

Ansätze zur Bestimmung von *Locality* für deutsche Webseiten

Raiko Eckstein, Andreas Henrich, Volker Lüdecke

Otto-Friedrich-Universität Bamberg

D-96045, Bamberg, Deutschland

mail@raikoeckstein.de, {andreas.henrich|volker.luedecke}@wiai.uni-bamberg.de

Abstract

Das geographische Information Retrieval (GeoIR) berücksichtigt bei Suchanfragen – insb. nach Webseiten – neben dem Inhalt von Dokumenten auch eine räumliche Komponente, um gezielt nach Seiten suchen zu können, die für eine spezifische Region bedeutsam sind. Dazu müssen GeoIR-Systeme den geographischen Kontext einer Webseite erkennen können und in der Lage sein zu entscheiden, ob eine Seite überhaupt regional-spezifisch („lokal“) ist oder einen rein informativen Charakter besitzt, der keinen geographischen Bezug besitzt.

Im Folgenden werden Ansätze vorgestellt, Merkmale lokaler Seiten zu ermitteln und diese für eine Einteilung von Webseiten in globale und lokale Seiten zu verwenden. Dabei sollen insbesondere die sprachlichen und geographischen Eigenschaften deutscher Webseiten berücksichtigt werden.

1 Einführung

Nach einer Studie von [Sanderson and Kohler, 2004] weisen etwa 20% aller Suchanfragen im Web einen geographischen Kontext auf. Das bedeutet, dass in diesen Fällen insbesondere Seiten gesucht werden, die einen bestimmten lokalen Bezug haben. Herkömmliche Suchdienste sind bei der Anfragebearbeitung auf die in der Anfrage verwendeten Begriffe beschränkt, die über boolesche Operatoren miteinander verknüpft werden können. Eine Suche nach „Hotel“, „bei“ und „Bamberg“ würde dementsprechend nur solche Webseiten liefern können, die alle Begriffe enthalten, unabhängig von den tatsächlichen geographischen Kontexten der jeweiligen Seiten. Das Ergebnis wäre, dass viele der Seiten im Suchergebnis keinen direkten Bezug zu Bamberg hätten, sondern beispielsweise zu Reisebüros oder Hotelsuchmaschinen gehörten. Auch würden Seiten von Hotels aus dem Umland von Bamberg dadurch in der Regel nicht gefunden, wenn dort der Begriff „Bamberg“ nicht enthalten ist. Die geographische Nähe würde also ebenfalls nicht berücksichtigt.

Spezialisierte geographische Suchmaschinen sollten den geographischen Aspekt des Informationswunsches eines Benutzers erkennen und in der Anfragebearbeitung explizit berücksichtigen können. Im Idealfall würden dem Benutzer die geographischen Kontexte der Ergebnisse seiner Suche z. B. mit Hilfe einer Karte veranschaulicht werden.

Das geographische Information Retrieval beschäftigt sich mit der Nutzbarmachung des geographischen Kontexts

von Dokumenten. Ein solcher Kontext besteht aus einem Ort oder einer Menge von Orten („Location“), auf die in einer Seite Bezug genommen wird, was auch als der geographische Fokus einer Seite bezeichnet wird. Daneben kann die *Locality* den Grad der Lokalität einer Seite angeben. Mit Hilfe von *Locality* soll entschieden werden können, ob der Inhalt eines Dokuments für eine lokal oder regional eingegrenzte Nutzergruppe relevant ist – wie z. B. bei einer Tourismus- oder Handwerkerseite – oder ob dieser als ortsunabhängig von Interesse und somit eher „global“ einzuordnen ist, wie es beispielsweise bei Bedienungsanleitungen der Fall ist.

Ein geographischer Suchdienst muss dementsprechend in der Lage sein, den geographischen Kontext einer Seite automatisiert zu erfassen und sowohl Location als auch *Locality* bei der Anfragebearbeitung zu berücksichtigen. In diesem Beitrag sollen Ansätze zur automatischen Bestimmung von *Locality* für deutsche Webseiten beschrieben werden. Dabei werden zunächst „white box“ Ansätze betrachtet, die mit dem Ziel verfolgt wurden, nachvollziehbare Anhaltspunkte für *Locality* zu erhalten. Anschließend werden auch lernende Verfahren wie Support Vector Machines (SVM) eingesetzt.

Im folgenden Abschnitt werden verwandte Arbeiten vorgestellt. In Abschnitt 3 werden unsere Untersuchungen zur Ermittlung spezifischer Merkmale von *Locality* beschrieben. Eine vergleichende Betrachtung und Bewertung der Ergebnisse erfolgt in Abschnitt 4. Schließlich wird eine Zusammenfassung und ein Ausblick auf weitere Arbeiten gegeben.

2 Verwandte Ansätze

In der Literatur wird beim geographischen Kontext nicht immer eindeutig zwischen Location und *Locality* unterschieden. Deswegen sollen in diesem Abschnitt verwandte Ansätze vorgestellt werden, die sich mit der Bestimmung des geographischen Kontexts bzw. Fokus befassen.

[Buyukkokten *et al.*, 1999] adressieren das Problem des *Geographical Scope*. Mit Hilfe der Registrierungsdaten einer Domain sowie anhand von identifizierten Telefonvorwahlen und Postleitzahlen auf einer Webseite wird versucht, eine geographische Ausdehnung zu ermitteln, auf die sich diese Webseite bezieht.

[McCurley, 2001] beschäftigt sich mit dem geographischen Indexieren von Webseiten und einer geographischen Navigation zwischen ihnen. Eine geographische Einordnung erfolgt bei diesem Ansatz anhand von Indikatoren, die durch Auswertung des Seiteninhalts gewonnen werden. Neben der Sprache, in der ein Dokument geschrieben ist, und die ein grober Anhaltspunkt für einen geographischen

Kontext sein kann, analysiert McCurley Adressen und Telefonnummern und beschäftigt sich mit den Problemen, die aus einer Vielzahl möglicher Formate entstehen. Schließlich wird versucht, weitere Informationen aus der Verlinkung von Webseiten zu gewinnen. Es wird hier versucht, jeder Seite einen geographischen Kontext zuzuordnen und dabei nicht zwischen lokalen und globalen Seiten unterschieden.

Einen anderen Weg beschreiten Amitay, Har'El, Sivan und Soffer in [Amitay *et al.*, 2004]. Das dort vorgestellte System *Web-A-Where* versucht, Webseiten einer bestimmten geographischen Region zuzuordnen und befasst sich in erster Linie mit der Auflösung von Mehrdeutigkeiten. Diese treten im GeoIR auf, wenn derselbe Ausdruck für unterschiedliche geographische Orte steht oder ein Ausdruck sowohl eine geographische als auch eine nicht-geographische Bedeutung tragen kann. Der in diesem System verwendete *Geotagger* extrahiert potentielle Toponyme aus Webseiten und ordnet diese einem Knoten in einer Taxonomie für die USA zu, die aus den Ebenen Stadt, Bundesstaat und Land besteht. Diese Einordnung dient anschließend zur Auflösung der genannten Mehrdeutigkeiten und zur Bestimmung eines geographischen Fokus.

