Bauhaus-Universität Weimar
Fakultät Medien

Diplomarbeit

Intrinsische Plagiaterkennung am Beispiel einer Artikelsammlung

Marion Kulig

Weimar, 16. August 2006

Gliederung

1. Motivation
2. Algorithmen zu Erkennung von Plagiatvergehen
3. Methoden zur Quantifizierung von Stil
4. Ein Korpus zur Evaluierung von Plagiaterkennungsalgorithmen
5. Evaluierung
6. Ausblick und Zusammenfassung

Motivation

Plagiat = Diebstahl fremden geistigen Eigentums

[Quelle: Plagiat Design]

[Quelle: Plagiat Literatur]

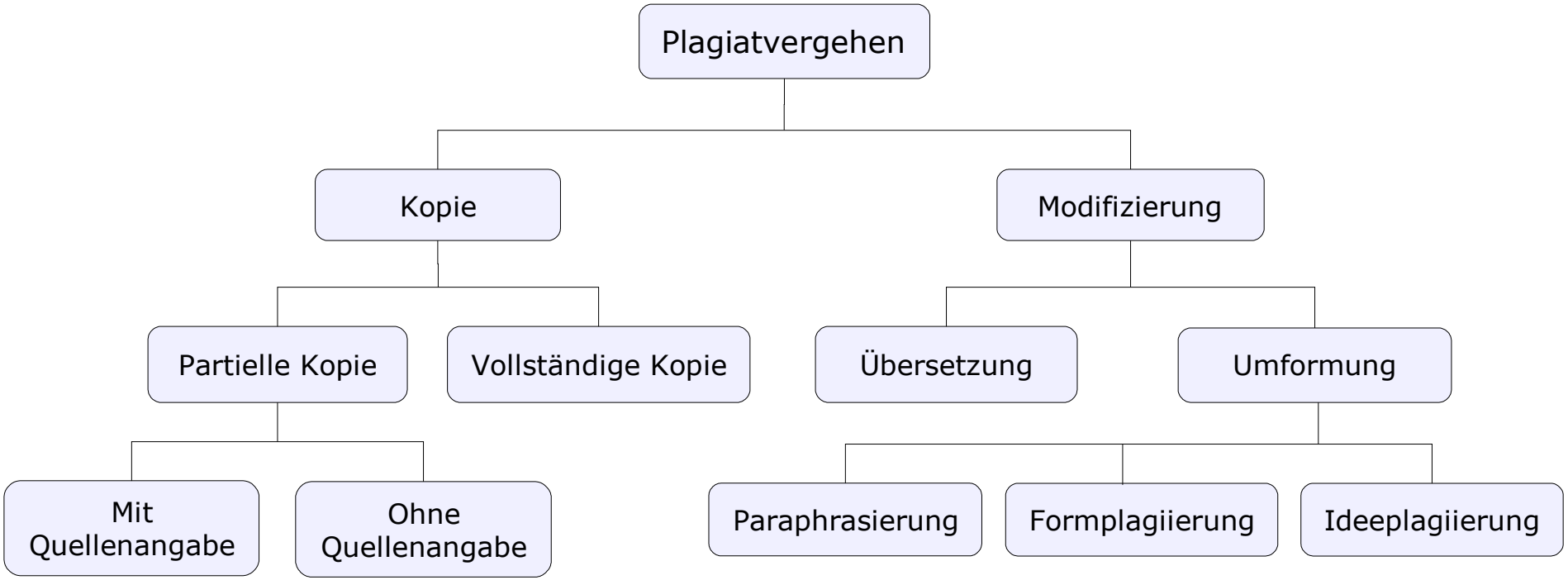
[Quelle: Plagiat Technik]

[Quelle: Plagiat Kunst]

Motivation

Plagiat = Diebstahl fremden geistigen Eigentums

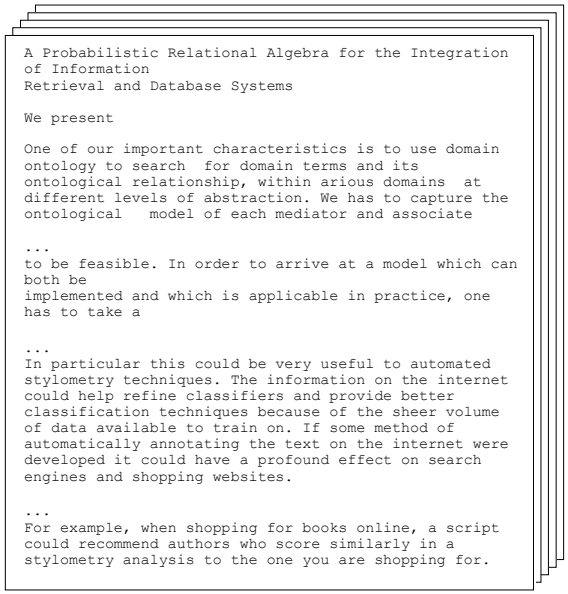
Taxonomie der Plagiate



Plagiatvergehen und deren Nachweis

Aufgabe: Finden von plagiierten Abschnitten

(n Korpusdokumente, c Abschnitte pro Dok.)


A Probabilistic Relational Algebra for the Integration of Information
Retrieval and Database Systems

We present

One of our important characteristics is to use domain ontology to search for domain terms and its ontological relationship, within arious domains at different levels of abstraction. We has to capture the ontological model of each mediator and associate

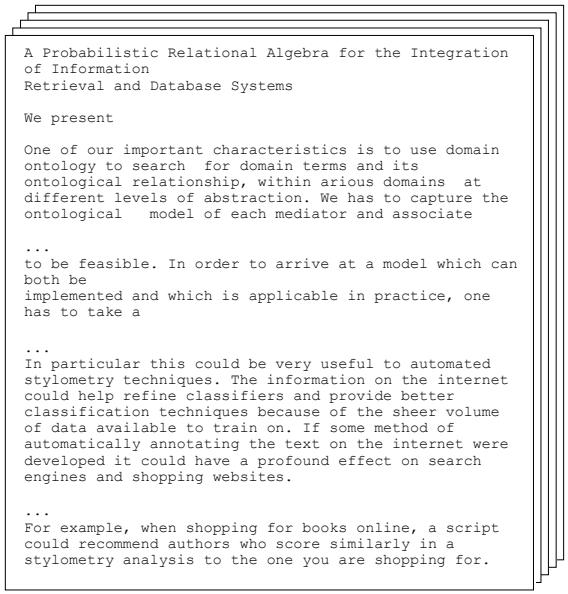
...
to be feasible. In order to arrive at a model which can both be
implemented and which is applicable in practice, one has to take a

...
In particular this could be very useful to automated stylometry techniques. The information on the internet could help refine classifiers and provide better classification techniques because of the sheer volume of data available to train on. If some method of automatically annotating the text on the internet were developed it could have a profound effect on search engines and shopping websites.

...
For example, when shopping for books online, a script could recommend authors who score similarly in a stylometry analysis to the one you are shopping for.

Korpus

Plagiatvergehen und deren Nachweis



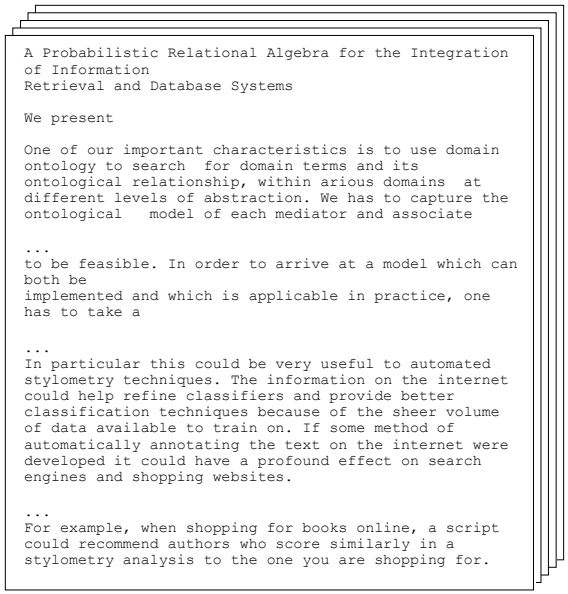
Korpus

Aufgabe: Finden von plagiierten Abschnitten

(n Korpusdokumente, c Abschnitte pro Dok.)

- Paarweiser Abschnittsvergleich $O(n \cdot c^2)$

Plagiatvergehen und deren Nachweis



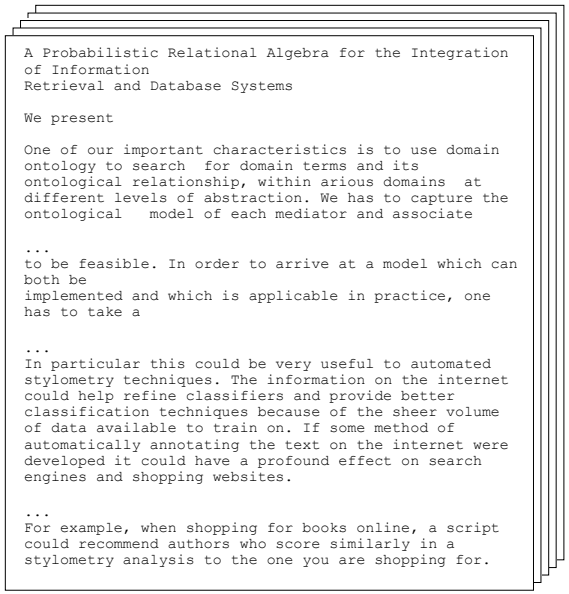
Korpus

Aufgabe: Finden von plagiierten Abschnitten

(n Korpusdokumente, c Abschnitte pro Dok.)

