Topic-Identifikation

Formalisierung, Analyse und neue Verfahren

Benno Stein und Sven Meyer zu Eißen

Unter Topic-Identifikation versteht man die Generierung sinnvoller und ausdrucksstarker Kurzbeschreibungen bzw. Label für Gruppen von Dokumenten. Topic-Identifikation spielt eine Schlüsselrolle in allen Anwendungen, in denen unüberwacht Kategorien, also Gruppen von Dokumenten gebildet werden: Eine automatisch erstellte Dokumentkategorisierung ist wertlos, wenn es nicht gelingt, Kategoriebezeichner abzuleiten, die jede Kategorie adäquat repräsentieren und sie gegenüber den anderen Kategorien abgrenzen kann.

In der Forschung zur unüberwachten Kategorisierung hat Topic-Identifikation eher wenig Beachtung gefunden. Unser Beitrag widmet sich dieser Lücke und motiviert Anwendungen, spezifiziert formale Aspekte, stellt neue Ansätze und Algorithmen vor und evaluiert existierende Verfahren.

1 Einführung und Grundlagen

Ziel der Topic-Identifikation ist die Suche bzw. Konstruktion von Bezeichnern, mit denen Kategorien in einer Dokumentkollektion adäquat beschrieben werden. In der Literatur wird dieses Problem auch als „Topic-Findung", „Label-Identifikation", „Cluster-Labeling", „Kategorie-Labeling" oder „Label-Mining" bezeichnet [11]. Weiterhin besteht auch eine Verwandtschaft zu dem Problem der Schlüsselwortbestimmung [12, 18] und zur Generierung von Textzusammenfassungen [25].

Algorithmen zur Schlüsselwortbestimmung lassen u.a. außer Acht, dass verschiedene Kategorien voneinander abzugrenzen sind: die besten Schlüsselworte jeder einzelnen Kategorie können zusammen genommen einen schwachen diskriminatorischen Charakter haben. Methoden zur Generierung von Textzusammenfassungen versuchen ein explizites oder implizites inhaltliches Modell zu erstellen – ein Ansatz, der Aufgrund der Kürze der Label für die Topic-Identifikation nicht in Frage kommt. In dem vorliegenden Beitrag wird den Besonderheiten und Herausforderungen der Topic-Identifikation Rechnung getragen:

- Es wird eine Systematik eingeführt, die anhand formaler Kriterien das Problem der Topic-Identifikation spezifiziert und einer quantifizierbaren Betrachtungsweise zugänglich macht.
- Existierende Ansätze zur Topic-Identifikation werden eingeordnet und es wird ein effizienter Label-Algorithmus vorgestellt.
- Die Leistungsfähigkeit des neuen Verfahrens wird sowohl in Hinblick auf die formalen Kriterien als auch hinsichtlich Precision und Recall mit existierenden Verfahren verglichen.
- Die größte Beschränkung der bislang entwickelten Verfahren zur Topic-Identifikation ist, dass sie ausschließlich Informationen aus den vorliegenden Kategorien auswerten. Wir stellen in diesem Papier einen Ansatz vor, der die Deskriptoren der DMOZ-Ontologie zur Topic-Identifikation verwendet und der aufgrund dieses externen Wissens in der Lage ist, Schwächen der existierenden Ansätze zu überwinden.

Die beiden folgenden Unterabschnitte motivieren die Bedeutung der Topic-Identifikation und diskutieren das Spektrum der Lösungsansätze unter einem interessanten Gesichtspunkt, der ursprünglich aus der Cluster-Analyse stammt: polythetische versus monothetische Verfahren.

Topic-Identifikation in der kategorisierenden Suche

Kategorisierende Suche ist die Anwendung von unüberwachten Klassifikationsverfahren für Retrieval-Aufgaben, bei denen eine große Anzahl von Dokumenten zurückgegeben wird. Ein prominentes Beispiel hierfür sind Internet-Suchmaschinen wie Google oder Lycos: Ausgehend von einer Anfrage liefern sie oft eine riesige Menge D von Dokumenten. Ziel der kategorisierenden Suche ist es, D sortiert als eine Menge von – a priori unbekannten – Kategorien darzustellen, so dass sich thematisch ähnliche Dokumente in einer Gruppe befinden.

Die Operationalisierung einer kategorisierenden Suche birgt eine Reihe von Herausforderungen, wobei Effizienz und Unüberwachtheit zu den wichtigsten gehören. Effizienz ist entscheidend, weil die Kategoriebildung quasi auf Knopfdruck zu geschehen hat; das Problem der Unüberwachtheit rührt daher, dass kein auf die Suchanfrage zugeschnittenes Klassifikationsschema zur Verfügung steht. Wir befinden uns in einer Phase vor dem Semantic Web – d. h., es gibt nur wenige Dokumente, die semantisch so annotiert sind, dass sie eine automatische Kategorisierung unterstützen. Jedoch werden mit der existierenden bzw. verfeinerten Cluster-Technologie mittlerweile beachtliche Ergebnisse bei der ad-hoc-Konstruktion von Kategorien erzielt [26].[1]

Aber, selbst eine inhaltlich hervorragend organisierte Dokumentkategorisierung bleibt wertlos, wenn es nicht gelingt, adäquate Kategoriebezeichner on-the-fly abzuleiten.

Labeling: polythetisch versus monothetisch

Auf den ersten Blick scheint die Bestimmung eines adäquaten Kategoriebezeichners einfacher zu sein, als die Kategoriebildung selbst – also diejenigen Dokumente zu finden, die zu einer Kategorie gehören: Die Kategoriebildung mittels Clustering basiert auf einem bestimmten Dokumentmodell und einem entsprechenden Ähnlichkeitsmaß. Ein Cluster (eine Kategorie) C lässt sich durch ein Meta-Dokument, wie z. B. den Cluster-Zentroid $\mathbf{c}$, repräsentieren, der mittels Durchschnittsbildung aus allen Modellen der Dokumente in C berechnet wird. Cluster-Labeling kann als eine Funktion $\tau(C)$ verstanden werden, die auf eine Menge von

[1] Mit seinen leistungsfähigen Konzepten zur Annotation und Anfrageformulierung auf Basis von RDF, RDFS und OWL könnte das Semantic Web mittel- bis langfristig die augenblicklich vorherrschenden Konzepte zur unüberwachten Klassifikation überflüssig machen. Dieser „semantisch reiche" Weg erfordert Korpora mit annotierten Dokumenten und auch den Zugriff auf so genannte höhere Ontologien [5, 7].

