Plagiaterkennung

Ausarbeitung für das Seminar Information Retrieval SS 2012

Christoph Rauwolf

Universität Koblenz-Landau

crauwolf@uni-koblenz.de

Abstract. Plagiaterkennung ist ein Bereich des Information Retrieval, der in der letzten Zeit immer mehr an Bedeutung gewonnen hat. Dabei gibt es zwei Ansätze, den intrinsischen und den extrinsischen, unter die sich die verschiedenen Verfahren einordnen lassen. In dieser Arbeit werden die beiden Ansätze, sowie ihre Abläufe schematisch erklärt und an bestimmte Aspekten beispielhaft ausgeführt. Außerdem wird sowohl auf ein intrinsisches, als auch auf ein extrinsisches Verfahren genauer eingegangen, um so einen besseren Einblick in die Thematik zu gewähren.

1 Einleitung

Gerade durch die in akademischen Bereichen zunehmende Nutzung des Internets und dem damit verbundenen Zugriff auf große Mengen von veröffentlichten Dokumenten kommt es in der heutigen Zeit immer häufiger zu Plagiatsfällen. Hierbei werden Passagen aus anderen Veröffentlichungen in wörtlicher, oder modifizierter Form übernommen und als eigene Leistung ausgegeben. Aufgrund der enormen Menge an veröffentlichten Arbeiten und potentiell in Fragen kommenden Quellen, kann dieser Entwicklung allerdings kaum noch durch menschliche Kontrolle effektiv entgegengewirkt werden. An diesem Punkt greift die algorithmischen Plagiaterkennung ein.

Die Verfahren, die es hierbei gibt, lassen sich in intrinsische und extrinsische Ansätze einteilen. Im **extrinsischen** Fall liegt ein zu untersuchendes Dokument (verdächtiges Dokument) vor, dass mit einer Menge von möglichen Quelldokumenten auf direkt, oder in modifizierter Form übernommene Passagen verglichen werden soll. Beim **intrinsischen** Fall hingegen soll nur ausgehend von nur einem Text (Quelldokument) entschieden werden, ob dieser von einem, oder von mehreren Verfassern geschrieben wurde. Wie diese Ansätze im einzelnen funktionieren und auf welchen Grundlagen sie beruhen, wird im weiteren Verlauf dieser Arbeit dargestellt.

Im Abschnitt 2 werden der intrinsische und der extrinsische Ansatz zur Plagiaterkennug und ihr prinzipieller Ablauf genauer ausgeführt. Im 3. Abschnitt werden beispielhaft einige intrinsische Methoden näher erklärt. Um auch die extrinsische Seite genauer zu beleuchten, wird im 4. Abschnitt das in Paper [2] von Efstathios

adfa, p. 1, 2011.

Stamatatos vorgestellte Verfahren beispielhaft erläutert. Für einen Einblick in den aktuellen Stand der Forschung, wird im 5. Abschnitt das in [3] beschriebene Verfahren, dass den PAN-Workshops[1] 2011 gewonnen hat, kurz vorgestellt. Am Ende folgt im 6. Abschnitt ein kurzes Fazit.

2 Allgemeines

Die beiden in diesem Absatz vorgestellten Ansätze beziehen sich auf unterschiedliche Situationen, bzw. unterschiedlich beschaffene Szenarien. In den meisten Fällen lassen sich Verfahren zur Plagiaterkennung entweder dem einen, oder dem anderen Ansatz zuordnen, wobei es auch Ausnahmen gibt (siehe Abschnitt 5), die in beiden Situationen angewendet werden können.

2.1 Intrinsische Plagiaterkennung

Bei einem **intrinsischen** Szenario soll ausgehend von nur einem Dokument (verdächtiges Dokument) herausgefunden werden, ob dieses von einem, oder von mehreren Verfassern geschrieben wurde (ähnlich wie bei der Autorenschaft Identifizierung). Dies kann unter anderem hilfreich sein, wenn nicht ohne weiteres eine Liste von möglichen Quelldokumenten erstellt werden kann. In diesem Fall kann das Quelldokument auf Brüche im Stil, Ausreißer in der Wortwahl und andere, vorwiegend strukturelle Eigenschaften hin untersucht werden.
Einige der Methoden aus diesem Ansatz können auch bei der Voranalyse von extrinsischen Verfahren genutzt werden, um die auf Plagiate zu untersuchenden Bereiche bereits im Vorfeld einzugrenzen.

