Bauhaus-Universität Weimar
Fakultät Medien
Studiengang Mediensysteme

N-Gramm-basierte Retrieval-Modelle in der Autorschaftsbestimmung

Bachelorarbeit

Michael Blersch

1. Gutachter: Prof. Benno Stein
Betreuer: Dipl.-Inf. Martin Potthast

Datum der Abgabe: 1. April 2010

Eidesstattliche Erklärung

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ahnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen worden ist. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Weimar, den 1. April 2010

Michael Blersch

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Diese Arbeit befasst sich mit dem Problem der Autorschaftsbestimmung von Texten. Dabei soll für einen anonymen Text festgestellt werden, wer der Verfasser ist. Verfahren die dieses Problem lösen sind zentraler Bestandteil in vielen Bereichen der Informatik, wie z. B. der Plagiaterkennung oder der computerlinguistischen Textforensik.

Die Plagiaterkennung ist mit der Autorschaftsbestimmung eng verwandt und beschäftigt sich mit der Erkennung von Abschnitten eines Textes, die von anderen Autoren stammen. Ein Abschnitt ist plagiiert, wenn keine Referenzierung des Abschnitts bezüglich der Quelle stattgefunden hat. Plagiatsvergehen werden immer häufiger begangen. Dies liegt an der zunehmenden Verbreitung von Texten im Internet und der Einfachheit, Abschnitte aus den Texten zu kopieren, um diese in der eigenen Arbeiten zu verwenden. Ein Verfahren zur Erkennung solcher Abschnitte ist beispielsweise die intrinsische Plagiaterkennung [SM07].

Im Bereich der Computerforenstik werden Informationen von Computersystemen, Netzwerken und Datenspeichern gesammelt und analysiert. Die Analyse konzentriert sich dabei auf das Auffinden von gerichtlich verwertbaren Beweisen von Computerverbrechen und die Rekonstruktion solcher Vergehen. Ein Anwendungsfall in der Textforensik ist die Analyse von Drohbriefen, die beispielweise in Form von E-Mails verfasst wurden [DVACM01], oder die Analyse von Quelltexten zu schädlichen Programmen [FSGK06]. Für beide Vergehen werden Beweise gesammelt, um die Autorschaft nachzuweisen.

Die Autorschaftsbestimmung gliedert sich in zwei Problemklassen, nämlich in das Problem der sog. Attribution und das Verifikationsproblem. Erstere definiert das Problem hinsichtlich der Zuordnung. Dabei sind Texte von verschiedenen Autoren vorhanden und es soll nun festgestellt werden, welcher dieser Autoren der Verfasser eines gegebenen anonymen Textes ist. Das Verifikationsproblem unterscheidet sich dadurch, dass von einem Autor Texte vorhanden sind und nun entschieden werden soll, ob weitere Texte von diesem Autor verfasst wurden oder nicht [SKS07].

Um Texte bezüglich der Autorschaft unterscheiden zu können, werden Verfahren verwendet, die auf einer vergleichenden Schreibstilanalyse basieren. Der Schreibstil eines Autors wird durch statistisch quantifizierbare Merkmale, die unmittelbar aus dem Text bestimmt werden, abgebildet. Die Merkmale werden als Stilmerkmale bezeichnet, wobei die Definition und Auswahl der Stilmerkmale maßgebend für die Unterscheidung von Autoren ist. Die Schwierigkeit, Schreibstile durch Stilmerkmale zu unterscheiden, liegt darin, ungleiche Autoren möglichst stark abzugrenzen und gleichzeitig Variationen in Schreibstilen von gleichen Autoren zu tolerieren.

Im Rahmen dieser Arbeit werden Verfahren zur Lösung des Attributionsproblems in der Autorschaftsbestimmung vorgestellt und untersucht. Diese Verfahren bezeichnet man als Retrieval-Modelle. Vorgestellt werden zwei existierende Modelle und ein neu entwickeltes Modell. Dabei soll untersucht werden, wie gut das neue Verfahren Autoren erkennen kann und evtl. eine Verbesserung zu der existierenden Methode darstellt. Für die Modelle ist entscheidend, in welcher Form der Schreibstil durch Stilmerkmale und durch eine modellabhängige Repräsentation abgebildet wird. Für die Entscheidung, welches der beiden Modelle die Schreibstile besser repräsentiert, werden Techniken aus dem maschinellen Lernen eingesetzt. Das zweite existierende Modell beschreibt einen Lösungsweg, mit dem es möglich ist, die Autorschaft eines Textes zu bestimmen, für den tausende an potentielle Autoren in Frage kommen. Das Modell wird in dieser Arbeit modifiziert und hinsichtlich der Modifikation untersucht.

2 Retrieval-Modelle zur Autorschaftsbestimmung

Information Retrieval (IR) bezeichnet ein interdisziplinäres Forschungsgebiet und verknüpft Bereiche der Informatik, Computerlinguistik und Informationswissenschaft miteinander. Das Ziel ist dabei, die Extraktion und Bereitstellung von Informationen aus unstrukturierten Daten, um den Informationsbedarf eines Nutzers zu befriedigen. Die Autorschaftsbestimmung ist dabei in das Forschungsgebiet, der automatischen Textkategorisierung im IR einzuordnen. Dazu zählen unter anderem die Spracherkennung und die inhaltliche Analyse von Texten. Forschungen im Bereich des inhaltsbasierten IR beschäftigen sich mit der Erfassung und Abgrenzung inhaltlicher Beschreibungen von Texten. Retrieval-Strategien aus diesem Bereich werden für die Autorschaftsbestimmung übertragen und auf die Problemstellung, Autoren aufgrund von Schreibstilen zu unterscheiden, angepasst. Um Schreibstile aus Texten zu ermitteln, sind sprachverarbeitende Verfahren und Werkzeuge aus der Computerlingusik von zentraler Bedeutung.

Im IR wird ein Dokument mit d bezeichnet und der von einem Benutzer eines IR-Systems verspürte Informationsbedarf mit q. Mengen von Dokumenten werden mit D, Mengen von Bedarfen mit Q bezeichnet. Ein Dokument ist für einen Menschen ein verstehbarer Text, die Repräsentation $\mathbf{d}$ eines Dokuments, ist für die Verarbeitung durch einen Computer aufbereitet. Der Informationsbedarf wird ebenfalls zu einer formalen Anfrage $\mathbf{q}$ aufbereitet. Die Art der Repräsentation orientiert sich dabei gänzlich an der zu lösenden Aufgabe, dem sog. Retrieval-Task, und dem zu bedienenden Informationsbedarf. Die Relevanz-Funktion $\rho_R(\mathbf{q}, \mathbf{d})$ quantifiziert die Relevanz von einer Anfragerepräsentation zu einer Dokumentrepräsentation eines Retrieval-Modells R. Der abstrakte Begriff des Retrieval-Modells bezeichnet eine generalisierte Zusammenfassung einer orthogonalen Repräsentationart und einer Relevanzfunktion. R ist aus dem Tupel $\langle \mathbf{Q}, \mathbf{D}, \rho_R \rangle$ definiert und beinhaltet Mengen von Anfragerepräsentationen $\mathbf{Q}$ und Mengen von Dokumentrepräsentationen $\mathbf{D}$ [Ste08c].

Ein Retrieval-Modell zur Bestimmung der Autorenschaft vergleicht eine Schreibstilrepräsentation $\mathbf{x}$ aus einem Dokument x eines unbekannten Autors, mit repräsentierten Schreibstilen aus Dokumenten von bekannten, nicht umstrittenen Autoren. Die Ergebnisse werden als Retrieval-Werte bezeichnet [SP07]. Sie ermöglichen das Ranken der Dokumente. Das Dokument aus dem der ähnlichste Schreibstil berechnet wird, nimmt dabei den höchsten Retrieval-Wert an. Der Autor von diesem Dokument wird dem Dokument x, für das die Autorschaft bestimmt werden soll, zugewiesen.

2.1 Retrieval-Evaluierung

Zur Bewertung eines Retrieval-Modells muss zunächst die Bedeutung des Begriffs der Relevanz geklärt werden. Definition nach Salton und McGill aus [SM]: *„Relevance is the correspondence in context between an information requirement statement (a query) and an article (a document), that is, the extent to which the article covers the material that is appropriate to the requirement statement.“* Er beschreibt damit, dass Relevanz dadurch entsteht, wenn die Information, die aus der Anfrage hervorgeht, mit der Information, die aus dem Dokument hervorgeht, in einem gemeinsamen Kontext korrespondiert und eine Überlappung der Informationen, bezüglich der Anfrage stattfindet. Die Frage der Relevanz wird somit binär beantwortet, wobei 1 relevant und 0 nicht relevant bedeutet. Daraus folgt die relationale Abbildung r der Dokumente D bekannter Autoren mit den Anfragen hinsichtlich der Autorschaft durch die Dokumente X von unbekannten Autoren auf die Werte 0 und 1.

$$r : D \times X \rightarrow \{0, 1\} \tag{2.1}$$

Die Relation bildet den „wahren“ Zusammenhang ab. Als Beispiel nehmen wir an wir wissen, dass Dokument d_1 und Dokument d_2 vom selben Autor geschrieben wurden. Somit bildet die Relation den Zusammenhang zwischen d_1 und d_2 auf den booleschen Wert 1 ab. Die Relevanz-Funktion ρ_R versucht r nachzubilden und stellt damit eine Annäherung an den „wahren“ Zusammenhang zwischen den realen Dokumenten dar. Um zu bewerten, wie gut ein Retrieval-Modell Relevanz erkennt, werden quantitative Qualitätsmaße verwendet. Für die folgenden Qualitätsmaße gilt:

Sei A die Menge der relevanten Dokumente und B die Menge der vom Modell gefundenen Dokumente, wobei U das Universum aller Dokumente im Modellversuch abbildet, dann definiert

Precision ein Maß zur Beurteilung der Genauigkeit des Modells und liefert den Anteil der relevanten Antworten in der Ergebnismenge und

$$P = \frac{|A \cap B|}{|B|}$$

Recall ein Maß zur Beurteilung der Vollsändigkeit der gelieferten Ergebnissmenge, es werden die gefundenen relevanten Antworten angegeben.

$$R = \frac{|A \cap B|}{|A|}$$

Zur Bestimmung der Ergebnismengen werden die vier Fälle betrachtet die bei der Zuordnung möglich sind. Die Unterscheidung erfolgt zum einen nach der Bewertung des Modells in positiv oder negativ, d. h. zwei Dokumente sind laut Modell vom selben Autor oder sie sind es nicht. Zum anderen ist die tatsächliche Übereinstimmung, die bekannt ist, entweder richtig oder falsch .

a Anzahl der richtig zugeordneten aus den als positiv bewerteten Autoren

b Anzahl der fälschlich zugeordneten aus den als positiv bewerteten Autoren

c Anzahl der fälschlich zugeordneten aus den als negativ bewerteten Autoren

d Anzahl der richtig zugeordneten aus den als negativ bewerteten Autoren

Zur Berechnung von Precision und Recall gilt somit $P = \frac{a}{a+c}$, $R = \frac{a}{a+b}$.

Die Precision kann durch die Verwendung eines Schwellenwerts gesteigert werden. Wenn die Zuordnung des Autors aufgrund dieser Insignifikanzschwelle nicht möglich ist, wird dieser als unbekannt eingestuft. Das kann sich positiv auf die Precision auswirken, da die Bewertung nicht der Definition von c entspricht. Dafür sinkt der Recall, indem insgesamt für weniger Autoren eine Entscheidung getroffen wurde.

2.2 Merkmale für den Schreibstil

Der Schreibstil eines Autors definiert sich in der computergestützten Autorschaftsbestimmung durch statistisch quantifizierbare Merkmale, durch die im Idealfall signifikante Unterschiede bei unterschiedlichen Autoren und insignifikante Unterschiede zwischen Dokumenten gleicher Autoren erkennbar sind. Weiterhin sollten die Merkmale durch die Autoren selbst nicht bewusst manipulierbar sein und keine

inhaltlichen Informationen transportieren. Im Laufe der Jahre wurden nach einer Schätzung von [Rud97] über 1.000 Stilmerkmale untersucht, die mehr oder weniger diesen Anforderungen genügen. Abbildung 2.1 zeigt eine mögliche Einteilung der Stilmerkmale nach [Sta09] in sprachliche Ebenen, die sich gleichzeitig komplexitätsbedingt abgrenzen. Unter Komplexität wird der Aufwand bezeichnet, der nötig ist, um die Stilmerkmale aus einem Text zu bestimmen und wie stark die Definition der Stilmerkmale den Bezug zu sprachwissenschaftlichen Theorien herstellt.

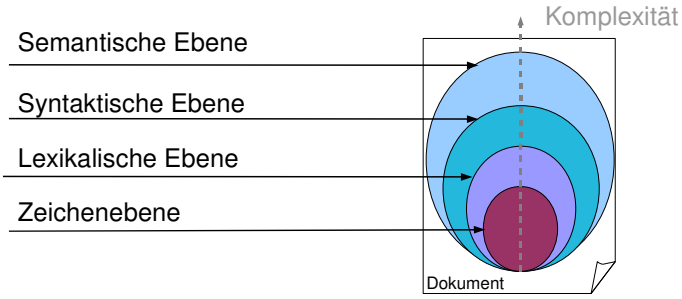


Abbildung 2.1: Einteilung der Stilmerkmale in vier sprachliche Ebenen.

Auf der semantischen Ebene wird ein Text in seiner Gänze, durch sinnmäßig verknüpfte Einheiten betrachtet. Es wird versucht, Zusammenhänge von Worten und Sätzen durch Hilfsstrukturen, wie z.B. Graphen oder vordefinierte Kontexteinheiten, abzubilden, um Autoren aufgrund wiederkehrender semantischer Relationen zu unterscheiden. Stilmerkmale auf syntaktischer Ebene quantifizieren wiederkehrende Zusammensetzungen von Wortarten (Part-of-Speech, POS), die Wortstellung im Satz oder gegebenenfalls Rechtschreibfehler, die ein strukturbedingtes autorspezifisches Muster ergeben. Merkmale auf diesen Ebenen können nur durch komplexe, sprachabhängigen Werkzeuge (POS-tagger, Semantic-Parser, Semantic-Splitter) extrahiert werden und sind von ihnen qualitativ abhängig. In [Sta09] wird außerdem eine anwendungsspeziefische Ebene aufgeführt. Hierzu zählen Merkmale, die aufgrund der vorliegenden Textform (E-Mail, Webseiten, Online-Foren) strukturelle Eigenheiten und Variationen bei sprachlichen Konventionen (Grußfloskel, Emoticons, spezielle Kurzformen) ermitteln. Diese werden hier als syntaktisch charakterisiert. Der Hauptkritikpunkt zur Verwendung dieser Merkmale besteht nach [Sta09] in der Feststellung, dass noch keine ausreichend genauen Werkzeuge zur semantischen Analyse zur Verfügung stehen: die Datenerhebung erfolgt zu ungenau. Syntaktische Merkmale bilden außerdem keine generalisierte Stilerfassung. Die Stilmerkmale sind eher sprachwissenschaftlich motiviert und entfernen sich von einer rein statistischen Vorgehensweise. Deutlich bessere Ergebnisse wurden nur in Kombination mit Stilmerkmalen niedrigerer Ebenen erzielt.

Auf der lexikalischen Ebene definieren sich die Merkmale durch die Zerlegung eines Textes in Wort- und Satzeinheiten. Die Zeichenebene reduziert den Text noch weiter auf alle vorkommenden, einzelnen Einheiten wie Buchstaben, Zahlen und Satzzeichen. Schreibstile werden durch die Häufigkeit dieser Einheiten im Text erfasst. Hierzu zählen außerdem Maße, welche die Wortvielfalt, die Lesbarkeit oder den Lesegrad von Texten auf einen skalaren Wert abbilden. Ein einfaches Verfahren zur Bestimmung der Wortvielfalt ist das Verhältnis der verschiedenen Wörter zur Gesamtanzahl aller Wörter eines Textes zu berechnen [Hol94]. Die Lesbarkeit und der Lesegrad untersuchen den inhaltlichen Anspruch und die Verständlichkeit eines Textes, mit überwiegend indexgebundenen Formeln. Der Lesegrad gibt die Anzahl an besuchten Schuljahren an, die nötig sind, um einen Text zu verstehen. Für die Maße bilden beispielsweise Wort-Satz und Silbenanzahl die Parameter, welche in ein vorgegebenes lineares Gleichungssystem eingesetzt werden und einen Wert ergeben. Dieser Wert referenziert im spezifischen Index je nach Methode die Lesbarkeit oder den Lesegrad [DuB04].