- Paarweiser Abschnittsvergleich $O(n \cdot c^2)$
- Hashing $O(n \cdot c)$

Plagiatvergehen und deren Nachweis


A Probabilistic Relational Algebra for the Integration of Information
Retrieval and Database Systems

We present

One of our important characteristics is to use domain ontology to search for domain terms and its ontological relationship, within arious domains at different levels of abstraction. We has to capture the ontological model of each mediator and associate

...
to be feasible. In order to arrive at a model which can both be
implemented and which is applicable in practice, one has to take a

...
In particular this could be very useful to automated stylometry techniques. The information on the internet could help refine classifiers and provide better classification techniques because of the sheer volume of data available to train on. If some method of automatically annotating the text on the internet were developed it could have a profound effect on search engines and shopping websites.

...
For example, when shopping for books online, a script could recommend authors who score similarly in a stylometry analysis to the one you are shopping for.

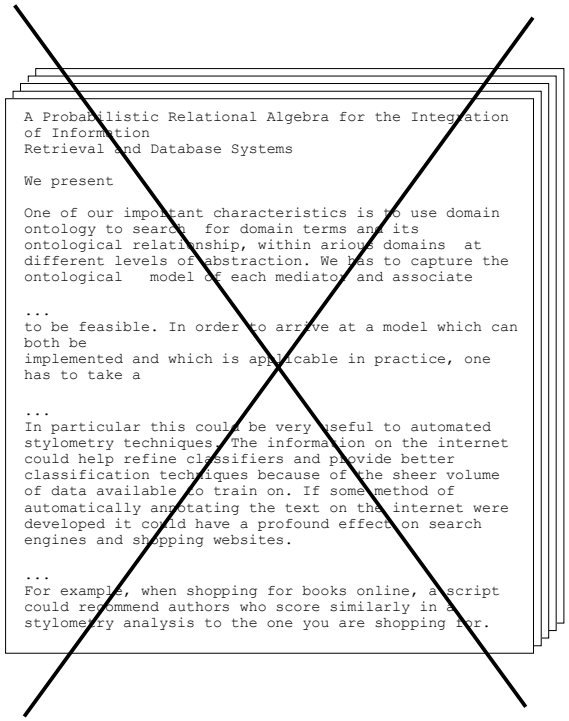
Korpus

Aufgabe: Finden von plagiierten Abschnitten

(n Korpusdokumente, c Abschnitte pro Dok.)

- Paarweiser Abschnittsvergleich $O(n \cdot c^2)$
- Hashing $O(n \cdot c)$
- Fuzzy-Fingerprinting $O(n \cdot c)$

Plagiatvergehen und deren Nachweis



Korpus

Aufgabe: Finden von plagiierten Abschnitten

(n Korpusdokumente, c Abschnitte pro Dok.)

- Paarweiser Abschnittsvergleich $O(n \cdot c^2)$
- Hashing $O(n \cdot c)$
- Fuzzy-Fingerprinting $O(n \cdot c)$

?

Stilanalyse

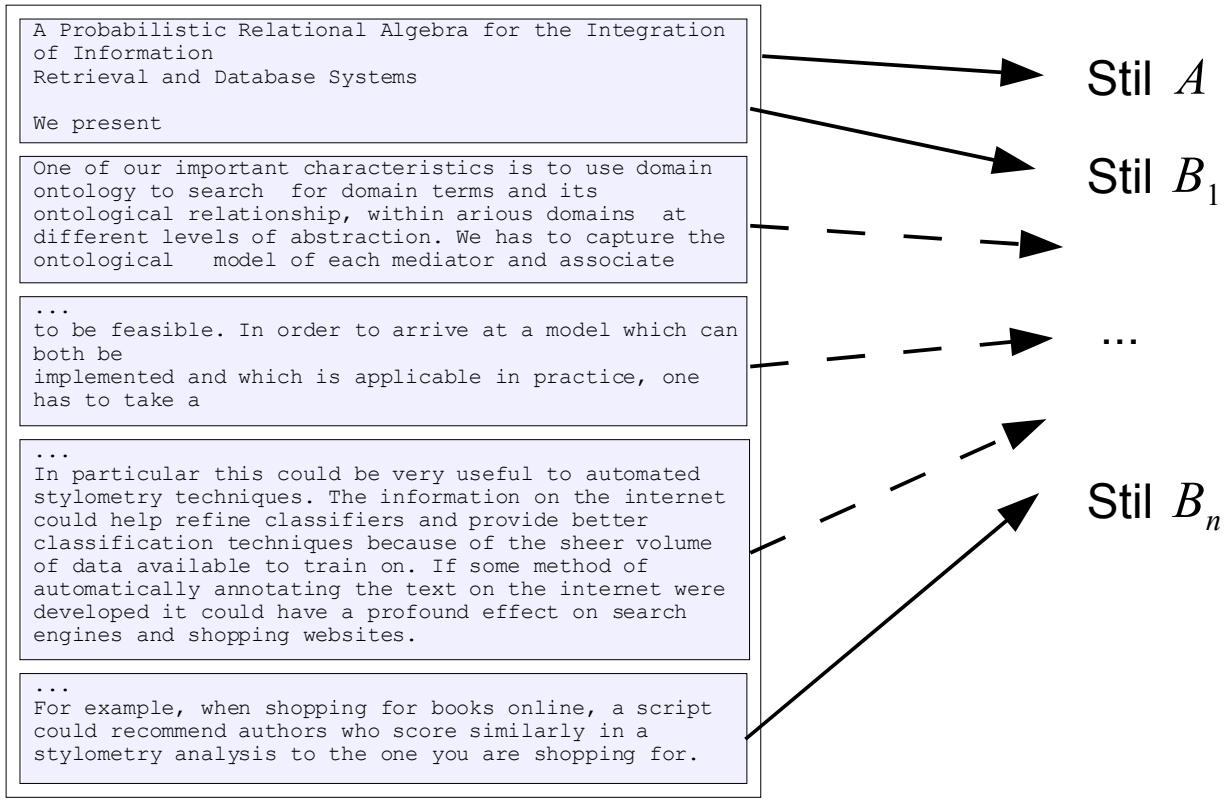
```
A Probabilistic Relational Algebra for the Integration
of Information
Retrieval and Database Systems

We present

One of our important characteristics is to use domain
ontology to search  for domain terms and its
ontological relationship, within arious domains  at
different levels of abstraction. We has to capture the
ontological   model of each mediator and associate

...
to be feasible. In order to arrive at a model which can
both be
implemented and which is applicable in practice, one
has to take a

...
In particular this could be very useful to automated
stylometry techniques. The information on the internet
could help refine classifiers and provide better
classification techniques because of the sheer volume
of data available to train on. If some method of
automatically annotating the text on the internet were
developed it could have a profound effect on search
engines and shopping websites.

...
For example, when shopping for books online, a script
could recommend authors who score similarly in a
stylometry analysis to the one you are shopping for.
```