Ding, Gravano und Shivakumar versuchen in [Ding *et al.*, 2000] auf ähnliche Weise, den geographischen Fokus von Webseiten zu bestimmen. Auch sie teilen das Gebiet der USA in eine dreistufige Hierarchie auf und betrachten die Linkstruktur sowie Toponyme im Seiteninhalt.

Zhang *et al.* versuchen in [Zhang *et al.*, 2006] Suchergebnisse unter Berücksichtigung von geographischer Ähnlichkeit zu ranken und nutzen dafür die Linkstruktur zwischen Seiten und eine lokal vorhandene Datenbasis mit Beispieleinträgen zum gesuchten Thema.

Markowetz, Brinkhoff und Seeger unterscheiden in [Markowetz *et al.*, 2004] zwischen Seiten von lokalem und globalem Interesse. Sie skizzieren in dieser Arbeit ihre Idee, diese *Locality* anhand ein- und ausgehender Links zu berechnen, ohne sie jedoch weiter zu verfolgen.

Den angeführten Ansätzen ist gemein, dass zwar eine geographische Zuordnung von Dokumenten bzw. Webseiten erfolgt, aber nicht betrachtet wird, ob diese Seiten überhaupt von lokaler Bedeutung sind.

Unserer Arbeit am ähnlichsten ist [Gravano *et al.*, 2003]. Dort werden Anfragen an Internetsuchmaschinen mit dem Ziel untersucht, sie lokalen oder globalen Informationsbedürfnissen zuzuordnen. Dazu kommen verschiedene Klassifizierungs-Mechanismen zum Einsatz, die auf Features angewendet werden, die aus Suchergebnissen von Google zu einer Anfrage ermittelt werden. Die verwendeten Features werden aus Häufigkeiten und Verteilungen von Toponymkandidaten im Text gebildet. Zudem wird ein C4.5 Klassifikator eingesetzt, um Webseiten in lokal oder global zu unterscheiden. Die Trainingsdaten bestehen aus 140 manuell klassifizierten Websites des Yahoo! directory, wobei Seiten, die unter Regional eingeordnet sind, als lokal bezeichnet werden und solche in allgemeinen Kategorien als global. Eine genauere Betrachtung oder Hinterfragung der *Locality* oder der Gründe für die Lokalität von Seiten findet jedoch nicht statt.

3 Merkmale von lokalen Webseiten

Im Folgenden werden Ansätze vorgestellt, mit denen die *Locality* von Webseiten automatisch bestimmt werden soll. Dabei werden die Besonderheiten der deutschen Sprache und der hierarchischen Gliederung Deutschlands explizit

	lokal	global	Stadt	BL	Land
Seiten	1756	1040	893	539	324
Anteil Toponyme	2,38%	0,75%	2,74%	2,34%	1,46%
mind. 1 Toponym	85,1%	70,7%	92,7%	76,6%	78,1%
Ø TelNr.	1,86	0,50	2,18	1,93	0,86
mind. 1 TelNr.	39,9%	24,0%	47,4%	32,8%	31,2%

Tabelle 1: Merkmale lokaler Webseiten

berücksichtigt. Unsere These ist, dass für eine Bestimmung des Lokalitätsgrades einer Seite verschiedene Merkmale, die sich aus einer Analyse des Seiteninhalts ergeben, als Indizien herangezogen und nicht ein einzelnes Kriterium verwendet werden sollte.

3.1 Locality

Eine nicht selten anzutreffende Sichtweise von *Locality* ist, dass eine Seite umso „lokaler“ ist, je enger die geographische Ausdehnung oder der geographische Fokus ist. Wird *Locality* als Maß für den Grad der Lokalität eines Dokumentes verstanden, wird deutlich, dass die Bestimmung von *Locality* kontextabhängig ist und damit auch nicht direkt mit der Größe eines geographischen Gebietes verbunden ist, sondern dass dieses im jeweiligen Kontext enthalten sein kann. Während globale Information keinen geographischen Bezugspunkt aufweist – oder dieser höchstens implizit durch die Sprache eines Dokumentes gegeben sein kann –, ist eine lokale Information stets für einen bestimmten geographischen Bereich von Interesse. So ist eine Seite, die das deutsche Autobahnnetz beschreibt, von genauso lokaler Bedeutung für das gesamte Bundesgebiet, wie ein Gasthaus von lokaler Bedeutung für ein sehr eingegrenztes Gebiet ist. Der Schluss, dass eine weite geographische Ausdehnung der *Locality* widerspricht, ist also bei diesem Verständnis von Lokalität nicht richtig.

Abschließend sei darauf hingewiesen, dass die Vielzahl möglicher Kontexte, in denen ein Dokument betrachtet werden kann, eine eindeutige Entscheidung über den Grad der Lokalität unmöglich macht. Als Beispiel sei eine Webseite einer Pizzeria genannt, die neben der Kontaktadresse mit Öffnungszeiten auch das Rezept der Hauspizza beinhaltet; aus Sicht eines Restaurantbesuchers ist dies eine lokale Seite, wohingegen ein Hobbykoch das Pizzarezept als globale Information gesucht haben könnte. In der Praxis treten darüber hinaus weitere, schwieriger zu bewertende Fälle auf.

3.2 Testkollektion

Den Untersuchungen lag eine selbst erstellte Testkollektion lokaler und nicht-lokaler Webseiten zu Grunde. Dafür wurden verschiedene Teilläste mit den Schwerpunkten Bamberg und Bayern des deutschen Regional Verzeichnisses des Open Directory¹ gecrawled und diese Webseiten danach manuell klassifiziert. Um der Problematik der geographischen Ausdehnung und der damit verbundenen unterschiedlichen Kontexte Rechnung zu tragen, wurde jedes Dokument als lokal für *Stadt*, *Bundesland* oder *Deutschland* oder als nicht-lokales Dokument bewertet. Seiten, bei denen keine plausible Einordnung möglich war, wurden verworfen. Für manche Untersuchungen wurden die 3 lokalen Klassen zu einer einzigen zusammengefasst. Insgesamt umfasst die Testkollektion ca. 3000

¹<http://dmoz.org/World/Deutsch/Regional/Europa/Deutschland/>

	$\alpha = 0$			$\alpha = 0,25$			$\alpha = 0,5$			$\alpha = 0,75$			$\alpha = 1$		
	Rec.	Prec.	F	Rec.	Prec.	F	Rec.	Prec.	F	Rec.	Prec.	F	Rec.	Prec.	F
Locality	46,6	70,9	56,3	60,7	78,4	68,4	70,8	84,1	76,9	74,5	83,2	78,6	74,5	82,5	78,3
No Locality	65,0	40,0	49,5	69,3	49,1	57,5	75,5	58,6	66,0	72,4	60,8	66,1	71,2	60,4	65,4
Accuracy	53,2			63,8			72,5			73,8			73,3		

Tabelle 2: Klassifizierung anhand absoluter Termhäufigkeiten

Webseiten. Für die Untersuchungen thematischer Zusammenhänge wurden nochmals ca. 3200 Webseiten manuell den Themen *Tourismus-*, *Behörden-*, *Finanz- und Wirtschaftsförderungs-* und *Gewerbeseiten*, sowie *Webauftritte lokaler Tageszeitungen* zugeordnet, wobei hier die geographische Ausdehnung nicht gesondert berücksichtigt wurde.