Abbildung 1: Kategoriebaum, der von der Suchmaschine Vivísimo für die Anfrage „Lassie“ auf der Basis von 150 Dokumenten konstruiert wurde. Die Kategoriebezeichner sind das Ergebnis eines polythetischen Labeling-Algorithmus.

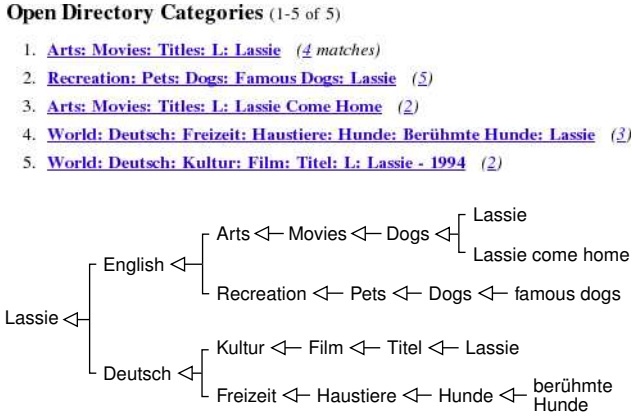

Abbildung 2: Suchergebnisse und Kategoriebezeichner der DMOZ -Hierarchie für die Anfrage „Lassie“. Die hohe Qualität der Kategoriebezeichner spiegelt sich sowohl in der sinnvollen Spezialisierungssemantik als auch in der Orthogonalität der Deskriptoren wider.

Deskriptortermen abbildet, üblicherweise eine Teilmenge der Indexterme aus $\mathbf{c}$. Folgt man der Cluster-Terminologie, so stellt die Bestimmung von $\tau(C)$ einen *polythetischen* Labeling-Ansatz dar: Der Labeling-Prozess beruht auf der gleichzeitigen Analyse mehrerer Merkmale, den Indextermen. Diese Beobachtung wurde auch von Sanderson gemacht [23].

Zahlreiche Analysen zeigen, dass Cluster-Algorithmen in der Lage sind, Kategorien zu identifizieren, die auch aus dem Blickwinkel menschlicher Editoren sinnvoll erscheinen [31, 14]. Während der polythetische Ansatz für die Cluster-Bildung, also für die Ähnlichkeitsanalyse zwischen Dokumenten, erfolgreich ist, so ist er im Zusammenhang mit Cluster-Labeling problematisch. Cluster-Label spielen die Rolle von Konzeptdeskriptoren in einer Konzepthierarchie: Idealerweise sollten die Label eines hierarchischen Clusterings eine Konzeptualisierung der Dokumente der unterliegenden Kollektion darstellen, d. h., sie sollten von einem ontologischen Standpunkt aus sinnvoll sein. Eine Konzepthierarchie lässt sich als das Ergebnis eines *monothetischen* Cluster-Algorithmus begreifen: In jedem Clustering-Schritt wird aus der verbliebenen Menge von Dokumenten D' das generischste Konzept t als Merkmal mit den Ausprägungen $t_1,\ldots,t_k$ aufgefasst, bzgl. derer die Dokumente in D' unterschieden werden.

Die Abbildungen 1 und 2 zeigen den Unterschied zwischen dem polythetischen Labeling-Paradigma auf der einen Seite und der Auswahl von Konzeptdeskriptoren auf der anderen anhand der Suchergebnisse von zwei prominenten Dokument-Retrieval-Anwendungen: Vivísimo und das DMOZ Open-Directory-Project; Suchanfrage ist „Lassie“.

Die Ausgabe von Vivísimo zeigt, dass sich mittels des polythetischen Ansatzes sinnvolle Ergebnisse erzielen lassen, es zeigt aber auch die Grenzen des Ansatzes: Klappt man den Kategoriebaum auf, so sieht man unverständliche und sich wiederholende Kategoriebezeichner sowie auch Unterkategorien, die keine Spezialisierungen ihrer Oberkategorie darstellen. Im Gegensatz zu Vivísimo werden Kategorien bei DMOZ von menschlichen Editoren gepflegt, und die Anfrageergebnisse spiegeln das ontologische Weltverständnis der Editoren wider.

Die *automatische Ableitung* eines authentischen Kategorisierungsschemas für eine gegebene Dokumentkollektion D ist sehr schwierig. In der Forschung wurden in diesem Zusammenhang die Kombination komplexer Techniken wie lokale Kontextanalyse (LCA), Subsumptionsanalyse, Kollokationsanalyse oder Latent-Semantic-Indexing (LSI) untersucht [24]. Die Ergebnisse sind meist unbefriedigend, die Laufzeitkomplexität für viele Anwendungen inakzeptabel. So sind heuristische und ad-hoc-Ansätze zur Topic-Identifikation, die im Kern auf eine möglichst geschickte Indextermauswahl abzielen, gängige Praxis bei den kategorisierenden Suchmaschinen.

2 Systematik zur Formalisierung von Label-Eigenschaften

Sei D eine Menge von Dokumenten. Eine Kategorisierung $\mathcal{C} \subseteq \{C \mid C \subseteq D\}$ von D ist eine Aufteilung von D in Mengen $C_1,\ldots,C_m$ mit $C_1 \cup \ldots \cup C_m = D$. $\mathcal{C}$ heißt exklusive Kategorisierung, falls $C_i \cap C_{j\neq i} = \emptyset$ für $C_i, C_j \in \mathcal{C}$ gilt; andernfalls heißt $\mathcal{C}$ nicht-exklusive Kategorisierung. Weiterhin wird davon ausgegangen, dass kein Klassifizierungsschema vorgegeben ist – also die Standardsituation für kategorisierende Suchmaschinen vorliegt, die $\mathcal{C}$ mit Hilfe eines Cluster-Algorithmus bestimmen.