Stilmerkmale aus der lexikalischen Ebene und der Zeichenebene bilden im überwiegenden Teil der Studien das Fundament zur Erfassung stilistischer Gemeinsamkeiten und Unterschiede. Die Auswahl und Repräsentation orientiert sich dabei an einer Vorgehensweise der inhaltsbasierten Textklassifikation. Ein Retrieval-Modell zur Berechnung der inhaltlichen Ähnlichkeiten zwischen Dokumenten ist das Vektorraummodell in Kapitel 2.5. Dazu werden aus einer Dokumentkollektion Indexterme als Vokabular bestimmt, auf deren Basis eine häufigkeitsabhängige, gewichtete Vektorrepräsentation der Texte bezüglich ihres Inhalts erfolgt. Für die Autorschaftsbestimmung bilden in diesem Zusammenhang die häufigsten Wörter (MFW), die direkt aus der Kollektion ermittelt werden, das Vokabular, um den Schreibstiel zu erfassen. Die Problemstellung liegt dabei in der Festlegung der Anzahl dieser Stilmerkmale. In [Bur92], einer früheren Studie, wurde eine geringe Anzahl von 100 MFW verwendet. Durch die Anwendung von Klassifikationsmethoden, die mit einer größeren Anzahl an Merkmalen besser umgehen können, wie zum Beispiel der Stützvektormethode, wurden in [KSBD07] 250 und in [Sta06] die 1.000 häufigsten Wörter verwendet. In [MGL+05] wurde die Anzahl noch weiter erhöht, indem alle Wörter verwendet wurden, die mindestens zweimal in der zu untersuchenden Textkollektion vorkommen. Eine weitere Möglichkeit zur Stilerfassung besteht darin, definierte Wortmengen (Konjuntionen, Präpositionen, Modalverben), die als Funktionsworte (FW) bezeichnet werden als Vokabular zu verwenden. Sie unterscheiden sich bezüglich der MFW somit in der Definition, nach welchen Wortarten die Häufigkeitsverteilung in den Texten erhoben wird und haben die Gemeinsamkeit, dass diese Worte ebenfalls sehr häufig vorkommen und daher als nicht bewusst manipulierbar gelten. Die Anzahl

variiert ebenfalls, in [AC05] wurden beispielsweise 150 Funktionsworte definiert, hingegen benutzten in [ZZ05] eine Menge von 365 FW.

Als besonders geeignet haben sich jedoch Character-n-Gramme herausgestellt. Ein gegebener Text wird als eine Folge von Zeichen betrachtet. Ein Character-n-Gramm besteht aus genau n Zeichen. Das erste besteht aus den ersten n Zeichen des Textes, wobei zur Extraktion des nächsten die Startposition um eine Stelle nach rechts verschoben wird. Die Anzahl an Character-n-Grammen hängt von den vorkommenden Zeichen in der Textkollektion und der Länge n ab. Die maximal möglichen Permutationen ergeben sich aus dem Zeichensatz, der mit n potenziert wird. Bei steigender Länge wird die Reduktion der gefundenen Character-n-Gramme relevanter, da nur wenige Zeichenkombination sehr häufig in allen Texten zu finden sind. In [KSA09] wurden mit einem Selektionsalgorithmus 10.000 Character-n-Gramme bestimmt. Daraus wurden die 1.000 häufigsten für die Stilrepräsentation verwendet. Dabei sind kurze Character-n-Gramme für englische Texte, mit einer Länge von zwei bis drei Zeichen nach [FH96] und [Gri07] sowie vier Zeichen in [KSA09] geeignet, um möglichst wenig inhaltliche Informationen miteinzubeziehen. Die optimale Länge für Character-n-Gramme ist nach [Sta09] sprachabhängig, da beispielsweise in Texten in deutscher oder griechischer Sprache die Wörter allgemein länger sind als im Englischen, wodurch längere Character-n-Gramme wahrscheinlich geeigneter sind. Studien von [Gri07] und [KSA09] erzielten im direkten Vergleich verschiedener lexikalischer Stilmerkmale mit Character-n-Grammen die besten Resultate.

2.3 Termgewichtung

Ein Termgewicht quantifiziert die Relevanz eines Merkmals aus einem Dokument durch einen Wert. Dieser Wert gibt an, wie wichtig dieses Merkmal für die Repräsentation des Dokuments ist. Die in 2.2 formulierten Bedingungen für Stilmerkmale erfordern eine abgrenzende Betrachtung dieser Relevanz, nämlich die Trennung zwischen der Erfassung inhaltlicher oder stilistischer Beschreibungen.

Im inhaltsbasierten IR werden die Merkmale als Indexterme bezeichnet. Die Auswahl an Indextermen soll es ermöglichen, inhaltliche Aspekte der Dokumente zu repräsentieren und eine klare Abgrenzung der Dokumente untereinander gewährleisten. Verknüpfungen zwischen thematisch ähnlichen Dokumenten müssen erkennbar sein. Um solche Merkmale festzulegen wird ein Prozess durchlaufen, den man als Indexierung bezeichent. In der Autorschaftsbestimmung werden dagegen Merkmale wie

Funktionsworte, Character-n-Gramme, Maße für die Lesbarkeit oder die Wortvielfalt verwendet, die durch unterschiedliche Häufigkeiten Schreibstile beschreiben und unterscheidbar machen.

Eines der bekanntesten Gewichtungsmaße für Indexterme ist $tfidf$, das sich in die Bestandteile der Termhäufigkeit tf und der inversen Dokumenthäufigkeit idf aufteilt. Für die Maße bezeichnet v ein Merkmal, das Vokabular V eine Menge an Merkmalen, mit $v_k \in V$ und $k = \{1, \ldots, |V|\}$.

Die Gewichtung $tf(v_k, d)$ misst für einen gegebenes Merkmale v_k die Auftrittshäufigkeit in einem Dokument d. Die Dokumenthäufigkeit $df(v_k, D)$ eines Merkmals ist die Anzahl der Dokumente in einer Dokumentkollektion D, die das Merkmal v_k enthalten. Die beiden Gewichtungen eignen sich für Stilmerkmale. Dies wird am Beispiel von Character-n-Grammen deutlich. Die Auswahl an Character-n-Grammen, kann durch die Berechnung der Dokumentfrequenz erfolgen, indem die häufigsten Character-n-Gramme aus der Dokumentkollektion ermittelt werden. Zur Repräsentation eignet sich dann eine tf-Gewichtung, indem die ausgewählten Merkmale dokumentabhängig gewichtet werden, um die unterschiedliche Häufigkeitsverteilung je Autor aufzuzeigen.

Ein Wort, welches in vielen Dokumenten vorkommt und somit eine hohe Dokumenthäufigkeit aufweist, bildet kein diskriminierendes Merkmal, um Dokumente inhaltlich zu unterscheiden. Dies wird am Beispiel von Funktionsworten deutlich. Für die inhaltliche Beschreibung gehört der überwiegende Teil der Funktionsworte zu der Kategorie Stoppworte. Unter Stoppworte werden Wortarten wie Artikel, Konjunktionen und Präpositionen verstanden, die in der Dokumentkollektion gleichverteilt und sehr häufig auftreten. Ein Selektionsschritt zur Auswahl von Indextermen wird in diesem Zusammenhang als Stoppwortelimination bezeichnet. Aus diesem Sachverhalt ergibt sich, dass die Bedeutung eines Wortes für die inhaltliche Abgrenzung eines Dokuments unter Berücksichtigung aller Dokumente, umgekehrt proportional zur Anzahl derjeniger Dokumente ist, die das Wort beinhalten. Für eine Dokumentkollektion aus $|D|$ Dokumenten ist die inverse Dokumentfrequenz idf eines Wortes v_k wie folgt definiert:

$$idf(v_k) = \log_2 \left(\frac{|D| + 1}{df(v_k, D) + 1} \right)$$

Zur Vermeidung einer zu starken Gewichtung von Worten, die eine sehr niedrige Dokumenthäufigkeit aufweisen, ist idf ein logarithmisches Maß.

Die $tfidf$ Gewichtung ergibt sich aus der Multiplikation der Termhäufigkeit tf mit der inversen Dokumenthäufigkeit idf und ist wie folgt definiert:

$$tfidf(v_k, d) = tf(v_k, d) \cdot idf(v_k)$$

Die Verknüpfung bewirkt, dass eine inhaltliche Beschreibung durch die häufigsten Worte in einem Dokument erfolgt und gewichtet diese Worte zusätzlich, um das Potential zur Abgrenzung anderer Dokumenten zu erhöhen.

2.4 Vektorrepräsentationen

Ein Dokument d wird anhand des Vokabulars V für eine internen Repräsentation in einem Retrieval-Modell in einen n-dimensionalen Vektor $\mathbf{d}$ mit $k = 1, \ldots, n$ und $n = |V|$ Dimensionen überführt. In der k-ten Dimension steht die gewichtete Auftrittshäufigkeit des Merkmals $v_k \in V$ aus dem Dokument.

Die beiden Dokumente d_1 und d_2 werden exemplarisch jeweils anhand von vier Indextermen, Funktionsworten und Character-3-Grammen repräsentiert. Die Repräsentationen $\mathbf{d}_1$ und $\mathbf{d}_2$ werden in Tabelle 2.1 gegenübergestellt und zeigen die unterschiedlichen Häufigkeiten der Merkmale.

d_1 : *"The hallway smelt of boiled cabbage and old rag mats."* [1]
d_2 : *"Tall and rather thin but upright, the Director advanced into the room."* [2]

Indexterme		Funktionsworte		Character-3-Gramme	
$\mathbf{d}_1$	$\mathbf{d}_2$	$\mathbf{d}_1$	$\mathbf{d}_2$	$\mathbf{d}_1$	$\mathbf{d}_2$
director 0	director 1	the 1	the 2	the 1	the 3
hallway 1	hallway 0	of 1	of 0	he_ 1	he_ 2
cabbage 1	cabbage 0	and 1	and 1	ed_ 1	ed_ 1
room 0	room 1	but 0	but 1	all 0	all 1

Tabelle 2.1: Darstellung von Vektorrepräsentationen für Dokumente d_1 und d_2.

[1] Auszug aus dem Roman "1984" von George Orwell
[2] Auszug aus dem Roman "Brave New World" von Aldous Huxley

2.5 Messung von Stilähnlichkeit

Das Vektorraummodell wurde von Salten u. a. in [SWY75] vorgestellt. Der Ansatz zur Berechnung der inhaltlichen Ähnlichkeit kann als generische Funktion verstanden werden. Es werden zwei Vekoren auf Grundlage des Skalarproduktes aus der linearen Algebra bezüglich des Winkels verglichen, mit der Bedingung, dass beide Vektoren im selben Vektorraum liegen müssen und die Vektorkomponenten aus positiven Werten bestehen. Das bedeuted, dass für zwei Repräsentationen der selben Repräsentationsart, die diese Bedingungen erfüllen, eine Ähnlichkeit berechnet werden kann. Diese Funktion wird als Kosinusähnlichkeit oder Kosinusmaß bezeichnet. Wenn Character-n-Gramme den Merkmalsraum aufspannen, wird der Anzatz nach Stamatatos in [Sta09] auch als Vektorraummodell bezeichnet und im Folgenden näher beschrieben.

Das Skalarprodukt ist für zwei Vektoren a und b wie folgt definiert:

$$a \cdot b \quad := \quad \|a\| \cdot \|b\| \cdot \cos\varphi$$

Der Winkel φ bezeichnet den Winkel zwischen a und b. Man versteht darunter das Bogenmaß, mit $\varphi \in [0, \pi]$. Wenn beide Vektoren nur aus positiven Werten bestehen liegt das Bogenmaßes im Intervall $[0, \frac{\pi}{2}]$. Die Kosinusähnlichkeit berechnet den Kosinus des Winkels zweier Stilrepräsentationen $\mathbf{d}_1$ und $\mathbf{d}_2$, von Dokument d_1 zu Dokument d_2 und ist wie folgt definiert:

$$\rho_{VSM} = \varphi_{cos}(\mathbf{d}_1, \mathbf{d}_2) = \frac{\mathbf{d}_1^T \cdot \mathbf{d}_2}{\|\mathbf{d}_1\| \cdot \|\mathbf{d}_2\|}$$

Dabei ist $\mathbf{d}^T$ der transponierte Vektor $\mathbf{d}$. Die verwendeten Vektoren enthalten gewichtete Aufrittshäufigkeiten der Merkmale und können somit keine negativen Ausprägungen haben. Bei einer hohen Ähnlichkeit von d_1 zu d_2 nähert sich $\varphi(d_1, d2)$ der maximalen Ähnlickeit mit $\cos(0) = 1$ an. Das bedeutet, dass der Winkel kleiner wird und der resultierende Wert größer, wobei die maximale Ähnlichkeit nur duch den Vergleich von identischen Dokumenten erreicht werden kann. Wenn die Repräsentationen orthogonal zueinender stehen, sind die Schreibstile maximal unähnlich durch $\cos(\frac{\pi}{2}) = 0$. Dieser Fall tritt nur ein, wenn die gemeinsame Schnittmenge der Stilmerkmale in d_1 und d_2 leer ist. Um den Einfluss der Textlänge auf die berechnete Ähnlichhkeit zu vermeiden werden die Vektoren normiert. Die Kosinusähnlichkeit für normierte Vektoren entspricht deren Skalarprodukt.

2.6 Projektionsmodell von Koppel

Koppel u. a. beschreiben in [KSA10] eine neue Methode, um die Ähnlichkeit zwischen Schreibstilen zu berechnen. In seiner Arbeit wird die Autorenanzahl auf 10.000 Autoren erhöht, wobei jeder Autor mit einem Textbeispiel vertreten ist. Die Textkollektion[3] besteht aus Blogs, wodurch eine Abgrenzung zu klassischen Studien der Autorschaftsbestimmung auf Basis literarischen Werken wie Büchern oder Gedichten [Bur02] stattfindet.

Der Ansatz basiert auf zwei grundlegenden Überlegungen. Erstens, um derartig viele Autoren effizient durch ein Modell unterscheiden zu können, sollte eine relativ einfache, nicht rechenintensive Methode verwendet werden, die für eine vergleichende Stilerfassung gut geeignet ist. Hierfür wird die Kosinusähnlichkeit zwischen einem Dokument eines unbekannten Autors zu Dokumenten bekannter Autoren, deren Schreibstile durch Character-4-Gramm-Vektoren repräsentiert werden, berechnet. Koppel u. a. beziehen sich dabei auf die Autoren in [KSAM06] und [LD08], die erwähnen, dass ähnlichkeitsbasierte Methoden im Vergleich zu Methoden aus dem maschinellen Lernen in diesem Zusmammenhang besser geeignet sind. Zweitens, diese von Koppel u. a. als "naiv" bezeichnete Methode verwendet eine festgelegte Menge an Character-4-Grammen zur Repräsentation des Schreibstils und erzielt gute Ergebnisse unter Berücksichtigung der großen Anzahl von Autoren. In seiner Arbeit wurden mit dieser Methode 1.000 Textbeispiele jeweils mit 10.000 Texten verglichen. 46% konnten so richtig zugeordnet werden. Das bedeutet, dass ein Teil der Autoren durch die verwendeten Character-4-Gramme differenzierbar sind. Für die Autorschaftsbestimmung ist eine Precision von 46 % jedoch inadequat. Das Ziel der neuen Methode ist, die Precsion zu steigern. Dazu wurde der Ansatz in [KSAM06] verwendet, indem die Antwort: "Autor unbekannt", in den Fällen zugelassen wird, in denen die Zuordnung als nicht verlässlich bzw. unsicher erscheint. Um die Sicherheit der Zuordnung zu erhöhen, wird der naive Ansatz n mal ausgeführt, wobei zur Repräsentation immer eine andere, zufällig ermittelte Character-n-Gramm-Menge verwendet wird. Dies entspricht der Konstruktion von n Vektorraummodellen, die sich durch die Auswahlmenge der Character-4-Grammen und darurch in den Repräsentationen unterscheiden. Koppel u. a. gehen davon aus, dass bei einer Variation der Character-n-Gramm-Mengen der berechnete Schreibstil des Dokuments vom selben Autor am häufigsten dem ähnlichsten Schreibstil entspricht. Schreibstile anderer Autoren können vereinzelt als ähnlicher bewertet werden. Es wird aber als unwahrscheinlich aufgefasst, dass ein anderer Autor konsistent über die verschiedenen Merk-

[3] Beschreibung und Download des Korpus auf `http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm`

malsmengen, die höchsten Ähnlichkeiten aufzeigt. Im Modell liefert jede Iteration bezüglich der Merkmalsmengen genau ein Ranking an Schreibstilen bekannter Autoren. Der ähnlichste Schreibstil wird über eine Menge an Rankings durch die Anzahl an Top-Platzierung berechnet. Die Sicherheit der Zuordnung ergibt sich durch die Festlegung, wie oft ein Autor die Top-Platzierung erreichen muss. Wird die Schwelle von diesem Autor unterschritten, dann ist der ähnlichste Autor dem Modell nicht bekannt.