Stilanalyse

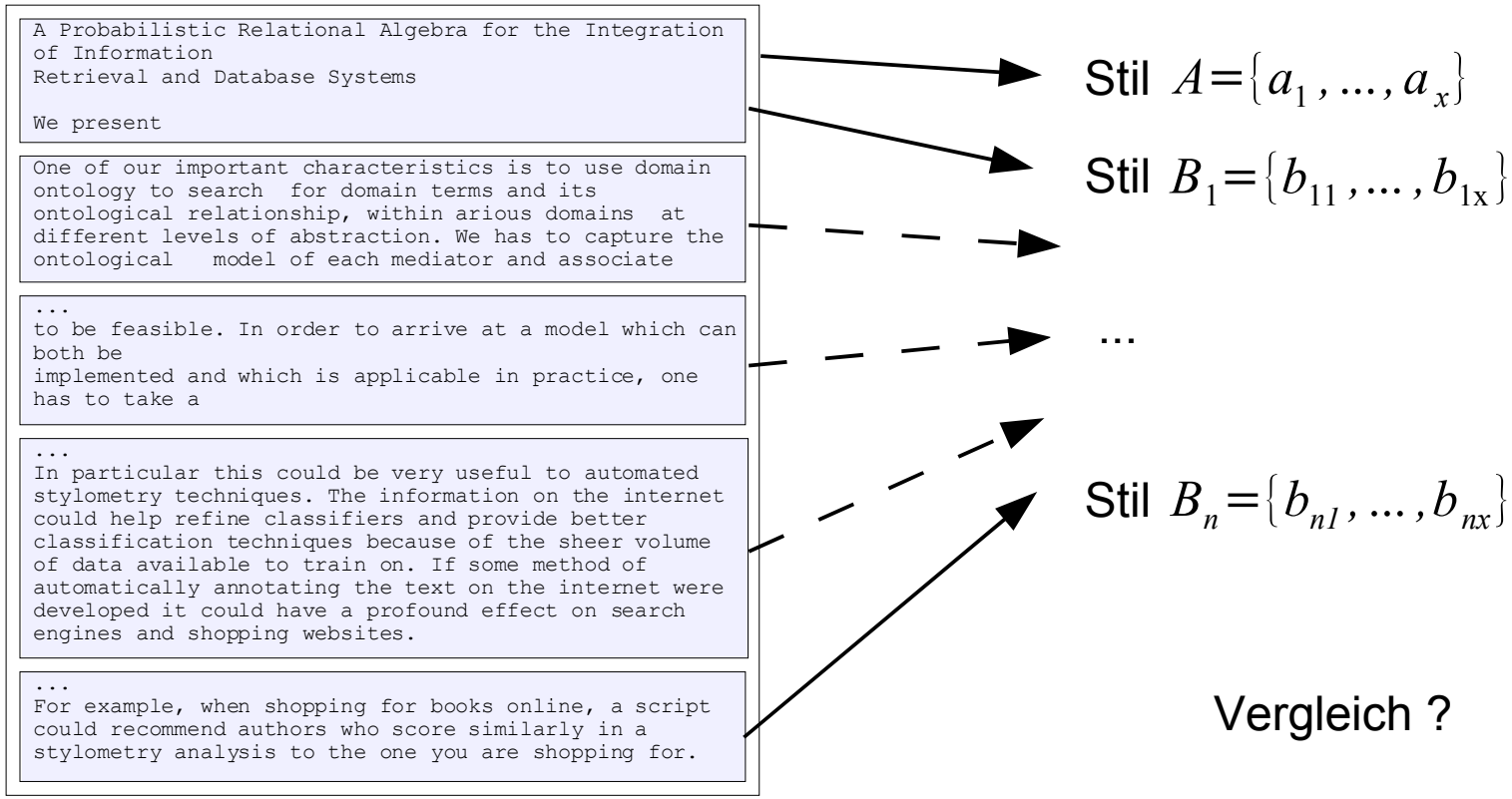
```
A Probabilistic Relational Algebra for the Integration
of Information
Retrieval and Database Systems

We present

One of our important characteristics is to use domain
ontology to search  for domain terms and its
ontological relationship, within arious domains  at
different levels of abstraction. We has to capture the
ontological   model of each mediator and associate

...
to be feasible. In order to arrive at a model which can
both be
implemented and which is applicable in practice, one
has to take a

...
In particular this could be very useful to automated
stylometry techniques. The information on the internet
could help refine classifiers and provide better
classification techniques because of the sheer volume
of data available to train on. If some method of
automatically annotating the text on the internet were
developed it could have a profound effect on search
engines and shopping websites.

...
For example, when shopping for books online, a script
could recommend authors who score similarly in a
stylometry analysis to the one you are shopping for.
```

→ Stil A

Stilanalyse



Stilanalyse



Methoden zur Quantifizierung von Stil

Stilmerkmale

- **Wortvielfalt**
  - Honoré's R (1979)

$$R = \frac{100 \cdot \log N}{1 - \frac{V_1}{V}}$$

  - Yule's K (1944)

$$K = \frac{10^4 \cdot \left(\sum_{i=1}^{V} i^2 V_i\right) - V}{V^2}$$

- **Lesbarkeit**
  - **Index** — Flesch's Reading Ease (1948)

$$R = 206{,}835 - \frac{1{,}015 \cdot W}{S} - \frac{84{,}6 \cdot T}{W}$$

  - **Lesegrad** — Gunning Fog Index (1952)

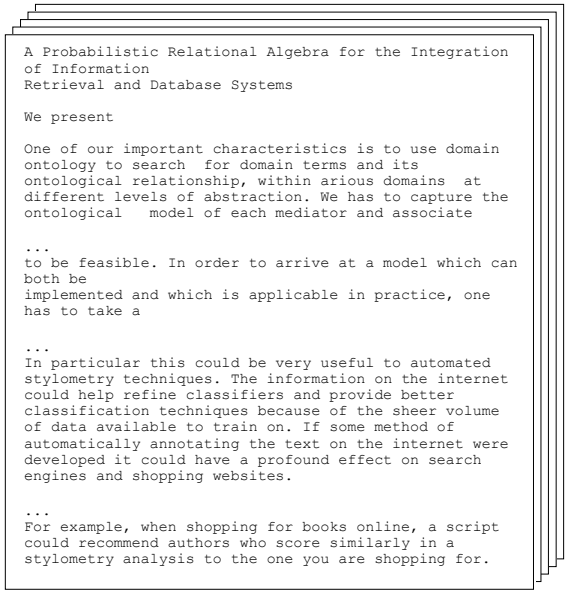
$$R = \left(\frac{W}{S} + \frac{100 \cdot W_{T \geq 3}}{W}\right) \cdot 0{,}4$$

- **Wortarten**

Übersicht Stilmerkmale

Maß	Stilmerkmal
einfache Stilmerkmale	Ø Satzlänge
	Ø Stoppwortanteil
	Ø Silbenanzahl pro Wort
Lesbarkeit	Flesch-Reading-Ease-Index FREI
	Flesch-Kincaid-Grade-Level FKGL
	Gunning-Fog-Index GFI
	Dale-Chall-Index DCI
Wortvielfalt	Honoré's R
	Yule's K
	Kullback-Leibler-Divergenz KLD
	Average-Word-Frequency-Class AWFC

Average-Word-Frequency-Class


A Probabilistic Relational Algebra for the Integration of Information
Retrieval and Database Systems

We present

One of our important characteristics is to use domain ontology to search for domain terms and its ontological relationship, within arious domains at different levels of abstraction. We has to capture the ontological model of each mediator and associate

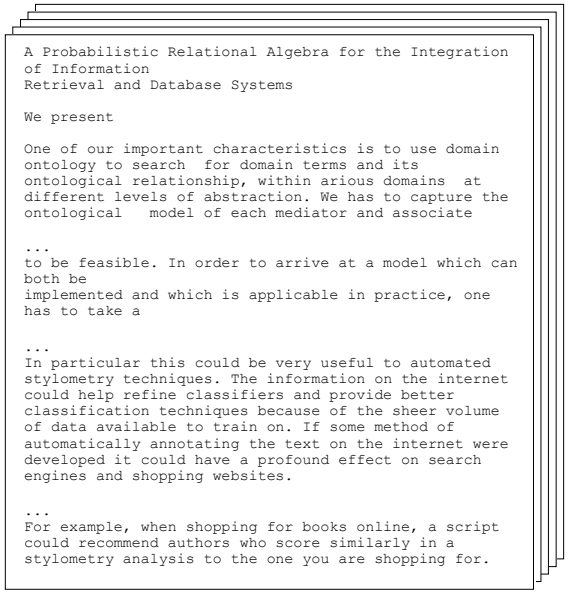
...
to be feasible. In order to arrive at a model which can both be
implemented and which is applicable in practice, one has to take a

...
In particular this could be very useful to automated stylometry techniques. The information on the internet could help refine classifiers and provide better classification techniques because of the sheer volume of data available to train on. If some method of automatically annotating the text on the internet were developed it could have a profound effect on search engines and shopping websites.

...
For example, when shopping for books online, a script could recommend authors who score similarly in a stylometry analysis to the one you are shopping for.

Korpus

Average-Word-Frequency-Class


A Probabilistic Relational Algebra for the Integration of Information
Retrieval and Database Systems

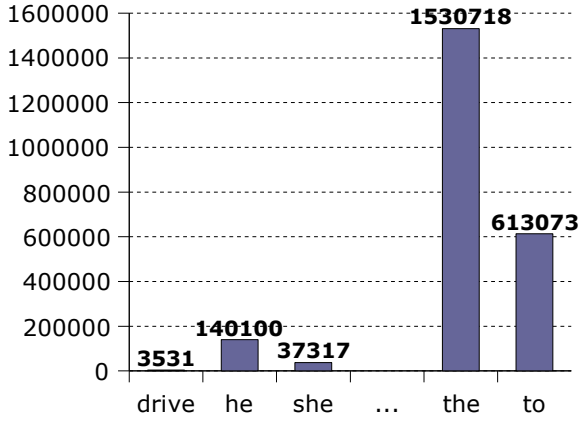
We present

One of our important characteristics is to use domain ontology to search for domain terms and its ontological relationship, within arious domains at different levels of abstraction. We has to capture the ontological model of each mediator and associate

...
to be feasible. In order to arrive at a model which can both be
implemented and which is applicable in practice, one has to take a

...
In particular this could be very useful to automated stylometry techniques. The information on the internet could help refine classifiers and provide better classification techniques because of the sheer volume of data available to train on. If some method of automatically annotating the text on the internet were developed it could have a profound effect on search engines and shopping websites.

