Diplomverteidigung

Neue Verfahren zu intrinsischen Plagiatanalyse

Franz Coriand

Bauhaus-Universität Weimar · Fakultät Medien
Content Management & Web Technologien

7. März 2008

Übersicht

1. Einführung
2. Methoden zur Plagiaterkennung
3. Verbesserung der intrinsischen Plagiatanalyse
4. Plagiatanalyse als One-Class-Klassifikation
5. Zusammenfassung

Neue Verfahren zur intrinsischen Plagiatanalyse

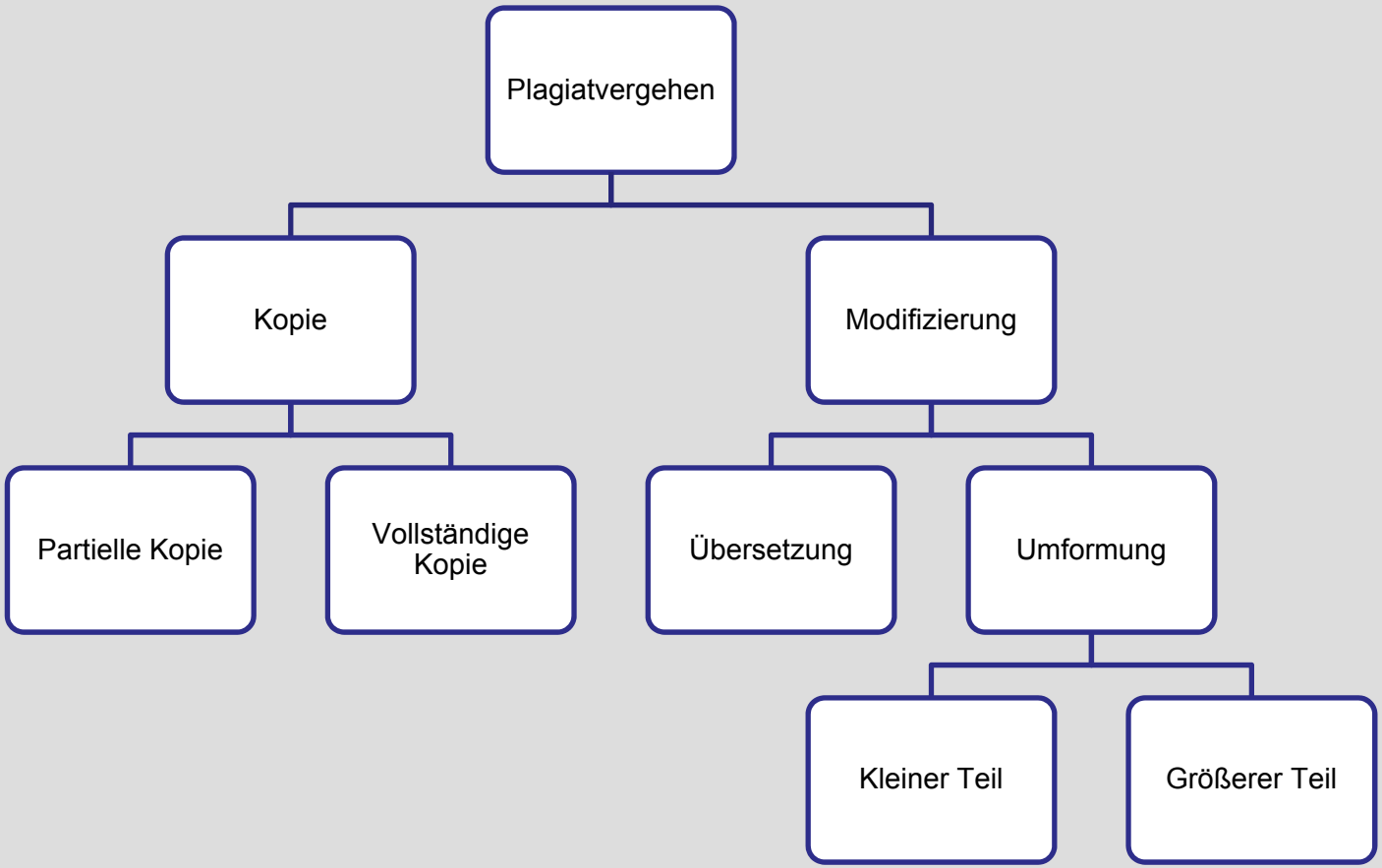
EINFÜHRUNG

Einführung

- Was ist ein Plagiat?
  - „[…] teilweise oder vollständige Übernahme eines […] Werkes unter Vorgabe eigener Urheberschaft.“ [1]
  - Allgemeine Definition – nicht auf das Gebiet der Literatur beschränkt
- Plagiat in Literatur?
  - Übernahme von Idee/Gedanken mit unzureichende bzw. fehlende Referenz/Quellenangabe
  - Problem u.a. im wissenschaftlichen Umfeld (Studienarbeiten, Dissertationen, …)