Die Objekte in D sind Abstraktionen der eigentlichen Dokumente und wurden auf Basis eines Dokumentmodells konstruiert [2]. Für die folgenden Überlegungen fassen wir eine Dokumentdatenstruktur $d \in D$ als eine geordnete Menge von Indexstermen $W_d = \{w_{d_1},\ldots,w_{d_n}\}$ auf, denen verschiedene Funktionen zugeordnet sind, die von W_d nach $\mathbf{R}^+$ abbilden. Beispiele für solche Funktionen sind die Termfrequenz $tf_d(w)$ oder die inverse Dokumentfrequenz $idf(w)$.

Sei $W = \bigcup_{d\in D} W_d$ die gesamte, $\mathcal{C}$ unterliegende Wortmenge, auch „Dictionary“ genannt. Dann bedeutet Topic-Identifikation die Bestimmung einer Funktion τ, die jeder Kategorie $C \in \mathcal{C}$ eine Menge $T_C \subset W$ zuordnet, in Zeichen: $\tau : \mathcal{C} \rightarrow W$ mit $\tau(C) = T_C$. τ wird als Labeling bezeichnet.

Für eine gegebene Kategorisierung $\mathcal{C}$ lassen sich verschiedene Eigenschaften und Constraints für ein Labeling τ formulieren. Die nachfolgende Liste ist bzgl. des Schwierigkeitsgrades in Hinblick auf eine automatische Umsetzung aufsteigend sortiert.

1. *Eindeutigkeit.* Keine zwei Label von zwei verschiedenen Kategorien verwenden einen Term gemeinsam:
$$\forall_{\substack{C_i,C_j\in\mathcal{C}\\ C_i\neq C_j}} : \tau(C_i) \cap \tau(C_j) = \emptyset$$

2. *Überdeckend.* Der Label einer Kategorie C enthält von jedem Dokument in C mindestens einen Term:
$$\forall_{C\in\mathcal{C}}\ \forall_{d\in C} : \tau(C) \cap W_d \neq \emptyset$$

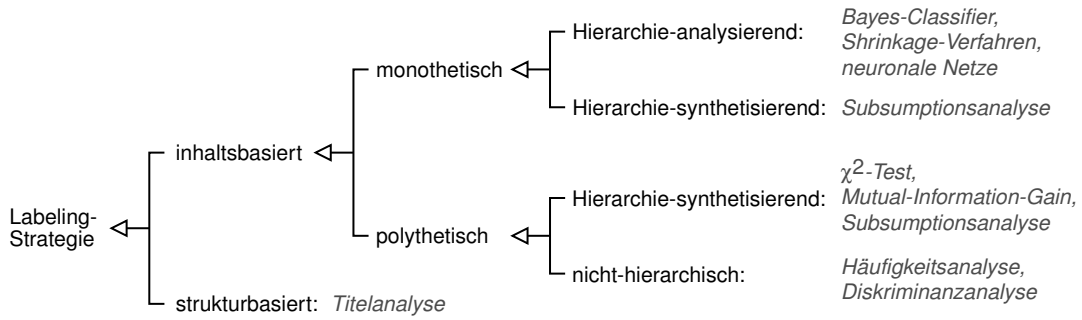

Abbildung 3: Klassifikationsschema und Einordnung von Labeling-Strategien für Dokumentmengen; Beispiele für die jeweilige Strategie sind an den Blättern angegeben

3. *Ausdrucksstärke bzw. Plausibilität.* Die Terme in dem Label einer Kategorie C gehören bezüglich der Dokumente in C zu den häufigsten:

$$\forall_{C\in\mathcal{C}}\ \exists_{w'\in\tau(C)}\ \forall_{d\in C}\ \forall_{\substack{w\in W_d\\ w\notin\tau(C)}}\ :\ tf_d(w)\leq tf_d(w'),$$

mit $tf_d(w)$ als Termfrequenz von Term w in Dokument d.

4. *Trennstärke bzw. Diskriminanz.* In dem Label einer Kategorie C gibt es einen Terms w', dessen relative Häufigkeit bezogen auf die Dokumente in C signifikant größer ist, als bei allen anderen Kategorien:

$$\forall_{\substack{C_i,C_j\in\mathcal{C}\\ C_i\neq C_j}}\ \exists_{w'\in\tau(C_j)}:\ \frac{1}{|C_i|}tf_{C_i}(w')\ll\frac{1}{|C_j|}tf_{C_j}(w'),$$

mit $W_C=\bigcup_{d\in C}W_d$ als Termmenge von Kategorie C und $tf_C(w)$ die Termfrequenz von w in Kategorie C.

5. *Zusammenhängend.* Um Idiome, Namen, zusammengesetzte Worte, etc. abzubilden, enthält der Label einer Kategorie C Terme, die in den Dokumenten von C unmittelbar aufeinander folgen:

$$\forall_{C\in\mathcal{C}}\ \exists_{\substack{w',w''\in\tau(C)\\ w'\neq w''}}\ \forall_{d\in C}\ \exists_{w_i,w_{i+1}\in W_d}:$$
$$w_i=w'\wedge w_{i+1}=w''$$

6. *Hierarchische Konsistenz.* Stellt eine Kategorie C_j eine Spezialisierung einer Kategorie C_i dar, in Zeichen $C_i\succ C_j$, so spiegelt sich diese Spezialisierung auch in den Termen der zugehörigen Label wider:

$$\forall_{\substack{C_i,C_j\in\mathcal{C}\\ C_i\neq C_j}}:\ C_i\succ C_j\Rightarrow P(w_i|w_j)=1\wedge P(w_j|w_i)<1,$$

mit $w_i\in\tau(C_i)$, $w_j\in\tau(C_j)$ sowie P als bedingte Wahrscheinlichkeit für das Auftreten der Worte w_i,w_j in den Dokumenten der Cluster C_i,C_j.