3.3 Termanalyse

Zunächst wurden die Termhäufigkeiten aller Dokumente mit dem Ziel untersucht, charakteristische Ausdrücke als Entscheidungshilfe für die Bestimmung von *Locality* zu gewinnen. Die Ergebnisse der Betrachtung zeigen, in welchen Domänen lokale Seiten gehäuft auftreten und lassen erkennen, dass eine Differenzierung nach geographischer Ausdehnung und die Berücksichtigung einer thematischen Klassifikation durchaus lohnenswert sein können. Als Terme wurden stoppwortbereinigte Wortstämme betrachtet.

Die Auswertung hat gezeigt, dass es notwendig ist, sowohl relative als auch absolute Unterschiede zwischen den Vorkommenshäufigkeiten lokaler und nicht-lokaler Seiten zu berücksichtigen. Beispielsweise kommt der Begriff *Gemeinde* in 21,5% der Stadt-lokalen vor und in 2,5% der nicht-lokalen Seiten, der Begriff *Heiligenstadt* hingegen in 6,18% der Stadt-lokalen und 0,1% der nicht-lokalen. Im Folgenden werden deshalb getrennt voneinander die absoluten Unterschiede (19% im ersten Fall) und die relativen Abweichungen der lokalen von den nicht-lokalen Seiten betrachtet (6180% im zweiten Fall, also ein gut 60mal häufigeres Vorkommen des Terms in lokalen Seiten), was im Weiteren als relativer Unterschied bezeichnet wird.

Auf lokalen Stadtseiten finden sich neben Orts- und Gemeinamen auch häufig Regionennamen sowie Regierungsbezirke. Viele Begriffe dieser Klasse können darüber hinaus dem Tourismus oder der Verwaltungsebene zugeordnet werden, wie *Sehenswürdigkeit*, *Hotel*, *Restaurant* oder *Rathaus*, *Bürgermeister* und *Gemeinderat*. Auf Bundesland-Ebene sind diese beiden Themenbereiche ebenfalls vertreten, allerdings durch andere Begriffe wie *Bayern*, *Veranstaltung*, *Mittelfranken*, *Oberfranken* etc.

Die oberste betrachtete Ebene *Deutschland* zeichnet sich unter anderem durch Terme verschiedener deutschlandweit operierender Organisationen aus, wie dem Deutschen Wetterdienst (*DWD*), dem Fahrradclub *ADFC* oder dem deutschen Sportbund. Die gefundenen Toponyme umfassen neben Stadtnamen und Bundesländern auch Gewässer und Flüsse wie die *Nordsee* und die *Donau*. Hoch gewichtet finden sich viele meteorologische Termini, die auf Wetterportale in der Testkollektion zurückzuführen sind. Dabei wurden hauptsächlich Haupteinstiegsseiten betrachtet, die in diese Klasse eingeordnet wurden. Erst auf den Unterseiten wird die geographische Ausdehnung auf Regionen und Städte eingeschränkt.

Die Termanalysen ergaben, dass bei der Berücksichtigung von geographischen Konzepten neben Orts- und Gemeinamen auch Bezeichnungen berücksichtigt werden sollten, die sowohl die administrative Ebene als auch unpräzise Regionen, wie beispielsweise die *Fränkische*

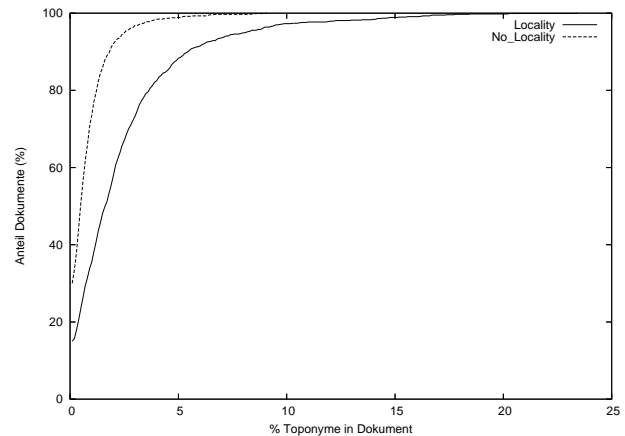


Abbildung 1: Approximative Verteilungsfunktion von Toponymen auf Seiten mit Locality

Schweiz und Gewässernamen, umfassen.

3.4 Inhaltsanalyse

Für die Gewinnung weiterer Merkmale wurden aus den Webseiten Telefonnummern und Toponyme, zu denen Orts- und Gemeinamen, Bundesländer sowie Postleitzahlen gehören, extrahiert. Gefundene Toponymkandidaten wurden dabei als Toponyme betrachtet und mögliche Mehrdeutigkeiten mit gleichlautenden nicht-geographischen Begriffen vernachlässigt. Nur besonders häufig auftretende Homonyme wurden mit Hilfe einer manuell erstellten Liste entfernt.

Als erstes Merkmal wurde der Anteil der Toponyme am Seiteninhalt betrachtet. Aus Tabelle 1 wird deutlich, dass ein Dokument nicht zwangsläufig eine lokale Bedeutung hat, wenn in dessen Inhalt geographische Begriffe auftreten: auch 70% der nicht-lokalen Seiten enthalten mindestens ein Toponym. Dennoch treten erwartungsgemäß Toponyme auf lokalen Seiten häufiger auf. Durchschnittlich sind 2,38% der Terme eines lokalen Dokuments Toponyme, gegenüber 0,75% bei nicht-lokalen Seiten. Abbildung 1 zeigt die Verteilungsfunktion der Toponymanteile beider Klassen. Der Toponymanteil ist demnach ein Indiz für eine lokale Webseite.

Bei Betrachtung der lokalen Klassen unterschieden nach deren geographischer Ausdehnung ist in Abbildung 2 erneut zu erkennen, dass jede der Klassen einen höheren Toponymanteil als nicht-lokale Dokumente aufweist. Allerdings wird die Abgrenzung der Klassen beispielsweise durch den sehr ähnlichen Verlauf von lokalen Stadt- und Bundeslandsseiten erschwert.

