CIMAWA

Entwicklung und Anwendung einer textbasierten Assoziations-Berechnungsmethode

genehmigte DISSERTATION

zur Erlangung des Grades eines Doktors

der Ingenieurwissenschaften

vorgelegt von

Dipl.-Wirt. Inform. Patrick Uhr

eingereicht bei der Naturwissenschaftlich-Technischen Fakultät

der Universität Siegen

Siegen 2014

gedruckt auf alterungsbeständigem holz- und säurefreiem Papier

1. Gutachter: Prof. Dr.-Ing. Madjid Fathi
2. Gutachter: Prof. Dr.-Ing. habil. Marcin Grzegorzek

Tag der mündlichen Prüfung: 24.09.2014

Danksagung

An erster Stelle bedanke ich mich bei meiner Familie für die uneingeschränkte, stetige Unterstützung in allen Phasen meiner akademischen Laufbahn.

Auch möchte ich mich bei Professor Madjid Fathi für die Möglichkeit bedanken, am Institut für Wissensbasierte Systeme und Wissensmanagement zu promovieren.

Professor Grzegorzek danke ich für die Denkanstöße in der Endphase meiner Promotion.

Für seinen fachlichen Rat in ungezählten Diskussionsrunden danke ich ganz besonders Herrn Dr. André Klahold für seine Unterstützung und Anregungen. Nicht unerwähnt bleiben dürfen meine geschätzten Kollegen am Institut WBS und im Besonderen Frau Susanne Dienst und Herrn Fazel Ansari, die meine Arbeit mit Sach- und Fachwissen bereicherten.

Zusammenfassung

Die vorliegende Arbeit stellt die Entwicklung und Anwendung einer textbasierten Assoziations - Berechnungsmethode vor. Das Verfahren trägt den Titel CIMAWA (Concept for the Imitation of the Human Ability of Word Association) und berechnet die Stärke der Beziehungen zwischen Worten. CIMAWA orientiert sich dabei an der menschlichen Wortassoziation und versucht diese möglichst exakt nachzubilden. Basierend auf großen Textsammlungen, werden statistische Auswertungen über gemeinsames Vorkommen von Worten und deren Häufigkeit dazu verwendet, die Stärke der Assoziationen zwischen Begriffen zu berechnen. Die Ergebnisse der CIMAWA-Berechnungen werden in mehreren Fallstudien mit Assoziationstests an menschlichen Probanden verglichen. Zusätzlich wurden aus der Literatur bekannte Assoziations – Berechnungsmethoden implementiert und in ihrer Leistungsfähigkeit bewertet.

Die detaillierte Erläuterung der Berechnung und die Herleitung der Parameter werden komplettiert durch die Darstellung der konzeptuellen Unterschiede zwischen den bekannten Berechnungsverfahren und CIMAWA. Die vielseitige Anwendbarkeit und praktische Relevanz der CIMAWA-Assoziationsberechnung wird durch vier Umsetzungen aus verschiedenen Anwendungsgebieten gezeigt.

Die erste Anwendung zeigt, wie CIMAWA zur Erkennung von Multi-Themenstrukturen in Textdokumenten eingesetzt wird. Die Metaanalyse von Textdokumenten im Instandhaltungsmanagement wird zum Gegenstand der zweiten Anwendung und die kontextbasierte Bereitstellung von Texten im Produktverbesserungsprozess ist im dritten Beispiel behandelt. Ein assoziatives Suchverfahren für die Wissensbasis von Unternehmen bildet die abschließende CIMAWA-Anwendung.

Abstract

The present work discusses the development and application of a novel method for text-based word association measuring. The method is entitled as CIMAWA which stands for the ‘Concept for the Imitation of the Human Ability of Word Association‘. CIMAWA calculates the strength of the relationship between words. Taking into account the human ability of word association as an archetype, CIMAWA is aimed at simulating the existing but not necessarily discovered associations. It applies statistical analysis to detect co-occurring terms and frequencies based on huge collections of texts, and uses the outcomes for the calculation of the strength of the relation.

CIMAWA is verified in several case studies, especially in comparison with free association tests of human test subjects. In addition the literatures of association measuring are reviewed and the most common methods are implemented and compared with CIMAWA’s outcomes. A detailed explanation of the calculation and the parameters are given, as well as a demonstration of the conceptual differences between CIMAWA and other measurement methods.

The multilateral areas of application and the practical adaptability are shown in four independent software applications. The first application shows how CIMAWA is utilized to detect multi-topic structures in text documents. The second discusses the meta-analysis of text documents in maintenance management. The third presents a CIMAWA based recommender system for text documents towards improving quality of industrial goods. Finally, the fourth application is developed for associative search engine in companies.

Inhaltsverzeichnis

1 Einleitung

1.1 Motivation und Problembeschreibung

Bereits in den 50er Jahren des vergangenen Jahrhunderts wurde über die negativen Auswirkungen des Informationszuwachses diskutiert [1], [2]. Dieser unter dem Begriff des ‘Information Overload‘ bekannt gewordene Sachverhalt verschärft sich zunehmend mit Verbreitung und Akzeptanz des world wide web.
Auf der Suche nach dem Ursprung der Informationsflut stößt man auf einen im ersten Moment in diesem Zusammenhang unerwarteten Namen: Johannes Gutenberg, den Erfinder des modernen Buchdrucks. Seit Mitte des 15ten Jahrhunderts war es möglich, Texte in bisher ungekannter Geschwindigkeit und Masse maschinell herzustellen. Viel später vereinfachten weit weniger spektakuläre Errungenschaften wie das Kohlepapier oder die Erfindung der ersten Fotokopierer die Vervielfältigung existenter Informationen [3]. Diese Entwicklung gipfelt (vorerst) in den digitalen Dokumenten, die mit einem Mausklick vervielfältigt und weitergegeben werden können.
Dazu passt eine der bekanntesten und am häufigsten zitierten Formulierungen des Trend- und Zukunftsforschers John Naisbitt, welche heute, obwohl mehr als 30 Jahre alt, aktueller als jemals zuvor erscheint:

“We are drowning in information and starving for knowledge.” [4]

Frei übersetzt beschreibt die Metapher die Probleme der Informationsflut anschaulich, denn die Separierung von relevanten und irrelevanten Informationen gestaltet sich im digitalen Zeitalter zunehmend problematisch.
Durch den beschriebenen Hintergrund motiviert entwickelte sich die Idee einen kleinen Beitrag zur Bewältigung der beschriebenen Problematik zu leisten. Nicht zuletzt durch das Interesse an automatisierten Analyseverfahren großer Textbestände, an uns herangetragen durch unsere Partner aus der Industrie, wurde die Notwendigkeit von innovativen Text Mining-Lösungen offen gelegt. Konsequenterweise fiel die Entscheidung, den Forschungsschwerpunkt der vorliegenden Arbeit auf Text Mining und die automatisierte Analyse großer Textmengen zu legen.
Die initial zu beantwortende Frage ist die nach dem Ausgangspunkt einer wissenschaftlichen Arbeit, die auf der einen Seite das globale Ziel der Eingrenzung der Informationsüberflutung verfolgt, und auf der anderen Seite konkrete Anwendungsszenarien und Problembeschreibungen der Industriepartner einbeziehen soll. Um beides miteinander in Einklang zu bringen, wurde ein zweigeteilter Fokus entwickelt.
Der erste Teilbereich konzentriert sich auf eine tiefgreifende Analyse von Texten, heruntergebrochen auf die Ebene der Worte und deren Beziehung zueinander. Von Menschen verfasste Texte sind in aller Regel keine willkürlich aneinandergereihten Begriffe, sondern sind vom Autor mit einer bestimmten Intention verfasst. Davon ausgehend, dass es sich bei

Texten nicht um zufällige Wortsammlungen handelt, sind weitere Schlüsse zulässig. Man kann feststellen, dass bestimmte Worte in bestimmten Kontexten häufiger Verwendung finden als in anderen. Demnach sollte es durch eine Textanalyse möglich sein, Informationen darüber zu gewinnen, wie Worte mit anderen Worten in Beziehung stehen. Die Berechnung von Wortassoziationen aus Textsammlungen ist das Ergebnis der Textanalyse und die Basis für verschiedene in dieser Arbeit entwickelte Applikationen in weitreichenden Anwendungsfeldern.
Der zweite Teilbereich baut auf den Ergebnissen des ersten auf und zeigt, wie die Informationen über die Assoziationen von Worten für unterschiedliche praxisnahe Anwendungsszenarien verwendet werden können.
Den anwendungsbezogenen Schwerpunkt legt die Arbeit auf die Nutzung und Wiederverwendung vorhandener textgebundener Wissenspotentiale, durch automatisierte Analyseverfahren auf Basis der entwickelten Assoziationsberechnungsmethode.

