



UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Fakultät für Elektrotechnik, Informatik und Mathematik
Institut für Informatik
Fachgebiet Wissensbasierte Systeme

Masterarbeit

zur Erlangung des Grades
Master of Science

Methoden zur sprachübergreifenden Plagiaterkennung

von

Maik Anderka

Leipziger Str. 1
34454 Bad Arolsen
manderka@upb.de

vorgelegt am 01. Oktober 2007 bei

Prof. Dr. Benno Stein
und Prof. Dr. Wilhelm Schäfer

betreut von Dipl.-Inf. Martin Potthast

Kurzfassung

Gegenstand dieser Arbeit ist die Erforschung und Entwicklung von Methoden zur sprachübergreifenden Erkennung von Plagiaten in Textdokumenten.

Das Problem der sprachübergreifenden Plagiaterkennung wird erstmals als Ganzes betrachtet. Dabei werden zwei Teilaufgaben unterschieden: „Heuristisches Retrieval“ und „detaillierte Analyse“. Für beide Teilaufgaben werden verschiedene Lösungsansätze vorgeschlagen. Der Schwerpunkt liegt auf der detaillierten Analyse, hierzu werden Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse entwickelt.

In dieser Arbeit wird ein neues Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse vorgestellt – die Cross-Language Explicit Semantic Analysis (CL-ESA). Der CL-ESA liegt ein sprachübergreifendes Konzeptraummodell zu Grunde, das es ermöglicht, den Inhalt von verschiedensprachigen Dokumenten, auf der Basis einer enormen Menge von externem Wissen, zu repräsentieren und zu vergleichen. Die CL-ESA ist eine Generalisierung der Explicit Semantic Analysis (ESA) ([Gabrilovich und Markovitch, 2007](#)) und verwendet als Wissensbasis die Online-Enzyklopädie Wikipedia.

Es wurde ein Softwaresystem entwickelt, das die CL-ESA für die Sprachen Deutsch und Englisch implementiert und anhand dessen umfangreiche Experimente zur Evaluierung der CL-ESA durchgeführt wurden. Dabei hat sich gezeigt, dass die CL-ESA bei der sprachübergreifenden Ähnlichkeitsanalyse sehr gute Ergebnisse erzielt und eine geringe Laufzeit besitzt, die mit der des Vektorraummodells vergleichbar ist. In 82,2% der Testfälle konnte die englische Übersetzung eines deutschen Dokuments in einer Menge von Dokumenten identifiziert werden und in 96% der Fälle lag die Übersetzung unter den zehn Dokumenten, die die höchste Ähnlichkeit zu dem deutschen Dokument aufwiesen.

Inhaltsverzeichnis

| | |
|--|-----------|
| Kurzfassung | iii |
| Notationen | xi |
| 1 Einleitung | 1 |
| 2 Information Retrieval (IR) | 5 |
| 2.1 Vektorraummodell und Dokumentindexierung | 7 |
| 2.2 Cross-Language Information Retrieval (CLIR) | 8 |
| 2.3 Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse | 9 |
| 3 Plagiaterkennung | 13 |
| 3.1 Grundlegende Vorgehensweise | 13 |
| 3.2 Verfahren zur monolingualen Plagiaterkennung | 16 |
| 3.2.1 Dokumentmodellvergleich | 16 |
| 3.2.2 Fingerprinting | 17 |
| 3.2.3 Fuzzy-Fingerprinting | 17 |
| 3.2.4 Stilanalyse | 18 |
| 4 Sprachübergreifende Plagiaterkennung | 21 |
| 4.1 Problemstellung | 21 |
| 4.2 Heuristisches Retrieval | 22 |
| 4.2.1 CLIR-basiertes heuristisches Retrieval | 23 |
| 4.2.2 MT-basiertes heuristisches Retrieval | 24 |
| 4.3 Detaillierte Analyse | 25 |
| 5 Konzeptraummodell | 27 |
| 5.1 Monolinguales Konzeptraummodell (CSM) | 28 |
| 5.1.1 Konzeptindexierung | 29 |
| 5.1.2 Explicit Semantic Analysis (ESA) | 29 |
| 5.2 Sprachübergreifendes Konzeptraummodell (CL-CSM) | 31 |

| | | |
|----------|---|-----------|
| 5.2.1 | Wörterbuchbasiert | 33 |
| 5.2.2 | Thesaurusbasiert | 34 |
| 5.2.3 | Parallelkorpusbasiert | 36 |
| 5.3 | Cross-Language Explicit Semantic Analysis (CL-ESA) | 37 |
| 6 | Details zur Implementierung | 39 |
| 6.1 | Konstruktion des Wikipedia-basierten CL-CSM | 39 |
| 6.1.1 | Extraktion der relevanten Artikel (Filtern) | 40 |
| 6.1.2 | Konstruktion eines bilingualen Wikipedia-Thesaurus | 42 |
| 6.1.3 | Indexierung | 43 |
| 6.2 | Repräsentation von Dokumenten | 44 |
| 6.2.1 | Dokumentindexierung | 45 |
| 6.2.2 | Berechnung der Konzeptvektoren | 46 |
| 7 | Experimentelle Auswertung der CL-ESA | 47 |
| 7.1 | Allgemeine Parameter | 47 |
| 7.2 | Monolinguale Evaluierung der ESA* | 49 |
| 7.3 | Evaluierung der sprachübergreifenden Ähnlichkeitsanalyse mittels CL-ESA | 50 |
| 7.3.1 | Evaluierung anhand von Wikipedia | 51 |
| 7.3.2 | Evaluierung anhand des Europarl-Korpus | 53 |
| 7.4 | Laufzeit der CL-ESA | 55 |
| 7.5 | Multilingualität der CL-ESA | 55 |
| 7.6 | Diskussion | 57 |
| 7.6.1 | Schlussfolgerungen | 57 |
| 7.6.2 | Vergleich der CL-ESA mit anderen Verfahren | 61 |
| 8 | Zusammenfassung und Ausblick | 63 |
| | Literaturverzeichnis | 67 |
| | Stichwortverzeichnis | 73 |

Abbildungsverzeichnis

| | | |
|-----|---|----|
| 2.1 | Prozess des Information Retrieval | 5 |
| 2.2 | Beispiel für das Vektorraummodell | 7 |
| 2.3 | Beispiel für das CLIR | 9 |
| 2.4 | Taxonomie der Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse . | 10 |
| 3.1 | Plagiaterkennung als ein Prozess aus drei Schritten | 14 |
| 3.2 | Taxonomie der Plagiatvergehen und mögliche Erkennungsmethoden . . . | 15 |
| 4.1 | UML-Aktivitätsdiagramm: Lösungsansätze für das heuristische Retrieval | 23 |
| 4.2 | UML-Aktivitätsdiagramm: Lösungsansatz für die detaillierte Analyse . . | 25 |
| 5.1 | Beispiel für das monolinguale Konzeptraummodell | 28 |
| 5.2 | Beispiel für ein (sprachübergreifendes) Konzept | 31 |
| 5.3 | Beispiel für das sprachübergreifende Konzeptraummodell | 32 |
| 5.4 | Beispiel für das wörterbuchbasierte sprachübergreifende Konzeptraummo- dell | 33 |
| 5.5 | Beispiel für das thesaurusbasierte sprachübergreifende Konzeptraummodell | 35 |
| 5.6 | Beispiel für das enzyklopädiebasierte sprachübergreifende Konzeptraum- modell | 37 |
| 6.1 | Konstruktion eines Wikipedia-basierten sprachübergreifenden Konzep- traummodells | 40 |
| 6.2 | UML-Aktivitätsdiagramm: Parsen eines Wikipedia-Dumps | 41 |
| 6.3 | Beispiel für Wikipedia-Artikel, die Paare bilden. | 42 |
| 6.4 | Auszug aus einem bilingualen Wikipedia-Thesaurus | 43 |
| 6.5 | UML-Aktivitätsdiagramm: Indexierung der Wikipedia-Artikel | 43 |
| 6.6 | Beispiel für die Erstellung eines invertierten Indexes | 44 |
| 6.7 | Beispiel für die Indexierung eines Dokuments, CL-ESA | 45 |
| 7.1 | Ergebnisse der monolingualen Ähnlichkeitsanalyse mittels ESA* | 49 |

| | | |
|-----|--|----|
| 7.2 | Laufzeit der CL-ESA | 55 |
| 7.3 | Anzahl der Wikipedia-Artikel in verschiedenen Wikipedia-Sprachversionen | 56 |
| 7.4 | Anzahl der Wikipedia-Artikel, die in verschiedenen Wikipedia-Sprachversionen Übersetzungspaare bilden | 56 |
| 7.5 | Ähnlichkeitsverteilung für die Dokumente des Europarl-Korpus | 59 |
| 7.6 | Ähnlichkeitsverteilung für die Artikel in Wikipedia | 59 |

Tabellenverzeichnis

| | | |
|-----|--|----|
| 5.1 | Relevanzkriterien für Wikipedia-Artikel | 30 |
| 6.1 | Informationen zu den verwendeten Wikipedia-Dumps | 40 |
| 7.1 | Relevanzkriterien, die in den Experimenten eingesetzt wurden | 48 |
| 7.2 | Anzahl der relevanten Artikel, entsprechend der Relevanzkriterien | 48 |
| 7.3 | Vergleich zwischen ESA und ESA* | 50 |
| 7.4 | Ergebnisse der Experimente zur Evaluierung der CL-ESA anhand von Wikipedia | 52 |
| 7.5 | Ergebnisse der Experimente zur Evaluierung der CL-ESA anhand des Europarl-Korpus | 54 |
| 7.6 | Bewertung der Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse | 61 |

Notationen

| Symbol | Bedeutung |
|--------------------------------|---|
| $\langle \cdot, \cdot \rangle$ | Kreuzprodukt |
| $\ \cdot\ _2$ | L_2 -Norm |
| \mathcal{D} | Dokumentsammlung bzw. Referenzkorpus |
| \mathcal{H} | Hashtabelle |
| \mathcal{K} | Wissensmenge |
| \mathcal{K}_j | Menge von Wissen in der Sprache L_j |
| \mathcal{L} | Menge von Sprachen |
| φ | Ähnlichkeitsfunktion |
| φ_{cos} | Kosinusähnlichkeit |
| c, c_i | Textabschnitt |
| \mathcal{C} | Menge von Textabschnitten |
| d, d_i | Dokument bzw. Artikel |
| d_q | Verdächtiges Dokument bzw. Beispieldokument |
| \mathbf{d}, \mathbf{d}_i | Dokumentmodell bzw. Vektorrepräsentation eines Dokuments d bzw. d_i |
| $[\mathbf{d}]_i$ | i -te Komponente des Vektors \mathbf{d} |
| D, D_i | Dokumentmenge |
| f, f_i | Merkmal bzw. Schlüsselwort |
| $h(c), h_\varphi(c)$ | Hashwert eines Textabschnitts c bzw. (Fuzzy-) Fingerprint |
| k, k_i | Konzept |
| \mathbf{k}, \mathbf{k}_i | Support-Vektor für das Konzept k bzw. k_i |
| $\mathbf{k}_i^{L_j}$ | Support-Vektor für das Konzept k_i in der Sprache L_j |
| K | Menge von Konzepten |
| L, L_j | (Natürliche) Sprache |
| q | Anfrage bzw. Wortanfrage |
| $tf(f, d)$ | Häufigkeit des Schlüsselwortes f in dem Dokument d |
| $tf \cdot idf$ | Termgewichtsmaß: Termhäufigkeit multipliziert mit der inversen Dokumenthäufigkeit |
| U | Universum von Hashwerten |
| V | Menge von Schlüsselwörtern bzw. Vokabular |
| V^{de}, V^{en} | Vokabular eines deutschen bzw. englischen invertierten Indexes |
| w, w_j | Wort |
| $w(f)$ | Gewicht des Schlüsselwortes f |

1 Einleitung

Das rasante Wachstum des World Wide Web führt dazu, dass dem Menschen eine unüberschaubare Menge von Informationen zu allen Themen zur Verfügung steht, die von jedem Ort und zu jeder Zeit abrufbar ist. Mittels Suchmaschinen wird der Unüberschaubarkeit entgegengewirkt, denn diese ermöglichen eine einfache und gezielte Suche nach relevanten Informationen. Die wachsende Leichtigkeit der Informationsbeschaffung ist für manche ein großer Anreiz zu plagieren. Doch auch die Leichtigkeit Plagiate mittels maschineller Suche zu entlarven, hat dazu geführt, dass das Problem der Plagiaterkennung in den letzten Jahren mehr und mehr diskutiert wurde.

Die Benutzung oder Übernahme von fremdem geistigen Eigentum, ohne korrekte Angabe der Originalquelle, wird als „Plagieren“ bezeichnet. Das Produkt, das dadurch entsteht, wird „Plagiat“ genannt.

Plagiate lassen sich in vielen verschiedenen Bereichen finden, beispielsweise in der Literatur, in Musikstücken, in wissenschaftlichen Arbeiten, in Software oder auch in abstrakten Dingen, wie etwa neuen Ideen oder wissenschaftlichen Erkenntnissen. Diese Arbeit befasst sich ausschließlich mit Plagiaten in Textdokumenten.

Besonders im akademischen Bereich und vor allem an Universitäten sind Plagiate ein ernst zu nehmendes Problem (Culwin und Lancaster, 2000). Für Studierende, die ausreichend kriminelle Energie mitbringen, ist es ein leichtes, in ihre Hausarbeiten oder wissenschaftlichen Ausarbeitungen fremde Inhalte einzufügen, denn Suchmaschinen und Online-Enzyklopädien liefern zu fast jedem Themengebiet eine Vielzahl von Informationen, die einfach – z. B. mittels „Kopieren & Einfügen“ – übernommen werden können. Dass Plagiate im akademischen und schulischen Bereich keine Seltenheit sind, zeigt nicht zuletzt die Tatsache, dass mittlerweile umfangreiche Online-Angebote existieren, die das Plagieren unterstützen, wie z. B. Cheat House¹. Um dem zunehmenden Problem des Plagierens entgegenzuwirken, ist es an vielen Universitäten in den USA und in Kanada

¹Cheat House: <http://www.cheathouse.com>.

inzwischen üblich, studentische Arbeiten auf Plagiate hin zu überprüfen und Plagiatvergehen entsprechend zu ahnden. Da es aufgrund der Vielzahl von Studierenden nicht möglich ist, jede Arbeit manuell nach Plagiaten zu durchsuchen, werden meist kommerzielle Softwaresysteme zur maschinellen Plagiaterkennung eingesetzt, wie z. B. Turnitin² oder Plagiarism Finder³.

Auch im kommerziellen Bereich spielt die Plagiaterkennung eine wichtige Rolle, etwa um Copyright-Verletzungen aufzudecken. Plagiate tauchen beispielsweise in der Literatur auf, in Nachrichten- bzw. Zeitungsartikeln, auf Webseiten oder in der Werbung. Ein Softwaresystem zur Plagiaterkennung, das speziell für den Einsatz im kommerziellen Bereich entwickelt wurde, ist z. B. iThenticate⁴.

Über die bekannten Softwaresysteme zur maschinellen Erkennung von Plagiaten, wie Turnitin, Plagiarism Finder oder iThenticate, sind keine technischen Informationen bekannt, da es sich um kommerzielle Anwendungen handelt. Mit Hinblick auf den aktuellen Stand der Forschung im Bereich der Plagiaterkennung ist jedoch anzunehmen, dass durch diese Systeme nur die Plagiate erkannt werden, die durch eine Eins-zu-eins-Kopie entstehen. Bei einer Eins-zu-eins-Kopie wird der Originaltext Wort für Wort übernommen. Die vermehrt auftretenden Plagiate sind allerdings diejenigen, die durch eine Modifikation des Originaltexts entstehen, z. B. durch eine Veränderung des Satzbaus oder durch den Austausch einzelner Wörter. Diese Art des Plagierens wird häufig angewandt, da die Gefahr, entlarvt zu werden, geringer ist, wenn der Originaltext leicht verändert wird. Ein Softwaresystem, das die Erkennung von Plagiaten ermöglicht, die durch Modifikation entstehen, ist Picapica.net⁵.

Eine weitere, weit verbreitete, Form von Plagiaten, die bisher von keiner Plagiaterkennungssoftware entlarvt werden, sind Plagiate, die durch eine Übersetzung des Originaltexts entstehen. Die meisten Studierenden hierzulande schreiben beispielsweise ihre Ausarbeitungen in deutscher Sprache, die entsprechende Literatur ist jedoch in vielen Disziplinen nur in Englisch verfügbar. Es ist sehr einfach, Texte oder Textpassagen manuell oder mit der Unterstützung von entsprechender Software zu übersetzen und in die eigene Arbeit einzufügen. Die Erkennung solcher Plagiate wird als „sprachübergreifende Plagiaterkennung“ bezeichnet. Aktuell sind keine Forschungsarbeiten bekannt, die eine Lösung für die sprachübergreifende Plagiaterkennung vorschlagen. Lediglich eine Arbeit

²Turnitin: <http://www.turnitin.com>.

³Plagiarism Finder: <http://m4-software.com>.

⁴iThenticate: <http://www.ithenticate.com>.

⁵Picapica.net (Plagiarism Indication by Computer-based Analysis): <http://www.picapica.net>.

(Pouliquen et al., 2003) bietet eine Lösung für ein Teilproblem; die Problematik der sprachübergreifenden Plagiaterkennung wird jedoch nicht vertieft.

Gegenstand dieser Arbeit ist die Erforschung und Entwicklung von Methoden zur maschinellen, sprachübergreifenden Erkennung von Plagiaten. Die sprachübergreifende Plagiaterkennung wird erstmals als Ganzes betrachtet und es werden verschiedene Verfahren zur Lösung vorgestellt. In diesem Zusammenhang wird ein neues Dokumentmodell – das Konzeptraummodell – vorgestellt, das es ermöglicht, den Inhalt von Dokumenten – basierend auf einer enormen Menge von Wissen – sprachübergreifend zu repräsentieren. Weiterhin wird ein neues Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse präsentiert, in dem ein Konzeptraummodell eingesetzt wird – die Cross-Language Explicit Semantic Analysis (CL-ESA).

Im Rahmen dieser Arbeit wurde ein Softwaresystem entwickelt, das die CL-ESA implementiert und zur sprachübergreifenden Plagiaterkennung eingesetzt werden kann. Das System wurde anhand von verschiedenen Experimenten evaluiert. Die Ergebnisse werden mit existierenden Verfahren aus dem Bereich der sprachübergreifenden Ähnlichkeitsanalyse verglichen.

Kapitel 2 gibt eine Einführung in die Grundlagen und die Terminologie des Information Retrieval sowie des Cross-Language Information Retrieval. Außerdem wird eine Übersicht über existierende Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse gegeben.

In Kapitel 3 wird die Plagiaterkennung erläutert und zwischen monolingualer und sprachübergreifender Plagiaterkennung unterschieden. Weiterhin erfolgt die Beschreibung der Problemstellung sowie grundlegender Techniken der monolingualen Plagiaterkennung.

Auf die sprachübergreifende Plagiaterkennung wird in Kapitel 4 eingegangen. Es werden die Problemstellung definiert und neue Verfahren zur Lösung vorgeschlagen.

In Kapitel 5 erfolgt die Erläuterung des Konzeptraummodells und der CL-ESA. Dabei werden verschiedene Arten von Konzeptraummodellen, denen unterschiedliche Wissensbasen zu Grunde liegen, beschrieben.

Kapitel 6 enthält Details zur Implementierung der CL-ESA.

Die Experimente zur Evaluierung der CL-ESA werden in Kapitel 7 beschrieben, deren Ergebnisse diskutiert und mit anderen Verfahren verglichen.

Eine Zusammenfassung der Arbeit und ein Ausblick erfolgt in Kapitel 8.

2 Information Retrieval (IR)

Neben allen Vorteilen, die das enorme Wachstum und die große Informationsvielfalt des World Wide Web mit sich bringen, wird das Auffinden von Informationen, die für einen Benutzer relevant sind, immer schwieriger. Das Fachgebiet Information Retrieval (IR)¹ befasst sich mit der Suche von Informationen in großen, unstrukturierten Datenmengen. Ziel ist es, den Informationsbedarf einer Person, durch ein Dokument zu befriedigen. Der Informationsbedarf wird häufig als, für ein Informationssystem geeignete, Anfrage spezifiziert. Ein Beispiel für ein IR-System sind Suchmaschinen, wie z. B. Google².

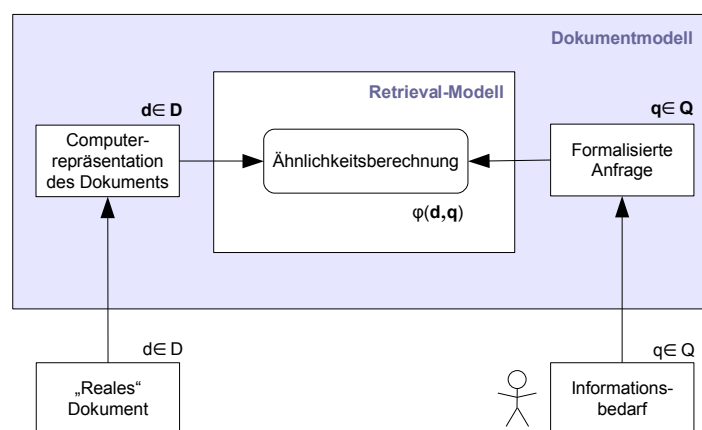


Abb. 2.1: Prozess des Information Retrieval (Stein et al., 2006). Das Ziel ist, einen Informationsbedarf q durch ein „reales“ Dokument d zu befriedigen. Um das Ziel mit der Unterstützung eines Informationssystems zu erreichen, werden q und d durch q und d abstrahiert. Anhand eines Ähnlichkeitsmaßes φ wird die Ähnlichkeit zwischen q und d bestimmt.

Wie das oben genannte Ziel mit der Unterstützung eines Informationssystems erreicht wird, ist in **Abb. 2.1** dargestellt. Im Information Retrieval wird zwischen einem „realen“ Dokument d und seiner Computerrepräsentation d unterschieden. Ein Informationsbedarf q und ein Dokument d werden durch q und d abstrahiert. Anhand eines Ähnlichkeitsmaßes φ wird die Ähnlichkeit zwischen q und d bestimmt. Die Ähnlichkeitsfunktion $\varphi(q, d)$ bildet q und d auf das Intervall $[0; 1]$ ab, wobei 0 keiner

¹„Information Retrieval“ bedeutet Informationsbeschaffung oder auch Informationswiedergewinnung.

²Google: <http://www.google.de>.

Ähnlichkeit und 1 der maximalen Ähnlichkeit zwischen \mathbf{q} und \mathbf{d} entspricht. Eine Computerrepräsentation zusammen mit einem Ähnlichkeitsmaß wird als Dokumentmodell bezeichnet.

Der Informationsbedarf bzw. die Anfrage kann auf unterschiedliche Art und Weise spezifiziert werden. Es wird unterschieden zwischen einer Wortanfrage und einem Dokument. Eine Wortanfrage besteht aus einer Menge von Schlüsselwörtern. Diese Form der Anfrage kommt z. B. bei Suchmaschinen zum Einsatz. In der Plagiatanalyse dagegen wird der Informationsbedarf durch ein (verdächtiges) Dokument spezifiziert. Dieses Prinzip wird als „Query by Example“ bezeichnet.³

Im Information Retrieval kann außerdem zwischen einer „geschlossenen“ und einer „offenen“ Retrieval-Situation unterschieden werden (Stein, 2007). Eine geschlossene Retrieval-Situation liegt vor, wenn die komplette Dokumentsammlung D , in der sich ein Dokument d befindet, im Voraus bekannt ist. In einer offenen Retrieval-Situation ist D im Voraus unbekannt.

Die Plagiaterkennung ist ein Anwendungsfall des Information Retrieval. Die Verfahren zur Plagiaterkennung basieren auf Techniken aus dem Bereich des IR. In diesem Kapitel wird daher eine Einführung in die Grundlagen sowie die Terminologie des IR geben. Dabei werden nur die Themen angesprochen, die im weiteren Verlauf der Arbeit von Bedeutung sind.

In Abschnitt 2.1 wird ein weit verbreitetes Dokumentmodelle – das Vektorraummodell – beschrieben und das Prinzip der Dokumentindexierung erläutert.

Mit einem Teilgebiet des IR, dem Cross-Language Information Retrieval (CLIR), befasst sich Abschnitt 2.2. Hierbei können sowohl die Anfrage q als auch das Dokument d in beliebigen Sprachen vorliegen.