2.2 Extrinsische Plagiaterkennung

Im **extrinsischen** Fall liegt ein zu untersuchendes Dokument (verdächtiges Dokument) vor, dass mit einer Menge von anderen Dokumenten (mögliche Quelldokumente) verglichen wird.
Gerade in diesem Fall ist es wichtig, dass der verwendete Algorithmus nicht nur zuverlässig, sondern auch effizient ist, da mitunter sehr viele Dokumente analysiert werden müssen. Eine solche Effizienzsteigerung kann zum Beispiel durch die Entfernung von Stoppwörtern[2] (en: "stopwords") erreicht werden, was auf der einen Seite zwar zu einer Steigerung der Geschwindigkeit führt, auf der andern Seite allerdings auch einen Verlust an strukturellen Informationen bedeutet.

---

[1] Der PAN-Workshop (Plagiarism Analysis, Authorship Identification, und Near-Duplicate Detection) ist der weltweit führende Wettbewerb zu den Themen Plagiaterkennung, Autorenschaft Identifizierung und Vandalismuserkennung. Er ist Teil der CLEF 2012 (vgl. [6],[7]).

[2] Stoppwörter sind Wörter, die sehr häufig in Texten vorkommen und fast keine Relevanz für die Erfassung von Inhalten haben (vgl. [4]).

Wenn dabei das verdächtige Dokument in einer anderen Sprache verfasst ist als die Quelldokumente, so nennt man diesen Spezialfall **Sprachübergreifende Plagiatierkennung** (en: "cross-lingual plagiarism detection"). Dieser Fall wird von den meisten Verfahren nicht direkt unterstützt. Es gibt allerdings auch extra dafür entwickelte Verfahren, die mit Übersetzungen oder ähnlichem arbeiten.

2.3 Prinzipieller Ablauf der Verfahren

Auch wenn die Ansätze und Verfahren sich sonst unterscheiden, laufen sie fast alle in den gleichen drei Stufen ab:

1. **Voranalyse/ Kandidatenauswahl (en: "Pre-Analysis Stage")**
   Um die Effizienz in den weiteren Phasen zu steigern, wird zunächst einmal eine Vorauswahl von weiter zu untersuchenden Passagen (bzw. Dokumenten) vorgenommen.
   Beim intrinsischen Ansatz findet dies alleine im verdächtigen Dokument statt (z.B. Abschnittsweise), während beim extrinsischen Fall in erster Linie die Menge der zu vergleichenden Quelldokumente reduziert wird. Die Passagen (bzw. die Quelldokumente), die in dieser Stufe als unauffällig "markiert" wurden, werden im Folgenden nicht weiter untersucht. Daher ist es wichtig eher mit einer hohen Toleranzschwelle zu arbeiten, um so keine Plagiate von Anfang an auszuschließen.

2. **Detailanalyse (en: "Classification Stage")**
   Die in der ersten Phase als auffällig markierten Passagen des zu untersuchenden Dokuments, bzw. die Menge der möglichen Quelldokumente, wird in dieser Phase einer genaueren Untersuchung unterzogen.
   Bei intrinsischen Verfahren wird hier versucht die unstimmigen Passagen im Dokument zu bestimmen, während beim extrinsischen Ansatz versucht wird die exakten Verbindungen zwischen dem verdächtigen und den in der ersten Phase "markierten" Quelldokumenten zu bestimmen.

3. **Nachkontrolle ("post-processing")**
   Die in der zweiten Phase entdeckten, möglichen Plagiat-Fälle werden in der letzten Phase noch einmal genauer kontrolliert und dann entweder endgültig markiert, oder aber verworfen.
   Je nach Verfahren kann hier auch bestimmt werden mit welcher Wahrscheinlichkeit es sich bei einer Passage um ein Plagiat handelt. Durch das Setzen eines Grenzwertes kann dann zusätzlich gefiltert werden, um eine höhere Präzision zu erzielen.

3 Methoden zur intrinsischen Plagiaterkennung

Der prinzipielle Ablauf für intrinsische Ansätze zur Plagiaterkennung ist bereits im obigen Abschnitt erläutert worden. Damit man sich einen besseren Eindruck von der Arbeitsweise dieser Verfahren machen kann, werden hier nun einige Methoden, die bei intrinsischen Verfahren Anwendung finden, genauer erklärt.
Die einzelnen Punkte entstammen dabei mit leichten Änderungen dem unter [1] zu findenden Paper der Autoren Stein, Lipka und Meyer zu Eisen. Da es an dieser Stelle aber nur darum geht einen Eindruck zu vermitteln, werden im Folgenden nur ausgewählte Methoden weiter erläutert und den jeweiligen Stufen im Ablauf zugeordnet.

