal.acm.org/citation.cfm?id=1150402.1150482

[Meyer zu Eißen 2007] Meyer zu Eissen, Sven: *On Information Need and Categorizing Search*. Version: Feb 2007. http://ubdata.uni-paderborn.de/ediss/17/2007/meyer_zu/

[Mihalcea 2007] Mihalcea, R.: Using wikipedia for automatic word sense disambiguation. In: *Proceedings of NAACL HLT* Bd. 2007, 2007

[Mihalcea u. Tarau 2004] Mihalcea, R ; Tarau, P: TextRank: Bringing order into texts. In: *Proceedings of EMNLP* (2004), Januar. http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Mihalcea.pdf

[Mishne 2006] Mishne, Gilad: Autotag: a collaborative approach to automated tag assignment for weblog posts. In: *Proceedings of the 15th international conference on World Wide Web*, ACM, 2006, 954

[Muthukrishnan u. a. 2008] Muthukrishnan, P. ; Gerrish, J. ; Radev, D.R.: Detecting multiple facets of an event using graph-based unsupervised methods. In: *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, Association for Computational Linguistics, 2008 (August), 609616

[Nagao u. Mori 1994] Nagao, Makoto ; Mori, Shinsuke: A new method of N-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese. In: *Proceedings of the 15th conference on Computational linguistics*. Kyoto, Japan : Association for Computational Linguistics, 1994, S. 611–615

[Nastase u. Strube 2008] Nastase, V ; Strube, M: Decoding Wikipedia categories for knowledge acquisition. In: *Proceedings of the AAAI* 8 (2008)

[Navigli 2009] Navigli, Roberto: Word sense disambiguation. In: *ACM Computing Surveys* 41 (2009), Nr. 2, 1–69. http://dx.doi.org/10.1145/1459352.1459355. – DOI 10.1145/1459352.1459355. – ISSN 03600300

[Nguyen u. a. 2009] Nguyen, C ; Phan, X ; Horiguchi, S: Web Search Clustering and Labeling with Hidden Topics. In: *ACM Transactions on Asian Language Information {...}* (2009), Januar. http://portal.acm.org/citation.cfm?id=1568292.1568295

[Nigam 2001] Nigam, Kamal P.: *Using unlabeled data to improve text classification.* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.63.7998&rep=rep1&type=pdf. Version: 2001

[Nupedia 2003] Nupedia: *Nupedia, the free encyclopedia.* http://web.archive.org/web/20030810153103/www.nupedia.com. Version: 2003

[Osinski u. a. 2004] Osinski, S. ; Stefanowski, J. ; Weiss, D.: Lingo: Search results clustering algorithm based on singular value decomposition. In: *Intelligent information processing and web mining: proceedings of the International IIS: IIPWM'04 Conference held in Zakopane, Poland, Mai 17-20, 2004*, Springer Verlag, 2004, 359

[Ponte u. Croft 1998] Ponte, J ; Croft, W: A language modeling approach to information retrieval. In: *Proceedings of the 21st annual international ACM {...}* (1998), Januar. http://portal.acm.org/citation.cfm?id=291008&dl=GUIDE,

[Ponzetto u. Strube 2007] Ponzetto, S.P. ; Strube, M.: Deriving a large scale taxonomy from Wikipedia. In: *Proceedings of the national conference on artificial intelligence* Bd. 22, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, November 2007, 1440

[Popescul u. Ungar 2000] Popescul, A ; Ungar, L: Automatic labeling of document clusters. In: *Unpublished manuscript* (2000), Januar. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.141&rep=rep1&type=pdf

[Porter 1980] Porter, M F.: An algorithm for suffix stripping. In: *Program* 14 (1980), Nr. 3, 130–137. http://portal.acm.org/citation.cfm?id=275705. ISBN 1–55860–454–5

[Radev u. a. 2002a] Radev, D.R. ; Hovy, Eduard ; McKeown, K.: Introduction to the special issue on summarization. In: *Computational Linguistics* 28

(2002), Nr. 4, 399–408. http://www.mitpressjournals.org/doi/abs/10.1162/089120102762671927

[Radev u. a. 2002b] Radev, D.R. ; Qi, Hong ; Wu, H. ; Fan, W.: Evaluating web-based question answering systems. In: *Ann Arbor* 1001 (2002), 48109. http://clair.si.umich.edu/~radev/papers/NSIR.pdf

[Radev u. a. 2000] Radev, Dragomir R. ; Jing, Hongyan ; Budzikowska, Malgorzata: Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In: *NAACL-ANLP 2000 Workshop on Automatic Summarization*. Morristown, NJ, USA : Association for Computational Linguistics, 2000, S. 21–30

[Rao u. a. 2002] Rao, D ; Deepak, P ; Khemani, D: Corpus Based Unsupervised Labeling of Documents. In: *aaai.org* (2002). https://www.aaai.org/Papers/FLAIRS/2006/Flairs06-063.pdf