Abschnitt 2.3 geht auf die sprachübergreifende Ähnlichkeitsanalyse von Dokumenten ein. Die Problemstellung der sprachübergreifenden Ähnlichkeitsanalyse entspricht der Problemstellung der sprachübergreifenden Plagiaterkennung.

³Das Dokument, das die Anfrage darstellt, wird als „Example“ bzw. „Beispieldokument“ bezeichnet.

2.1 Vektorraummodell und Dokumentindexierung

Ein im Information Retrieval weit verbreitetes Dokumentmodell ist das Vektorraummodell (Vector Space Model, VSM) nach [Salton et al. \(1975\)](#).

Ein Dokument d wird durch einen m -dimensionalen Vektor \mathbf{d} repräsentiert. Jede Dimension des dadurch aufgespannten Vektorraums entspricht einem Merkmal f , aus einer Menge von Merkmalen $V = \{f_1, \dots, f_m\}$. Der i -te Eintrag des Vektors \mathbf{d} quantifiziert die Wichtigkeit des Merkmals f_i in Bezug auf den Inhalt des Dokuments d . Häufig werden Worte als Merkmale verwendet, in diesem Fall spricht man von einer „bag-of-words“-Repräsentation. Der Vektor \mathbf{d} wird dann als „Wortvektor“ bezeichnet. Ein Beispiel für das Vektorraummodell ist in **Abb. 2.2** dargestellt.

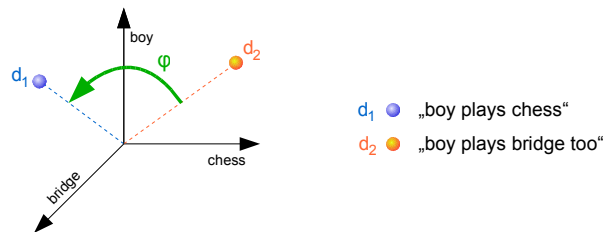


Abb. 2.2: Beispiel für das Vektorraummodell ([Meyer zu Eißgen et al., 2005](#)). Die Dokumente d_1 und d_2 werden in einem Vektorraum, der durch die Worte „boy“, „chess“ und „bridge“ aufgespannt wird, repräsentiert. Die Worte bilden das Vokabular $V = \{\text{boy}, \text{chess}, \text{bridge}\}$. Als Ähnlichkeitsmaß φ ist die Kosinusähnlichkeit eingezeichnet.

Um alle Dokumente einer Dokumentmenge D in demselben Vektorraum zu modellieren, werden alle Wörter der Dokumente zu einem gemeinsamen Vokabular V zusammengefasst. Häufig wird dabei auf so genannte Stoppwörter verzichtet. Dies sind Wörter, die nicht aussagekräftig für den Inhalt eines Dokuments sind, wie beispielsweise Artikel, Konjunktionen oder Präpositionen. Des Weiteren werden die Wörter auf ihre Stammformen gebracht, z. B. wird sowohl aus „bedient“, als auch aus „bedienten“ das Wort „bedienen“. Dieses Vorgehen wird als „Stemming“ bezeichnet. Ein bekanntes Stemming-Verfahren ist Porters Stemming-Algorithmus, [Porter \(1980\)](#).

Zur Bestimmung der Wichtigkeit eines Worts, in Bezug auf den Inhalt eines Dokuments, werden so genannte Termgewichtsmaße eingesetzt. Ein weit verbreitetes Termgewichtsmaß ist $tf \cdot idf$ ([Salton und McGill, 1983](#)), das sich aus dem Produkt der Termhäufigkeit (tf , Term Frequency) und der inversen Dokumenthäufigkeit (idf , Inverse Document Frequency) errechnet. Das Gewicht $w(f_i)$ für ein Wort f_i bzgl. des Dokuments $d \in D$

berechnet sich wie folgt:

$$w(f_i) = tf(f_i, d) \cdot \log \frac{|D|}{df(f_i)}$$

Wobei $tf(f_i, d)$ die Häufigkeit von f_i in d ist und $df(f_i)$ die Anzahl der Dokumente aus D , in denen f_i vorkommt. Das Termgewichtsmaß gibt solchen Wörtern eine hohe Bewertung, die in wenigen Dokumenten häufig vorkommen und daher gut geeignet sind, um diese Dokumente von anderen abzugrenzen.

Das Vektorraummodell ermöglicht es, die inhaltliche Ähnlichkeit zweier Dokumente d_1 und d_2 , anhand ihrer Vektorraumrepräsentationen \mathbf{d}_1 und \mathbf{d}_2 zu bestimmen. Es existieren verschiedene Ähnlichkeitsmaße für Wortvektoren. Das populärste Ähnlichkeitsmaß ist die Kosinusähnlichkeit. Der Grad der Ähnlichkeit zwischen zwei Wortvektoren d_1 und d_2 wird über den Kosinus des Winkels zwischen den Vektoren \mathbf{d}_1 und \mathbf{d}_2 bestimmt. Je spitzer der Winkel zwischen zwei Wortvektoren ist, desto größer ist die Ähnlichkeit der beiden Dokumente (siehe Abb. 2.2). Die Kosinusähnlichkeit ist wie folgt definiert:

$$\varphi_{\cos}(\mathbf{d}_1, \mathbf{d}_2) = \frac{\langle \mathbf{d}_1, \mathbf{d}_2 \rangle}{\|\mathbf{d}_1\|_2 \cdot \|\mathbf{d}_2\|_2}$$

Wobei $\langle \cdot, \cdot \rangle$ das Kreuzprodukt zweier Vektoren ist und $\|\cdot\|_2$ die L_2 -Norm eines Vektors, also seine Länge.

2.2 Cross-Language Information Retrieval (CLIR)

Die Aufgabenstellung beim Cross-Language Information Retrieval (CLIR)⁴ ist dieselbe wie beim Information Retrieval (siehe Abb. 2.1), mit dem Unterschied, dass sowohl das Dokument als auch der Informationsbedarf in beliebigen Sprachen vorliegen können.

Die Forschung im Bereich des CLIR konzentriert sich hauptsächlich auf die Retrieval-Situation, in der der Informationsbedarf in Form einer Wortanfrage, die durch eine Person spezifiziert wird, vorliegt. Für die Wortanfrage in einer Sprache L wird in einer Dokumentmenge D' , in einer Sprache L' , nach Dokumenten gesucht, deren Inhalt zu dem Thema der Wortanfrage passt. Im Fall der sprachübergreifenden Plagiaterkennung wird der Informationsbedarf jedoch durch ein Dokument d_q , anstatt durch eine Wortanfrage,

⁴In einigen früheren Arbeiten wird CLIR auch als „Multilingual Information Retrieval“, „Translingual Information Retrieval“ oder „sprachübergreifendes Information Retrieval“ bezeichnet.

definiert. Es wird in D' nach Dokumenten gesucht, die inhaltlich zu d_q passen. Diese Retrieval-Situation wird als „sprachübergreifende Ähnlichkeitsanalyse“ bezeichnet. Zur sprachübergreifenden Ähnlichkeitsanalyse wurde bisher nur sehr wenig Forschung betrieben. In **Abb. 2.3** sind beide Retrieval-Situationen anhand eines Beispiels dargestellt.

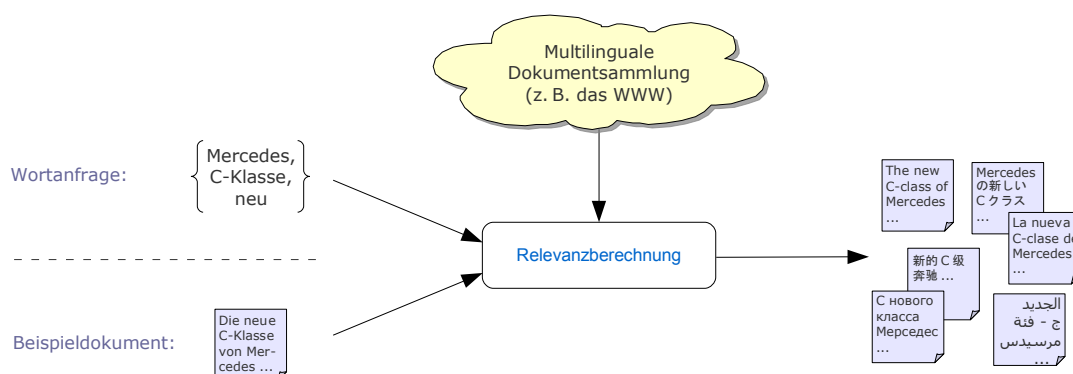


Abb. 2.3: Beispiel für das CLIR mit zwei unterschiedlichen Arten von Anfragen: Wortanfrage und Beispieldokument. In einer multilingualen Dokumentsammlung werden Dokumente in beliebigen Sprachen identifiziert, die inhaltlich zu der deutschen Wortanfrage bzw. dem deutschen Beispieldokument passen.

Der Schwerpunkt dieser Arbeit liegt in der sprachübergreifenden Plagiaterkennung. Die Problemstellung der sprachübergreifenden Plagiaterkennung entspricht der Problemstellung der sprachübergreifenden Ähnlichkeitsanalyse. Im folgenden Abschnitt 2.3 werden existierende Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse vorgestellt.

Es gibt eine Vielzahl von Verfahren und früheren Arbeiten zum CLIR, für die Retrieval-Situation, in der der Informationsbedarf in Form einer Wortanfrage spezifiziert wird. Für weitere Informationen hierzu wird auf [Oard und Dorr \(1996\)](#) sowie [Hull und Grefenstette \(1996\)](#) verwiesen.

2.3 Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse

Es kann grundsätzlich zwischen zwei Ansätzen zur sprachübergreifenden Ähnlichkeitsanalyse unterschieden werden:

1. Basierend auf Parallelkorpora.

Ein Parallelkorpus stellt Dokumente in den Sprachen L und L' bereit, die paarweise entweder Übersetzungen voneinander sind oder inhaltlich dasselbe Thema beschreiben.

2. Basierend auf Wörterbüchern bzw. Thesauri.

Ein Wörterbuch sowie ein multilingualer Thesaurus bieten die Möglichkeit, Wörter oder Konzepte von der Sprache L in die Sprache L' zu übersetzt.

Weiterhin werden die Ansätze dahingehend unterschieden, ob sie nur in einer geschlossenen Retrieval-Situation angewandt werden können oder auch für den Einsatz in einer offenen Retrieval-Situation geeignet sind. In **Abb. 2.4** sind die unterschiedlichen Ansätze zur sprachübergreifenden Ähnlichkeitsanalyse in einer Taxonomie dargestellt. Die Taxonomie enthält außerdem alle bisher bekannten Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse. Die Verfahren werden im Folgenden beschrieben.

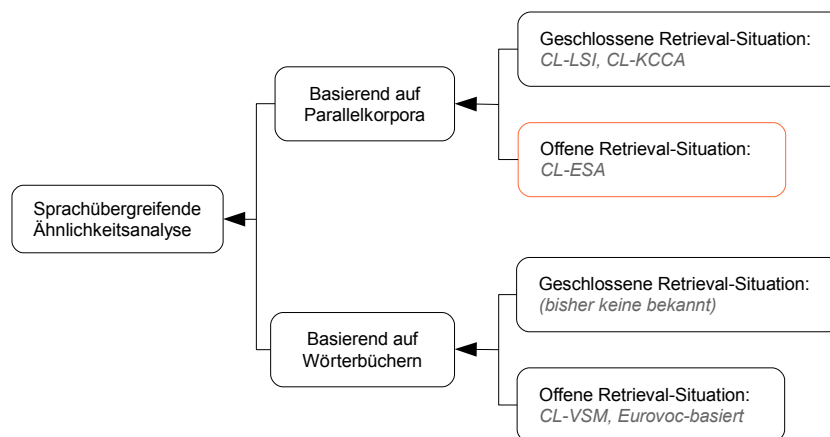


Abb. 2.4: Taxonomie der Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse (Potthast et al., 2007). Das Verfahren, das in dieser Arbeit vorgestellt wird, ist hervorgehoben.

Cross-Language Latent Semantic Indexing (CL-LSI) (Littman et al., 1998) ist eine Generalisierung des, aus dem monolingualen IR bekannten, Latent Semantic Indexing (LSI). CL-LSI basiert auf einem bilingualen Parallelkorpus, in den Sprachen L und L' , der eine Menge von Dokumentpaaren enthält. Ein Dokumentpaar besteht aus zwei Versionen des gleichen Dokuments, in den Sprachen L und L' . Jedes Dokumentpaar wird zu einem einzigen Dokument vereint. Zwei Wörter, in den Sprachen L und L' , die häufig zusammen in einem solchen Dokument vorkommen, können als ähnlich angesehen werden. Diese Dokumente werden verwendet um mittels Singulärwertzerlegung einen bilingualen LSI-Raum zu trainieren. In dem LSI-Raum können sowohl Dokumente in L , als auch Dokumente

in L' repräsentiert und miteinander verglichen werden. Mittels CL-LSI werden sehr gute Ergebnisse bei der sprachübergreifenden Ähnlichkeitsanalyse erreicht, allerdings kann CL-LSI aufgrund des hohen Rechenaufwands nur für kleine Dokumentsammlungen angewandt werden.

Bei der Cross-Language Kernel Canonical Correlation Analysis (CL-KCCA) (Vinokourov et al., 2003) wird für beide Dokumentmengen eines bilingualen Parallelkorpus, für die Sprachen L und L' , jeweils ein Vektorraum erstellt. Mittels KCCA werden Korrelationen zwischen den beiden Vektorräumen gelernt, mit der Annahme, dass jede Korrelation einer Aussage über die semantische Ähnlichkeit zwischen den Vektorräumen entspricht. Daraus ergeben sich sprachübergreifende Merkmale, durch die Dokumente, sowohl in L , als auch in L' , repräsentiert werden können und somit eine sprachübergreifende Ähnlichkeitsanalyse dieser Dokumente ermöglicht wird. Die CL-KCCA liefert sehr gute Ergebnisse bei der sprachübergreifenden Ähnlichkeitsanalyse, allerdings gilt auch hier, dass die CL-KCCA aufgrund des hohen Rechenaufwands nur für kleine Dokumentsammlungen praktikabel ist.

Die Cross-Language Explicit Semantic Analysis (CL-ESA) – die in Kapitel 5 vorgestellt wird – ist eine Generalisierung der Explicit Semantic Analysis (ESA) (Gabrilovich und Markovitch, 2007). Im Gegensatz zum CL-LSI und zur CL-KCCA kann die CL-ESA auch für große Dokumentkollektionen angewandt werden.

In dem sprachübergreifenden Vektorraummodell (Cross-Language Vector Space Model, CL-VSM) entspricht jede Dimension einem Paar, bestehend aus einem Schlüsselwort in der Sprache L und einem Schlüsselwort in der Sprache L' . Dokumente in L bzw. L' werden durch die Schlüsselwörter in L bzw. L' repräsentiert. Die Paare entsprechen sprachübergreifenden Merkmalen und ermöglichen eine sprachübergreifende Ähnlichkeitsanalyse der Dokumente. Durch ein bilinguales Wörterbuch können die Paare definiert werden (Levow et al., 2005). Sie können jedoch auch andere Informationen, außer Wörter, enthalten (Steinberger et al., 2004), z. B. Eigennamen, Zeitangaben, Zahlen oder Ortsangaben. Diese Informationen werden aus Namensverzeichnissen, geografischen Lexika bzw. Ortsregistern (Gazetteers) oder Thesauri gewonnen.

Das Eurovoc-basierte Verfahren ähnelt dem CL-VSM, allerdings werden hierbei die Deskriptoren des Eurovoc-Thesaurus⁵ als sprachübergreifende Merkmale zur Repräsentation von Dokumenten verwendet (Pouliquen et al., 2003).

⁵Eurovoc-Thesaurus: <http://europa.eu/eurovoc>.

Jedes der zuvor beschriebenen Verfahren basiert auf einem Dokumentmodell, das externes Wissen nutzt, um daraus sprachübergreifende Merkmale bzw. Konzepte zu gewinnen, durch die die Dokumente repräsentiert werden. Ein solches Dokumentmodell wird als „Konzeptraummodell“ bezeichnet. Die Verfahren verwenden verschiedene Varianten von Konzeptraummodellen, die sich in den zu Grunde liegenden Wissensbasen unterscheiden, z. B. wörterbuch-, thesaurus- oder parallelkorpusbasiertes Konzeptraummodell. Auf das Konzeptraummodell sowie die genannten Varianten wird in Kapitel 5 detailliert eingegangen. In Abschnitt 7.6.2 werden die Verfahren anhand verschiedener Eigenschaften miteinander verglichen.

3 Plagiaterkennung

Es gibt unterschiedliche Beweggründe für die Suche nach Plagiaten in Dokumenten. Beispielsweise im akademischen Bereich, um zu prüfen, ob eine Abschlussarbeit in eigenständiger Leistung erbracht wurde, oder zu kommerziellen Zwecken, etwa zur Erkennung von Copyright-Verletzungen, z. B. in der Literatur.

Wie lassen sich Plagiate in Dokumenten erkennen? Für Menschen ist es im Allgemeinen nicht schwer, einen Text oder einen Textabschnitt als Plagiat zu entlarven. Einem geübten Leser fällt z. B. eine plötzliche Abweichung des Schreibstils auf, die auf einen plagiierten Textabschnitt hindeuten kann. Aber auch die Erfahrung bzw. das Vorwissen eines Lesers ermöglichen es, Plagiate zu erkennen, beispielsweise indem festgestellt wird: „das habe ich doch schon mal irgendwo gelesen“. Die manuelle Suche nach Plagiaten ist allerdings sehr zeitaufwendig und daher nicht rentabel. Der Schwerpunkt dieser Arbeit liegt in der maschinellen Plagiaterkennung, d. h. in der Erkennung von Plagiaten durch ein Softwaresystem. Die Verfahren zur maschinellen Plagiaterkennung sind dem oben beschriebenen menschlichen Vorgehen nachempfunden.

In Abschnitt 3.1 wird die Vorgehensweise bei der Plagiaterkennung genauer erläutert und in drei Teilprobleme unterteilt. Außerdem werden verschiedene Plagiatvergehen sowie spezialisierte Verfahren zur Erkennung der daraus resultierenden Plagiate beschrieben und der Unterschied zwischen monolingualer und sprachübergreifender Plagiaterkennung verdeutlicht. Auf die Verfahren zur monolinguale Plagiaterkennung wird in Abschnitt 3.2 eingegangen.

3.1 Grundlegende Vorgehensweise

Der Ausgangspunkt bei der Plagiaterkennung ist ein verdächtiges Dokument, das auf Plagiate hin überprüft werden soll. Hierbei werden zwei Herangehensweisen unterschieden:

1. Globale Dokumentanalyse.

Es wird überprüft, ob es sich bei dem gesamten Dokument um ein Plagiat handelt.

2. Lokale Dokumentanalyse.

Es wird überprüft, ob das verdächtige Dokument plagierte Textabschnitte enthält.

Bei der lokalen Dokumentanalyse findet ein so genanntes Chunking statt, d.h. das verdächtige Dokument wird in eine Menge von Textabschnitten unterteilt. Im Fall einer globalen Dokumentanalyse findet kein Chunking statt, es wird das komplette Dokument analysiert. Im weiteren Verlauf dieser Arbeit wird nur die lokale Dokumentanalyse betrachtet, da sich diese auf die globale Dokumentanalyse verallgemeinern lässt, indem angenommen wird, dass das gesamte Dokument aus einem einzigen Textabschnitt besteht.

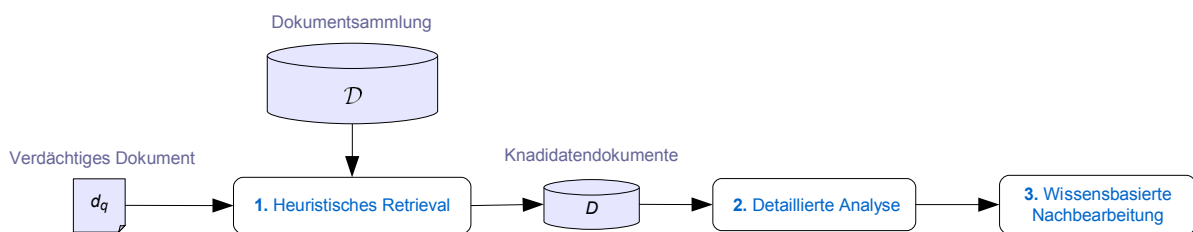


Abb. 3.1: Plagiaterkennung als ein Prozess bestehend aus den drei Schritten: Heuristisches Retrieval, detaillierte Analyse und wissensbasierte Nachbearbeitung. (Stein et al., 2007)

Das Ziel der Plagiaterkennung ist es, in einem verdächtigen Dokument d_q einen plagiierten Textabschnitt c_q zu erkennen, indem in einer Dokumentsammlung \mathcal{D} das Originaldokument, aus dem der Textabschnitt c_q plagiiert wurde, wiedergefunden wird. Häufig wird die Dokumentsammlung \mathcal{D} als „Referenzkorpus“ bezeichnet.¹ Die Plagiaterkennung kann in drei Schritte unterteilt werden (Stein et al., 2007), siehe **Abb. 3.1**.

1. Heuristisches Retrieval.

Aus einer Dokumentsammlung \mathcal{D} wird anhand von Heuristiken eine Menge D von Kandidatendokumenten bestimmt.

2. Detaillierte Analyse.

Die möglicherweise plagiierten Textabschnitte der Kandidatendokumente werden mit den Textabschnitten des verdächtigen Dokuments d_q verglichen.

3. Wissensbasierte Nachbearbeitung.

Die Textabschnitte aus den Kandidatendokumenten, die eine hohe Ähnlichkeit

¹Ein Referenzkorpus ist zur Plagiaterkennung nicht zwingend notwendig. Beispielsweise können mittels Stilanalyse plagierte Textabschnitte auch ohne Referenzkorpus erkannt werden. Die Stilanalyse wird in Abschnitt 3.2.4 beschrieben.

zu den Textabschnitten aus d_q aufweisen, werden genauer untersucht, z. B. wird überprüft, ob es sich um ein Plagiate oder ein korrektes Zitat handelt.

Es existieren verschiedene Formen von Plagiaten. Meyer zu Eißel und Stein (2006) haben unterschiedliche Plagiatvergehen in einer Taxonomie zusammengefasst, siehe **Abb. 3.2**. In der Taxonomie wird jedem Plagiatvergehen eine spezialisierte Erkennungsmethode zugeordnet, die – im Hinblick auf den Prozess der Plagiaterkennung in Abb. 3.1 – in dem zweiten Schritt „detaillierte Analyse“ eingesetzt werden. Grundsätzlich kann zwischen zwei Arten von Plagiatvergehen unterschieden werden: Ein Originaldokument wird eins-zu-eins kopiert, d. h. der Text wird Wort für Wort übernommen oder ein Originaldokument wird modifiziert, z. B. durch eine Veränderung des Satzbaus oder durch den Austausch von Wörtern. Weiterhin wird unterschieden, ob ein großer oder kleiner Teil eines Originaldokuments plagiiert wird und ob zur Plagiaterkennung ein Referenzkorpus zur Verfügung steht.

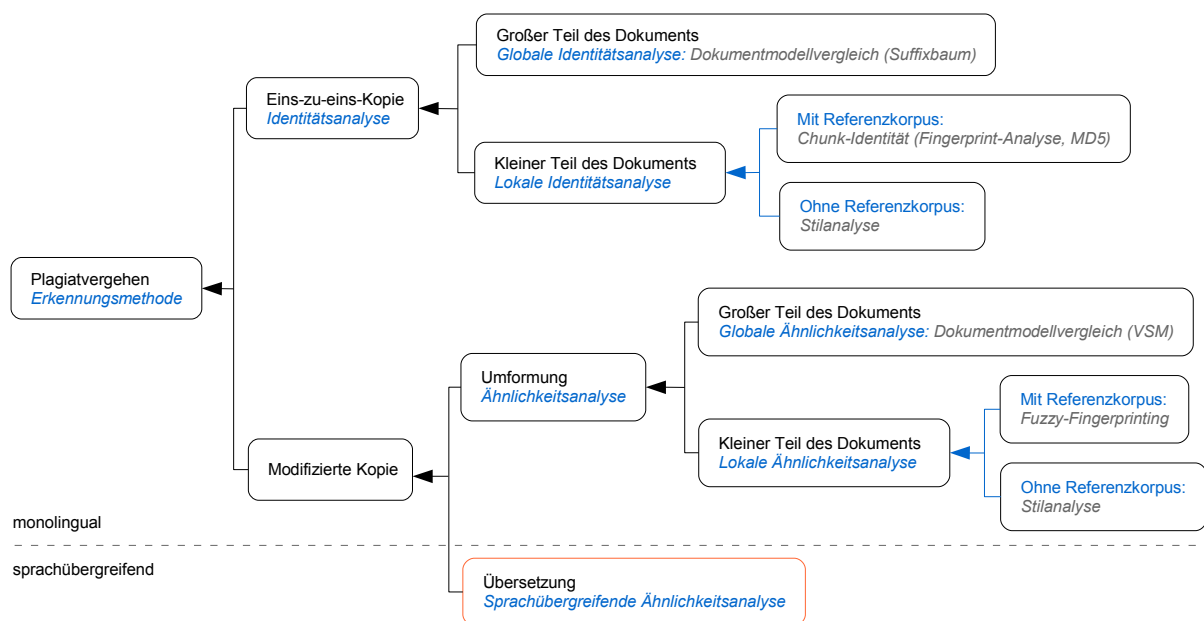


Abb. 3.2: Taxonomie der Plagiatvergehen, mit jeweils entsprechenden Erkennungsmethoden und konkreten Verfahren (Meyer zu Eißel und Stein, 2006). Für alle Arten von Plagiatvergehen, mit Ausnahme des sprachübergreifenden (hervorgehoben), existieren entsprechende Verfahren. In dieser Arbeit werden Verfahren zur sprachübergreifenden Plagiaterkennung vorgestellt.