1.2 Zielsetzung

Die vorliegende Arbeit setzt sich im Wesentlichen zwei Ziele. Zum einen die Entwicklung eines Verfahrens zur Berechnung von Wortassoziationen und zum anderen die Nutzbarmachung und praktische Anwendung dieser Wortbeziehungen.
Für die Assoziationsberechnung wird ein Verfahren namens CIMAWA (Concept for the Imitation of the Human Ability of Word Association) entwickelt, vorgestellt und getestet. Dieses setzt sich zum Ziel, Beziehungen zwischen Worten zu berechnen und die Stärke der Beziehung in Form eines numerischen Wertes auszudrücken. Als Vorbild für das entwickelte Verfahren dient die menschliche Wortassoziation. Jeder Mensch assoziiert Begriffe mit anderen, und diese unterbewussten Beziehungen zwischen Worten bilden die Zielgröße der CIMAWA-Assoziationsberechnung. Beispiele für besonders eingängige Wortassoziationen beim Menschen sind ‘Fisch‘ und ‘Wasser‘, ‘Universität‘ und ‘Professor‘ oder ‘gut‘ und ‘böse‘. Mit CIMAWA sollen diese Wortassoziationen auf Textbasis berechnet werden.
Als Gradmesser für die Güte der berechneten Ergebnisse dienen Assoziationstests, bei denen menschliche Probanden aufgefordert sind ihre Assoziationen zu bestimmten Begriffen anzugeben. Ergebnis dieser Befragungen sind Assoziationslisten, die angeben, welche Begriffe wie stark mit welchen anderen assoziiert sind. CIMAWA ist bestrebt, eben diese beim Menschen beobachteten Assoziationen so präzise wie möglich nachzubilden.
Wie in der vorliegenden Arbeit gezeigt wird, hängen die berechneten Assoziationen stark von der Größe und der Zusammensetzung der zugrunde liegenden Textbasis ab. Dieser Sachverhalt verhilft CIMAWA zu weitreichenden Anwendungsfeldern. Denn es stellt sich heraus, dass die berechneten Assoziationen auf einer stark spezialisierten Textsammlung zugleich sehr spezialisierte Assoziationsergebnisse liefern. Die entwickelten Applikationen auf Basis von CIMAWA zeigen den Nutzen einer solchen Assoziationsanalyse und tragen so dazu bei, aus der Menge der digital gespeicherten Dokumente wertvolle Informationen zu extrahieren.

Mit dem Ziel die berechneten Assoziationen in praktischen Anwendungsszenarien einzusetzen, enthält die vorliegende Arbeit die assoziationsbasierten Methoden, die während der letzten Jahre im Rahmen meiner Tätigkeit am Institut für Wissensbasierte Systeme entwickelt wurden.

1.3 Abgrenzung zu anderen Verfahren

Neben dem in dieser Arbeit entwickelten CIMAWA-Ansatz, sind aus der Literatur weitere mathematisch-statistische Assoziationsmaße bekannt. Der folgende Abschnitt grenzt CIMAWA von anderen Verfahren ab und arbeitet Alleinstellungsmerkmale heraus.
Die mathematisch-statistischen Assoziationsmaße zur Berechnung von Wortassoziationen basieren allesamt auf den gleichen Kenngrößen, gewonnen aus Analysen großer Textsammlungen. Zu nennen sind hier das gemeinsame Vorkommen von Worten innerhalb definierter Textfenster, sowie deren Häufigkeit in der Textsammlung. Trotz der Fundierung auf identischen Kennzahlen, gibt es grundsätzliche Unterschiede in der Konzeption der Assoziationsmaße. Bei dem Vergleich der konzeptuellen Unterschiede wird der innovative Charakter von CIMAWA deutlich.
Die meisten Verfahren zur Berechnung von Wortassoziationen definieren die Beziehung zwischen zwei Worten als symmetrisch. Das bedeutet, sie bestimmen einen einzigen numerischen Wert, um die Stärke der Wortassoziation eines Wortpaares abzubilden.
Eine andere Herangehensweise wird von den so genannten asymmetrischen Berechnungsverfahren gewählt. Diese berechnen pro Wortpaar zwei numerische Werte, jeweils einen für jede Assoziationsrichtung. Diese Konzeption verspricht eine exaktere Darstellung insbesondere derer Wortassoziationen, bei denen es signifikante Unterschiede zwischen den Assoziationsrichtungen gibt.
Das Alleinstellungsmerkmal von CIMAWA und somit der Unterschied zu symmetrischen und asymmetrischen Verfahren, besteht in der Einbeziehung beider Assoziationsrichtungen in einen Assoziationswert. Aufgrund dieser Kombination ist CIMAWA das einzige hybride Verfahren der Assoziationsberechnung, da es weder rein symmetrisch noch rein asymmetrisch arbeitet. Dieser konzeptuelle Unterschied grenzt CIMAWA grundsätzlich von allen anderen dem Autor bekannten Verfahren ab.
Die bis dato geleistete Forschungsarbeit am Institut für Wissensbasierte Systeme im Bereich der automatisierten Textanalyse, fokusiert die semantische Ähnlichkeit von Texten und die Berechnung von Schlüsselworten in Textdokumenten. Die entwickelten Verfahren können in Kombination mit dem CIMAWA-Ansatz in unterschiedlichen Applikationsszenarien zum Einsatz gebracht werden. CIMAWA ist jedoch das erste und einzige am Lehrstuhl entwickelte Verfahren, das Wortbeziehungen automatisiert bewertet.

1.4 Aufbau der Arbeit

Die vorliegende Arbeit beinhaltet 7 Kapitel. Nach dem einführenden ersten Kapitel mit Motivation, Problemstellung und Zielsetzung schließt sich Kapitel 2 mit den vorbereitenden Grundlagen an.
Darin wird zunächst eine Begriffsbestimmung des Text Mining-Begriffs vorgenommen, sowie die gängigsten Definitionen einander gegenübergestellt. Die für das Verständnis der Arbeit nötigen sprachstatistischen Grundlagen werden ebenfalls im zweiten Kapitel gelegt. Darüber hinaus ist die menschliche Wortassoziation ein Hauptthema dieses Abschnitts, wobei neben den psychologischen Grundlagen und Assoziationstests aus dem Bereich der Linguistik auch eigens erstellte Testreihen zur Charakteristik der menschlichen Wortassoziation einfließen.
Das dritte Kapitel trägt den Titel ‘Simulation von Wortassoziationen‘ und stellt die aus der Literatur bekannten Assoziationsberechnungsverfahren vor. Sämtliche in diesem Kapitel analysierten Verfahren wurden implementiert und mit der Eigenentwicklung CIMAWA verglichen (siehe Kapitel 5). Zur besseren Vergleichbarkeit der Verfahren wird zunächst eine einheitliche Definition der Simulationsparameter vorgenommen, um darauf aufbauend eine einheitliche Formelsammlung aller Methoden bereitzustellen. Des Weiteren werden die Simulationsverfahren basierend auf ihrer Konzeption als symmetrisch oder asymmetrisch kategorisiert.
Das vierte Kapitel ist der Eigenentwicklung CIMAWA gewidmet. Am Anfang wird CIMAWA in Relation zu den in Kapitel 3 vorgestellten Verfahren eingeordnet. Hierzu werden insbesondere die Konzeptionen verglichen, um deutlich zu machen, inwiefern die Verfahren voneinander abzugrenzen sind. Eine detaillierte Beschreibung des Aufbaus des CIMAWA-Ansatzes schließt sich an, wobei die eigentlichen Berechnung und die Parameterbeschreibung ein eigenes Unterkapitel einnimmt.
In Kapitel 5 werden sämtliche in Kapitel 3 und 4 vorgestellten Assoziationsberechnungsverfahren gegeneinander getestet. Hierzu wurden insgesamt drei Fallstudien durchgeführt, die zusammen genommen acht getrennt voneinander zu betrachtende Testreihen beinhalten. Die Fallstudien und Testreihen sind dabei so aufgebaut, dass sie durch die Verwendung unterschiedlicher Umgebungsparameter, wie Korpusgröße oder Korpuszusammensetzung, aussagekräftige Ergebnisse zur Güte der Assoziationsberechnung zulassen. Eine umfangreiche Zusammenfassung und Interpretation der dargestellten Ergebnisse runden dieses Kapitel ab.
Im sechsten Kapitel stellt die Arbeit in vier Unterkapiteln ebenso viele Anwendungsbeispiele der CIMAWA-Assoziationsberechnung vor.
In Kapitel 6.1 wird eine Methode namens ‘Associative Gravity‘ vorgestellt, in der CIMAWA dazu verwendet wird Multi-Themenstrukturen in Textdokumenten zu erkennen und darzustellen.
Der nächste Abschnitt 6.2 zeigt, wie CIMAWA im Anwendungsfeld des Instandhaltungsmanagement eingesetzt werden kann, um mit dem sogenannten ‘Association Mapping‘ Strukturen in Dokumenten der Instandhaltung zu identifizieren und zu visualisieren.

