Technologien zur Wiederverwendung von Texten aus dem Web

Martin Potthast

Bauhaus-Universität Weimar
martin.potthast@uni-weimar.de

Abstract: Texte aus dem Web können einzeln oder in großen Mengen wiederverwendet werden. Ersteres wird Textwiederverwendung und letzteres Sprachwiederverwendung genannt. Zunächst geben wir einen Überblick darüber, auf welche Weise Text und Sprache wiederverwendet und wie Technologien des Information Retrieval in diesem Zusammenhang angewendet werden können. In der übrigen Arbeit werden dann eine Reihe spezifischer Retrievalaufgaben betrachtet, darunter die automatische Erkennung von Textwiederverwendungen und Plagiaten, der Vergleich von Texten über Sprachen hinweg, sowie die Wiederverwendung des Webs zur Verbesserung von Suchergebnissen und zur Unterstützung des Schreibens von fremdsprachigen Texten.

1 Einleitung

Etwas wiederzuverwenden bedeutet, es nach seiner ersten Verwendung einem neuen Zweck zuzuführen. Die Wiederverwendung uns umgebender Dinge ist ein alltäglicher Vorgang. Dennoch wird nur selten davon gesprochen, einen Text wiederzuverwenden. Stattdessen spricht man von Zitaten, Übersetzungen, Paraphrasen, Metaphrasen, Zusammenfassungen, Textbausteinen und nicht zuletzt Plagiaten. Sie alle können mit dem Begriff der „Textwiederverwendung" umschrieben und darunter angeordnet werden (siehe Abbildung 1). Einen Text ein zweites Mal zu verwenden ist nichts ungewöhnliches, sondern fester Bestandteil des Schreibens in vielen Genres. Es ist jedoch noch weitgehend unbekannt, wie weit verbreitet die Wiederverwendung von Text heute ist. Das liegt vor allem daran, dass die nötigen Werkzeuge fehlen, dieses Phänomen im großen Stil zu betrachten.

Im Web stehen große Mengen Text zur freien Verfügung. Abgesehen davon, sie einzeln wiederzuverwenden, besteht eine weitere Möglichkeit darin, sie insgesamt wiederzuverwenden, um eine bestimmte Aufgabe (automatisch) zu erledigen. Man spricht in diesem Zusammenhang auch von Sprachwiederverwendung. Da Texte im Web auf unzählige Weisen mit anderen Objekten in Verbindung stehen, liegt hierin großes Potenzial, neue Aufgaben zu finden, die durch geschickte Sprachwiederverwendung besser gelöst werden können. Aus Sicht der Informatik befassen wir uns daher mit folgenden Forschungsfragen:

- Wie und in welchem Umfang können Textwiederverwendungen erkannt werden?
- Welche Aufgaben können durch Sprachwiederverwendung unterstützt werden?

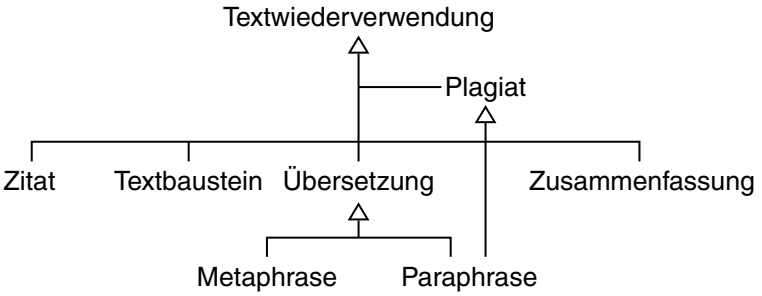

Abbildung 1: Taxonomie bekannter Formen der Textwiederverwendung

1.1 Beiträge

Der vorliegende Text ist eine zusammenfassende, paraphrasierte Übersetzung der in Englisch verfassten Dissertation „Technologies for Reusing Text from the Web" [Pot11]. Die Dissertation ist in zwei Teile gegliedert. Im ersten Teil präsentieren wir Technologien zur Erkennung von Textwiederverwendungen und leisten folgende Beiträge: (1.) Ein einheitlicher Überblick über projektionsbasierte- und einbettungsbasierte Fingerprintingverfahren zur Erkennung fast identischer Texte, sowie die erstmalige Evaluierung einer Reihe dieser Verfahren auf den Revisionshistorien der Wikipedia. (2.) Ein neues Modell zum sprachübergreifenden, inhaltlichen Vergleich von Texten. Das Modell kommt ohne Wörterbücher oder Übersetzungsverfahren aus, sondern benötigt nur eine Menge von Pärchen themenverwandter Texte. Wir vergleichen das Modell in mehreren Sprachen mit herkömmlichen Modellen. (3.) Die erste standardisierte Evaluierungsumgebung für Algorithmen zur Plagiatserkennung. Sie besteht aus Maßen, die die Erkennungsleistung eines Algorithmus' quantifizieren und einem großen Korpus von Plagiaten. Die Plagiate wurden automatisch generiert sowie manuell, mit Hilfe von Crowdsourcing, erstellt. Darüber hinaus haben wir drei internationale Wettbewerbe veranstaltet, in denen insgesamt 32 Forschergruppen ihre Erkennungsansätze gegeneinander antreten ließen.