[Rauber u. Merkl 1999] Rauber, A ; Merkl, Dieter: LabelSOM: On the labeling of self-organizing maps. In: *Proc. International Joint Conference on Neural {...}* (1999), Januar. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.36.3931&rep=rep1&type=pdf

[Rosenberg u. Hirschberg 2007] Rosenberg, Andrew ; Hirschberg, Julia: V-Measure : A conditional entropy-based external cluster evaluation measure. In: *Computational Linguistics* (2007), Nr. Juni, S. 410–420

[Sacco 2000] Sacco, Giovanni M.: Dynamic Taxonomies: A Model for Large Information Bases. In: *IEEE Trans. on Knowl. and Data Eng.* 12 (2000), Nr. 3, S. 468–479. http://dx.doi.org/http://dx.doi.org/10.1109/69.846296. – DOI http://dx.doi.org/10.1109/69.846296. – ISSN 1041–4347

[Salton 1971] Salton, G: *The SMART Retrieval System—Experiments in Automatic Document Processing*. Upper Saddle River, NJ, USA : Prentice-Hall, Inc., 1971 http://portal.acm.org/citation.cfm?id=1102022

[Salton u. a. 1975] Salton, G. ; Wong, A. ; Yang, C. S.: A Vector Space Model for Automatic Indexing. In: *Communications of the ACM* 18 (1975), November, Nr. 11, 613–620. http://dx.doi.org/10.1145/361219.361220. – DOI 10.1145/361219.361220. – ISSN 00010782

[Salton u. Yang 1973] SALTON, G. ; YANG, C.S.: On the specification of term values in automatic indexing. In: *Journal of documentation* (1973). http://www.emeraldinsight.com/Insight/ViewContentServlet?contentType=Article&Filename=/published/emeraldfulltextarticle/pdf/2780290401.pdf

[Sanderson u. Croft 1999] SANDERSON, M ; CROFT, Bruce: Deriving concept hierarchies from text. In: *Proceedings of the 22nd annual international ACM {...}* (1999), Januar. http://portal.acm.org/citation.cfm?id=312679&dl=GUIDE,

[Santorini 1990] SANTORINI, B.: Part-of-speech tagging guidelines for the Penn Treebank Project. In: *University of Pennsylvania, 3rd Revision, 2nd Printing* (1990). http://www.personal.psu.edu/faculty/x/x/xxl13/teaching/sp07/apling597e/resources/Tagset.pdf

[Schonhofen 2006] SCHONHOFEN, Peter: Identifying Document Topics Using the Wikipedia Category Network. In: *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*. Washington, DC, USA : IEEE Computer Society, 2006. – ISBN 0–7695–2747–7, S. 456–462

[SIGIR 2006] SIGIR: *SIGIR 2006: Workshop on Faceted Search*. August 2006

[Spärck Jones 1972] SPÄRCK JONES, Karen: A statistical interpretation of term specificity and its application in retrieval. In: *Journal of Documentation* 60 (1972), Nr. 5, 493–502. http://dx.doi.org/10.1108/00220410410560573. – DOI 10.1108/00220410410560573. – ISSN 0022–0418

[Stefanowski u. Weiss 2003] STEFANOWSKI, Jerzy ; WEISS, Dawid: Carrot and Language Properties in Web Search Results Clustering. In: *AWIC*, 2003, S. 240–249

[Stefanowski u. Weiss 2007] STEFANOWSKI, Jerzy ; WEISS, Dawid: Comprehensible and Accurate Cluster Labels in Text Clustering. In: *RIAO*, 2007

[Stein u. Meyer zu Eißen 2004] STEIN, Benno ; MEYER ZU EISSEN, Sven: Topic Identification: Framework and Application. In: TOCHTERMANN, Klaus (Hrsg.) ; MAURER, Hermann (Hrsg.): *Proceedings of the 4th International Conference on Knowledge Management (I-KNOW 04), Graz, Austria*. Graz, Austria : Know-Center, Juli 2004 (Journal of Universal Computer Science). – ISSN 0948–6968, S. 353–360

[Stein u. Meyer zu Eißen 2007] STEIN, Benno ; MEYER ZU EISSEN, Sven: Topic-Identifikation: Formalisierung, Analyse und neue Verfahren. (2007), S. 1–7

[Syed u. a. 2008] Syed, Z ; Finin, T ; Joshi, A: Wikipedia as an ontology for describing documents. In: *Proceedings of the second international conference {...}* (2008), Januar. http://www.aaai.org/Papers/ICWSM/2008/ICWSM08-024.pdf

[Tomokiyo u. Hurst 2003] Tomokiyo, T. ; Hurst, M.: A language model approach to keyphrase extraction. In: *Proceedings of the ACL Workshop on Multiword Expressions*, 2003, 3440

[Treeratpituk u. Callan 2006] Treeratpituk, P ; Callan, J: Automatically labeling hierarchical clusters. In: *Proceedings of the 2006 international conference on {...}* (2006), Januar. http://portal.acm.org/citation.cfm?id=1146598.1146650