Das Kapitel 6.3 weitet das zuvor beschriebene Konzept auf das Anwendungsfeld des Produktlebenszyklusmanagement aus. Unter Verwendung der berechneten Assoziationen werden Textempfehlungen bereitgestellt, die den Produktentwickler bei der Bewältigung seiner Aufgaben unterstützen.
Das letzte Anwendungsbeispiel in Kapitel 6.4 zeigt, wie CIMAWA in einer assoziativen Dokumentensuchmaschine im unternehmerischen Umfeld eingesetzt werden kann. Ziel ist die Reduzierung der Recherchezeiten für Informationen und die Erhöhung der Wiederverwendungsquote von abgelegten Dokumenten.
Kapitel 7 fasst abschließend die wichtigsten Erkenntnisse der Arbeit zusammen und beschreibt im Ausblick Chancen und Herausforderungen in der zukünftigen Nutzung von CIMAWA.

2 Grundlagen

2.1 Der Text Mining Begriff

Wie die meisten zum Thema Text Mining erschienenen wissenschaftlichen Artikel beginnt auch diese Einführung mit der Beantwortung der Frage nach dem Grund für das steigende Interesse der Wissenschaftgemeinschaft an diesem Themengebiet [5]. Die Antwort liegt nach Meinung zahlreicher Autoren in der entstandenen und sich weiter verschärfenden Problematik der Informationsüberflutung [6], [7], [8]. Diese Informationsflut besteht für Mehler und Wolf [9] hauptsächlich aus online verfügbaren Dokumenten im Inter- und Intranet. Text Mining verspricht dieser Entwicklung in Ansätzen entgegenzuwirken, wobei Feldmann und Sanger prägnant zusammenfassen:

"*Text Mining is a* […] *research area that tries to solve the information overload problem by using techniques from data mining, machine learning, natural language processing (NLP), information retrieval (IR), and knowledge management.*" [10]

Die computerbasierte automatisierte Verarbeitung von Texten stellt Hearst in den Vordergrund in dem sie definiert:

"*Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources.*" [11]

Andere Definitionen zielen auf den Anwender und die Text Mining Applikationen ab, wobei insbesondere der Prozess der Entscheidungsfindung bzw. deren Unterstützung in den Fokus gerückt wird. So führen Aggarwal und Zhai aus:

"*Text Mining can be regarded as going beyond information access to further help users to analyze and digest information and facilitate decision making. There are also many applications of text mining where the primary goal is to analyze and discover any interesting patterns, including trends and outliers, in text data* [...]." [12]

Diese anwendugsbezogene Definition wird präzisiert durch Chen, indem er darlegt:

"Text mining performs various searching functions, linguistic analysis and categorizations." [13]

Abgesehen von der bloßen Anzahl an Dokumenten und der resultierenden Informationsflut, impliziert bereits die Namensgebung die zentrale Ausrichtung des Text Mining Forschungszweiges. Texte sind der Rohstoff aus dem Linguisten mit Hilfe mannigfaltiger Analysemethoden Kapital zu schlagen versuchen. Dabei bezeichnen Sommerfeld, Starke und

Hackel in [14] die Bedeutung der Texte und wie sich diese aufbauen als Textsemantik, und die Textsyntax geht der Frage nach, wie die Textbedeutung syntaktisch ausgedrückt werden kann.
Informationstheoretisch handelt es sich bei einem Text zunächst um aneinandergereihte Zeichen. Die Basis zur Verarbeitung solcher Zeichen wurde bereits sehr früh in der Informationstheorie (vgl. Shannon und Weaver [15]) gelegt. Heyer, Quasthoff und Wittig verstehen Zeichen als:

„[…] *Elemente eines endlichen, geordneten Zeichenvorrats* […]." [16]

die im Falle der Texte in einem Alphabet definiert sind. Durch die von der Grammatik vorgegebenen Ordnungsregeln lassen sich die Zeichen zu Zeichenketten zusammenstellen. Darüber hinaus beschreiben Heyer, Quasthoff und Wittig, wie sich Zeichen zu einer Nachricht akkumulieren und wann eine Nachricht zu einer Information für den Empfänger wird:

„*Eine nach vorher festgelegten Regeln zusammengestellte, endliche Folge von Zeichen und Zuständen, die eine Information vermittelt, bezeichnet man als Nachricht. Eine Nachricht zusammen mit ihrer Bedeutung für den Empfänger ist eine Information.*" [16]

Um aus der erhaltenen Information Wissen zu generieren, muss der Empfänger diese Information mit anderen vernetzen, das bedeutet, er bringt selbst gemachte Erfahrungen oder andere Informationen mit der erhaltenen in Zusammenhang, kombiniert mehrere Informationsobjekte und erzeugt so 'Wissen'. Ohne den letzten Schritt der Veredelung werden die Informationen als sinnhafte Datenobjekte verstanden und als 'Content' [16] bezeichnet.
Vor diesem informationstheoretischen Hintergrund liegt es beim Empfänger, oder im Falle von Texten beim Leser, diesen über die reine Datenebene zu Informationen und schließlich zu Wissen zu veredeln. Dahinter verbirgt sich eine generelle Annahme: Texte sind (in den allermeisten Fällen) der Versuch eines Autoren eine wie auch immer zu interpretierende Information festzuhalten und weiterzugeben. Treffend fasst Hearst zusammen:

"*Text expresses a vast, rich range of information, but encodes this information in a form that is very difficult to decipher automatically.*" [17]

Setzt man entsprechende Sprachkenntnisse und Vorwissen voraus ist es dem Menschen generell möglich eine Information bei entsprechendem Zeitaufwand aus dem Text zu extrahieren. Text Mining versucht im Speziellen aufgrund der steigenden Anzahl an verfügbaren Texten, sinnhafte Informationen automatisch zu extrahieren und nutzbar zu machen.

Die Weitergabe von Wissen mittels Texten stellen Heyer, Quasthoff und Wittig in den Vordergrund indem sie festhalten:

„*Seit den Anfängen der Schriftsprache in unserem Kulturkreis dient Text dazu, Wissen festzuhalten, zu bearbeiten und weiterzugeben.*“ [16]

Demzufolge dienen Texte seit jeher als Speicher menschlichen Wissens, wobei nach [18] an dieser Stelle einschränkend die generelle Frage nach der Versprachlichung oder Externalisierung von Wissen zu thematisieren ist. Zwar ist es im Allgemeinen nicht möglich, das gesamte gesammelte Wissen eines Menschen in Worte zu fassen (Erfahrungswissen, erlernte Handlungsabläufe, handwerkliche Fähigkeiten etc.), wohldefinierte Teilbereiche des menschlichen Wissensfundus, in denen dies möglich ist lassen sich jedoch sehrwohl benennen (Enzyklopädien, wissenschaftliche Artikel, Fachbücher, technische Berichte etc.). Auf die zuvor genannten Dokumente zielen Text Mining-Verfahren ab und so kann zusammenfassend geschlussfolgert werden, dass sich aus den in Texten kodifizierten Wissensbeständen die Potentiale des Text Mining ergeben.
Ziele und Anwendungen des Text Mining Forschungsgebiets sind auf Basis der gegebenen Umgebungsparameter im digitalen Zeitalter weitestgehend einheitlich definiert. Wie bei den meisten vergleichsweise jungen Disziplinen, existiert jedoch auch beim Text Mining (noch) keine einheitliche Benennung des Forschungsgebiets. Unterschiedliche Namensgebungen, dessen was in der vorliegenden Arbeit unter dem Oberbegriff Text Mining verstanden wird, fassen Mehler und Wolff in [9] zusammen:

- Text Mining [19], [20]
- Knowledge Discovery in Textual Databases [21]
- Text Knowledge Engineering [22]
- Knowledge Discovery in Texts [23]
- Text Data Mining [17], [24]
- Textual Data Mining [25]