Im zweiten Teil präsentieren wir auf Sprachwiederverwendung basierende Technologien für drei verschiedene Retrievalaufgaben: (4.) Ein neues Modell zum medienübergreifenden, inhaltlichen Vergleich von Objekten aus dem Web. Das Modell basiert auf der Auswertung der zu einem Objekt vorliegenden Kommentare. In diesem Zusammenhang identifizieren wir Webkommentare als eine in der Forschung bislang vernachlässigte Informationsquelle und stellen die Grundlagen des Kommentarretrievals vor. (5.) Zwei neue Algorithmen zur Segmentierung von Websuchanfragen. Die Algorithmen nutzen Web n-Gramme sowie Wikipedia, um die Intention des Suchenden in einer Suchanfrage festzustellen. Darüber hinaus haben wir mittels Crowdsourcing ein neues Evaluierungskorpus erstellt, das zwei Größenordnungen größer ist als bisherige Korpora. (6.) Eine neuartige Suchmaschine, genannt NETSPEAK, die die Suche nach geläufigen Formulierungen ermöglicht. NETSPEAK indiziert das Web als Quelle für geläufige Sprache in der Form von n-Grammen und implementiert eine Wildcardsuche darauf. Im Folgenden werden die Beiträge genauer beschrieben und eine Auswahl an Ergebnissen präsentiert.

2 Erkennung von Textwiederverwendungen

Für ein Dokument, dessen Originalität in Frage steht, bestehe die Aufgabe darin, alle aus anderen Dokumenten wiederverwendeten Passagen zu identifizieren. Dazu gibt es drei Ansätze: (1.) Die Suche nach Originaldokumenten. (2.) Die Prüfung, ob das Dokument vom angeblichen Autor geschrieben wurde. (3.) Die Prüfung, ob alle Passagen des Dokuments vom gleichen Autor geschrieben wurden. Mit dem ersten Ansatz werden die Schritte, die der Autor des verdächtigen Dokuments zum Auffinden von Texten zur Wiederverwendung gehen musste, nachvollzogen. Die beiden anderen Ansätze basieren darauf, Autoren anhand ihres Schreibstils auseinander zu halten. Im Grunde lassen sich jedoch alle drei Ansätze darauf reduzieren, das verdächtige Dokument (passagenweise) mit anderen zu vergleichen, wobei nach einer „überraschenden" Gleichförmigkeit in Syntax oder Semantik gesucht wird. Bestimmte syntaktische Ähnlichkeiten zeigen gleiche Autoren an, wohingegen semantische Ähnlichkeiten ein mögliches Original entlarven. Im Idealfall würde das verdächtige Dokument mit allen anderen verfügbaren Dokument auf diese Weise verglichen, in der Praxis zwingt der nötige Aufwand aber zur Einschränkung auf wenige Kandidaten. Deshalb müssen diese Kandidaten mit Bedacht gewählt werden, um die Chance auf einen Treffer zu maximieren, sofern es etwas zu treffen gibt.

Mit Hilfe maßgeschneiderter Technologien ist es möglich, den Umfang solcher Untersuchungen erheblich zu steigern und sie zu beschleunigen. In [SMP07] haben wir zu diesem Zweck einen allgemeinen Retrievalprozess vorgeschlagen (siehe Abbildung 2), der die beiden oben diskutierten Schritte (Kandidatenretrieval und detaillierter Vergleich) um einen dritten ergänzt. Die wissensbasierte Nachverarbeitung dient dazu, falsch positive Erkennungen zu vermeiden, korrekte Zitate zu erkennen und Textmodifikationen, die durch den Autor des verdächtigen Dokuments eventuell gemacht wurden, zu visualisieren. All das soll die Bearbeitung solcher Fälle so einfach wie möglich gestalten.

Dieser Erkennungsprozess funktioniert ähnlich für alle in Abbildung 1 gezeigten Formen der Textwiederverwendung, aber es gibt keine allumfassende Lösung. Daher konzentrieren wir uns hier auf Zitate, Textbausteine und Übersetzungen. Ein weiterer Schwerpunkt ist die Evaluierung von Implementierungen dieses Prozesses und die hierfür nötigen Werkzeuge.

