Intrinsische Plagiatserkennung und Autorenerkennung mittels Grammatikanalyse[1]

Michael Tschuggnall[2]

Abstract: Durch die hohe und ständig steigende Anzahl an frei verfügbaren Textdokumenten wird es immer leichter, Quellen für mögliche Plagiate zu finden, während es auf der anderen Seite für automatische Erkennungstools aufgrund der großen Datenmengen immer schwieriger wird, diese zu erkennen. In dieser Arbeit wurden verschiedene Algorithmen zur intrinsischen Plagiatserkennung entwickelt, welche ausschließlich das zu prüfende Dokument untersuchen und so das Problem umgehen, externe Daten heranziehen zu müssen. Dabei besteht die Grundidee darin, den Schreibstil von Autoren auf Basis der von ihnen verwendeten Grammatik zur Formulierung von Sätzen zu untersuchen, und diese Information zu nutzen, um syntaktisch auffällige Textfragmente zu identifizieren. Unter Verwendung einer ähnlichen Analyse wird diese Idee auch auf das Problem, Textdokumente automatisch Autoren zuzuordnen, angewendet. Darüber hinaus wird gezeigt, dass die verwendete Grammatik auch ein unterscheidbares Kriterium darstellt, um Informationen wie das Geschlecht und das Alter des Verfassers abzuschätzen. Schlussendlich werden die vorherigen Analysen und Resultate verwendet und so adaptiert, dass Anteile von verschiedene Autoren in einem gemeinschaftlich verfassten Text automatisch erkannt werden können.

1 Einführung

Durch die Entwicklungen im Bereich der elektronischen Datenverarbeitung und der Zugänglichkeit des World Wide Web steigt die Anzahl der öffentlich zugänglichen Textdokumente täglich. Neben Online-Bibliotheken wie z.B. Project Gutenberg[3], welche Millionen von elektronischen Büchern zum freien Download anbieten, wird Text auch massiv über soziale Medien verbreitet. Im Gegensatz zu letzteren, wo der Inhalt und die Autoren meist leicht klassifizierbar sind, stellt die unsachgemäße Wiederverwendung von Textfragmenten vor allem im akademischen Bereich ein ernsthaftes Problem dar. Die Erkennung solcher Plagiatsfälle kann mit recht einfachen Mitteln erfolgen, wenn ganze Textbausteine ohne oder mit nur geringfügiger Veränderung aus öffentlich zugänglichen und populären Quellen wie z.B. Wikipedia übernommen wurden. Andererseits ist es bereits wesentlich schwieriger, Plagiate zu erkennen, wenn der Originaltext stark umstrukturiert wurde oder auch wenn die Quelle gar nicht (elektronisch) verfügbar ist. Vor allem in letzteren Fällen ist eine dokumentinterne Analyse des Schreibstils unvermeidbar.

Während es für menschliche Leser oft leicht ist, Änderungen des Schreibstils zu identifizieren, ist dies für computerbasierte Algorithmen deutlich schwerer. Beispielsweise erkennen Betreuer von wissenschaftlichen Arbeiten Plagiate recht häufig, weil gewisse

[1] Englischer Titel der Dissertation: "Intrinsic Plagiarism Detection and Author Analysis By Utilizing Grammar"
[2] Institut für Informatik, Universität Innsbruck, michael.tschuggnall@uibk.ac.at
[3] Project Gutenberg, `http://www.gutenberg.org`, besucht im Februar 2015

Sätze oder Absätze *"anders"* sind und nicht ins Gesamtbild passen. Die hier meist intuitiv herangezogenen Kenngrößen dienen zum Erkennen von plötzlichen Stiländerungen und inkludieren Merkmale wie z.B. der Art der Verwendung von Vokabular, der (durchschnittlichen) Satzlänge oder der Komplexität der verwendeten Grammatik.