Die Taxonomie ist in monolinguale und sprachübergreifende Plagiatvergehen unterteilt. Für alle Arten der monolingualen Plagiatvergehen existieren entsprechende Verfahren zur (monolingualen) Plagiaterkennung. Für die Erkennung von Plagiaten, die durch eine Übersetzung eines Originaldokuments in eine andere Sprache entstehen, wurde bisher noch keine konkrete Lösung vorgeschlagen. Die Erkennung solcher Plagiate wird als

„sprachübergreifende Plagiaterkennung“ bezeichnet² und ist Gegenstand dieser Arbeit. Im folgenden Kapitel 4 werden verschiedene Lösungsansätze zur sprachübergreifenden Plagiaterkennung vorgestellt.

Zunächst werden jedoch in Abschnitt 3.2 die Verfahren zur monolingualen Plagiaterkennung – die in Abb. 3.2 genannt werden – beschrieben.

3.2 Verfahren zur monolingualen Plagiaterkennung

Die im Folgenden beschriebenen Verfahren entsprechen – mit Bezug auf Abb. 3.1 – dem zweiten Schritt (detaillierte Analyse) im Prozess der Plagiaterkennung. Der Ausgangspunkt ist ein verdächtiges Dokument d_q und eine Menge von Kandidatendokumenten D .

3.2.1 Dokumentmodellvergleich

Das Prinzip wird an dem Vektorraummodell verdeutlicht. Um zu überprüfen, ob es sich bei einem Textabschnitt c_q des verdächtigen Dokuments d_q um ein Plagiat handelt, werden für c_q sowie für alle Textabschnitte c_x der Kandidatendokumente die Wortvektorrepräsentationen \mathbf{c}_q bzw. \mathbf{c}_x erstellt. Zwischen \mathbf{c}_q und allen \mathbf{c}_x werden die Ähnlichkeiten, z. B. anhand der Kosinusähnlichkeit, bestimmt. Falls ein \mathbf{c}_x eine hohe Ähnlichkeit zu \mathbf{c}_q aufweist, wird davon ausgegangen, dass es sich bei dem Textabschnitt c_q mit hoher Wahrscheinlichkeit um ein Plagiat des Textabschnitts c_x handelt.

Angenommen die Menge der Kandidatendokumente beinhaltet n Dokumente und die durchschnittliche Anzahl von Textabschnitten pro Dokument beträgt k , so ergibt sich eine Laufzeitkomplexität von $O(n \cdot k^2)$. Die quadratische Laufzeit ist jedoch kein Nachteil, da durch das vorherige heuristische Retrieval (der erste Schritt in Abb. 3.1) sichergestellt wird, dass n sehr klein ist.

²Sprachübergreifende Plagiaterkennung wird auch „Cross-Lingual Plagiarism Detection“, „Multilingual Plagiarism Detection“ oder „Multilingual Copy Detection“ genannt.

3.2.2 Fingerprinting

Für jeden Textabschnitt c_q des verdächtigen Dokuments d_q und für jeden Textabschnitt c_x der Kandidatendokumente werden die Fingerprints berechnet. Der Fingerprint $h(c)$ eines Textabschnitts c besteht aus einer kleinen Menge von Zahlen und wird anhand einer Hashfunktion h berechnet – hierzu wird meist der MD5-Algorithmus (Rivest, 1992) eingesetzt. Die Fingerprints aller Textabschnitte c_x werden in einer Hashtabelle \mathcal{H} gespeichert.

Um zu überprüfen, ob ein Textabschnitt c_q ein Plagiat ist, wird mit dem Fingerprint $h(c_q)$ die Hashtabelle \mathcal{H} angefragt. Falls es zu einer Hashkollision – d. h. der Hashwert existiert bereits in \mathcal{H} – mit einem Fingerprint $h(c_x)$ kommt, sind die beiden Textabschnitte d_q und d_x mit hoher Wahrscheinlichkeit gleich. Es gilt:

$$h(c_q) = h(c_x) \Rightarrow c_q = c_x$$

Eine Hashkollision zwischen den Fingerprints zweier Textabschnitte kann somit als ein Indikator für ein Plagiat angesehen werden.

Mittels Fingerprinting können nur Eins-zu-eins-Kopien erkannt werden, denn bereits die Veränderung eines einzelnen Zeichens in einem Textabschnitt erzeugt einen völlig anderen Fingerprint. Plagiierte Textabschnitte können durch Fingerprinting in linearer Zeit gefunden werden. Die Laufzeitkomplexität beträgt $O(n \cdot k)$, wobei n die Anzahl der Kandidatendokumente ist und k die durchschnittliche Anzahl von Textabschnitten in einem Dokument.

Es existieren viele Arbeiten, in denen Fingerprinting zur Plagiaterkennung eingesetzt wird. Frühere Arbeiten, die Details und Variationen zu Fingerprinting beschreiben sind Brin et al. (1995), Heintze (1996), Finkel et al. (2002) sowie Hoad und Zobel (2003). Eine bekannte Anwendung, in der Fingerprinting zur Plagiaterkennung eingesetzt wird, ist z. B. das SCAM Projekt³ (Shivakumar und Garcia-Molina, 1995).

3.2.3 Fuzzy-Fingerprinting

Das Fuzzy-Fingerprinting (Stein, 2005) ist eine Variante des Fingerprinting, mit speziellen Eigenschaften. Mittels Fuzzy-Fingerprinting kann in einer Menge C , die alle Textab-

³SCAM: Stanford Copy Analysis Method.

schnitte der Kandidatendokumente enthält, in quasi konstanter Zeit ein Textabschnitt $c_x \in C$ identifiziert werden, der eine hohe Ähnlichkeit zu einem Textabschnitt c_q des verdächtigen Dokuments besitzt.

Es wird eine spezielle Hashfunktion $h_\varphi : C \rightarrow U$ eingesetzt, die die Menge C auf ein Universum U von Hashwerten abbildet und folgende Eigenschaft besitzt:

$$h_\varphi(c_1) = h_\varphi(c_2) \Rightarrow \varphi(\mathbf{c}_1, \mathbf{c}_2) \geq 1 - \epsilon, \quad \text{mit } c_1, c_2 \in C, 0 < \epsilon \ll 1$$

$h_\varphi(c)$ ist der Fuzzy-Fingerprint eines Textabschnitts c und $\varphi(\mathbf{c}_1, \mathbf{c}_2)$ ist eine Ähnlichkeitsfunktion.

Die Fuzzy-Fingerprints $h_\varphi(c_x)$ der Textabschnitte $c_x \in C$ werden in einer Hashtabelle gespeichert. Eine Hashkollision zwischen dem Fuzzy-Fingerprint $h_\varphi(c_q)$ und einem Fuzzy-Fingerprint $h_\varphi(c_x)$ kann als Indiz für eine hohe Ähnlichkeit zwischen c_q und c_x angesehen werden und deutet darauf hin, dass c_q ein Plagiat von c_x ist.

Im Gegensatz zum Fingerprinting können mittels Fuzzy-Fingerprinting auch Plagiate erkannt werden, die durch Modifikation entstanden sind. Experimente von [Stein und Meyer zu Eißen \(2006\)](#) haben außerdem gezeigt, dass Fuzzy-Fingerprints eine durchschnittliche Chunkgröße von 100 Wörtern erlauben, während MD5-Fingerprints nur mit einer Chunkgröße von drei bis zehn Wörtern akzeptabel arbeiten. Mittels Fuzzy-Fingerprinting kann daher die Erkennung von Plagiaten stark verbessert und gleichzeitig die Fingerprint-Datenbasis verkleinert werden. In [Stein und Potthast \(2006\)](#) wird Fuzzy-Fingerprinting mit einem weiteren Konstruktionsprinzip für Hashfunktionen, dem Locality-Sensitive-Hashing, verglichen. Anhand verschiedener Experimente kann gezeigt werden, dass Fuzzy-Fingerprinting bei der Ähnlichkeitssuche dem Locality-Sensitive-Hashing überlegen ist.

Für detailliertere Informationen zum Fuzzy-Fingerprinting siehe [Stein \(2005\)](#) sowie [Stein und Meyer zu Eißen \(2006\)](#).

3.2.4 Stilanalyse

Die Stilanalyse ist ein Verfahren zur intrinsischen Erkennung von Plagiaten. Intrinsisch bedeutet „von innen her kommend“.⁴ Dementsprechend sind Verfahren zur intrinsischen

⁴Wikipedia, <http://de.wikipedia.org/wiki/Intrinsisch>.

Plagiaterkennung dazu in der Lage, plagierte Textabschnitte zu identifizieren, indem ausschließlich das verdächtige Dokument selbst analysiert wird – ein Referenzkorpus wird nicht benötigt.

Über bestimmte quantifizierbare Stilmerkmale kann der individuelle Schreibstil eines Autors berechnet werden. Stilmerkmale sind z. B. die durchschnittliche Anzahl der Wörter pro Satz oder die Wortvielfalt. Wird in einem verdächtigen Dokument festgestellt, dass der Schreibstil in einem Textabschnitt von dem Schreibstil des restlichen Dokuments abweicht, so ist dies ein Indiz dafür, dass es sich bei dem Textabschnitt mit hoher Wahrscheinlichkeit um ein Plagiat handelt.

Die Stilanalyse kann jedoch nicht zur globalen Dokumentanalyse eingesetzt werden. Der Grund dafür liegt darin, dass es nicht möglich ist ein komplett plagiirtes Dokument zu erkennen – bei dem z. B. der eigene Name auf ein fremdes Werk gesetzt wird –, da die Stilanalyse auf der Berechnung von Stilunterschieden innerhalb eines Dokuments beruht. Daher ist es auch nicht möglich, Dokumente zu analysieren, die von mehreren Autoren verfasst wurden, da in solchen Dokumenten natürlicherweise Stilunterschiede vorhanden sind. Aufgrund der intrinsischen Vorgehensweise ist es allerdings auch möglich, Plagiate zu erkennen, deren Originalquelle nicht digital verfügbar ist, z. B. wenn es sich bei der Quelle um ein Buch handelt.

In [Meyer zu Eißel et al. \(2007\)](#) sowie in [Stein und Meyer zu Eißel \(2007\)](#) wird die Stilanalyse detailliert beschrieben. Außerdem werden unterschiedliche Stilmerkmale vorgestellt und untersucht.

4 Sprachübergreifende Plagiaterkennung

Bisher wurde keine Lösung für das Problem der sprachübergreifenden Plagiaterkennung vorgeschlagen. In einigen früheren Arbeiten wird die sprachübergreifende Plagiaterkennung erwähnt und als offenes Problem identifiziert (Clough, 2003; Steinberger und Pouliquen, 2003; Steinberger et al., 2004; Meyer zu Eißel und Stein, 2006). Weiterhin erwähnen Pouliquen et al. (2003), dass ihr Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse auch zur sprachübergreifenden Plagiaterkennung eingesetzt werden kann. Es werden jedoch keine Experimente dazu durchgeführt und dieser Ansatz wird auch in neueren Arbeiten nicht weiter verfolgt. Jede der oben genannten Arbeiten betrachtet nur ein Teilproblem der sprachübergreifenden Plagiaterkennung, der gesamte Prozess wird nicht weiter ausgeführt.

In diesem Kapitel erfolgt eine ganzheitliche Betrachtung des Retrieval-Problems „sprachübergreifende Plagiaterkennung“. Die Problemstellung der sprachübergreifenden Plagiaterkennung wird in Abschnitt 4.1 definiert und es werden zwei grundsätzliche Teilaufgaben unterschieden. In Abschnitt 4.2 sowie in Abschnitt 4.3 werden für jede dieser Teilaufgaben verschiedenen Lösungsansätze vorgeschlagen.

4.1 Problemstellung

Das Ziel der sprachübergreifenden Plagiaterkennung ist das Erkennen von Plagiaten, die – mit Bezug auf die Taxonomie der Plagiatvergehen – durch Übersetzung entstehen (siehe Abschnitt 3.2).

Der Ausgangspunkt ist ein verdächtiges Dokument d_q in der Sprache L . Weiterhin existiert eine Dokumentsammlung \mathcal{D}' in der Sprache L' . Ein Textabschnitt c_q aus d_q wird

als Plagiat entlarvt, wenn in \mathcal{D}' das Originaldokument, aus dem c_q plagiiert wurde, wiedergefunden wird. Daraus ergeben sich zwei Teilaufgaben (vgl. Abb. 3.1):

1. Heuristisches Retrieval:

Aus \mathcal{D}' wird eine Menge D' von Kandidatendokumenten gewonnen, die Textabschnitte enthalten, die eine hohe Ähnlichkeit zu Textabschnitten aus d_q besitzen.

2. Detaillierte Analyse:

Die Textabschnitte der Kandidatendokumente werden mit den Textabschnitten von d_q verglichen. Hierzu müssen Methoden eingesetzt werden, die die Bestimmung der Ähnlichkeit zwischen zwei Textabschnitten in L und L' ermöglichen. Wird eine hohe Ähnlichkeit zwischen einem Textabschnitt c'_x eines Kandidatendokuments und einem Textabschnitt c_q festgestellt, so handelt es sich bei c_q mit großer Wahrscheinlichkeit um ein Plagiat von c'_x .

In den folgenden Abschnitten werden für jede Teilaufgabe verschiedene Lösungsansätze vorgestellt. Der Schwerpunkt dieser Arbeit liegt auf der detaillierten Analyse.

4.2 Heuristisches Retrieval

In dem UML-Aktivitätsdiagramm in **Abb. 4.1** sind zwei unterschiedliche Lösungsansätze für die Teilaufgabe „heuristisches Retrieval“ dargestellt:

1. Heuristisches Retrieval basierend auf Techniken des CLIR.
2. Heuristisches Retrieval basierend auf maschineller Übersetzung (Maschine Translation, MT).

In beiden Lösungsansätzen wird aus dem verdächtigen Dokument d_q eine Wortanfrage q' in der Sprache L' erstellt, mit der ein Schlüsselwortindex angefragt wird, der die Dokumentsammlung \mathcal{D}' repräsentiert. Bei dem Schlüsselwortindex handelt es sich um einen invertierten Index (Witten et al., 1999), der jedes Schlüsselwort der Dokumente aus \mathcal{D}' auf die Dokumente aus \mathcal{D}' abbildet, in denen es vorkommen. Dadurch werden die Dokumente $d' \in \mathcal{D}'$ identifiziert, in denen die Schlüsselwörter der Wortanfrage q' vorhanden sind bzw. die eine hohe Ähnlichkeit zu q' besitzen. Die Dokumente d' bilden die Menge der Kandidatendokumente D' .

Im Folgenden wird auf jeden der beiden Lösungsansätze detailliert eingegangen.

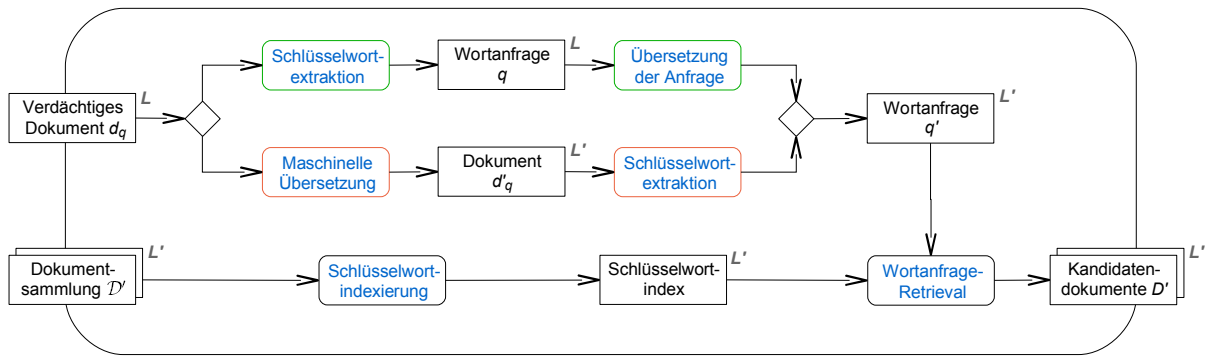


Abb. 4.1: Das UML-Aktivitätsdiagramm zeigt die Lösungsansätze für das heuristische Retrieval zur sprachübergreifenden Plagiaterkennung. Es sind zwei verschiedene Lösungsansätze dargestellt: 1. basierend auf CLIR-Techniken (grün) und 2. basierend auf maschineller Übersetzung (rot). In beiden Fällen wird für die Dokumentensammlung \mathcal{D}' ein Schlüsselwortindex erstellt und mit der Wortanfrage q' angefragt.

4.2.1 CLIR-basiertes heuristisches Retrieval

Die grundsätzliche Idee dieses Ansatzes ist es, robuste und gut erforschte Verfahren aus dem Bereich des CLIR (siehe Abschnitt 2.2) zum heuristischen Retrieval einzusetzen. Die Forschung im Bereich des CLIR konzentriert sich hauptsächlich auf die Retrieval-Situation, in der der Ausgangspunkt für die Suche eine Wortanfrage ist. Dementsprechend existieren für diese Situation eine Vielzahl von Verfahren, die akzeptable Ergebnisse liefern. In dem hier vorliegenden Fall ist der Ausgangspunkt jedoch ein Dokument – das verdächtige Dokument d_q .

Um CLIR-Verfahren einsetzen zu können, werden aus dem verdächtigen Dokument d_q Schlüsselwörter extrahiert, die eine Wortanfrage q bilden. Zur Schlüsselwortextraktion kann z. B. $tf \cdot idf$ eingesetzt werden, allerdings müssen für die Berechnung des idf -Wertes Trainingsdokumente vorhanden sein. Ein weiteres Verfahren, das zur Schlüsselwortextraktion keine Trainingsdokumente benötigt und vergleichbare Ergebnisse zu $tf \cdot idf$ liefert, basiert auf Co-Occurrence-Statistiken (Matsuo und Ishizuka, 2004).

Zur Übersetzung von Wortanfragen existieren viele unterschiedliche Verfahren. Eine einfache Technik ist, jedes Schlüsselwort der Anfrage durch eine entsprechende Übersetzung aus einem bilingualen Wörterbuch zu substituieren. Bei mehrdeutigen Wörtern ist dieses Vorgehen jedoch problematisch, da die korrekte Bedeutung der Wörter nicht erkannt werden kann. Die Bedeutung eines Worts geht aus dem Kontext hervor, in dem das Wort benutzt wird. In einer Wortanfrage sind allerdings kaum Kontextinformationen enthalten – vor allem, wenn die Anfrage nur aus wenigen Wörtern besteht. In diesem Fall wird ein Verfahren namens Query Expansion eingesetzt, um die Anfrage mit

zusätzlichen Wörtern, die ihrem wahrscheinlichen Kontext entstammen, anzureichern und so die Übersetzung der gesamten Anfrage zu erleichtern. [McNamee und Mayfield \(2002\)](#) erläutern in welchen Situationen es sinnvoll ist, Query Expansion einzusetzen und vergleichen verschiedene Query-Expansion-Techniken miteinander. Ein weiteres Verfahren, um die Qualität der Anfrage zu verbessern, ist das so genannte Relevance Feedback. Beim Relevance Feedback wird das Ergebnis der Anfrage analysiert und daraufhin die ursprüngliche Anfrage reformuliert, mit dem Ziel, dass die neue Anfrage relevantere Ergebnisse liefert. Für detaillierte Informationen zu Relevance Feedback siehe [Orengo und Huyck \(2006\)](#). Ein wörterbuchbasiertes Verfahren zur Übersetzung von Wortanfragen, in dem Query Expansion eingesetzt wird und das dazu in der Lage ist, Mehrdeutigkeiten aufzulösen, wird von [Levow et al. \(2005\)](#) beschrieben.

4.2.2 MT-basiertes heuristisches Retrieval

Das verdächtige Dokument d_q wird mittels maschineller Übersetzung in die Sprache L' übersetzt und aus der Übersetzung d'_q , analog zu dem CLIR-basierten Ansatz, durch eine Schlüsselwortextraktion die Anfrage q' erstellt.

Das Hauptaugenmerk dieses Lösungsansatzes liegt auf der kompletten Übersetzung des verdächtigen Dokuments. Hierzu kann auf existierende Anwendungen aus dem Bereich der maschinellen Übersetzung zurückgegriffen werden.¹ Eine umfangreiche Zusammenstellung aller bekannten MT-Systeme bietet [Hutchins \(2007\)](#). Das National Institute of Standards and Technology (NIST)² hat im November 2006 einen internationalen Vergleichstest von 48 MT-Systemen – sowohl kommerziellen als auch frei verfügbaren – durchgeführt. Detaillierte Informationen zu den Tests und der Evaluierung sowie die Ergebnisse sind in [NIST \(2006\)](#) zu finden. In den meisten Tests schneidet der Online-Übersetzungsservice von Google³ als bester ab und ist bei allen Tests unter den besten drei.

Heutige MT-Systeme können keine qualitativ hochwertigen Übersetzungen liefern ([Hutchins, 2005](#)). Für den Erfolg dieses Lösungsansatzes ist dies jedoch nicht von entscheidender Bedeutung, da zur Erstellung einer Wortanfrage keine exakte Übersetzung – die z. B. die Grammatik oder die Zeitform der Wörter berücksichtigt – benötigt wird.

¹Für detaillierte technische Informationen zu Machine Translation siehe [Brown et al. \(1990\)](#) und [Hutchins \(2005\)](#).

²National Institute of Standards and Technology: <http://www.nist.gov>.

³Online-Übersetzungsservice von Google: http://www.google.com/language_tools.

Daher kann anstelle eines MT-Systems auch eine so genannte Pseudo-Übersetzung (Kishida und Kando, 2005) eingesetzt werden, bei der jedes Wort des verdächtigen Dokuments durch eine entsprechende Übersetzung, aus einem bilingualen Wörterbuch, substituiert wird.

4.3 Detaillierte Analyse

Der Lösungsansatz für die detaillierte Analyse zur sprachübergreifenden Plagiaterkennung ist in dem UML-Aktivitätsdiagramm in **Abb. 4.2** dargestellt. Sowohl das verdächtige Dokument d_q , als auch die Kandidatendokumente der Menge D' werden mittels Chunking in Textabschnitte unterteilt. Daraus resultieren zwei Mengen von Textabschnitten C bzw. C' . Für alle möglichen Paare, bestehend aus einem Textabschnitt aus C und einem Textabschnitt aus C' , werden die Ähnlichkeiten berechnet. Hierzu werden Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse eingesetzt (siehe Abschnitt 2.3). Eine hohe Ähnlichkeit zwischen den Textabschnitten eines Paares deutet auf ein Plagiat hin.

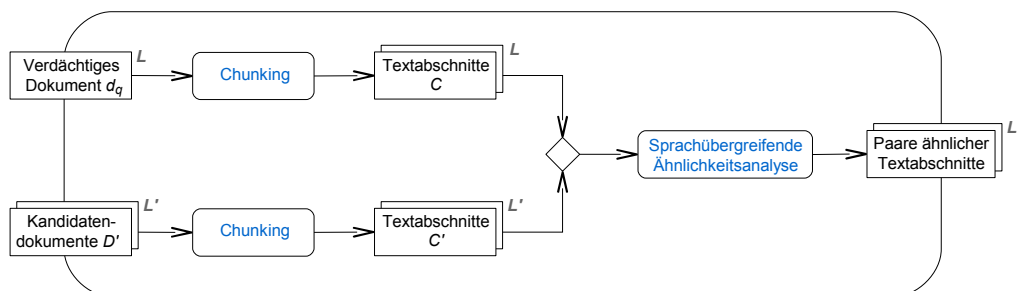


Abb. 4.2: Das UML-Aktivitätsdiagramm zeigt den Lösungsansatz für die detaillierte Analyse zur sprachübergreifenden Plagiaterkennung.