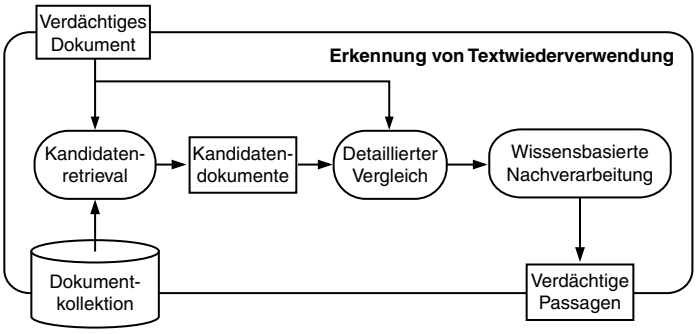

Abbildung 2: Retrievalprozess zur Erkennung von Textwiederverwendungen.

2.1 Fingerprinting zur Erkennung Fast Identischer Texte

Wörtliche Zitate und Textbausteine haben gemeinsam, dass der wiederverwendete Text sehr ähnlich zum jeweiligen Original ist, es aber dennoch Unterschiede geben kann. Beispielsweise werden so wiederverwendete Texte üblicherweise umformatiert, in Zitaten Kommentare eingefügt oder Auslassungen gemacht, in Textbausteinen variable Teile angepasst, und wenn die Wiederverwendung mit der Absicht zu plagiieren geschieht, kleine Modifikationen am Text vorgenommen, um diese Tatsache zu verschleiern. In der Literatur werden diese Formen der Textwiederverwendung daher auch mit dem Begriff „fast identische Texte" umschrieben. Algorithmen zur Erkennung dieser Formen der Wiederverwendung müssen daher robust gegenüber solchen Unterschieden sein.

Eine Klasse von Verfahren, die diese Eigenschaft mitbringt und gleichzeitig sublineare Retrievalzeit ermöglicht, heißt Fingerprinting. Fingerprinting basiert auf Hashing und berechnet für alle Dokumente einer Kollektion einen Fingerprint bestehend aus einer kleinen Zahl von Hashwerten. Anders als mit traditionellen Hashfunktionen werden die Hashwerte so kodiert, dass ähnliche Texte denselben Hashwert erhalten. Im Rahmen unserer Forschung haben wir erstmals das gemeinsame Schema herausgearbeitet, nach dem alle Fingerprintingverfahren arbeiten (siehe Abbildung 3). Kern dieser Verfahren ist die Einbettung hochdimensionaler Dokumentrepräsentationen in niedrigdimensionale Räume. Die anschließende Berechnung von Hashwerten ist gleichzusetzen mit einer Raumpartitionierung, die ähnlichen Dokumenten gleiche Raumabschnitte zuordnet. Zur Dimensionsreduktion in Linearzeit der Dokumentlänge werden Projektion und Einbettung eingesetzt.

Wir haben fünf Fingerprintingverfahren evaluiert und festgestellt, dass das projektionsbasierte Supershingling am besten abschneidet. Ein Problem bei der Evaluierung ist das Fehlen eines Referenzkorpus. Wir schlagen hierfür die Revisionshistorien von Wikipedia-Artikeln vor, die zahlreiche sehr ähnliche Texte aufweisen. Ein überraschendes Ergebnis ist die Tatsache, dass die niedrigdimensionalen Vektoren, die das Fuzzy-Fingerprinting-Verfahren durch Einbettung erzeugt, in Standardretrievalexperimenten ähnlich gut abschneiden wie hochdimensionale Vektorraummodelle: Unabhängig vom Fingerprinting erlaubt dieses Verfahren also die Erzeugung sehr kompakter Dokumentrepräsentationen.

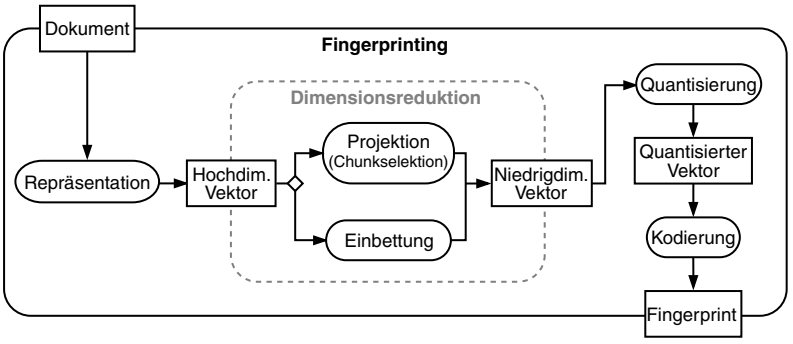

Abbildung 3: Prozess zur Fingerprinterzeugung, der allen Fingerprintingverfahren zugrunde liegt.