Einführung

- Arten des Plagiarismus [2]:



Neue Verfahren zur intrinsischen Plagiatanalyse

METHODEN ZUR PLAGIATERKENNUNG

Methoden zur Plagiaterkennung

- Analyseverfahren teilen sich in zwei wesentliche Teilgebiete

  1. Verfahren mit externen Informationen

  2. Verfahren ohne externen Informationen

Methoden zur Plagiaterkennung

- Verfahren mit externen Informationen
  - Ausgangspunkt:
    - Zu untersuchendes Dokument
    - Dokumentkollektion (Referenzkorpus) vorhanden
  - Ziel:
    - Suche nach Textstellen im Dokument, die sich im Korpus wiederfinden lassen
  - Verfahren:
    - Hashing-Verfahren → Finden von identischen Textabschnitten
    - Fuzzy-Fingerprinting → Finden von veränderten Textabschnitten

Methoden zur Plagiaterkennung

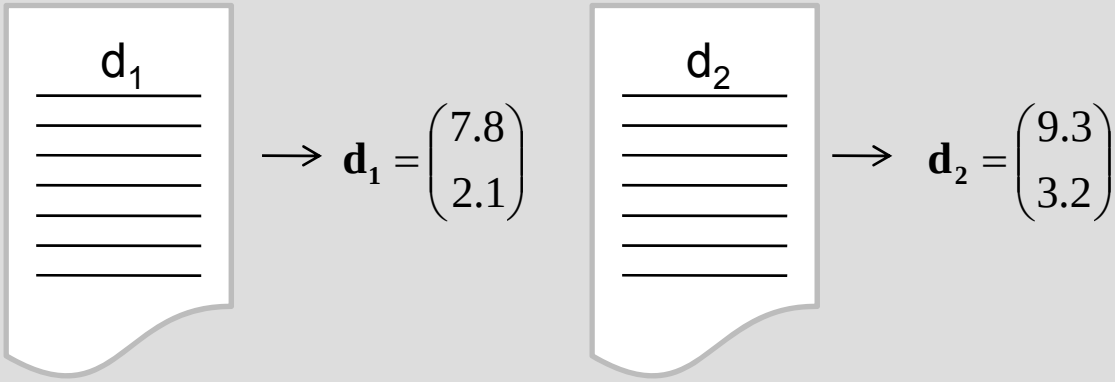
- Verfahren ohne externe Informationen = „Intrinsische Plagiatanalyse“
  - Ausgangspunkt:
    - Zu untersuchendes Dokument – KEIN Vergleichskorpus!
  - Grundidee:
    - Analyse der Abweichungen des Schreibstil von Textabschnitten zum Gesamtdokument
  - Ziel:
    - Beantwortung der Frage, ob ein Dokument plagiierte Textabschnitte enthält sowie ggf. Finden der Stellen
    - Primär nicht entscheidend, ob plagiierter Textabschnitt identisch kopiert oder verändert ist

Methoden zur Plagiaterkennung

- Strukturelle oder mathematische Erfassung eines Dokuments
- Datenstruktur: *n*-dimensionaler Vektor (Dokumentmodell)
- Instanz nennt man Feature-Vektor