Die systematisch algorithmische Analyse solcher Schreibstiländerungen zum Erkennen von Plagiaten in Textdokumenten wird meist als intrinsische Plagiatserkennung bezeichnet. Hierbei werden neben den oben genannten Charakteristika noch viele weitere lexialische, syntaktische oder semantische Kenngrößen herangezogen, um Texte zu untersuchen. Diese Dissertation [Tsc14] befasst sich mit dem Thema der intrinsischen Plagiatserkennung und entwickelt Lösungsansätze, die stilistische Änderungen innerhalb eines Dokuments aufgrund der verwendeten Grammatik des Autors und somit mögliche Plagiatsfälle erkennen können. Darüber hinaus wird auch die systematische Anwendung der Ansätze in verwandten Problemstellungen, wie z.B. der automatischen Autorenerkennung, erforscht. Konkret wurden folgende Forschungsfragen bearbeitet und gelöst:

1. Kann die ausschließliche Analyse von Grammatik mögliche Plagiate in Textdokumente finden?
2. Ist es möglich, Texte aufgrund grammatikalischer Strukturen den entsprechenden Autoren zuzuweisen?
3. Lässt die verwendete Grammatik Schlüsse auf Metainformationen wie das Alter oder das Geschlecht des Autors zu?
4. Können Mehrautorenwerke so zerlegt werden, dass die einzelnen Fragmente den entsprechenden Autoren zugeordnet werden können?

Im Folgenden wird eine Zusammenfassung der Dissertation gegeben. Nachdem in Kapitel 2 eine kurze Einführung in die grammatikalischen Strukturen von Autoren gegeben wird, werden die entwickelten Algorithmen zur intrinsischen Plagiatserkennung erläutert. Aufgrund der vorgegebenen Länge dieser Zusammenfassung wird dabei nur die grundlegende Idee näher erklärt, wobei alle anderen Themen nur oberflächlich rekapituliert werden. Kapitel 3 zeigt in diesem Sinne kurz, wie die Ansätze so angepasst werden können, um für die automatische Autorenerkennung eingesetzt werden zu können. Des weiteren werden in Kapitel 4 Modifikationen gezeigt, um sowohl Alter und Geschlecht des Verfassers abzuschätzen als auch um die einzelnen Urheber von Mehrautorenwerken zu finden. Schlussendlich wird in Kapitel 5 eine nochmalige Zusammenfassung und ein wissenschaftlicher Ausblick gegeben.

2 Intrinsische Plagiatserkennung

Grundsätzlich wird bei der Plagiatserkennung unterschieden zwischen externen und intrinsischen Verfahren [M. 11]. Erstere ziehen beliebige Datenquellen wie z.B. eigens angelegte Textdatenbanken oder Webseiten heran und versuchen dann durch Vergleichsverfahren, Kopien zu finden. Häufig angewandte Techniken inkludieren dabei n-Gramme

oder standardmäßige Information-Retrieval-Techniken wie z.B. gemeinsame Substrings, häufig kombiniert mit Machine-Learning-Techniken. Im Gegenzug dazu analysieren intrinsische Verfahren ausschließlich das zu untersuchende Dokument und versuchen hier, durch gefundene Stiländerungen mögliche Plagiate aufzudecken. Hierbei werden ebenfalls n-Gramme verwendet [Sta11], aber auch Vokabular-Untersuchungen [OLRV11], Komplexitätsanalysen [SM09] oder Vorkommen von Rechtschreib-/Tippfehlern [KS03] werden herangezogen.

Die im Folgenden zusammengefassten Plag-Inn Algorithmen[4] wurden als intrinsische Plagiatsverfahren entwickelt und basieren auf der Grundannahme, dass sich Autoren in ihrem grammatikalischen Stil unterscheiden, und dass dieser als signifikanter Indikator zur Differenzierung herangezogen werden kann. Beispielsweise kann der englische Satz[5]

(1) *The strongest rain ever recorded in India shut down the financial hub of Mumbai, officials said today.*

auch formuliert werden als

(2) *Today, officials said that the strongest Indian rain which was ever recorded forced Mumbai's financial hub to shut down.*

Die Sätze sind semantisch äquivalent, unterscheiden sich aber signifikant in ihrer Syntax. Die Grammatikbäume, welche sich durch die Verwendung der entsprechenden Strukturregeln der englischen Sprache ergeben, sind in Abbildung 1 dargestellt. Die Knoten bezeichnen dabei sog. *Part-of-Speech (POS)*-Tags und klassifizieren so z.B. Verben (VB), Adjektive (JJ), Nomenphrasen (NP) oder Adverbphrasen (ADVP).