2.2 Erkennung von Sprachübergreifender Textwiederverwendung

Übersetzungen sind eine große Herausforderung für die automatische Erkennung von Textwiederverwendung. Das gilt insbesondere für den detaillierten Vergleich eines verdächtigen- mit einem Kandidatendokument des in Abbildung 2 dargestellten Erkennungsprozesses. Anders als innerhalb einer Sprache, kann man sich hier nicht auf syntaktische Überlappungen verlassen. Die Semantik eines Textes zu modellieren und sie automatisch von einer in eine andere Sprache zu überführen, erfordert die Zusammenstellung von Übersetzungswörterbüchern oder parallelen Korpora von Übersetzungen, um damit einen maschinellen Übersetzer zu trainieren. Solche Ressourcen sind schwer zu beschaffen und maschinelles Übersetzen für sich ist ein noch ungelöstes Forschungsproblem.

Wir schlagen das Modell CL-ESA zum sprachübergreifenden Textvergleich vor. Es kommt ohne maschinelle Übersetzung aus und beruht einzig auf *vergleichbaren* Korpora. Das sind Sammlungen von Dokumenten, so dass zu einem Thema in zwei oder mehr Sprachen ein Dokument vorliegt. Die Dokumente können unabhängig voneinander entstanden sein, was ihre Beschaffung bedeutend erleichtert. Die Wikipedia ist zum Beispiel ein vergleichbares Korpus. Mit Hilfe des Modells können zwei verschiedensprachige Dokumente wie folgt verglichen werden: Jedes Dokument wird zunächst mit den Dokumenten des Korpus verglichen, die in seiner Sprache vorliegen, und die Ähnlichkeitswerte aufgezeichnet. Wenn die Ähnlichkeitswerte des einen Dokuments mit denen des anderen übereinstimmen, so sind sich beide sehr ähnlich. Der Grad der Übereinstimmung über alle vergleichbaren Dokumente des Korpus erlaubt eine stufenlose sprachübergreifende Ähnlichkeitsmessung.

CL-ESA wurde mit zwei herkömmlichen Modellen auf Paarungen der Sprachen Englisch, Deutsch, Spanisch, Französisch, Niederländisch und Polnisch verglichen. Mehr als 100 Mio. Vergleiche wurden berechnet und CL-ESA hat sich dabei als konkurrenzfähig erwiesen. Abbildung 4 zeigt das Verhalten von CL-ESA abhängig von seiner Dimensionalität (Zahl vergleichbarer Dokumente). Beim Ranking vergleichbarer Dokumente kann CL-ESA nahezu perfekten Recall erreichen. Außerdem können auch mehrere syntaktisch unabhängige Sprachen gleichzeitig repräsentiert und untereinander verglichen werden.

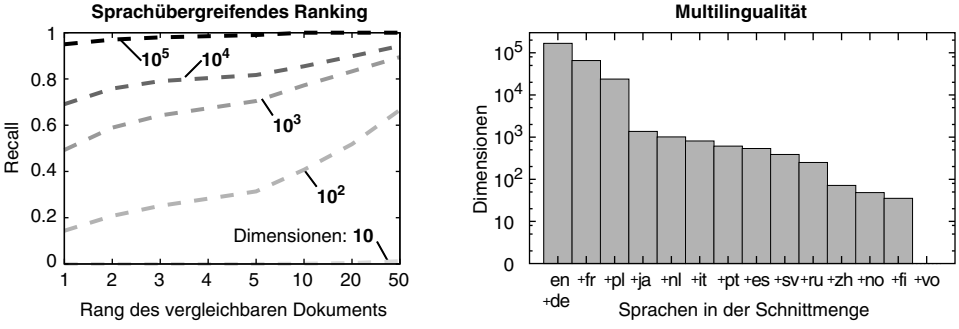

Abbildung 4: Links: CL-ESAs dimensionsabhängige Effektivität im sprachübergreifenden Ranking; Rechts: CL-ESAs Dimensionalität, je mehr Sprachen der Wikipedia gleichzeitig verwendet werden.

2.3 Evaluierung von Verfahren zur Erkennung von Plagiaten

Das Fehlen einer standardisierten Evaluierungsumgebung ist ein schwerwiegendes Problem in empirischer Forschung, da Ergebnisse nicht über Papiere hinweg verglichen oder reproduziert werden können. Wir haben die verfügbare Literatur (205 Papiere) zur Erkennung von Textwiederverwendung und zur Plagiatserkennung in dieser Hinsicht untersucht und festgestellt dass 46% keine Vergleichsverfahren heranziehen, 80% kein freies Korpus verwenden, 77% mit weniger als 1000 Dokumenten evaluieren und insgesamt wenig aussagekräftige Erfolgsmaße eingesetzt werden.