Neben diesem Auszug aus der Namensgebung des Forschungsfeldes und den zum Teil differierenden Ansichten der Autoren bezüglich dessen Umfang, existiert ebenso keine einheitliche Definition des Text Mining-Begriffs selbst (vgl. [26], [27], [28]). Zusätzlich zu den bereits aufgeführten Definitionen des Text Mining-Begriffs zu Beginn dieses Kapitels, soll an dieser Stelle der enge Bezug zur Forschungsdisziplin des Data Mining herausgearbeitet werden. Deshalb wird im Folgenden ein Auszug an Definitionen zitiert, die die dargestellten Definitionen um diese Facette ergänzen:

“[...] *text data mining is a process of exploratory data analysis* [29], [30] *that leads to the discovery of heretofore unknown information, or to answer to questions for which the answer is not currently known.*” [17]

" *Text mining can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools. In a manner analogous to data mining, text mining seeks to extract useful information from data sources through the identification and exploration of interesting patterns. In the case of text mining* [...]*, the data sources are document collections, and interesting patterns are found* [...] *in the unstructured textual data in the documents in these collections."* [10]

"*We define text mining to be data mining on text data. Text mining is all about extracting patterns and associations previously unknown from large text databases.*" [31]

Bei Betrachtung dieser Text Mining Definitionen fällt der unmittelbare Bezug zum Data Mining Begriff auf, der in zahlreichen Definitionen Verwendung findet. Übereinstimmend ist den allermeisten Publikationen zu diesem Thema zu entnehmen, dass Text Mining ein Teilgebiet des Data Mining darstellt [13]. Data Mining wiederum bezeichnet einen Schritt im Knowledge Discovery in Databases (KDD) Prozess (vgl. [32]). Abstrakt ausgedrückt fasst KDD den Gesamtprozess zusammen, in dem nützliches Wissen aus Daten extrahiert wird [32]. Coronel, Morris und Rob beschreiben Data Mining als:

"*A process that employs automated tools to analyze data in a data warehouse and other sources and to proactively identify possible relationships and anomalies.*" [33]

Die Autoren beschreiben weiter, dass die Data Mining Werkzeuge zwar in der Lage sind, die Beziehungen und Anomalien automatisch zu extrahieren, dennoch kommen sie nicht ohne Interaktion mit dem Nutzer aus. Zu den vom Nutzer zu erfüllenden Aufgaben gehört unter anderem die Problemdefinition, die Auswahl der Daten sowie der Anstoß der Analyse.
In ihrem in dieser Arbeit bereits mehrfach zitierten Artikel mit dem Titel "Untangling Text Data Mining" [17] klassifiziert Marti A. Hearst unter anderem Data Mining, Information Retrieval, Computational Linguistics und Text (Data) Mining. Die für eine Differenzierung der Begrifflichkeiten wichtigsten Ausführungen werden im Folgenden kurz zusammengefasst. Hearst ist der Ansicht, dass der 'Mining' Begriff an sich keine angemessene Metapher für die Methoden darstellt, die man gemeinhin unter Data Mining zusammenfasst, da Mining die Extraktion von wertvollen 'Nuggets' aus ansonsten wertlosem Gestein impliziert. Wörtlich genommen sollte Data Mining in diesem Fall völlig neue und bislang unbekannte Fakten aus einer Bestandsdatenbank extrahieren können. Stattdessen wird argumentiert, dass Data Mining zwar Trends und Muster (semi)automatisch extrahiert, aber keine bislang völlig unbekannten Fakten zu Tage fördert. Einen Gegensatz dazu will Hearst beim, wie sie es nennt, Text (Data) Mining gefunden haben, denn bei der Analyse von Texten kann die 'Mining for Nuggets' Metapher tatsächlich Anwendung finden und Antworten gefunden werden, bei denen bislang nicht einmal die Frage selbst bekannt war.
Eine ähnliche Argumentationskette bildet die Autorin beim Vergleich zwischen 'Information Retrieval' oder wie sie es nennt 'Information Access' und Text (Data) Mining. Das Ziel des Information Retrieval ist es, dem Nutzer diejenigen Dokumente zur Verfügung zu stellen, die er bezogen auf den aktuellen Kontext benötigt. Anders ausgedrückt geht es um die Identifikation einer Information in Form eines Dokuments in einer Menge von Dokumenten.

Da die Information in diesem Dokument keine Neuheit darstellt, denn sie ist zumindest dem Autoren des Dokuments bekannt, kann die Mining Metapher auch an dieser Stelle nicht unmittelbar angewendet werden.
Folgt man der gebräuchlichen Begriffsauslegung von Text Mining und versteht es als eine Art Data Mining in Textsammlungen, so gibt es bereits einen Namen für diese Forschung: Corpus-based Computational Linguistics [34]. Hierbei werden Statistiken basierend auf großen Textsammlungen erhoben und nach Mustern untersucht. Gefundene Muster können für verschiedene Anwendungen im natural language processing (NLP) eingesetz werden. Beispiele hierfür sind ‘part-of-speech tagging‘ und bilinguale Wörterbücher [34]. Die folgende Tabelle 1 fasst die beschriebenen Unterschiede zusammen.

Tabelle 1. Klassifikation von Data Mining und Text (Data) Mining nach Hearst [17]

	Finding Patterns	Finding Nuggets	
		Novel	Non-Novel
Non-textual data	Standard Data Mining	?	Database queries
Textual data	Computational Linguistics	real Text Data Mining	Information Retrieval

Mit einem eingängigen Beispiel, was Hearst unter ‘real Text Data Mining‘ oder der Antwort auf eine unbekannte Frage versteht, können die Unterschiede weiter verdeutlicht werden. Der Informatiker Don R. Swanson wurde bekannt durch seine Analysen auf medizinischen Textsammlungen, in denen er versuchte Beziehungen zwischen Sachverhalten herzuleiten, die bislang völlig unbekannt waren. In einem Anwendungsfall konnten allein durch die Analyse von biomedizinischen Artikeln Rückschlüsse auf einen Zusammenhang von Magnesiummangel und Migräne-Kopfschmerzen gezogen werden. Später wurde dieser Zusammenhang tatsächlich in einer klinischen Studie nachgewiesen und bildet so ein plakatives Beispiel für Unterscheidung zwischen Computational Linguistics auf der einen, und real Text (Data) Mining auf der anderen Seite.
Trotz nachvollziehbarer Argumentation ist der Differenzierungsansatz nach Hearst bis heute weniger gebräuchlich und in Teilen umstritten (vgl. dazu [35]). Deshalb wird in dieser Arbeit die meist gebräuchliche Definition zugrunde gelegt, die Text Mining als einen Teilbereich des Data Mining versteht.

2.2 Sprachstatistische Grundlagen

Wie bereits zuvor herausgearbeitet, handelt es sich bei Texten in aller Regel um Wortfolgen, die mit einer bestimmten Intention des Autors verfasst wurden. Daher ist davon auszugehen, dass hinter der Wortfolge eines Textes, die Intention des Verfassers steckt, eine Aussage zu vermitteln. Um diese Aussagen automatisiert extrahieren bzw. nutzen zu können, werden sprachstatistische Verfahren angewendet, die in erster Linie auf der Häufigkeitsverteilung und statistischen Abhängigkeiten zwischen Worten basieren [16]. Deshalb sollen in diesem Abschnitt die wichtigsten sprachstatistischen Gesetzmäßigkeiten, Grundlagen und Begriffe vorgestellt werden.

2.2.1 Zipfsches Gesetz

Bekannt wurde John Kingsley Zipf durch seine Untersuchungen der Häufigkeit von Wortformen in mehreren Sprachen. Zipf erkannte dabei als Erster den Zusammenhang zwischen Rang und Häufigkeit einer Wortform und postulierte, dass die natürliche Sprache dem ‘Principle of Least Effort‘ folgt und so die am häufigsten verwendeten Worte einer Sprache meist sehr kurze Funktionswörter sind [16]. In [36] stellte Zipf fest, dass die Liste der Worte, absteigend sortiert nach absoluten Häufigkeiten, multipliziert mit deren Rang, in etwa konstant ist. Mit der textabhängigen Konstanten *k*, dem Rang *r* und der absoluten Häufigkeit *n* gilt:

Formel 1. Zusammenhang von Rang und Häufigkeit nach Zipf

$$r * n \approx k$$

Im Zusatz besagt das Zipfsche Gesetz, dass die Häufigkeit eines Wortes umgekehrt proportional zu seinem Rang ist, was wiederum bedeutet, dass relativ wenige Worte einen großen Anteil aller Texte ausmachen [37].
Die folgende Abbildung 1 visualisiert den von Zipf erkannten Zusammenhang zwischen der absoluten Häufigkeit der Worte auf der Y-Achse und des Ranges der Worte auf der X-Achse. Als Basis für die folgende Auswertung dient das Projekt ‘Deutscher Wortschatz‘ der Universität Leipzig mit einem Umfang von ca. 18 Mio. Wortinstanzen.