Als weiteres Merkmal wurde das Vorkommen von Telefonnummern betrachtet. Während etwa nur 24% der nicht-lokalen Seiten eine solche enthalten, sind dies bei lokalen etwa 40%, bei Stadt-lokalen Seiten sogar ca. 47%. Mehr als 90% der nicht-lokalen Seiten enthalten nicht mehr als eine Telefonnummer, gegenüber 73% bei lokalen Seiten. Es erscheint plausibel, diese Merkmale als Indiz für Lokalität zu

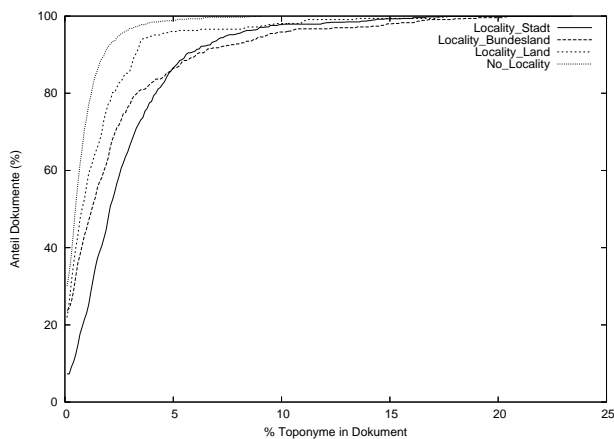


Abbildung 2: Approximative Verteilungsfunktion von Toponymen auf Seiten mit Locality (mit geographischer Ausdehnung)

verwenden, wenn sie auch nicht sehr stark differenzieren.

Daneben wurden lokale Seiten auch thematisch klassifiziert und anhand derselben Kriterien untersucht. Dabei werden signifikante Unterschiede zwischen den Themenbereichen deutlich. So weisen Tourismuseiten mit 2,99% erwartungsgemäß den größten Toponymanteil auf, wohingegen Finanz- und Wirtschaftsförderungsseiten mit durchschnittlich 0,61% Toponymen sogar unter den Werten nicht-lokaler Seiten (0,75%) liegen.

Es erscheint daher lohnenswert, die Merkmale einer Webseite in Abhängigkeit von einer thematischen Klassifikation zu bewerten. So ist beispielsweise auch zu erkennen, dass Telefonnummern auf Finanz- und Wirtschaftsförderungsseiten ein wichtiges Indiz für *Locality* darstellen, da auf 86,5% dieser Seiten mindestens eine auftrat. Dagegen beinhalten nur 13,2% der betrachteten Seiten lokaler Tageszeitungen eine Telefonnummer.

Die Zahlen zeigen, dass keines dieser Merkmale allein dazu geeignet ist, *Locality* festzustellen. Vielmehr stützen sie die These, dass dafür eine Vielzahl von Indizien herangezogen werden muss.

3.5 Klassifikation nach Wortvorkommen

Es soll nun versucht werden, die Ergebnisse der Termanalyse aus Abschnitt 3.3 dazu zu verwenden, Webseiten in lokal und nicht-lokal einzuordnen. Dabei sollen zwei Merkmale betrachtet werden: zum einen der reine Anteil der lokalen Terme in Relation zu der Gesamtwortanzahl eines Dokuments, zum anderen der gewichtete Anteil lokaler Terme, bei dem solche Terme höher gewichtet werden, die nach der Analyse aus Abschnitt 3.3 häufiger in lokalen Seiten vorkommen. Um dabei die Dominanz der häufigsten Terme abzuschwächen, wurden die Gewichte rangerhaltend mittels des Logarithmus zur Basis 10 angepasst. Mit Hilfe des Parameters α werden diese beiden Merkmale relativ zueinander gewichtet.

Dafür werden folgende Definitionen verwendet:

- n_D Anzahl der Wörter in Dokument D
- n_{lD} Anzahl der lokalen Terme in Dokument D
- w_{ck} Gewicht des Terms k in Klasse c
- tf_{dk} Vorkommenshäufigkeit des Terms k in Dokument d
- α Gewichtungsfaktor [0;1]

Die w_{ck} werden den Termauswertungen entnommen. Der *score*, der den Grad der *Locality* für eine der betrachteten Klasse angibt, wird mit Hilfe der untenstehenden Formel berechnet.

$$score = \alpha \cdot \frac{n_{lD}}{n_D} + (1 - \alpha) \cdot \frac{\sum_{k \in N_{lD}} \log_{10}(w_{ck}) \cdot tf_{dk}}{top(k)}$$

$$top(k) = \sum \text{k-größten Gewichte } w_{ck}$$

Nachfolgend werden die Ergebnisse der Klassifizierung anhand von Wortvorkommen vorgestellt, die durch die Berechnung von *score*-Werten erreicht wurden. Die Auswertungen wurden sowohl mit den relativen als auch den absoluten Unterschieden der Wortvorkommen zwischen lokalen und nicht-lokalen Seiten durchgeführt. Zur Messung des Einflusses der beiden Merkmale wurde der Parameter α variiert, um die Gewichtung der zwei Merkmale auf der Testkollektion zu untersuchen. Dabei wurden für die geographische Dimension weitere 437 und für die thematischen Klassen 758 Dokumente als Testmenge klassifiziert.

Da eine isolierte Betrachtung von *Precision* bzw. *Recall* nicht aussagekräftig genug erscheint, wird in [van Rijsbergen, 1979] das *F₁-Measure* vorgestellt. Das *F₁-Measure* beschreibt das harmonische Mittel zwischen *Recall* und *Precision*.

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Daneben gibt die *Accuracy* den Anteil korrekter Klassifizierungsvorgänge an.

Zunächst werden die absoluten Wortvorkommenshäufigkeiten bei variierendem α betrachtet, und es wird nur zwischen lokalen und nicht-lokalen Seiten unterschieden. Die Ergebnisse sind in Tabelle 2 dargestellt. Es lässt sich erkennen, dass die Ergebnisse mit größerem α besser werden.

Das beste Ergebnis nach dem F-Maß lässt sich bei einem α von 0,75 erreichen. Dabei werden 74,5% der relevanten lokalen Dokumente aus der Testmenge erkannt, und nur 17,0% der dieser Klasse zugeordneten Dokumente falsch eingeordnet.

Dieselbe Betrachtung wurde noch einmal anhand der relativen Wortvorkommen durchgeführt. Die Ergebnisse davon sind in Tabelle 3 zu sehen.

Schließlich wurde zusätzlich die Einteilung in Stadt-, Bundesland- und Deutschland-lokal berücksichtigt. In Tabelle 4 sind die entsprechenden Ergebnisse zu finden. Während bei Stadt-lokalen Seiten eine Einbeziehung charakteristischer Wörter eine leichte Verbesserung bewirkt, trifft dies auf die anderen *Locality*-Klassen nicht zu.