2.1 Der Plag-Inn Algorithmus

Die Grundidee des Algorithmus besteht nun darin, etwaige Differenzen in den Grammatikbäumen wie in Abbildung 1 dargestellt zu quantifizieren, um so Irregularitäten in der verwendeten Syntax zu finden. Konkret besteht der Plag-Inn Algorithmus aus den fünf folgenden Schritten:

1. Als erstes wird das zu prüfende Dokument in einzelne Sätze zerlegt.
2. Für jeden Satz wird anschließend ein Grammatikbaum erzeugt. Da der Ansatz ausschließlich die verwendete Grammatik untersucht und nicht das Vokabular (d.h. die konkreten Wörter), werden die Blätter der Bäume ignoriert.
3. Im nun folgenden Schritt wird die Edit-Distanz zwischen jedem Paar von Grammatikbäumen berechnet, d.h. es wird quantifiziert, wie sehr sich die Struktur der

[4] Plag-Inn steht für *Plag*iarism Detection *Inn*sbruck

[5] Beispiel von der Stanford Parser Website, `http://nlp.stanford.edu/software/lex-parser.shtml`, besucht im Februar 2015

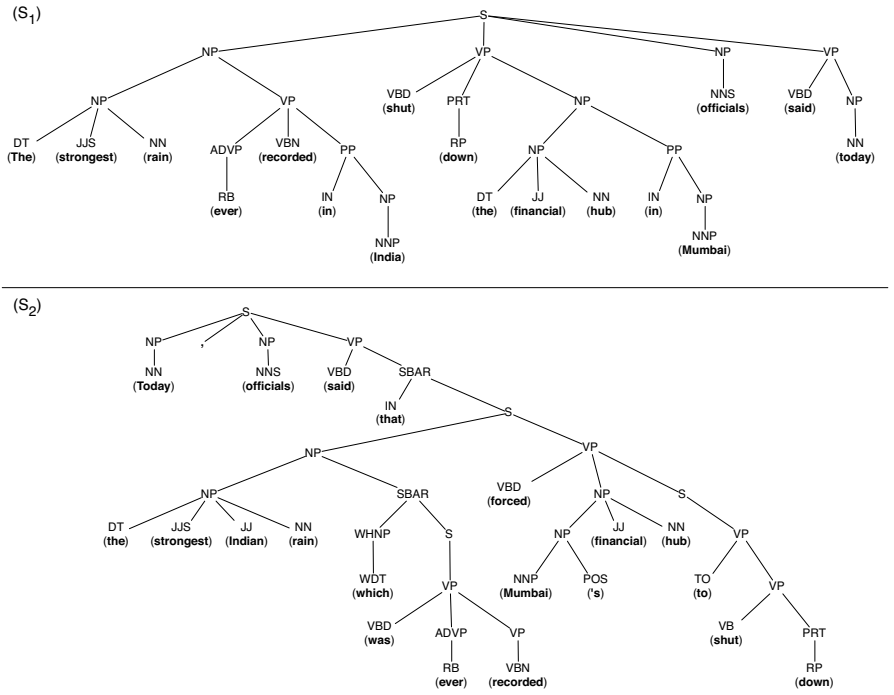

Abb. 1: Grammatikbäume der Sätze (1) und (2).

Bäume unterscheiden. Hierbei wird auf das Konzept von pq-Grammen bzw. der pq-Gramm-Distanz [ABG10] zurückgegriffen, welche als eine Art n-Gramme für Bäume interpretiert werden können. Ein pq-Gramm besteht dabei aus p vertikalen Knoten und q horizontalen Knoten, wobei etwaige fehlende Knoten mit einem $*$ aufgefüllt werden (falls z.B. weniger als q horizontale Knoten zur Verfügung stehen). Beispielsweise kann mit $p = 2$ und $q = 3$ aus dem mittleren Ast des Baums (1) der Abbildung 1 folgendes pq-Gramm extrahiert werden: `[S-VP-VBD-PRT-NP]` (2 nach unten, 3 horizontal). Um die Distanz zwischen zwei Bäumen zu berechnen, müssen alle möglichen pq-Gramme extrahiert werden, so z.B. auch `[S-NP-*-*-NNS]`, welches sich aus dem rechten Nachbarn des eben betrachteten Astes ergibt. Schlussendlich wird die Distanz zwischen zwei Bäumen berechnet, indem die Mengen an pq-Grammen nach bestimmten Regeln verglichen werden.