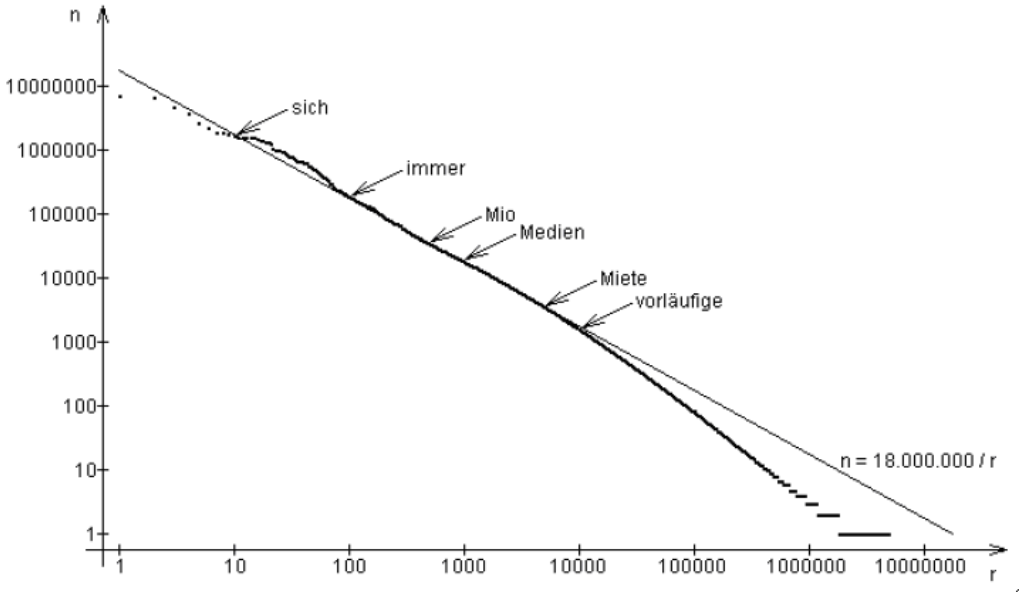


Abbildung 1. Zusammenhang zwischen Rang und Worthäufigkeit [16]

Für beide Achsen gilt eine logarithmische Skalierung, wobei der Verlauf zeigt, dass die Worte annäherungsweise auf einer Gerade liegen. Wird ein idealer Korpus vorausgesetzt bzw. fände das Zipfsche Gesetz exakt Anwendung, würden die abgetragenen Worte in Abbildung 1 eine Gerade bilden. Es zeigt sich jedoch, dass in den Randbereichen deutliche Abweichungen zu

verzeichnen sind, welche sich laut Klahold [37] bei weniger großen und thematisch fragmentierten Korpora weiter verstärken.
Erwähnung finden soll an dieser Stelle, dass Benoît B. Mandelbrot (vgl. [38]) eine Anpassung der Zipfschen Formel vorschlägt, die eine Krümmung der Geraden aus Abbildung 1 in den Randbereichen vornimmt und sich so dem beobachteten Verlauf annähert.
Sinnvoll genutzt werden kann das Zipfsche Gesetz zur Abschätzung von textbezogenen Variablen. Dazu gehört die Abschätzungen der Vergrößerung des Vokabulars einer Textsammlung bei entsprechender Erhöhung der Textmenge, die Abschätzung der Anzahl bestimmter Wortformen bei gegebenen Umgebungsparametern oder andere Umfangsschätzungen bezüglich des Vokabulars (vgl. [16]).

2.2.2 Kookkurrenz: Das gemeinsame Auftreten von Wortformen

Kookkurrenz oder englisch ‘cooccurrence‘ aus Sicht des Text Mining, ist die Bezeichnung für das gemeinsame Auftreten zweier Wortformen innerhalb eines definierten Textfensters. Formell ausgedrückt beschreibt die Kookkurrenz

> „[…] *das Miteinandervorkommen sprachlicher Einheiten in derselben Umgebung (z.B. im Satz).“* [39]

Je nach Größe und Art der Umgebung können mehrere Arten der Kookkurrenz benannt werden. Textfenster können definiert werden als Satz-, Absatz-, oder Gesamtdokument. Dementsprechend werden diese auch als Satz-, Absatz-, und Dokumentkookkurrenzen bezeichnet. Die gängigste Vorgehensweise für die Berechnung ist jedoch die fensterbasierte Kookkurrenz fester Größe. Dabei wird ein Textfenster mit fester Anzahl von Worten definiert, das rechts und links von dem zu untersuchenden Wort aufgebaut wird. Ein Spezialfall der fensterbasierten Kookkurrenzen bildet dabei die Nachbarschaftskookkurrenz, diese beschreibt das Vorkommen zweier Wortformen in unmittelbarer Nachbarschaft und einer Fenstergröße ±1. Eigennamen, Redewendungen und Substantive mit beschreibenden Adjektiven sind Beispiele für solche Nachbarschaftskookkurrenzen.
Die Festlegung der Fenstergröße ist stets an die Zielvorgabe der Analyse gekoppelt. Wählt man eine sehr kleine Fenstergröße, so werden die häufig miteinander vorkommenden Wortformen in der Tendenz auch eine starke semantische Beziehung aufweisen. Die Ergebnisse sind von hoher Qualität. Allerdings bleiben durch die Wahl einer kleinen Fenstergröße solche Beziehungen unentdeckt, die außerhalb der Reichweite des kleinen Textfensters liegen. Mit der Vergrößerung des Textfensters erhöht sich demzufolge die Ergebnismenge. Das bedeutet, dass in der Breite mehr semantische Wortbeziehungen aufgedeckt werden, aber gleichzeitig die Anzahl der unbrauchbaren Wortpaare in der Ergebnismenge steigt. Die erzielte quantitative Verbesserung bedeutet in der Folge Einbußen in qualitativer Hinsicht.
Nach Festlegung der Fenstergröße wird für jede Wortform berechnet, wie häufig diese mit anderen Wortformen im gewählten Textfenster vorkommt. Ergebnis ist eine Matrix mit den

Kookkurrenzen, die in ihrer Größe mit dem Umfang der zugrunde liegenden Textbasis variiert.
Tabelle 2 zeigt exemplarisch eine solche Kookkurrenzmatrix. Zu entnehmen ist, wie häufig die aufgeführten Worte innerhalb der Textbasis und dem gewählten Textfenster gemeinsam auftreten. Definitionsbedingt sind solche Matrizen stets symmetrisch aufgebaut. Unabhängig von der Art der Präsentation der Kookkurrenzergebnisse, ob in Form einer Matrix für eine Textsammlung oder in Listenform für eine bestimmte Wortform, bleibt die Frage nach der Interpretation der berechneten Statistik.

Tabelle 2. 4x4 Kookkurrenzmatrix

	rot	blau	gelb	Grün
rot	-	11	6	67
blau	11	-	9	3
gelb	6	9	-	0
grün	67	3	0	-

Insbesondere bei großen Textsammlungen sind die Kookkurrenzstatistiken sehr umfangreich und enthalten derart viele Wortpaare, dass die Frage nach den wichtigsten Kookkurrenzen nicht unbeantwortet bleiben darf. Die wichtigsten Kookkurrenzen sind dabei nicht zwangsläufig diejenigen mit der höchsten Frequenz, also diejenigen die am häufigsten im Text auftreten. Aufgrund der Frequenzen der einzelnen Worte in einer Textsammlung kommen rein statistisch gesehen manche Wortverbindungen häufiger vor als andere. Anders ausgedrückt müssen zwei Wortformen nur häufig genug im Text vorkommen, um entsprechend häufig miteinander vorzukommen. Klahold definiert daher den Begriff der Kollokation mit Bezug auf Kookkurrenz eher allgemein, indem er ausführt:

„*Eine Kollokation liegt vor, wenn die Kookkurrenz* […] *zweier Worte* […] *signifikant über dem Durchschnittswert* […] *aller Wortpaare liegt.*“ [37]

Gesucht wird demzufolge nach Methoden um Kollokationen im Text ausfindig zu machen. Diese Methoden werden in der Literatur als Signifikanzmaße [16] oder Assoziationsberechnungsverfahren (engl. association measures) [40], [41] bezeichnet.
Bereits sehr früh wurde mit der Entwicklung solcher Methoden begonnen (vgl. [42]). Um die signifikanten Kookkurrenzen zu extrahieren, nehmen die Verfahren in Gänze eine statistische Interpretation der berechneten Kookkurrenzdaten vor. Ein umfassender Überblick der in dieser Arbeit verwendeten Berechnungsverfahren liefert Kapitel 3 der vorliegenden Arbeit.