Aus diesen Betrachtungen scheint ersichtlich, dass der Anteil Toponyme in einem Dokument ein wichtigeres Kriterium als das Vorkommen charakteristischer Terme ist. Gleichwohl bringt die Berücksichtigung dieser Terme einen leichten Anstieg der Klassifikationsgenauigkeit. Weitere Verbesserungen sind zu erwarten, wenn diese Liste (manuell) optimiert wird. Beispielsweise sind dort noch alle Toponyme enthalten, die zum Teil sehr ortsspezifisch sind. Eine Unterscheidung von *Locality* in Klassen unterschiedlicher geographischer Ausdehnung scheint hier zunächst weniger erfolgversprechend. Untersuchungen, für jede Klasse eigene Wortlisten zu erstellen, fanden noch nicht statt und erscheinen ohne eine Verbesserung des Verfahrens hinsichtlich einer Verbesserung der Ergebnisqualität auch fraglich.

	$\alpha = 0$			$\alpha = 0,25$			$\alpha = 0,5$			$\alpha = 0,75$			$\alpha = 1$		
	Rec.	Prec.	F	Rec.	Prec.	F	Rec.	Prec.	F	Rec.	Prec.	F	Rec.	Prec.	F
Locality	55,7	84,7	67,2	64,8	85,0	73,5	66,8	84,7	74,7	68,5	85,0	75,8	66,8	84,7	74,7
No Locality	81,6	50,2	62,2	79,1	55,1	65,0	77,9	56,2	65,3	77,9	57,5	66,2	77,9	56,2	65,3
Accuracy	64,9			69,9			70,7			71,8			70,7		

Tabelle 3: Klassifizierung anhand relativer Unterschiede der Termhäufigkeiten

	$\alpha = 0$			$\alpha = 0,25$			$\alpha = 0,5$			$\alpha = 0,75$			$\alpha = 1$		
	Rec.	Prec.	F	Rec.	Prec.	F	Rec.	Prec.	F	Rec.	Prec.	F	Rec.	Prec.	F
Stadt	38,8	61,0	47,4	58,1	83,3	68,5	72,1	92,1	80,9	72,1	92,1	80,9	73,6	89,6	80,9
Bundesland	77,7	42,4	54,9	87,2	50,6	64,1	88,3	58,0	70,0	89,4	60,9	72,4	87,2	63,1	73,2
Land	38,7	38,2	38,4	44,0	54,1	48,5	44,0	55,9	49,3	46,5	58,3	51,9	46,7	63,6	53,9
No Locality	43,6	54,2	48,4	55,8	61,5	58,5	63,2	65,2	64,2	66,3	66,7	66,5	69,3	66,5	67,9
Accuracy	48,4			61,0			67,7			69,4			70,5		

Tabelle 4: Klassifizierung anhand absoluter Unterschiede der Termhäufigkeiten, nach geographischer Ausdehnung

Für eine Berücksichtigung unterschiedlicher thematischer Kontexte wurden die Seiten der Testkollektion fünf verschiedenen thematischen Klassen zugeordnet. Anschließend wurden dieselben Untersuchungen mit relativen und absoluten Wortvorkommen durchgeführt.

Die Ergebnisse zeigten, dass zwischen den einzelnen thematischen Klassen große Unterschiede in der Anwendbarkeit dieses Verfahrens bestehen. Insbesondere Behördenseiten weichen hier von den Charakteristika anderer Themenklassen ab. Für die weitere Optimierung der Erkennungsrate lokaler Seiten kann es demnach durchaus hilfreich sein, zunächst den thematischen Kontext eines Dokumentes zu erfassen, um anschließend dessen Besonderheiten bei den Ausprägungen der betrachteten Merkmale zu berücksichtigen.

3.6 Berücksichtigung weiterer Merkmale

Man kann nun versuchen, die gefundenen Indizien für *Locality* aus Abschnitt 3.4 zusätzlich zu den Gewichtungen der Wortvorkommen für die lokalen Klassen Abschnitt 3.5 zu verwenden.

Die Untersuchung beschränkt sich momentan auf die binäre Entscheidung, ob ein Dokument in die Klasse der lokalen oder nicht-lokalen Dokumente gehört. Dabei wird die geographische Testkollektion und deren Vokabular verwendet.

Es werden Auswertungen durchgeführt, die bei der *score*-Berechnung den Toponymanteil und die Anzahl der Erwähnungen von Telefonnummern berücksichtigen. Dabei wird zuerst jeweils eins der drei Indizien verwendet und mit der Wortvorkommenmethode kombiniert, wobei der Einfluss eines Indizes über einen Parameter β variiert.

$$score_{combined} = \beta \cdot score_{\alpha=0,75} + (1 - \beta)w_{Indiz}$$

Aus Abschnitt 3.4 stehen für die verschiedenen untersuchten Indizien Verteilungsfunktionen zur Verfügung. Es soll möglich sein, die Gewichtungen für eine Ausprägung eines Indizes dafür je nach Stärke des Indizes zu variieren. Ist bekannt, dass 30% der lokalen Webseiten drei Telefonnummern aufweisen und dieser Anteil bei nicht-lokalen nur bei 5% liegt, soll diese Verteilung mit in die Gewichtung eingehen.

Für die Bestimmung dieser Gewichte für die Indizien wurden verschiedene Ansätze evaluiert. Anfangs wurde als Gewichtung die relative Dokumentenhäufigkeit einer Klasse verwendet, bei der das Indiz bei einer bestimmten Ausprägung vorkommt. Die Ergebnisse für die Klasse der lo-

kalen Seiten waren nach dem F-Maß schlechter als bei der reinen Auswertung nach den Wortvorkommen. Einzig bei Gewichtung der Wortvorkommen zu 90% und zu 10% des Toponymanteils lassen sich die Basisergebnisse einstellen. Bei den nicht-lokalen Dokumenten lässt sich eine leichte Steigerung des F-Maßes erreichen. Aufgrund dieser Ergebnisse insbesondere für die Klasse der lokalen Dokumente wurde nach anderen Gewichtungsmöglichkeiten für Indizien gesucht, die auf *Locality* hinweisen.

Im nächsten Schritt wurde die approximative Verteilungsfunktion der verschiedenen Indizien in den betrachteten Klassen verwendet. Als Gewichtung kam folgende Berechnung zum Einsatz:

$$w(X = x) = 1 - \hat{F}_X^*(x)$$

Die Gewichte eines Indizes in einem Dokument mit der Ausprägung x ergeben sich aus der Differenz von 1 mit dem Wert der approximativen Verteilungsfunktion an der Stelle $X = x$. Die Gewichtung eines Indizes stellt also den Anteil der Dokumente dar, die das Indiz x -mal oder mehr aufweisen.

Als Basisdaten kommen die Ergebnisse der Klassifizierung der Seiten der geographischen Taxonomie zum Einsatz, die bei Betrachtung der absoluten Differenzen der Vorkommenshäufigkeiten in Abschnitt 3.3 errechnet wurden. Als Referenzwerte werden die besten Ergebnisse, die bei Betrachtung des *score* erzielt worden sind, verwendet. Die Auswertungen sind in den Tabellen 5 und 6 dargestellt. Zuerst wird der Einfluss des Toponymanteils auf das Ergebnis der *Locality*-Berechnung untersucht.