Daher haben wir eine von Grund auf neue Evaluierungsumgebung erschaffen, die aus großen Korpora von Plagiaten und neu entwickelten Erfolgsmaßen besteht. Da reale Plagiate schwer zu bekommen sind,[1] wurden Plagiate sowohl künstlich als auch manuell erzeugt. Das Korpus wurde in drei Versionen erstellt,[2] die jede mehr als 25 000 Dokumente mit jeweils zwischen 60- und 90 000 künstlich generierten Plagiaten enthalten. Es wurden eine Reihe von Parametern variiert und Heuristiken eingesetzt, die das Vorgehen eines Plagiators nachbilden und Textmodifikationen vornehmen, die die automatische Erkennung erschweren. Es wurden auch Übersetzungsplagiate von deutschen und spanischen Texten ins Englische generiert. Zusätzlich wurden erstmals mit Hilfe von Amazons Mechanical Turk über 4000 Plagiatsfälle von mehr als 900 Teilnehmern manuell erzeugt.

Weiterhin haben wir vier Erfolgsmaße für Verfahren zur Plagiatserkennung erforscht und entwickelt, die bisher unberücksichtigte Erfolgsaspekte messen: anstatt auf Dokumentebene messen sie auf Passagenebene und berücksichtigen sowohl das verdächtige als auch das Originaldokument sowie die Eindeutigkeit der Erkennung. Gemessen werden Precision, Recall und Granularität der Erkennung plagiierter Passagen. Letzteres ist die durchschnittliche Häufigkeit mit der ein- und derselbe Fall erkannt wird. Das Maß Plagdet kombiniert diese drei, um die Bildung einer Rangfolge von Verfahren zu ermöglichen.

Sowohl das Korpus als auch die Maße wurden erfolgreich im Rahmen dreier internationaler Wettbewerbe zur Plagiatserkennung eingesetzt.[3] Insgesamt sind 32 Forschergruppen aus aller Welt mit ihren Algorithmen in den Disziplinen externe und intrinsische Plagiatserkennung angetreten. Externe Erkennung meint die Suche nach Originalen für ein gegebenes verdächtiges Dokument, intrinsische Erkennung die oben erwähnte Möglichkeit, Plagiate anhand von Schreibstiländerungen im verdächtigen Dokument zu identifizieren. Abbildung 5 zeigt die in den Wettbewerben erzielten Ergebnisse. Keiner der Algorithmen hat alle in den Korpora versteckten Plagiate erkannt und nur wenige erreichen unter allen Maßen gute Bewertungen. Precision erscheint weniger schwer zu erreichen als Recall. Die Werte sind über die Jahre hinweg nicht direkt vergleichbar, da die Korpora zunehmend schwerer konfiguriert wurden. In 2010 wurden externe und intrinsische Plagiatserkennung als gemeinsame Aufgabe abgehalten. Viele neue Ideen wurden im Rahmen der Wettbewerbe getestet und eine deutliche Entwicklung war zu beobachten. Dennoch stecken Algorithmen zur Erkennung von Textwiederverwendung immernoch in den Kinderschuhen.

[1] Es gibt außerdem rechtliche und ethische Probleme, reale Plagiate als Teil von Korpora zu veröffentlichen. Die kürzlich bekannt gewordenen Fälle unter deutschen Politikern stellen keinen repräsentativen Ausschnitt dar.

[2] Das „PAN Plagiarism Corpus" ist frei verfügbar unter: `http://www.webis.de/research/corpora`

[3] Die Wettbewerbe haben im Rahmen unserer Workshopreihe PAN stattgefunden: `http://pan.webis.de`

Externe Plagiatserkennung

	gro	kas	bas	pal	zec	sch	per	val	mal	all
PAN 2009	0.70	0.61	0.60	0.30	0.19	0.14	0.06	0.03	0.02	0.01
	0.74	0.56	0.67	0.67	0.61	0.75	0.66	0.01	0.03	0.37
	0.66	0.70	0.63	0.44	0.37	0.53	0.10	0.46	0.60	0.01
	1.00	1.02	1.11	2.33	4.44	19.43	5.40	1.01	6.78	2.83

	kas	zou	muh	gro	obe	rod	per	pal	sob	got	mic	cos	naw	gup	van	sua	alz	ift
PAN 2010	0.80	0.71	0.69	0.62	0.61	0.59	0.52	0.51	0.44	0.26	0.22	0.21	0.21	0.20	0.14	0.06	0.02	0.00
	0.94	0.91	0.84	0.91	0.85	0.85	0.73	0.78	0.96	0.51	0.93	0.18	0.40	0.50	0.91	0.13	0.35	0.60
	0.69	0.63	0.71	0.48	0.48	0.45	0.41	0.39	0.29	0.32	0.24	0.30	0.17	0.14	0.26	0.07	0.05	0.00
	1.00	1.07	1.15	1.02	1.01	1.00	1.00	1.02	1.01	1.87	2.23	1.07	1.21	1.15	6.78	2.24	17.31	8.68

	grm	gro	obe	coo	rod	rao	pal	naw	gho
PAN 2011	0.56	0.42	0.35	0.25	0.23	0.20	0.19	0.08	0.00
	0.94	0.81	0.91	0.71	0.85	0.45	0.44	0.28	0.01
	0.40	0.34	0.23	0.15	0.16	0.16	0.14	0.09	0.00
	1.00	1.22	1.06	1.01	1.23	1.29	1.17	2.18	2.00