7. *Irredundanz.* Je zwei Terme eines Labels stellen keine Synonyme dar:

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

Bemerkungen. (i) Offensichtlich handelt es sich bei diesen Eigenschaften um ideale Constraints, die in der Praxis nur approximiert werden können. Das bezieht sich insbesondere auf die „für alle"-Forderungen hinsichtlich der Dokumente in den Eigenschaften 2, 3 und 5, die in realen Dokumentkollektionen als „möglichst viele" zu interpretieren sind oder für die eine Zahl von Ausnahmen zugelassen ist. (ii) Der in der Suchmaschine Vivísimo realisierte Cluster-Ansatz beruht auf der Datenstruktur der Suffix-Bäume [30]. Diese Datenstruktur wird zur Ableitung von Labeln verwendet, welche die Zusammenhangseigenschaft 5 erfüllen. (iii) Ohne externes Wissen ist es praktisch unmöglich, für ein Labeling τ die Eigenschaft 7 der Irredundanz herzustellen.

3 Ansätze zur Topic-Identifikation

Dieser Abschnitt präsentiert ein Klassifikationsschema existierender Ansätze von einem operationalen Standpunkt. Abbildung 3 zeigt dieses Schema; es enthält sowohl Strategien, die speziell für automatisches Labeling bzw. Topic-Identifikation entwickelt wurden, als auch allgemeine Verfahren, die sich unmittelbar hierfür einsetzen lassen.

Auf der obersten Ebene wird zwischen inhaltsbasierten und strukturbasierten Labeling-Strategien unterschieden. Inhaltsbasierte Strategien analysieren die Terme eines Dokumentes, d. h., sie verwenden ein gängiges Dokumentmodell des Information Retrieval. Die strukturbasierten Strategien analysieren ein Dokument hinsichtlich besonderer Textelemente wie Titel, Schlüsselwortliste oder typographischer Akzente.

Auf der zweiten Ebene lassen sich inhaltsbasierten Labeling-Strategien weiter in monothetische und polythetische Ansätze aufteilen; ihre grundsätzlichen Unterschiede zusammen mit ihren Vor- und Nachteilen wurden in Abschnitt 1 erörtert. Polythetisches Labeling wird in den meisten Systemen zur automatischen Dokumentkategorisierung eingesetzt und ist die vorherrschende Strategie zur Topic-Identifikation [30, 31, 11, 26]. Falls keine hierarchische Ordnung (Hyponymie) unterstellt werden kann, stützen sich polythetische Label-Konstruktionsverfahren auf Term- und Dokumentfrequenzen, auf Quantisierungsfehler und auf die Kombinationen solcher Merkmale. Unter der Hierarchieannahme kommen χ^2-Tests, Verfahren des Mutual-Information-Gain oder Subsumptionsanalysen zum Einsatz [21, 23].

Auf der dritten Ebene können monothetische Ansätze weiter hinsichtlich der ihnen zu Grunde liegenden Informationsquelle aufgeteilt werden: Hierarchie-analysierende Verfahren verwenden externes Wissen wie z. B. eine existierende Ontologie oder eine Taxonomie, um Label-Information abzuleiten. Demgegenüber verwenden Hierarchie-synthetisierende Verfahren das (interne) Wissen der vorliegenden Dokumentkollektion; d. h., sie versuchen, eine Konzepthierarchie mittels eines eng verzahnten Kategorisierungs- und Labeling-Prozesses abzuleiten [23]. Zu den polythetischen Verfahren gehören unter anderem häufigkeitsanalysierende Methoden, die auf der Bestimmung von „frequent itemsets" beruhen [4, 3, 13].

Topic-Identifikation auf Basis von externem Klassifizierungswissen stellt einen neuen Ansatz dar und wird in Abschnitt 5 behandelt.

Heuristische Topic-Identifikation mit WCC

Das im Folgenden vorgestellte Verfahren „Weighted Centroid Covering" (WCC) ist ein universell einsetzbares Verfahren zur Topic-Identifikation. Neben hoher Performanz war der Leitgedanke beim Entwurf dieses Algorithmus eine Maximierung der in Abschnitt 2

vorgestellten Eigenschaften 1-4. WCC basiert auf dem polythetischen, nicht-hierarchischen Paradigma.

Wie zuvor bezeichne D eine Menge von Dokumenten bzgl. einer Wortmenge W, $w \in W$ ein Wort, $\mathcal{C} = \{C_1, \ldots C_m\}$ ein Clustering bzw. eine Kategorisierung von D und $tf_C(w)$ die Termfrequenz von Wort w in Cluster (Kategorie) $C \in \mathcal{C}$. Weiterhin sei $termRank : W \times \mathbf{N} \to \mathcal{C}$ mit $termRank(w, i) \mapsto C$ eine Funktion, die einen Cluster C liefert, in dem Wort w am i-thäufigsten ist. So bezeichnen z. B. $termRank(w, 1)$ und $termRank(w, m)$ diejenigen Cluster, in denen w am häufigsten bzw. am seltensten vorkommt.

Der Algorithmus besteht aus zwei Teilen. Zunächst werden für jedes Wort w die k Cluster bestimmt, in denen w am häufigsten vorkommt. Entsprechende 3-Tupel aus Wort, Termfrequenz und Cluster (insgesamt $k \cdot |W|$ Tupel) werden in einer Liste $\mathcal{T}$, nach Termfrequenz absteigend sortiert, eingefügt. Im zweiten Teil des Algorithmus werden jedem Cluster l verschiedene Terme zugeordnet, wobei jeder Cluster in jeder Runde genau ein Wort erhält (Round-Robin). Bedingt durch die Top-Down-Verarbeitung der Tupel in $\mathcal{T}$ werden die häufigsten Worte der gewichteten Cluster-Zentroiden überdeckt.

Algorithmus: WCC, Weighted Centroid Covering

Input: $\mathcal{C}$ Clustering $\mathcal{C}$.
l Anzahl der Terme eines Labels.
k Anzahl der Vorkommen desselben Terms in verschiedenen Labeln.

Output: τ Labeling.