Die nächste Auswertung befasst sich mit dem Einfluss von Telefonnummern auf einer Seite und kombiniert diesen mit den Ergebnissen der Wortvorkommenmethode. Auch hier ist eine leichte Verbesserung der Erkennungsraten festzustellen.

In einer letzten Untersuchung wird eine Kombination der drei Indizien mit der Wortvorkommenmethode durchgeführt. Dabei wurde die Gewichtung zwischen 0,5 und 0,1 variiert. Der Anteil der Indizien wurde auf jedes Einzelindiz zu gleichen Teilen verteilt. Die hierbei erzielten Ergebnisse sind in Tabelle 7 wiedergegeben. Es ist zu erkennen, dass die F-Maßwerte der lokalen Klasse bei jeder betrachteten Gewichtung über denen der Basismethode liegen.

3.7 Klassifikation mit lernenden Verfahren

Alternativ zur oben vorgestellten Methode fand zusätzlich eine Klassifikation nach lokalen Webseiten mit lernenden

	Toponyme 0,5			Toponyme 0,25			Toponyme 0,1		
	$\alpha = 0,75$			$\alpha = 0,75$			$\alpha = 0,75$		
	Rec.	Prec.	F-Maß	Rec.	Prec.	F-Maß	Rec.	Prec.	F-Maß
Locality	85,9	73,8	79,4	80,2	80,5	80,3	76,2	82,9	79,4
No Locality	44,2	63,2	52,0	64,4	64,0	64,2	71,2	62,0	66,3
Accuracy	71,2			74,4			74,4		

Tabelle 5: Kombination von Toponym- und Wortvorkommen

	Tel.Nr. 0,5			Tel.Nr. 0,25			Tel.Nr. 0,1		
	$\alpha = 0,75$			$\alpha = 0,75$			$\alpha = 0,75$		
	Rec.	Prec.	F-Maß	Rec.	Prec.	F-Maß	Rec.	Prec.	F-Maß
Locality	82,6	79,4	80,9	78,5	82,4	80,5	75,8	83,1	79,3
No Locality	60,7	65,6	63,1	69,3	63,8	66,5	71,8	61,9	66,5
Accuracy	74,8			75,3			74,4		

Tabelle 6: Kombination von Telefonnummern- und Wortvorkommen

Verfahren statt. Dabei wurden ein Naive Bayes Klassifikator (vgl. [Rish, 2001]) und Support Vector Machines (vgl. [Boser *et al.*, 1992]) verwendet. Beide Verfahren eignen sich für die Klassifikation von Texten (vgl. [Lewis and Ringuette, 1994], [Joachims, 1998], [Yang, 1999]). Für die Klassifizierung nach dem Bayes Theorem wird die Java-Bibliothek Classifier4J² verwendet. Beim Training mit dem Bayes Klassifikator ist anzumerken, dass die verwendete Bibliothek keine Mehrklassenklassifizierung unterstützt, d. h. beim Lernvorgang ist darauf zu achten, zu der positiven Trainingsdatenmenge auch negative Trainingsbeispiele anzugeben. Es muss explizit beschrieben werden, was das Charakteristische an Dokumenten mit *Locality* ist und wie die Komplementärmenge der nicht-lokalen Webseiten aussieht. Als Support Vector Machine kam die Bibliothek LIBSVM³ zum Einsatz (vgl. [Chang and Lin, 2001]). Zur Klassifizierung wurden ein *linearer* und ein *radial basis function (rbf)* Kernel verwendet. Die Mehrklassenklassifizierung greift auf die *One-Against-One* Methode zurück (vgl. [Hsu and Lin, 2002] und [Chin, 1999]). Dazu werden bei der Klassifizierung von k Klassen $k(k-1)/2$ binäre Probleme gelöst. Die Entscheidung, in welche Klasse ein Dokument gehört, wird über die *Max Wins* Strategie bestimmt. Als Ergebnisklasse wird diejenige gewählt, für die bei den binären Entscheidungen am häufigsten eine Klassenzugehörigkeit bestimmt wurde.

Zunächst wird die Klassifizierung nach *lokalen* und *nicht-lokalen* Webseiten betrachtet. Die Ergebnisse sind in Tabelle 8 dargestellt. Die Support Vector Machine mit *rbf* Kernel erreicht bei Betrachtung des F-Maßes beider Klassen die besten Werte.

Tabelle 9 stellt die Ergebnisse der Klassifizierung nach den drei untersuchten geographischen Ausdehnungen dar. Auch hier ist die Support Vector Machine am erfolgreichsten. Die Klassifizierung von Dokumenten funktioniert bei der Klasse der lokalen Dokumente auf Stadtebene mit beiden Kernen sehr gut.

Bei der Klasse der lokalen Seiten auf Bundeslandebene sinken die Ergebnisse bei allen drei Verfahren, auf Landesebene verschlechtern sich die Klassifizierungsergebnisse sogar signifikant.

Betrachtet man nur die Rate der richtig klassifizierten Dokumente bezogen auf die Gesamtzahl der relevanten Dokumente dieser Klasse ergibt sich der höchste Recall von 87,8% bei dem Naive Bayes Klassifikator, der

allerdings mit einer Precision von 55,0% am schlechtesten abschneidet, d. h. es werden zuviele Dokumente anderer Klassen dieser Klasse zugeordnet. Begründet werden kann dies durch die Bestimmung der Klasse beim verwendeten Naive Bayes Klassifikator. Ein Dokument wird einer (lokalen) Positivklasse, erst dann zugeordnet, wenn die errechnete Wahrscheinlichkeit der Klassenzugehörigkeit größer gleich 80% beträgt. Deshalb fallen Dokumente, die eigentlich zu einer Klasse gehören mitunter in die Klasse der nicht-lokalen Dokumente (Negativklasse), da die Einzelwahrscheinlichkeit für eine Zuordnung nicht ausreicht. Die Support Vector Machine betrachtet alle Klassen und errechnet die Wahrscheinlichkeiten so, dass diese addiert 100% ergeben. Als Ergebnisklasse wird die mit der höchsten Wahrscheinlichkeit gewählt.

3.8 Berücksichtigung thematischer Kontexte

Darüber hinaus wurden die lernenden Verfahren auch verwendet, um Seiten direkt thematisch lokalen Seiten zuzuordnen. Dabei wurde jede Seite entweder als lokal und einem bestimmten thematischen Gebiet zugeordnet oder als nicht-lokal eingestuft. Die Klasse der nicht-lokalen Dokumente enthält demnach Webseiten aus allen Themenbereichen, die nicht als von lokalem Interesse eingestuft wurden. Die Ergebnisse der Untersuchung zeigt Tabelle 10. Auffällig ist das schlechte Abschneiden der Gewerkekategorie, für die offenbar andere Merkmale zum Tragen kommen als in den anderen Klassen.