Jede berechnete Distanz zwischen zwei Grammatikbäumen, d.h. zwischen zwei Sätzen, wird nun eine Distanzmatrix eingetragen. Die Visualisierung der Matrix eines Beispieldokuments mit ca. 800 Sätzen ist in Abbildung 2 dargestellt. Auf den x-/y-Achsen sind jeweils die einzelnen Sätze abgebildet, während die Höhe der z-Achse der jeweiligen Distanz entspricht. Hier ist deutlich ersichtlich, dass die Unterschiede im Bereich der Sätze um Nr. 200 signifikant höher sind als an anderen Positionen, was auf ein mögliches Plagiat hindeutet.

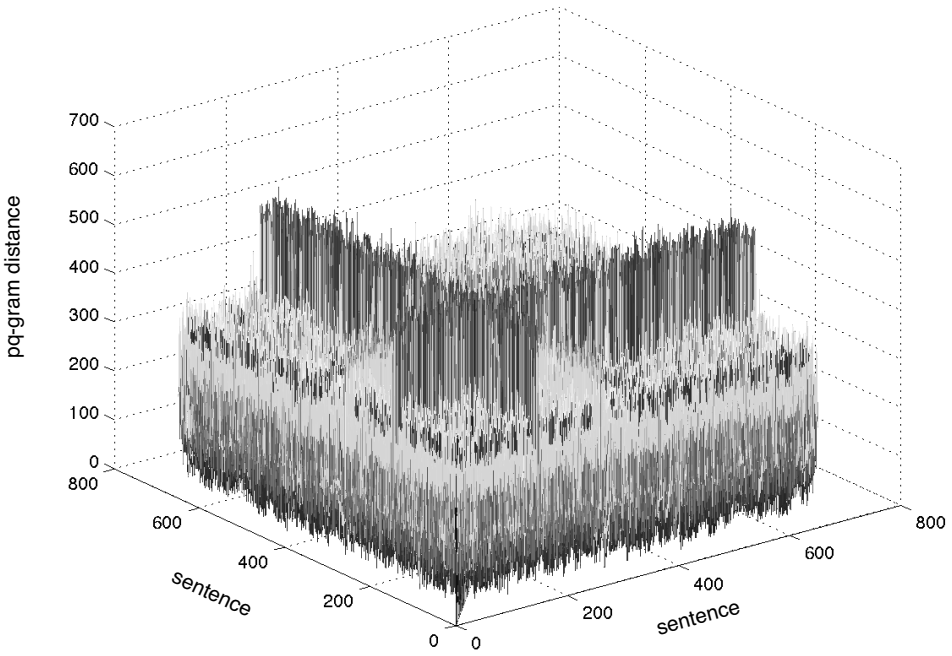

Abb. 2: Visualisierung einer Distanzmatrix eines Beispieldokuments.

4. Um diese für einen menschlichen Betrachter schon gut zu identifizierenden Unterschiede algorithmisch zu finden, wird nun die durchschnittliche Distanz jedes Satzes im Vergleich zu allen anderen Sätzen berechnet. Unter Zuhilfenahme einer Gauß-Verteilung und vordefinierten Schwellwerten werden verdächtige Sätze identifiziert.

5. Für die Berechnung der finalen Aussage über möglichen Plagiarismus im zu prüfenden Dokument wird zuletzt noch ein eigens entwickelter Selektionsalgorithmus angewandt, welcher vorher identifizierte Sätze zu Abschnitten zusammenfasst, nicht markierte Sätze unter gewissen Umständen hinzufügt oder markierte Sätze ggf. wieder entfernt.