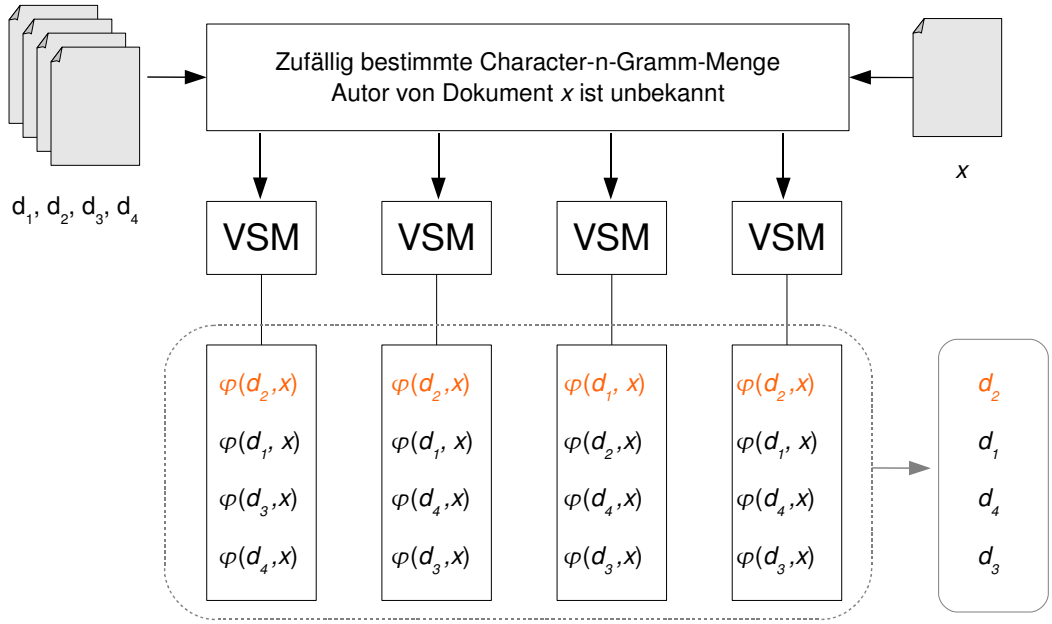


Abbildung 2.2: Die Abbildung zeigt den Prozess des neuen Modells am Bespiel von vier Dokumenten unterschiedlicher Autoren. Im paarweisen Vergleich mit Dokument x des unbekannten Autors, wird der Autor von Dokument d_2 als Autor für x zugewiesen. Der Schreibstil dieses Dokuments wurde dreimal, von vier Durchgängen am ähnlichsten zu x bewertet.

In dieser Arbeit wird der Ansatz erweitert, indem die Berechnung des ähnlichsten Schreibstils aus den Ergebnismengen als Ranking-Fusionierungsproblem aufgefasst wird, wobei die optimale Zusammenfassung der Mengen gefunden werden muss um die Precision und den Recall zu steigern. Dabei werden zwei Methoden, nämliche die sog. rank based fusion und die sog. score based fusion, betrachtet. Erstere ist eine Methode zur Aggregation von Rängen bzw. der Positionen von Dokumenten in den Rankings. Die zweite Methode bezieht sich auf die Zusammenfassung der erzielten Ähnlichkeitswerte [RS03]. Die beiden Methoden können auch zu einem Hybrid zusammengefasst werden. Durch die Fusionierung ist die Top-Platzierung eines

Dokuments des gleichen Autors in den jeweiligen Rankings nicht mehr das alleinige Kriterium. Es genügt, wenn das Dokument über viele Rankings konsistent hoch bewertet wird, d.h. in jedem Durchgang wird eine hohe Ähnlichkeit ermittelt, aber nicht zwingend die höchste. Für die Schreibstile aus Dokumenten anderer Autoren wird angenommen, dass die Ähnlichkeiten und Positionen in den Rankings stark variieren und im Durchschnitt schlechter sind.

Für das erweiterte Modell bezeichnet x ein Dokument eines unbekannten Autors und $d \in D$ ein Dokument aus der Menge von Dokumenten, mit $|D|$ bekannten Autoren. V ist die Menge aller Character-n-Gramme aus D, und $V^{'} \subset V$ eine zufällige Auswahl an $|V'|$ Elementen. Ein Ranking wird mit S bezeichnet und $\mathcal{S}, S \in \mathcal{S}$, bildet eine Menge aus verschiedenen Rankings. Das Ranking $S_{x|V'}$ von x bezüglich V' ist eine Sequenz und enthält die Kosinusähnlichkeiten $\varphi(\mathbf{d}, \mathbf{x})$ der durch V' repräsentierten Dokumente $\mathbf{x}$ und $\mathbf{d} \in \mathbf{D}$ mit $S_{x|V'} = (\varphi(\mathbf{d}_1, \mathbf{x}), \ldots, \varphi(\mathbf{d}_{|\mathbf{D}|}, \mathbf{x}))$. Für Dokument d_i, mit $i \in [1, |D|]$ bildet $S_{x|V'}(d_i)$ die Position des Dokuments im Ranking ab, mit $\varphi(\mathbf{d}_i, \mathbf{x}) > \varphi(\mathbf{d}_j, \mathbf{x}) \rightarrow S_{x|V'}(d_i) < S_{x|V'}(d_j)$. Die Menge aller Rankings $\mathcal{S}_x$ bezüglich x beinhaltet n Rankings $|\mathcal{S}_x| = n$, mit $\mathcal{S}_x = \{S_{x|V_1'}, \ldots, s_{x|V_n'}\}$. Die Gewichtung $\omega_S(d)$ ist die Fusionierungsmethode. Die Art unterscheidet sich wie folgt in

rank based fusion: $$\omega_S^r(d_i) = 1 - \frac{s(d_i) - 1}{|\mathbf{D}|}$$

score based fusion: $$\omega_S^s(d_i) = \varphi(\mathbf{d}_i, \mathbf{x})$$

rank & score based: $$\omega_S^{rs}(d_i) = \alpha \cdot \omega_S^r + \beta \cdot \omega_S^s$$

Dabei bezeichnen α und β Gewichtungsfaktoren mit $\alpha, \beta \in [0, 1]$. Die Fusionierungsmethode wird über $\mathcal{S}$ angewendet, wobei $f_x(d_i)$ den resultierenden Fusionswert bezüglich d_i zu x aus allen Rankings $\mathcal{S}_x$ ausdrückt und wie folgt definiert ist:

$$f_x(d_i) = \sum_{S \in \mathcal{S}_x} \omega_S(d_i)$$

Das fusionierte Ranking S_{f_x} enthält alle Fusionswerte $f_x(d_1), \ldots, f_x(d_{|D|})$. Der Autor des Dokuments d_i mit dem ähnlichsten Schreibstil erzielt den maximalen Wert durch die Fusionierung. Für die Insignifikanzschwelle σ gilt $\sigma \in [0, n]$. Ist der maximle Wert kleiner als σ so kann der Autor von x nicht bestimmt werden.

$$\text{Autor von } x := \begin{cases} \text{Autor von } d_i \text{ mit } f_x(d_i) = S_{f_x}(1), & \text{wenn} f_x(d_i) > \sigma \\ \text{unbekannt, sonst} \end{cases}$$

Für die Aggregation der Platzierungen werden nach obiger Formel die Position im Intervall $(0, 1]$ normalisiert, wobei Platz 1 mit 1.0 den maximalen Wert annimmt. Bei n Iterationen erhält im Idealfall das Dokument d, welches von dem selben Autor geschrieben wurde wie x, einen Fusionswert $f_x(d) = n$. Die Kosinusähnlichkeit berechnet Werte aus demselben Intervall. Der maximale Wert liegt nahe bei n, da $\varphi(\mathbf{d}, \mathbf{x}) = 1.0$ praktisch nur von identischen Dokumenten erzielt werden kann.

2.7 ESA-Modell zur Autorschaftsbestimmung

Explicit Semantic Analysis (kurz ESA) bezeichnet ein weiteres Retrieval-Modell, welches die inhaltliche Ähnlichkeit von Dokumenten in einem Vektorraum berechnet. Im Gegensatz zum Vektorraummodell, in dem die thematische Nähe von Dokumenten im direkten Vergleich ermittelt wird, zieht das ESA-Modell eine externe Informationsquelle für eine „semantische Analyse“ hinzu. Wenn wir einen Text lesen und inhaltlich erschließen, findet automatisch ein Abgleich mit dem vorhandenen Wissen und eigenen Erfahrungen statt. Wir greifen dabei auf Vorwissen zurück, um sinnmäßige Verknüpfungen zu bilden und das Gelesene zu kontextualisieren. Der ESA Ansatz versucht, die Art, wie die menschliche Organisation von Wissen funktioniert, anzugleichen. Das Modell greift diesen Gedanken auf und bildet sein Hintergrundwissen durch externen Dokumente. Die Online-Enzyklopädie Wikipedia[4] eignet sich laut Gabrilovich und Markovitch [GM07] als geeignete Basis, um das Vorwissen in Form von Wikipediaartikeln zu generieren. Die Artikel werden als Konzepte bezeichnet, da sie einen gewissen Sachverhalt oder Gegenstand ausführlich beschreiben. Um den Inhalt eines Dokuments zu beurteilen, vergleicht das Modell dazu die Kosinusähnlichkeit zu den vorhanden Artikel. Dadurch entsteht ein Vektor, den die Autoren als Interpretationsvektor bezeichnen, in dem in jeder Dimension die inhaltliche Ähnlichkeit des Dokuments zu einem Konzept steht. Um zwei Dokumente zu vergleichen, werden zwei Interpretationsvektoren berechnet. Die Ähnlichkeit zwischen Interpretationsvektoren wird wieder mit der Kosinusähnlichkeit ermittelt.

In dieser Arbeit wird der ESA-Ansatz erstmalig für die Autorschaftsbestimmung eingesetzt und untersucht. Hierfür werden statt Interpretationsvektoren Stilähnlichkeitsvektoren eines Dokuments durch Character-3-Gramme bezüglich der Stilähnlichkeiten zu Wikipediaartikeln berechnet. Durch den Stilähnlichkeitsvektor soll das Potential zur Abgrenzung verschiedener Schreibstile erhöht werden. Für zwei gegebene Dokumente desselben Autors wird angenommen, dass sehr ähnliche Schreibstile, zu den überwiegend gleichen Konzepten gefunden werden. Beide Dokumente

[4] Online Enzyklopädie Wikipedia `http://en.wikipedia.org/wiki/Main_Page`

verweisen also auf eine Menge von Konzepten durch hohe Stilähnlichkeiten. Die Kosinusähnlichkeit für zwei Stilähnlichkeitsvektoren steigt, wenn hohe Werte in den gegnübergestellten Vektorkomponenten auftreten und sinkt, wenn deren Werte stark unterschiedlich sind.

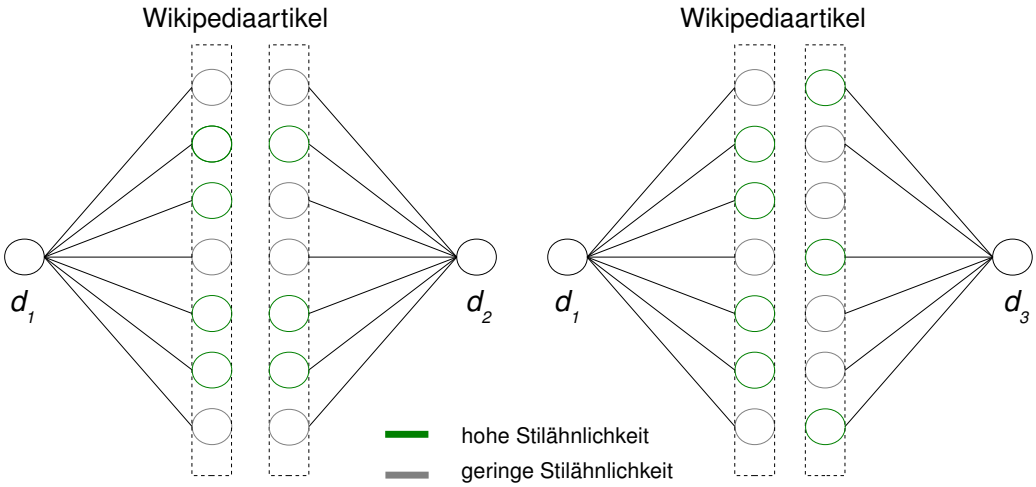


Abbildung 2.3: Die Abbildung zeigt den Zusammenhang zu ähnlichen und unähnlichen Wikipediaartikeln, wobei zwei Dokumente d_1 und d_2 vom selben Autor mehr Übereinstimmungen aufweisen im Vergleich zu den Dokumenten d_1 und d_3 von unterschiedlichen Autoren.

Für das ESA-Modell bestehen somit die wesentlichen Schritte aus der Konstruktion von Stilähnlichkeitsvektoren und der anschließende Vergleich, um die Stilähnlichkeit zwischen zwei Dokumente zu erfassen.

Seien d und x zwei Dokument. $\mathbf{d}$ und $\mathbf{x}$ die Repräsentationen des Schreibstils durch Character-4-Gramme im Vektorraummodell. D^* bezeichnet eine Menge externer Dokumente und M_{D^*} eine Stildokumentmatrix, wobei jede Spalte j einer Repräsentation $\mathbf{d}_j$ des Schreibstiels eines externen Dokuments $d_j \in D^*$ entspricht. Für alle Dokumente gilt eine normalisierte Vektorrepräsentation, mit $\|\mathbf{x}\| = \|\mathbf{d}\| = \|\mathbf{d}_j\| = 1$. Der Stilähnlichkeitsvektor $\mathbf{d}_s$ für d und $\mathbf{x}_s$ für x ist wie folgt definiert:

$$\mathbf{d}_s = M_{D^*}^T \cdot \mathbf{d} \text{ und } \mathbf{x}_s = M_{D^*}^T \cdot \mathbf{x}$$

Dabei ist $M_{D^*}^T$ die transponierte Matrix von M_{D^*} [Ste].

Für den Vergleich von einem Dokument d_i, aus einer Menge von Dokumente D mit $d_i \in D$ und $i = \{1, \ldots, |D|\}$, zu einem Dokument x gilt die Kosinusähnlichkeit zwischen $\mathbf{d}_{si}$ und $\mathbf{x}_s$, wobei das Dokument des Autors mit dem ähnlichsten Schreibstil bezüglich x den höchsten Wert annimmt:

$$\rho_i = \varphi(\mathbf{d}_{si}, \mathbf{x}_s) \text{ , Autor von } x := \max \rho_i$$

Die verwendeten externen Dokumente bilden eine konstante Menge an Indexdokumenten, die in diesem Modell auf verschiedene Schreibstile abbilden. Die Autoren schlagen vor, die Indexdokumente in Form eines invertierten Indexes abzulegen. Die Datenstruktur ermöglicht einen effizienten, hashwertbasierten Zugriff auf die jeweilige Auftrittshäufigkeit der Character-4-Gramme in den Dokumenten. Jedes Character-4-Gramm $v_k \in V$ mit $k = \{1, \ldots, |V|\}$ bildet einen Schlüssel ab. Ein Eintrag im invertierten Index bestehen aus einem v_k und einer Liste $P(v_k)$ mit $P(v_k) = (d_1(v_k), \ldots, d_{|D^*|}(v_k))$. Die Liste wird als Postlist für den Schlüssel v_k bezeichnet und liefert das Gewicht von v_k aus den Konzepten D^*. Somit bildet eine Postlist genau eine Zeile der Stildokumentmatrix ab. Die Einträge der Liste quantifizieren, wie stark die Schreibstile aus den externen Dokumenten mit v_k im Zusammenhang stehen. Tabelle 2.2 zeigt beispielhaft, die Struktur eines invertierten Indexes.