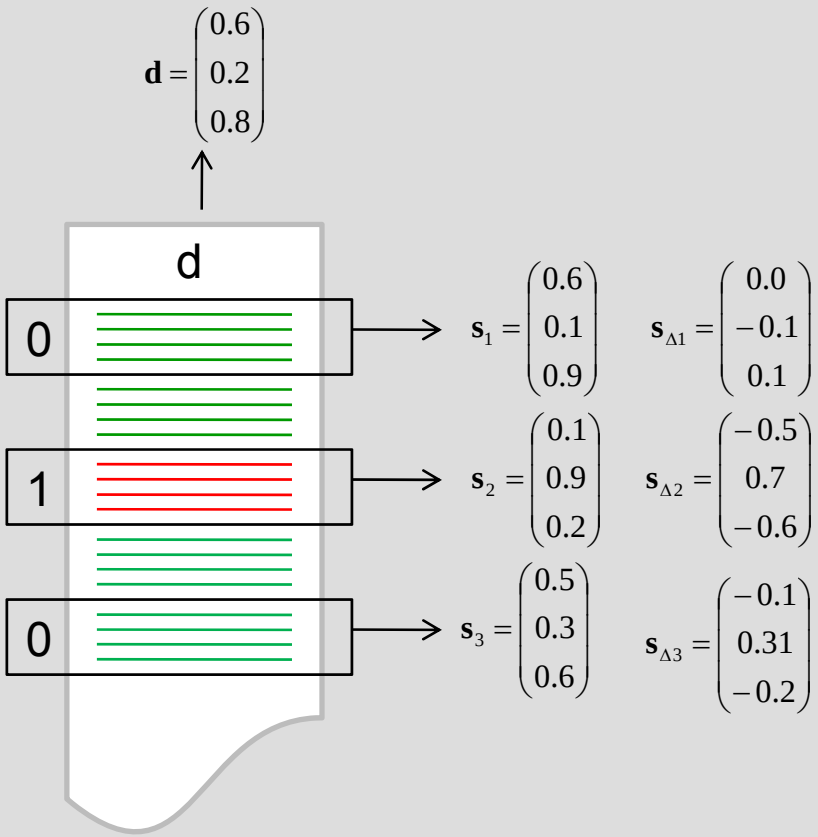
$$\mathbf{d} = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

$$\mathbf{d} = \begin{pmatrix} \text{Wörter je Satz} \\ \text{Silben je Wort} \end{pmatrix}$$



$$\rightarrow \mathbf{d}_1 = \begin{pmatrix} 7.8 \\ 2.1 \end{pmatrix} \qquad \rightarrow \mathbf{d}_2 = \begin{pmatrix} 9.3 \\ 3.2 \end{pmatrix}$$

Methoden zur Plagiaterkennung



- Finden einer Hypothese für eine möglichst genaue Vorhersage:

$$f : \mathbf{s}_\Delta \mapsto \{0,1\}$$

- Funktion *f* durch Lernverfahren des Maschinellen Lernens bestimmt

Methoden zur Plagiaterkennung

- Vorhandenes Dokumentmodell im „PicaPica“-Projekt
  - 29-dimensionaler Vektor:
    - Durchschnittliche Wortanzahl je Satz
    - Durchschnittliche Stoppworthäufigkeiten
    - Durchschnittliche Silbenanzahl je Wort
    - Honoré-Reichhaltigkeitsmaß
    - Wiener Sachtextformel
    - …

Methoden zur Plagiaterkennung

- „PicaPica“-Klassifizierungsergebisse:

Klassifikator	Recall	Precision	F-Measure
BayesNet	70,85%	76,51%	73,57%

  - Recall:
    - Anteil der gefundenen relevanten Dokumente → Vollständigkeit eines Suchergebnisses
  - Precision:
    - Anteil der relevanten Dokumente in der Ergebnismenge → Genauigkeit eines Suchergebnisses
  - F-Measure:
    - Harmonisches Mittel aus Recall und Precision

Neue Verfahren zur intrinsischen Plagiatanalyse

VERBESSERUNG DER INTRINSISCHEN ANALYSE

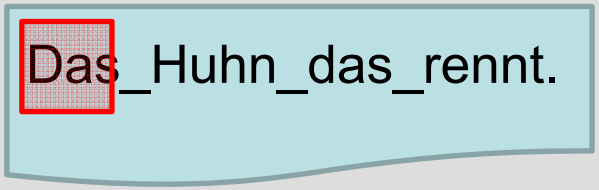
Verbesserung der intrinsischen Analyse

- Grundidee:
  - Verwendung von Objekt-*n*-Grammen zur Analyse des Plagiatverdachts
  - Definition *n*-Gramm:

    Sei Σ ein endliches Alphabet und sei *n* eine positive ganze Zahl. Dann ist ein *n*-Gramm ein Wort ω über dem Alphabet Σ, d.h. $\omega = (\omega_1, \ldots, \omega_n) \in \Sigma^n$.