In dieser Arbeit wird ein neues Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse vorgestellt – die Cross-Language Explicit Semantic Analysis (CL-ESA). Der CL-ESA liegt ein wissensbasiertes, sprachübergreifendes Dokumentmodell zu Grunde, das es ermöglicht, sowohl die Textabschnitte aus C , als auch die aus C' zu repräsentieren und miteinander zu vergleichen. Das Dokumentmodell wird als „Konzeptraummodell“ bezeichnet. Im folgenden Kapitel 5 wird das Konzeptraummodell und die CL-ESA detailliert beschrieben.

5 Konzeptraummodell

Das Konzeptraummodell (Concept Space Model, CSM) ist ein Dokumentmodell, das es erlaubt – basierend auf einer enormen Menge von externem Wissen – die Semantik eines Dokuments zu modellieren. Im Gegensatz zum klassischen Vektorraummodell wird ein Dokument nicht durch die Wörter repräsentiert, die in dem Dokument vorkommen, sondern durch semantische Konzepte¹, die aus externem Wissen gewonnen werden und den Inhalt des Dokuments beschreiben.

Das Konzeptraummodell kann eingesetzt werden, um die inhaltliche Ähnlichkeit zwischen zwei Dokumenten zu bestimmen. Ein Verfahren, in dem ein Konzeptraummodell eingesetzt wird ist die Explicit Semantic Analysis (ESA) (Gabrilovich und Markovitch, 2007). Bei der monolingualen Ähnlichkeitsanalyse von Dokumenten, ist die ESA den bisherigen Standardtechniken, wie z. B. dem Vektorraummodell oder der Latent Semantic Analysis (LSA), überlegen (Gabrilovich und Markovitch, 2007).

Das Prinzip des Konzeptraummodells wird in Abschnitt 5.1 detailliert beschrieben, zunächst nur für die monolinguale Situation. Es wird erläutert, wie ein Dokument auf der Basis von Konzepten modelliert wird und wie mittels Konzeptindexierung aus externem Wissen Konzepte gewonnen werden. Außerdem wird die ESA beschrieben.

In Abschnitt 5.2 wird das sprachübergreifende Konzeptraummodell vorgestellt. Dabei handelt es sich um ein neues Dokumentmodell, das eine Erweiterung des monolingualen Konzeptraummodells ist und eine sprachübergreifende Repräsentation von Dokumenten ermöglicht. Es werden verschiedene Varianten von sprachübergreifenden Konzeptraummodellen erläutert, denen unterschiedliche Wissensbasen zu Grunde liegen.

In Abschnitt 5.3 wird die Cross-Language Explicit Semantic Analysis (CL-ESA) beschrieben. Die CL-ESA ist ein neues Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse,

¹Ein Konzept ist eine gedankliche Zusammenfassung (Vorstellung) von Gegenständen und Sachverhalten, die sich durch gemeinsame Merkmale auszeichnen.
(Wikipedia, <http://de.wikipedia.org/wiki/Konzept>.)

das auf einem sprachübergreifenden Konzeptraummodell basiert und eine Generalisierung der ESA darstellt. Im Gegensatz zur ESA können durch die CL-ESA verschiedene sprachige Dokumente modelliert und miteinander verglichen werden.

5.1 Monolinguales Konzeptraummodell (CSM)

Ein monolinguales Konzeptraummodell basiert auf einer Menge von Konzepten $K = \{k_1, \dots, k_m\}$. Die Konzepte spannen einen m -dimensionalen Konzeptraum auf, wobei jedes Konzept $k_i \in K$ einer Dimension entspricht. Ein Dokument d wird durch einen m -dimensionalen, gewichteten Konzeptvektor \mathbf{d} modelliert. Der i -te Eintrag von \mathbf{d} entspricht der Ähnlichkeit zwischen dem Dokument d und dem Konzept $k_i \in K$. Ein Beispiel für das monolinguale Konzeptraummodell ist in **Abb. 5.1** dargestellt.

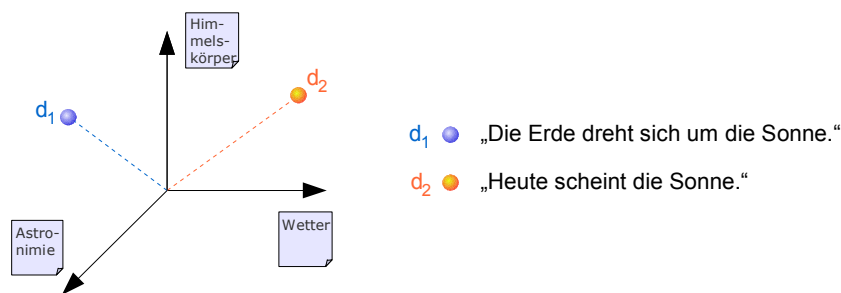


Abb. 5.1: Beispiel für das monolinguale Konzeptraummodell. Die Konzepte „Astronomie“, „Wetter“ und „Himmelskörper“, die jeweils durch ein Dokument beschrieben werden, spannen einen Konzeptraum auf, in dem die Dokumente d_1 und d_2 repräsentiert werden.

Da ein Konzept an sich abstrakt ist, wird eine Menge von Wissen \mathcal{K} benötigt, die die Konzepte beschreibt. Das Wissen in \mathcal{K} muss in einer geeigneten Form repräsentiert werden. Für jedes Konzept $k_i \in K$ wird ein so genannter Support-Vektor \mathbf{k}_i erstellt, der das Wissen aus \mathcal{K} , das k_i beschreibt, operationalisiert und k_i repräsentiert. Bei der Wissensmenge \mathcal{K} kann es sich z. B. um eine Dokumentsammlung D handeln, so dass jedes Konzept $k_i \in K$ durch ein Dokument $d_i \in D$ beschrieben wird (vgl. Abb. 5.1). Ein Support-Vektor \mathbf{k}_i kann in diesem Fall durch die Indexierung von d_i erstellt werden.

Die Konzepte des Konzeptraummodells sowie die entsprechenden Support-Vektoren können durch den Einsatz eines Verfahrens namens Konzeptindexierung aus einer unstrukturierten Dokumentsammlung gewonnen bzw. erstellt werden. Die Konzeptindexierung wird im folgenden Abschnitt 5.1.1 beschrieben.

In Abschnitt 5.1.2 wird die Explicit Semantic Analysis (ESA) beschrieben. Dabei handelt es sich um ein Verfahren, in dem ein monolinguales Konzeptraummodell zur Ähnlichkeitsanalyse von Dokumenten eingesetzt wird. Die Konzepte des monolingualen Konzeptraummodells werden hierbei durch die Artikel der Online-Enzyklopädie Wikipedia definiert, die auch als Wissensbasis für die Erstellung der Support-Vektoren dient. Alternativ zu der Konzeptindexierung werden die Konzepte bei der ESA aus einer manuell erstellten, strukturierten Wissensbasis gewonnen.

5.1.1 Konzeptindexierung

Durch Konzeptindexierung (Karypis und Han, 2000) kann aus einer Dokumentsammlung eine Menge von Konzepten mit entsprechenden Support-Vektoren gewonnen werden.

Der Ausgangspunkt ist eine unstrukturierte Dokumentsammlung D , die als Wissensbasis dient. Mittels Clustering-Verfahren² wird die Dokumentsammlung D in m disjunkte Teilmengen unterteilt, $D = D_1 \cup \dots \cup D_m$, so dass jede Teilmenge $D_i \subset D$ inhaltlich ähnliche Dokumente enthält. Jede Menge D_i definiert ein Konzept k_i , das durch den Inhalt der Dokumente in D_i beschrieben wird.

Der Support-Vektor \mathbf{k}_i , der ein Konzept k_i repräsentiert, entspricht dem Zentroidvektor der Dokumente in der Menge D_i und wird wie folgt berechnet:

$$\mathbf{k}_i = \frac{1}{|D_i|} \sum_{d \in D_i} \mathbf{d}$$

d ist ein Dokument aus der Menge D_i und \mathbf{d} ist die Wortvektorrepräsentation des Dokuments d .

5.1.2 Explicit Semantic Analysis (ESA)

Die Explicit Semantic Analysis (ESA) (Gabrilovich und Markovitch, 2007) ist ein Verfahren zur Ähnlichkeitsanalyse von Dokumenten, in dem ein monolinguales Konzeptraummodell eingesetzt wird. Die Konzepte des monolingualen Konzeptraummodells werden durch die Artikel der Online-Enzyklopädie Wikipedia definiert.

²Karypis und Han (2000) verwenden zur Gruppierung der Dokumente rekursive Bisektion.

Sei D eine Menge von Wikipedia-Artikeln. Ein Artikel $d_i \in D$ beschreibt ein Konzept $k_i \in K$ des Konzeptraummodells. Für jeden Artikel $d_i \in D$ wird die Wortvektorrepräsentation \mathbf{k}_i erstellt, die dem Support-Vektor für das Konzept k_i entspricht.

Zur Berechnung des Konzeptvektors \mathbf{d} für ein Dokument d wird zunächst die Wortvektorrepräsentation $\hat{\mathbf{d}}$ von d erstellt. Um den i -ten Eintrag $[\mathbf{d}]_i$ von \mathbf{d} zu bestimmen, wird die Kosinusähnlichkeit zwischen dem Wortvektor $\hat{\mathbf{d}}$ und dem Support-Vektor \mathbf{k}_i berechnet:

$$[\mathbf{d}]_i = \varphi_{\cos}(\hat{\mathbf{d}}, \mathbf{k}_i)$$

Die Repräsentation eines Dokuments d' in dem Konzeptraum erfolgt auf die gleiche Weise. Die inhaltliche Ähnlichkeit zwischen d und d' wird über die Kosinusähnlichkeit $\varphi_{\cos}(\mathbf{d}, \mathbf{d}')$ zwischen \mathbf{d} und \mathbf{d}' bestimmt.

Gabrilovich und Markovitch (2007) können zeigen, dass die ESA den Standardverfahren, wie dem Vektorraummodell und der Latent Semantic Analysis, bei der monolingualen Ähnlichkeitsanalyse von Dokumenten überlegen ist. In ihren Experimenten werden die Ergebnisse der drei Verfahren mit menschlichen Ähnlichkeitsbewertungen korreliert, die als Gold-Standard dienen. Dabei kann mit der ESA eine Korrelation von 0,72 erreicht werden. Im Gegensatz dazu erreicht das Vektorraummodell eine Korrelation von 0,5 und die Latent Semantic Analysis eine Korrelation von 0,6.

Durch den Einsatz von Wikipedia als Wissensbasis können mittels ESA Dokumente mit beliebigem Inhalt repräsentiert werden. Die aktuelle englische Version von Wikipedia beinhaltet über zwei Millionen Artikel und deckt damit so gut wie alle Themenbereiche ab. Jedoch nicht alle Wikipedia-Artikel sind geeignet, um ein Konzept eines Konzeptraummodells zu definieren. In **Tab. 5.1** sind verschiedenen Heuristiken bzw. Relevanzkriterien aufgeführt, um relevante Artikel herauszufiltern. Gabrilovich und Markovitch verwenden eine englische Wikipedia-Version, bestehend aus 1.187.839 Artikeln, von denen 241.393 den Relevanzkriterien aus der Tabelle entsprechen.

Relevanzkriterien bzw. Heuristiken

Der Artikel besitzt mehr als n Wörter, die keine Stoppwörter sind.

Der Artikel besitzt mehr als k eingehende und ausgehende Links.

Der Artikel beschreibt kein spezifisches Datum.

Bei dem Artikel handelt es sich nicht um eine „Weiterleitung“ oder „Begriffsklärung“.

Tab. 5.1: Relevanzkriterien für Wikipedia-Artikel (Gabrilovich und Markovitch, 2007). Ein Artikel wird als relevant angesehen, wenn er alle Kriterien erfüllt. „Weiterleitungen“ und „Begriffsklärungen“ (im Englischen „Disambiguation“) sind spezielle Wikipedia-Artikel, die lediglich auf andere Artikel verweisen. In Gabrilovich und Markovitch (2007) wird $n = 100$ und $k = 5$ verwendet.

5.2 Sprachübergreifendes Konzeptraummodell (CL-CSM)

Das sprachübergreifende Konzeptraummodell (Cross-Language Concept Space Model, CL-CSM) ist eine Generalisierung des monolingualen Konzeptraummodells, die es ermöglicht, verschiedensprachige Dokumente zu vergleichen.

Für eine sprachübergreifende Repräsentation von Dokumenten werden sprachübergreifende Merkmale benötigt. Sprachübergreifende Merkmale beschreiben ein Dokument unabhängig von einer bestimmten Sprache. In einem Konzeptraummodell werden die Dokumente durch Konzepte repräsentiert. Konzepte sind von Natur aus sprachübergreifend. Ein Beispiel hierfür ist in **Abb. 5.2** dargestellt. Ein Konzept kann somit als ein sprachübergreifendes Merkmal angesehen werden.

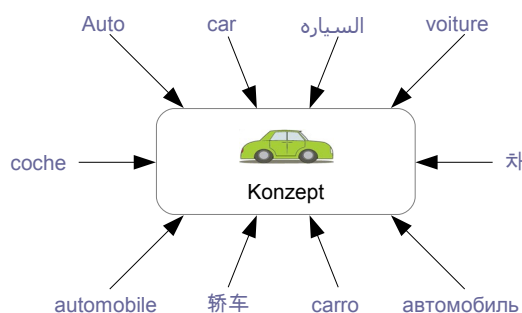


Abb. 5.2: Beispiel für ein (sprachübergreifendes) Konzept. Das Konzept selbst ist unabhängig von einer bestimmten Sprache, jedoch existiert in jeder Sprache eine Entsprechung des Konzepts, z. B. als Wort oder als Thema für ein Dokument.

Wie beim monolingualen Konzeptraummodell, wird auch bei der sprachübergreifenden Variante durch eine Menge von Konzepten $K = \{k_1, \dots, k_m\}$ ein Konzeptraum aufgespannt, in dem die Dokumente repräsentiert werden. Sei $\mathcal{L} = \{L_1, \dots, L_n\}$ die Menge der Sprachen, die von einem sprachübergreifenden Konzeptraummodell unterstützt werden.³ Die Grundlage des sprachübergreifenden Konzeptraummodells bildet eine multilinguale Wissensbasis $\mathcal{D} = \{\mathcal{K}_1, \dots, \mathcal{K}_{|\mathcal{L}|}\}$, die Wissen in jeder Sprache aus \mathcal{L} bereitstellt. Jede Wissensmenge $\mathcal{K}_j \in \mathcal{D}$ enthält Wissen, dass die Konzepte der Menge K in der Sprache $L_j \in \mathcal{L}$ beschreibt. Aus den Wissensmengen $\mathcal{K}_j \in \mathcal{D}$ werden die Support-Vektoren $\mathbf{k}_i^{L_j}$ erstellt, die die Konzepte $k_i \in K$ in jeder Sprache $L_j \in \mathcal{L}$ repräsentieren.

³Für ein monolinguales Konzeptraummodell gilt also $|\mathcal{L}| = 1$.

In **Abb. 5.3** ist ein Beispiel dargestellt, das das Prinzip des sprachübergreifenden Konzeptraummodells veranschaulicht.

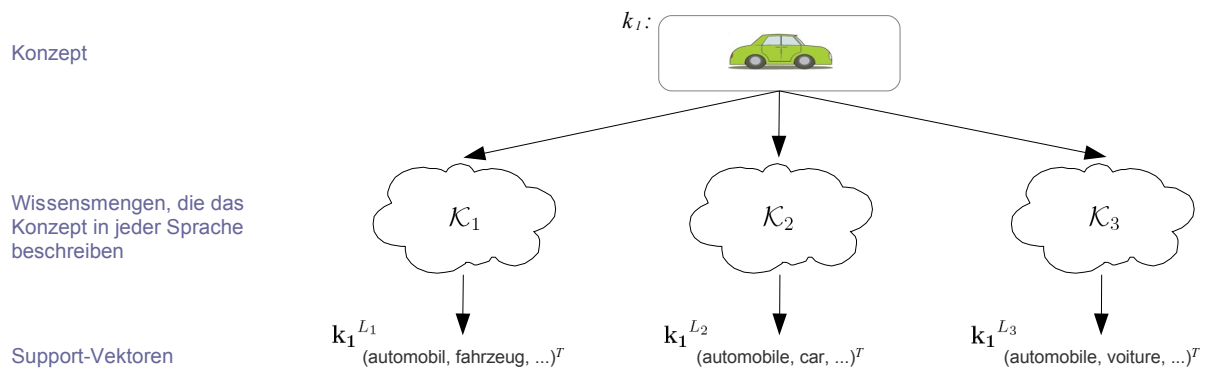


Abb. 5.3: Beispiel für das sprachübergreifende Konzeptraummodell. Die Menge der unterstützten Sprachen ist $\mathcal{L} = \{L_1, L_2, L_3\}$, mit $L_1 = \text{Deutsch}$, $L_2 = \text{Englisch}$ und $L_3 = \text{Französisch}$. Für jede Sprache existiert eine Wissensmenge, die das Konzept beschreibt. Aus den Wissensmengen werden die Support-Vektoren erstellt, die das Konzept in jeder Sprache repräsentieren.

Der Konzeptvektor \mathbf{d} für ein Dokument d wird analog zu dem monolingualen Konzeptraummodell – wie in Abschnitt 5.1 beschrieben – berechnet. Zur Berechnung von \mathbf{d} werden allerdings die Support-Vektoren benutzt, die die Konzepte in der Sprache von d repräsentieren. Um den i -ten Eintrag $[\mathbf{d}]_i$ des Konzeptvektors \mathbf{d} eines Dokuments d in der Sprache $L_j \in \mathcal{L}$ zu bestimmen, wird die Kosinusähnlichkeit zwischen der Wortvektorrepräsentation $\hat{\mathbf{d}}$ von d und dem Support-Vektor $\mathbf{k}_i^{L_j}$ berechnet. Hierzu kommen ausschließlich robuste Techniken des IR zum Einsatz, nämlich das klassische Vektorraummodell in Kombination mit der Kosinusähnlichkeit.

Um die Konzepte und die entsprechenden Support-Vektoren, die die Konzepte in verschiedenen Sprachen beschreiben, zu gewinnen, können unterschiedliche multilinguale Wissensbasen eingesetzt werden. Es existieren multilinguale Wissensbasen, die entweder implizit Konzepte definieren oder aus denen Konzepte extrahiert werden können. Die entsprechenden Support-Vektoren werden auf der Grundlage des Wissens konstruiert, das durch die multilinguale Wissensbasis in verschiedenen Sprachen bereitgestellt wird. Wie ein sprachübergreifendes Konzeptraummodell – basierend auf einer multilingualen Wissensbasis – realisiert wird, wird anhand der folgenden Wissensbasen erläutert:

- Multilinguales Wörterbuch (Abschnitt 5.2.1).
- Multilingualer Thesaurus (Abschnitt 5.2.2).
- Parallelkorpus (Abschnitt 5.2.3).

5.2.1 Wörterbuchbasiert

Die Wissensbasis \mathcal{D} des sprachübergreifenden Konzeptraummodells ist ein multilinguales Wörterbuch. Der Einfachheit halber wird das Prinzip zunächst an einem bilingualen Wörterbuch verdeutlicht. In **Abb. 5.4** ist ein Beispiel dargestellt.

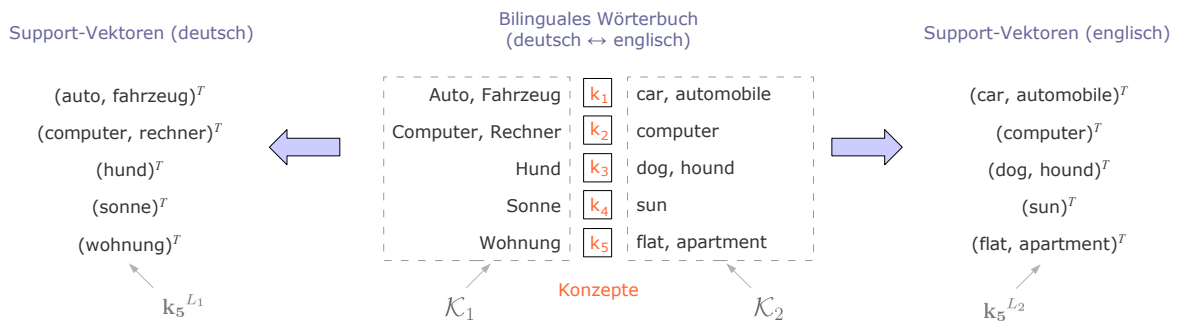


Abb. 5.4: Beispiel für das wörterbuchbasierte sprachübergreifende Konzeptraummodell. Das bilinguale Wörterbuch definiert die Konzepte k_1 bis k_5 . Die Menge der unterstützten Sprachen ist $\mathcal{L} = \{L_1, L_2\}$, mit $L_1 = \text{Deutsch}$ und $L_2 = \text{Englisch}$. Für jede Sprache L_j existiert eine Wissensmenge \mathcal{K}_j , die aus den Wörtern, in der Sprache L_j , des Wörterbuchs besteht, und die Konzepte beschreibt, mit $i \in \{1, \dots, 5\}$ und $j \in \{1, 2\}$. Aus den Wissensmengen \mathcal{K}_j werden die Support-Vektoren $\mathbf{k}_i^{L_j}$ erstellt.

Ein bilinguales Wörterbuch für die Sprachen L_1 und L_2 ordnet jedem Wort in L_1 ein bzw. mehrere Wörter in L_2 zu. Die Wörter, die einander zugeordnet werden, definieren ein Konzept, das die Bedeutung der Wörter zusammenfasst. Dieses Prinzip lässt sich für ein multilinguales Wörterbuch entsprechend generalisieren.

Die Menge der Sprachen \mathcal{L} , die von einem wörterbuchbasierten sprachübergreifenden Konzeptraummodell unterstützt wird, entspricht einer Teilmenge der Sprachen des zu Grunde liegenden multilingualen Wörterbuchs. Jede Wissensmenge $\mathcal{K}_j \in \mathcal{D}$, die die Konzepte in der Sprache $L_j \in \mathcal{L}$ beschreibt, besteht aus den Wörtern, in der Sprache L_j , des Wörterbuchs. Die Support-Vektoren $\mathbf{k}_i^{L_j}$, die ein Konzept $k_i \in K$ in der Sprache $L_j \in \mathcal{L}$ repräsentieren, werden aus den Wörtern der Wissensmenge \mathcal{K}_j , die k_i beschreiben, erstellt (siehe Abb. 5.4).

Es ist möglich, viele der meistgesprochenen Sprachen zu unterstützen, da mittlerweile eine Vielzahl von elektronischen Wörterbüchern existiert, die sehr umfangreich sind und immer mehr Sprachen beinhalten, wie z. B. Dictionary.com⁴ oder LEO⁵.

⁴Dictionary.com: <http://dictionary.reference.com>.

⁵Leo: <http://dict.leo.org>.

Ein grundsätzlicher Nachteil bei der Verwendung von Wörterbüchern als Wissensbasis für ein sprachübergreifendes Konzeptraummodell besteht darin, dass meist nur wenig Wissen über die Konzepte vorhanden ist. Ein Konzept wird häufig durch wenige Wörter definiert, was dazu führt, dass die Support-Vektoren nur eine geringe Menge an Informationen enthalten (vgl. Abb. 5.4). Ein weiteres Problem ergibt sich durch mehrdeutige Wörter. Die englische Übersetzung für das deutsche Wort „Blatt“ kann z. B. „blade“, „leaf“ und „newspaper“ sein, ein Konzept, das durch diese vier Wörter definiert wird ist allerdings wenig aussagekräftig. Dies erhöht den Aufwand zur Erstellung des Konzeptraummodells, da zusätzliche Verfahren zur Auflösung von Mehrdeutigkeiten eingesetzt werden müssen.

Das wörterbuchbasierte sprachübergreifende Konzeptraummodell ist mit dem CL-VSM (siehe Abschnitt 2.3) vergleichbar.

5.2.2 Thesaurusbasiert

Die Wissensbasis \mathcal{D} des sprachübergreifenden Konzeptraummodells bildet ein multilingualer Thesaurus. Ein Thesaurus definiert Konzepte implizit und stellt zu jedem Konzept spezielles Wissen bereit, wie z. B. Beziehungen zu anderen Konzepten, synonyme Begriffe oder eine kurze Beschreibung des Konzepts. In einem multilingualen Thesaurus ist dieses Wissen in verschiedenen Sprachen verfügbar. Ein Beispiel für das thesaurusbasierte sprachübergreifende Konzeptraummodell ist in **Abb. 5.5** dargestellt.