V	d_1	d_2	d_3	$\cdots$	$d_{\lvert D^*\rvert}$
v_1	0.4	0.1	0.3	$\cdots$	0.2
v_2	0.1	0.5	0.2	$\cdots$	0.1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$v_{\lvert V\rvert}$	0.1	0.2	0.4	$\cdots$	0.3

Tabelle 2.2: Strukturdarstellung eines invertierten Indexes. Jede Spalte j in der Tabelle enthält die Auftrittshäufigkeit eines Character-n-Gramms $v_k \in V$ aus einem Dokument d_j mit $j = \{1, \ldots, |D^*|\}$.

Zur Konstruktion des Stilähnlichkeitsvektors für ein Dokument d wird mit jedem $v_k \in d$ auf den Index zugegriffen. Das Gewicht $d(v_k)$ wird mit jedem Element der Postlist $p(v_k)$, multipliziert. Aus dem Index wird die Postlist eines Character-n-Gramms nur dann angefragt, wenn $d(v_k) > 0$ ist. Die Postlisten werden elementweise addiert. Für den Stilähnlichkeitsvektor $\mathbf{d}_s$ zu Dokument d gilt dann:

$$\mathbf{d}_s = \sum_{v_k \in d}^{|V|} d(v_k) \cdot p(v_k)$$

3 Überwachtes Lernen

Maschinelles Lernen bezeichnet automatisierte Lernverfahren, die für eine bestimmte Aufgabe nach einem Modell bzw. Lernparadigma Erfahrungswerte generieren. Die erlernten Erfahrungswerte bilden die Grundlage für Vorhersagen hinsichtlich der spezifischen Aufgabenstellung. Überwachtes Lernen bezeichnet ein Lernparadigma, welches für Aufgaben im Bereich der automatischen Klassifikation verwendet wird.

Der gesamte Prozess einer Klassifikation durchläuft zwei Phasen (Abb. 3.1). Dafür wird die Menge an Dokumenten D in eine Trainingsmenge D' und eine Testmenge D'' aufgeteilt, mit $D = D' \cup D''$ und $D' \cap D'' = \emptyset$. In Phase 1 dient die Trainingsmenge einem Lernalgorithmus als Eingabedaten. D' wird in Merkmalsvektoren $\mathbf{d} \in \mathbf{D}'$ überführt, die ein Lernalgorithmus verwendet um Beziehungen zwischen Merkmalsausprägungen in der Trainingsmenge und den Klassen C zu finden. Daraus ergibt sich ein gelerntes Modell, welches in der zweiten Phase durch D'', mit $\mathbf{D}''$ Merkmalsvektoren evaluiert wird und eine resultierende Genauigkeit ergibt. Die Aufteilung wird dabei durch die Evaluierungsmethode bestimmt. Die Genauigkeit, mit der der Lernalgorithmus die Klassifikation mittels des gelernten Modells auf den Testdaten durchführt, kann durch verschiedene Gütemaße bestimmt werden. Hierfür eigenen sich die in 2.1 vorgestellten Maße, Precision und Recall.

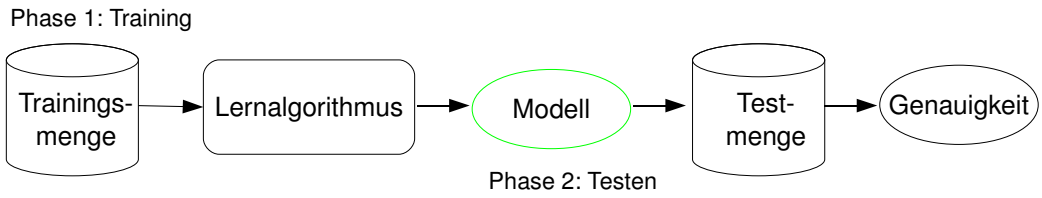


Abbildung 3.1: Phasendarstellung eines maschinellen Lernvorgangs aus [Bin07].

3.1 Formale Grundlagen

Bei einer Klassifikation wird zunächst angenommen, dass ein idealer Klassifikator γ, der auf realen Objekten arbeitet, existiert und die Objekte den Klassen C eindeutig zuordnen kann. Die Objekte werden durch eine Modellbildungsfunktion α in einen Merkmalsraum A überführt. Die Aufgabe ist nun, den idealen Klassifikator durch Techniken des maschinellen Lernens zu approximieren. Der Klassifikator $\mathcal{C}$ bezeichnet den zu lerneden Klassifikator auf dem Merkmalsraum. Die Approximationsfunktion für $\mathcal{C}$ wird als Zielfunktion y bezeichnet. Daraus lassen sich die drei Schritte ableiten, die zur Konstruktion eines Klassifikators benötigt werden:

(1) Überführung der Objekte in den Merkmalsraum

(2) Erstellung einer Trainingsmenge für den Lernalgorithmus

(3) Approximation eines Klassifikators durch Techniken des maschinellen Lernens

In [Sta09] wird das Problem der Autorschaftsbestimmung als "multi-class single-label text categorization task" nach Sebastiani in [Seb02] aufgefasst. Die zu klassifizierenden Objekte sind Dokumente D. Die Aufgabe besteht darin, zu jedem Tupel $\langle d_i, c_j \rangle \in D \times C$, mit $C = \{c_1, \ldots, c_{|C|}\}$ und $|C| > 2$ vordefinierter Klassen, welche die Menge an Autoren bilden, einen Wert aus $\{0, 1\}$ zuzuweisen. Für jedes Dokument d_i wird entschieden, ob dieses Dokument von c_j geschrieben wurde, mit $\langle d_i, c_j \rangle \rightarrow 1$, oder nicht, mit $\langle d_i, c_j \rangle \rightarrow 0$. Single label bezeichnet den Sachverhalt, dass genau ein $c_j \in C$ zugeordnet werden kann. Die Zielfunktion y beschreibt den Lernalgorithmus um die relationale Abbildung $y : D \times C \rightarrow \{0, 1\}$ zu ermöglichen.

Die Überführung in den Merkmalsraum bedeutet eine Vektorrepräsentation $\mathbf{d} = \mathbf{d} \cup \{c_j\}$ eines Dokuments d mit der Klassenzugehörigkeit c_j anhand von Merkmalsausprägungen $A_k = \{a_1, \ldots, a_{|A_k|}\}$ aus der Merkmalsmenge mit $A \in \{A_1, \ldots, A_{|A|}\}$ Merkmalen, die den Raum aufspannen. Die Repräsentation wird als Merkmalsvektor bezeichnet.

Die Modellbildungsfunktion α kann beispielsweise durch die Vektorrepräsentation eines Dokuments im Vektorraummodell durch Character-n-Gramm-Vektoren erfolgen, oder durch die in Kapitel 2.2 beschriebenen Stilähnlichkeitsvektoren im ESA-Modell.

3.2 Attributselektion

Die Qualität eines automatischen Lernverfahrens hängt zentral von dem verwendeten Lernalgorithmus ab. Dabei wird jeder Lernalgorithmus maßgebend von den verwendeten Attributen, die eine Unterscheidbarkeit ermöglichen, beeinflusst. Ein Verfahren, welches beurteilt, wieviel Information die Attribute für die Unterscheidung zwischen verschiedenen Klassen beisteuern, ist das sog. information gain. Das Verfahren misst den Informationsunterschied in Bits, der durch Anwesenheit oder Abwesenheit eins Attributs entsteht.

Basierend auf der informationstheoretischen Entropie nach Shannon [Lin91] setzt sich das information gain der Trainingsmenge D' aus der Entropie $H(D')$ und der Entropie der Attribute $H_{A_k}(D')$, mit $A = \{A_1, \dots A_{|A|}\}$ und $k \in [1, |A|]$, zusammen. Das gain eines Attributs A_k ist wie folgt definiert:

$$\text{gain}(D', A_k) \quad = \quad H(D') - H_{A_k}(D')$$

Die Entropie $H(D')$ ermittelt den Informationsgehalt zur Verteilung der Trainingsmenge bezüglich der Klassenzugehörigkeit $c_j \in C$ und ist wie folgt definiert:

$$H(D') \quad = \quad -\sum_{j=1}^{|C|} P(c_j) \log_2 P(c_j)$$

Die Entropie misst die Verteilung der Instanzen bezüglich ihrer Klassenzugehörigkeit. Je höher dieser Wert ist, desto gleichmäßiger sind Instanzen der jeweiligen Klassen verteilt. Bei einem kleinen Wert steigt die Reinheit (purity) der Verteilung in einer Menge, gleichzeitig sinkt die Unreinheit (inpurity). Abbildung 3.2 zeigt das Verhalten für den 2-Klassenfall mit der Wahrscheinlichkeit P und der Entropie H. Die Entropie ist am größten, wenn gleich viele Bespiele beider Klasse vorhanden sind.

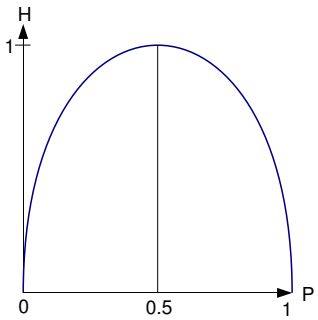


Abbildung 3.2: Illustration der Entropie im 2-Klassenfall aus [Ste08a]

Zur Bestimmung des Informationsgehalts der Attribute wird jedes Attribut evaluiert, indem die Trainingsmenge in disjunkte Teilmengen für jede mögliche Attributausprägung partitioniert wird. Auch hier werden die Attribute hinsichtlich der impurity untersucht. Für ein Attribut $A_k \in A$, sind $|A_k|$ Ausprägungen vorhanden. Die Partitionierung erfolgt in die Teilmengen $\{D'_1, \ldots, D'_{|A_k|}\}$, mit $D'_i = \{D_i | i \in A_k\}$. Der Informationsgewinn eines Attributs A_k aus einer Trainingsmenge D', ist dann wie folgt definiert:

$$H_{A_k}(D') = \sum_{i=1}^{|A_k|} \frac{|D'_i|}{|D'|} \cdot H(D'_i)$$

Information gain wird als impurity-Funktion aufgefasst, da jedes Attribut untersucht wird, ob es die impurity durch eine Partitionierung reduzieren kann. Um das Verfahren auf Attribute anzuwenden, die kontinuierliche Werte annehmen, werden Methoden zur Diskretisierung der Attribute für alle auftretenden Werte aus den Trainingsbeispielen angewandt. Die resultierenden Intervallgrenzen werden dann wie diskrete Ausprägungen behandelt.

3.3 Naive Bayes

Der Naive Bayes Klassifikator stammt aus dem Bereich der statistischen Lernverfahren und leitet die Klassenzugehörigkeit eines Objekts von bedingten Wahrscheinlichkeiten ab. Den Ausgangspunkt bilden die bedingte Wahrscheinlichkeit, die totalen Wahrscheinlichkeit sowie der Satz von Bayes, der als a-Posteriori-Wahrscheinlichkeit eines Ereignisses bezeichnet wird. Zur Definition wurden die Aufzeichnungen von [Ste08b] verwendet und angepasst.

Sei $(\Omega, \mathcal{P}(\Omega), P)$ ein Wahrscheinlichkeitsraum mit dem Ereignis B, und $A \in P(\Omega)$ mit $P(A) > 0$, dann ist die bedingte Wahrscheinlichkeit wie folgt definiert:

$$P(B|A) = \frac{P(B) \cap P(A)}{P(A)} = \frac{P(A|B) \cdot P(B)}{P(A)}$$

Sei $(\Omega, \mathcal{P}(\Omega), P)$ ein Wahrscheinlichkeitsraum mit den disjunkten Ereignissen $\Omega = B_1 \cap \ldots \cap B_{|B|}$, $P(B_j) > 0, j = 1, \ldots, |B|$ und $A \in P(\Omega)$ mit $P(A) > 0$, dann

ist die totale Wahrscheinlichkeit $P(A)$ und der Satz von Bayes $P(B_j|A)$ wie folgt definiert:

$$\begin{aligned} P(A) &= \sum_{j=1}^{|B|} P(A|B_j) \cdot P(B_j) \\ P(B_j|A) &= \frac{P(A|B_j) \cdot P(B_j)}{\sum_{j=1}^{|B|} P(A|B_j) \cdot P(B_j)} = P(B_j) \cdot \frac{P(A|B_j)}{P(A)} \end{aligned}$$

Zur Klassifikation lässt sich durch den Satz von Bayes bestimmen, mit welcher Wahrscheinlichkeit $P(C{=}c_j|d)$, ein Dokument d zur Klasse $c_j \in C$ zuzuordnen ist. Das Dokument d wird durch den Merkmalsvektor $\mathbf{d} = \langle A_1{=}a_1, \ldots, A_{|A|}{=}a_{|A|}\rangle$ repräsentiert, mit einer Menge an diskreten Attributen $A_1, \ldots, A_{|A|}$. Dabei treten die Ausprägungen $a_1, \ldots, a_{|A|}$ der Attribute A in $\mathbf{d}$ auf, wobei a_k eine mögliche Merkmalsausprägung des Attributs A_k, mit $k \in [1, \ldots |A|]$ darstellt. Zur Annäherung wird die bedingte Wahrscheinlichkeit $P(A|C{=}c_j)$ abgeschätzt. Dazu wird angenommen, dass unter Bedingung der Klasse c_j keine Korrelation zwischen den Attributwerten besteht. Das Auftreten der Werte $a_1, \ldots, a_{|A|}$ wird als stochastisch unabhängig bewertet. Diese Annahme wird als Naive Bayes Assumption bezeichent und ist wie folgt definiert:

$$P(A_1{=}a_1, \ldots, A_{|A|}{=}a_{|A|}|C{=}c_j) \overset{NB}{=} \prod_{k=1}^{|A|} P(A_k{=}a_k|C{=}c_j)$$

Für einen Naive Bayes Klassifikator haben die a-Priori-Wahrscheinlichkeiten der Merkmalsvektoren $P(A_1{=}a_1, \ldots, A_{|A|}{=}a_{|A|})$ keinen Einfluss auf das Klasssifikationsergebnis. Sie sind für jede Klasse gleich hoch. Die Ausgangsformel 3.1 reduziert sich auf die a-Priori-Wahrscheinlichkeiten der Klassen und die bedingten Wahrscheinlichkeit der Attributwerte. Die Wahrscheinlichkeiten können aus der Trainingmenge mit den Dokumenten D' berechnet werden:

$$\begin{aligned} P(C{=}c_j) &= \frac{|(D'|C{=}c_j)|}{|D'|} \\ P(A_k{=}a_k|C{=}c_j) &= \frac{|(D'|A_k{=}a_k \in c_j)|}{|(D'|C{=}c_j)|} \end{aligned}$$

Um ein Dokument d aus der Testmenge zu klassifizieren wir die Klasse, mit der größten Wahrscheinlichkeit berechnet:

$$c_{NB} = \arg\max_{c_j \in C} P(C{=}c_j) \prod_{k=1}^{|A|} P(A_k{=}a_k|C{=}c_j)$$

3.4 Entscheidungsbaum

Entscheidungsbäume benutzen Zerlegungsstrategien, um die Trainingsdaten in eine generalisierte Datenstruktur in die Form eines endlichen Baumes zu überführen. Jeder Knoten entspricht einem Attribut und partitioniert den Merkmalsraum der Trainingsmenge. Dabei entspricht ein Pfad den Attributtests bezüglich der Attributausprägungen. Jeder Blattknoten eines Baumes repräsentiert die Ausprägung einer Klasse. Für eine gegebene Instanz aus der Testmenge erfolgt die Klassifikation durch eine Traversierung des Baumes, indem Konjunktionen bezüglich Knoten und Merkmalsausprägungen der Instanz erfüllt werden.

Abbildung 3.3 zeigt exemplarisch die räumliche Darstellung ermittelter Werte für zwei Stilmerkmale A_1 und A_2 aus einer Trainingsmenge mit Dokumenten von zwei Autoren c_1 und c_2. Da es sich um kontinuierliche Werte handelt, werden zur Diskretisierung Intervalle für beide Attribute ermittelt, die in der Abbildung durch Separierungslinien abgebildet werden. Rechts in der Abbildung wird der Entscheidungsbaum anhand der aufgeteilten Trainingsmenge gezeigt.