Verbesserung der intrinsischen Analyse

- Verständnisbeispiel mit Buchstaben:



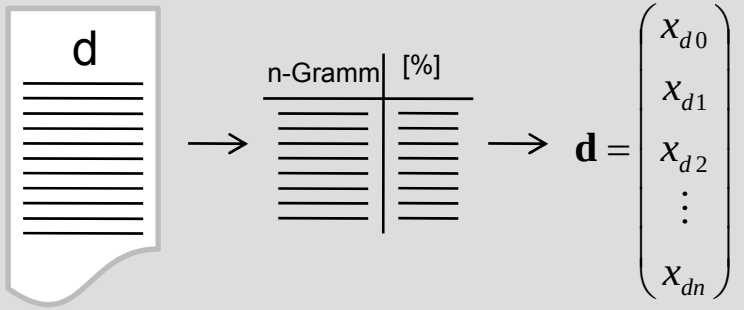
- Adaption auf Wörter und Wortklassen

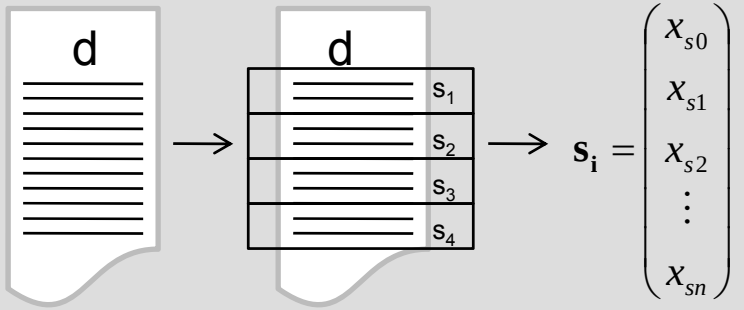
Buchstaben-3-Gramm	Abs. Häufigkeit	Rel. Häufigkeit
das	2	12%
as_	2	12%
s_h	1	6%
_hu	1	6%
huh	1	6%
uhn	1	6%
hn_	1	6%
n_d	1	6%
_da	1	6%
s_r	1	6%
_re	1	6%
ren	1	6%
enn	1	6%
nnt	1	6%
nt.	1	6%

Verbesserung der intrinsischen Analyse

- Verarbeitungsalgorithmus:

1. *n*-Gramm-Tabelle von d erstellen → Feature-Vektor **d**
2. Zerlegung von d in gleich große Textabschnitte s
3. *n*-Gramm-Tabellen für Textabschnitte s_i erstellen → Feature-Vektoren $\mathbf{s_i}$
4. Differenzbildung von **d** und $\mathbf{s_i}$ → Differenzvektor $\mathbf{s_{\Delta i}}$





$$\mathbf{s_{\Delta i}} = \begin{pmatrix} g(x_{d0}, x_{s0}) \\ g(x_{d1}, x_{s1}) \\ g(x_{d2}, x_{s2}) \\ \vdots \\ g(x_{dn}, x_{sn}) \end{pmatrix} \text{ mit } g(a,b) = \frac{a-b}{a} \; (a \neq 0)$$

Verbesserung der intrinsischen Analyse

- Evaluation von Anwendungsparametern:
  - Gesuchte Parameter:
    - Größe von Textabschnitten
    - *n*-Grammgröße der Merkmale
    - Sinnvolle Anzahl von Features
    - Laufzeitverhalten von *n*-Grammen und Klassifikationen
  - Grundlage:
    - Trainingskorpus (aus 70 Dissertationen wurden 690 Plagiate und 70 Originale erzeugt)

Verbesserung der intrinsischen Analyse

- Parameterergebnisse:
  - Größe von Textabschnitten:
    - 500 Wörter
  - N-Grammgröße:
    - Buchstaben-3-Gramme
    - Wort-2-Gramme
    - Wortklassen-2-Gramme
  - Feature-Anzahl:
    - 250 häufigsten Objekt-*n*-Gramme
  - Laufzeitverhalten:
    - Buchstaben- bzw. Wort-*n*-Gramme um Faktor 30 bzw. 60 schneller als Wortklassen-*n*-Gramme

Verbesserung der intrinsischen Analyse

- Klassifizierungsergebnisse für BayesNet:

Modell	Recall	Precision	F-Measure	Δ-Recall	Δ-Precision
PicaPica	70,85%	76,51%	76,57%		
Buchstaben	84,10%	88,95%	86,46%	13,25%	12,44%
Wörter	81,45%	83,07%	82,25%	10,60%	6,56%
Wortklassen	78,25%	81,89%	80,03%	7,40%	5,38%
Buchstaben + Wörter	88,50%	90,54%	89,51%	17,65%	14,03%
Buchstaben + Wortklassen	85,85%	89,29%	87,54%	15,00%	12,78%
Wörter + Wortklassen	87,25%	86,69%	86,97%	16,40%	10,17%
Buchstaben + Wörter +Wortklassen	89,30%	90,61%	89,95%	18,45%	14,10%
Pica + Buchstaben	85,15%	88,28%	86,69%	14,30%	11,77%
Pica + Wörter	86,30%	85,11%	85,70%	15,45%	8,60%
Pica + Wortklassen	80,35%	84,27%	82,26%	9,50%	7,76%
Pica + Buchstaben + Wörter	89,20%	90,19%	89,69%	18,35%	13,68%
Pica + Buchstaben + Wortklassen	86,60%	89,79%	88,16%	15,75%	13,28%
Pica + Wörter + Wortklassen	86,70%	86,70%	86,70%	15,85%	10,19%
Pica + Buchstaben + Wörter + Wortklassen	88,90%	90,76%	89,82%	18,05%	14,25%

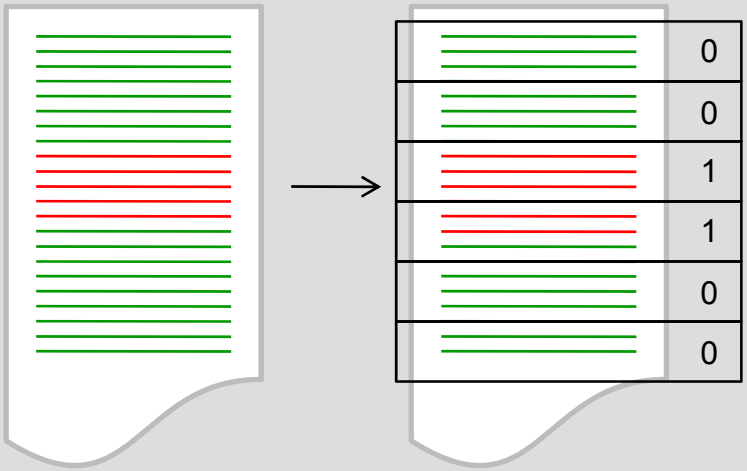
Verbesserung der intrinsischen Analyse

- Grundproblematik:
  - Intrinsische Analyse klassifiziert spezielle Textabschnitte
  - Verallgemeinerung der Fragestellung:
    - Enthält ein Dokument ein Plagiat?
  - Trotz hohe Precision für Textabschnitte → „falsch positiv“-Klassifizierung möglich → niedrige Precision für Dokumentklassifikation

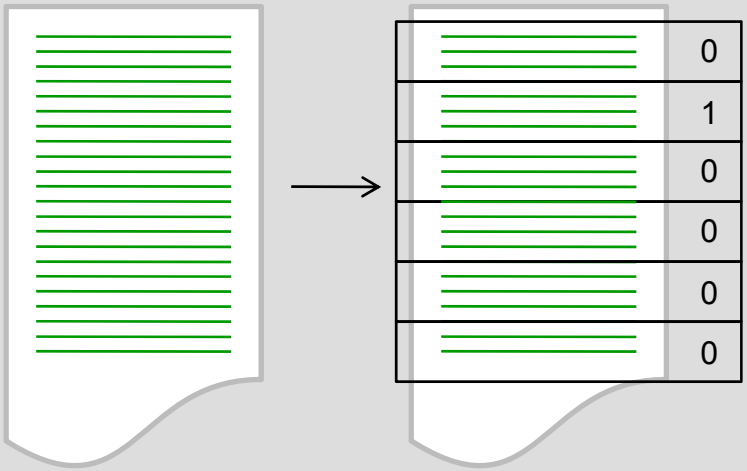
Modell	Recall	Precision	F-Measure
Pica + Buchstaben + Wörter	96,7%	51,8%	67,4%
Pica + Buchstaben + Wörter + Wortklassen	96,7%	58,0%	72,5%

Verbesserung der intrinsischen Analyse

- Idealer Fall:
  - Dokument enthält plagiierte Abschnitte → diese werden erkannt



- Problematischer Fall:
  - Dokument enthält keine plagiierte Abschnitte → Verfahren klassifiziert jedoch Textabschnitt als Plagiat