[Tseng 1998] Tseng, Yuen-hsien: Multilingual Keyword Extraction for Term Suggestion. In: *Processing* (1998), S. 377–378

[Tunkelang 2009] Tunkelang, Daniel: *Faceted Search*. Morgan {&} Claypool Publishers, 2009 (SynThese Lectures on Information Concepts, Retrieval, and Services)

[Turney 2000] Turney, P.D.: Learning algorithms for keyphrase extraction. In: *Information Retrieval* 2 (2000), Nr. 4, 303–336. http://www.springerlink.com/index/T742610M31G64X52.pdf

[Wayne 2000] Wayne, C: Multilingual topic detection and tracking: Successful research enabled by corpora and {}. In: *Language Resources and Evaluation Conference ({}* (2000), Januar. http://gandalf.aksis.uib.no/non/lrec2000/pdf/168.pdf

[Weiss 2006] Weiss, D.: *Descriptive clustering as a method for exploring text collections*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.74.3051&rep=rep1&type=pdf. Version: 2006

[Wikipedia 2010a] Wikipedia, The Free E.: *Wikipedia Statistics: Article count (official)*. http://stats.wikimedia.org/EN/TablesArticlesTotal.htm. Version: 2010

[Wikipedia 2010b] Wikipedia, The Free E.: *Wikipedia:Category names (Namenskonventionen)*. http://en.wikipedia.org/wiki/Wikipedia:Naming_conventions_(categories). Version: 2010

[de Winter u. de Rijke 2007] Winter, W de ; Rijke, M de: Identifying facets in query-biased sets of blog posts. In: *ICWSM* (2007), Januar. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.74.9124&rep=rep1&type=pdf

[Witten u. a. 1999] Witten, Ian H. ; Paynter, Gordon W. ; Frank, Eibe ; Gutwin, Carl ; Nevill-Manning, Craig G.: KEA: practical automatic keyphrase extraction. In: *DL '99: Proceedings of the fourth ACM conference on Digital libraries*. New York, NY, USA : ACM, 1999. – ISBN 1–58113–145–3, S. 254–255

[Wu u. a. 2006] Wu, Harris ; Zubair, Mohammad ; Maly, Kurt: Harvesting social knowledge from folksonomies. In: *Proceedings of the seventeenth conference on Hypertext and hypermedia - HYPERTEXT '06* (2006), 111. http://dx.doi.org/10.1145/1149941.1149962. – DOI 10.1145/1149941.1149962. ISBN 1595934170

[Yang 1999] Yang, Y: An evaluation of statistical approaches to text categorization. In: *Information retrieval* 1 (1999), Nr. 1, 69–90. http://www.springerlink.com/index/X3N6633584015P59.pdf

[Yang u. Pedersen 1997] Yang, Y ; Pedersen, J: A comparative study on feature selection in text categorization. In: *MACHINE LEARNING-INTERNATIONAL WORKSHOP {...}* (1997), Januar. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.32.9956&rep=rep1&type=pdf

[Yang u. Wilbur 1996] Yang, Y. ; Wilbur, J.: Using corpus statistics to remove redundant words in text categorization. In: *Journal of the American Society for Information Science* 47 (1996), Nr. 5, 357–369. http://www.aquaphoenix.com/presentations/candidacy/yang96.pdf

[Yarowsky 1992] Yarowsky, David: Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In: *International Conference On Computational Linguistics* (1992). http://portal.acm.org/citation.cfm?id=992133.992140

[Zamir u. Etzioni 1998] Zamir, O ; Etzioni, O: Web document clustering: A feasibility demonstration. In: *Proceedings of the 21st annual international ACM {...}* (1998), Januar. http://portal.acm.org/citation.cfm?id=290941.290956&dl=ACM&dl=ACM&type=series&idx=290941&par...

[Zamir u. Etzioni 1999] Zamir, O ; Etzioni, O: Grouper: a dynamic clustering interface to Web search results. In: *Computer Networks-the International Journal of {...}* (1999), Januar. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.115.7735&rep=rep1&type=pdf

[Zhang 2002] Zhang, Dong: *Towards Web Information Clustering.* 2002

A Analyse zur Ermittlung von Themen

Dieses Kapitel vervollständigt die Auswertung zur Ermittlung von Themen in den Testkorpora E1 bis E4 aus Kapitel 8.2.1. Es werden folgende Cluster-Labeling-Verfahren untersucht: k-Means (Weighted Centroid Covering), Lingo und Descriptive k-Means.

k-Means (Weighted Centroid Covering)

k-Means erzeugt jeweils die zuvor durch k festgelegte Anzahl an Kategorien im jeweiligen Testkorpus. Die Ergebnisse sind in Abbildung A.1 auf Seite 136 in der linken Hälfte zusammengefasst. Zusätzlich wird für jeden Korpus experimentell ein zweites k, $k \in [1; 20]$, durch Auswahl des besten F-Measure Wertes ermittelt. Die Ergebnisse hierzu sind in der rechten Hälfte der Abbildung A.1 zu sehen.