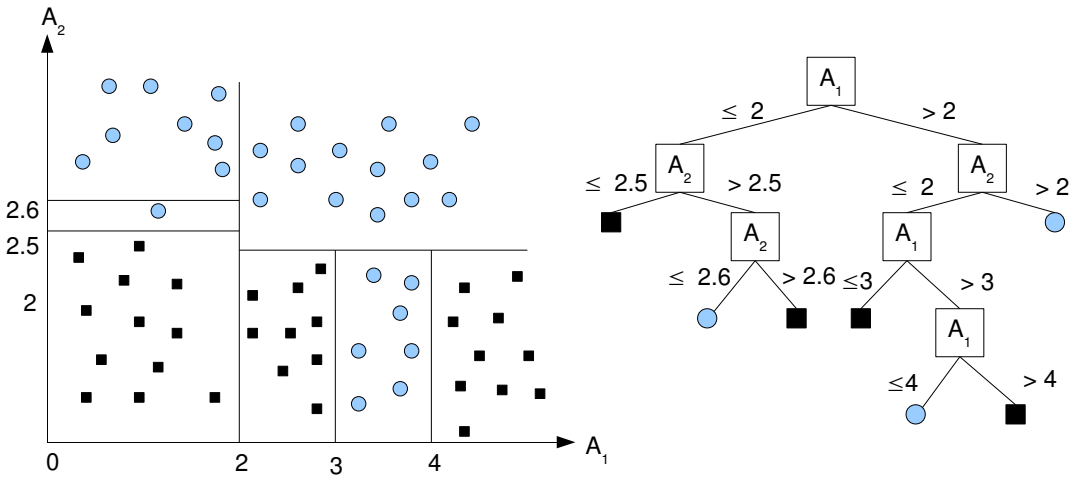


Abbildung 3.3: Links: Räunliche Partitionierung von Trainingsdaten, anhand der Ausprägungen von Stilmerkmalen A_1 und A_2. Rechts: Ein Entscheidungsbaum bezüglich der Aufteilung der Trainingsmenge nach [Bin07].

Zu einem Datensatz können mehrere strukturelle Alternativen gefunden werden, die versuchen, den selben Sachverhalt abzubilden, jedoch den Merkmalsraum anders aufteilen. Die Aufteilung hängt von der Wahl der impurity-Funktion ab, die durch die Berechnung des information gain aus 3.1 abgebildet werden kann. Die Anforderung an die Funktion ist, Attribute zu selektieren, wodurch eine minimale Teilmenge an

Regeln erzeugt werden kann, die hinreichend für eine Klassifikation geeignet sind. Die Funktion soll die impurity der Teilmengen minimieren, das bedeutet, dass die jeweiligen Zerlegungen pro Ebene des Baumes auf möglichst wenig verschiedene Klassen zutreffen. Ein Attributtest muss die bestmöglichste Separierung bezüglich der Klassenzugeörigkeit darstellen.

3.5 Stützvektormethode

Die Stützvektormethode (engl. Support Vector Machine, SVM) bezeichnet ein lineares Lernsystem zur binären Klassifikation von Objekten. Die Idee ist, die Objekte in einem Raum zu betrachten, und eine Hyperebene zu finden, welche die Objekte in zwei Halbräume aufteilt, die somit als positiv oder negativ bewertet werden können.

Eine Ebene wird im Raum durch das Skalarprodukt zwischen der normalen w der Ebene und einem Vektor sowie dem Abstand b der Ebene beschrieben. Zur Konstruktion des Klassifikators soll eine lineare Funktion der Form $f(\mathbf{d}) = \langle w \cdot \mathbf{d} \rangle + b$ bestimmt werden, wobei f der Ebene entspricht und $\mathbf{d}$ die Vektorrepräsentation bezüglich eines Dokuments d ist. Dafür werden die Dokumente D' aus der Trainingsmenge verwendet. Ein Vektor $\mathbf{d}_i$ wird, bezogen auf die Normale, als positiv bezeichnet, falls dieser vor der Hyperebene liegt und negativ falls dieser dahinter liegt.

$$c_i = \begin{cases} +1, & \text{wenn } \langle w \cdot \mathbf{d}_i \rangle + b \geq 0 \\ -1, & \text{wenn } \langle w \cdot \mathbf{d}_i \rangle + b < 0 \end{cases}$$

Der maschinelle Lernvorgang bezieht sich darauf, die optimal trennende Hyperebene $\mathcal{H}$ aus einer Menge an potentiellen Hyperebene zu finden, die den Abstand zwischen den Unterräume maximiert. Zur Annäherung werden auf beiden Seiten parallel verlaufende Hyperebenen $\mathcal{H}_+$ und $\mathcal{H}_-$ hinzugefügt.

$$\begin{aligned} \mathcal{H}_+ = \langle w \cdot \mathbf{d}_i \rangle + b &\geq +1 \\ \mathcal{H}_- = \langle w \cdot \mathbf{d}_i \rangle + b &= -1 \end{aligned}$$

Dadurch wird eine Art Korridor gebildet, wobei die Vektoren auf beiden Seiten ermittelt werden, die die jeweilige Hyperebenen tangieren und den Korridor maximal verbreitern. Diese Vektoren bilden eine diskriminierende Untermenge für den Klassifikator und werden als Stützvektoren bezeichnet.

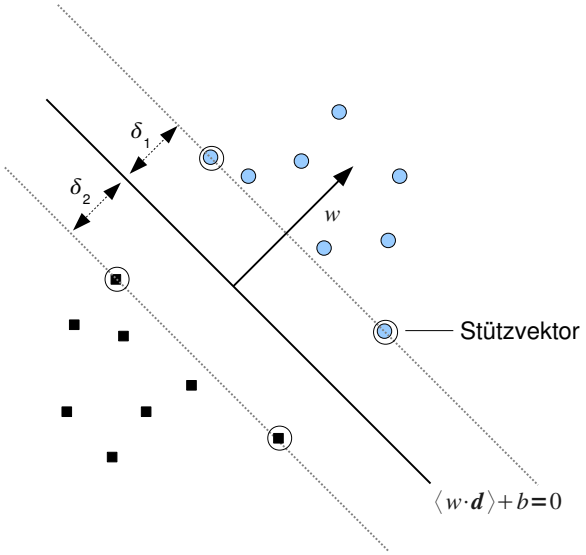


Abbildung 3.4: Die Grafik stammt aus [Bin07] und zeigt die räumliche Darstellung der Positionen von Vektoren aus zwei Klassen, die durch eine Ebene getrennt werden.

Die Euklidische Distanz ermittelt den Abstand δ eines Punktes zu einer Ebene und berechnet sich durch einen Punkt p, der Normalen w und dem Abstand b der Ebene. Für die Hyperebenen ist der maximale Abstand von der Distanz δ_1 eines nächstgelegenen Vektors $\mathbf{d}_1$, der $\mathcal{H}_+$ tangiert, sowie der Distanz δ_2 eines nächstgelegenen Vektors $\mathbf{d}_2$, der $\mathcal{H}_-$ tangiert, abhängig und berechnet sich durch das Gleichsetzen und umformen der beiden Distanzen:

$$\begin{aligned} \delta_1(\mathcal{H}_+, \mathbf{d}_1) &= \frac{\langle w \cdot \mathbf{d}_1 \rangle + b - 1}{||w||} \\ \delta_2(\mathcal{H}_-, \mathbf{d}_2) &= \frac{\langle w \cdot \mathbf{d}_2 \rangle + b + 1}{||w||} \\ \delta_1 + \delta_2 &= \frac{\langle w \cdot (\mathbf{d}_1 - \mathbf{o}_2) \rangle}{||w||} = \frac{2}{||w||} \end{aligned}$$

Um den Abstand oder die Trennspanne zu maximieren, formuliert man das Optimierungsproblem hinsichtlich der Minimierung des Betrages der Normalen der Hyperebene wie folgt:

$$\text{minimiere} : \frac{\langle w \cdot w \rangle}{2}$$

$$\text{unter den Nebenbedingungen} : c_i(\langle w \cdot \mathbf{d}_i \rangle + b) \geq 1, i = 1, \ldots, |D'|$$

Das Problem wird mit Hilfe der Lagrangefunktion $L(w, b, \alpha)$ zusammengefasst, wobei α den Vektor der Lagrange-Multiplikatoren α_i für $i = \{1, \ldots, |D'|\}$ abbildet:

$$L(w, b, \alpha) = \frac{1}{2}\langle w \cdot w \rangle - \sum_{i=1}^{n} \alpha_i [(c_i \langle w \cdot \mathbf{d}_i \rangle) - 1]$$

Hierfür sollen die Parameter für w und b minimiert und die Lagrange-Multiplikatoren α_i maximiert werden. Durch die Formulierung der Bedingungen für die partiellen Ableitungen, erhalten wir folgende Eigenschaften für die Berechnung eines Normalenvektors:

$$\frac{\partial}{\partial b} L(w, b, \alpha) \stackrel{!}{=} 0 \quad \text{und} \quad \frac{\partial}{\partial w} L(w, b, \alpha) \stackrel{!}{=} 0$$
$$\sum_{i=1}^{|D'|} \alpha_i c_i = 0 \quad \text{und} \quad w = \sum_{i=1}^{|D'|} \alpha_i c_i \mathbf{d}_i$$

Durch Einsetzen und Umformen der Lagrangefunktion erhält man das sog. duale Optimierungsproblem $L_{\mathcal{D}}$. Die Lösung ergibt die optimierten Lagrange-Multiplikatoren α_i, wodurch der Normalenvektor w bestimmt werden kann:

$$L_{\mathcal{D}}(\alpha) = \sum_{i=1}^{|D'|} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{|D|} c_i c_j \alpha_i \alpha_j \langle \mathbf{d}_i \cdot \mathbf{d}_j \rangle \text{ unter Bedingung}$$
$$\alpha_i > 0 \text{ und } \sum_{i=1}^{'D'} \alpha_i c_i = 0$$

Für die Lage der Hyperebene sind nur die Stützvektoren relevant, da für alle anderen Vektoren die Multiplikatoren α_i Null sind. Für die finale maximale Trennspanne mit der Menge an Stützvektoren S gilt:

$$\langle w \cdot \mathbf{d} \rangle + b = \sum_{i \in S} c_i \alpha_i \langle \mathbf{d}_i \cdot \mathbf{d}_i \rangle + b = 0$$

wobei zur Klassifikation eines Dokuments $\mathbf{x} \in \mathbf{D'}$ folgende Entscheidungsfunktion angewendet wird:

$$c_{SVM} = \text{sign}(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b) = \text{sign}\left(\sum_{i \in S} c_i \alpha_i \langle \mathbf{d}_i \cdot \mathbf{x} \rangle \right)$$

Bislang wurde gezeigt, dass eine Separierung der Daten durch eine lineare Funktion erfolgt. In den meisten Fällen kann jedoch eine gute Trennung nur durch eine nichtlineare Funktion erreicht werden [Bin07]. Um den Ansatz weiter verwenden zu können, wird ein Vektor $\mathbf{d}$ durch eine nicht lineare Abbildungsfunktion ϕ in einen

höher dimensionalen Raum F projiziert, in dem die lineare Trennung wieder möglich ist.

$$\phi : A \to F \ , \ \mathbf{d} \mapsto \phi(\mathbf{d})$$

Die Dimension wird ausgehend von der Anzahl der Attributen vergrößert und kann abhängig von der Abbildungsfunktion enorm anwachsen. Jedoch kann auf die explizite Verwendung der projizierten Vektoren verzichtet werden, wenn eine sog. Kernelfunktion K verwendet wird, die sich wie ein Skalarprodukt in F verhält. Die Skalarprodukte werden durch Kernel-Funktion ersetzt.

$$K(\mathbf{d}_i, \mathbf{d}_j) = \langle \phi(\mathbf{d}_i) \cdot \phi(\mathbf{d}_j) \rangle$$

Um die Stützvektormethode für ein Problem mit mehr als zwei Klassen anwenden zu können, werden einerseits Maßnahmen verwendet die das Problem zu einer Kombination von mehreren Zwei-Klassen-Problemen umformen, oder es wird versucht einen direkten Lösungsansatz zu finden. Beide Ansätze werden in dieser Arbeit nicht weiter verfolgt.

4 Evaluierung

In diesem Kapitel wird die Evaluierung der vorgestellten Modelle beschrieben. Die Gliederung erfolgt in drei Abschnitten. Zunächst werden die verwendeten Textkorpora aufgeführt. Abschnitt 4.2 und 4.3 beschreiben die Experimente und Ergebnisse für das ESA-Modell und das modifizierte Projektionsmodell hinsichtlich einer spezifischen Fragestellung:

Für das ESA-Modell wird untersucht, ob die Berechnung von Stilähnlichkeitsvektoren die Gemeinsamkeiten und Unterschiede der Schreibstile von Autoren erhöht und dadurch eine genauere Zuordnung und Unterscheidung von Schreibstilen im Vergleich zum Vektorraummodell möglich ist. Die Evaluierung der Modelle erfolgt durch Klassifikationsmethoden, die in Kapitel 3 vorgestellt wurden. Dazu werden drei Klassifikatoren verwendet, die jeweils mit den modellspezifischen Repräsentationen von Schreibstilen trainiert werden. Die Evaluierung erfolgt durch die Gegenüberstellung der erzielten Ergebnisse von Klassifikatoren für beide Modelle. Dabei werden die folgenden Parameter hinsichtlich der Auswirkung auf den Recall der Klassifikation untersucht:

- Variation der Auswahlmengen von Character-n-Grammen
- Variation der Repräsentation
- Variation der Indexgröße und der Indexkollektion im ESA-Modell
- Variation von Textlängen

Für das Projektionsmodell wird untersucht, wie sich die Fusionierungsmethoden auf den Recall und die Precision auswirken. Zusätzlich werden Character-n-Gramm-Repräsentationen hinsichtlich deren Eignung für den Schreibstilvergleich im Projektionsmodell untersucht:

- Variation der Auswahlmengen von Character-n-Grammen
- Variation der Länge von Character-n-Grammen

4.1 Korpora

Zur Evaluierung der vorgestellten Modelle werden vier Korpora verwendet. Jeder enthält eine Zusammenstellung aus englischen Texten von unterschiedlichen Quellen. Alle Texte sind im Web frei verfügbar und können heruntergeladen werden.

Der Koppel-Buchkorpus besteht aus Werken von neun Autoren, wobei jeder Autor mit je zwei Büchern vertreten ist. Die Autoren und deren Bücher wurden nach Angaben von Koppel u. a. in [KSA09] und [KS04] zusammengestellt. Die Bücher sind im 19. und 20. Jahrhundert entstanden und bilden ein breites Spektrum an literarischen Genres ab. Darunter befinden sich auch Werke der Bronte Schwestern, die in mehreren Studien verwendet wurden, unter anderem in [Gam04] und [KPCT03].

Gruppe	Autor	Buch	Wortanzahl
Amerikanische Schriftsteller	Hawthorne	Dr. Grimshawe's Secret	9232
		House of Seven Gables	105126
	Melville	Moby Dick	218639
		Redburn	120539
	Cooper	The Last of the Mohicans	148650
		The Spy	148189
Amerikanische Essayisten	Thoreau	Walden	117194
		A Week on Concord	116820
	Emerson	English Traits	65510
		Conduct Of Life	68124
Britische Theater-autoren	Shaw	Pygmalion	34939
		Getting Married	58968
	Wilde	An Ideal Husband	34301
		Woman of No Importance	23708
Bronte Schwestern	Anne	Tenant Of Wildfell Hall	168782
		Agnes Grey	69285
	Charlotte	Jane Eyre	188309
		The Professor	89376

Tabelle 4.1: Die Tabelle zeigt die verwendeten Bücher.

Der Gutenberg-Korpus wurde für diese Arbeit zusammengestellt und umfasst 7085 Texte von Project Gutenberg[5]. Die Anzahl an Texten entspricht dabei der Anzahl an Autoren. Die Texte bestehen mindestens aus 3.000 Wörtern.

[5]Webseite Project Gutenberg `http://www.gutenberg.org/wiki/Main_Page`

Die Wikipedia-Indexkollektion ist ein sog. XML-Dump. Wikipedia stellt unterschiedliche offline-Kollektionen von Wikipediaartikeln im XML-Format zur Verfügung. Die Artikelsammlung umfasst 2 Millionen Artikel aus dem Jahr 2008[6].