Die Menge der Sprachen \mathcal{L} , die von einem thesaurusbasierten sprachübergreifenden Konzeptraummodell unterstützt wird, entspricht einer Teilmenge der Sprachen des zu Grunde liegenden multilingualen Thesaurus. Das Wissen, das ein multilingualer Thesaurus in einer Sprache $L_j \in \mathcal{L}$ bereitstellt, bildet die Wissensmenge $\mathcal{K}_j \in \mathcal{D}$, für alle $j \in \{1, \dots, |\mathcal{L}|\}$. Die Support-Vektoren $\mathbf{k}_i^{L_j}$, die ein Konzept k_i in jeder Sprache \mathcal{K}_j repräsentieren, werden aus dem Wissen der Wissensmenge \mathcal{K}_j , das das Konzept k_i beschreibt, erstellt (siehe Abb. 5.5).

Mittlerweile existieren multilinguale Thesauri, die mehr als 20 Sprachen umfassen. Ein umfangreicher und frei verfügbarer Thesaurus ist z. B. der Eurovoc-Thesaurus⁶, die aktuelle Version 4.2 erscheint in 21 Sprachen.

⁶Eurovoc-Thesaurus: <http://europa.eu/eurovoc>.

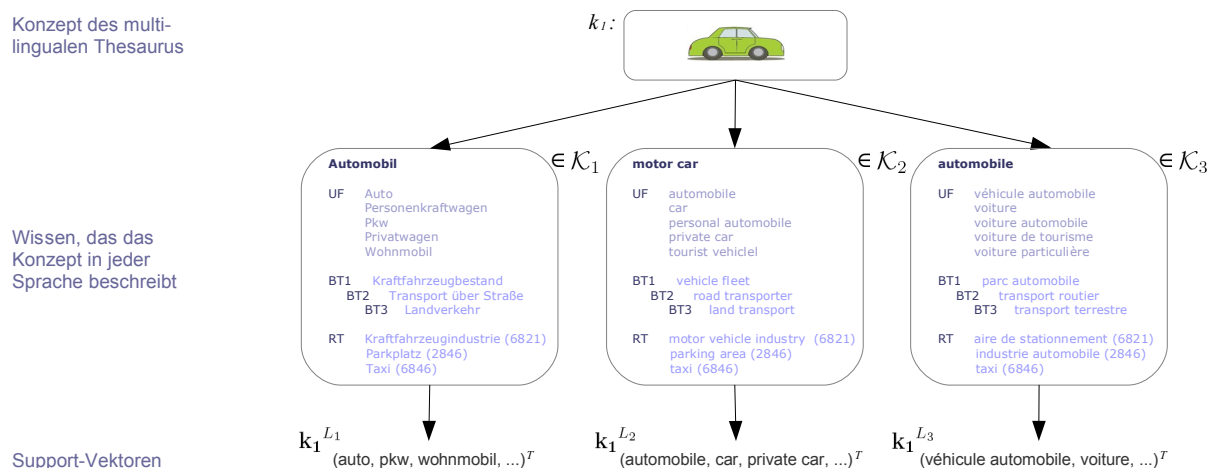


Abb. 5.5: Beispiel für das thesaurusbasierte sprachübergreifende Konzeptraummodell. Die Menge der unterstützten Sprachen ist $\mathcal{L} = \{L_1, L_2, L_3\}$, mit $L_1 = \text{Deutsch}$, $L_2 = \text{Englisch}$, $L_3 = \text{Französisch}$. Zu dem Konzept k_1 existiert in jeder Sprache $L_j \in \mathcal{L}$ eine Wissensmenge \mathcal{K}_j , mit $j \in \{1, 2, 3\}$. Die Support-Vektoren $k_1^{L_j}$ werden aus dem Wissen, der Wissensmengen \mathcal{K}_j das k_1 beschreibt, erstellt. Dargestellt ist ein Auszug aus dem Eurovoc-Thesaurus. Zu sehen sind pro Sprache synonyme Wörter (UF), übergeordnete Begriffe (BT) und assoziierte Begriffe (RT).

Ein Nachteil des thesaurusbasierten sprachübergreifenden Konzeptraummodells besteht darin, dass in einem Thesaurus meist wenig Wissen über die Konzepte vorhanden ist. Ein Konzept wird häufig durch wenige Wörter definiert (vgl. Abb. 5.5), so dass die Support-Vektoren nur eine geringe Menge an Informationen enthalten. Von [Steinberger et al. \(2002\)](#) wird jedoch ein Verfahren beschrieben, um einen multilingualen Thesaurus mit zusätzlichem Wissen anzureichern. Auf der Basis von Trainingsdokumenten, die den Konzepten des multilingualen Thesaurus manuell zugeordnet wurden, werden für jedes Konzept zusätzliche Wörter in jeder Sprache gelernt, die das Konzept beschreiben. Dieses Verfahren ist jedoch im Allgemeinen nicht praktikabel, da meist keine geeigneten Trainingsdokumente vorhanden sind und eine manuelle Erstellung sehr aufwändig ist.

Ein weiterer Nachteil, der sich bei der Verwendung eines Thesaurus als Wissensbasis für ein Konzeptraummodell ergibt, ist die Domänenabhängigkeit. Nahezu alle z. Z. verfügbaren Thesauri decken nur eine bestimmte Domäne bzw. einen bestimmten Themenbereich ab. Dadurch können nur die Dokumente gut repräsentiert werden, deren Inhalt aus dem jeweiligen Themenbereich stammt.

Ein thesaurusbasiertes sprachübergreifendes Konzeptraummodell wird in dem Eurovoc-basierten Ansatz zur sprachübergreifenden Ähnlichkeitsanalyse (siehe Abschnitt 2.3) eingesetzt.

5.2.3 Parallelkorpusbasiert

Ein Parallelkorpus ist eine multilinguale Dokumentsammlung, in der jedem Dokument d , in einer Sprache L , ein entsprechendes Dokument, in einer Sprache L' , das entweder eine Übersetzung von d ist oder inhaltlich dasselbe Thema beschreibt, zugeordnet ist.⁷ Die Granularität der Zuordnung kann unterschieden werden in, eine Zuordnung auf Dokument-, Paragraph-, Satz- oder Wortebene. Dabei gilt, je feiner die Granularität, desto schlechter ist die Verfügbarkeit entsprechender Parallelkorpora.

Es existieren so gut wie keine umfangreichen Parallelkorpora, die auf Wort- oder Satzebene zugeordnet sind. Ein bekannter Parallelkorpus, der elf Sprachen umfasst und pro Sprache ca. 500 Dokumente enthält, die auf Dokumentebene zugeordnet sind, ist der Europarl-Korpus (Koehn, 2005). Ein weiterer Parallelkorpus ist der JRC-Acquis-Korpus, der von Steinberger et al. (2006) erstellt wurde. Der JRC-Acquis-Korpus umfasst 20 Sprachen und besteht aus bis zu 8.000 Dokumenten pro Sprache, deren Paragraphen einander zugeordnet sind. Laut Steinberger et al. ist dies der größte Parallelkorpus dieser Art.

Für ein sprachübergreifendes Konzeptraummodell wird jedoch eine große Zahl an Dokumenten benötigt, die inhaltlich sehr viele verschiedene Themenbereiche abdecken. Eine Alternative zu den klassischen Parallelkorpora sind multilinguale Enzyklopädien, wie z. B. Wikipedia. Das Prinzip eines enzyklopädiebasierten sprachübergreifenden Konzeptraummodells wird im Folgenden erläutert.

Enzyklopädiebasiertes CL-CSM

Als Wissensbasis \mathcal{D} des sprachübergreifenden Konzeptraummodells kommt eine multilinguale Enzyklopädie zum Einsatz. Eine Enzyklopädie definiert Konzepte und stellt zu jedem Konzept einen Artikel bereit, der es präzise beschreibt. In einer multilingualen Enzyklopädie sind die Artikel in verschiedenen Sprachen verfügbar. Ein Beispiel für eine multilinguale Enzyklopädie ist Wikipedia. In **Abb. 5.6** wird das enzyklopädiebasierte sprachübergreifende Konzeptraummodell anhand eines Beispiels veranschaulicht.

Die Menge der Sprachen \mathcal{L} , die unterstützt wird, entspricht einer Teilmenge der Sprachen der zu Grunde liegenden multilingualen Enzyklopädie. Die Wissensmengen $\mathcal{K}_j \in \mathcal{D}$ bestehen aus den Artikeln, in der Sprache $L_j \in \mathcal{L}$, der multilingualen Enzyklopädie.

⁷In einigen früheren Arbeiten wird in diesem Zusammenhang zwischen einem „parallelen Korpus“ und einem „vergleichbaren Korpus“ unterschieden. Der hier verwendete Begriff „Parallelkorpus“ umfasst beide Situationen.

Die Support-Vektoren $\mathbf{k}_i^{L_j}$, die ein Konzept k_i in einer Sprache $L_j \in \mathcal{L}$ repräsentieren, werden aus den Artikeln der Enzyklopädie erstellt, die das Konzept k_i in der Sprache L_j beschreiben. Die Support-Vektoren werden durch eine Indexierung des entsprechenden Artikels erstellt.

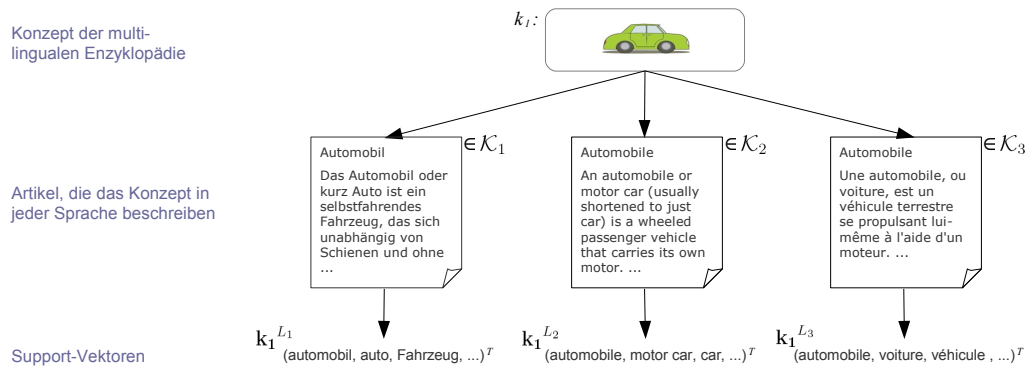


Abb. 5.6: Beispiel für das enzyklopädiebasierte sprachübergreifende Konzeptraummodell. Die Menge der unterstützten Sprachen ist $\mathcal{L} = \{L_1, L_2, L_3\}$, mit $L_1 = \text{Deutsch}$, $L_2 = \text{Englisch}$, $L_3 = \text{Französisch}$. Zu dem Konzept k_1 existiert in jeder Sprache $L_j \in \mathcal{L}$ ein Artikel aus der Wissensmenge \mathcal{K}_j , aus dem der Support-Vektoren $\mathbf{k}_1^{L_j}$ erstellt wird, mit $j \in \{1, 2, 3\}$. Dargestellt ist ein Auszug aus den Wikipedia-Artikeln für das Konzept „Auto“ in Deutsch, Englisch und Französisch.

Ein Vorteil des enzyklopädiebasierten sprachübergreifenden Konzeptraummodells ist, dass viel Wissen über die Konzepte zur Verfügung steht, da die Artikel, die die Konzepte beschreiben – je nach Enzyklopädie – sehr umfangreich sind. Die Konzepte einer Enzyklopädie können außerdem als qualitativ hochwertig angesehen werden, da sie manuell erstellt wurden. Weiterhin decken große Enzyklopädien, wie z. B. Wikipedia, so gut wie alle Themengebiete ab, so dass die Domänenabhängigkeit bei einem enzyklopädiebasierten Konzeptraummodell sehr gering ist.

Bisher ist keine Arbeit bekannt, in der eine multilinguale Enzyklopädie als Basis für eine sprachübergreifende Repräsentation von Dokumenten dient. In dem folgenden Abschnitt wird ein neues Verfahren vorgestellt, das diese Lücke schließt – die Cross-Language Explicit Semantic Analysis (CL-ESA).

5.3 Cross-Language Explicit Semantic Analysis (CL-ESA)

Die Cross-Language Explicit Semantic Analysis (CL-ESA) ist ein neues Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse, dem ein enzyklopädiebasiertes

sprachübergreifendes Konzeptraummodell zu Grunde liegt. Als Enzyklopädie kommt Wikipedia zum Einsatz. Die CL-ESA ist eine Generalisierung der ESA. Im Gegensatz zur ESA können durch die CL-ESA verschiedensprachige Dokumente repräsentiert und miteinander verglichen werden.

Die Menge $K = \{k_1, \dots, k_n\}$ der Konzepte wird durch Wikipedia-Artikel definiert. Sei $\mathcal{L} = \{L_1, \dots, L_n\}$ die Menge der Sprachen, die von der CL-ESA unterstützt wird. Die Wissensbasis $\mathcal{D} = \{D_1, \dots, D_{|\mathcal{L}|}\}$ besteht aus Mengen von Wikipedia-Artikeln, wobei jede Menge $D_j \in \mathcal{D}$ Artikel in der Sprache $L_j \in \mathcal{L}$ enthält. Dabei gilt, dass der i -te Artikel d_i aus jeder Menge $D_j \in \mathcal{D}$ das i -te Konzept $k_i \in K$ in der Sprache L_j beschreibt.

Zur Repräsentation eines Dokuments d , in einer Sprache $L_j \in \mathcal{L}$, durch den Konzeptvektor \mathbf{d} , wird die ESA eingesetzt, mit der Menge $D_j \in \mathcal{D}$ als Wissensbasis. Analog wird ein Dokument d' , in der Sprache $L_k \in \mathcal{L}$, durch den Konzeptvektor \mathbf{d}' repräsentiert. Die inhaltliche Ähnlichkeit zwischen d und d' wird unabhängig von der Sprache der Dokumente über die Kosinusähnlichkeit $\varphi_{\cos}(\mathbf{d}, \mathbf{d}')$ zwischen den Konzeptvektoren \mathbf{d} und \mathbf{d}' bestimmt.

Im folgenden Kapitel 6 wird auf die Implementierung der CL-ESA eingegangen.

6 Details zur Implementierung

Im Rahmen dieser Arbeit wurde die CL-ESA für die Sprachen Deutsch und Englisch realisiert. Eine Erweiterung um zusätzliche Sprachen ist mit geringem Aufwand möglich. Im Folgenden werden Details zur Implementierung beschrieben.

Zunächst wird in Abschnitt 6.1 auf die Konstruktion des Wikipedia-basierten sprachübergreifenden Konzeptraummodells eingegangen. In Abschnitt 6.2 wird die Repräsentation von Dokumenten erläutert.

6.1 Konstruktion des Wikipedia-basierten CL-CSM

Der CL-ESA liegt ein Wikipedia-basiertes sprachübergreifendes Konzeptraummodell zu Grunde. In diesem Abschnitt wird die Konstruktion des Konzeptraummodells für die Sprachen Deutsch und Englisch erläutert. Das Vorgehen ist in **Abb. 6.1** dargestellt und kann in drei Schritte unterteilt werden, die für jede Sprache ausgeführt werden:

- Extraktion der relevanten Artikel aus Wikipedia (Filtern).
- Konstruktion eines bilingualen Thesaurus aus den relevanten Artikeln.
- Indexierung der relevanten Artikel und Erstellung eines invertierten Indexes.

Das sprachübergreifende Konzeptraummodell besteht aus einem bilingualen Thesaurus sowie aus einem invertierten Index, der die deutschen Artikel enthält und einem invertierten Index, der die englischen Artikel enthält. Sowohl der deutsche als auch der englische invertierte Index definieren ein Vokabular V^{de} bzw. V^{en} , das aus den Schlüsselwörtern besteht, die in dem entsprechenden invertierten Index enthalten sind (vgl. Abb. 6.1). Für den deutschen invertierten Index gilt z. B. $V^{de} = \{f_1, \dots, f_n\}$.

In den folgenden Abschnitten werden die drei Schritte zur Konstruktion eines Wikipedia-basierten sprachübergreifenden Konzeptraummodells detailliert beschrieben.

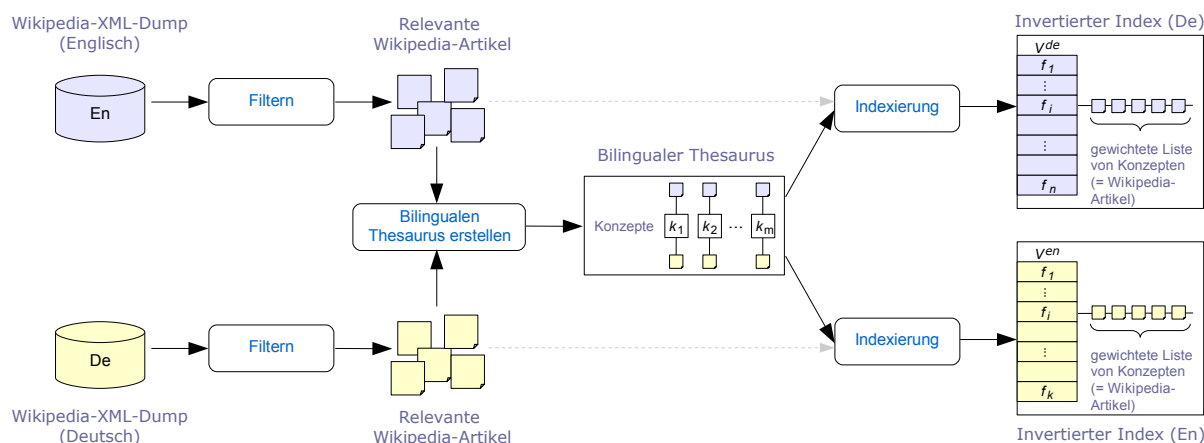


Abb. 6.1: Konstruktion eines Wikipedia-basierten sprachübergreifenden Konzeptraummodells für die Sprachen Deutsch und Englisch. Das Konzeptraummodell besteht aus dem bilingualen Thesaurus sowie aus dem deutschen und englischen invertierten Index.

6.1.1 Extraktion der relevanten Artikel (Filtern)

Der gesamte Inhalt von Wikipedia wird in Form von Dumps zum Download bereitgestellt.¹ Bei einem Dump handelt es sich um eine Datei im XML-Format, die verschiedene Inhalte von Wikipedia enthält. Die Dumps sind pro Sprachversion von Wikipedia verfügbar.

In dieser Arbeit werden die `pages-articles`-Dumps verwendet, die den kompletten Text der Wikipedia-Artikel in der aktuellen Version enthalten. Informationen zu den Dumps, aus denen das sprachübergreifende Konzeptraummodell konstruiert wird, sind in **Tab. 6.1** dargestellt.

| Wikipedia-Dump | Größe komprimiert | Größe entpackt |
|------------------------------------|-------------------|----------------|
| Deutscher XML-Dump vom 03.09.2007 | ca. 930 MB | ca. 3,7 GB |
| Englischer XML-Dump vom 02.08.2007 | ca. 2,7 GB | ca. 12,2 GB |

Tab. 6.1: Informationen zu den verwendeten Wikipedia-Dumps.

Aufgrund der Größe der XML-Dateien müssen effiziente Verfahren eingesetzt werden, um die relevanten Artikel zu extrahieren. Zum Parsen der XML-Dateien wird ein SAX-Parser² verwendet. Der SAX-Parser ist ein event-basierter Parser, der es ermöglicht, nur die XML-Tags zu betrachten, die für die jeweilige Anwendung von Bedeutung sind. In dem UML-Aktivitätsdiagramm in **Abb. 6.2** sind die grundlegenden Schritte dargestellt,

¹Wikipedia Downloads: <http://download.wikipedia.org/>.

²Simple API for XML: <http://www.saxproject.org>.

die während des Parsens eines Wikipedia-Dumps für jeden Artikel durchgeführt werden.

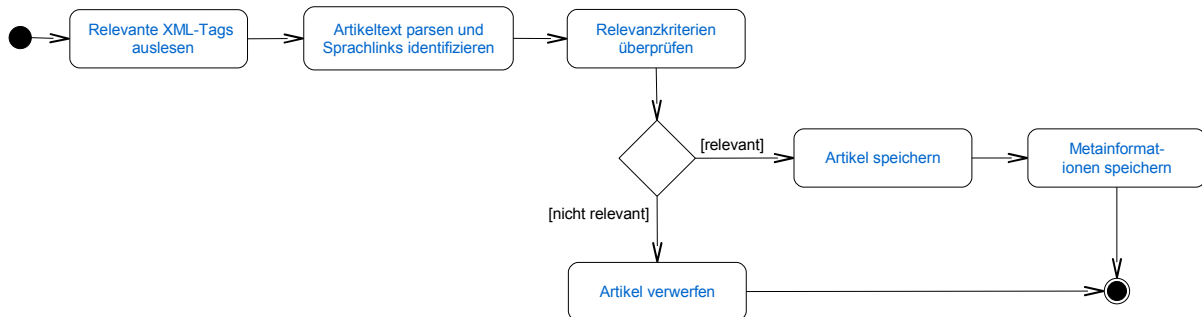


Abb. 6.2: Das UML-Aktivitätsdiagramm zeigt die grundlegenden Schritte, die während des Parsens eines Wikipedia-Dumps für jeden Artikel durchgeführt werden.

Für diese Anwendung sind nur die Id, der Titel und der Text eines Artikels von Interesse. Diese Informationen können direkt aus entsprechenden XML-Tags ausgelesen werden (`<revision-id>`, `<title>` und `<text>`).

Der Text der Artikel liegt in einer speziellen Wikipedia-Syntax vor, die zusätzlich verschiedenen HTML-Elemente enthält. Dementsprechend wird der Artikeltext selbst noch einmal geparkt und „bereinigt“. Dabei werden die Sprachlinks ausgelesen, die im Artikeltext durch einen speziellen Ausdruck in der Wikipedia-Syntax definiert sind. Sprachlinks in einem Wikipedia-Artikel verweisen auf weitere Wikipedia-Artikel, die dasselbe Konzept in einer anderen Sprache beschreiben. Ein Sprachlink, der z. B. auf den englischen Artikel „dog“ verweist, hat in der Wikipedia-Syntax die folgende Form: `[[en:dog]]`.

Aufgrund der gewonnenen Informationen werden die Relevanzkriterien überprüft. Dabei gilt, dass ein deutscher Artikel einen englischen Sprachlink besitzen muss und ein englischer Artikel einen deutschen Sprachlink. Dies ist notwendig, um ein sprachübergreifendes Konzeptraummodell für die Sprachen Deutsch und Englisch zu konstruieren. Zusätzlich werden weitere Relevanzkriterien überprüft (siehe Abschnitt 5.1.2).

Falls der Artikel die geforderten Relevanzkriterien nicht erfüllt, wird er verworfen. Andernfalls wird der Text des Artikels in einer Datei abgespeichert, mit der Artikel-Id als Dateiname. Zusätzlich werden zu jedem Artikel Metainformationen gespeichert, wie der Titel, die Sprache, die Sprachlinks und die Titel der Artikel, auf die die Sprachlinks verweisen.

6.1.2 Konstruktion eines bilingualen Wikipedia-Thesaurus

Der Ausgangspunkt für die Konstruktion des bilingualen Wikipedia-Thesaurus ist eine Menge D^{de} von relevanten Wikipedia-Artikeln in Deutsch und eine Menge D^{en} von relevanten Wikipedia-Artikeln in Englisch (vgl. Abb. 6.1).

Bei der Auswahl der Artikel wird sichergestellt, dass die relevanten englischen Artikel einen deutschen Sprachlink besitzen und die relevanten deutschen Artikel einen englischen Sprachlink. Falls der englische Sprachlink eines Artikels $d_1 \in D^{de}$ auf den Artikel $d_2 \in D^{en}$ zeigt und der deutsche Sprachlink des Artikels d_2 auf den Artikel d_1 , so bilden die Artikel d_1 und d_2 ein Paar.

Es wird bei der Auswahl der Artikel jedoch nicht sichergestellt, dass die Vereinigung der beiden Mengen D^{de} und D^{en} nur Paare enthält. Lediglich die Artikel in einer Teilmenge von D^{de} und die in einer Teilmenge von D^{en} sind Paare, siehe **Abb. 6.3**. Der Grund dafür ist, dass in vielen Fällen ein Artikel eines Paares nicht alle Relevanzkriterien erfüllt und daher nicht extrahiert wird, während der andere Artikel des Paares alle Kriterien erfüllt und in die Menge der relevanten Artikel aufgenommen wird.