Sämtliche Schwellwerte und Parameter wurden auf mehrere Weisen optimiert und auf einem State-of-the-Art Testkorpus des PAN-Workshops [M. 11] der Universität Weimar getestet. Evaluationsergebnisse zeigen, dass der Algorithmus auf diesem Datenset einen F-Score von über 35% erreicht, was einem sehr hoher Wert für intrinsische Verfahren entspricht. Weiters konnte erforscht werden, dass der Algorithmus bei Normallängen-Dokumenten von ca. 100-200 Sätzen (etwa ein wissenschaftliches Paper) im Vergleich zu Buchlänge-Dokumenten noch deutlich an Zuverlässigkeit gewinnt und bis zu 50% F-Score erzielt.

2.2 Plag-Inn Varianten

Im Rahmen der Dissertation wurden noch zwei weitere Varianten des Plag-Inn Algorithmus entwickelt, welche auf der originalen Idee aufbauen, sich aber in mehreren Details unterscheiden. Im *POS-Plag-Inn* Ansatz wird auf die Auswertung von Grammatikbäumen verzichtet und anstatt dessen nur linearisierte Folgen von Part-of-Speech-Tags verarbeitet. Für diese Sequenzen werden dann mithilfe von adaptierten Algorithmen aus der Genetik (Sequenz-Alinierung) die Edit-Distanzen berechnet, welche analog zur ursprünglichen Idee verarbeitet werden.

Weiters wurde im *PQ-Plag-Inn* Algorithmus eine Variante entwickelt, welche nicht mehr einzelne Sätze miteinander vergleicht, sondern Profile aus pq-Grammen erstellt. Hier wird für das gesamte Dokument ein Profil erstellt, welches die meist verwendeten pq-Gramme und deren Häufigkeiten enthält und über Sliding Windows mit einzelnen Textabschnitten verglichen wird. Signifikant ”andere” Abschnitte werden als mögliches Plagiat identifiziert.

Die Varianten wurden ebenfalls ausgiebig getestet und optimiert, und es zeigt sich, dass der POS-Ansatz in etwa die selbe Performanz wie der Grundalgorithmus liefert, während die Variante mit den Profilen hingegen nochmal eine deutliche Verbesserung des F-Scores bringt.

3 Automatische Autorenerkennung

Die Problemstellung der automatischen Autorenerkennung kann recht einfach formuliert werden: Weise einem Textdokument unbekannter Urheberschaft einem bekannten Autor zu, oder anders formuliert: gegeben sei ein Textdokument - wer hat es geschrieben? Die Zahl der möglichen Autoren wird dabei meist so weit wie möglich eingeschränkt (auf z.B. drei oder maximal 20), und für jeden Kandidaten existieren verifizierte Schriftstücke, mit denen das unbekannte Dokument verglichen werden kann. Aktuelle Ansätze verwenden eine Reihe von Kenngrößen (oft mehr als 100), welche sehr häufig mit Machine-Learning-Algorithmen verarbeitet werden [Sta09]. Konkret werden wie bei der intrinsischen Plagiatserkennung stilistische Analysen durchgeführt, die lexikalische (z.B. n-Gramme [Gri07]), syntaktische (z.B. POS-Trigramme [HF07]) oder andere Eigenschaften (z.B. Komprimierungsraten [MWH05]) untersuchen. Die Genauigkeit der Ansätze beläuft sich dabei je nach verwendeten Testdaten auf 70-95%.

In der Dissertation wurde der entwickelte Ansatz zur Plagiatserkennung so adaptiert, dass er für die Autorenerkennung verwendet werden kann. Aufgrund der Evaluationen, welche für den Profilansatz die besten Ergebnisse lieferten, wurde auch hier mit Profilen gearbeitet. Zur Berechnung eines Autorenprofils wird wieder auf Grammatikbäume zurückgegriffen, woraus alle möglichen pq-Gramme extrahiert werden. Die Vorkommenshäufigkeit der einzelnen pq-Gramme und die globale Reihung bilden schlussendlich die Basis für die weitere Berechnung. Ein Beispiel ist in Tabelle 1 angegeben.

pq-Gramm	Häufigkeit [%]	Reihung
NP-NN-*-*	4.07	1
NP-DT-*-*	2.94	2
NP-NNS-*-*	2.90	3
...	...	...

Tab. 1: Beispiel eines Profils zur Autorenerkennung (mit $p = 2, q = 2$).