4 Evaluierung

Abschließend soll eine zusammenfassende und kritische Betrachtung der erzielten Ergebnisse sowohl bei den Textanalysen als auch den Termanalysen stattfinden. Außerdem werden die Ergebnisse der automatischen Klassifizierung nach den beiden lernenden Verfahren – Naive Bayes Klassifikator und Support Vector Machine – dem Ansatz gegenübergestellt, *Locality* anhand von einzelnen Merkmalen zu bestimmen.

4.1 Indizien für Locality

Die Textanalysen aus Abschnitt 3.4 haben ergeben, dass sich lokale Webseiten in ihrem Seiteninhalt von nicht-lokalen unterscheiden. Dies trifft sowohl auf verschiedene untersuchte Indizien zu als auch für das verwendete Vokabular. Je nach betrachtetem Indiz – Telefonnummern oder Toponymanteil – sind die Unterschiede bei den Hinweisen auf *Locality* unterschiedlich stark ausgeprägt. Bei Betrachtung von lokalen und nicht-lokalen Dokumenten kann

²<http://classifier4j.sourceforge.net/>

³<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

	Kombination 0,5			Kombination 0,25			Kombination 0,1		
	$\alpha = 0,75$			$\alpha = 0,75$			$\alpha = 0,75$		
	Rec.	Prec.	F-Maß	Rec.	Prec.	F-Maß	Rec.	Prec.	F-Maß
Locality	83,2	78,5	80,8	77,9	82,6	80,1	75,8	83,1	79,3
No Locality	58,3	65,5	61,7	69,9	63,3	66,5	71,8	61,9	66,5
Accuracy	74,4			75,1			74,4		

Tabelle 7: Klassifizierungsergebnisse der Kombination lokaler Indizien

	Naive Bayes			SVM (linear kernel)			SVM (rbf kernel)		
	Recall	Precision	F-Maß	Recall	Precision	F-Maß	Recall	Precision	F-Maß
Locality	77,9	82,6	80,2	84,7	81,5	83,1	86,9	81,5	84,1
No Locality	70,5	64,0	67,1	65,4	70,3	67,8	63,8	72,7	68,0
Accuracy	75,3			77,8			78,7		

Tabelle 8: Klassifizierungsergebnisse durch lernende Verfahren

vor allem der Toponymanteil ein guter Hinweis auf lokale Webseiten sein. Der Schluss auf die Tatsache, dass lokale Webseiten durchschnittlich mehr Toponyme aufweisen als nicht-lokale Seiten, ist somit möglich.

Das Kriterium der Telefonnummernanzahl auf einer Webseite ist nicht so trennscharf wie der Toponymanteil. Ersichtlich wird das durch den hohen Anteil der Webseiten beider Klassen, die keine Telefonnummer aufweisen.

Unterteilt man die lokale Klasse nach der geographischen Ausdehnung, sind auch Unterschiede bei der Verteilung der verschiedenen Indizien festzustellen. Allerdings fällt hier die Abgrenzung der einzelnen Klassen mitunter schwer, da die Verteilung der verschiedenen Klassen nicht immer trennscharf ist.

Es erscheint schwer möglich, für alle untersuchten Klassen und Taxonomien Heuristiken zu formulieren, anhand derer man eine Klassifizierung nach *Locality* vornehmen bzw. eine auf anderen Kriterien beruhende verbessern kann. Bei Beschränkung auf zwei Klassen erscheint die Formulierung einfacher Heuristiken nach dem Schema noch relativ einfach realisierbar zu sein.

4.2 Klassifizierung durch lernende Verfahren

Bei den lernenden Verfahren erzielte die Support Vector Machine mit rbf Kernel insgesamt die besten Ergebnisse, die auch besser waren als diejenigen, die das Verfahren anhand von Einzelmerkmalen erzielen konnte.

Aus diesem Grund wurde zuletzt noch versucht, diese Merkmale auch durch die SVM berücksichtigen zu lassen. Dazu wurden diese Merkmale den Features für die SVM hinzugefügt. In den ersten Versuchen ließen sich hier jedoch nur bei Mitberücksichtigung des Toponymanteils geringe Verbesserungen erzielen, wodurch die Accuracy um 0,3 erhöht werden konnte. Weitere Optimierungen könnten hier in der Zukunft die Klassifikationsgenauigkeit weiter erhöhen.

4.3 Geschwindigkeit

Bei der reinen Bewertung der Klassifikationsqualität scheint die SVM das geeignetste Verfahren zur Bestimmung von *Locality* zu sein. Dass eine Betrachtung von Merkmalen von Lokalität dennoch eine Daseinsberechtigung hat, zeigt sich spätestens bei einer Berücksichtigung der Verarbeitungsgeschwindigkeit. Die Bestimmung von *Locality* allein anhand von Wortvorkommen war bei den Untersuchungen um den Faktor 4-5 schneller als durch den Einsatz von SVMs, und etwa 11mal schneller als mit Hilfe des Naive Bayes Klassifikators.

5 Zusammenfassung und Ausblick

In dieser Arbeit wurde gezeigt, dass lokale Webseiten bestimmte Merkmale aufweisen, die auch dazu verwendet werden können, über die *Locality* einer Webseite zu entscheiden.

Als wesentliches Merkmal wurde dabei das Vorkommen von Toponymen im Seiteninhalt identifiziert und die Annahme bestätigt, dass viele Toponyme auf eine Seite von lokaler Bedeutung hinweisen. Gleichwohl konnte auch gezeigt werden, dass das Vorkommen von Toponymen allein nicht ausreicht, um eine Webseite als lokal ansehen zu können.

Daneben scheint auch das Vokabular ein geeignetes Kriterium für Lokalität zu sein. Lernende Verfahren konnten hier erwartungsgemäß die besten Ergebnisse aufweisen, wobei ein einfacherer Ansatz ebenfalls zu ordentlichen Ergebnissen kommt und dabei wesentlich schneller arbeiten kann.

Die Ergebnisse zeigen, dass es grundsätzlich möglich ist, Webseiten einerseits nach dem Kriterium der *Locality* und andererseits nach der geographischen Ausdehnung einer Seite – im Sinne des *Geographical Scopes* – zu klassifizieren. Auch unterscheiden sich Webseiten aus unterschiedlichen Themenbereichen in ihren Merkmalsausprägungen, was sich in mehrstufigen Klassifikationsvorgängen ausnutzen ließe.