Abbildung 1: Beispielhafte Stufenübersicht eines intrinsischen Verfahrens. Quelle: vgl. [1]

1. Stufe: Voranalyse
In dieser Phase geht es darum eine Vorauswahl kritischer Passagen aus dem verdächtigen Dokument zu treffen, die dann weiter untersucht werden. Dazu kann der Text auf verschiede Aspekte hin untersucht werden:

- Als erstes findet eine "**Verunreinigungsabschätzung**" (en: "impurity assessment") statt, bei der unter Berücksichtigung von Länge, Positionen und Formatierung nach Unstimmigkeiten gesucht wird. Auch Informationen über die Art des Textes (wissenschaftlicher Text, Hausarbeit, ...), sein Genre (Roman, Sachbuch, ...) und die herausgebende Institution spielen bei dieser Voruntersuchung eine wichtige Rolle.

- Unter Berücksichtigung der oben genannten Aspekte wird dann eine passende **Zerlegungsstrategie** (en: "decomposition strategy") gewählt. Diese kann von der einfachen Einteilung in Blöcke mit gleicher Wortzahl, über die Berücksichtigung von strukturellen Grenzen (wie zum Beispiel Satzgrenzen), bis hin zur Berücksichtigung der Textgliederung (also Abschnitte, Unterkapitel, Kapitel) gehen. Auch Tabellen, Fußnoten und Zitate können berücksichtigt werden, genauso wie inhaltliche, oder sogar stilistische Beziehungen innerhalb des Dokuments.

- Als letztes erfolgt in dieser Stufe die **Erstellung eines Stil-Modells** (en: "style model construction") unter Berücksichtigung der Informationen aus den vorangegangenen Schritten. Dies kann je nach Zerlegungsstrategie auf Grundlage von autorspezifischen Charakteristika wie Wortwahl oder Komplexität, durch die Verwendung von n-Grammen [3] (en: "n-grams"), oder durch viele andere Verfahren geschehen. Die einzelnen Attribute müssen dabei unterschiedlich stark gewichtet werden.

2. Stufe: Detailanalyse

In dieser Stufe werden die als auffällig markierten Passagen anhand des erstellten Stilmodells auf "Stilaußenseiter" (en: "style outlier") hin untersucht. Die Verfahren, die dazu verwendet werden, lassen sich in drei Klassen einteilen:

- **Häufigkeitsbasierende** Methoden (en: "density methods"): Betrachtet man die Häufigkeit des Auftretens bestimmter Wörter in den zu untersuchenden Abschnitten und dem Dokument insgesamt, so lässt sich über die relative Häufigkeitsverteilung direkt auf die Wahrscheinlichkeit für einen Plagiat-Fall schließen.

- **Zusammenhangsbasierende** Methoden (en: "boundary methods"): Diese betrachten Informationen über die Zusammenhänge von Objekten, die aus den Abständen der Objekte untereinander gewonnen werden (strukturbezogene Informationen).

- **Rekonstruktionsmethoden** (en: "reconstrution methods"): In der ersten Stufe wurde bereits ein Dokument-Stilmodell erstellt. Die Idee der Rekonstruktionsmethode ist nun, für jede zu untersuchende Passage ebenfalls ein Stilmodell nach den gleichen Kriterien zu erstellen. Dieses wird dann mit den Dokument-Stilmodell vergleichen. Je größer die Abweichung ist, desto wahrscheinlicher stammt der jeweilige Abschnitt von einem anderen Autor.

3. Stufe: Nachkontrolle

Hier werden die in der zweiten Stufe markierten Passagen noch einmal genauer überprüft um den Plagiat-Verdacht entweder zu bestätigen, oder aber zu verwerfen. Die hierbei verwendeten Methoden unterscheiden sich darin, ob sie auf einzelne Abschnitte angewendet werden können, oder aber die Informationen aus dem gesamten Text benötigen.

- Die Verfahren, die sich lediglich auf die **einzelnen Abschnitte** beziehen, können die Fehlerwahrscheinlichkeit durch bestimmte Arten von heuristischen Filtern minimieren, indem sie zum Beispiel als solches gekennzeichnete Zitate herausfiltern. Diese könnten sonst als Plagiatsfälle markiert werden und würden so das Ergebnis verfälschen.

---

[3] Unter einem n-Gramm versteht man die Aufteilung eines Textes in einzelne Fragmente, von denen jeweils n-Stück zu einem n-Gramm zusammengefasst werden (vgl. [4]).

- Die Methoden, die Informationen über den **kompletten Text** benötigen, versuchen die Fehlerwahrscheinlichkeiten durch "Meta-Lernen" zu verringern. Hierbei wird der Text in zwei Mengen aufgeteilt. Die erste Menge D1 bilden alle als Ausreißer (en: "outlier") markieren Abschnitte und die zweiten Menge D2 alle übrigen [Stein und Meyer zu Eisen]. Mögliche Ansätze für Meta-Lernen sind:

- Vertrauensbasierte **Mehrheitsentscheidungen** (en: "confidence-based majority decisions"): Von einem geschätzten Fremdanteil im verdächtigen Dokument ausgehend, wird ein Grenzwert bestimmt. Überschreitet die Anzahl der Ausreißer diesen Grenzwert, so wird davon ausgegangen, dass der Text von mehr als einem Autor geschrieben wurde.