Zusammengefasst werden Autoren wie folgt erkannt und zugewiesen: (1.) Berechne ein Grammatikprofil für jeden Autorkandidaten, (2.) berechne ein Grammatikprofil für das zu untersuchende Dokument und (3.) weise dem Dokument aufgrund der Ähnlichkeit zu bereits bekannten Profilen einem der Autoren zu.

Die Zuweisung im letzten Schritt wurde dabei auf mehrfache Weise untersucht: einerseits mit Distanzmetriken und andererseits mit bekannten Machine-Learning Algorithmen. Im ersten Fall wird die Distanz zwischen jedem Autorprofil und dem Dokumentprofil berechnet, und der Autor mit der geringsten Distanz wird zugewiesen. Zur Berechnung der Distanzen wurden mehrere aus der Literatur bekannte Metriken herangezogen und für den Umgang mit pq-Gramm-Profilen entsprechend modifziert.

Im Falle der Machine-Learning-Algorithmen wurde auf bekannte Ansätze wie etwa Entscheidungsbäume, Naive Bayes oder k-Nearest-Neighbors gesetzt. Diesen Methoden werden die berechneten Grammatikprofile der Autoren übermittelt, welche als Trainingsdaten dienen. Anschließend wird das Dokumentprofil übermittelt, woraufhin die Algorithmen eine Zuordnung aufgrund der bekannten Profile berechnen.

Beide Varianten wurden parameteroptimiert und auf vier unterschiedlichen Datensätzen mit unterschiedlicher Anzahl von möglichen Autoren ausgiebig getestet. Es zeigt sich, dass die Distanzmetrikberechnung etwas schlechter funktioniert als der Machine-Learning-Ansatz, jedoch beide ausgesprochen gute Resultate liefern. Mit einer Genauigkeit von über 75-90% konnten sowohl Datensätze mit wenigen als auch mit mehreren Kandidaten zugewiesen werden, wobei einzelne Datensätze sogar eine Genauigkeit von 100% Genauigkeit erreichten. Dies ist insbesondere deshalb ein ausgesprochen gutes Resultat, weil im Vergleich zu anderen Ansätzen hier nur eine einzige Kenngröße herangezogen wurde, und zwar die verwendete Grammatik der Autoren.

4 Weitere Einsatzgebiete

Aufgrund der guten Resultate aus den intrinsischen Plagiatserkennungs- und Autorenerkennungsansätzen wurden in der Arbeit zwei weitere, verwandte Problemstellungen behandelt, d.h. es wurde überprüft, ob die reine Analyse von Grammatik auch hier gute Ergebnisse liefern kann. Hierbei wurde einerseits untersucht, ob durch Grammatik Metainformationen wie das Alter oder das Geschlecht des Autors abgeleitet werden können, und andererseits auch, ob eine Grammatikanalyse hilft, Trennlinien in Mehrautorenwerken zu finden.

4.1 Erkennung von Alter und Geschlecht

In aktuellen sog. *Profiling*-Ansätzen wird versucht, möglichst viel Information aus einem gegebenen Textstück zu extrahieren. Dies umfasst häufig das Alter und Geschlecht des Autors (z.B. [AKPS09]), aber auch Daten wie der kulturelle Hintergrund, der Ausbildungsgrad oder psychologische Einstufungen wie etwa Intro-/Extrovertiertheit (z.B. [NRJ13].

In der Dissertation wurde versucht, sowohl das Geschlecht als auch das Alter (aufgeteilt in drei Gruppen) des Autors eines gegebenen Schriftstücks aufgrund der verwendeten Grammatik automatisch zu erkennen. Dabei wurde ähnlich wie bei der Autorenerkennung wieder auf pq-Gramm-Profile zurückgegriffen, welche mit Machine-Learning-Algorithmen verarbeitet wurden. Die Evaluation auf einem häufig verwendeten, großen Testdatenset von mehreren tausenden Web-Blogs ergab auch hier sehr gute Ergebnisse: Das Geschlecht konnte mit nahezu 70% Genauigkeit erkannt werden, und das Alter mit knapp über 60%. Auswertungen im Detail ergaben, dass die Grammatikanalyse ausgesprochen gut zwischen Personen im Alter von 10-20 und Personen von 20-30 unterscheiden kann, allerdings Probleme bei der Trennung zwischen letzteren und Personen über 30 hat.