E1 Die vier thematisch verschiedenen Kategorien MySQL, Antibiotika, A Space Odyssey und Bluebirds werden eindeutig erkannt. Die Cluster-Label bestehen hier aus nur einem einzigen Term, der allerdings ausreichend repräsentativ für das jeweilige Cluster ist.

E2 k-Means grenzt ebenso wie Topical k-Means die inhaltlich ähnlichen Kategorien schlechter voneinander ab. Nur die Kategorien Killer's Kiss (*london*) sowie Eyes Wide Shut (*shut*) werden eindeutig erkannt. Im Vergleich zu anderen Clustering-Labeling Verfahren erzeugt k-Means zwei Cluster mit weniger als fünf Dokumenten: *juli* und *class*. Beide besitzen keine Aussagekraft und lassen keine Rückschlüsse auf die im Cluster enthaltenen Dokumente zu. Cluster-Labeling-Verfahren wie Topical k-Means berücksichtigen dagegen Cluster mit weniger als fünf Dokumenten nicht.

Ebenso wie Topical k-Means fasst k-Means alle Kategorien in einem gemeinsamen Cluster *greatest* zusammen. Das Cluster-Label steht vermutlich für *greatest films*. Hier zeigt sich das Problem datenzentrierter Verfahren. Cluster-Label sind durch Stemming

und Reduktion auf einzelne Terme weniger verständlich als Cluster-Label von Topical k-Means, Lingo oder Descriptive k-Means.

E3 Die Ausreißer-Kategorie Antibiotika wird eindeutig erkannt. Es fällt auf, dass Dokumente der Kategorie PostgreSQL in nahezu allen Clustern auftreten. Unter anderem auch fälschlicherweise im Cluster für die Kategorie Antibiotika. Internetseiten über PostgreSQL sind thematisch ähnlich zu anderen Datenbank-Kategorien, so dass Dokumente in allen Clustern vorkommen können. Dennoch war Topical k-Means in der Lage, PostgreSQL als eigenständiges Cluster zu erkennen. Es gab nur geringfügige Überschneidungen mit dem Cluster *MySQL*.

E4 Auffällig ist, dass Cluster, die in E2 und E3 erkannt wurden, nun nicht mehr identifiziert werden. Hierzu zählen unter anderem *shut* und *london*. Jedoch wird nun die Kategorie Vertigo repräsentiert. Diese wurde in E2 nicht eindeutig erkannt.

k-Means gelingt eine Generalisierung der Film-Kategorie durch das Cluster *film*. Es sind keine Dokumente der Datenbank-Kategorie enthalten.

Lingo

Lingo erzeugt für die Testkorpora E1 bis E4 zwischen 6 und 15 Cluster. Die Ergebnisse sind in Abbildung A.2 auf Seite 137 zusammengefasst.

E1 Für den Testkorpus E1 werden alle vier nicht verwandten Themengebiete eindeutig identifiziert. Im Unterschied zu Topical k-Means ist hierbei die Anzahl der Cluster mit 9 gegenüber 18 deutlich geringer. Somit bekommt ein Nutzer einen schnelleren Überblick über den Inhalt einer Dokumentkollektion.

E2 Lingo erkennt die verschiedenen Filme besser als alle bisher betrachteten Cluster-Labeling-Verfahren. Erstmalig werden die Filme Vertigo, Eyes Wide Shut, Rear Window und Psycho durch eigenständige Cluster repräsentiert. Interessanterweise wird die größte Kategorie, A Space Odyssey, nicht erkannt. Dokumente dieser Kategorie verteilen sich auf mehrere Cluster, wobei dem Cluster *Kubrick* die meisten Dokumente zugewiesen sind. Dieses fasst zudem ebenfalls wie Topical k-Means alle Filme von Stanley Kubrick zusammen.

E3 Die Kategorien MySQL, PostgreSQL als auch die Ausreißer-Kategorie Antibiotika werden eindeutig identifiziert. Dagegen werden die Kategorien DB2 und Data Warehousing nicht erkannt.

E4 Im Unterschied zu E3 wird Data Warehousing im Testkorpus E4 durch ein Cluster repräsentiert. Die verschiedenen Filme werden in einem Cluster *film* zusammengefasst. Auch hier ist Lingo wie bereits die vorangegangenen Verfahren in der Lage zu generalisieren. Die einzelnen Cluster beinhalten jeweils nur Dokumente der Kategorie Film oder Datenbanken, so dass beide Themenbereiche voneinander separiert werden.

Beispielhaft für Lingo ist die Tendenz zu kleinen Clustern, sowie der große Anteil an über 80 Dokumenten, die keinem Cluster zugewiesen wurden.

Descriptive k-Means

Descriptive k-Means erzeugt für die Testkorpora E1 bis E4 zwischen 42 und 118 Cluster. Die Ergebnisse für die Experimente sind in Abbildung A.3 auf Seite 138 zusammengefasst.