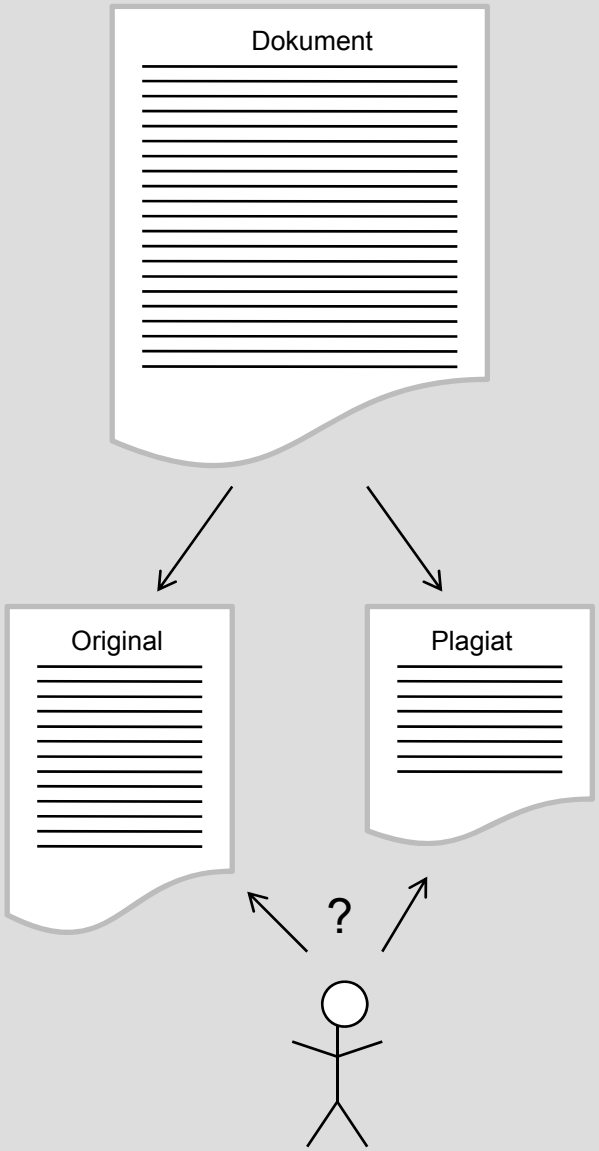


Neue Verfahren zur intrinsischen Plagiatanalyse

PLAGIATANALYSE ALS ONE-CLASS-KLASSIFIKATION

Plagiatanalyse als One-Class-Klassifikation

- Ausgangspunkt:
  - Zwei Textmengen, erzeugt durch intrinsischen Analyse → Textstellen vom Typ Plagiat und Original
- Ziel:
  - Klärung der Frage, ob beide Textmengen (eventuell doch) vom selben Autor stammen
- Umsetzung:
  - Klärung durch Algorithmus vorgestellt von Moshe Koppel [3] → Anwendung als Post-Prozess



Plagiatanalyse als One-Class-Klassifikation

- Verfahren nach Koppel:
  - Erstellen eines Feature-Vektors mit häufig verwendeten Wörter der beiden Textmengen
  - Iteratives Entfernen der individuellen Wörter aus Vektor