Obwohl andere Ansätze noch etwas bessere Resultate bei der Geschlechts- und Alterserkennung aufweisen können, sind die Ergebnisse der Arbeit ausgesprochen gut, da wiederum ausschließlich die Grammatik untersucht wurde und auf weitere Kenngrößen verzichtet wurde.

4.2 Zerlegung von Mehrautorendokumenten

Schlussendlich wird die Arbeit noch abgerundet, indem versucht wurde, Einzelbeiträge aus einem gemeinschaftlich geschriebenen Dokument zu filtern. Die Arbeitsweise ist hierbei sehr ähnlich zu den vorigen Ansätzen, d.h. es wurde wieder mit Grammatikbäumen, Profilen und Machine-Learning-Algorithmen gearbeitet. Da es allerdings möglich sein sollte, auch ohne vorher bekannte Proben von den Autoren Einzelbeiträge zu finden, konnten in diesem Fall keine Klassifizierer wie z.B. Naive Bayes verwendet werden. Deshalb wurde auf Clustering-Algorithmen wie z.B. K-Means zurückgegriffen, um grammatikalisch ähnliche Textpassagen automatisiert zu gruppieren.

Die Evaluation wurde auf vier verschiedenen Testdatensätzen durchgeführt, wobei zwei davon aus bekannten literarischen Werken selbst zusammengestellt wurden, und zwei Datensätze vom Autorenerkennungsansatz übernommen und für diese spezielle Problemstellung adaptiert wurden. Die Anzahl der mitwirkenden Autoren der Dokumente war dabei im Bereich von 2-4, wobei einigen Clustering-Algorithmen die korrekte Zahl der Cluster vorgegeben wurde und anderen diese Entscheidung selbst überlassen wurde. Gemittelt über alle Datensätze konnte eine Genauigkeit von 63% erzielt werden, wobei Einzelergebnisse von bis zu 89% erreicht werden konnten. Die globale Einstufung der Performanz dieses Ansatzes ist dabei schwer zu treffen, da in der Literatur nur sehr wenig bis gar keine vergleichbaren Methoden existieren.

5 Zusammenfassung und Ausblick

Textueller Plagiarismus ist ein häufig auftretendes Problem in der modernen vernetzten Gesellschaft, vor allem durch die leichte Zugänglichkeit von Millionen von Textdokumenten. Als eine mögliche Gegenmaßnahme wurden in dieser Dissertation drei intrinsische Plagiatserkennungsalgorithmen entwickelt, welche ausschließlich das zu prüfende Dokument untersuchen und keine externen Vergleiche durchführen. Die Grundidee, dass sich Autoren in der Verwendung ihrer Grammatik signifikant unterscheiden, diese aber meist unbewusst verwenden und somit ungewollte Fingerabdrücke hinterlassen, wurde systematisch verfolgt und in mehreren Varianten validiert. Evaluationen und Optimierungen liefern sehr gute Ergebnisse und deuten darauf hin, dass Plagiate tatsächlich durch Grammatikanalysen gefunden werden können. Des weiteren wurden die Ideen für die verwandten Problemstellungen automatische Erkennung von Autoren, Extrahieren von Alter und Geschlecht und Zerlegung von Mehrautorenwerken übernommen. Messungen ergaben ebenfalls sehr gute Resultate und bestätigen, dass eine reine Analyse der verwendeten Grammatikstrukturen auch in diesen Bereichen signifikante Verbesserungen bringen und sogar als selbständige Ansätze funktionieren können.

Erweiterungen der Arbeit sind in vielen Bereichen denkbar. So wurden z.B. ausschließlich englische Texte untersucht, wobei vermutet werden kann, dass die Ansätze auch in anderen Sprachen ähnlich gute Ergebnisse liefern. Zudem wäre es möglich, dass in grammatikalisch komplexeren Sprachen wie z.B. Deutsch aufgrund der mannigfaltigeren Formulierungsmöglichkeiten von Sätzen sogar noch bessere Resultate erzielt werden können. Nachdem in der Arbeit gezeigt wurde, dass allein Grammatik genug Aussagekraft besitzt, um Autoren zu unterscheiden, wäre eine weitere wichtige Ausbaumöglichkeit, die entwickelten Ansätze um andere häufig verwendete Kenngrößen zu erweitern bzw. diese zu kombinieren. Ein wichtiger Startpunkt wäre hierbei die Integration von Analysen auf Wort- bzw. Vokabularebene, da aus verwandten Arbeiten bekannt ist, dass dies ebenfalls zu guten Ergebnissen führt. Die Kombination aus Grammatik- und Wortanalyse könnte die Algorithmen noch deutlich verbessern. Dies gilt es in weiterführenden Arbeiten zu validieren.