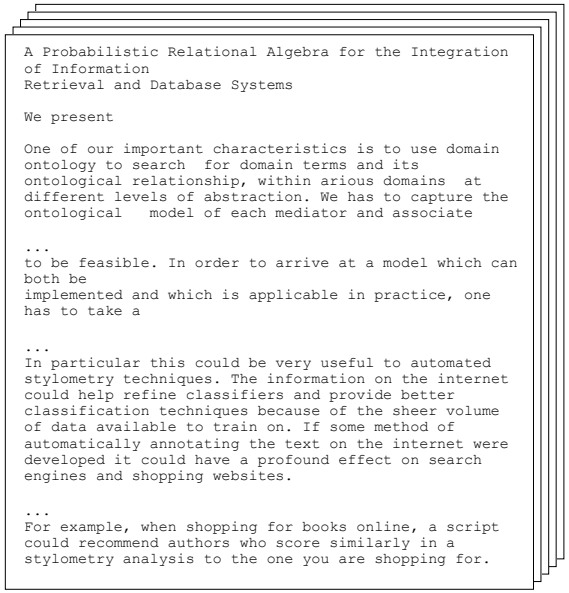
...
For example, when shopping for books online, a script could recommend authors who score similarly in a stylometry analysis to the one you are shopping for.



Korpus

Worthäufigkeiten jedes Wortes

Average-Word-Frequency-Class


A Probabilistic Relational Algebra for the Integration of Information
Retrieval and Database Systems

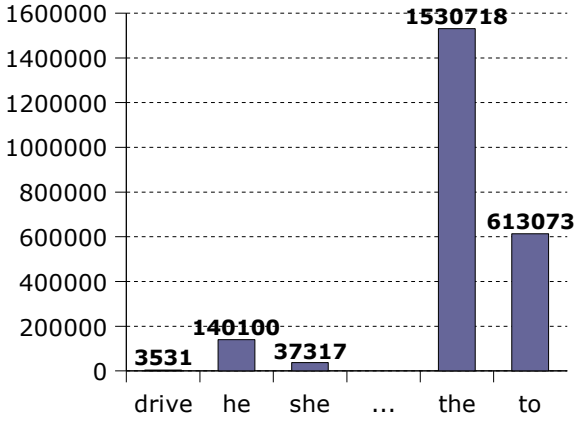
We present

One of our important characteristics is to use domain ontology to search for domain terms and its ontological relationship, within arious domains at different levels of abstraction. We has to capture the ontological model of each mediator and associate

...
to be feasible. In order to arrive at a model which can both be
implemented and which is applicable in practice, one has to take a

...
In particular this could be very useful to automated stylometry techniques. The information on the internet could help refine classifiers and provide better classification techniques because of the sheer volume of data available to train on. If some method of automatically annotating the text on the internet were developed it could have a profound effect on search engines and shopping websites.

...
For example, when shopping for books online, a script could recommend authors who score similarly in a stylometry analysis to the one you are shopping for.



the	0
to	1
he	3
she	5
drive	8
...	
polyethnicity	18
...	
...	19

Korpus

Worthäufigkeiten jedes Wortes

Häufigkeitsklasse

Average-Word-Frequency-Class

Maß für die durchschnittliche Worthäufigkeitsklasse eines Dokuments

C ... Textkorpus

$f(w)$ ... Worthäufigkeit eines Wortes $w \in C$

$c(w)$ ... Häufigkeitsklasse eines Wortes

$$c(w) = \lfloor \log_2(f(w^*)/f(w)) \rfloor$$

w^* ... häufigstes Wort in C, hier das Wort „the“ mit entsprechender Häufigkeitsklasse 0. Das seltenste hat die 18.

Ein Korpus zur Erkennung von Plagiaterkennungsalgorithmen

- 100 Dokumente aus ACM-Bibliothek im PDF-Format
- Aus 3 Themengebieten (Information Retrieval, CSCW, und Plagiarism)
- Übertragung ins XML-Format
- Einfügen von plagiierten Abschnitten (copied / modified)
- Kennzeichnung des Ursprungs

```
<?xml version="1.0" encoding="UTF-8"?>
<document xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:noNamespaceSchemaLocation="collection_schema.xsd"
 documentSource="http://portal.acm.org/citation.cfm?doid=239041.239045">

 A Probabilistic Relational Algebra for the Integration of
 Information Retrieval and Database Systems NORBERT FUHR and
 THOMAS RO" LLEKE University of Dortmund

 We present
 ...

 <inserted
   source="http://portal.acm.org/citation.cfm?doid=375212.375229"
   type="modified">
   One of our important characteristics is to use domain ontology to
   search for domain terms and its ontological relationship, within
   arious domains at different levels of abstraction. We has to
   capture the ontological model of each mediator and associate
   ...

 </inserted>

 ...
 to be feasible. In order to arrive at a model which can both be
 implemented and which is applicable in practice, one has to take a
 ...

 <inserted source="..." type="...
   ...
 </inserted>

 ...
 ...
 ...

 ...

</document>
```

Fragestellungen

1. Können plagiierte Abschnitte intrinsisch durch den Einsatz von Stilmerkmalen erkannt werden?

Fragestellungen

1. Können plagiierte Abschnitte intrinsisch durch den Einsatz von Stilmerkmalen erkannt werden?
2. Wie gut können plagiierte Abschnitte erkannt werden?

Fragestellungen

1. Können plagiierte Abschnitte intrinsisch durch den Einsatz von Stilmerkmalen erkannt werden?
2. Wie gut können plagiierte Abschnitte erkannt werden?
3. Gibt es signifikante Unterschiede zwischen verschiedenen Themengebieten? Ist eine Themenabhängigkeit vorhanden?

Fragestellungen

1. Können plagiierte Abschnitte intrinsisch durch den Einsatz von Stilmerkmalen erkannt werden?
2. Wie gut können plagiierte Abschnitte erkannt werden?
3. Gibt es signifikante Unterschiede zwischen verschiedenen Themengebieten? Ist eine Themenabhängigkeit vorhanden?
4. Ab welchem plagiierten Anteil können plagiierte Abschnitte zuverlässig bestimmt werden?

Fragestellungen

1. Können plagiierte Abschnitte intrinsisch durch den Einsatz von Stilmerkmalen erkannt werden?
2. Wie gut können plagiierte Abschnitte erkannt werden?
3. Gibt es signifikante Unterschiede zwischen verschiedenen Themengebieten? Ist eine Themenabhängigkeit vorhanden?
4. Ab welchem plagiierten Anteil können plagiierte Abschnitte zuverlässig bestimmt werden?
5. Gibt es Stilmerkmale, die besonders gut und besonders schlecht funktionieren?

Fragestellungen

1. Können plagiierte Abschnitte intrinsisch durch den Einsatz von Stilmerkmalen erkannt werden?
2. Wie gut können plagiierte Abschnitte erkannt werden?
3. Gibt es signifikante Unterschiede zwischen verschiedenen Themengebieten? Ist eine Themenabhängigkeit vorhanden?
4. Ab welchem plagiierten Anteil können plagiierte Abschnitte zuverlässig bestimmt werden?
5. Gibt es Stilmerkmale, die besonders gut und besonders schlecht funktionieren?
6. Wie stabil funktionieren die Stilmerkmale in Abhängigkeit der Textlänge?

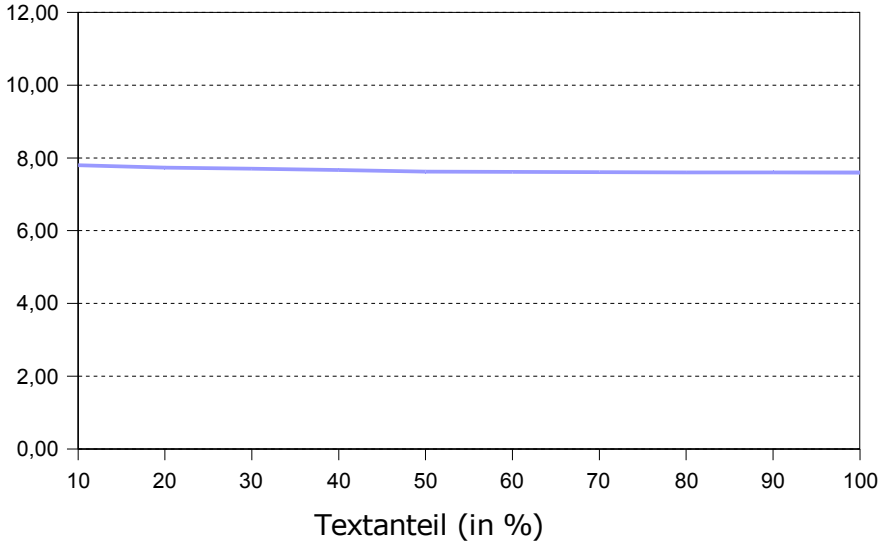
Stabilität der Stilmerkmale

Stilmerkmale müssen auf Textabschnittsebene zuverlässig funktionieren.