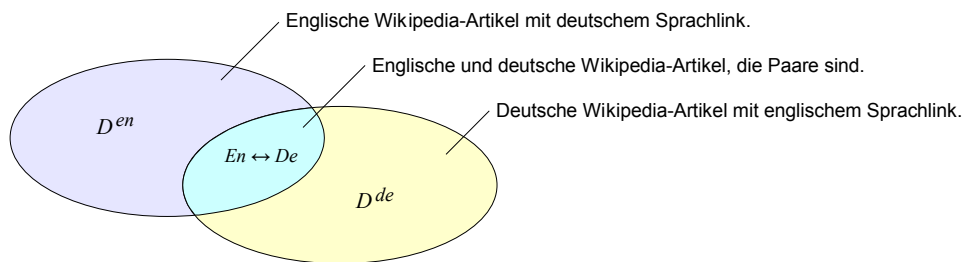


Abb. 6.3: Beispiel für Wikipedia-Artikel, die Paare bilden.

Nur Artikel, die ein Paar bilden sind geeignet, ein Konzept des sprachübergreifenden Konzeptraummodells zu definieren. Bei der Konstruktion des bilingualen Thesaurus werden die Artikelpaare aus den Mengen D^{de} und D^{en} identifiziert und jedem Paar eine eindeutige Konzept-Id zugeordnet. Der bilinguale Thesaurus ist eine Datenstruktur, die diese Informationen verwaltet. In **Abb. 6.4** ist ein Beispiel eines bilingualen Thesaurus dargestellt.

| De | Titel | Artikel-Id | Konzept-Id | Artikel-Id | Titel | En |
|----|-------------------------------------|------------|------------|------------|--|----|
| | „Wettlauf um Afrika“ | 35648205 | 23933 | 147857042 | „Scramble for Africa“ | |
| | „Steven Spielberg“ | 36250012 | 23934 | 148919784 | „Steven Spielberg“ | |
| | „Party“ | 36271723 | 23935 | 148970179 | „Party“ | |
| | „Das Rätsel der unheimlichen Maske“ | 35987116 | 23936 | 148177281 | „The Phantom of the Opera (1962 film)“ | |

Abb. 6.4: Auszug aus einem bilingualen Wikipedia-Thesaurus für die Sprachen Deutsch und Englisch. Der bilinguale Thesaurus verwaltet die Konzepte, die jeweils eine eindeutige Konzept-Id besitzen. Jedem Konzept ist ein deutscher und ein englischer Artikel zugeordnet, die das Konzept beschreiben.

6.1.3 Indexierung

Der bilinguale Thesaurus enthält die relevanten Artikel, die die Konzepte des sprachübergreifenden Konzeptraummodells definieren. Aus jedem dieser Artikel muss ein Support-Vektor erstellt werden, der das entsprechende Konzept repräsentiert (vgl. Abschnitt 5.2). Hierzu werden die Artikel indexiert. Für jeden Artikel wird eine Wortvektorrepräsentation erstellt, die dem Support-Vektor entspricht. In dem UML-Aktivitätsdiagramm in **Abb. 6.5** sind die grundlegenden Schritte des Indexierungsprozesses dargestellt. Die Indexierung erfolgt jeweils für die deutschen und die englischen Artikel, die in dem bilingualen Thesaurus enthalten sind.

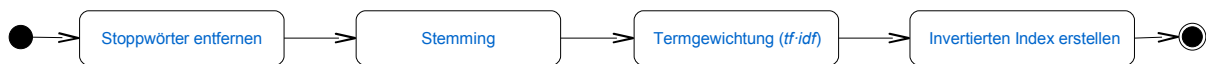


Abb. 6.5: Das UML-Aktivitätsdiagramm zeigt die grundlegenden Schritte des Indexierungsprozesses.

Zunächst werden die Stoppwörter aus den Artikeln entfernt. Weiterhin kommt ein Stemmer zum Einsatz, um die Wörter auf ihre Stammformen zu bringen. Hier wird der Snowball-Stemmer³ verwendet. Dabei handelt es sich um einen sprachabhängigen Stemmer, der für viele europäische Sprachen verfügbar ist. Die resultierenden Schlüsselwörter werden mittels $tf \cdot idf$ gewichtet. Der idf -Wert bezieht sich auf die Menge der Artikel in der jeweiligen Sprache, die in dem bilingualen Thesaurus enthalten sind.

Da die Menge der zu indexierenden Artikel sehr groß ist, muss eine effiziente Datenstruktur verwendet werden, um einerseits die Wortvektorrepräsentationen zu speichern und andererseits, einen schnellen Zugriff zur Laufzeit zu ermöglichen. Daher wird aus den gewichteten Wortvektoren der Artikel ein invertierter Index erstellt und abgespeichert.

³Snowball-Stemmer: <http://snowball.tartarus.org>.

Invertierter Index

Das Prinzip des invertierten Indexes ist in **Abb. 6.6** dargestellt. Der invertierte Index ermöglicht einen Zugriff über die Schlüsselwörter. Zu jedem Schlüsselwort existiert eine Liste von Tupeln, die so genannte Postliste. Die Tupel in der Postliste eines Schlüsselworts enthalten die Ids der Artikel, in denen das Schlüsselwort vorkommt, und den $tf \cdot idf$ -Wert, der aussagt, wie relevant der Artikel für das Schlüsselwort ist.

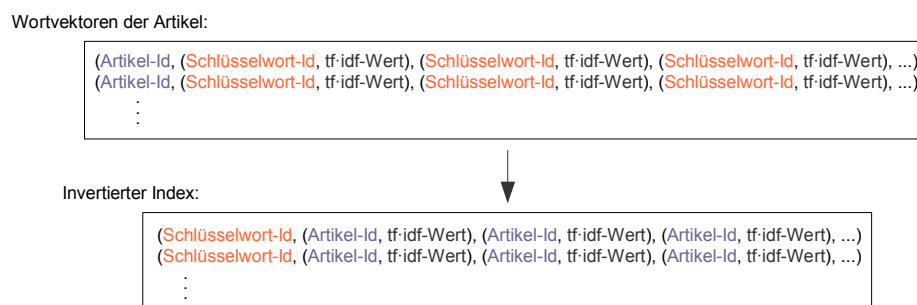


Abb. 6.6: Beispiel für die Erstellung eines invertierten Indexes aus den Wortvektorrepräsentationen der Artikel.

Realisiert wird der invertierte Index durch einen in C++ implementierten Hashindex, der im Rahmen des Projekts „Wikipedia in the Pocket“ ([Potthast, 2007](#)) entwickelt und für den Einsatz zu diesem Zweck modifiziert wurde.

Der Hashindex stellt u. a. die Funktion `lookup` bereit, die es ermöglicht, die komplette Postliste für ein Schlüsselwort aus dem invertierten Index auszulesen.

6.2 Repräsentation von Dokumenten

Folgende Schritte werden durchgeführt, um ein Dokument d durch das sprachübergreifende Konzeptraummodell zu modellieren:

1. d wird indexiert und durch die Schlüsselwörter des Vokabulars, des invertierten Indexes in der Sprache des Dokuments, repräsentiert.
2. Aus der Wortvektorrepräsentation von d wird, basierend auf dem sprachübergreifenden Konzeptraummodell, ein Konzeptvektor berechnet, der den Inhalt des Dokuments sprachübergreifend repräsentiert.

Auf den ersten Schritt wird in Abschnitt 6.2.1 eingegangen. Der zweite Schritt wird in Abschnitt 6.2.2 erläutert.

6.2.1 Dokumentindexierung

In **Abb. 6.7** ist ein Beispiel für die Indexierung eines deutschen Dokuments dargestellt.

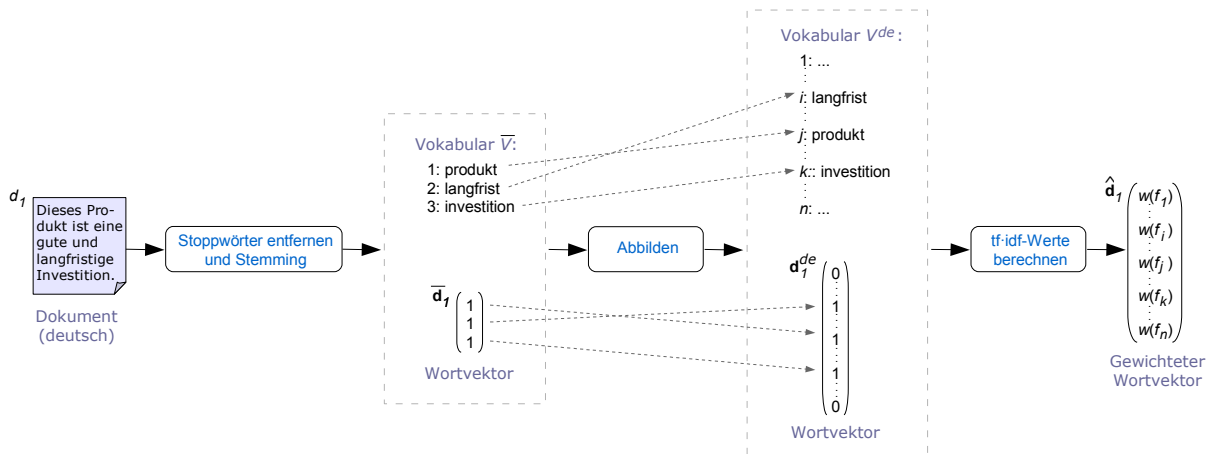


Abb. 6.7: Beispiel für die Indexierung eines deutschen Dokuments. Es werden die Stoppwörter entfernt und Stemming angewandt. Daraus resultiert ein Vokabular \bar{V} und eine entsprechende Wortvektorrepräsentation $\bar{\mathbf{d}}_1$ des Dokuments d_1 mit einfachen Häufigkeiten. Das Vokabular \bar{V} wird auf das Vokabular V^{de} des deutschen invertierten Indexes aus dem sprachübergreifenden Konzeptraummodell abgebildet und ein entsprechender $|V^{de}|$ -dimensionaler Wortvektor \mathbf{d}_1^{de} erstellt. Die Schlüsselwörter des Wortvektors \mathbf{d}_1^{de} werden mittels $tf \cdot idf$ gewichtet und so der Wortvektor $\hat{\mathbf{d}}_1$ erstellt.

Um ein Dokument d zu indexieren, werden die Stoppwörter aus d entfernt und ein Snowball-Stemmer eingesetzt. Die daraus resultierenden Schlüsselwörter bilden ein Vokabular \bar{V} . Basierend auf dem Vokabular \bar{V} wird ein Wortvektor $\bar{\mathbf{d}}$ für das Dokument d erstellt, dessen Einträge die Häufigkeiten der Schlüsselwörter aus \bar{V} in d sind. Die Schlüsselwörter des Vokabulars \bar{V} werden auf die Schlüsselwörter des Vokabulars V^{de} bzw. V^{en} , des invertierten Indexes in der Sprache von d , abgebildet. Es wird ein $|V^{de}|$ -dimensionaler bzw. $|V^{en}|$ -dimensionaler Wortvektor \mathbf{d}^{de} bzw. \mathbf{d}^{en} erstellt, dessen Einträge an den entsprechenden Stellen die Häufigkeiten von $\bar{\mathbf{d}}$ enthalten und sonst 0. Die Schlüsselwörter des Wortvektors \mathbf{d}^{de} bzw. \mathbf{d}^{en} werden mittels $tf \cdot idf$ gewichtet. Dabei werden die df -Werte verwendet, die sich aus dem invertierten Index ergeben. Beispielsweise entspricht der df -Wert für ein Schlüsselwort $f_i \in V^{de}$ der Anzahl von Tupeln in der Postliste des invertierten Indexes an der Stelle f_i (vgl. Abb. 6.6).

6.2.2 Berechnung der Konzeptvektoren

Um den i -ten Eintrag des Konzeptvektors \mathbf{d} für ein Dokument d zu berechnen wird die Ähnlichkeit zwischen der Wortvektorrepräsentation $\bar{\mathbf{d}}$ von d und dem Support-Vektor, der das Konzept k_i repräsentiert, berechnet (vgl. Abschnitt 5.2).

Der invertierte Index ermöglicht einen Zugriff auf die Support-Vektoren über die Schlüsselwörter. Ein Konzeptvektor wird iterativ nach **Algorithmus 1** berechnet.

Algorithmus 1 : Berechnung eines Konzeptvektors \mathbf{d} für ein Dokument d .

Invertierter Index ii ;

Vokabular V ;

Falls $Sprache(d) == \text{Deutsch}$, dann

$$V = V^{de};$$

$ii = \text{deutscher invertierter Index}$;

Falls $Sprache(d) == \text{Englisch}$, dann

$$V = V^{en};$$

$ii = \text{englischer invertierter Index}$;

Für alle Schlüsselwörter $f_i \in V$

$$postliste = ii.lookup(f_i);$$

Für alle Konzepte k_i des bilingualen Thesaurus

$$[\mathbf{d}]_i = [\mathbf{d}]_i + [\hat{\mathbf{d}}]_j \cdot postliste(i);$$

7 Experimentelle Auswertung der CL-ESA

In diesem Kapitel wird die CL-ESA anhand von verschiedenen Experimenten evaluiert.

Zunächst werden in Abschnitt 7.1 allgemeine Parameter, die für alle Experimente in diesem Kapitel gelten, definiert und beschrieben.

In Abschnitt 7.2 wird untersucht, ob die Ergebnisse, die durch die eigene Implementierung der ESA erzielt wurden, mit denen von [Gabrilovich und Markovitch \(2007\)](#) vergleichbar sind. Zur besseren Unterscheidung wird die eigene Implementierung im Folgenden als ESA* bezeichnet. Die Evaluierung, der sprachübergreifenden Ähnlichkeitsanalyse mittels CL-ESA, erfolgt in Abschnitt 7.3. In den Experimenten wurde die CL-ESA eingesetzt, um in einem Parallelkorpus Übersetzungen von Dokumenten zu identifizieren. Als Testkorpora dienten der Europarl-Korpus und Wikipedia. Verschiedene Experimente, die zur Bestimmung der Laufzeit der CL-ESA durchgeführt wurden, werden in Abschnitt 7.4 beschrieben. In Abschnitt 7.5 wird die Multilingualität der CL-ESA untersucht.

Eine Diskussion der Ergebnisse aller Experimente und ein Vergleich der CL-ESA mit anderen Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse erfolgen in Abschnitt 7.6.

7.1 Allgemeine Parameter

Relevanzkriterien

Die Relevanzkriterien, die in den Experimenten zur Auswahl der Wikipedia-Artikel eingesetzt wurden, sind in **Tab. 7.1** aufgeführt.

| Nr. | Relevanzkriterien |
|-----|---|
| 1 | Ein Artikel muss mehr als 100 Wörter enthalten, die keine Stoppwörter sind. |
| 2 | Bei dem Artikel handelt es sich nicht um eine „Weiterleitung“ oder eine „Begriffsklärung“. |
| 3 | Der Titel des Artikels ist keine Zahl und kein Datum. |
| 4 | Deutsche Artikel müssen einen englischen Sprachlink besitzen und englische Artikel einen deutschen. |
| 5 | Die deutschen und englischen Artikel müssen jeweils Paare bilden (siehe Abschnitt 6.1.2). |

Tab. 7.1: Die Relevanzkriterien 1 bis 4 wurden in den monolingualen Experimenten eingesetzt, in den sprachübergreifenden galt zusätzlich das Kriterium 5.

Anzahl der relevanten Artikel

Die Anzahl der Artikel, die den jeweiligen Relevanzkriterien entsprechen und für die Experimente aus den Wikipedia-Dumps extrahiert wurde, ist in **Tab. 7.2** dargestellt.

| Wikipedia-Dump | Relevante Artikel nach Kriterium 1, 2, 3, 4 | Relevante Artikel nach Kriterium 1, 2, 3, 4, 5 |
|------------------------------------|---|--|
| Deutscher XML-Dump vom 03.09.2007 | 182.278 | 109.267 |
| Englischer XML-Dump vom 02.08.2007 | 161.514 | 109.267 |

Tab. 7.2: Die Tabelle zeigt die Anzahl der Artikel aus dem deutschen und dem englischen Wikipedia-Dump, die die Relevanzkriterien 1 bis 4 erfüllen sowie die Anzahl der Artikel, die zusätzlich das Relevanzkriterium 5 erfüllen.

Dimensionalität der Konzeptraummodelle

Die Experimente wurde für verschiedene sprachübergreifende Konzeptraummodelle mit unterschiedlicher Dimensionalität durchgeführt. Die Dimension eines Konzeptraummodells entspricht der Anzahl der Artikel, mit denen es erstellt wird. Die entsprechenden Artikel wurden zufällig aus der Menge aller relevanten Artikel ausgewählt.

Für die sprachübergreifenden Experimente wurden aus der deutschen und der englischen Menge der relevanten Artikel je 500 zufällig ausgewählte Artikel nicht in die Konstruktion der Konzeptraummodelle miteinbezogen. Diese Artikel dienten in den Experimenten als Testartikel. Somit ergibt sich für die sprachübergreifenden Experimente eine Dimensionalität von 108.767.

7.2 Monolinguale Evaluierung der ESA*

Es wird untersucht, ob die eigene Implementierung ESA* vergleichbare Ergebnisse erzielt, wie die ESA (Gabrilovich und Markovitch, 2007). Hierzu wurden die Experimente von Gabrilovich und Markovitch wiederholt.

Als Testkorpus kam eine Dokumentsammlung, bestehend aus 50 englischsprachigen Nachrichtenartikeln, zum Einsatz, die von Lee et al. (2005) erstellt wurde. Die Dokumentsammlung enthält menschliche Ähnlichkeitsbewertungen für alle 1.225 möglichen Paarungen der 50 Dokumente. Die Ähnlichkeitsbewertung für ein Paar ergibt sich aus dem Mittelwert von acht bis zwölf menschlichen Beurteilungen, für die Ähnlichkeit zwischen den beiden Dokumenten des Paares.

Alle 50 Dokumente des Testkorpus wurden mittels ESA* in einem Konzeptraum repräsentiert. Innerhalb des Konzeptraums wurde die Ähnlichkeit zwischen allen Dokumentpaaren anhand der Kosinusähnlichkeit berechnet. Daraus ergaben sich 1.225 Ähnlichkeitswerte, die mittels Spearmans Rangkorrelationskoeffizient mit den menschlichen Bewertungen korreliert wurden. Die Ergebnisse sind, abhängig von der Dimensionalität des Konzeptraummodells, in **Abb. 7.1** dargestellt.

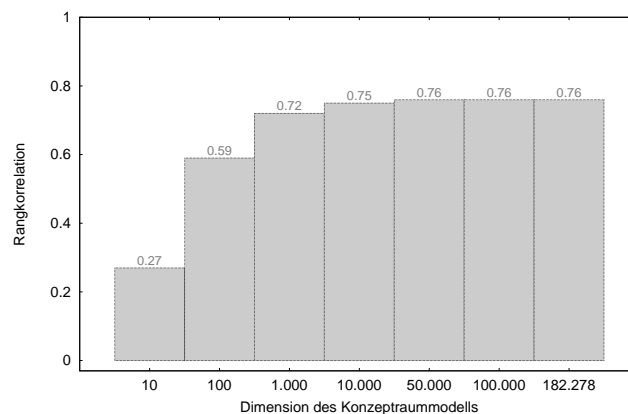


Abb. 7.1: Ergebnisse der monolingualen Ähnlichkeitsanalyse mittels ESA*. Dargestellt ist die Rangkorrelation mit menschlichen Ähnlichkeitsbewertungen.

In **Tab. 7.3** wird das beste Ergebnis der ESA* dem der ESA gegenübergestellt. Weiterhin sind in der Tabelle die Ergebnisse der Standardverfahren aus dem Bereich der monolingualen Ähnlichkeitsanalyse – des Vektorraummodell und der Latent Semantic Analysis (LSA) – dargestellt.

| Verfahren | Korrelation |
|--|-------------|
| Vektorraummodell (Lee et al., 2005) | 0,5 |
| LSA (Lee et al., 2005) | 0,6 |
| ESA (Gabrilovich und Markovitch, 2007) | 0,72 |
| ESA* (eigene Implementierung) | 0,76 |

Tab. 7.3: Dargestellt sind die Resultate des Vektorraummodells, der LSA, der ESA und der eigenen Implementierung ESA*. Alle Experimente wurden auf dem gleichen Testkorpus (Lee et al., 2005) durchgeführt. Die Ergebnisse sind Korrelationen mit menschlichen Ähnlichkeitsbewertungen.

7.3 Evaluierung der sprachübergreifenden Ähnlichkeitsanalyse mittels CL-ESA

In den, im Folgenden beschriebenen Experimenten wurde untersucht, ob sich die CL-ESA zur sprachübergreifenden Ähnlichkeitsanalyse eignet. Die sprachübergreifende Ähnlichkeitsanalyse mittels CL-ESA wurde anhand von zwei Parallelkorpora evaluiert: Wikipedia und dem Europarl-Korpus¹ (Koehn, 2005).

Die Parallelkorpora stellen Dokumente in Deutsch und Englisch bereit, die paarweise entweder Übersetzungen voneinander sind (Europarl-Korpus) oder inhaltlich dasselbe Thema beschreiben (Wikipedia). Ein solches Paar wird im Folgenden als „Übersetzungspaar“ bezeichnet und ein Dokument eines Übersetzungspaares als „Übersetzungspartner“.

Um die CL-ESA zu evaluieren, wurde in den beiden Dokumentmengen D^{de} und D^{en} , die die deutschen und englischen Dokumente der bilingualen Parallelkorpora enthalten, nach Übersetzungspaaren gesucht. Es wurde eine Kreuzvalidierung – ausgehend von der Menge D^{de} – durchgeführt, indem für ein Dokument $d \in D^{de}$ anhand der CL-ESA die Ähnlichkeiten zu allen Dokumenten $d' \in D^{en}$ berechnet wurden. Für die Ähnlichkeiten aller Paarungen wurden die Ränge bestimmt – die höchste Ähnlichkeit besitzt den Rang 1, die zweithöchste den Rang 2 usw. Die CL-ESA liefert ein perfektes Ergebnis, wenn die Ähnlichkeit des Übersetzungspaares, bestehend aus d und dem Übersetzungspartner $d' \in D^{en}$, den Rang 1 besitzt.

Die Experimente zur Evaluierung der CL-ESA anhand von Wikipedia und des Europarl-Korpus sowie deren Ergebnisse werden in den folgenden Abschnitten beschrieben. Die Experimente wurden für sprachübergreifende Konzeptraummodelle mit den Dimensionen 10, 100, 1.000, 10.000, 50.000 und 108.767 durchgeführt.

¹Es wurde die Version „v2“ des Europarl-Korpus eingesetzt: <http://www.statmt.org/euoparl>.

7.3.1 Evaluierung anhand von Wikipedia

Zur Konstruktion der Konzeptraummodelle wurden der jeweiligen Dimensionalität entsprechend viele Artikel in Deutsch und in Englisch aus der Menge der relevanten Wikipedia-Artikel ausgewählt. Aus den restlichen Artikeln wurden zufällig 500 deutsche und 500 englische Artikel bestimmt, mit denen die Kreuzvalidierung durchgeführt wurde. Die Ergebnisse sind in **Tab. 7.4** dargestellt.

Die Zeilen entsprechen den Dimensionen und beinhalten jeweils die Ergebnisse für ein Konzeptraummodell mit der entsprechenden Dimensionalität.

Die erste Spalte gibt Antwort auf die Frage, in wie viel Prozent der Fälle der Übersetzungspartner gefunden wird (Rang 1) und in wie viel Prozent der Fälle er unter den ersten n Rängen ist. Beispielsweise wurde der Übersetzungspartner bei einer Dimension von 108.767 in 82,2% der Fälle gefunden und lag in 95,6% der Fälle unter den 10 Artikeln mit dem höchsten Rang.

Die Werte in der zweiten Spalte beziehen sich auf die Fälle, in denen die Ähnlichkeit des Übersetzungspaares nicht Rang 1 besitzt. Es wird die Frage beantwortet, wie hoch in diesen Fällen die Ähnlichkeit zwischen dem Artikel mit Rang 1 und dem Übersetzungspartner ist. Bei einer Dimension von 108.767 lag beispielsweise in 19,1% der Fälle, in denen das Übersetzungspaar nicht Rang 1 besaß, die Ähnlichkeit zwischen dem Artikel mit Rang 1 und dem Übersetzungspartner in dem Intervall von 0,7 bis 0,8.