- **Demaskierung** (en. "unmasking") :

  Das Prinzip dahinter ist das folgende: Zu Beginn des Verfahrens wird das gesamte verdächtige Dokument in zwei Gruppen von Passagen aufgeteilt. Im Laufe des Verfahrens wird dann bestimmt, ob die Passagen der beiden Gruppen vom gleichen Autor geschrieben wurden.
  Dazu werden in mehreren Iterationsstufen immer wieder von beiden Gruppen die Elemente entfernt, die sich am meisten unterscheiden. Auf diese Weise wird der Gesamtstil der beiden Gruppen immer mehr einander angenähert (angenommen der Stil wird durch die Wortwahl bestimmt). Wenn es nach einer bestimmten Anzahl von Durchläufen immer noch möglich ist die Stile der beiden Gruppen zu unterscheiden, so stammen sie von unterschiedlichen Autoren, weil in beiden Gruppen noch genug stiltragende Elemente vorhanden sind.
  Wären die beiden Gruppen von nur einem Autor geschrieben worden, so wären nach gleicher Anzahl von Durchläufen vermutlich nicht mehr genug Informationen über den Stil vorhanden, da die entfernten Elemente ja den Stil der beiden Gruppen in gleicher Weise betreffen. Es wären folglich nicht mehr möglich den Stil der beiden Gruppen auseinander zu halten, so wie es bei verschiedenen Autoren der Fall wäre.

4 Beispiel für ein extrinsisches Verfahren zur Plagiaterkennung

Am Beispiel des in [2] vorgestellten Verfahrens von Efstathios Stamatatos soll in diesem Abschnitt die Arbeitsweise eines extrinsischen Verfahrens zur Erkennung von Plagiaten vorgestellt werden.

4.1 Prinzipielle Funktionsweise

Die Grundidee hinter dem Verfahren ist es, die strukturellen Eigenschaften des Dokuments näher zu untersuchen, statt sich auf die inhaltliche Aspekte zu konzentrieren. Die Informationen über die strukturelle Beschaffenheit des Dokuments werden dabei

über das Auftreten einer kleinen Liste von **Stoppwörter** (en: "stopwords") gewonnen. Mit Hilfe von Stoppwort n-Grammen (en: "stopword n-gramms") können lokale syntaktische Gemeinsamkeiten zwischen Dokumenten erkannt werden. Dieses Verfahren kann sich mit den aktuellsten Verfahren messen und liefert auch bei besonders schwer zu erkennenden Plagiat-Fällen noch brauchbare Resultate. Ein solcher Fall liegt beispielsweise dann vor, wenn die übernommenen Passagen stark verändert wurden, viele der Wörter durch Synonyme ersetzt wurden, oder die kopierten Passagen recht kurz sind.

4.2 Ablauf

Der im 2. Abschnitt vorgestellte Ablauf für Verfahren gilt auch hier und läuft in den gleichen drei Stufen ab. Im Vorfeld wird das Dokument allerdings noch für die weitere Verarbeitung anders repräsentiert.

- **Vorbereitung:**
  Das Dokument wird als Menge von Stoppwort n-Grammen (SWNG) dargestellt. Anhand einer vorgegebenen Liste von Stoppwörtern wird der gesamte Text bis auf diese Wörter reduziert. Alle anderen Wörter werden verworfen. Aus den verbleibenden Stoppwörtern werden dann die n-Gramme gebildet. Die Menge aller dieser SWNG bildet, geordnet nach ihren ersten Auftreten, das sogenannte "Profil" des Dokuments.

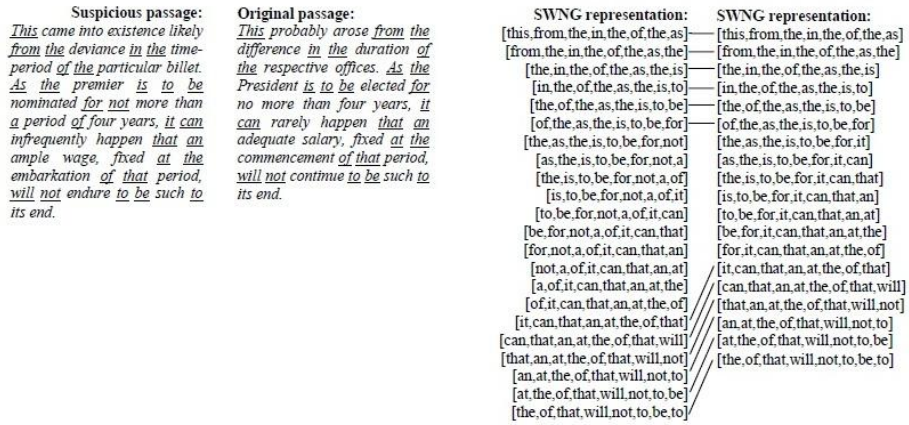

Abbildung 2: Beispiel für die Repräsentation zweier Texte als Stoppwort 8-Gramme. Gleiche 8-Gramme sind mit Linien verbunden und stellen potentielle Plagiat-Fälle dar. Quelle: [2]