E1 Descriptive k-Means trennt ebenfalls die vier Themen klar voneinander ab. Es ist auffällig, dass die Kategorie Antibiotika erst sehr spät identifiziert wird. Dieses wird damit begründet, dass durch die geringere Anzahl der Dokumente im Cluster *antibiotic resistant infections* die Bewertung für das Cluster schlechter ausfällt als für die anderen Kategorien in E1.

E2 Neben Lingo erkennt auch Descriptive k-Means viele Film-Kategorien separat. Ausschließlich ein Cluster für die Kategorie Rear Window wird nicht identifiziert. Cluster-Label wie *Bill Hartford Tom Cruise*, *Marion Norman* und *Scottie Stewart* lassen allerdings nicht ohne weiteres auf die entsprechenden Kategorien Eyes Wide Shut, Psycho und Vertigo schließen. Hier sind die Cluster-Label von Lingo stärker.

Zusätzlich zur Erkennung der einzelnen Filme führt Descriptive k-Means eine Generalisierung durch. Dies zeigt sich unter anderem in dem Cluster *Director Stanley Kubrick*, welches alle drei Filme des Regisseurs enthält.

E3 Descriptive k-Means erkennt die Ausreißer-Kategorie Antibiotika. Es werden bis auf ein Cluster dieselben Cluster wie für E1 erzeugt. Ebenso wie Lingo erkennt das Verfahren DB2 nicht als eigenständiges Cluster. Dies könnte damit zusammenhängen, dass DB2 keine Nominalphrase darstellt und somit nicht als Cluster-Label berücksichtigt wird.

Topical k-Means zeigte allerdings, dass für DB2 Cluster-Label wie *LUW Performance*, *Optim Performance Management* und *IBM mainframe* ermittelt werden können.

E4 Das Verfahren ist in der Lage, Film- und Datenbank-Kategorien voneinander zu trennen und zu generalisieren. Ein Cluster *Kubrick film* fasst erneut alle drei Filme von Stranley Kubrick unter einem anderen Cluster-Label zusammen. Es zeigt sich allerdings, dass keine Filme von Alfred Hitchcock identifiziert werden. Auch wird separat nur der Film A Space Odyssey vom Regisseur Kubrick ermittelt. Dieses ist erneut mit der geringen Anzahl an Dokumenten in den Film-Kategorien zu begründen. Descriptive k-Means bevorzugt größere Cluster.

In allen Experimenten für Descriptive k-Means zeigt sich, dass viele Cluster nahezu dieselben Dokumente abdecken. Es kommt zu einer verstärkten Überlappung der Cluster. Dieses tritt beispielsweise in E3 für die Cluster *Business Data Quality*, *data Business Data Quality*, *data quality software* und *data quality solutions* auf. Im Vergleich dazu erzeugt Topical k-Means nur Cluster, die sich stärker voneinander unterscheiden: *Data Quality Software Solutions*, *Data quality management solutions* als auch *data cleansing software* oder *data quality analysis*. Die Redundanz der Cluster-Label und somit auch in der Überdeckung der Dokumente[1] ist einerseits damit zu begründen, dass bei Descriptive k-Means nicht nur die allgemeinsten beziehungsweise spezifischsten Phrasen als Cluster-Label verwendet werden. Letzteres ist dagegen bei Topical k-Means der Fall. Desweiteren wird bei Topical k-Means jedes weitere Cluster-Label verworfen, welches keine neuen Dokumente abdeckt. Dies geschieht bei Descriptive k-Means nicht.

[1] Es wird ein monothetisches Clustering durchgeführt, so dass das Cluster-Label in jedem Dokument des Clusters enthalten ist.