WCC$(\mathcal{C}, l, k)$

1. $\mathcal{T} = \emptyset$;
   FOREACH C IN $\mathcal{C}$ DO $\tau(C) = \emptyset$;
2. FOREACH w IN W DO
     FOR $i = 1$ TO k DO
       $C = termRank(w, i)$;
       $f = tf_C(w)$;
       insert $\langle w,\ f,\ C \rangle$ into $\mathcal{T}$;
     ENDFOR
   ENDDO
3. Sort $\mathcal{T}$ with descending term frequencies;
4. FOR $round = 1$ TO l DO
     $j = 1$;
     WHILE not all clusters got a term DO
       let $t_j = \langle w,\ f,\ C \rangle$ be jth tuple of $\mathcal{T}$;
       IF C got no new term this round THEN
         $\tau(C) = \tau(C) \cup \{w\}$;
         delete t_j from $\mathcal{T}$;
       ENDIF
       $j = j + 1$;
     ENDWHILE
   ENDFOR
5. RETURN τ;

Bemerkungen. Ein von WCC generiertes Labeling τ erfüllt die Eigenschaft 1 für $k = 1$. Durch die Sortierung von $\mathcal{T}$ gemäß Termfrequenz wird die Erfüllung von Eigenschaft 2 angestrebt. Die Sortierung in Kombination mit der Cluster-weisen Round-Robin-Strategie zielt auf die Optimierung der Eigenschaften 3 und 4 ab. Durch die Verwendung der absoluten Termfrequenz, tf_C, ist implizit verankert, dass größere Kategorien bevorzugt mit Termen überdeckt werden. Die Bestimmung der Tupel für $\mathcal{T}$ einschließlich ihrer Sortierung ist in $O(k \cdot |W| \cdot \log(k \cdot |W|))$; die Label-Zuweisung ist in $O(l \cdot k \cdot |W|)$. Weil k und l durch kleine Konstanten beschränkt sind, ist die Gesamtkomplexität von WCC in $O(|W| \cdot \log(|W|))$.

4 Analyse von Algorithmen zur Topic-Identifikation

Die empirische Analyse von Algorithmen zur Topic-Identifikation ist schwierig, da keine entsprechenden Testkollektionen existieren. Wichtigster Grund dafür ist der Erstellungsaufwand. Theoretisch wäre Kategorisierungen automatisch durch Cluster-Analysen generierbar, jedoch beeinflusst die Wahl eines Algorithmus wie auch seine Parametrisierung die Qualität der erzeugten Kategorien erheblich [27].

Bislang wurde die Qualität von Algorithmen zur Topic-Identifikation durch Benutzerstudien gemessen, bei denen Menschen für eine Menge von generierten Labelings einer Kategorisierung eine Rangfolge angeben [21, 17]. Wenngleich diese Studien Hinweise auf die relative Leistungsfähigkeit der Algorithmen geben, so sind sie weder reproduzierbar noch im Sinne von Precision und Recall interpretierbar. Als Ausweg schlagen wir hier Statistiken vor, die auf Grundlage unserer Systematik ableitbar sind und mit denen sich der Erfüllungsgrad gewünschter Label-Eigenschaften aus Abschnitt 2 quantifizieren lässt:

1. *Eindeutigkeit.*

$$f_1(\tau) = 1 - \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{|\ \tau(C_i) \cap \tau(C_j)\ |}{|\ \tau(C_i) \cup \tau(C_j)\ |}$$

   $f_1(\tau)$ nimmt den Wert 1 an, falls alle Terme aller Label unterschiedlich sind; hierbei ist k die Anzahl der Kategorien in $\mathcal{C}$. Je näher der Wert bei 0 liegt, desto mehr überschneiden sich die Terme unterschiedlicher Label.

2. *Überdeckend.*

$$f_2(\tau) = \frac{1}{k} \sum_{C \in \mathcal{C}} \frac{1}{|C|} \sum_{d \in C} \frac{|\ \tau(C) \cap W_d\ |}{|\ \tau(C)\ |}$$

   Je näher $f_2(\tau)$ bei 1 liegt, desto besser überdecken die Label-Terme die Dokumente der zugehörigen Kategorie. Ein Wert nahe 0 zeigt an, dass die Kategorie-Label in keinem oder nur wenigen Dokumenten einer Kategorie vorkommen.

3. *Plausibilität.*

$$f_3(\tau) = 1 - \frac{1}{k} \sum_{C \in \mathcal{C}} \operatorname*{argmin}_{w' \in \tau(C)} \frac{1}{|C|} \sum_{d \in C} \frac{1}{|W_d|} \sum_{\substack{w \in W_d \\ w \notin \tau(C)}} \frac{tf_d(w)}{tf_d(w')}$$

   $f_3(\tau)$ strebt gegen das Maximum 1, falls für jeden Cluster ein ausdrucksstarker Term existiert. Man beachte, dass f_3 für schlecht gewählte Label negativ werden kann.

4. *Trennstärke.*

$$f_4(\tau) = 1 - \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \operatorname*{argmin}_{w' \in \tau(C_j)} \frac{|C_j|}{|C_i|} \frac{tf_{C_i}(w')}{tf_{C_j}(w')}$$

   Je näher $f_4(\tau)$ bei 1 liegt, desto stärker ist die Diskriminanz; mit kleiner werdenden Werten sinkt auch die Diskriminanzkraft von τ. Auch hier kann der Funktionswert negativ werden.

Zur Ermittlung der Werte dieser Funktionen wurde eine Kollektion D mit 150 wissenschaftlichen Artikeln der digitalen Bibliothek CiteSeer[2] erstellt und mehrfach eine zufällig konstruierte

[2] http://citeseer.ist.psu.edu

Menge $D' \subset D$, $|D'| = 80$, durch k-Means geclustert und die Cluster mit verschiedenen Verfahren gelabelt. Tabelle 1 stellt die gemittelten Werte der Statistiken für das neu entwickelte Verfahren WCC aus Abschnitt 3, für das Verfahren von Popescul et al. [21] und für das Schlüsselwort-Extraktionsverfahren RSP [29] gegenüber.

	WCC	Popescul	RSP
f_1 *(Eindeutigkeit)*	0.96	1.00	1.00
f_2 *(Überdeckend)*	0.78	0.60	0.51
f_3 *(Plausibilität)*	0.74	0.47	0.61
f_4 *(Trennstärke)*	0.98	1.00	0.89
Durchschnitt	0.87	0.77	0.75

Tabelle 1: Erfüllung der Eigenschaften 1-4 für die Verfahren WCC, Popescul und RSP. Pro Cluster (Kategorie) wurden vier Label-Terme generiert.