In der dritten Spalte ist die Verteilung der Ähnlichkeiten, die Rang 1 besitzen, dargestellt. Beispielsweise bei einer Dimension von 108.767, lagen 55% der Ähnlichkeiten mit Rang 1 zwischen 0,7 und 0,8.

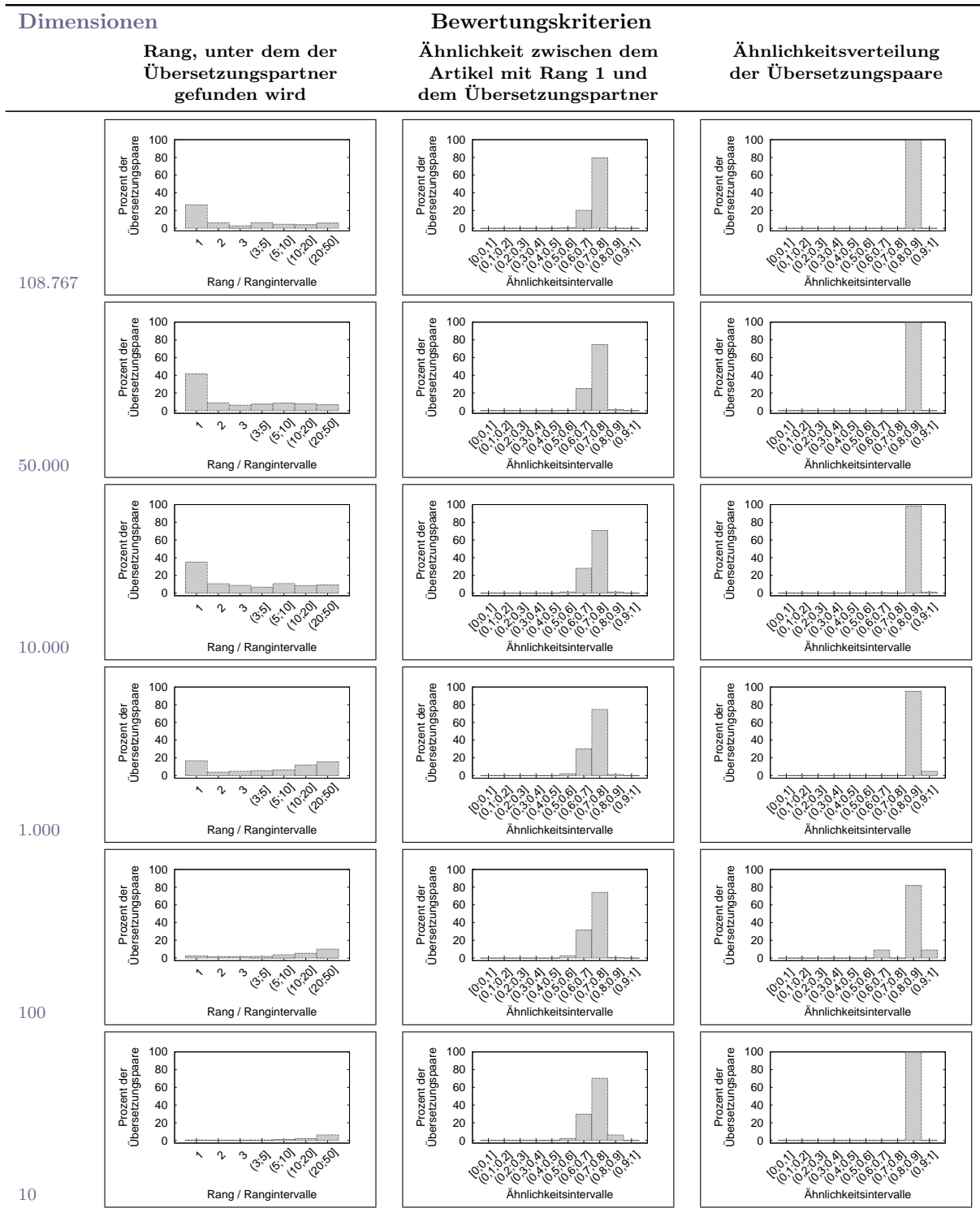
| Dimensionen | Bewertungskriterien | Ähnlichkeitsverteilung der Übersetzungspaare |
|---|---|--|
| Rang, unter dem der Übersetzungspartner gefunden wird | Ähnlichkeit zwischen dem Artikel mit Rang 1 und dem Übersetzungspartner | |
| 108.767 | | |
| 50.000 | | |
| 10.000 | | |
| 1.000 | | |
| 100 | | |
| 10 | | |

Tab. 7.4: Ergebnisse der Experimente zur Evaluierung der CL-ESA anhand von Wikipedia. Die Zeilen beinhalten jeweils die Ergebnisse für ein Konzeptraummodell mit der entsprechenden Dimensionalität.

7.3.2 Evaluierung anhand des Europarl-Korpus

Der Europarl-Korpus beinhaltet jeweils 488 deutsche und englische Dokumente. In den Experimenten wurden dieselben sprachübergreifenden Konzeptraummodelle eingesetzt, wie bei der Evaluierung mittels Wikipedia. Für die Kreuzvalidierung wurden alle Dokumente des Korpus benutzt. Die Ergebnisse sind – analog zu der Evaluierung mittels Wikipedia – in **Tab. 7.5** dargestellt.

Die große Abweichung zu den Ergebnissen, die anhand von Wikipedia erzielt wurden, ist durch die Beschaffenheit des Europarl-Korpus bedingt. In Abschnitt 7.6 wird genauer auf dieses Problem eingegangen und erläutert, warum der Europarl-Korpus für die Evaluierung von Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse ungeeignet ist.



Tab. 7.5: Ergebnisse der Experimente zur Evaluierung der CL-ESA anhand des Europarl-Korpus. Die Zeilen beinhalten jeweils die Ergebnisse für ein Konzeptraummodell mit der entsprechenden Dimensionalität.

7.4 Laufzeit der CL-ESA

Es wurde untersucht, welche Laufzeit sich für die CL-ESA ergibt und wie diese im Verhältnis zu der des klassischen Vektorraummodells steht.

Die Laufzeit der CL-ESA entspricht der Zeit, die benötigt wird, um ein Dokument durch das sprachübergreifende Konzeptraummodell zu repräsentieren. Um die Laufzeit zu bestimmen, wurden 500 zufällig ausgewählte deutsche Wikipedia-Artikel sowie 500 zufällig ausgewählte englische Wikipedia-Artikel durch das sprachübergreifende Konzeptraummodell repräsentiert. Pro Artikel wurde die dafür benötigte Zeit gemessen. Die Laufzeit ergibt sich aus dem Mittelwert aller Zeiten. Die Experimente wurden auf einem aktuellen Desktop-Rechner² durchgeführt. Die Ergebnisse sind in **Abb. 7.2**, abhängig von der Dimensionalität des sprachübergreifenden Konzeptraummodells, dargestellt.

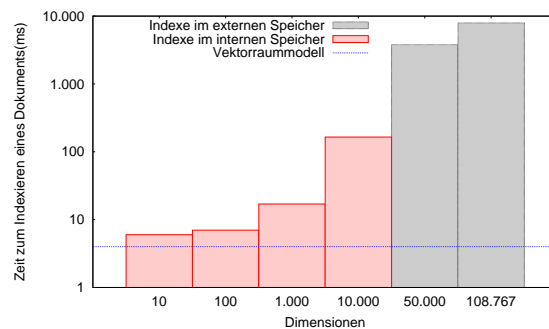


Abb. 7.2: Laufzeit der CL-ESA. Für jede Dimension ist die Zeit in Millisekunden dargestellt, die benötigt wird, um ein Dokument durch das jeweilige sprachübergreifende Konzeptraummodell zu repräsentieren. Es wird unterschieden, ob die invertierten Indexe aufgrund ihrer Größe im Hauptspeicher gehalten werden können oder in den externen Speicher ausgelagert werden müssen. Zum Vergleich ist die Zeit dargestellt, die mit dem Vektorraummodell benötigt wird.

Bis zu einer Dimension von 10.000 konnten die invertierten Indexe im Hauptspeicher gehalten werden. Bei einer Dimension von 50.000 und 108.767 wurden die Indexe in den externen Speicher ausgelagert.

7.5 Multilingualität der CL-ESA

Es soll die Frage beantwortet werden, wie viele Sprachen durch die CL-ESA unterstützt werden können. Dazu wird die Multilingualität von Wikipedia untersucht, da diese in

²CPU: Intel Core 2 Duo, 2 GHz; Arbeitsspeicher: 1024 MB.

direktem Zusammenhang zu der Multilingualität der CL-ESA steht.

Aktuell ist Wikipedia in 253 Sprachen verfügbar. Die Anzahl der Artikel ist in den verschiedenen Sprachversionen jedoch sehr unterschiedlich, siehe **Abb. 7.3**. Zur Konstruktion eines sprachübergreifenden Konzeptraummodells, das die Menge der Sprachen \mathcal{L} unterstützt, sind nur die Wikipedia-Konzepte relevant, die durch einen Artikel in jeder Sprache aus \mathcal{L} beschrieben werden (vgl. Abschnitt 5.3). In **Abb. 7.4** ist für die 15 größten Wikipedia-Sprachversionen die Anzahl der relevanten Konzepte dargestellt.

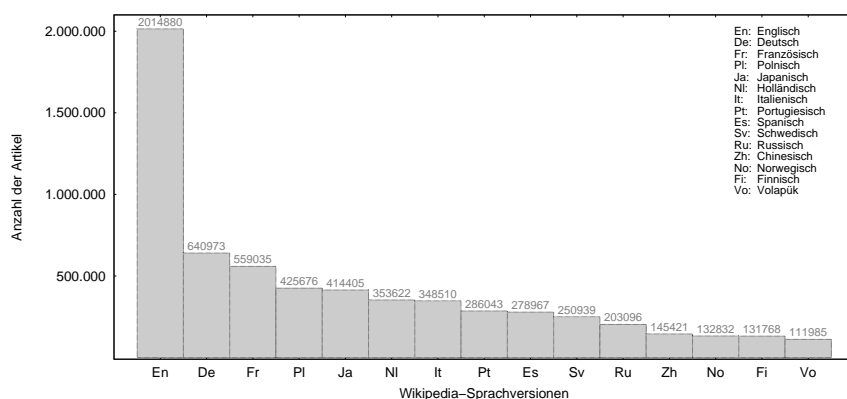


Abb. 7.3: Dargestellt ist die Anzahl der Artikel für die 15 größten Sprachversionen von Wikipedia. (http://meta.wikimedia.org/wiki/Complete_list_of_language_Wikipedias_available, Zahlen von September 2007.)

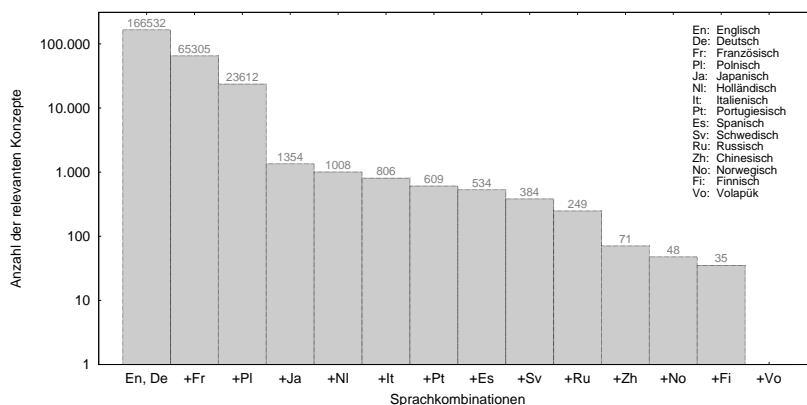


Abb. 7.4: Für verschiedene Sprachkombinationen ist die Anzahl der Wikipedia-Konzepte dargestellt, die durch einen Artikel in jeder der Sprachen beschrieben werden. Ausgehend von den Sprachversionen Englisch und Deutsch werden sukzessive weitere hinzugefügt. Die Reihenfolge entspricht der Anzahl der Artikel, siehe Abb. 7.3.

7.6 Diskussion

Im folgenden Abschnitt 7.6.1 erfolgt die Bewertung und Diskussion der zuvor präsentierten Ergebnisse. In Abschnitt 7.6.2 wird die CL-ESA anhand von bestimmten Eigenschaften mit anderen Verfahren aus dem Bereich der sprachübergreifenden Ähnlichkeitsanalyse verglichen.

7.6.1 Schlussfolgerungen

Dimensionalität und Laufzeit

Mit steigender Dimensionalität des sprachübergreifenden Konzeptraummodells nimmt die Genauigkeit sowohl bei der monolingualen (siehe Abb. 7.1), als auch bei der sprachübergreifenden Ähnlichkeitsanalyse (siehe Tab. 7.4) zu. Dies lässt sich dadurch erklären, dass durch eine größere Anzahl von Dimensionen bzw. Konzepten die Inhalte der Dokumente genauer repräsentiert werden können. Ab einer Dimensionalität von 10.000 werden die Resultate nur geringfügig besser. Daraus lässt sich schließen, dass bereits 10.000 Konzepte ausreichend sind, um Dokumente für den Zweck der Ähnlichkeitsanalyse zu repräsentieren.

Auch die Laufzeit der CL-ESA nimmt mit steigender Dimensionalität des sprachübergreifenden Konzeptraummodells zu (siehe Abb. 7.2). Für die Dimensionen, bei denen die invertierten Indexe in den Hauptspeicher passen, liegt die Laufzeit nur minimal über der des Vektorraummodells.

Einen guten Kompromiss zwischen Laufzeitkomplexität und Genauigkeit stellt eine Dimensionalität von 10.000 dar. Die durchschnittliche Zeit, die hierbei zur Repräsentation eines Dokuments benötigt wird, beträgt 165 Millisekunden und die Ergebnisse bei der Ähnlichkeitsanalyse liegen nur geringfügig unter denen, die bei der maximal möglichen Dimensionalität erreicht werden.

Monolinguale Evaluierung der ESA*

Der Vergleich mit den menschlichen Bewertungen (siehe Tab. 7.3) zeigt, dass die ESA* (mit einer Korrelation von 0,76) bei der monolingualen Ähnlichkeitsanalyse sowohl den

Standardverfahren aus diesem Bereich – dem Vektorraummodell (0,5) und der LSA (0,6) – als auch der ESA (0,72) überlegen ist.

Die Verbesserung im Vergleich zur ESA kann durch den Einsatz des sprachlinkbasierten Relevanzkriteriums (Kriterium 4 in Abschnitt 7.1) erklärt werden. Dadurch ist es möglich, aussagekräftige Konzepte zu gewinnen.

Wikipedia vs. Europarl

Die Ergebnisse der Experimente zur sprachübergreifenden Ähnlichkeitsanalyse mittels CL-ESA (Tab. 7.4 und Tab. 7.5) zeigen, dass sich die Resultate zwischen der Evaluierung mittels Wikipedia und der Evaluierung anhand des Europarl-Korpus stark unterscheiden. Während bei der Evaluierung mittels Wikipedia vielversprechende Ergebnisse erzielt werden, sind die Ergebnisse bei der Evaluierung anhand des Europarl-Korpus unbefriedigend. Dies ist jedoch durch die Beschaffenheit des Europarl-Korpus bedingt.

Die Dokumente des Europarl-Korpus besitzen untereinander eine sehr hohe Ähnlichkeit. Dies zeigt u. a. die Ähnlichkeitsverteilung der Übersetzungspaare in Tab. 7.5, denn so gut wie alle Ähnlichkeitswerte liegen in dem Intervall $(0,8; 0,9]$. In **Abb. 7.5** ist die Ähnlichkeitsverteilung für alle Dokumente des Europarl-Korpus dargestellt. Auch hier zeigt sich, dass zwischen den Dokumenten eine hohe Ähnlichkeit besteht. Diese Tatsache lässt den Schluss zu, dass der Europarl-Korpus für die Evaluierung von Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse ungeeignet ist, da sich darin die Ähnlichkeiten der Übersetzungspaare kaum von den Ähnlichkeiten jedes beliebigen Dokumentpaares unterscheiden.

Wikipedia-Artikel hingegen besitzen untereinander eine sehr geringe Ähnlichkeit, denn jeder Artikel beschreibt ein Konzept, das in der gesamten Enzyklopädie nur einmal enthalten ist. In **Abb. 7.6** ist die Ähnlichkeitsverteilung für die Artikel aus den Wikipedia-Versionen, die in den Experimenten verwendet werden, dargestellt. Wikipedia ist somit für die Evaluierung von Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse geeignet.

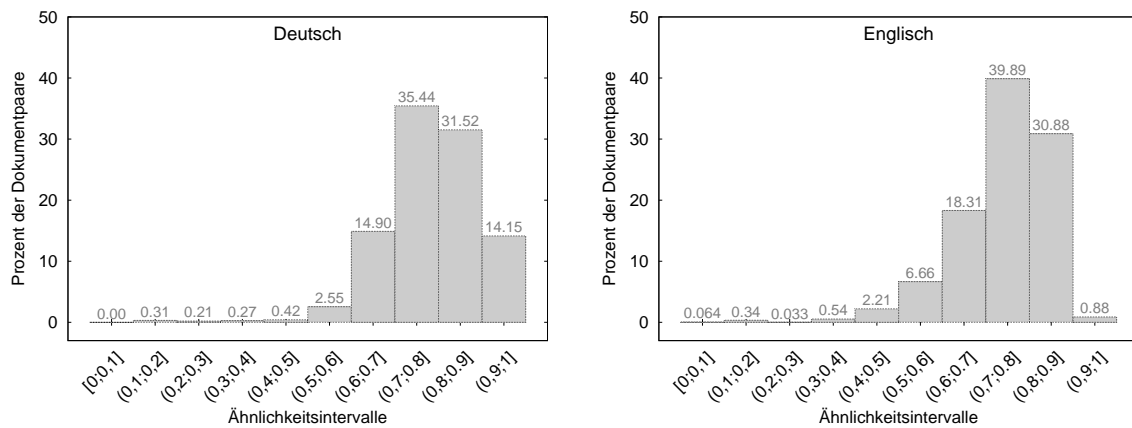


Abb. 7.5: Ähnlichkeitsverteilung für die Dokumente des Europarl-Korpus. Dargestellt sind die Ähnlichkeiten zwischen allen möglichen Dokumentpaaren aus der Menge der deutschen Dokumente (links) und der Menge der englischen Dokumente (rechts). Die Ähnlichkeiten werden anhand des Vektorraummodells in Kombination mit der Kosinusähnlichkeit bestimmt.

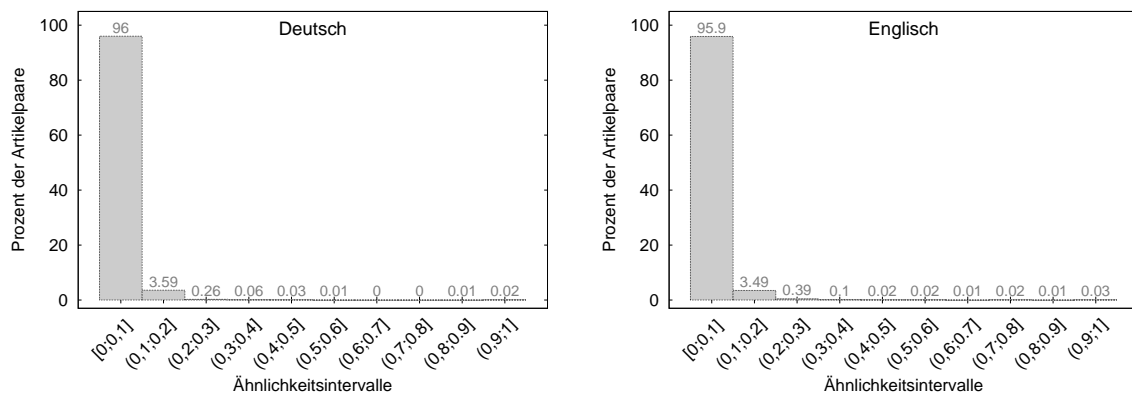


Abb. 7.6: Ähnlichkeitsverteilung für die Artikel in Wikipedia. Dargestellt sind die Ähnlichkeiten zwischen allen möglichen Artikelpaaren aus einer Testmenge, bestehend aus 10.000 zufällig ausgewählten deutschen Artikeln (links), und einer Testmenge, bestehend aus 10.000 zufällig ausgewählten englischen Artikeln (rechts). Die Ähnlichkeiten werden anhand des Vektorraummodells in Kombination mit der Kosinusähnlichkeit bestimmt.

Sprachübergreifende Ähnlichkeitsanalyse

Mittels CL-ESA kann in 82,2% der Testfälle der Übersetzungspartner eines deutschen Wikipedia-Artikels in einer Menge von englischen Artikel identifiziert werde. Dieser Wert wird durch ein sprachübergreifendes Konzeptraummodell mit der Dimensionalität 108.767 erzielt. Bei einer Dimensionalität von 10.000 werden immer noch in 78,2% der Testfälle die Übersetzungspartner identifiziert.

In über 60 % der Fälle, in denen das Übersetzungspaar nicht gefunden wird, besteht eine hohe Ähnlichkeit zwischen dem Artikel mit Rang 1 und dem Übersetzungspartner. Das heißt, dass in diesen Fällen ein Artikel gefunden wird, der mit großer Wahrscheinlichkeit ein inhaltlich ähnliches Konzept beschreibt.

Zur sprachübergreifenden Plagiaterkennung ist es zudem sinnvoll, nicht ausschließlich den Artikel mit Rang 1 zu betrachten. Bei einer Dimensionalität von 108.767 ist der Übersetzungspartner in 93 % der Fälle unter den ersten fünf Artikeln, mit der höchsten Ähnlichkeit, und in 96 % der Fälle unter den besten zehn. Bei einer Dimensionalität von 10.000 ist der Übersetzungspartner in 91,4 % der Fälle unter den ersten fünf Artikeln und in 95 % der Fälle unter den besten zehn. Dies zeigt, dass die CL-ESA ein vielversprechendes Verfahren ist, um Plagiate sprachübergreifend zu identifizieren.

Multilingualität

Die Multilingualität der CL-ESA entspricht der Multilingualität von Wikipedia. Je mehr Sprachen von der CL-ESA unterstützt werden, umso kleiner wird die maximal mögliche Dimensionalität des sprachübergreifenden Konzeptraummodells, da die Anzahl der relevanten Wikipedia-Konzepte – diejenigen, die durch je einen Artikel in jeder der Sprachen beschrieben werden – mit steigender Anzahl der Sprachen abnimmt (siehe Abb. 7.4). Um akzeptable Ergebnisse bei der sprachübergreifenden Ähnlichkeitsanalyse zu erzielen, wird mindestens eine Dimensionalität von 10.000 benötigt (vgl. Abb. 7.4).

Die maximal mögliche Dimensionalität ist stark von den Sprachen, die miteinander kombiniert werden, abhängig. In Abb. 7.4 werden die Sprachen entsprechend der Größe der jeweiligen Wikipedia-Versionen kombiniert. Dabei zeigt sich, dass beispielsweise bei der Hinzunahme von Japanisch die Anzahl der relevanten Konzepte stark abnimmt. Es ist anzunehmen, dass die Anzahl der relevanten Konzepte innerhalb von Sprachräumen (z. B. germanisch, lateinisch, asiatisch), geographischen Regionen (z. B. Europa, Asien, Südamerika) oder Kulturräumen deutlich höher ist.

Weiterhin wird die Multilingualität der CL-ESA bedingt durch das enorme Wachstum von Wikipedia ständig verbessert.

7.6.2 Vergleich der CL-ESA mit anderen Verfahren

In **Tab. 7.6** werden die CL-ESA sowie die Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse, die in Abschnitt 2.3 beschrieben werden, anhand der folgenden Eigenschaften bewertet.

| Verfahren | Bewertungskriterien | | | | | |
|-----------------|----------------------------------|----------|---------------------------------|-----------------------------|------------------------|--------------------------|
| | Multilingualität (# Sprachen) | Laufzeit | Skalierbarkeit (# Dokumente) | Verfügbarkeit Ressourcen | Retrieval- Qualität | Domänenab- hängigkeit |
| CL-LSI | 3 | hoch | 10 ⁴ | sehr schlecht | sehr gut | total |
| CL-KCCA | 2 | hoch | 10 ⁴ | schlecht | sehr gut | total |
| CL-ESA | 14 | mittel | Web | gut | gut | gering |
| CL-VSM | 2 | gering | Web | gut | schlecht | keine |
| Eurovoc-basiert | 21 | mittel | Web | schlecht | mittel | mittel |

Tab. 7.6: Bewertung der Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse. Oben sind die parallelkorpusbasierten Verfahren dargestellt, unten die wörterbuch- bzw. thesaurusbasierten.