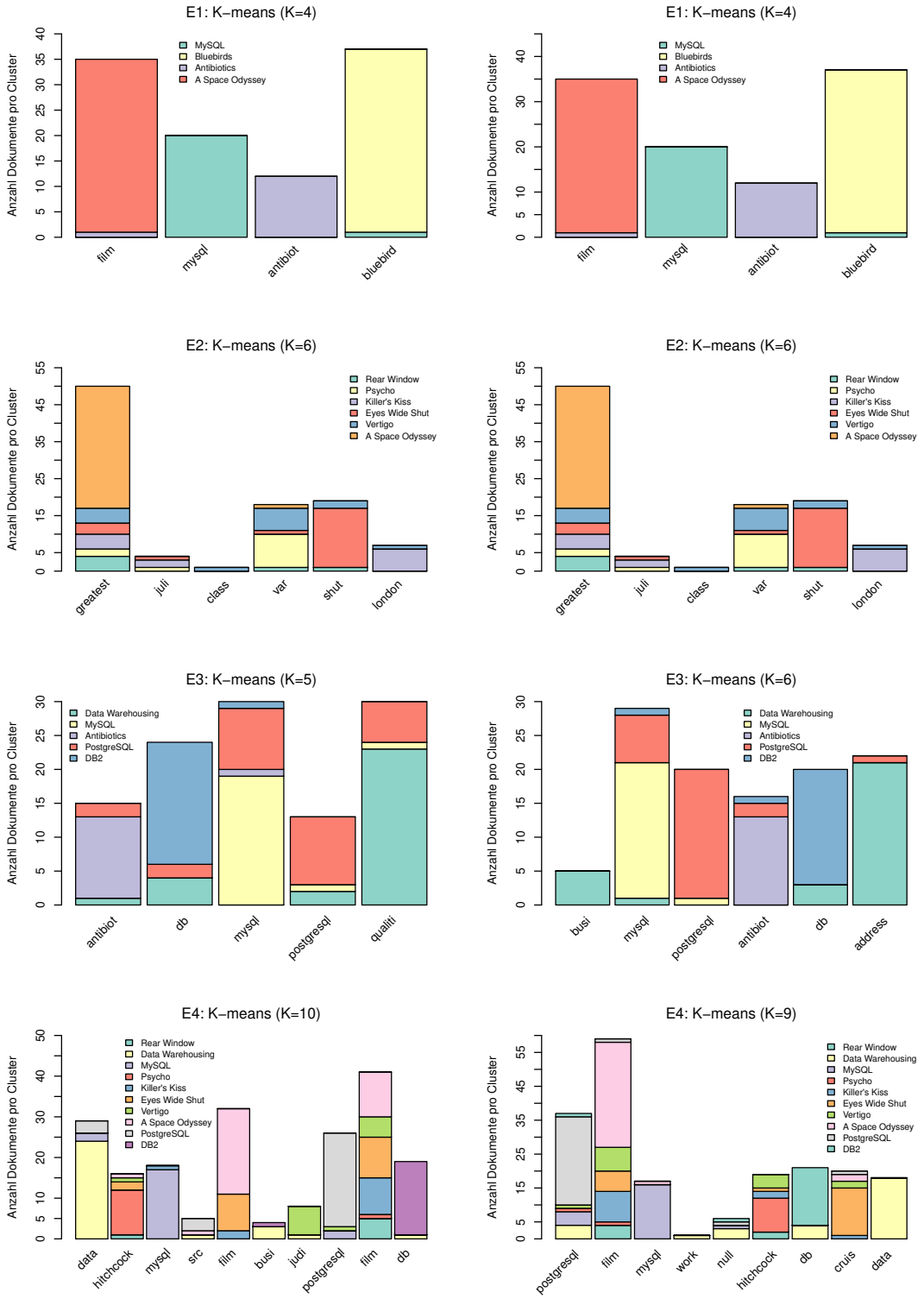

Abbildung A.1 *: Cluster-Labeling für die Korpora E1 bis E4 mittels Weighted Centroid Covering.*