Nimmt man die von Menschen vergebenen Schlüsselworte als Maßstab, so läßt sich mit Precision und Recall die Qualität einer Überdeckung dieser Schlüsselworte quantifizieren. In unserer Analyse wurden hierfür die Vereinigungsmengen der tatsächlichen Schlüsselworte aller Dokumente pro Kategorie ermittelt und mit den generierten Labeln der Verfahren WCC, Popescul und RSP verglichen. Die Abbildungen 4 und 5 zeigen Precision- und Recall-Werte der Verfahren für ein repräsentatives Clustering abhängig von der Anzahl der erzeugten Label-Terme. Auffallend ist die hohe Precision von Popesculs Verfahren und von WCC. Hier können Schlüsselwort-Extraktionsverfahren nicht mithalten, da sie nicht auf eine Diskriminierung zwischen Dokumenten abzielen.

5 Externe Topic-Identifikation

Im Idealfall sollten die Label eines Clusterings eine Konzeptualisierung der Dokumente der unterliegenden Kollektion darstellen. Mit den vorgestellten Ansätzen zur Topic-Identifikation ist das nur in Ausnahmefällen erreichbar. Technisch gesehen ist zwar aus einer Dokumentmenge D mit jedem hierarchischen Cluster-Algorithmus ein Kategoriebaum konstruierbar [10], aber ein hierauf basierendes Labeling τ wird von einer semantisch sinnvollen Taxonomie weit entfernt sein. Insbesondere lassen sich die Eigenschaften 6 (hierarchische Konsistenz) und 7 (Irredundanz) der Systematik aus Abschnitt 2 nicht mit einer hierarchischen Cluster-Analyse erfüllen [11].

Die genannten Schwächen lassen sich durch die Nutzung externen Klassifikationswissens aus einer Referenzkategorisierung überwinden. Diese Idee wird hier als *externe* Topic-Identifikation bezeichnet: Sei $\mathcal{C} = C_1, \ldots, C_m$ die von einem Cluster-Algorithmus erzeugte Kategorisierung einer Menge D von Suchergebnissen. Weiterhin sei $\mathcal{O} = O_1, \ldots, O_l$ eine Referenzkategorisierung einer Menge von Dokumenten $D_\mathcal{O}$ und $\tau_\mathcal{O}$ ein Labeling von $\mathcal{O}$, das eine ontologische Sicht auf $D_\mathcal{O}$ realisiert. Externe Topic-Identifikation lässt sich dann wie folgt operationalisieren:

1. Jede Kategorie $C \in \mathcal{C}$ wird ihrer ähnlichsten Entsprechung $O \in \mathcal{O}$ zugeordnet. Falls die Zuordnung eindeutig ist, wird $\tau_\mathcal{O}(O)$ als Kategoriebezeichnung gewählt.
2. Für Kategorien, die ohne Entsprechung in $\mathcal{O}$ sind, werden mit einem polythetischen, nicht-hierarchischen Labeling-Algorithmus wie WCC die Kategoriebezeichner konstruiert.

Externe Topic-Identifikation nutzt Kategoriebeschreibungen und -zusammenhänge aus, die von Menschen u. a. in Hinblick auf die Eigenschaften 1-7 sorgfältig konstruiert wurden. Die Herausforderung bei der externen Topic-Identifikation ist somit nicht die Optimierung dieser Eigenschaften, sondern die Konstruktion eines hierarchischen Klassifizierers, der eine sinnvolle Zuordnung der Kategorien aus $\mathcal{C}$ zu den Knoten einer Ontologie $\mathcal{O}$ realisiert. Im einfachsten Fall kann dies durch eine Nächste-Nachbarn-Suche zwischen den Zentroidvektoren der Cluster $C \in \mathcal{C}$ und denen der Kategorien $O \in \mathcal{O}$ geschehen. Es bietet sich jedoch an, den Mengencharakter der zu klassifizierenden Kategorien auszunutzen und Multipfad- und Subsumptionstechniken einzusetzen. Beachtete Arbeiten zur hierarchischen Textklassifikation finden sich in [16, 19, 20, 9, 28, 22, 6]. Die Ausnutzung von Klassifizierungsergebnissen zur Konstruktion von Labeln ist neu und wurde bislang nicht in Arbeiten zur Textklassifikation diskutiert; gleichwohl ist die Aufgabenstellung mit der des Ontologie-Matchings vergleichbar und eventuell mit den dort angewandten Verfahren lösbar [8].

Auf Basis der DMOZ-Ontologie haben wir umfangreiche Analysen zur Machbarkeit des skizzierten Ansatzes durchgeführt [1]. Eine vollständige Diskussion der untersuchten Fragestellungen, angewandten Technologien und durchgeführten Analysen ist hier nicht möglich, wichtige Ergebnisse sind nachfolgend dargestellt.

Die DMOZ-Ontologie ist eine thematisch hierarchisch organisierte Dokumentkollektion, die von Menschen gepflegt wird und die öffentlich zugänglich ist.[3] Die in unseren Analysen verwendete Version vom Juni 2006 hat knapp 710 000 Kategorien, in die über 4,7 Millionen Dokumente eingeordnet sind. Die Meta-Beschreibung der Ontologie liegt als RDF-Datei vor und hat einen Umfang von 2GB. Die Ontologie hat 15 Toplevel-Kategorien und eine Tiefe von 14, wobei der Großteil der Kategorien sich in den Ebenen 4 - 7 befindet; in diesem Bereich liegen auch über 90% der Dokumente. Abbildung 6 zeigt die Charakteristik dieser Verteilung am Beispiel der Toplevel-Kategorie „Science“.

Unsere Analysen fallen in zwei Gruppen:

[3] http://www.dmoz.org

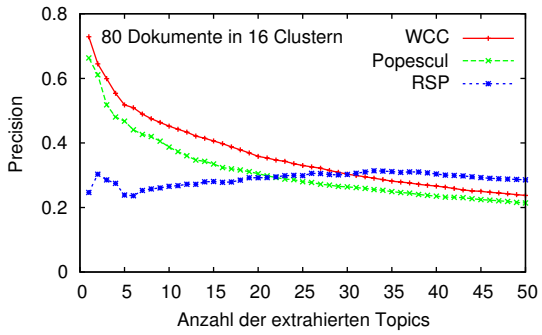

Abbildung 4: Precision der untersuchten Verfahren abhängig von der Anzahl der erzeugten Label-Terme.