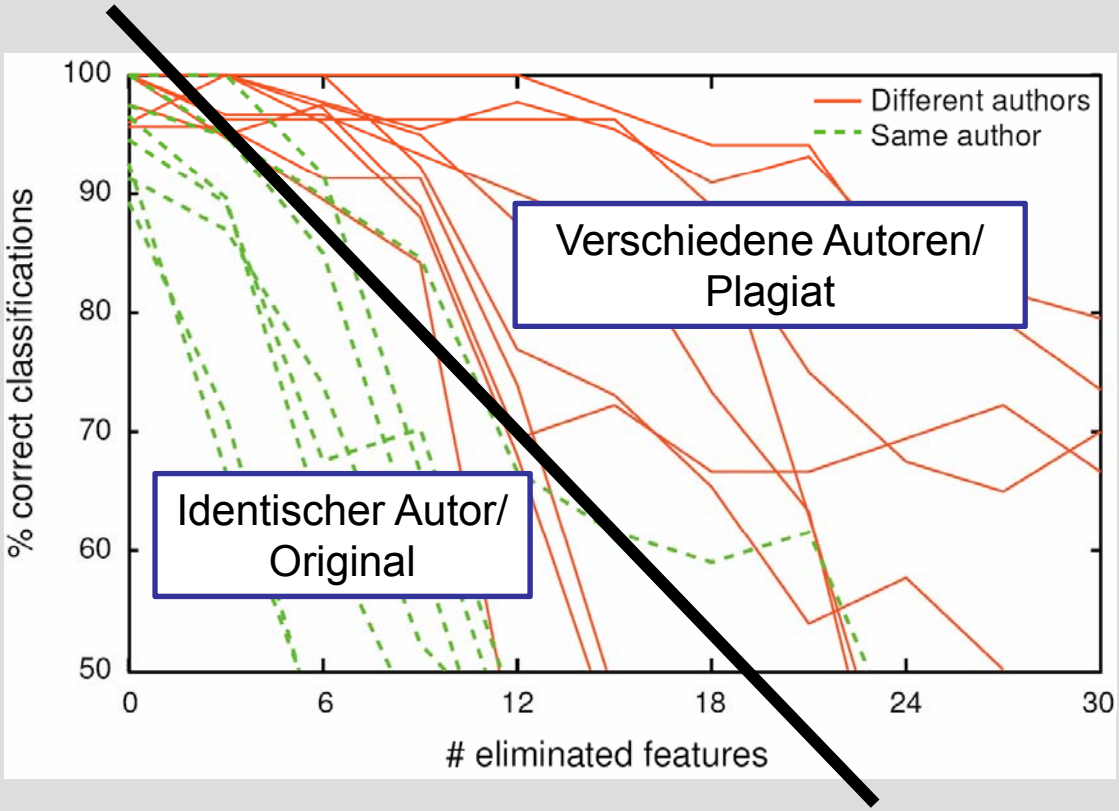
d_1: Pippi und Herr Nielson spielen mit den anderen Kindern in der Villa.

d_2: Karlsson fliegt zu seinem Haus, wo die anderen Kinder auf ihn warten.

  - Offenlegung von unbewussten Satz- und Wortkonstruktionen → Lernen
  - Klassifizierungsergebnisse nach jedem Entfernungsschritt? → Lernkurve entsteht

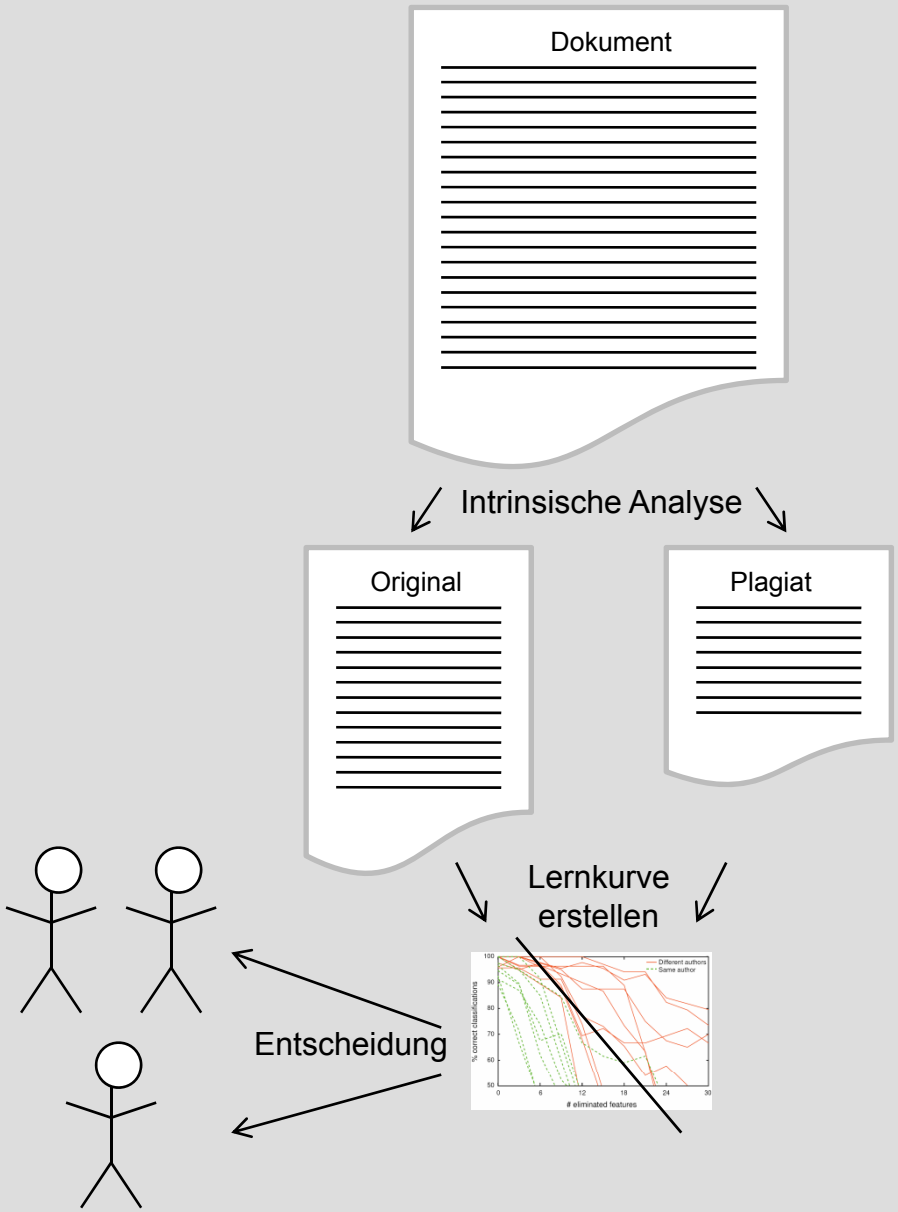
Plagiatanalyse als One-Class-Klassifikation