Multilingualität. Anzahl der Sprachen, die von dem Verfahren unterstützt werden. Die Anzahl ist bei allen Verfahren von der zu Grunde liegenden Wissensbasis abhängig. Dasselbe gilt für die Sprachen die unterstützt werden.

Laufzeit. Zeit, die benötigt wird, um das Dokumentmodell zu erstellen.

Skalierbarkeit. Maximale Anzahl der Dokumente, die durch das Dokumentmodell repräsentiert werden können.

Verfügbarkeit der Ressourcen. Verfügbarkeit der zu Grunde liegenden Wissensbasen.

Retrieval-Qualität. Genauigkeit, die die Verfahren bei der sprachübergreifenden Ähnlichkeitsanalyse erzielen.

Domänenabhängigkeit. Abhängigkeit von einem bestimmten Themenbereich, d. h. es können nur die Dokumente gut repräsentiert werden, deren Inhalt sich mit dem jeweiligen Themenbereich befasst. Die Domänenabhängigkeit ergibt sich aus der zu Grunde liegenden Wissensbasis.

Wie die Tabelle zeigt, füllt die CL-ESA die Lücke zwischen dem CL-LSI sowie der CL-KCCA auf der einen Seite und dem CL-VSM sowie dem Eurovoc-basiert Ansatz auf der anderen Seite. Erstere erzielen eine sehr gute Retrieval-Qualität, allerdings sind so gut wie keine entsprechenden und umfangreichen Parallelkorpora verfügbar, so dass nur wenige Sprachen unterstützt werden und eine sehr hohe Domänenabhängigkeit besteht. Weiterhin sind die Verfahren, aufgrund des hohen Rechenaufwands, nur für kleine Dokumentsammlungen praktikabel. Das CL-VSM und der Eurovoc-basiert Ansatz hingegen besitzen eine hohe Skalierbarkeit sowie eine akzeptable Laufzeit und sind kaum

domänenabhängig. Allerdings kann durch diese Verfahren keine hohe Retrieval-Qualität erreicht werden.

Wie in den Experimenten gezeigt wurde, erzielt die CL-ESA vielversprechende Retrieval-Ergebnisse und besitzt gleichzeitig eine gute Laufzeit. Mit Wikipedia steht außerdem eine sehr umfangreiche, multilinguale Wissensbasis zur Verfügung, so dass die CL-ESA eine sehr geringe Domänenabhängigkeit besitzt und eine hohe Multilingualität.

8 Zusammenfassung und Ausblick

Gegenstand dieser Arbeit ist die Erforschung und Entwicklung von Methoden zur sprachübergreifenden Plagiaterkennung.

Bisher ist keine Arbeit bekannt, in der eine Lösung für das Problem der sprachübergreifenden Plagiaterkennung vorgeschlagen wurde. In diesem Bereich wurde bisher nur wenig Forschung betrieben, die sich zudem ausschließlich auf Teilprobleme beschränkt. In dieser Arbeit wird das Retrieval-Problem „sprachübergreifende Plagiaterkennung“ erstmals als Ganzes betrachtet. Es werden verschiedene Verfahren zur Lösung vorgeschlagen.

Das Ziel der sprachübergreifenden Plagiaterkennung ist es, ein Plagiat zu entlarven, indem das Originaldokument, aus dem plagiiert wurde, in einer Dokumentsammlung wiedergefunden wird. Es werden zwei grundsätzliche Teilaufgaben unterschieden: 1. „heuristisches Retrieval“, dabei werden in einer Dokumentsammlung Kandidaten für Originaldokumente identifiziert und 2. „detaillierte Analyse“, hierbei wird die inhaltliche Ähnlichkeit zwischen den Kandidatendokumenten und einem verdächtigen Dokument bestimmt. Falls eine hohe Ähnlichkeit festgestellt wird, wurde mit großer Wahrscheinlichkeit ein Plagiat gefunden.

Für beide Teilaufgaben werden entsprechende Verfahren zur Lösung vorgestellt. Der Schwerpunkt dieser Arbeit liegt auf der detaillierten Analyse. Hierzu muss die inhaltliche Ähnlichkeit zwischen Dokumenten, die in unterschiedlichen Sprachen vorliegen, bestimmt werden. Diese Problemstellung wird als „sprachübergreifende Ähnlichkeitsanalyse“ bezeichnet.

Zur sprachübergreifenden Ähnlichkeitsanalyse existieren bisher nur wenige Lösungsansätze. Es wird eine Übersicht über alle bekannten Verfahren gegeben und eine Bewertung bzw. ein Vergleich, aufgrund verschiedener Eigenschaften, vorgenommen.

In dieser Arbeit wird ein neues Verfahren zur sprachübergreifenden Ähnlichkeitsanalyse vorgestellt – die Cross-Language Explicit Semantic Analysis (CL-ESA). Der CL-ESA liegt ein sprachübergreifendes Konzeptraummodell zu Grunde, das es ermöglicht, den Inhalt von verschiedensprachigen Dokumenten, auf der Basis einer enormen Menge von externem Wissen, zu repräsentieren und zu vergleichen. Die CL-ESA ist eine Generalisierung der Explicit Semantic Analysis (ESA) (Gabrilovich und Markovitch, 2007) und verwendet als Wissensbasis die Online-Enzyklopädie Wikipedia.

Die CL-ESA nutzt die Multilingualität von Wikipedia, um unter Verwendung der ESA ein Dokument, in einer beliebigen Sprache L , in einem Vektorraum zu repräsentieren, der mit dem Vektorraum eines Dokuments, in einer anderen Sprache L' , das ebenfalls durch die ESA repräsentiert wurde, vergleichbar ist. Die inhaltliche Ähnlichkeit zwischen den Dokumenten wird anhand der Kosinusähnlichkeit bestimmt.

Im Rahmen dieser Arbeit wurde ein Softwaresystem entwickelt, das die CL-ESA für die Sprachen Deutsch und Englisch realisiert. Anhand dessen wurden umfangreiche Experimente zur Evaluierung der CL-ESA durchgeführt.

Die Ergebnisse von Gabrilovich und Markovitch (2007) konnten durch die eigene Implementierung der ESA bestätigt werden. Dabei wurde zudem eine leichte Verbesserung der Ergebnisse erzielt. Der Grund dafür liegt darin, dass ein neues Kriterium zur Auswahl von Wikipedia-Artikeln verwendet wurde, das auf Sprachlinks in den Artikeln basiert.

In den Experimenten wurde die CL-ESA zur sprachübergreifenden Ähnlichkeitsanalyse eingesetzt, indem in einem bilingualen Parallelkorpus für ein Dokument d in einer Sprache L die Übersetzungspartner d' in der Sprache L' gesucht wird. Dabei wird ausgehend von d anhand der CL-ESA die inhaltliche Ähnlichkeit zu allen Dokumenten des Parallelkorpus in der Sprache L' berechnet. Falls für d' die höchste Ähnlichkeit bestimmt wird, wurde das Übersetzungspaar gefunden. Als Testkorpus kamen der Europarl-Korpus und Wikipedia zum Einsatz.

In den Experimenten wurde gezeigt, dass die Ergebnisse stark von der Dimensionalität – d. h. der Anzahl der Konzepte – des Konzeptraummodells, das der CL-ESA zu Grunde liegt, abhängig sind. Mit steigender Anzahl der Dimensionen nimmt die Retrieval-Qualität zu. Bei einer Dimensionalität von 108.767 wurde der Übersetzungspartner eines Dokuments in 82,2% der Testfälle identifiziert. In 96% der Fälle lag der Übersetzungspartner unter den zehn Dokumenten mit der höchsten Ähnlichkeit. Weiterhin wird für die Fälle, in denen der Übersetzungspartner nicht identifiziert wurde,

gezeigt, dass das Dokument, für das die CL-ESA die höchste Ähnlichkeit bestimmt hat, eine hohe Ähnlichkeit zu dem Übersetzungspartner besitzt und daher mit großer Wahrscheinlichkeit ein inhaltlich ähnliches Konzept beschreibt.

Des Weiteren wurde gezeigt, dass die Laufzeit der CL-ESA sehr gering ist und mit steigender Dimensionalität des Konzeptraummodells zunimmt. Bei kleinen Dimensionen (bis 1.000) ist die Zeit, die benötigt wird, um ein Dokument durch das Konzeptraummodell zu repräsentieren, mit der Zeit, die beim klassischen Vektorraummodell benötigt wird, vergleichbar.

Die Anzahl der Sprachen, die von der CL-ESA unterstützt werden, ist von der Multilingualität von Wikipedia abhängig. Mit zunehmender Anzahl der Sprachen nimmt die maximal mögliche Dimensionalität des Konzeptraummodells ab. Die höchste Dimensionalität kann bei den Sprachen Deutsch und Englisch erreicht werden (109.267). Zudem wurde gezeigt, dass die Sprachen der 14 größten Wikipedia-Versionen unterstützt werden können.

Die Experimente liefern die Rahmenbedingungen, um die CL-ESA an die Anforderungen verschiedenster sprachübergreifender Retrieval-Aufgaben anzupassen. Falls beispielsweise eine hohe Retrieval-Qualität benötigt wird, können Dokumente mit der maximalen Dimensionalität repräsentiert werden, allerdings werden in diesem Fall nur zwei Sprachen unterstützt. Wenn die Anforderungen eine schnelle Laufzeit und eine hohe Multilingualität sind, dann wird eine niedrige Dimensionalität, z. B. 100, gewählt, daraus folgt jedoch eine geringere Retrieval-Qualität. Einen guten Kompromiss, zwischen Retrieval-Qualität und Laufzeit, stellt eine Dimensionalität von 10.000 dar. Dabei werden immer noch 78,2 % der Übersetzungspartner identifiziert und die Zeit zur Repräsentation eines Dokuments beträgt 165 Millisekunden.

Weitere Forschungsarbeit sollte investiert werden, um die Multilingualität von Wikipedia genauer zu untersuchen. Es ist anzunehmen, dass die Anzahl der Konzepte, die in einer bestimmten Kombination von Wikipedia-Sprachversionen durch jeweils einen Artikel in jeder Sprache beschrieben werden, höher ist, wenn ausschließlich Wikipedia-Versionen aus bestimmten Sprachräumen (z. B. germanisch, lateinisch, asiatisch), geographischen Regionen (z. B. Europa, Asien, Südamerika) oder Kulturräumen kombiniert werden. Wenn dies der Fall ist, dann ist es möglich, die entsprechenden Sprachen durch Konzeptraummodelle mit einer hohen Dimensionalität zu unterstützen. Weiterhin müssen im Fall der Plagiaterkennung identifizierte Plagiate genauer untersucht werden, um z. B. Plagiate von korrekten Zitaten zu unterscheiden. Eine offene Frage ist außerdem, wie sich

die CL-ESA bei anderen Sprachkombinationen, außer Deutsch und Englisch, verhält. Dies sollte in weiteren Experimenten untersucht werden.

Literaturverzeichnis

- Brin, Sergey, Davis, James und Garcia-Molina, Hector: Copy detection mechanisms for digital documents. In: *SIGMOD '95*. ACM Press, 1995, S. 398–409. ISBN 0-89791-731-6. 3.2.2
- Brown, Peter F., Cocke, John, Pietra, Stephen Della, Pietra, Vincent J. Della, Jelinek, Frederick, Lafferty, John D., Mercer, Robert L. und Roossin, Paul S.: A statistical approach to machine translation. In: *Computational Linguistics*, Band 16(2):S. 79–85, 1990. 1
- Clough, Paul: Old and new challenges in automatic plagiarism detection. National Plagiarism Advisory Service; <http://ir.shef.ac.uk/cloughie/index.html> [Zugriff: Juli], 2003. 4
- Culwin, Fintan und Lancaster, Thomas: A review of electronic services for plagiarism detection in student submissions. In: *Proceedings of the 1st LTSN I&CS Annual Conference*. Heriot-Watt University Edinburgh, 2000, S. 54–61. 1
- Finkel, Raphael A., Zaslavsky, Arkady, Monostori, Krisztian und Schmidt, Heinz: Signature extraction for overlap detection in documents. In: Oudshoorn, Michael J. (Hg.) *Proceedings of the 25th Australian conference on Computer Science (ACSC2002)*. ACS, Melbourne, Australia, 2002, Band 4, S. 59–64. 3.2.2
- Gabrilovich, Evgeniy und Markovitch, Shaul: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI), Hyderabad, Indien*. 2007, S. 1606–1612. (document), 2.3, 5, 5.1.2, 5.1, 7, 7.2, 7.2, 8
- Heintze, Nevin: Scalable document fingerprinting. In: *Proceedings of the Second USENIX Workshop on Electronic Commerce, Oakland, California*. 1996. 3.2.2
- Hoad, Timothy C. und Zobel, Justin: Methods for identifying versioned and plagiarized documents. In: *J. Am. Soc. Inf. Sci. Technol.*, Band 54(3):S. 203–215, 2003. ISSN 1532-2882. 3.2.2

- Hull, David A. und Grefenstette, Gregory: Querying across languages: a dictionary-based approach to multilingual information retrieval. In: *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. Zurich, Switzerland. 1996, S. 49–57. 2.2
- Hutchins, John: Current commercial machine translation systems and computer-based translation tools: system types and their uses. In: *International Journal of Translation*, Band 17(1–2), 2005. 4.2.2, 1
- Hutchins, John: Compendium of translation software: Directory of commercial machine translation systems and computer-aided translation support tools, 2007. URL: <http://www.hutchinsweb.me.uk/Compendium.htm> [Zugriff: Juli 2007] (Bisher wurde keine gedruckte Version veröffentlicht.). 4.2.2
- Karypis, George und Han, Eui-Hong: Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval and categorization. Technical report tr-00-0016, University of Minnesota, 2000. 5.1.1, 2
- Kishida, Kazuaki und Kando, Noriko: A hybrid approach to query and document translation using a pivot language for cross-language information retrieval. In: Peters, Carol et. al. (Hg.) *CLEF '05, Lecture Notes in Computer Science*. Springer, 2005, Band 4022, S. 93–101. ISBN 3-540-45697-X. 4.2.2
- Koehn, Philipp: Europarl a multilingual corpus for evaluation of machine translation. 2005. 5.2.3, 7.3
- Lee, Michael D., Pincombe, B. und Welsh, M.: An empirical evaluation of models of text document similarity. In: Bara, B. G., Barsalou, L. W. und Bucciarelli, M. (Hg.) *Proceedings of the 27th Annual Meeting of the Cognitive Science Society, CogSci2005*. 2005, S. 1254–1259. 7.2, 7.2, 7.3
- Levow, Gina-Anne, Oard, Douglas W. und Resnik, Philip: Dictionary-based techniques for cross-language information retrieval. In: *Inf. Process. Manage*, Band 41(3):S. 523–547, 2005. 2.3, 4.2.1
- Littman, Michael L., Dumais, Susan T. und Landauer, Thomas K.: Automatic cross-language information retrieval using latent semantic indexing. In: G. Grefenstette (Hg.) *Cross Language Information Retrieval*. Kluwer, 1998. 2.3
- Matsuo, Y. und Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. In: , Band 13(1):S. 157–169, 2004. 4.2.1
- McNamee, Paul und Mayfield, James: Comparing cross-language query expansion techniques by degrading translation resources. In: *SIGIR '02: Proceedings of the 25th annual*

- international ACM SIGIR conference on Research and development in information retrieval*. ACM Press, New York, NY, USA, 2002, S. 159–166. ISBN 1-58113-561-0. 4.2.1
- Meyer zu Eißén, Sven und Stein, Benno: Intrinsic plagiarism detection. In: M. Lalmas and A. MacFarlane and S. Rüger and A. Tombros and T. Tsikrika and A. Yavlinisky (Hg.) *Advances in Information Retrieval: Proceedings of the 28th European Conference on IR Research, ECIR 2006, London, 3936 of Lecture Notes in Computer Science*. Springer, 2006, S. 565–2006. ISBN 3-540-33347-9. 3.1, 3.2, 4
- Meyer zu Eißén, Sven, Stein, Benno und Kulig, Marion: Plagiarism detection without reference collections. In: Decker, Reinhold und Lenz, Hans J. (Hg.) *Advances in Data Analysis*. Springer, 2007, S. 359–366. ISBN 978-3-540-70980-0. 3.2.4
- Meyer zu Eißén, Sven, Stein, Benno und Potthast, Martin: The suffix tree document model revisited. In: Klaus Tochtermann und Hermann Maurer (Hg.) *Proceedings of the 5th International Conference on Knowledge Management (I-KNOW 05)*. Journal of Universal Computer Science, Graz, Österreich, 2005, S. 596–603. 2.2
- NIST, National Institute of Standards and Technology: Nist 2006 machine translation evaluation official results. http://www.nist.gov/speech/tests/mt/doc/mt06eval_official_results.html [Zugriff: Juli 2007], 2006. 4.2.2
- Oard, Douglas W. und Dorr, Bonnie J.: A survey of multilingual text retrieval. Technischer Bericht, College Park, MD, USA, 1996. 2.2
- Orengo, Viviane Moreira und Huyck, Christian R.: Relevance feedback and cross-language information retrieval. In: *Inf. Process. Manage*, Band 42(5):S. 1203–1217, 2006. 4.2.1
- Porter, Martin F.: An algorithm for suffix stripping. In: *Program*, Band 14(3):S. 130–137, 1980. 2.1
- Potthast, Martin: Wikipedia in the pocket - indexing technology for near-duplicate detection and high similarity search. In: Clarke, Charles, Fuhr, Norbert, Kando, Noriko, Kraaij, Wessel und de Vries, Arjen (Hg.) *30th Annual International ACM SIGIR Conference*. ACM, 2007, S. 909–909. ISBN 987-1-59593-597-7. 6.1.3
- Potthast, Martin, Stein, Benno und Anderka, Maik: A wikipedia-based multilingual document model for cross-language plagiarism detection. 2007. In Vorbereitung. 2.4
- Pouliquen, Bruno, Steinberger, Ralf und Ignat, Camelia: Automatic identification of document translations in large multilingual document collections. In: *Proceedings of the*

- International Conference Recent Advances in Natural Language Processing (RANLP '03)*. Borovets, Bulgaria, 2003, S. 401–408. 1, 2.3, 4
- Rivest, Ronald L.: The md5 message-digest algorithm. <http://people.csail.mit.edu/rivest/Rfc1321.txt> [Zugriff: Juli 2007], 1992. 3.2.2
- Salton, G. und McGill, M. J.: *Introduction to modern information retrieval*. McGraw-Hill, 1983. ISBN 0070544840. 2.1
- Salton, G., Wong, A. und Yang, C. S.: A vector space model for automatic indexing. In: *Commun. ACM*, Band 18(11):S. 613–620, 1975. 2.1
- Shivakumar, Narayanan und Garcia-Molina, Hector: Scam: A copy detection mechanism for digital documents. In: *Proceedings of 2nd International Conference in Theory and Practice of Digital Libraries (DL '95)*, Austin, Texas. 1995. 3.2.2
- Stein, Benno: Fuzzy-fingerprints for text-based information retrieval. In: Klaus Tochtermann und Hermann Maurer (Hg.) *Proceedings of the 5th International Conference on Knowledge Management (I-KNOW 2005)*. Journal of Universal Computer Science, Graz, Österreich, 2005, S. 572–579. 3.2.3
- Stein, Benno: Principles of hash-based text retrieval. In: Charles Clarke, Noriko Kando Wessel Kraaij Arjen de Vries, Norbert Fuhr (Hg.) *30th Annual International ACM SIGIR Conference*. ACM, 2007, S. 527–534. ISBN 987-1-59593-597-7. 2
- Stein, Benno und Meyer zu Eißén, Sven: Intrinsic Plagiarism Analysis with Meta Learning. In: Stein, Benno, Koppel, Moshe und Stamatatos, Efstathios (Hg.) *SIGIR Workshop Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07)*. CEUR-WS.org, 2007. 3.2.4
- Stein, Benno und Meyer zu Eißén, Sven: Near Similarity Search and Plagiarism Analysis. In: Spiliopoulou, M., Kruse, R., Borgelt, C., Nürnberger, A. und Gaul, W. (Hg.) *From Data and Information Analysis to Knowledge Engineering*. Springer, 2006, S. 430–437. ISBN 1431-8814. 3.2.3
- Stein, Benno, Meyer zu Eissen, Sven und Potthast, Martin: Syntax versus semantics: Analysis of enriched vector space models. In: Stein, Benno und Kao, Odej (Hg.) *Third International Workshop on Text-Based Information Retrieval (TIR 06)*. University of Trento, Italy, 2006, S. 47–52. ISSN 1613-0073. 2.1
- Stein, Benno, Meyer zu Eissen, Sven und Potthast, Martin: Strategies for retrieving plagiarized documents. In: Charles Clarke, Noriko Kando Wessel Kraaij Arjen de Vries, Norbert Fuhr (Hg.) *30th Annual International ACM SIGIR Conference*. ACM, 2007, S. 825–826. ISBN 987-1-59593-597-7. 3.1

- Stein, Benno und Potthast, Martin: Hashing-basierte indizierung: Anwendungsszenarien, theorie und methoden. In: *Proceedings of the Workshop Information Retrieval 2006 of the Special Interest Group Information Retrieval (FGIR) in conjunction with Lernen – Wissensentdeckung – Adaptivität 2006 (LWA '06)*. 2006, S. 159–166. ISSN 0941-3014. 3.2.3
- Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D. und Varga, D.: The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In: *Proceedings of 5th International Conference on Language Resources and Evaluation, LREC06*. 2006, S. 1–6. 5.2.3
- Steinberger, Ralf, Bruno, Pouliquen und Ignat, Camelia: Exploiting multilingual nomenclatures and language-independent text features as an interlingua for cross-lingual text analysis applications. In: *Proceedings of the 4th Slovenian Language Technology Conference. Information Society 2004 (IS '04), Ljubljana, Slovenia*. 2004. 2.3, 4
- Steinberger, Ralf und Pouliquen, Bruno: Cross-lingual indexing. Technischer Bericht, Final Report for the IPSC Exploratory Research Project. JRC Internal Note, 30 pages, 2003. 4
- Steinberger, Ralf, Pouliquen, Bruno und Hagman, Johan: Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. In: Gelbukh, A. (Hg.) *Computational Linguistics and Intelligent Text Processing, Third International Conference (CICLing '02), Lecture Notes in Computer Science*. Springer, 2002, Band 2276, S. 415–424. ISBN 3-540-43219-1. 5.2.2
- Vinokourov, A., Shawe-Taylor, J und Cristianini, N.: Inferring a semantic representation of text via cross-language correlation analysis. In: *NIPS-02: Advances in Neural Information Processing Systems. MIT Press*. 2003, S. 1473–1480. 2.3
- Witten, Ian H., Moffat, Alistair und Bell, Timothy C.: *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishers, San Francisco, CA, 1999. ISBN 1-55860-570-3. 4.2

Stichwortverzeichnis

- tf · idf*, 7
- Ähnlichkeitsmaß, 5
- Übersetzungspaar, 50
- Übersetzungspartner, 50
- Bilingualer Wikipedia-Thesaurus, 42
- Chunking, 14
- Cross-Language Explicit Semantic Analysis (CL-ESA), 37
- Cross-Language Information Retrieval (CLIR), 8
- Detaillierte Analyse (monolinguale Plagiaterkennung), 14
- Detaillierte Analyse (sprachübergreifende Plagiaterkennung), 22
- Dokumentmodell, 6
- Explicit Semantic Analysis (ESA), 29
- Fingerprint, 17
- Fuzzy-Fingerprinting, 17
- Geschlossenen Retrieval-Situation, 6
- Globale Dokumentanalyse, 14
- Heuristisches Retrieval (monolinguale Plagiaterkennung), 14
- Heuristisches Retrieval (sprachübergreifende Plagiaterkennung), 22
- Information Retrieval (IR), 5
- Invertierter Index, 44
- Konzeptindexierung, 29
- Konzeptraummodell (CSM), 27
- Kosinusähnlichkeit, 8
- Lokale Dokumentanalyse, 14
- Offene Retrieval-Situation, 6
- Parallelkorpus, 36
- Plagiat, 1
- Plagiierten, 1
- Relevanzkriterien für Wikipedia-Artikel, 30
- Sprachübergreifende Ähnlichkeitsanalyse, 9
- Sprachübergreifende Plagiaterkennung, 21
- Sprachübergreifendes Konzeptraummodell (CL-CSM), 31
- Stemming, 7
- Stoppwörter, 7
- Thesaurus, 34
- Vektorraummodell (VSM), 7