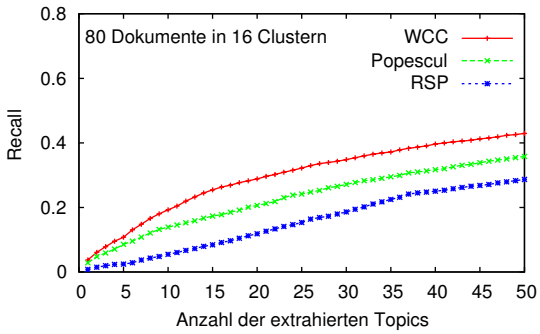

Abbildung 5: Recall der untersuchten Verfahren abhängig von der Anzahl der erzeugten Label-Terme.

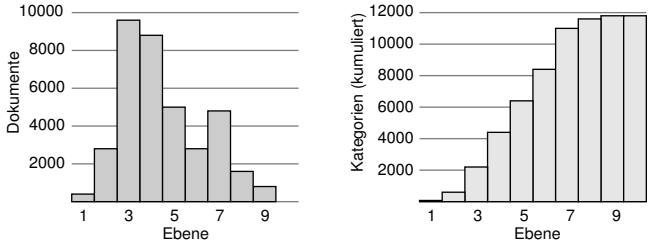

Abbildung 6: Verteilung der Dokumente (links) und der Unterkategorien (rechts, kumuliert) von „Science“. Die Charakteristik der Verteilungen ist typisch für Toplevel-Kategorien in DMOZ.

1. In der *Domänen-spezifischen* Kategorisierung wird sich auf eine einzelne Toplevel-Kategorie beschränkt. Dieses Szenario modelliert Situationen, in denen a-Priori-Wissen über die zu labelnden Cluster vorliegt.
2. In der *freien* Kategorisierung wird ein Großteil der Toplevel-Kategorien berücksichtigt, jedoch nur bis zur Tiefe 2. Dieses Szenario ist aus Labeling-Sicht besonders interessant, weil Cluster-Algorithmen bzgl. der Label-Qualität im Bereich des Allgemeinwissens ihre größte Schwäche besitzen.

Die Tabellen 2 und 3 zeigen Klassifikationsergebnisse für die Domänen-spezifische und die freie Kategorisierung. Die Zahlen zur korrekten Klassifizierung beziehen sich auf einzelne Dokumente eines Cluster $C \in \mathcal{C}$; d. h. kombiniert man die Klassifikationsergebnisse der Dokumente in C, so lassen sich – abhängig von $|C|$ – noch bessere Ergebnisse erzielen. Darüber hinaus kann die Klassifikationsqualität auf Kosten des Spezialierungsgrades weiter verbessert werden.

Kollektion	Kategorien Tiefe	Anzahl	Dokumente	Merkmale	korrekt
Science3	3	14	1500 (propagiert)	100	81%
				500	84%
Science4	4	24	1500 (propagiert)	100	79%
				500	83%
Science5	5	39	1500	100	68%
				500	77%

Tabelle 2: Domänen-spezifische Kategorisierung: Klassifikationsergebnisse bei der Zuordnung von Dokumenten aus der „Science“-Domäne in die DMOZ-Ontologie. Die Testkollektionen decken zwischen 17 (Sciene3) und 44 (Science5) Kategorien ab.

Technischer Hintergrund. Unsere Analysen beschränken sich auf englischsprachige Dokumente. Die eingesetzten Klassifikationstechniken umfassen Support Vector Machines, Naive Bayes mit den Varianten Shrinkage, Complement und Multinomial sowie die Konstruktion von Ensembles mit dem Adaboost- und einem Subsumptions-Verfahren. Für die Experimente zur freien Kategorisierung wurde darüber hinaus die Multipfad-Klassifikation angewandt. Bei den Merkmalen handelt es sich ausschließlich um Terme, die in den Dokumenten vorkommen. Als Auswahlkriterium dienten Information Gain, χ-Quadrat und Odds Ratio.

Die in den Tabellen gezeigten Klassifikationsergebnisse wurden mit einem Complement-Naive-Bayes-Verfahren erzielt; die Merkmalsselektion geschah auf Basis von Odds-Ratio, wobei jeder Experimentvariante 10 Durchläufe mit eigener Trainings- und Testmenge zugrunde lagen. Die Ergebnisse demonstrieren das Potenzial dieses Ansatzes, der insbesondere im Bereich der freien Kategorisierung als Ergänzung zu den Algorithmen der internen Topic-Identifikation dienen kann. Jedoch setzen Aufwand und Umfang der notwendigen Klassifikationstechnologie dem breiten Einsatz der externen Topic-Identifikation zur Zeit enge Grenzen.

Kollektion	Kategorien Tiefe 1	Tiefe 2	Dokumente	Merkmale	korrekt
TL3	3	15	2250	100	72%
				200	83%
				500	88%
TL10	10	50	7500	100	33%
				200	48%
				500	63%

Tabelle 3: Freie Kategorisierung: Klassifikationsergebnisse bei der Zuordnung von beliebigen DMOZ-Dokumenten in die DMOZ-Ontologie. Die Kollektion TL3 umfasst 3, die Kollektion TL10 sogar 10 Toplevel-Domänen.

6 Zusammenfassung und Ausblick

Topic-Identifikation ist die Generierung einer Liste von Termen (= Label), um eine Menge von Dokumenten möglichst gut zu charakterisieren. Bei der kategorisierenden Suche werden mit jeder Anfrage eine Menge $\mathcal{C}$ solcher Dokumentmengen zurück geliefert, und die Konstruktion von aufeinander abgestimmten Kategoriebezeichnern (= Labeling τ) ist eine zentrale Aufgabe. Es stellt sich die Frage, wann ein Labeling τ gut ist bzw. welche Eigenschaften es idealerweise erfüllen sollte.