- **1.Stufe: Voranalyse (Kandidaten Auswahl):**
  Ausgehend von der SWNG-Darstellung werden gemeinsame n-Gramme zwischen den verdächtigen- und den Quelldokumenten gesucht. Dabei ist es wichtig, dass die Länge der n-Gramme richtig gewählt wird. Ist sie zu groß gewählt, so werden

kaum Gemeinsamkeiten erkannt (in der ersten Phase sollte lieber zu viel als zu wenig erkannt werden). Ist sie zu klein, werden zu viele Gemeinsamkeiten erkannt und das Verfahren wird ineffizient.

Auch muss berücksichtigt werden, dass bestimmte **Ähnlichkeiten zufällig** auftreten können. Dies ist meistens dann der Fall, wenn die Passagen bestimmte, sehr häufig auftretende Stoppwörter enthalten. Um diese Fälle auszusortieren, wird eine Zusatzbedingung eingeführt. Hierbei werden die sechs häufigsten Stoppwörter (im englischen: the, of, and, a, in, to) in einer Menge *C* zusammengefasst und in die weitere Bewertungsmethode mit einbezogen. Zwei n-Gramme der Länge *n* gelten demnach, unter Verwendung des "Zufallsfilters", erst dann als ähnlich, wenn sie weniger als *n*-1 Elemente aus *C* enthalten und weniger als *n*-2 davon ohne Unterbrechung aufeinander folgen.

- **2. Stufe: Klassifizierung (Zusammenhänge bestimmen):**
  In diesem Schritt werden die genauen Zusammenhänge von den in der ersten Stufe als ähnlich markierten Dokumenten bestimmt. Dazu wird versucht, aus gleichen SWNGen möglichst lange Abfolgen zu erzeugen, die mit den Textabschnitten übereinstimmen. Falls die als Plagiat markierte Passage wörtlich übernommen wurde, kommt in beiden Dokumenten sogar die gleiche Abfolge von SWNGen in der selben Reihenfolge vor. Die Gemeinsamkeiten der beiden Dokumente werden dann in einem Streudiagrammen als diagonale Linien dargestellt, die wiederum erkannt werden können (siehe Abbildung 3). Wenn der Plagiat-Abschnitt allerdings stark verändert wurde, kommt es zu Störungen, wie zum Beispiel Rauschen oder Spalten[4] (en: "gaps") zwischen den Linien. Die Stärke dieser Störungen hängt von der gewählten Länge der n-Gramme ab. Je länger diese sind, desto störungsanfälliger sind sie auch und daher sollte für diese Stufe eine deutlich geringere Länge gewählt werden, als es in der ersten der Fall war.

[4] Mit Spalten sind in diesem Kontext die Abstände innerhalb der diagonalen Linien gemeint. Eine andere passende Übersetzung wäre Abrisse.

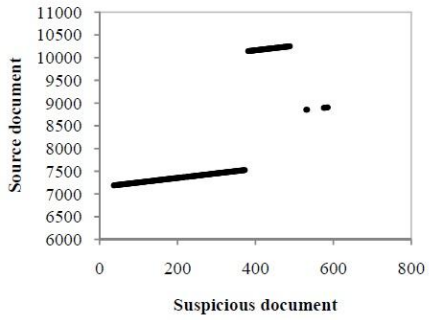

Abbildung 3: Streudiagramm für in beiden Dokumenten gleiche n-Gramme. Hier wurden Passagen unverändert übernommen und liegen im Text direkt beieinander. Quelle: [2].

Auch in dieser Stufe kann es zu zufälligen Ähnlichkeiten zwischen den n-Grammen kommen. Ähnlich wie in der ersten Stufe können diese wieder durch ihren großen Anteil von sehr häufigen Stoppwörtern erkannt werden. Das Auswahlkriterium darf allerdings nicht so streng sein wie das der ersten Stufe, weil sonst zu viele Spalten entstehen können. Auch unterschiedliche Reihenfolgen beim Auftreten gleicher n-Gramme in beiden Dokumenten können fälschlicher Weise als Spalten gedeutet werden.