2.3 Menschliche Wortassoziation

Um eine wie auch immer geartete technische Implementierung der menschlichen Wortassoziation umzusetzen, ist es unumgänglich zunächst eine Analyse dieser Wortassoziation aus psychologischer Sicht vorzunehmen. Der folgende Abschnitt beleuchtet daher diesen Aspekt und gibt Einblicke in Gegebenheiten, die als Grundvoraussetzung in die Implementierung inkludiert werden müssen.
Die Psychologie versteht unter Assoziation die Verbindung von Bewusstseinsinhalten (Erinnerung, Begriff etc.). Das bedeutet, dass das Auftreten des einen das Auftreten eines anderen, mit ihm assoziierten Inhalts nach sich zieht [43]. Schon Platon thematisierte in seinem Werk 'Phaidon' einen ersten theoretischen Ansatz zum Konzept der Assoziation. Aristoteles erklärt in „Über das menschliche Gedächtnis und die Erinnerung" das Zustandekommen von Assoziationen mit insgesamt drei Assoziationsprinzipien. Demnach rufen Gedächtnisinhalte eben solche hervor, die dem Ursprungsinhalt ähnlich, entgegengesetzt oder in räumlichen und zeitlichen Zusammenhang gebracht werden. Speziell das Assoziationsprinzip der räumlichen und zeitlichen Kontiguität findet seither breite Zustimmung. Diese erfahrungsabhängigen Assoziationen bilden die Grundlage aller geistigen Vorgänge und bilden den Grundgedanken der Assoziationspsychologie [43].
Die Verbindung zwischen Assoziationen und Worten stellt Klahold her, indem er ausführt:

„Unter Assoziation versteht man die Verknüpfung zweier oder mehrerer gedanklicher Inhalte durch einen Menschen. Auf Worte bezogen stellt die Wortassoziation Beziehungen zwischen unterschiedlichen Worten auf Basis deren Bedeutung her." [37]

Für Jahn sind Assoziationen die Verbindung verschiedener Vorstellungen, wobei eine Vorstellung eine andere aus dem Unbewusstsein ins Bewusstsein zieht und unter Assoziation so das Aneinanderreihen und Verbinden der Vorstellungen bei ihrem Bewusstwerden zu verstehen ist [44]. Andere Autoren wie Hartley [45] und Mill [46] sind der Auffassung, dass sich die Verknüpfung einander ähnelnder Objekte auf das Kontiguitätsprinzip zurückführen lässt, denn ähnliche Objekte werden häufig gleichzeitig oder in unmittelbarer Folge wahrgenommen [47]. Diese Sichtweise geht zurück auf James und seine Formulierung:

*"Objects once
experienced together tend to become associated in the imagination, so that when any one of them is thought of, the others are likely to be thought of also, in the same order of sequence or coexistence as before. This statement we may name the law of mental association by contiguity."* [48]

Seidensticker schlussfolgert in ähnlicher Weise und führt aus, dass Dinge in unmittelbarer zeitlicher Nähe nach dem Assoziationsgesetz miteinander verknüpft sind [49].
Zur Stützung der vorgenannten Thesen sei an dieser Stelle auf den Psychologen Iwan Pawlow verwiesen, der mit dem Klassischen Konditionieren als Erster das Erlernen von Assoziationen untersuchte. In seinen Versuchen an Hunden bot Pawlow vor der Fütterung der Tiere einen

neutralen Stimulus in Form des Erklingens einer Glocke dar. Die natürliche Reaktion der Hunde auf das Darbieten von Futter ist erhöhte Speichelabsonderung. Nach entsprechender Wiederholung des Experiments, also der gemeinsamen Präsentation von Glocke und Futter, konnte eine sogenannte konditionierte Reaktion beobachtet werden: Die Tiere reagierten auch ohne Futterangebot und alleinig auf das Erklingen der Glocke mit erhöhter Speichelabsonderung. Die Ergebnisse dieses Experiments deuten darauf hin, dass zeitliche Nähe zwischen Ereignissen die Bildung von entsprechenden Assoziationen zwischen diesen Ereignissen hervorrufen kann.

Für Seidensticker können die menschlichen Wortassoziationen jedoch nicht allein auf die klassische Konditionierung zurückgeführt werden, weil nicht alle Wortbedeutungen auf unbedingte Reflexe zurückzuführen sind [49]. Vielmehr handelt es sich bei dem Erlernen von Wortassoziationen um eine „Konditionierung höherer Ordnung“ [49], in dem bestimmte Wörter häufiger mit bestimmten anderen Wörtern in einem Text vorkommen. Je häufiger ein Individuum ein Wort in unmittelbarer Nähe zu einem anderen liest, desto stärker wird die Assoziation zwischen diesen Worten. Vermindert sich das gemeinsame Vorkommen der Worte jedoch, so kann diese Assoziation, analog zum klassischen Konditionierungsexperiment von Pawlow, auch wieder verlernt werden.

In diesem Zusammenhang beschäftigte sich Staats in [50] mit dem initialen Bilden von Wortassoziationen bei Kindern. Der sogenannten „*operanten Konditionierung*“ [50] kommt dabei eine zentrale Rolle beim Erlernen von Wortbedeutungen zu. Demnach erlernen Kleinkinder die ersten Wörter, indem sie aufgenommene Worte zu imitieren versuchen. Sind diese Imitationen erfolgreich, erfolgt in der Regel eine Bestärkung der Kinder durch ihre Eltern. Misslingt allerdings der Versuch den Klang des aufgenommenen Wortes widerzugeben, so bleibt diese Bestärkung in der Regel aus.

In einem späteren Entwicklungsstadium lernen die Kinder, Bilder und Aktivitäten (‘Fahrrad‘ und ‘fahren‘; ‘Wasser‘ und ‘trinken‘) nach dem Prinzip der operanten Konditionierung zu verknüpfen [49]. Noch weiter in seiner Entwicklung fortgeschritten können den Kindern anstelle der Bilder oder Gegenstände lediglich Worte dargebracht werden, worauf die Kinder im Idealfall mit der „richtigen“ Assoziation antworten. Als Beispiel hierfür kann das zuvor verwendete Wortpaar erneut angeführt werden: Nennt man dem Kind den Stimulus ‘Fahrrad‘ und antwortet dieses mit dem erwarteten ‘fahren‘, so wird das Kind für seine Antwort gelobt. Durch die Bestärkung des Kindes wird die erlernte Wortassoziation gefestigt und in diesem Sinne gelernt.

Laut Seidensticker setzt sich das Erlernen von Wortassoziationen auch in fortgeschrittenem Alter fort [49]. Insbesondere beim Lesen von Texten bilden sich Wortassoziationen dadurch, dass Worte in einem gemeinsamen Kontext, also in unmittelbarer Nähe zueinander wahrgenommen werden. Hat ein Leser ein Wortpaar bereits mehrfach in ähnlichem Kontext wahrgenommen, so erwartet er diese Kombination auch zukünftig vorzufinden. Praktisch bedeutet das, dass er nach dem Lesen des ersten Wortes einer entsprechend erlernten Wortassoziation auch das zweite Wort erwartet. Wird die Erwartung erfüllt, so wird dies als Bestärkung dieser Wortassoziation wahrgenommen und gewertet. Sollte die Erwartung jedoch wiederholt nicht erfüllt werden, so wird sich diese Wortassoziation abschwächen bzw. sie wird im beschriebenen Sinne verlernt. Demnach ist die nicht Erfüllung der Erwartung gleichbedeutend mit dem Ausbleiben der Bestärkung bei der operanten Konditionierung.

Insbesondere die aus der Literatur heraus begründeten und hier beschriebenen Regeln zur Bildung von Wortassoziationen von Individuen, während des Lesens von Texten, dient als Grundlage für die technische Simulation der menschlichen Wortassoziation. Im Speziellen die statistischen Werte in Bezug zum gemeinsamen Vorkommen von Worten zielen auf die beobachteten Gesetzmäßigkeiten der Bildung von Wortassoziationen beim Menschen ab.