Um diese Frage quantifizierbar zu beantworten und um Algorithmen gemäß ihrer Paradigmen einordnen zu können, wurde eine Systematik entwickelt. Weiterhin wurde ein effizienter heuristischer Algorithmus zur Topic-Identifikation vorgestellt, der vier Eigenschaften dieser Systematik optimiert und der existierenden Verfahren überlegen ist.

Zukünftige Ansätze zur Topic-Identifikation müssen externes Wissen heranziehen, um ihre wesentlichen Schwächen, die bei der hierarchischen Konsistenz und bei der Irredundanz liegen, beheben zu können. Wir haben argumentiert und experimentell gezeigt, wie eine Hierarchie nach dem Vorbild von DMOZ Verwendung finden kann, um eine ontologische Sicht auf Suchergebnisse abzuleiten. Aus heutiger Sicht kombiniert das ideale Verfahren zur Topic-Identifikation Vorschläge, die auf Basis einer höheren Ontologie gewonnen werden, mit einem internen Verfahren, das die Eigenschaften 1 - 5 der vorgestellten Systematik optimiert. Die wohl kritischste Frage in diesem Zusammenhang ist die nach der Verfügbarkeit sinnvoller und „vollständiger“ höherer Ontologien.

Literatur

[1] M. Anderka and N. Lipka. Hierarchische Textklassifikation als Verfahren zur Topic-Identification. Master's thesis, Paderborn University, July 2006.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[3] F. Beil, M. Ester, and X. Xu. Frequent term-based text clustering. In *Proceedings of KDD '02*, 2002. ACM Press.

[4] C. Clifton and R. Cooley. Topcat: Data mining for topic identification in a text corpus. In *Proceedings of PKDD'99*, 1999.

[5] J. Davies, D. Fensel, and F. van Harmelen. *Towards the Semantic Web: Ontology-Driven Knowledge Management*. John Wiley & Sons, New York, 2003.

[6] I. S. Dhillon, S. Mallela, and R. Kumar. Enhanced word clustering for hierarchical text classification. In *Proceedings of KDD'02*. ACM Press, 2002.

[7] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin, and J. Y. Zien. SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation. In *Proceedings of WWW'03*, May 2003.

[8] A. Doan, J. Madhavan, R. Dhamankar, P. Domingos, and A. Halevy. Learning to match ontologies on the semantic web. *The VLDB Journal*, 12(4):303–319, 2003.

[9] S. T. Dumais and H. Chen. Hierarchical Classification of Web Content. In *Proceedings of SIGIR-00*, pages 256–263, Athens, Greece, 2000. ACM Press, New York, US.

[10] A. El-Hamdouchi and P. Willett. Comparison of Hierarchic Agglomerative Clustering Methods for Document Retrieval. *The Computer Journal*, 32(3):220–227, 1989.

[11] L. Ertöz, M. Steinbach, and V. Kumar. Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach. In *Proceedings of Text Mine '01, Workshop on Data Mining*, 2001.

[12] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning. Domain-Specific Keyphrase Extraction. In *Proceedings of IJCAI'99*, pages 668–673. Morgan Kaufmann Publishers Inc., 1999.

[13] B. C. M. Fung, K. Wang, and M. Ester. Hierarchical document clustering using frequent itemsets. *Proceedings of ICDM*. SIAM, 2003.

[14] E.-H. Han, G. Karypis, and V. Kumar. Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification. In *Proceedings of PAKDD*, 2001.

[15] X. He, C. H. Q. Ding, H. Zha, and H. D. Simon. Automatic Topic Identification Using Webpage Clustering. In *Proceedings of ICDM'01*, November 2001.

[16] D. Koller and M. Sahami. Hierarchically Classifying Documents Using Very Few Words. In *Proceedings of ICML-97*, 1997.

[17] K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and R. Krishnapuram. A hierarchical monothetic document clustering algorithm for summarization and browsing search results. In *Proceedings of WWW 2004*, 2004.

[18] K. Lagus and S. Kaski. Keyword Selection Method for Characterizing Text Document Maps. In *Proceedings of ICANN'99*, 1999.

[19] A. K. McCallum, R. Rosenfeld, T. M. Mitchell, and A. Y. Ng. Improving Text Classification by Shrinkage in a Hierarchy of Classes. In *Proceedings of ICML-98*, 1998.

[20] D. Mladenic. *Machine Learning on non-homogeneous, distributed text data*. PhD thesis, University of Ljubljana, 1998.

[21] A. Popescul and L. H. Ungar. Automatic Labeling of Document Clusters. Unpublished manuscript. `http://citeseer.nj.nec.com/popescul00automatic.html`, 2000.

[22] M. E. Ruiz and P. Srinivasan. Hierarchical Text Categorization Using Neural Networks. *Information Retrieval*, 5(1):87–118, 2002.

[23] M. Sanderson and W. B. Croft. Deriving Concept Hierarchies from Text. In *Research and Development in Information Retrieval*, August 1999.

[24] R. Soto. Learning and performing by exploration: label quality measured by latent semantic analysis. In *Proceedings of the CHI'99*. ACM Press, 1999.

[25] K. Sparck-Jones. Automatic Summarising: Factors and Directions. In *Advances in automatic text summarisation*. MIT Press, Cambridge, MA, 1998.

[26] B. Stein and S. Meyer zu Eißen. Document Categorization with MajorClust. In *Proceedings of WITS 02*. Technical University of Barcelona, December 2002.

[27] B. Stein, S. Meyer zu Eißen, and F. Wißbrock. On Cluster Validity and the Information Need of Users. In *Proceedings of AIA '03), Benalmádena, Spain*, 2003. ACTA Press.

[28] A. Sun and E.-P. Lim. Hierarchical text classification and evaluation. In *Proceedings of ICDM'01*, November 2001.

[29] Y.-H. Tseng. Multilingual keyword extraction for term suggestion. In *Proceedings of SIGIR '98*, 1998.

[30] O. Zamir and O. Etzioni. Web Document Clustering: A Feasibility Demonstration. In *SIGIR'98*, pages 46–54, University of Washington, Seattle, USA, 1998.

[31] Y. Zhao and G. Karypis. Criterion Functions for Document Clustering: Experiments and Analysis. Technical Report 01-40, Univercity of Minnesota, Feb 2002.