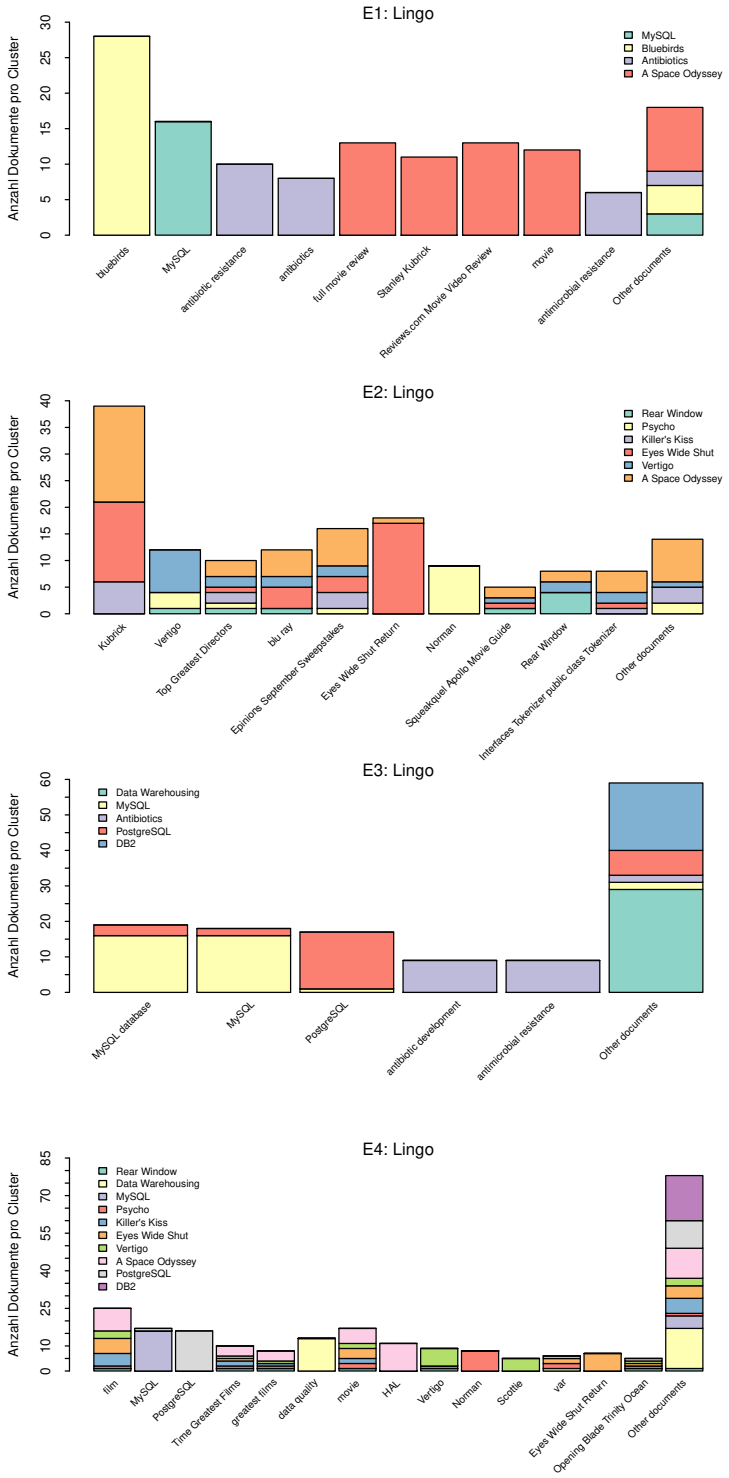

Abbildung A.2 *: Cluster-Labeling für die Korpora E1 bis E4 mittels Lingo.*

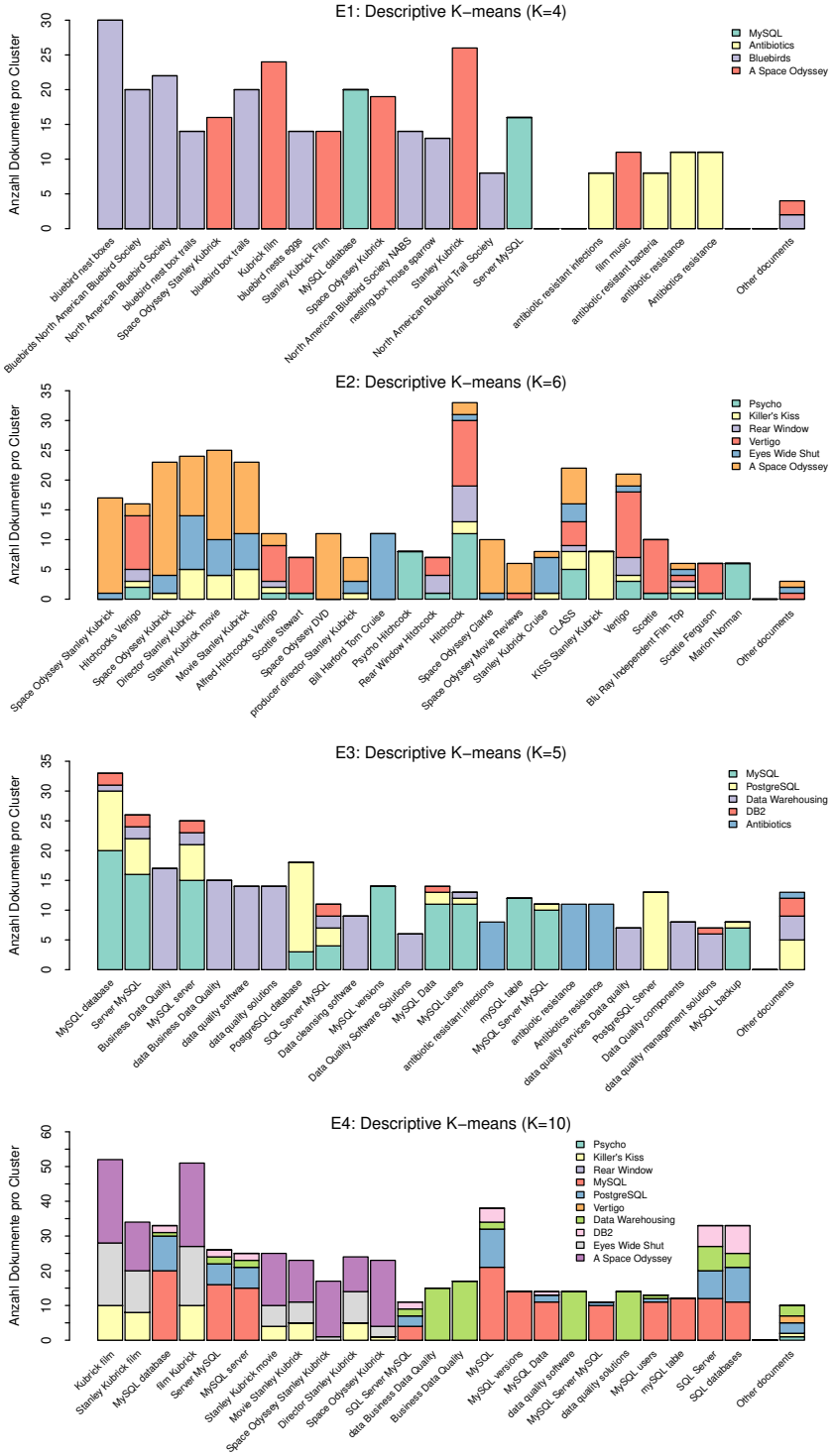

Abbildung A.3 *: Cluster-Labeling für die Korpora E1 bis E4 mittels Descriptive k-Means. Fehlende Cluster in den Abbildungen geben an, dass mehrere Cluster ausgelassen wurden. Das Cluster „Other documents“ wird immer aufgeführt.*