Intrinsische Plagiatserkennung

	sta	hag	zec	sea
Plagdet	0.25	0.20	0.18	0.12
Precision	0.23	0.11	0.20	0.10
Recall	0.46	0.94	0.27	0.56
Granularität	1.38	1.00	1.45	1.70

	obe	sta	kes	aki	rao
Plagdet	0.33	0.19	0.17	0.08	0.07
Precision	0.31	0.14	0.11	0.07	0.08
Recall	0.34	0.41	0.43	0.13	0.11
Granularität	1.00	1.21	1.03	1.05	1.48

Abbildung 5: Ergebnisse dreier Wettbewerbe zur Erkennung von Plagiaten (PAN 2009-2011). Pro Tabelle entspricht jede Spalte einem Teilnehmer. Die Spalten sind nach dem erzielten Plagdet-Wert sortiert. Die Zellenschattierung visualisiert die erzielten Erfolge: Je dunkler, desto besser.

3 Retrievalaufgaben Mittels Sprachwiederverwendung Lösen

Textwiederverwendung bezeichnet die Wiederverwendung einzelner Texte, wohingegen die Wiederverwendung großer Mengen von Texten als Sprachwiederverwendung bezeichnet wird. Ein Übersetzer verwendet Texte beispielsweise einzeln wieder, da sie zumeist unabhängig von anderen übersetzt werden. Ein Linguist hingegen verwendet viele Texte wieder, um anhand der Vorkommen eines Wortes all seine Bedeutungen zu erfassen. Die Wiederverwendung eines Textes geschieht also linear und unabhängig von anderen Texten. Die Wiederverwendung von Sprache dagegen geschieht durch das Ausnutzen bestimmter Texteigenschaften, die viele Texte miteinander teilen, um anhand ihrer Ausprägungen eine Aufgabe zu erfüllen. Es können jedoch nicht bloß linguistische Aufgaben durch die (automatische) Wiederverwendung von Sprache gelöst werden, sondern ungezählte andere: Prominente Beispiele im Information Retrieval sind Wikipedia und Web-n-Gramme. Die Wikipedia ist inzwischen eine weit verbreitete Informationsquelle, nicht nur für Menschen, sondern auch zur Informierung wissensbasierter Modelle und Algorithmen (ein umfassender Überblick ist in [MMLW09] zu finden). Ähnlich erfolgreich werden Web-n-Gramme (Wortfolgen der Länge n und ihre Häufigkeit im Web) eingesetzt. Die Zahl der Aufgaben, die durch Sprachwiederverwendung ganz oder teilweise gelöst werden können ist nicht ersichtlich, da Texte im Web in unzähligen Relationen mit anderen Dingen stehen. Im Rahmen unserer Forschung haben wir zwei neue Ansätze erforscht, um mit Hilfe von Sprachwiederverwendung Aufgaben des Information Retrieval zu lösen: Es handelt sich zum Einen um ein Modell zum inhaltlichen Vergleich von Texten und Objekten aller Mediengattungen und zum Anderen um Algorithmen zum Einfügen intendierter Anführungszeichen in Websuchanfragen. Weiterhin haben wir einen Webdienst entwickelt, der es erlaubt, alle Texte im Web als Formulierungshilfe zu verwenden.

3.1 Ein Modell zum Medienübergreifenden Vergleich von Objekten

Das Web besteht nicht bloß aus Texten, sondern aus Medienobjekten aller Art und selbstverständlich werden auch diese von den Nutzern des Webs gesucht. Suchmaschinen stehen daher vor dem Problem, textuelle Suchanfragen mit Objekten anderer Mediengattungen zu vergleichen. Traditionelle Ansätze hierfür basieren darauf, Korpora bestehend aus Multimediaobjekten auszuwerten, die von Hand mit Metainformationen über ihren Inhalt ausgezeichnet wurden. Mit diesen Daten werden maschinelle Lernverfahren trainiert, die eine Abbildung der Inhaltsangaben auf medienspezifische Charakteristiken erlernen sollen, um so Suchanfragen beantworten zu können. Korpora dieser Machart sind gegenwärtig nur in kleinem Rahmen verfügbar, was die Forschung an dieser Aufgabe behindert.