Die Gutenberg-Indexkollektion ist eine Zusammenstellung aus 10.000 Büchern von Project Gutenberg. Hier sind Autoren auch mehrfach mit verschieden Werken vertreten. Die Textlänge wurde nicht eingeschränkt.

4.2 Evaluierung des ESA Modells

Um die Experimente vergleichbar durchzuführen, orientiert sich die Auswahl der Klassifikatoren und die Stilrepräsentation im Vektorraummodell an den Experimenten von Koppel u. a. in [KSA09]. Im Vektorraummodell werden Schreibstile durch Character-3-Gramm-Vektoren repräsentiert und im ESA-Modell durch Stilähnlichkeitsvektoren. Die Klassifikatoren verwenden die Repräsentationen als Merkmalsvektoren. Die verwendeten Klassifikatoren sind Naive Bayes, J48 (Entscheidungsbaum) und SMO (Stützvektormethode) aus dem framework Weka[7]. Dabei sind die Parametereinstellungen der Klassifikatoren für alle Experimente identisch[8].

Zur Klassifikation besteht die Trainingsmenge D aus Textabschnitten gleicher Länge des jeweils ersten Buches der Autoren aus dem Koppel-Buchkorpus. Jeder Textabschnitt entspricht einem Dokument $d \in D$. Die Länge eines Abschnitts wird durch die Anzahl an Wörtern $|w|$ bestimmt. Für die Testmenge D' wird das gleiche Verfahren zur Aufteilung der restlichen Bücher verwendet. Die Basislinie für jedes Experiment wird durch die Autorenanzahl mit $1 : 9$ vorgegeben. Die Experimente zeigen die erzielten Recall-Werte der Klassifikatoren. Der Recall gibt in diesen Experimenten an, wie oft der richtige Autor eines Dokuments aus der Testmenge im Verhältnis zur Anzahl aller Dokumente aus der Trainings-und Testmenge erkannt wird.

4.2.1 Vokabular und Termgewichtung

In den Experimenten werden die Mengen von Character-3-Grammen, die das Vokabular abbilden, variiert und die Unterschiede zwischen den Termgewichtungsmaßen tf und $tfidf$ als Stilrepräsentation untersucht.

[6] XML-Dump enwiki-20080312
`http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Database_analysis/enwiki-20080312`

[7] machine learning tool Weka3`http://www.cs.waikato.ac.nz/ml/weka/`

[8] weka.classifiers.trees.J48-U-M2, weka.classifiers.bayes.NaiveBayes
weka.classifiers.functions.SMO-C1.0-L0.001-P1.0E-12-N0-V-1-W1-K

Jedes Dokument $d \in D$ besteht aus $|w| = 500$, jedes Dokument $d \in D'$ besteht aus $|w| = 1.500$ Wörtern. Die Anzahl der Dokumente in der Trainingsmenge beträgt $|D| = 2.000$. Die Testmenge besteht aus $|D'| = 1.500$ Dokumenten. Das Vokabular V beinhaltet die 10.000 häufigsten Character-3-Gramme aus D. Das Vokabular V' wird nach Angaben von Koppel u. a. in [KSA09] bestimmt, indem aus den 10.000 häufigsten Character-3-Grammen 1.000 mit dem höchste information gain[9] ausgewählt werden. Der ESA-Index besteht aus 10.000 Artikeln der Wikipedia-Indexkollektion.

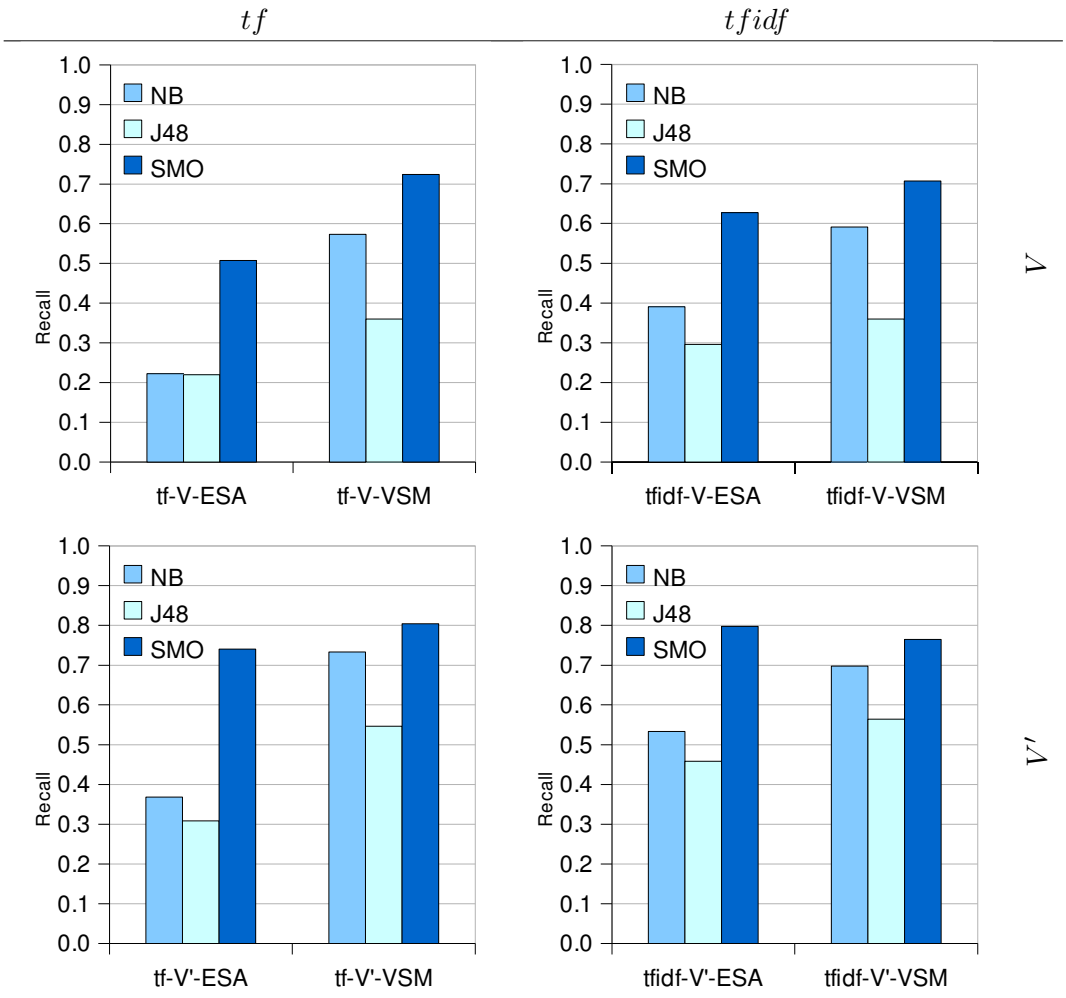


Tabelle 4.2: Ergebnisse der Klassifikatoren. Eine Zeile zeigt zwei Grafiken mit den ermittelten Recall-Werten bezüglich unterschiedlichen Stilrepräsentationen anhand eines Vokabulars. Eine Spalte zeigt zwei Grafiken mit Recall-Werten bezüglich einer Stilrepräsentation anhand verschiedener Vokabularien.

[9] weka.attributeSelection.InfoGainAttributeEval

Für beide Repräsentationsarten ist eine Steigerung der Recall-Werte je Klassifikator durch die zusätzliche Selektion der Character-3-Gramme erkennbar. Die besten Werte werden durch die Repräsentation (tf-V'-VSM), mit $73,3\,\%$, $54.65\,\%$ und $80,44\,\%$, je Klassifikator erzielt und entsprechen den Ergebnissen von Koppel. Die Überlegenheit der Stützvektormethode zeigt sich vor allem darin, dass mit hochdimensionalen Merkmalsvektoren aufgrund von V der Recall weniger stark fällt. Im Vergleich dazu sind die Klassifikatoren NB und J48 nicht geeignet, mit hochdimensionalen Vektoren vorteilhaft umzugehen. Der SMO Klassifikator kann mit großen Merkmalsvektoren deutlich besser umgehen und produziert weniger Fehler in der Zuordnung und bildet somit die beste Klassifikationsmethode in allen Experimenten ab. Im Hinblick auf das ESA-Modell sind die beiden Modelle somit nur mit dem SMO Klassifikator vergleichbar, da die ESA-Repräsentation aus 10.000 Stilähnlichkeiten besteht.

Im Modellvergleich sind Stilähnlichkeitsvektoren zur Klassifikation der Autoren deutlich schlechter geeignet als die Stilrepräsentationen im VSM. Bei fast allen Parametervariationen können weniger Autoren zugeordnet werden. Die Ausnahme bildet $tfidf$-V'-ESA. Dabei liegt die größte ermittelte Genauigkeit für Stilähnlichkeitsvektoren durch den SMO Klassifikator bei $79.79\,\%$. Dieses Ergebnis ist somit die beste Annäherung an die Ergebnisse von Koppel.

Die Ergebnisse bezüglich der Gewichtungen sind überraschend. Für Character-3-Gramm-Vektoren geht hervor, dass durch tf die Schreibstile besser repräsentiert werden. Wenn diese den Schreibstil genauer erfassen, dann sollten auch in den Dimensionen der Stilänlichkeitsvektoren die Referenzierung bezüglich ähnlicher und unähnlicher Schreibstile zu Wikipediaartikeln genauer erfolgen. Die Stilähnlichkeitsvektoren sind für die Klassifikatoren jedoch durch $tfidf$ besser zu unterscheiden. Im anschließenden Experiment wird die Verteilung der Stilähnlichkeiten zu der Indexkollektion durch das VSM direkt bestimmt - mit den selben Parametervariationen wie in den vorherigen Experimenten.

Für das ursprüngliche ESA-Modell ist bekannt, dass bei Indexdokumenten im Vergleich zu den zu untersuchenden Texten überwiegend Unähnlichkeiten auftreten. Nur ein geringer Anteil referenziert einen Text mit einer hohen inhaltlichen Ähnlichkeit. Für die Stilähnlichkeiten zu Wikipediaartikeln sollte die Verteilung nach Stilähnlichkeit die selben Charakteristiken aufzeigen. In diesem Zusammenhang werden die Stilähnlichkeiten eines Dokuments zu allen Wikipediaartikeln aus dem Index mit Character-3-Gramm-Vektoren durch die Kosinusähnlichkeit berechnet. In den Experimenten wird bestimmt, wieviel Indexdokumente eine Stilähnlichkeit aus den Intervallen $i = \{[0.0, 0.1], \ldots, [0.9, 1.0]\}$, zu einem Dokument erzielen. Der Vorgang wird für alle Dokumente wiederholt und die Ergebnisse gemittelt.

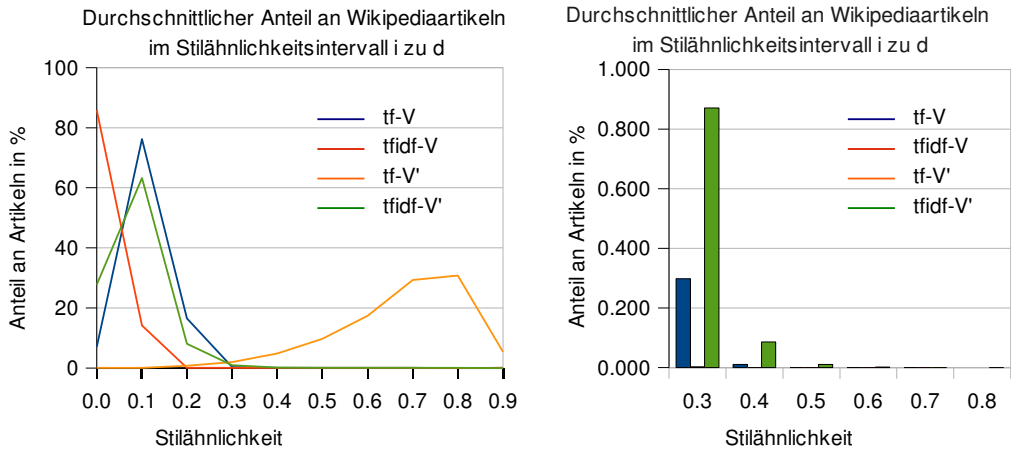


Tabelle 4.3: Die linke Grafik zeigt den durchschnittlichen Anteil an Wikipediaartikeln in einem Stilähnlichkeitsintervall. In der rechten Grafik wurde die Skalierung hinsichtlich der Dokumentmenge angepasst.

In der Grafik ist sichtbar, dass die ermittelten Ähnlichkeiten von den Gewichtungen und den Vokabularien V und V' abhängig sind. Dabei entsprechen alle Ergebnisse, bis auf tf, V' den Erwartungen, indem der überwiegende Teil von Wikipediaartikeln keine hohe Stilähnlichkeiten zu einem Dokument aufzeigen.

Problematisch ist, dass für die Repräsentationen tf-V und $tfidf$-V sowie $tfidf$-V' so gut wie keine hohen Stilähnlichkeiten zu den Artikeln gemessen wird. Hier wird ein Zusammenhang bezüglich der schlechten Klassifikationsergebnisse durch Stilähnlichkeitsvektoren vermutet. Die Vektorkomponenten eines Stilähnlchkeitsvektors könnten dadurch nur unähnliche Schreibstile enthalten, wobei sich diese Werte gleichzeitig kaum unterscheiden. Die rechte Grafik in Tabelle 4.3 zeigt, dass im Durchschnitt der Schreibstil von ca. 80 aus 100 Artikeln im Ähnlichkeitsintervall $[0.2, 0.3]$ liegen. Unter diesen Umständen sind Stilähnlichkeitsvektoren durch einen Klassifikator viel schwieriger zu unterscheiden als Repräsentationen im VSM.

Für die Repräsentation $tfidf$-V' wurden mehr Artikel gefunden, die im Schreibstil etwas ähnlicher sind. Dadurch könnten sich die Vektorkomponenten untereinander mehr unterscheiden und wären für die Klassifikatoren geeigneter.

Das Ergebnis bezüglich der Repräsentation tf-V' verhält sich gegenläufig zu den anderen. Es kann beobachtet werden, dass der überwiegende Teil der Artikel extrem hohe Stilähnlichkeiten aufzeigt. Daraus wird angenommen dass die Stilähnlichkeitsvektoren mit dieser Repräsentation ebenfalls kaum zu unterscheiden sind, da die Artikel bezüglich des Schreibstils insgesamt viel zu ähnlich bewertet werden.

Zusammengefasst zeigen die Ergebnisse, dass Wikipediaartikel entweder viel zu hohe oder viel zu geringe Stilähnlichkeiten zu Dokumenten der Gutenberg-Textkollektion erziehlen. Dadurch könnten Wikipediaartikel auch als ungeeignete Textbasis für die Konstruktion von Stilähnlichkeitsvektoren sein, unter der Annahme, dass sich die Artikel selbst im Schreibstil kaum unterscheiden.

4.2.2 Indexvariationen und Ensemble Entscheidung

Im folgenden werden zunächst Variationen zu internen Parametern des ESA-Modells untersucht. Dabei wird die Indexkollektion des ESA-Modells ausgetauscht und danach der Dokumentraum des Indexes erhöht, indem mehr Dokumente verwendet werden. Die Ergebnisse werden dann gegenübergestellt. In einem weiteren Experiment wird die Autoschaft durch eine Ensemble Entscheidung bestimmt, die anhand von Teilergebnissen aus Klassifikationen getroffen wird.

Für die Indexkollektion des ESA-Modells wird untersucht, ob die Gutenberg-Indexkollektion eine geeignetere Textbasis im Vergleich zur Wikipedia-Indexkollektion darstellt. Diese Kollektion könnte mehr Texte enthalten, die sich stärker voneinander abgrenzen. Wikipediaartikel könnten aus mehreren Gründen zueinander stilistisch sehr ähnlich sein. Artikel werden oft von vielen Autoren mehrmals nachgebessert, dadurch könnte der Schreibstil stark durch die Kollaborationen geprägt sein stellt somit nicht das Werk eines Autors, sondern vieler Autoren dar. Artikel könnten auch dadurch im Stil sehr ähnlich sein, da sie alle das selbe Ziel verfolgen. Die Artikel liefern Wikipediabeiträge, in denen z.B ein Sachverhalt erklärt wird, wobei die Form und der Stil sich durch dieses Ziel ergibt. Für Bücher des Gutenberg-Korpus sind diese beide Faktoren nicht gegeben.