B Empirische Studie zu quantifizierten Label-Eigenschaften

Tabelle B.1 vervollständigt die durch die internen Validierungsmaße f_1 bis f_6 ermittelten Cluster-Label für Kategorien des Referenzkorpus.

ODP-Kategorie	Besten 5 Phrasen	Schlechteste 5 Phrasen
Antibiotics	used antibiotics (5,909)	Technology (-3,773)
	other antibiotics (5,889)	queries (-0,591)
	Antibiotics Health (5,889)	project (1,045)
	antibiotics Antibiotics (5,875)	Print (1,045)
	Antibiotics Work (5,875)	time (1,389)
Bluebirds	other bluebirds (5,971)	user (-2,045)
	bluebird nesting boxes (5,97)	Navigation (0,5)
	Bluebirding information (5,969)	password (2,045)
	nesting cavity (5,969)	fungi (2,5)
	Cavity Birds (5,967)	roots (2,864)
Psycho	Psycho psycho (5,879)	User (-0,955)
	Bates Motel Norman (5,75)	TOPIC (-0,955)
	Marion Crane Janet Leigh (5,75)	mail (-0,909)
	shower scene Hitchcock (5,75)	list (0,482)
	Martin Balsam (5,745)	release (0,5)
Rear Window	James Stewart Grace Kelly (5,8)	mail (-6,5)
	Cornell Woolrich story (5,75)	database (-6,136)
	neighbors apartment (5,727)	April (-3,773)
	detective friend (5,667)	Company (-1,682)

ODP-Kategorie	Besten 5 Phrasen	Schlechteste 5 Phrasen
	Raymond Burr (5,667)	Click (-1,5)
Vertigo	Scottie Stewart (5,818)	team (-2,682)
	film Scottie (5,8)	June (-2,045)
	San Francisco detective (5,8)	interface (-0,864)
	Novak Barbara (5,8)	Company (-0,773)
	Alfred Hitchcocks Vertigo (5,758)	database (-0,636)
A Space Odyssey	Keir Dullea (5,944)	servers (1,182)
	Space Odyssey Stanley Kubrick (5,932)	Wednesday (1,682)
	Space Odyssey Kubrick (5,932)	celebrities Links (1,909)
	space Kubricks (5,931)	following review useful See (2,273)
	Dullea William Sylvester Gary (5,923)	manual (2,409)
Eyes Wide Shut	eye Kubrick (5,919)	Operation (-5,773)
	Eyes Wide Shut Eyes (5,916)	download (-5,136)
	films Eyes (5,885)	PDF (-0,5)
	film Cruise (5,869)	system (0,833)
	Ziegler Pollack (5,833)	affiliates (0,864)
Killer's Kiss	Frank Silvera (5,901)	Organism (-2,409)
	KISS Stanley Kubrick (5,9)	sample (-0,682)
	Kiss Kubrick (5,891)	database (-0,636)
	MOVIE KILLER (5,844)	order (-0,545)
	crime films (5,782)	migration (-0,045)
Data Warehousing	data quality data (5,952)	Harry (0,591)
	data Business Data Quality (5,938)	worry (1,591)
	match data (5,92)	Review (2,49)
	AddressAbility Software (5,916)	way (2,652)
	information quality (5,909)	years (2,696)
DB2	IBM IBM (5,902)	Night (1,455)
	IBM mainframes (5,891)	AUDIO (1,773)
	IBM Information (5,891)	beds (2,227)
	International User Group (5,875)	Yahoo (2,318)

ODP-Kategorie	Besten 5 Phrasen	Schlechteste 5 Phrasen
	User Group IBM (5,857)	time (2,352)
MySQL	use MySQL (5,928)	Summer (-0,227)
	Home MySQL (5,917)	awards (1,136)
	MySQL Server MySQL (5,909)	Tom (1,409)
	MySQL versions version (5,9)	myset (1,5)
	Mysql connection (5,879)	love (1,841)
PostgreSQL	PostgreSQL version (5,939)	water (-0,045)
	PostgreSQL installations (5,917)	leaders (0,136)
	download PostgreSQL (5,917)	East (0,682)
	PostgreSQL Tools (5,889)	leader (1,136)
	Procedural Language (5,889)	Tom (1,409)

Tabelle B.1 Durch quantifizierte Validierungsmaße experimentell bestimmte Cluster-Label für alle Kategorien der Testkorpora aus Kapitel 8.1.1.