2.3.1 Assoziationsexperimente

Eine intuitive Vorgehensweise um menschliche Wortassoziationen zu studieren ist das Durchführen von entsprechenden Experimenten. Solche Wortassoziationsexperimente sind in der Psychologie und der Linguistik weit verbreitet [51], [52], [53]. Eine spezielle Form der Assoziationsexperimente sind die sogenannten freien Assoziationstests (FAT). Im Gegensatz zu anderen Assoziationstests, bei denen die Assoziation zwischen zwei oder mehreren vorgegebenen Wörtern bewertet werden muss, werden bei den FATs lediglich Begriffe vorgegeben und die Testpersonen antworten frei. Die Grundstruktur dieser freien Experimente ist dabei sowohl in der Durchführung als auch in der Auswertung sehr ähnlich. Nelson et al. [54] beschreiben die Aufgabe der Probanden in den FATs wie folgt:

"Participants were asked to write the first word that came to mind that was meaningfully related or strongly associated to the presented word [...]."

Den Probanden wird zunächst ein Wort oder Begriff präsentiert, das sogenannte Stimuluswort oder Reizwort. Die Teilnehmer sind angehalten auf dieses Stimuluswort mit dem ersten Wort zu antworten, das ihnen in den Sinn kommt. Für diese Antwort wird den Probanden in der Regel nur eine sehr kurze Zeitspanne gewährt, da die Antwort möglichst aus der Intuition heraus gegeben werden soll. Durch diese Vorgehensweise wird sichergestellt, dass die Wortassoziationen möglichst unverfälscht aufgenommen werden können. Da den Testpersonen keine Zeit gegeben wird, über mögliche Alternativantworten nachzudenken, können die gefundenen Wortassoziationen als intuitiv bezeichnet werden. Die gegebene Antwort wird als Responsewort oder als assoziative Antwort auf das Stimuluswort bezeichnet. Die Ergebnisse solcher Assoziationsexperimente werden in Häufigkeitstabellen für jedes Stimuluswort festgehalten (siehe Tabelle 4) und als Assoziationsnormen bezeichnet. Das meistgenannte Responsewort für einen Stimulus wird Primärantwort genannt [49]. Diese Primärantworten der FATs besitzen große Bedeutung für die vorliegende Arbeit, da diese Antworten die stärkste Assoziation für das entsprechende Reizwort darstellt. Die mathematisch-statistischen Verfahren zur Assoziationsberechnung, inklusive dem in dieser Arbeit entwickelten Ansatz, versuchen genau dieses am stärksten mit dem Stimuluswort assoziierte Responsewort vorherzusagen. Insbesondere in Kapitel 5, in dem verschiedene Fallstudien zur menschlichen Wortassoziation dargestellt sind, wird auf die Vorhersage dieser Primärantworten aus den Assoziationstests eingegangen.
Eine bezüglich der Quantität der befragten Probanden und der bewerteten Stimulusworte umfangreichsten Studien wurde von Nelson et al. in [53] veröffentlicht. An dieser über mehr als zwei Jahrzehnte andauernden Studie nahmen über 6.000 Personen teil, wobei insgesamt

5.019 Stimulusworte zu bewerten waren [53]. Jedes Wort wurde im Durchschnitt 150 Probanden gezeigt und jedem Teilnehmer zwischen 100 und 200 Worte vorgelegt. Am Ende der Studie waren mehr als 750.000 assoziative Antworten auf die ausgegebenen Stimulusworte zu verzeichnen [53]. Die meistgenutzte Assoziationsnorm im englischsprachigen Raum stammt jedoch von Kent und Rosanoff und wurde in [51] veröffentlicht. Deren Studie umfasst insgesamt 100 Stimulusworte. Eben diese Stimulusworte wurden von Russell und Meseck ins Deutsche übersetzt, um damit eine eigenständige Studie durchzuführen. Die Ergebnisse zu dieser Studie wurden in [52] veröffentlicht. Probanden des Experiments waren 331 Schüler und Studenten aus Süddeutschland. Eine vollständige Liste der verwendeten Stimulusworte der Studie findet sich in Tabelle 3.

Tabelle 3. Stimuluswörter der Studie von Russell und Meseck [52]

Adler	Essen	Junge	Musik	schwarz
ängstlich	Fenster	Kalt	Nadel	schwer
arbeiten	Fluss	Käse	Obst	Soldat
Arzt	Frau	Kind	Ofen	Sorge
Baby	Freude	Kohl	Ozean	Spinne
Bad	Fuß	kommandieren	pfeiffen	Stadt
Bequemlichkeit	Gedächtnis	König	Priester	Stiel
Berg	Gelb	Kopf	Quadrat	Straße
Bett	Gerechtigkeit	Krankheit	Rauh	Stuhl
Bibel	Gesundheit	Kurz	Religion	süß
Bitter	Glatt	Lampe	Rot	Tabak
Blau	Grün	Lang	Ruhig	Teppich
Blüte	Hammelfleisch	langsam	Salz	tief
Brot	Hammer	Laut	Sauer	Tisch
Bürger	Hand	Licht	Schaf	träumen
Butter	Hart	Löwe	Schere	weich
Dieb	Haus	Mädchen	schlafen	weiß
dunkel	Häuschen	Magen	Schmetterling	Whisky
durstig	Hoch	Mann	Schnell	wünschen
Erde	Hungrig	Mond	Schön	Zorn

Jedes der in Tabelle 3 aufgeführten Stimulusworte wurde den Probanden der Studie vorgelegt. Diese gaben jeweils das aus ihrer Sicht am stärksten assoziierte Wort an. Im Ergebnis wurde für jedes der Worte eine Liste mit den assoziativen Antworten erstellt. Nachfolgende Tabelle 4 visualisiert exemplarisch die Studienergebnisse von Russel und Meseck für das Stimuluswort ‘Butter‘. Aufgeführt sind in diesem Fall die zehn häufigsten Antworten der Probanden sowie die Anzahl der Probanden die diese Antwort gaben.

Tabelle 4. Assoziative Antworten zum Stimulus 'Butter' [52]

Stimuluswort	Antwort	Anzahl Probanden
Butter	Brot	60
	weich	44
	Milch	32
	Margarine	27
	Käse	20
	Fett(e)	16
	gelb	14
	Butterbrot	8
	Dose	6
	essen	6

Demzufolge sind Tabelle 4, neben der Primärantwort, auch die weniger assoziierten Begriffe zum gegebenen Stimulus zu entnehmen. Auf den Stimulus ‘Butter‘ antworteten 60 der Teilnehmer mit dem Wort ‘Brot‘, was diesen Begriff zur Primärantwort in diesem Experiment macht.

2.4 Untersuchungen zur Charakteristik der menschlichen Wortassoziation

Als Grundlage für eine eigene Umsetzung zur Simulation der menschlichen Wortassoziation soll zunächst ein Überblick zur Charakteristik dieser gegeben werden. Dabei wird insbesondere untersucht, welche Grundregeln zur Beschreibung der Wortassoziation beim Menschen in der Literatur beschrieben sind. Demzufolge unterscheidet man in der Linguistik zwischen zwei grundsätzlichen Wortrelationen: Zum einen die paradigmatischen und syntagmatischen und zum anderen die symmetrischen und asymmetrischen Relationen. Erstere Kategorisierung geht in seinem Ursprung auf den Schweizer Linguisten Ferdinand de Saussure zurück, der entsprechende Beziehungen zwischen Einzelelementen beschreibt und dessen Ergebnisse unter anderem in [55] publiziert wurden. Bei Letzteren steht die Assoziationsstärke zwischen Worten und deren Richtung im Vordergrund.
In die Analyse fließen sowohl die Ergebnisse aus der Literatur, als auch die Ergebnisse einer eigens zu diesem Zweck durchgeführten Serie von Experimenten an freiwilligen Probanden ein. Den beiden erwähnten Relationsarten sind die nächsten Abschnitte gewidmet.