Ein sehr wichtiger Parameter des ESA-Modells ist die Anzahl an Dokumenten des Indexes, die den Dokumentraum aufspannen. Für die Anzahl an Indexdokumenten wird angenommen, dass eine Vergrößerung des Dokumentraums von Vorteil ist. Dadurch kann mehr stilistische Information über ähnliche und unähnliche Dokumente in die Stilähnlichkeitsvektoren aufgenommen werden.

Die ESA-Indexkollektion wird im ersten Experiment durch die Projekt Gutenberg Indexkollektion ersetzt, wobei die Dimension des Dokumentraums aus 10.000 Dokumenten nicht verändert wird. Im zweiten Experiment wird die Dimesion der ESA-Indexkollektion auf 100.000 Dokumente erhöht, wobei der ESA-Index wieder durch Wikipediaartikel abgebildet werden. Die Aufteilung und Zusammensetzung der Trainings- und Testmenge aus Abschnitt 4.2.1 wird beibehalten.

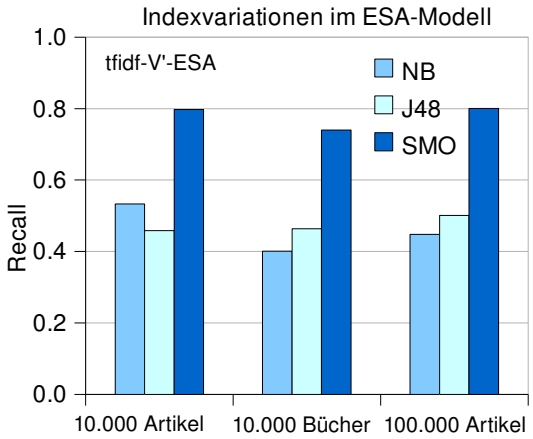


Abbildung 4.1: Indexvariationen

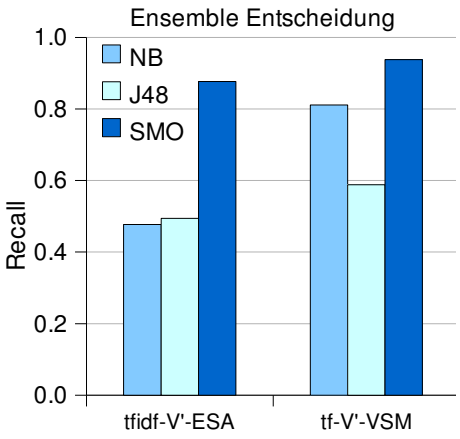


Abbildung 4.2: Ensemble Entscheidung

Abbildung 4.1 zeigt die Grafik bezüglich der erzielten Recall-Werte der Klassifikatoren für veränderte Parameter der ESA-Indexkollektion.

Es kann beobachtet werden, dass der Austausch der Indexkollektion keinen positiven Effekt auf die erzielten Recall-Werte der Klassifikatoren hat. Durch die Verwendung von Büchern als ESA-Indexkollektion werden sogar weniger Autoren richtig zugeordnet. Dadurch deutet sich eine Abhängigkeit zur Indexkollektion an. Der anfangs vermutete Zusammenhang wird jedoch nicht bestätigt.

Die Vergrößerung des Dokumentraumes ist ebenfalls keine günstige Parametereinstellung für das ESA-Modell. Der SMO Klassifikator steigert sich unmerklich von 80.0 % auf 81.0 % um 1.0125 %. In Anbetracht des Aufwands muss der kleine Vorteil weiter relativiert werden. Für einen Stilähnlichkeitsvektor müssen Stilähnlichkeiten zu der 10-fachen Anzahl an Indexdokumenten berechnet werden. Die Komplexität ist in $O(N^2)$, mit jeweils N Autoren in der Trainings- und Testmenge. Damit wird die Anzahl an 100.000 Artikeln nicht als eine günstige Parameterwahl bewertet.

In den Experimenten zur Ensemble Entscheidung wird untersucht, ob der überwiegende Teil von Schreibstilen aus Dokumenten eines Autors diesem Autor zugeordnet werden können. Die Aufteilung der Testmenge bezieht sich auf ein Ensemble von Dokumenten eines Autors. Dazu wird die Testmenge wieder durch die Aufteilung der jeweils zweiten Bücher der Autoren bestimmt.

Zur Klassifikation wir das zweite Buch aus dem Koppel-Buchkorpus eines jeden Autors in ein Teilbuch zerlegt. Ein Teilbuch ist dabei ein zufällig gewählter Abschnitt mit 10.000 Wörtern. Jedes Teilbuch wird in 20 Abschnitte aufgeteilt mit je 500

Worten. Diese Abschnitte werden klassifiziert. Die Entscheidung für die Autorschaft erfolgt für jedes Teilbuch durch eine Mehrheitsentscheidung der einzelnen Ergebnisse bezüglich der Autorenzuordnung. Die Zusammensetzung der Trainingsmenge wird nicht verändert und enthält 500 Worte je Dokument.

Abbildung 4.2, zeigt die Recall-Werte der der Klassifikatoren bezüglich einer Ensemble Entscheidung für je ein Teilbuch von neun Autoren. Die Stilrepräsentationen im VSM ist für die Klassifikatoren im Vergleich zu Stilähnlichkeitsvektoren des ESA-Modells wieder besser geeignet. Die Ergebnisse sind nicht überraschend, da die Genauigkeiten in den Zuordnungen aller Dokumente, die in den Experimenten in 4.2.1 für die Repräsentation im VSM ermittelt wurden, höher sind. Dadurch werden auch Dokumente eines Autors bei einer Ensemble Entscheidung öfters richtig zugeordnet und erzeugen so höhere Genauigkeiten in der Zuordnung von Teilbüchern eines Autors.

4.2.3 Textlänge

In diesen Experimenten wird der Zusammenhang zwischen Stilähnlichkeiten und der Länge von Texten untersucht. Bisher wurde die Textlänge durch die Experimente in [KSA09] mit $|w| = 500$ Worten pro Dokument festgelegt. Bei steigender Textlänge werden höhere Recall-Werte der Klassifikatoren für beide Repräsentationsarten angenommen, da die Häufigkeitsverteilungen von Character-3-Grammen je Autor größere Unterschiede aufweisen sollten. Stilähnlichkeitsvektoren könnten dadurch besonders profitieren, indem die Stilähnlichkeiten zu Wikipediaartikeln genauer erfasst werden. Dadurch könnten Unterschiede und Gemeinsamkeiten durch die Vektorkomponenten genauer abgebildet werden.

Die Zusammensetzung der Trainingsmenge wird nicht verändert und enthält $|w| = 500$ Worte je Dokument. Die Testmenge wird dreimal neu zusammengestellt und enthält jeweils Dokumente mit $|w| = 500$, $|w| = 2.000$ und $|w| = 5.000$ Worten. Die Klassifikatoren klassifizieren die Testmengen für die Stilrepräsentaionen im Vektorraummodell und im ESA-Modell.

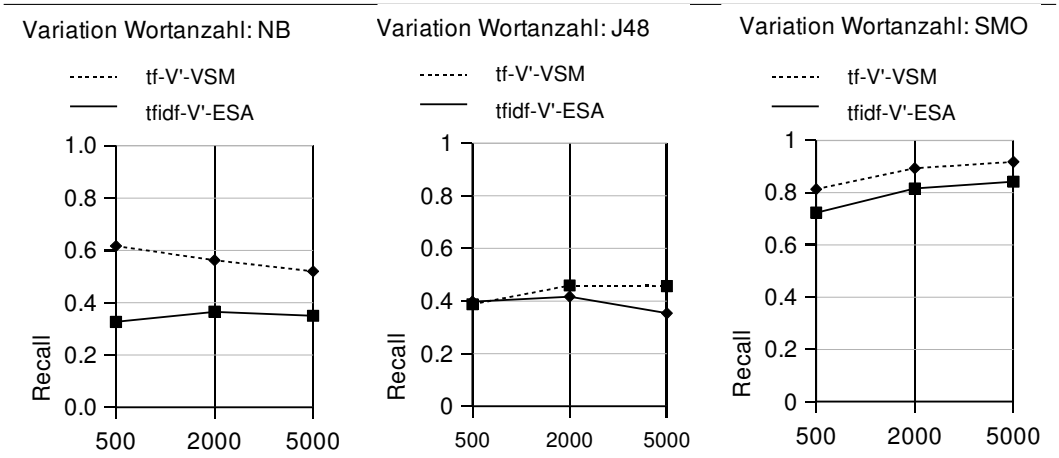


Tabelle 4.4: Die Tabelle zeigt die Ergebnisse je Klassifikator in drei Grafiken. Die Recall-Werte werden abhängig von der Textlänge in der Form von parallelen Koordinaten gegenübergestellt. Jede Grafik zeigt auf der x-Achse die Anzahl an Wörtern je Dokument aus der Testmenge. Die y-Achsen zeigen die erzielten Recall-Werte der Klassifikatoren.

Es ist zu beobachten, dass längere Texte in der Testmenge nicht generell zu bessern Ergebnissen führen. Der SMO Klassifikator bildet die Ausnahme. Bei steigender Textlänge können durch beiden Repräsentationsarten mehr Autoren richtig zugeordnet werden. Durch Stilähnlichkeitsvektoren aus längeren Texten werden im Vergleich zu den Repräsentationen im VSM weniger Autoren richtig zugeordnet. Ungewöhnlich ist, dass die anderen beiden Klassifikatoren sehr unterschiedlich auf die Repräsentationsarten und die Textlängen reagieren. Wenn sich die Länge der Dokumente der Trainingsmenge von der Länge der Dokumente aus der Testmenge zu stark unterscheiden, kann der Recall nicht erhöht werden.

4.3 Evaluierung des Projektionsmodells

Zur Evaluierung des Projektionsmodells hinsichtlich der Fusionierungsmethoden werden zunächst die Parameter vorgestellt, die für alle Experimente identisch sind.

Als Dokumentkollektion wird der Gutenberg-Korpus verwendet. Jeder Text wird in ein Dokument x und ein Dokument d zerlegt. Die Dokumente enthalten nicht überlappende Abschnitte des zerlegten Textes. Dabei besteht x aus 500 und d aus 1.500 Wörtern. Die Menge X enthält 1.000 Dokumente, wobei für jedes Dokument $x \in X$ die Autorschaft bestimmt werden soll. Die Menge D enthält 7.085 Dokumente,

wobei jedes Dokument $d \in D$ durch das Verfahren des Projektionsmodells mit x verglichen wird.

Koppel u. a. schlagen vor, 100 Iterationen für den Stilvergleich von Dokumenten durchzuführen. Je Fusionierungsmethode wird der Fusionswert $fx(d)$ bezüglich x aus den 100 Rankings berechnet. Der Recall und die Precision werden anhand eines Schwellenwertes σ berechnet. Dabei werden die Schwellen variiert. Für jede Schwelle σ mit $\sigma = \{0, \ldots, 100\}$ werden Precision und Recall ermittelt. Die Werte ergeben eine spezifische Recall-Precision-Kurve je Fusionierungsmethode. Für den Hybrid werden die Gewichtungsfaktoren α und β mit $\alpha, \beta = 0.5$ zur Berechnung des Durchschnitts verwendet.

Weiter wird untersucht, welche Länge der Character-n-Gramme sich als vorteilhaft erweist, um den Schreibstil zu repräsentieren. Dazu wird zusätzlich das Vokabular V der verwendeten Character-n-Grammen variiert, wobei die Anzahl an zufälligen Character-n-Grammen V' mit $|V'| = 10.000$ nicht verändert wird.

4.3.1 Fusionierungsmethoden

In den Experimenten wird die Zusammenfassung der Rankings durch die score fusion, die rank fusion unde den Hybrid untersucht. Dabei werden unterschiedliche Stilrepräsentationen verwendet, die sich wie folgt unterscheiden:

- C3G: V beinhaltet alle Character-3-Gramme
- C4G: V beinhaltet alle Character-4-Gramme
- C4G-100T: V beinhaltet die 100.000 häufigsten Character-4-Gramme
- C5G-100T: V beinhaltet die 100.000 häufigsten Character-5-Gramme

In vorangegangenen Experimenten wurde das Vokabular auch aus allen Character-5-Grammen des Korpus gebildet. Die Versuche haben gezeigt, dass dadurch so gut wie keine Autoren richtig zugeordnet werden konnten. Bei Character-3-Grammen ist eine Reduktion nicht nötig, da die Anzahl an verschiedenen Character-3-Grammen unter 100.000 liegt. Aus dem Gutenberg-Korpus wurde folgende Anzahlen an verschiedenen Character-n-Grammen gemessen.

- Character-3-Gramme: 95.330
- Character-4-Gramme: 521.096
- Character-5-Gramme: 1.934.831

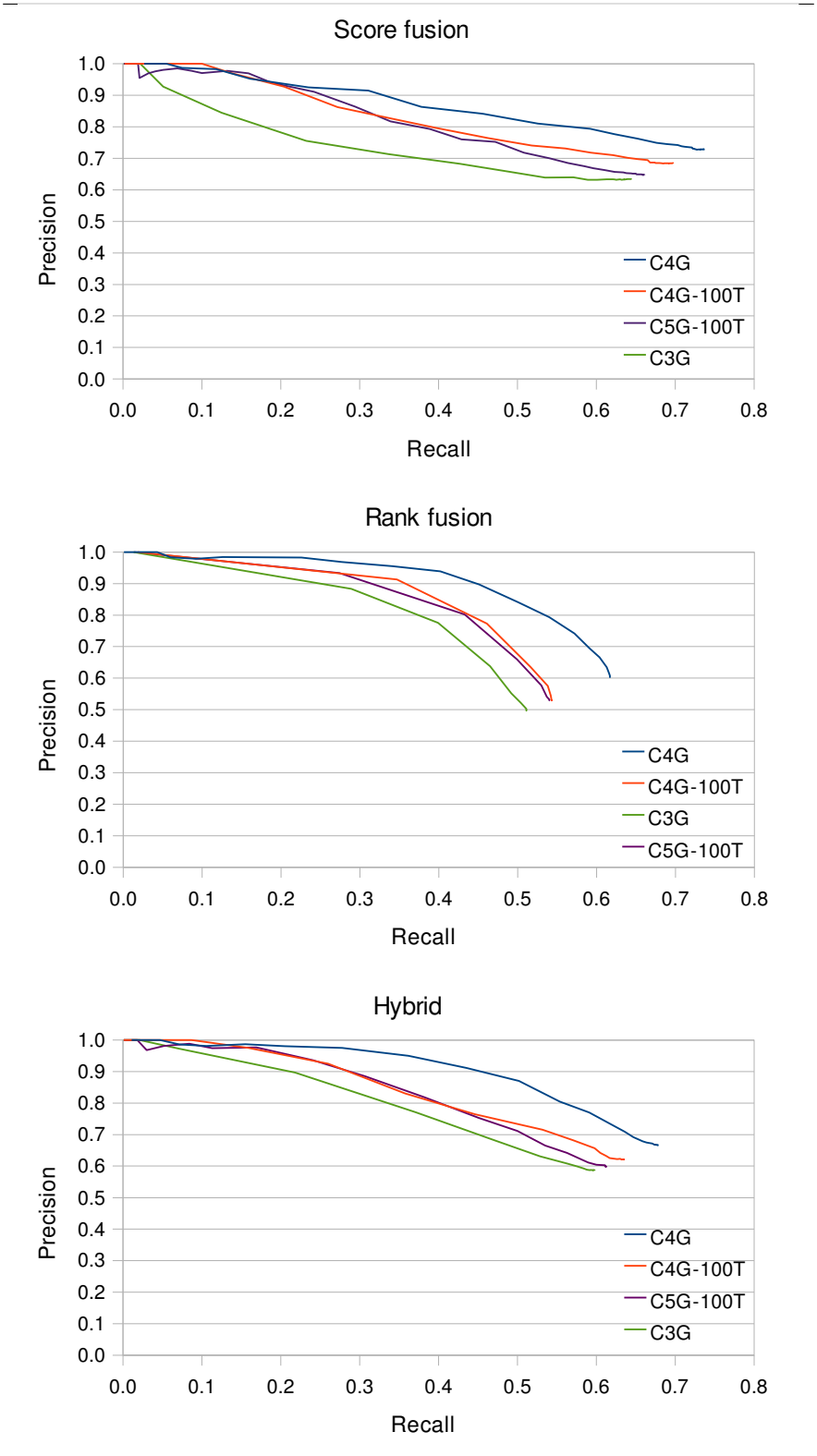
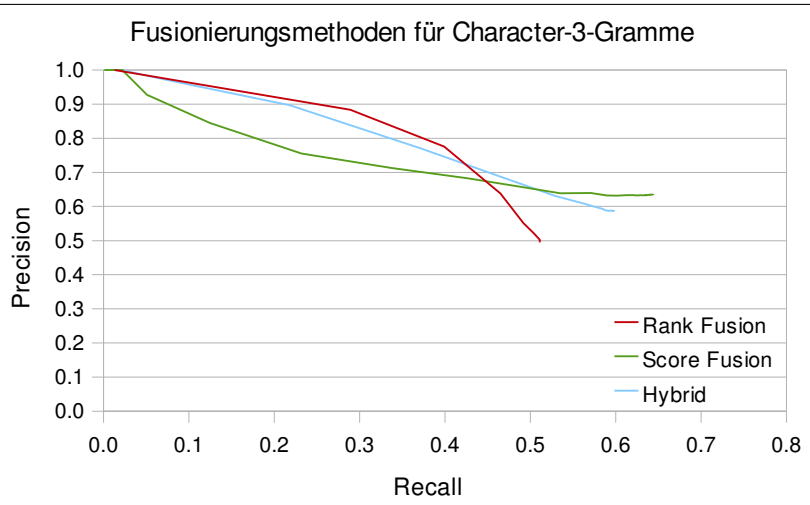
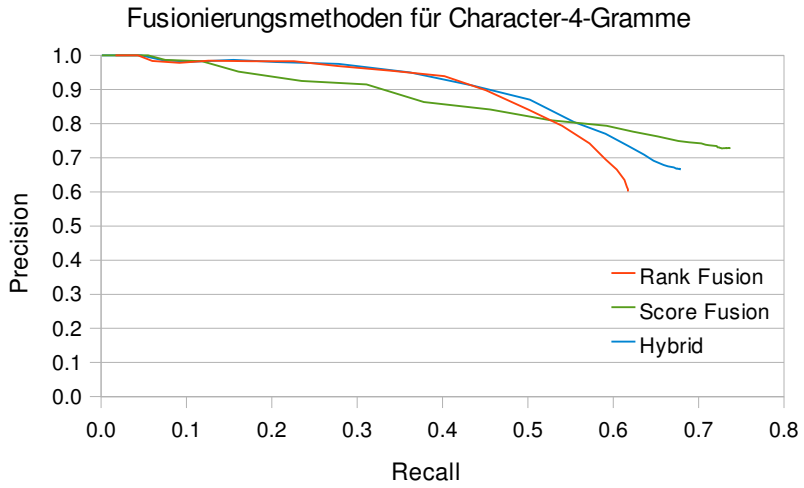