Wir haben Webkommentare zu Multimediaobjekten als eine weitgehend vernachlässigte Quelle von Informationen identifiziert. Darauf aufbauend schlagen wir ein neues Modell zum medienübergreifenden Vergleich von Objekten aller Mediengattungen vor, das auf Webkommentaren beruht. Das Modell verwendet alle Kommentare zu einem Objekt als Ersatz für eine inhaltliche Beschreibung wieder, um so mit textbasierten Standardmodellen die Kommentare zweier Objekte insgesamt miteinander zu vergleichen. Wird eine inhaltliche Übereinstimmung der Kommentare gemessen, so liegt mit hoher Wahrscheinlichkeit auch eine inhaltliche Übereinstimmung der kommentierten Objekte vor.

Das Modell wurde evaluiert, indem 6000 YouTube-Videos auf diese Weise mit rund 18 000 Artikeln der Nachrichtenseite Slashdot verglichen wurden. Aus den sich so ergebenen Paarungen wurden die 100 manuell ausgewertet, deren gemessene Ähnlichkeit am höchsten war und festgestellt, dass 91 mindestens ein verwandtes und 36 dasselbe Thema aufwiesen (siehe Tabelle 1). Außerdem wurden stichprobenartig weitere 150 Paarungen mit geringeren Ähnlichkeiten ausgewertet und festgestellt, dass thematische Übereinstimmungen ab einer gemessenen Kommentarähnlichkeit von 0.4 im Vektorraummodell gehäuft auftreten. Diese Ergebnisse lassen den Schluss zu, dass die Annahme, auf der unser Modell fußt, zutrifft und dass es zum medienübergreifenden Vergleich von Objekten eingesetzt werden kann. Die einzige Einschränkung dabei ist, dass mindestens rund 100 Kommentare vorhanden sein müssen.

Tabelle 1: Themenvergleich der 100 als am ähnlichsten erkannten Video-Artikel Paare.

Themenvergleich	**Anteil**	**Ähnlichkeit**				**∅ Zahl der Kommentare**	
		min.	∅	max.	σ	Slashdot	YouTube
gleich	36%	0.71	0.78	0.91	0.06	53	927
verwandt	55%	0.71	0.76	0.91	0.04	81	683
ungleich	9%	0.72	0.78	0.87	0.05	104	872
Σ	100%	0.71	0.77	0.91	0.05	74	790

3.2 Algorithmen zur Segmentierung von Suchanfragen

Die vorrangige Art von Suchanfragen an Websuchmaschinen sind Schlüsselwortanfragen. Obwohl die meisten Suchmaschinen auch fortgeschrittenere Anfrageoperatoren und -facetten anbieten, mit denen ein Nutzer das Gesuchte klarer umschreiben kann, werden diese kaum benutzt: Nur 1.12% der Suchanfragen enthalten solche Operatoren [WM07]. Eine der Möglichkeiten besteht darin, Anführungszeichen in Anfragen einzufügen, um so Phrasen als unteilbar zu markieren. Die Suchmaschine kann mit dieser Information die Precision der Suchergebnisse erhöhen, da Dokumente, die die Phrasen nicht enthalten, verworfen werden können. Da die überwältigende Mehrheit der Nutzer einer Suchmaschine, diese Option nicht nutzt, wird ein erhebliches Potential verschenkt.

Wir stellen einen neuen Algorithmus vor, der automatisch Anfragen segmentiert (also Anführungszeichen an geeignete Stellen einer Suchanfrage einfügt). Der Algorithmus basiert auf der Annahme, dass Phrasen, die hinreichend häufig im Web vorkommen, wichtige Konzepte sind, die es lohnt in Anführungszeichen zu setzen. Es werden zunächst alle Segmentierungen der Anfrage aufgezählt und dann jede Segmentierung gewichtet: Das Gewicht errechnet sich aus der Länge aller enthaltenen Phrasen und ihrer Häufigkeit im Web. Damit längere Phrasen eine Chance gegenüber den tendenziell häufiger vorkommenden, kürzeren haben, werden die Gewichte geeignet normalisiert. Am Ende wird die Segmentierung gewählt, deren Gewicht am höchsten ist. Um die Häufigkeit einer Phrase im Web effizient zu ermitteln, verwenden wir das Google-n-Gramm-Korpus, das die Häufigkeit aller im Jahr 2006 im Web vorkommenden Phrasen der Länge $n \leq 5$ Wörter enthält [BF06].