Literaturverzeichnis

[ABG10] N. Augsten, M. Böhlen und J. Gamper. The pq-gram Distance Between Ordered Labeled Trees. *ACM Transactions on Database Systems*, 35(1):4, 2010.

[AKPS09] S. Argamon, M. Koppel, J. Pennebaker und J. Schler. Automatically Profiling the Author of an Anonymous Text. *Communications of the ACM*, 52(2):119–123, 2009.

[Gri07] J. Grieve. Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, 22(3):251–270, 2007.

[HF07] G. Hirst und O. Feiguina. Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Literary and Linguistic Computing*, 22(4):405–417, 2007.

[KS03] M. Koppel und J. Schler. Exploiting Stylistic Idiosyncrasies for Authorship Attribution. In *Proc. of the 18th Joint Conf. on Artificial Intelligence*, Jgg. 69, Seiten 72–80, 2003.

[M. 11] M. Potthast et al. Overview of the 3rd International Competition on Plagiarism Detection. In *Notebook Papers of the 5th Evaluation Lab on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN)*, Amsterdam, The Netherlands, 2011.

[MWH05] Y. Marton, N. Wu und L. Hellerstein. On Compression-Based Text Classification. In *Advances in Information Retrieval*, Seiten 300–314. Springer, 2005.

[NRJ13] J. Noecker, M. Ryan und P. Juola. Psychological Profiling Through Textual Analysis. *Literary and Linguistic Computing*, 28(3):382–387, 2013.

[OLRV11] G. Oberreuter, G. L'Huillier, S. Ríos und J. Velásquez. Approaches for Intrinsic and External Plagiarism Detection. In *Notebooks of the 5th Eval. Lab on Uncovering Plagiarism, Authorship and Social Software Misuse*, Amsterdam, The Netherlands, 2011.

[SM09] L. Seaward und S. Matwin. Intrinsic Plagiarism Detection Using Complexity Analysis. In *Proceedings of the 25th Conference of the Spanish Society for Natural Language Processing*, Seite 56, 2009.

[Sta09] E. Stamatatos. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, March 2009.

[Sta11] E. Stamatatos. Intrinsic Plagiarism Detection Using Character n-gram Profiles. In *Notebook Papers of the 5th Evaluation Lab on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN)*, Amsterdam, The Netherlands, 2011.

[Tsc14] M. Tschuggnall. *Intrinsic Plagiarism Detection and Author Analysis By Utilizing Grammar*. Dissertation, Inst. of Computer Science, University of Innsbruck, December 2014.

Michael Tschuggnall wurde am 26. März 1982 in Hall i.T. in Österreich geboren und hat seinen Lebensmittelpunkt seither in Telfs. Nach dem Schuleinstieg 1988 in die Volks- und Hauptschule in Telfs besuchte er 1996 bis 2001 die Höhere Technische Lehranstalt (HTL) für Wirtschaftsingenieurwesen, mit dem Ausbildungszweig Betriebsinformatik. Nach dem Abschluss der Matura mit Auszeichnung startete er 2001 das Informatikstudium an der Universität Innsbruck. Nach 4-jähriger Unterbrechung konnte das Bachelorstudium 2008 und das anschließende Masterstudium 2010 jeweils mit ausgezeichnetem Erfolg abgeschlossen werden. Die Masterarbeit wurde bereits in der Gruppe Datenbanken- und Informationssysteme, geleitet von Prof. Günther Specht, geschrieben, wo eine direkt folgende Anstellung als Universitätsassistent nach 3,5 Jahren 2014 zur Promotion zum PhD führte. Begleitend zur universitären Ausbildung war er auf selbständiger Basis laufend an der Entwicklung diverser Software für Web und Mobiltelefone tätig.