Tabelle 4.5: Die Tabelle zeigt die Ergebnisse zu drei Fusionierungsmethoden. Für jede Methode sind vier Recall-Precision-Kurven abgebildet. Die Kurven zeigen die Ergebnisse bei einer Variation der Schreibstilrepräsentation.





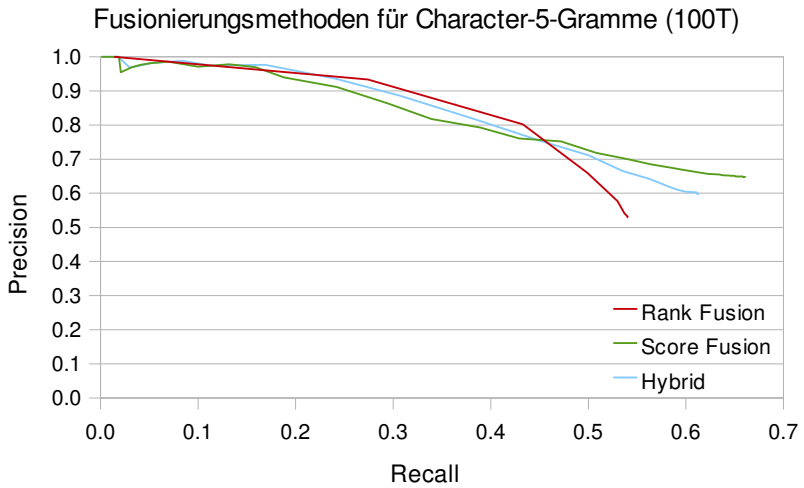


Tabelle 4.6: Die Tabelle zeigt die Ergebnisse zu drei Stilrepräsentationen. Für jede Repräsentation sind drei Fusionierungsmethoden durch Recall-Precision-Kurven abgebildet.

Die Ergebnisse zeigen, dass Character-4-Gramme zur stilistischen Beschreibung und Unterscheidung von tausenden Autoren am besten geeignet sind. Dabei ist zu beobachten, dass eine deutliche Verbesserung zu Character-5-Grammen nur dann erfolgt, wenn die Repräsentation durch die zufällige Auswahl aus allen Character-4-Grammen je Iterationsschritt stattfindet. Wenn jedoch die Anzahl an verschiedenen Character-4-Grammen, z.B. durch die Verwendung einer anderen Textkollektion, stark anwächst, wird vermutet, dass die Reduktion der Character-4-Gramme wieder sinnvoll sein könnte. Eine Möglichkeit, den Selektionsschritt zu vermeiden, könnte dadurch entstehen, den Zeichensatz für Character-n-Gramme einzuschränken. Es ist überraschend, dass durch Character-3-Gramme in diesen Experimenten die wenigsten Autoren richtig zugeordnet werden. Dies könnte ein Hinweis darauf sein, dass Character-3-Gramme bessere Ergebnisse liefern, wenn die Autorenanzahl und damit die Dokumentmenge deutlich kleiner ist. Im folgenden werden die Ergebnisse zu den Fusionierungsmethoden anhand der Character-4-Gramm-Repräsentationen (C4G) weiter besprochen.

Die Ergebnisse bezüglich der Ähnlichkeitsfusionierung zeigen, dass sich diese Methode günstig auf den maximalen Recall auswirkt. Wenn die Schwelle auf 0 gesetzt wird, können 70 % von 1.000 Autoren aus über 7.000 Autoren richtig zugeordnet werden. Ein Dokument, welches vom selben Autor verfasst wurde, erzielt häufiger den höchsten Fusionswert im fusionierten Ranking. Im Vergleich dazu werden mit der Methode der Rangfusionierung 60 % richtig zugeordnet. Für das Projektionsmodell steht jedoch die Sicherheit der Zuordnung im Vordergrund. In diesem Sinne ist die Rangfusionierung besser geeignet. Die Sicherheit wird durch diese Methode entscheidend erhöht. Wenn die Positionen im Ranking betrachtet werden, dann führt das zu weniger falschen Entscheidungen. Ein Dokument, welches vom selben Autor verfasst wurde und einen definierten Schwellenwert durch die Fusionierung der Positionen erfüllt und dabei den höchsten Fusionswert annimmt, wird dadurch mit einer höheren Sicherheit als ähnlichstes Dokument erkannt. Bei einem Schwellenwert von $\sigma = 90$ wird eine Precision von 95 % mit einem Recall von 40 % durch die Rangfusionierung erzielt.

Der Hybrid aus Ähnlichkeits-und Rangfusionierung zeigt bei Character-4-Grammen ebenfalls gute Precision-Recall-Werte mit $\alpha, \beta = 0.5$. Der Unterschied zur Rangfusionierung wird erst bei niedrigen Schwellenwerten für σ deutlicher, indem die Werte weniger stark abfallen. Der Vorteil dieser Methode ist, dass die höheren Recall-Werte der Ähnlichkeitsfusionierung einbezogen werden. Die Ergebnisse zu den anderen Repräsentationsarten zeigen jedoch, dass bei größeren Unterschieden der Werte zwischen Ähnlichkeits- und Rangfusionierung die Ergebnisse schlechter werden. Wenn

diese sich zu stark unterscheiden und somit die Kurven jeweils eine zu starke konvexe bzw. konkave Charakteristik aufweisen, wird lediglich der Durchschnitt zwischen den beiden Methoden gemessen. Der Hybrid sollte weiterhin mit unterschiedlichen Gewichtungen getestet werden, um eine klare Entscheidung zu treffen, ob diese Methode die Sicherheit der Zuordnung von Autoren eindeutig steigern kann.

5 Zusammenfassung und Ausblick

In dieser Arbeit wurden Verfahren zu dem Problem der Autorschaftsbestimmung vorgestellt. In der Autorschaftsbestimmung soll bestimmen werden, wer der Verfasser eines anonymen Textes ist. Der Einstiegspunkt in diesen Forschungsbereich ist die Schreibstilanalyse. Dabei wird der Schreibstil eines Autors durch statistisch quantifizierbare Merkmale definiert. Zu diesen Stilmerkmalen wurde zunächst eine Übersicht gegeben und eine Einteilung in sprachlichen Ebenen vorgenommen. Die Übersicht vermittelt dabei, dass es keine triviale Aufgabe ist, passende Stilmerkmale zu finden. Die Mehrheit der Studien haben gezeigt, dass Stilmerkmale aus der lexikalischen Ebene und im besonderen Character-n-Gramme aus der Zeichenebene Merkmale bieten, die den Stil eines Autors besser erfassen können. Retrieval-Modelle der Autorschaftsbestimmung verwenden die Stilmerkmale und Repräsentieren die Schreibstile modellabhängig für einen Vergleich.

Das ESA-Modell wurde in dieser Arbeit erstmalig als Stilähnlichkeitsmodell vorgestellt und untersucht. Dieser Ansatz wurde mit einer klassischen Methode, dem Vektorraummodell gegenübergestellt. Dabei verwenden beiden Modellen Character-3-Gramme um Schreibstile zu erfassen. Die modellspezifische Stilrepräsentationen wurden durch Klassifikatoren Evaluiert. Dabei zeigte sich die Überlegenheit des Vektorraummodells durch Character-3-Gramm-Repräsentationen in allen Experimenten. Es konnte kein Parameter ermittelt werden, der eine entscheidende Verbesserung oder zumindest eine Angleichung an die guten Ergebnisse der Character-3-Gramm-Repräsentationen bewirken konnte.

Die Ursache dafür konnte durch die Experimente nicht geklärt werden. Somit bleibt nur die Vermutung, dass die Vektorkomponenten von Stilähnlichkeitsvektoren homogene Stilähnlichkeiten zu Referenzdokumenten aufweisen.

Das vorgestellte Projektionsmodell ist ein interessanter Vorschlag um die Autorschaftsbestimmung auch auf tausende Autoren anzuwenden. Für das Modell wurde die Zusammenfassung der Rankings durch drei Methoden untersucht. Dabei hat sich herausgestellt, dass die Rangfusionierung besser geeignet ist um sichere Entscheidungen zu treffen. Für den Hybrid aus Ähnlichkeits- und und Rangfusionierung

trifft dies ebenfalls zu. Dazu sollten jedoch der Einfluss durch unterschiedliche Gewichtungsfaktoren näher untersucht werden um die Methode besser einschätzen zu können.

Literaturverzeichnis

[AC05] A. Abbasi and H. Chen. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75, 2005.

[Bin07] Liu Bing. *Web Data Mining.* Springer, Department of Computer Science University of Illinois at Chicago, 2007.

[Bur92] JF Burrows. Not Unles You Ask Nicely: The Interpretative Nexus Between Analysis and Information. *Literary and Linguistic Computing*, 7(2):91, 1992.

[Bur02] J. Burrows. 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267, 2002.

[DuB04] W.H. DuBay. The principles of readability. *Impact Information*, pages 1–76, 2004.

[DVACM01] O. De Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4):55–64, 2001.

[FH96] RS Forsyth and DI Holmes. Feature-finding for test classification. *Literary and Linguistic Computing*, 11(4):163, 1996.

[FSGK06] G. Frantzeskou, E. Stamatatos, S. Gritzalis, and S. Katsikas. Effective identification of source code authors using byte-level information. In *Proceedings of the 28th international conference on Software engineering*, page 896. ACM, 2006.

[Gam04] M. Gamon. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of COLING*, volume 4, page 611, 2004.

[GM07] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 6–12, 2007.

[Gri07] Jack Grieve. Quantitative Authorship Attribution: An Evaluation of Techniques. *Lit Linguist Computing*, 22(3):251–270, 2007.

[Hol94] D.I. Holmes. Authorship attribution. *Computers and the Humanities*, 28(2):87–106, 1994.

[KPCT03] V. Keselj, F. Peng, N. Cercone, and C. Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING*, volume 3, pages 255–264. Citeseer, 2003.

[KS04] M. Koppel and J. Schler. Authorship verification as a one-class classification problem. *CProceedings of the twenty-first international conference on Machine learning*, 2004.

[KSA09] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *J. Am. Soc. Inf. Sci. Technol.*, 60(1):9–26, 2009.

[KSA10] M. Koppel, J. Schler, and S. Argamon. Authorship attribution in the wild. *Language Resources and Evaluation*, pages 1–12, 2010.

[KSAM06] M. Koppel, J. Schler, S. Argamon, and E. Messeri. Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, page 660. ACM, 2006.

[KSBD07] M. Koppel, J. Schler, and E. Bonchek-Dokow. Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, 8:1261–1276, 2007.

[LD08] K. Luyckx and W. Daelemans. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 513–520. Association for Computational Linguistics, 2008.

[Lin91] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[MGL+05] D. Madigan, A. Genkin, D.D. Lewis, S. Argamon, D. Fradkin, and L. Ye. Author identification on the large scale. In *Proc. of the Meeting of the Classification Society of North America*. Citeseer, 2005.

[RS03] M.E. Renda and U. Straccia. Web metasearch: rank vs. score based rank aggregation methods. In *Proceedings of the 2003 ACM symposium on Applied computing*, page 846. ACM, 2003.

[Rud97] J. Rudman. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31(4):351–365, 1997.

[Seb02] F. Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[SKS07] Benno Stein, Moshe Koppel, and Efstathios Stamatatos. Plagiarism analysis, authorship identification, and near-duplicate detection pan'07. *SIGIR Forum*, 41(2):68–71, 2007.

[SM] G. Salton and MJ McGill. Introduction to modern information retrieval. 1983.

[SM07] Benno Stein and Sven Meyer zu Eissen. Intrinsic Plagiarism Analysis with Meta Learning. In Benno Stein, Moshe Koppel, and Efstathios Stamatatos, editors, *SIGIR Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07)*, pages 45–50. CEUR-WS.org, 2007.

[SP07] B. Stein and M. Potthast. *Construction of Compact Retrieval Models.* In: DOMINICH , S. und F. K ISS (Hrsg.): Studies in Theory of Information Retrieval, Foundation for Information Society, 2007.

[Sta06] E. Stamatatos. Authorship attribution based on feature set subspacing ensembles. *International Journal of Artificial Intelligence Tools*, 15(5):823–838, 2006.

[Sta09] E. Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.

[Ste] M.A.N.L.B. Stein. Evaluating Cross-Language Explicit Semantic Analysis and Cross Querying at TEL@ CLEF 2009.

[Ste08a] B. Stein. Maschinelles lernen : Unit. splitting, 2008. `http://www.uni-weimar.de/medien/webis/teaching/lecturenotes/machine-learning/unit-de-decision-trees-splitting.pdf` besucht im März 2010.

[Ste08b] B. Stein. Maschinelles lernen: bayesien learning, 2008. `http://www.uni-weimar.de/medien/webis/teaching/lecturenotes/machine-learning/unit-de-bayesian-learning.pdf` besucht im März 2010.

[Ste08c] B Stein. Unit. modelle und prozesse im ir, 2008. `http://www.uni-weimar.de/medien/webis/teaching/lecturenotes/`

`information-retrieval/unit-de-retrieval-models.pdf`besucht im Januar 2010.

[SWY75] G. Salton, A. Wong, and CS Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):620, 1975.

[ZZ05] Y. Zhao and J. Zobel. Effective and scalable authorship attribution using function words. *Lecture Notes in Computer Science*, 3689:174, 2005.

A Anhang

A.1 Implementierte Retrieval-Modelle

Der Java-Programmcode befindet sich auf einer beiliegenden CD.
In dem Ordner *authorship* befindet sich ein Eclipse-Projekt, welches den kompletten Programmcode zu den vorgestellten Retrieval-Modellen sowie den Programmcode zu den Experimenten enthält.