Weitere Verbesserungen der Klassifikationsgenauigkeit bei Beibehaltung einer hohen Verarbeitungsgeschwindigkeit, die für einen Indexierungsprozess einer Web-Suchmaschine unabdingbar ist, sind das Ziel weiterer Untersuchungen. Dabei sollen weitere Merkmale erkannt und betrachtet werden, wie z. B. die Linkstruktur zwischen lokalen Webseiten oder erkannte Adressangaben, sowie insbesondere die gewonnenen Erkenntnisse aus dieser Arbeit genutzt werden, um Merkmale geeignet mit Klassifikationen zu kombinieren, um die aussagekräftigsten Indizien für bestimmte Kontexte heranziehen zu können.

Letztlich soll als Ergebnis der Bestimmung von Lokalität ein nicht-binärer Wert stehen, der den Grad der Lokalität oder die Wahrscheinlichkeit für Lokalität angibt. Ein solcher Wert kann in weiteren Prozessschritten der Anfragebearbeitung genutzt werden: einerseits könnten von vornherein nur solche Dokumente überhaupt für eine ortsbezogene Suche betrachtet werden, die ein Mindestmaß an Lokalität aufweisen, andererseits ist auch eine Integration in ein Rankingverfahren ohne weiteres möglich. Kombinierte thematische und geographische Rankings, wie sie beispielsweise in [Vaid *et al.*, 2005] und [Martins *et al.*, 2005] betrachtet

	Naive Bayes			SVM (linear kernel)			SVM (rbf kernel)		
	Recall	Precision	F-Maß	Recall	Precision	F-Maß	Recall	Precision	F-Maß
Stadt	78,7	76,8	77,7	87,7	87,7	87,7	86,9	88,3	87,6
Bundesland	46,1	77,4	57,7	71,9	76,2	74,0	75,3	77,9	76,6
Land	8,57	60,0	15,0	47,1	71,7	56,9	50,0	76,1	60,3
No Locality	87,8	55,0	67,7	76,3	64,3	69,8	77,6	65,4	71,0
Accuracy	64,1			73,9			75,3		

Tabelle 9: Klassifizierungsergebnisse durch lernende Verfahren nach geographischen Ausdehnungen

	Naive Bayes			SVM (linear kernel)			SVM (rbf kernel)		
	Recall	Precision	F-Maß	Recall	Precision	F-Maß	Recall	Precision	F-Maß
Tourismus	72,7	62,4	67,2	71,6	78,0	74,6	75,4	80,2	77,7
Behörden	50,0	80,5	61,7	65,6	88,0	75,1	65,2	89,8	75,6
Tageszeitungen	40,2	97,1	56,9	82,9	100,0	90,7	81,7	100,0	89,9
Finanzen	44,4	92,3	60,0	81,5	81,5	81,5	88,9	82,7	85,7
Gewerbe	13,0	75,0	22,2	37,0	77,3	50,0	37,0	58,6	45,3
No Locality	92,7	45,9	61,4	87,8	51,4	64,9	86,0	52,0	64,8
Accuracy	61,2			72,6			73,1		

Tabelle 10: Vergleich der Klassifizierungsergebnisse bei verschiedenen thematischen Klassen

werden, lassen sich somit um einen Lokalfaktor erweitern und wären damit nicht länger auf die inhaltliche Ähnlichkeit und eine geographische Nähe beschränkt.

Literatur

- [Amitay *et al.*, 2004] Einat Amitay, Nadav Har’El, Ron Sivan, and Aya Soffer. Web-a-where: geotagging web content. In *SIGIR ’04*, pages 273–280, New York, NY, USA, 2004. ACM Press.
- [Boser *et al.*, 1992] Bernhard E. Boser, Isabelle Guyon, and Vladimir Vapnik. A training algorithm for optimal margin classifiers. In *Computational Learning Theory*, pages 144–152, 1992.
- [Buyukkokten *et al.*, 1999] Orkut Buyukkokten, Junghoo Cho, Hector Garcia-Molina, Luis Gravano, and Narayanan Shivakumar. Exploiting geographical location information of web pages. In *WebDB (Informal Proceedings)*, pages 91–96, 1999.
- [Chang and Lin, 2001] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001.
- [Chin, 1999] K. Chin. Support vector machines applied to speech pattern classification. Master’s thesis, University of Cambridge, 1999.
- [Ding *et al.*, 2000] Junyan Ding, Luis Gravano, and Narayanan Shivakumar. Computing geographical scopes of web resources. In *26th Intl. Conf. on Very Large Databases, VLDB 2000*, Cairo, Egypt, September 10–14 2000.
- [Gravano *et al.*, 2003] Luis Gravano, Vasileios Hatzivasiloglou, and Richard Lichtenstein. Categorizing web queries according to geographical locality. *CIKM’03*, 2003.
- [Hsu and Lin, 2002] C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines. In *IEEE Transactions on Neural Networks*, number 13, pages 415–425, 2002.
- [Joachims, 1998] Thorsten Joachims. Text categorization with support vector machines: learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conf. on Machine Learning*, number 1398, pages 137–142, Chemnitz, DE, 1998. Springer, Heidelberg.
- [Lewis and Ringuette, 1994] David D. Lewis and Marc Ringuette. A comparison of two learning algorithms for text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 81–93, SIGIR 2004, 1994.
- [Markowetz *et al.*, 2004] Alexander Markowetz, Thomas Brinkhoff, and Bernhard Seeger. Geographic information retrieval. *Proceedings of the 3rd Intl. Workshop on Web Dynamics, WWW2004, New York, NY, USA*, 2004.
- [Martins *et al.*, 2005] Bruno Martins, Mario J. Silva, and Leonardo Andrade. Indexing and ranking in geo-ir systems. In *GIR ’05: Proceedings of the 2005 workshop on Geographic information retrieval*, pages 31–34, New York, NY, USA, 2005. ACM Press.
- [McCurley, 2001] Kevin S. McCurley. Geospatial mapping and navigation of the web. In *WWW ’01: Proceedings of the 10th intl. conf. on World Wide Web*, pages 221–229, New York, NY, USA, 2001. ACM Press.
- [Rish, 2001] Irina Rish. An empirical study of the naive bayes classifier. In *IJCAI-01 workshop on Empirical Methods in AI*, 2001.
- [Sanderson and Kohler, 2004] Mark Sanderson and Janet Kohler. Analyzing geographic queries. In *Proceedings of the ACM SIGIR Workshop on Geographic Information Retrieval*, Sheffield, UK, 2004.
- [Vaid *et al.*, 2005] Subodh Vaid, Christopher B. Jones, Hideo Joho, and Mark Sanderson. Spatio-textual indexing for geographical search on the web. In *SSTD*, pages 218–235, 2005.
- [van Rijsbergen, 1979] C.J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2nd edition, 1979.
- [Yang, 1999] Yiming Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1/2):69–90, 1999.
- [Zhang *et al.*, 2006] Jianwei Zhang, Yoshiharu Ishikawa, Sayumi Kurokawa, and Hiroyuki Kitagawa. Localrank: A prototype for ranking web pages with database considering geographical locality. *Lecture Notes in Computer Science*, Volume 3841:1209–1213, 2006.