Ein anderer Problemfall tritt auf, wenn in einem verdächtigen Dokument viele übernommene Passagen in kurzem Abstand aufeinander folgen, was nicht notwendiger Weise auch im Quelldokument der Fall sein muss. Genauso können zwei originale Passagen im gleichen Quelldokument einen sehr kleinen Abstand zueinander haben, während sie als Plagiatsfälle im verdächtigen Dokument einen großen Abstand zueinander haben.
Um dem entgegenzuwirken wird zunächst eine Anfangsmenge von Passagenzusammenhängen maximaler Länge im verdächtigen Dokument gesucht, wobei kleinere Spalte zugelassen werden. Dann werden die zugehörigen Passagen aus dem Quelldokument bestimmt. Wenn eine dieser Passagen im Quelldokument nicht homogen ist (d.h. größere Lücken aufweist), wird sie in kleinere Passagen aufgeteilt. Letztendlich werden dann die Passagenzusammenhänge im verdächtigen Dokument auf Grundlage dieser kleineren Passagen bestimmt.

- **3. Stufe: Nachkontrolle:**
  In dieser Stufe sollen die bisher gefundenen und als Plagiat markierten Passagen noch einmal überprüft, und je nach Grad des Plagiats bewertet werden. Dabei darf es keinen Unterschied machen, ob eine Passage stark verändert, oder wörtlich übernommen wurde. Dieser Ähnlichkeitswert wird berechnet, indem zunächst einmal von den beiden zu vergleichenden Passagen je ein Profil von Buchstaben n-

Grammen erstellt wird. Zu diesem Zweck werden im Vorfeld alle Sonderzeichen entfernt und Groß- in Kleinbuchstaben umgewandelt. Der Ähnlichkeitswert ist dann der Quotient aus der Größe der Schnittmenge der beiden Profile und der Größe des größeren Profils.

Ein mögliches Problem in dieser Stufe können kurze Passagen sein, die in beiden Dokumenten auftreten, jedoch keine Plagiate sind. Dies können zum Beispiel Sprichwörter, Redewendungen, Zitate, Bibelverse und vieles mehr sein. Da sie wie gesagt meistens recht kurz sind, können sie durch die Einführung eines Schwellwertes für die minimale Passagenlänge herausgefiltert werden. Dieser muss anhand von Versuchen bestimmt werden, wobei das vorliegende Textgenre, etc. zu beachten ist. Mit dieser Methode vermeidet man zwar einige Fehlentdeckungen von Plagiaten, allerdings verliert man auch die kurzen Passagen, in denen tatsächlich Plagiat-Fälle vorliegen.

5 Gewinner des PAN-Workshops 2011

In diesem Abschnitt wird das in [3] beschriebene Verfahren von Oberreuter, L'Huillier, Ríos und Velásquez vorgestellt, dass den PAN-Workshop 2011 gewonnen hat. Diesem Verfahren gelang es in der Kategorie intrinsische Verfahren mit weitem Abstand den ersten Platz und in der Kategorie extrinsische Verfahren den dritten Platz zu belegen. So wurde es in der Summe zum Sieger des gesamten Wettbewerbs ernannt.
Besonders die Leistung des intrinsischen Verfahrens ist bemerkenswert, weil es zwar in erster Linie für die Verwendung bei intrinsischen Situationen konzipiert wurde, aber sogar bei extrinsischen Situationen noch recht gute Ergebnisse liefert. Folglich ist das Verfahren in der Lage ohne Berücksichtigung der potentiellen Quelldokumente nur das verdächtige Dokument zu untersuchen und trotzdem noch viele Plagiat-Fälle aufzufinden.

5.1 Methode für extrinsische Plagiaterkennung

Das Verfahren läuft in zwei Stufen ab. Ähnlich wie bei anderen Verfahren wird auch hier in der ersten Stufe versucht die Menge der potentiellen Quelldokumente einzuschränken. In erster Linie werden zu diesem Zweck alle Stoppwörter entfernt und aus den verbleibenden Wörtern Wort 4-Gramme erstellt. Weisen das verdächtige Dokument und ein Quelldokument mindestens zwei gleiche Wort 4-Gramme auf, so wird es für die nächste Phase markiert und andernfalls für die weiteren Untersuchungen verworfen.
In der zweiten Stufe findet dann eine intensivere Untersuchung der markierten Dokumente statt. Zu diesem Zweck werden zu den Texten Wort tri-Gramme gebildet, ohne die Stoppwörter vorher zu entfernen. Das Verfahren ist nicht dazu gedacht auch sprachübergreifende Plagiat-Fälle zu behandeln.