Bedingungen:

1. Die Merkmale sollten unabhängig vom verwendetem Textumfang möglichst konstante Werte liefern.

Idealisierte Darstellung



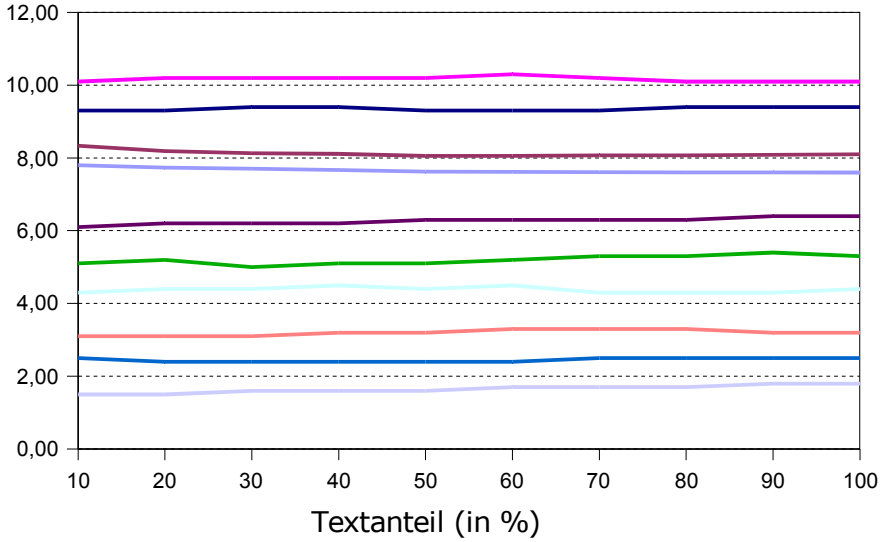
Stabilität der Stilmerkmale

Stilmerkmale müssen auf Textabschnittsebene zuverlässig funktionieren.

Bedingungen:

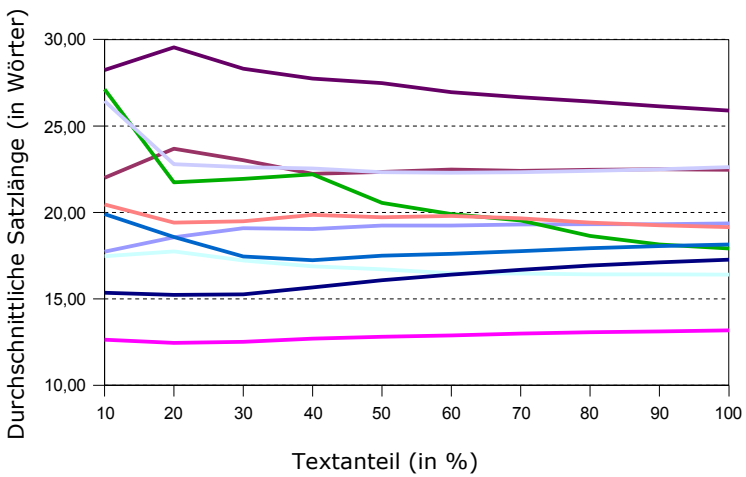
1. Die Merkmale sollten unabhängig vom verwendetem Textumfang möglichst konstante Werte liefern.
2. Die Unterschiede von Werten verschiedener Texte sollten möglichst groß sein.