- Klassifizierungsergebnisse einer Trainingsmenge [4]:



Plagiatanalyse als One-Class-Klassifikation

- Anwendung des Verfahrens auf Ergebnisse der intrinsischen Analyse



Plagiatanalyse als One-Class-Klassifikation

- Verfahren kann keine neuen Plagiatsfälle entdecken
- Verfahren bekräftigt oder verwirft die Annahme, dass beide Textmengen von verschiedenen Autoren stammen

→Anzahl der als Plagiat klassifizierten Dokumente wird geringer, aber präziser (Precision steigt)

Plagiatanalyse als One-Class-Klassifikation

- Anwendungsergebnisse mit BayesNet:

Modell	Recall	Precision	F-Measure	Δ-Recall	Δ-Precision
Pica + Buchstaben + Wörter	26,7%	72,7%	39,0%	-70,0%	20,9%
Pica + Buchstabe + Wörter + Wortklassen	20,0%	75,0%	31,6%	-76,7%	17,0%

  - Recall (Vollständigkeit des Suchergebnisses) sinkt
  - Precision (Genauigkeit des Suchergebnisses) steigt

Neue Verfahren zur intrinsischen Plagiatanalyse

ZUSAMMENFASSUNG

Zusammenfassung

- Entwicklung neuer Dokumentmodelle zur intrinsischen Plagiatanalyse:
  - Buchstaben-*n*-Gramm
  - Wort-*n*-Gramm
  - Wortklassen-*n*-Gramm
- Evaluation der Modelle sowie deren Kombinationen
  - Verbesserung der Klassifizierungsergebnisse im Vergleich zum „PicaPica“-Dokumentmodell bei Verwendung von BayesNet-Klassifikator

Zusammenfassung

- Evaluation der Modelle sowie deren Kombinationen
  - Wortklassen-*n*-Gramme sind nicht als Modell geeignet, da Laufzeitverhalten zu schlecht bei gleichen Klassifizierungsergebnissen
- Anwendung und Evaluierung einer One-Class-Klassifizierung als Post-Prozess
  - Wesentliche Verbesserung der Precision der Klassifikationsergebnisse bei sinkendem Recall-Wert

Neue Verfahren zur intrinsischen Plagiatanalyse

ENDE

Quellen

[1] Zeitverlag (Hrsg.): Die Zeit, Das Lexikon in 20 Bänden. Zeitverlag Gerd Bucerius GmbH & Co. KG, 2005. – ISBN 3-422-17571-0

[2] Meyer zu Eissen, Sven; Stein, Benno; Kulig, Marion: Plagiarism Detection without Reference Collections. In: Decker, Reinhold (Hrsg.); Lenz, Hans J. (Hrsg.): Advances in Data Analysis, Springer, 2007. – ISBN 978-33540-70980-0. S. 359-366

[3] Koppel, Moshe; Schler, Jonathan: Authorship verification as a one-class classification problem. In ICML `04: Proceedings of the twenty-first international conference on Machine learning. New York, NY, USA: ACM Press, 2004. – ISBN 1-58113-828-5, S. 62

[4] Stein, Benno; Meyer zu Eisen, Sven: Intrinsic Plagiarism Analysis with Meta-Learning. In: Stein, Benno (Hrsg.); Koppel, Moshe (Hrsg.); Stamatatos, Efstathios (Hrsg.): SIGIR Workshop Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07), CEUR-WS.org, Juli 2007