5.2 Methode für intrinsische Plagiaterkennung

Da es in dieser Arbeit in erster Linie um das allgemeine Verständnis für die verschiedenen Verfahren geht werden an dieser Stelle, so wie es auch in [3] gemacht wurde, zunächst einmal die dem Verfahren zugrundeliegende Idee näher erläutert:

- Um das Mitwirken mehrerer Autoren an einem Dokument feststellen zu können muss man in der Lage sein, den Schreibstil des Textes bestimmen zu können.
- Mit Hilfe von Wort n-Grammen kann man den "lokalen Stil" kleinerer Passagen erfassen und ihn mit dem Dokumentstil vergleichen (der ebenfalls durch n-Gramme erfasst wird). Dabei wird davon ausgegangen, dass sich der Gesamtstil auch in den kleineren Passagen wiederfinden muss, sofern diese von selben Autor geschrieben wurden.

Aus diesen Gedanken ergibt sich folgende Idee für das Verfahren:

Wenn es im Dokument autorspezifische Wörter gibt, dann lassen sich diese auch in allen Passagen wiederfinden, die von diesem Autor verfasst wurden.

Ablauf des Verfahrens:

Als erster Schritt werden aus dem Dokument alle Zahlen und Zeichen entfernt, die nicht im Bereich von a-z liegen, wobei Großbuchstaben in Kleinbuchstaben umgewandelt werden. Als zweites werden Wort-uni-Gramme gebildet unter Einbezug aller nicht-numerischen Wörter, wobei im Vorfeld keine Stoppwörter entfernt werden.
Im nächsten Schritt wird für jedes vorhandene Wort eine Häufigkeitsanalyse durchgeführt und ein Häufigkeitsvektor v erstellt. Danach wird das gesamte Dokument in eine Menge von Passagen C aufgeteilt. Die einzelnen Passagen c ∈ C werden dabei mit Hilfe eines Sliding-Windows [5] der Größe m erzeugt, das über das komplette Dokument läuft. Im Anschluss wird für jede dieser Passagen c ein neuer Häufigkeitsvektor v_c erstellt, der in den weiteren Schritten verwendet wird, um den Stil der einzelnen Passagen mit dem des Gesamtdokuments zu vergleichen.

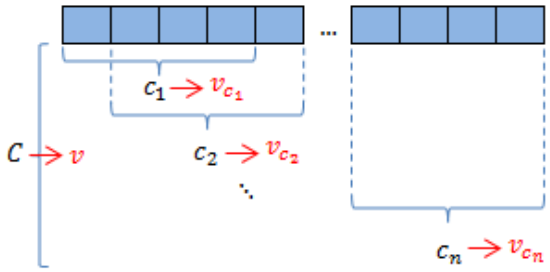

Abbildung 4: Einteilung in Gruppen c und Erstellung von Häufigkeitsvektoren v.

[5] Ein Bereich (Fenster) fester Größe, der eine Liste elementweise durchläuft. In jedem Schritt wandert der Bereich ein Element weiter, bis die gesamte Liste durchlaufen ist.

Dieser Gesamtstil ist der Durchschnitt der Unterschiede zwischen dem Gesamtdokument und jeder einzelnen Passage. Dieser Unterschied lässt sich vereinfacht mit v - v_c für alle c ∈ C berechnen (der genaue Ablauf wird im folgenden Abschnitt: "Algorithmus" beschrieben).

Im letzten Schritt wird jede Passage, hinsichtlich ihres Unterschiedes zum Gesamtdokumentstil bewertet (sog. Stilfunktion). Dabei wird jede Passage nur in Bezug auf die in ihr vorkommenden Wörter mit dem gesamten Dokument verglichen. Eine Passage gilt dann als auffällig, wenn dieser Unterschied einen vorher in Versuchsreihen bestimmten Grenzwert δ unterschreitet.

Damit entspricht das Verfahren dem intuitiven Denken. Wenn bestimmte Worte nur in manchen Passagen verwendet werden, führt der Vergleich dieser Passage mit dem gesamten Dokument zu einem geringen Wert, da die Häufigkeit dieser Worte in dieser Passage ähnlich der Häufigkeit der Wörter im gesamten Dokument wäre.

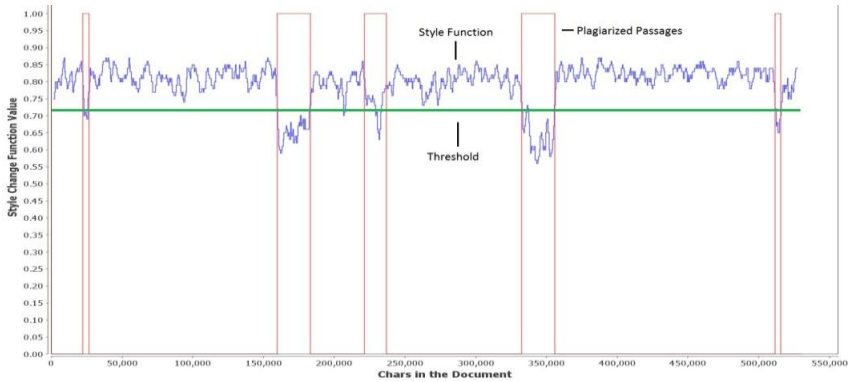

Abbildung 5: Beispiel für die vorgestellte intrinsische Plagiaterkennung. Es wird die Entwicklung der Stil-Funktion bei fortlaufendem Sliding-Window dargestellt. Quelle: [3].