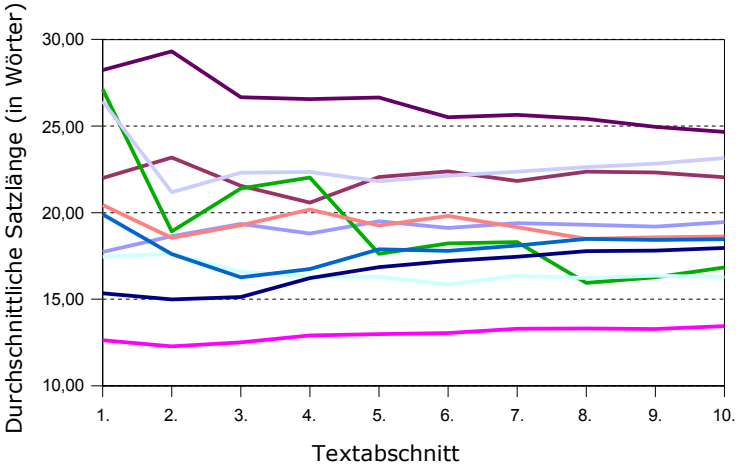
Idealisierte Darstellung

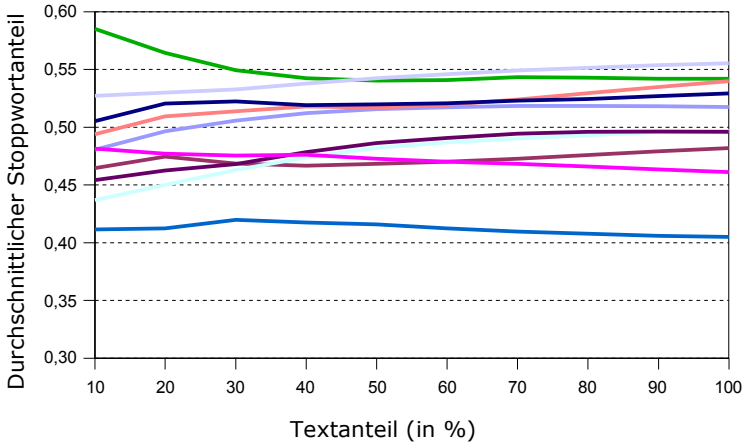


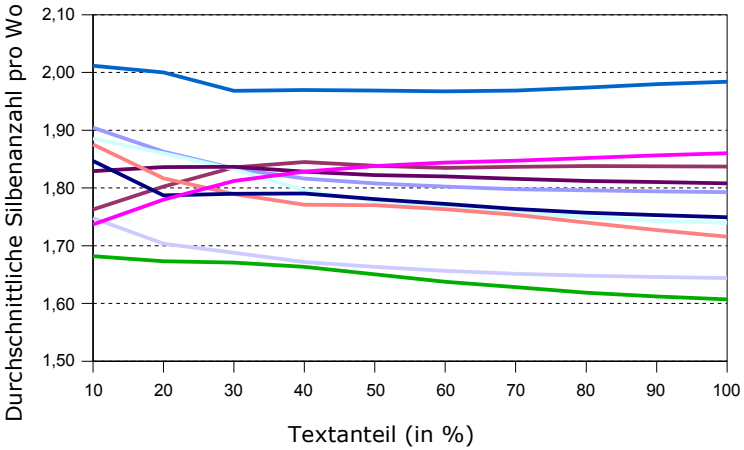
Stabilität der Stilmerkmale

1. Ergebnisse: einfache Stilmerkmale



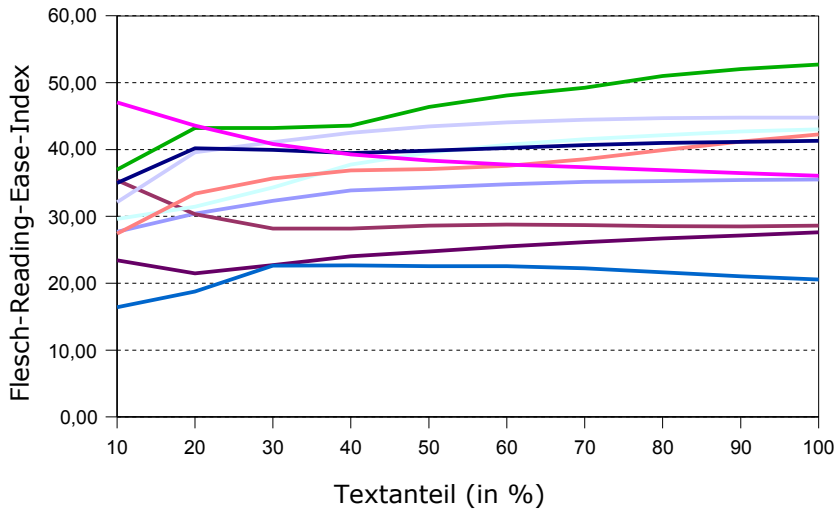


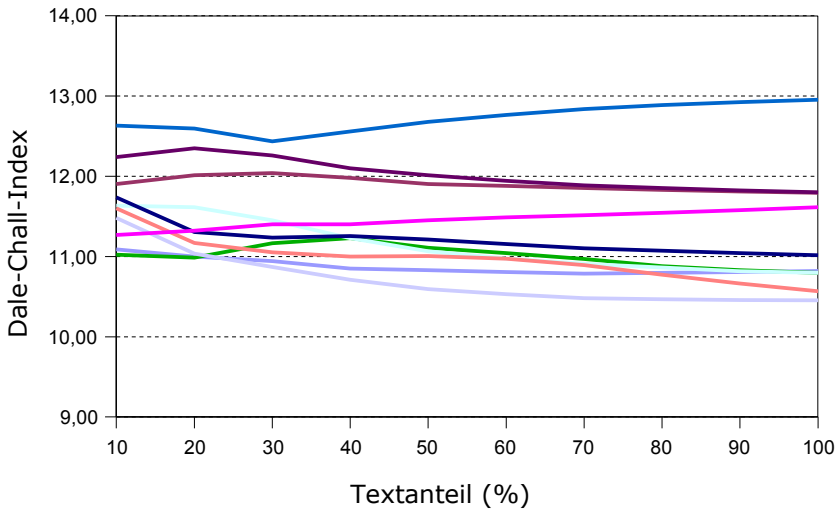




Stabilität der Stilmerkmale

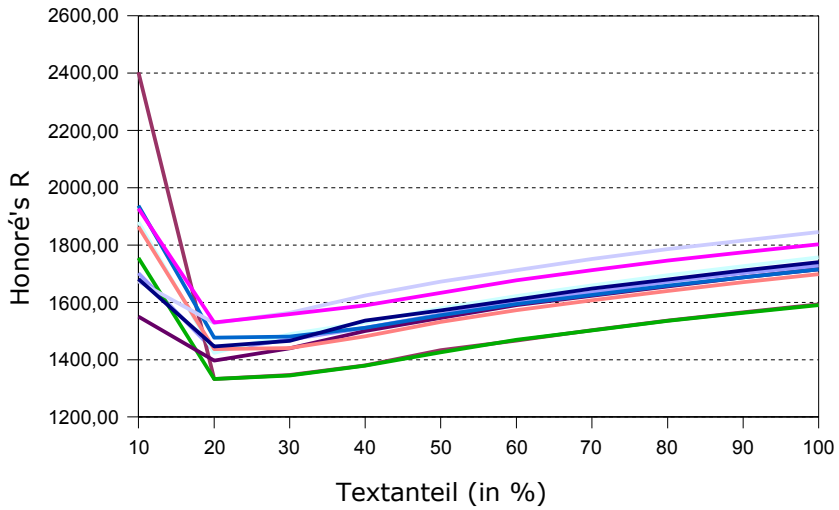
2. Ergebnisse: Lesbarkeitsformeln

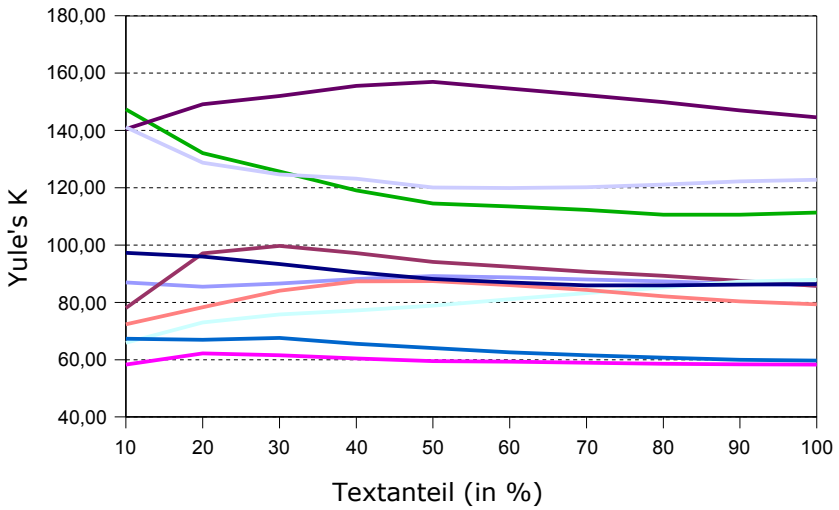




Stabilität der Stilmerkmale

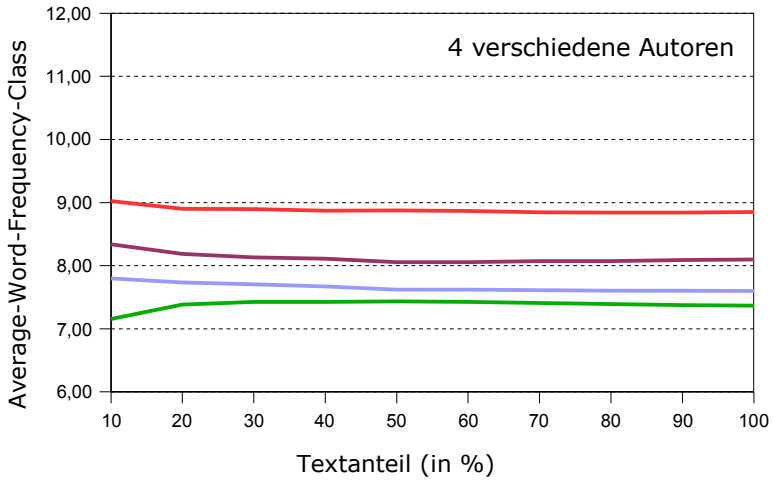
3. Ergebnisse: Maße zur Berechnung der Wortvielfalt

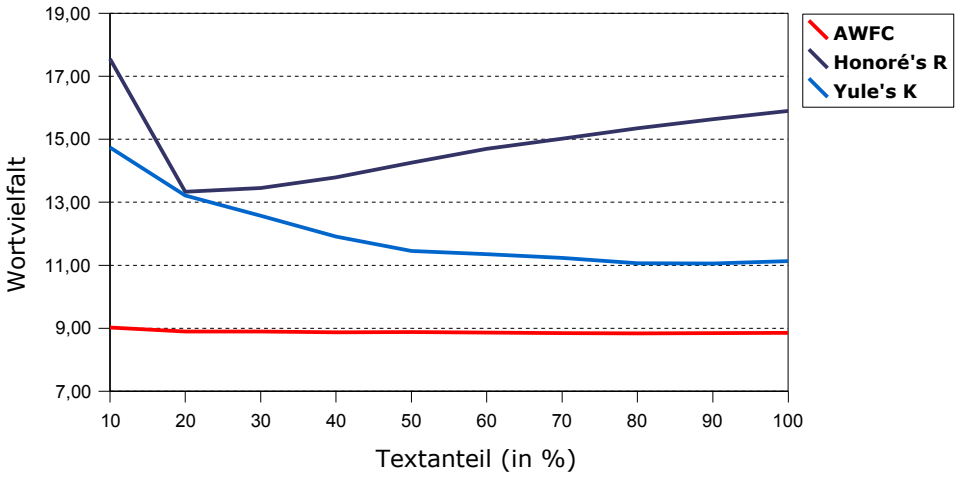




Stabilität der Stilmerkmale

4. Ergebnisse: Average-Word-Frequency-Class





Experiment – Plagiaterkennung

Generierung von Features:

a ... Stilmerkmalswert eines Dokuments

b ... Stilmerkmalswert eines Abschnitts des selben Dokuments

Feature F :

$$f(a,b)=2\cdot\left(\frac{a}{a+b}\right)-1$$

[-1,+1]

Experiment – Plagiaterkennung

Generierung von Features:

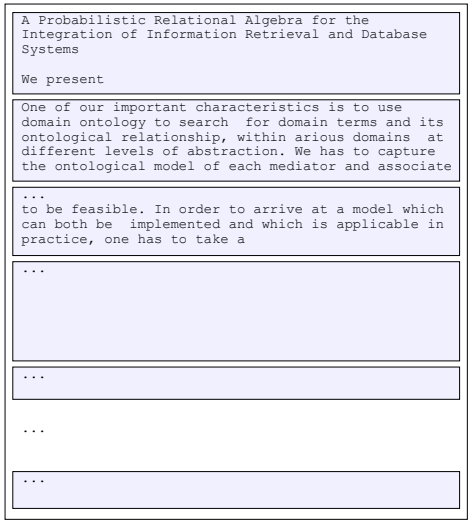
a ... Stilmerkmalswert eines Dokuments

b ... Stilmerkmalswert eines Abschnitts des selben Dokuments

Feature F :

$$f(a,b)=2\cdot\left(\frac{a}{a+b}\right)-1$$

[-1,+1]



Dokument

Experiment – Plagiaterkennung

Generierung von Features:

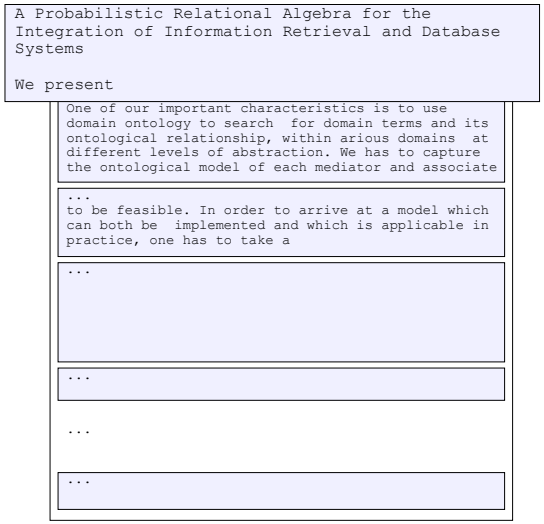
a ... Stilmerkmalswert eines Dokuments

b ... Stilmerkmalswert eines Abschnitts des selben Dokuments

Feature F :

$$f(a,b)=2\cdot\left(\frac{a}{a+b}\right)-1$$

[-1,+1]



Dokument

Experiment – Plagiaterkennung

Generierung von Features:

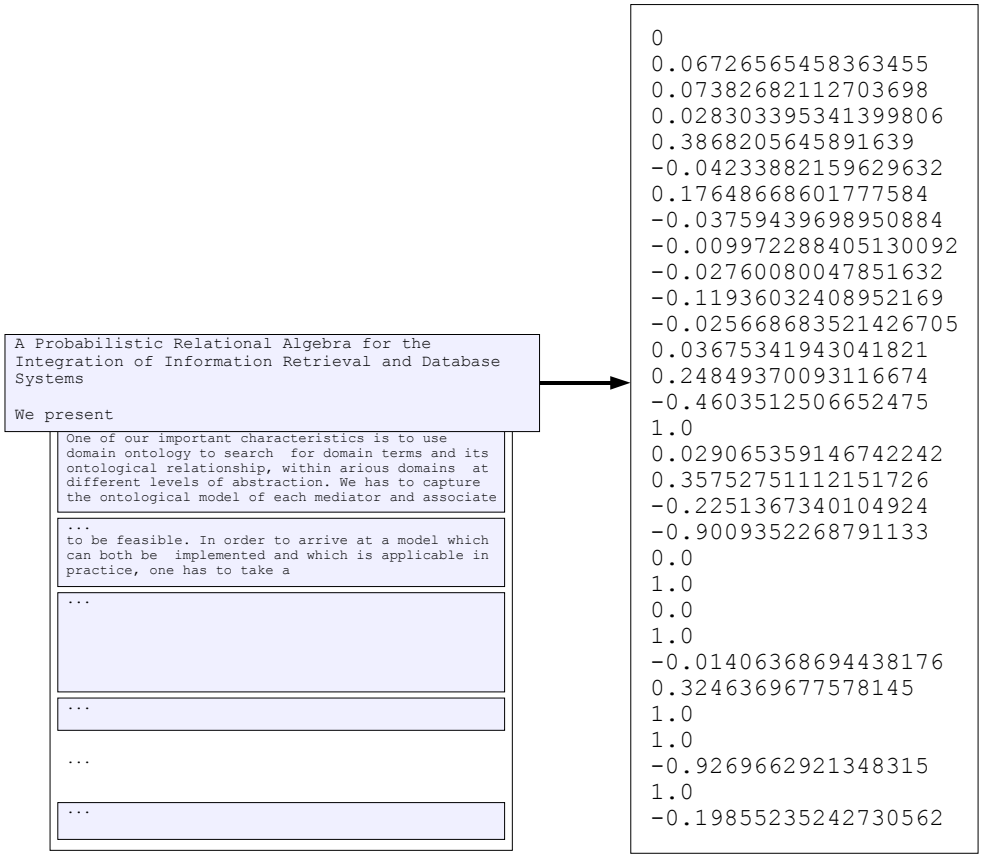
a ... Stilmerkmalswert eines Dokuments

b ... Stilmerkmalswert eines Abschnitts des selben Dokuments

Feature F :

$$f(a,b)=2\cdot\left(\frac{a}{a+b}\right)-1$$

[-1,+1]



Experiment – Plagiaterkennung

- 3200 Korpus-Dokumente
- ca. 70 Textabschnitte pro Dokument
- => 224.000 Feature-Datensätze

Experiment – Plagiaterkennung

- 3200 Korpus-Dokumente
- ca. 70 Textabschnitte pro Dokument
- => 224.000 Feature-Datensätze
- ca. 600 verschiedene plagiierte Abschnitte

Thema	Anzahl Abschnitte	davon plagiiert
CSCW	68704	4704
Information Retrieval	99552	4449
Plagiarism	54640	3760
Gesamt	**222896**	**12913**

Precision und Recall

Vorhergesagte Gruppenzugehörigkeit	Tatsächliche Gruppenzugehörigkeit: Plagiat	Tatsächliche Gruppenzugehörigkeit: kein Plagiat
Plagiat	**a)** Richtig Positiv	**b)** Falsch Positiv
kein Plagiat	**c)** Falsch Negativ	**d)** Richtig Negativ

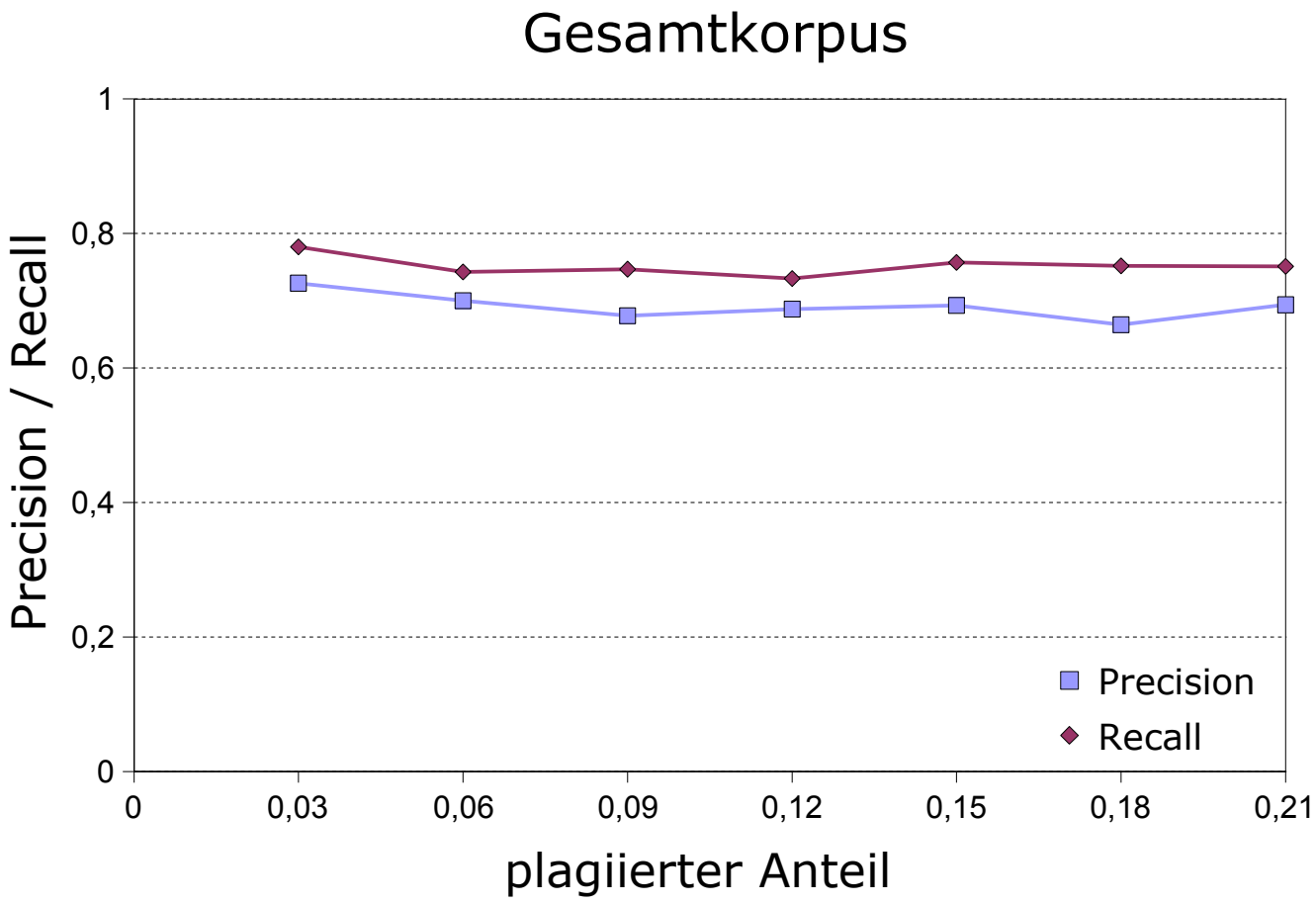
Precision:

$$P = \frac{a}{(a+b)}$$

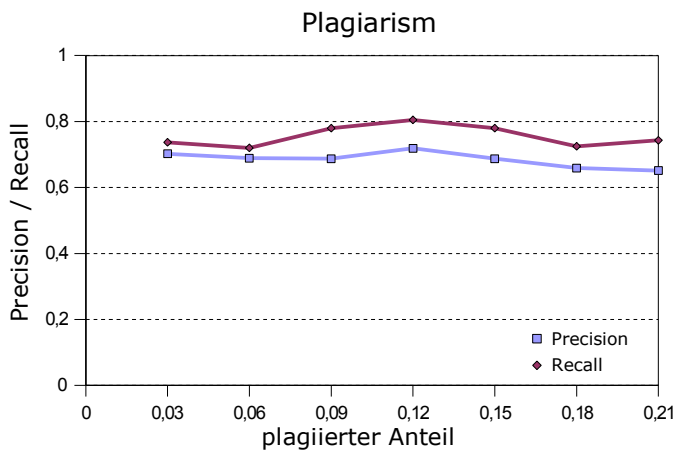
Recall:

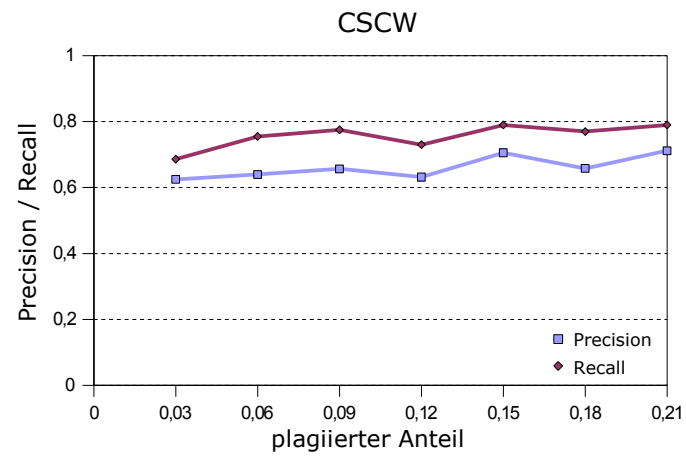
$$R = \frac{a}{(a+c)}$$

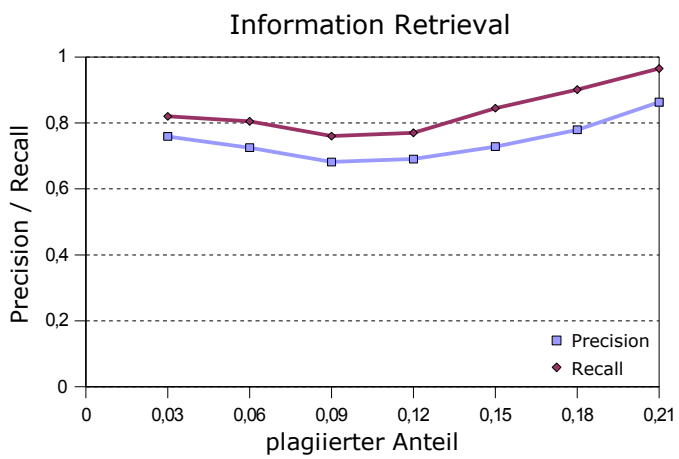
Ergebnisse


Gesamtkorpus
Precision / Recall
1
0,8
0,6
0,4
0,2
0
0
0,03
0,06
0,09
0,12
0,15
0,18
0,21
plagiierter Anteil
Precision
Recall

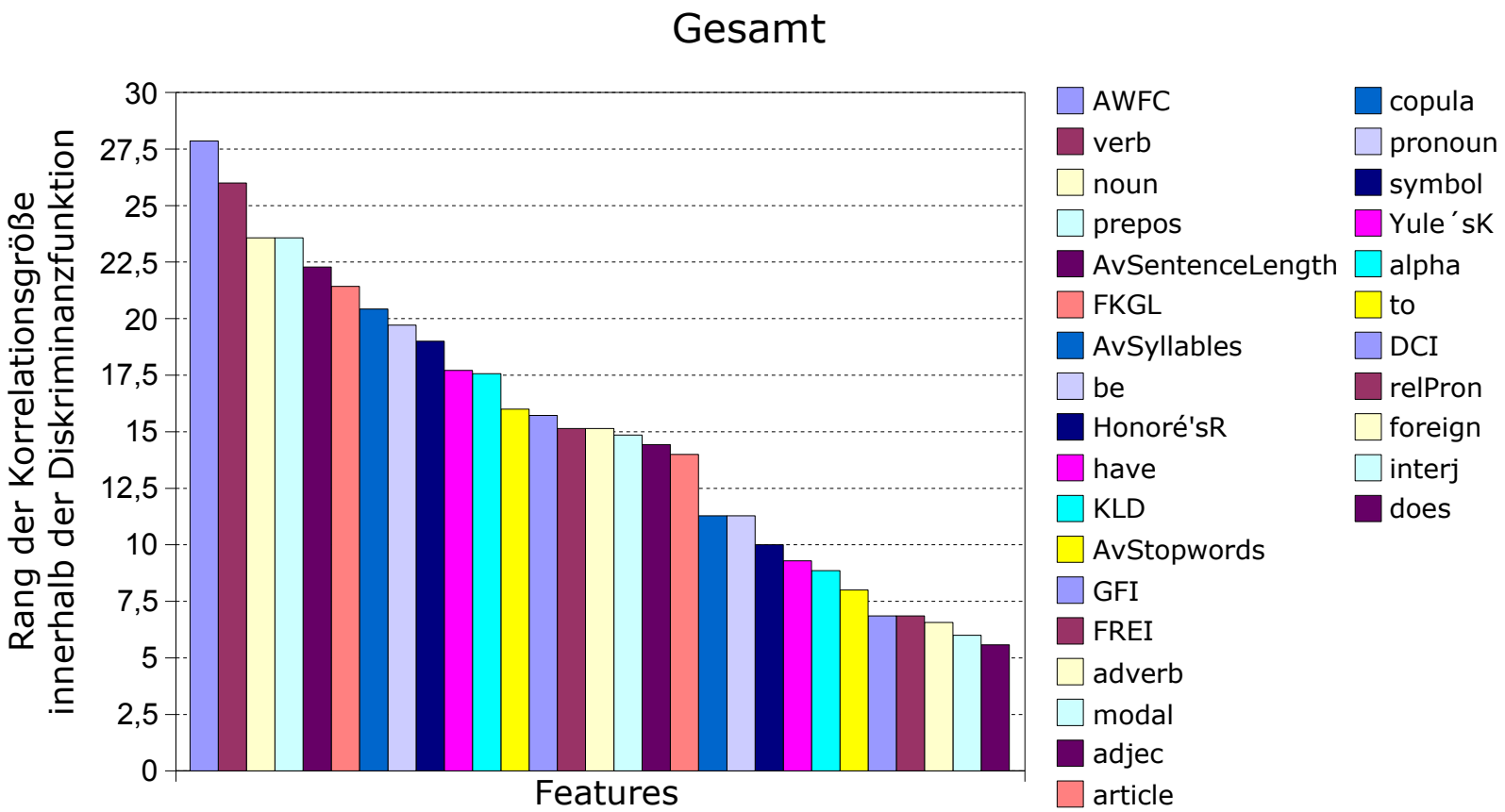
Ergebnisse


Plagiarism
Precision / Recall
1
0,8
0,6
0,4
0,2
0
0
0,03
0,06
0,09
0,12
0,15
0,18
0,21
plagiierter Anteil
Precision
Recall


CSCW
Precision / Recall
1
0,8
0,6
0,4
0,2
0
0
0,03
0,06
0,09
0,12
0,15
0,18
0,21
plagiierter Anteil
Precision
Recall


Information Retrieval
Precision / Recall
1
0,8
0,6
0,4
0,2
0
0
0,03
0,06
0,09
0,12
0,15
0,18
0,21
plagiierter Anteil
Precision
Recall

Ergebnisse


Gesamt
Rang der Korrelationsgröße innerhalb der Diskriminanzfunktion
30
27,5
25
22,5
20
17,5
15
12,5
10
7,5
5
2,5
0
Features
AWFC
verb
noun
prepos
AvSentenceLength
FKGL
AvSyllables
be
Honoré'sR
have
KLD
AvStopwords
GFI
FREI
adverb
modal
adjec
article
copula
pronoun
symbol
Yule´sK
alpha
to
DCI
relPron
foreign
interj
does

Zusammenfassung und Ausblick

- Experimente zur intrinsischen Plagiaterkennung unter Verwendung von Stilmerkmalen lieferten vielversprechende Ergebnisse.
- Precision 70 %, Recall 75%
- Anwendungsmöglichkeit liegt in der Vorauswahl verdächtiger Abschnitte, deren Ursprung dann gefunden werden muss.

Vielen Dank für Ihre Aufmerksamkeit

Literaturverzeichnis

Plagiat Technik: *http://www.plagiarius.com/img/2006_09_s.jpg*

Plagiat Design: *http://www.plagiarius.com/img/2006_03_s.jpg*

Plagiat Kunst: *http://www.hinternet.de/musik/intrview/images/plagiari.jpg*

Plagiat Literatur: *http://www.forumakad.pl/archiwum/99/4/images/image004.jpg*

Meyer zu Eißen, B. Stein: *Intrinsic Plagiarism Detection.* Proceedings of the European Conference on Information Retrieval (ECIR-06), 2006

Precision und Recall

Beispiel: Precision = 80 %,

Recall = 20%

A Probabilistic Relational Algebra for the Integration of Information
Retrieval and Database Systems

We present

One of our important characteristics is to use domain ontology to search for domain terms and its ontological relationship, within arious domains at different levels of abstraction. We has to capture the ontological model of each mediator and associate

...
to be feasible. In order to arrive at a model which can both be
implemented and which is applicable in practice, one has to take a

...
In particular this could be very useful to automated stylometry techniques. The information on the internet could help refine classifiers and provide better classification techniques because of the sheer volume of data available to train on. If some method of automatically annotating the text on the internet were developed it could have a profound effect on search engines and shopping websites.

...
For example, when shopping for books online, a script could recommend authors who score similarly in a stylometry analysis to the one you are shopping for.

Plagiatanalyse:

10 des Plagiats verdächtige Abschnitte gefunden

Precision: 8 von diesen 10 sind plagiiert. 2 sind nicht plagiiert.

Recall: Die 8 gefundenen entsprechen 20 % aller im Text plagiierten Abschnitte (insgesamt 40).