In einer groß angelegten Evaluierung haben wir unseren Algorithmus mit acht weiteren aus der Literatur verglichen. Zu diesem Zweck haben wir ein bislang häufig verwendetes Korpus zur Anfragesegmentierung verwendet, das aus 500 Anfragen besteht. Da dieses Korpus einige Konstruktionsschwächen aufweist und nicht repräsentativ ist, haben wir ein neues Korpus mit mehr als 50 000 Anfragen erstellt, das die Längen- und Häufigkeitsverteilung echter Anfragelogdateien repräsentiert. Zu jeder Anfrage wurden via Amazons Mechanical Turk zehn Personen befragt, an welchen Stellen sie Anführungszeichen einsetzen würden. Auf beiden Korpora übertrifft unser Algorithmus die anderen: Er fügt am ehesten Anführungszeichen dort ein, wo auch Menschen es tun würden, und ist gleichzeitig bedeutend einfacher zu realisieren als die bisherigen Verfahren.

3.3 Ein Werkzeug zur Schreibunterstützung

Die meiste Zeit beim Schreiben verbringt man damit, herauszufinden, wie man etwas schreiben möchte, nicht was. Gute Formulierungen für einen Sachverhalt zu finden, entscheidet darüber, wie gut die Zielgruppe eines Textes ihn versteht. Gerade deutsche Wissenschaftler stehen diesbezüglich vor dem Problem, dass der Diskurs vieler Disziplinen in Englisch stattfindet. Die meisten Deutschen verfügen aber nicht über das Vokabular oder das Sprachgefühl eines Englisch-Muttersprachlers. Die Suche nach Worten und Formulierungen wurde allerdings bis jetzt nicht hinreichend unterstützt.

Wir haben NETSPEAK entwickelt, eine Suchmaschine für geläufige englische Formulierungen.[4] Netspeak indiziert das Web in Form von n-Grammen und ermöglicht eine Wildcardsuche darauf. Suchanfragen bestehen aus kurzen Formulierungen, in die der Nutzer Wildcards dort eingefügt, wo Unsicherheit über die üblicherweise verwendeten Wörter besteht. Die zur Anfrage passenden n-Gramme werden gesucht und die Ergebnisse nach ihrer Häufigkeit im Web sortiert. Auf diese Weise können geläufige von ungebräuchlichen Formulierungen unterschieden werden. Die NETSPEAK zu Grunde liegende Hypothese ist die, dass neben der Korrektheit eines Textes auch die Verwendung geläufiger Formulierungen wichtig ist. Das bringt den Vorteil leichter Verständlichkeit mit sich und schränkt das Fehlerpotenzial ein wenig ein, gerade beim Schreiben in einer fremden Sprache.

4 Ausblick

Wie sähe die Welt aus, wenn alle Textwiederverwendungen im Web offen zutage lägen? Plagiarismus, der „böse Zwilling" der Textwiederverwendung, wäre sinnlos. Doch darüber hinaus würde ein Netzwerk zwischen Webdokumenten ersichtlich, das, anders als das Hyperlinknetzwerk, den Einfluss eines Textes auf andere sichtbar machen würde. Ein solches Netzwerk könnte als weiteres Relevanzsignal in der Websuche dienen, aber auch dazu, Reputation und Honorare zum Urheber eines Textes weiterzuleiten. Im gesamten Web wird dies vermutlich kaum realisierbar sein, in spezifischen Genres, wie der Wissenschaft, ist das aber durchaus denkbar.

Literatur

[BF06] Thorsten Brants und Alex Franz. Web 1T 5-gram Version 1. Linguistic Data Consortium LDC2006T13, 2006.

[MMLW09] Olena Medelyan, David Milne, Catherine Legg und Ian H. Witten. Mining Meaning from Wikipedia. *Int. J. Hum.-Comput. Stud.*, 67:716–754, 2009.

[Pot11] Martin Potthast. *Technologies for Reusing Text from the Web*. Dissertation, Fakultät Medien, Bauhaus-Universität Weimar, 2011.

[SMP07] Benno Stein, Sven Meyer zu Eißen und Martin Potthast. Strategies for Retrieving Plagiarized Documents. In *Proceedings of SIGIR 2007*.

[WM07] Ryen W. White und Dan Morris. Investigating the Querying and Browsing Behavior of Advanced Search Engine Users. In *Proceedings of SIGIR 2007*

Martin Potthast wurde im April 1981 in Steinheim geboren und vollendete seine schulische Laufbahn im Jahr 2000 am Gymnasium St. Xaver in Bad Driburg. Nach dem Zivildienst nahm er 2001 das Studium der Informtik an der Universität Paderborn auf und erhielt Anfang 2005 den Bachelortitel. Mitte 2006 vollendete er das Studium als Diplom-Infomatiker. Seitdem hat er am Lehrstuhl für Content Managment und Web-Technologien der Bauhaus-Universität Weimar promoviert und seine Dissertation Ende 2011 verteidigt.

[4] NETSPEAK ist frei verfügbar unter `http://www.netspeak.org`.