Algorithmus:

Zur besseren Übersicht ist hier noch einmal der Algorithmus in des intrinsischen Verfahrens in ähnlicher Form, wie er in [3] beschrieben ist, formaler dargestellt. Die mit einer # beginnenden Zeilen sind dabei ergänzende Kommentare, die im Original nicht zu finden sind.

```
Eingaben: C, v, m, δ
# C = Dokument in Gruppen eingeteilt
# v = Gesamtes Dokument
# m = Größe der Gruppen
# δ = Toleranz Grenzwert

# Berechnung der einzelnen Passagenstile/der Unterschiede
# zum Gesamtstil
for c ∈ C do
  d_c ← 0
  erstellen von v_c anhand der Wort-Häufigkeiten[6] in c
  for wort w ∈ v_c do
    d_c ← d_c + |häufigkeit(w,v) − häufigkeit(w,v_c)| / |häufigkeit(w,v) + häufigkeit(w,v_c)|
  end for
end for


# Berechnung des Gesamtdokumentstils
dokumentstil ← 1/|C| ∑_{c∈C} d_c


# Vergleichen der einzelnen Passagenstile mit dem
# Gesamtdokumentstil mit Grenzwert δ als Toleranz
for c ∈ C do
  if d_c < dokumentstil − δ then
    markiere Passage c als Außenseiter und potentielles
    Plagiat
  end if
end for
```

[6] Der Begriff Wort-Häufigkeit (en:" term frequency") ist hier unter dem Zusatz zu betrachten, dass unter anderem mit eingerechnet wird, dass bestimmte Wörter im allgemeinen häufiger vorkommen als andere.

6 Fazit

In dieser Arbeit wurde die prinzipielle Funktionsweise von intrinsischen- (nur auf einem Dokument basierenden) und extrinsischen (auf einem verdächtigen und einer Liste von potentiellen Quelldokumenten basierenden) Verfahren zur Plagiaterkennung vorgestellt. Außerdem wurden einige Verfahren an ausgewählten Beispielen näher betrachtet, um grundlegende Ideen zu erläutern und eine bessere Vorstellung der Funktionsweisen zu vermitteln.
Wie zu Beginn bereits erwähnt, ist Plagiaterkennung heutzutage mehr denn je ein wichtiges Thema, da es beinahe unmöglich ist der steigenden Anzahl von Plagiatfällen nur durch menschliche Kontrolle entgegenzuwirken. Ich bin mir sicher, dass man mit der algorithmischen Plagiaterkennung hierzu ein gutes Werkzeug geschaffen hat, dessen Potential noch längst nicht ganz ausgeschöpft ist. Schaut man sich die Entwicklung der alleine beim PAN-Workshop vorgestellten Verfahren an so stellt man fest, dass in den letzten Jahren ein stetiger Fortschritt zu verzeichnen ist. Ich bin überzeugt davon, dass es in den nächsten Jahren noch weitere Fortschritte geben wird und schon dieses Jahr einige neue Entwicklungen im Rahmen des nächsten PAN-Workshops im September 2012 vorgestellt werden.

Quellen

[1] Stein, Benno ans Lipka, N. & Meyer zu Eissen, S. Meta Analysis within Authorship Verification DEXA '08: 19th International Workshop on Database and Expert Systems Applications, IEEE Computer Society, 2008, 34-39

[2] Stamatatos, E. Plagiarism detection based on structural information Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, 2011, 1221-1230

[3] Gabriel Oberreuter, Gaston L'Huillier, Sebastián A. Ríos, and Juan D. Velásquez. Approaches for Intrinsic and External Plagiarism Detection - Notebook for PAN at CLEF 2011.

[4] Stoppwort
http://de.wikipedia.org/wiki/Stoppwort
Version vom 15.12.2011, Abgerufen am 21.05.2012

[5] N-Gramm
http://de.wikipedia.org/wiki/N-Gramm
Version vom 25.04.2012, Abgerufen am 21.05.2012

[6] Overview of PAN 2012
http://pan.webis.de/
Abgerufen am 27.05.2012

[7] Benno Stein: PAN 2011
http://www.uni-weimar.de/medien/webis/research/events/pan-11/pan11-talks/pan11-introduction.pdf
Abgerufen am 27.05.2012