2.4.1 Syntagmatische und paradigmatische Beziehungen

Wie bereits angedeutet geht die Unterscheidung zwischen syntagmatischen und paradigmatischen Beziehungen auf Ferdinand de Saussure zurück. In „Grundfragen der allgemeinen Sprachwissenschaft“ [55] wird im Wortlaut von assoziativen und syntagmatischen Beziehungen gesprochen. Der Begriff assoziativ wurde später von dem dänischen Linguisten Louis Hjelmslev durch paradigmatisch ersetzt (vgl. [56]). Seither hat sich diese Bezeichnung durchgesetzt und ist die heute gängige Bezeichnung. Nichtsdestotrotz ist Ferdinand de Saussure als Urheber der syntagmatischen und paradigmatischen Beziehungen zu benennen.
Das wichtigste Kriterium für das Vorliegen einer syntagmatischen Beziehung zwischen zwei Elementen ist ihr gemeinsames Vorkommen. Dabei werden in erster Linie inhaltliche Zusammenhänge abgebildet. Beispiele hierfür sind Wortformen, die typischerweise in unmittelbarer Nachbarschaft in einem Text bzw. in Sätzen benutzt werden:

- Eigennamen (San Francisco; Los Angeles; Universität Siegen etc.)
- Nomen-Verb Kombinationen (Taxi fahren, Haus bauen, Diplomarbeit schreiben etc.)
- Personennamen (Angela Merkel, James Bond etc.) oder
- Feste Wendungen (Guten Morgen, Fröhliche Weihnachten etc.).

Syntagmatische Beziehungen sind folglich aus ihrer Kombinierbarkeit und der Möglichkeit aus ihnen Sätze zu bilden definiert.
Paradigmatische Beziehungen hingegen werden als solche bezeichnet, wenn die Zeichen in ähnlichem Kontext vorkommen. Sind zwei Begriffe (Zeichen oder Worte) gegeneinander austauschbar, so stehen diese in einer paradigmatischen Beziehung zueinander. Paradigmatische Beziehungen erfassen typischerweise Zusammenhänge zwischen Elementen, die sich in syntaktischer, semantischer oder logischer Hinsicht ähnlich sind [16]. Typische Beispiele für solche Beziehungen sind synonym verwandte Begrifflichkeiten (rennen, laufen) oder noch allgemeiner solche, die auch ohne semantische Ähnlichkeit gegeneinander ausgetauscht werden können.

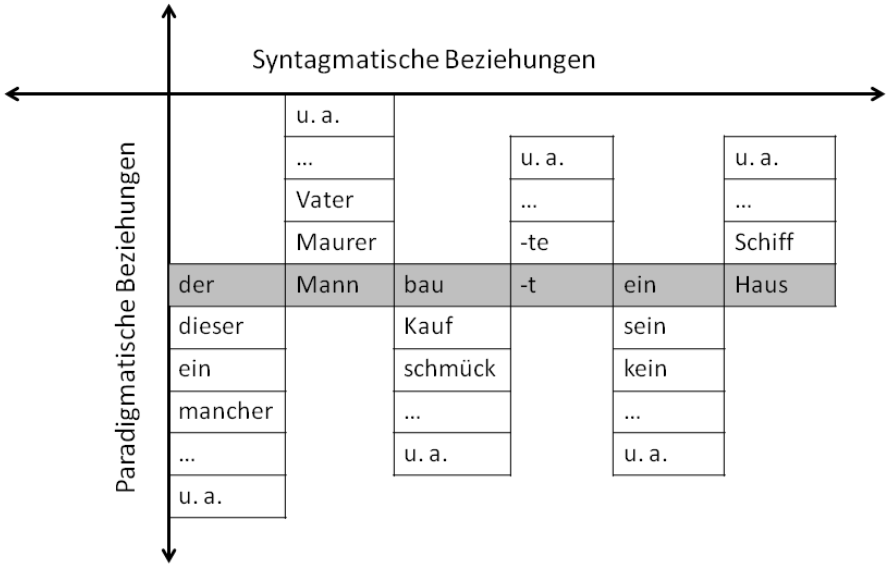


Abbildung 2. Schematische Darstellung eines Beispiels [57]

Eine Übersicht mit entsprechenden Beispielen liefert Ulrich in [57]. Die oben dargestellte Abbildung 2 entstammt [57] und erläutert die Beziehung zwischen den verschiedenen Sprachelementen. Die syntagmatische Relation ist dargestellt in der horizontalen Ebene und beschreibt die Beziehung eines Elements zu den Elementen in seiner Umgebung. Die vertikale Ebene hingegen beschreibt die paradigmatischen Beziehungen, wobei hier die Beziehungen der Elemente im Fokus stehen, die substituierbar sind und sich gegenseitig ausschließen.
Zusammenfassend bringen Manning und Schütze den Unterschied der Beziehungsarten auf den Punkt, indem sie ausführen:

"All elements that can be replaced for each other in a certain syntactic position [...] *are members of one paradigm. In contrast, two words bear a syntagmatic relationship if they can form a phrase (or syntagma)* [...]*."* [58]

Syntagmatische Beziehungen zwischen Wortpaaren finden sich demnach im lokalen Kontext in unmittelbarer Nachbarschaft, paradigmatische Beziehungen hingegen definieren sich in einem globalen Kontext durch Austauschbarkeit.

2.4.2 Symmetrische und asymmetrische Beziehungen

Neben der Unterscheidung zwischen syntagmatischen und paradigmatischen Beziehungen findet derzeit parallel eine Diskussion über die Symmetrie der Wortassoziation beim Menschen statt. Der Symmetriebegriff bezieht sich hierbei auf die Stärke der Assoziation zwischen zwei oder mehreren Worten. Sind die Worte in etwa gleich stark miteinander assoziiert, so kann von einer symmetrischen Wortassoziation gesprochen werden. Gibt es Unterschiede in der Assoziationsstärke, so geht man von einer asymmetrischen Beziehung aus. Von großer Bedeutung ist in diesem Zusammenhang der Begriff der Assoziationsrichtung. Bei jedem Wortpaar ($wort_2$, $wort_1$) sind die zwei Assoziationsrichtungen zu unterscheiden: ($wort_1$ → $wort_2$) und ($wort_2$ → $wort_1$). Erstgenanntes Wort entspricht in dieser Schreibweise dem Stimuluswort und letztgenanntes Wort der assoziativen Antwort. Tauscht man das Stimuluswort gegen die assoziative Antwort, so erhält man die Ergebnisse für die gegengerichtete Assoziation. Aus dem Vergleich zwischen den Stärken der Assoziationsrichtungen ergeben sich Hinweise auf Symmetrie oder Asymmetrie des vorliegenden Wortpaares, wobei gilt: Je größer die Differenz in der Bewertung der unterschiedlichen Richtungen, desto stärker die Hinweise auf eine asymmetrische Wortassoziation.
Ein fiktives Beispiel soll im Folgenden zum Verständnis der Assoziationsrichtungen und der Unterscheidung zwischen symmetrischen und asymmetrischen Wortassoziationen beitragen. Würde man Menschen in unseren Breitengraden die Frage stellen, welches Wort sie mit dem Wort ‘Mango‘ assoziieren, so würde vermutlich ein erheblicher Prozentsatz der Befragten darauf die assoziative Antwort ‘Obst‘ geben. Dies entspricht einem Versuchsaufbau bei den FATs, in dem ‘Mango‘ das Stimuluswort und ‘Obst‘ die vermutete assoziative Antwort ist. Kehrt man die Frage um und bittet die Probanden um ihre erste Assoziation zu dem Begriff ‘Obst‘, so würde die Antwort höchstwahrscheinlich nur bei einem sehr geringen Prozentsatz ‘Mango‘ lauten. Es gibt folglich Wortpaare, deren Beziehung nicht symmetrisch ist, da die Assoziation in die eine Richtung (Mango → Obst) deutlich stärker zu sein scheint als die Assoziation in der umgekehrten Richtung (Obst → Mango). An dieser Stelle verlassen wir das fiktive Beispiel um uns anzuschauen, welche anderen Autoren sich mit dieser Thematik auseinandergesetzt haben und zu welchen Ergebnissen sie gekommen sind.
Michelbacher et al. untersuchen in [59] die Zusammenhänge zwischen syntagmatischen, paradigmatischen sowie symmetrischen und asymmetrischen Beziehungen und geben Beispiele für die verschiedenen Kategorien. Die folgende Tabelle 5 zeigt exemplarische Wortpaare der Schnittmengen.

Tabelle 5. Beispiele für unterschiedliche Wortbeziehungen [59]

	Paradigmatisch	syntagmatisch
Symmetrisch	(bad, good)	(epileptic, seizure)
Asymmetrisch	(bird, canary)	(christmas, decorations)

Für die charakteristischen Beispiele von asymmetrischen und symmetrischen Beziehungen zwischen Worten greifen die Autoren auf die bereits zuvor Erwähnte und von Nelson et al. zusammengestellten Wortassoziationsnormen [53] zurück. Die Teilnehmer hatten dabei stets

die Aufgabe auf ein ihnen präsentiertes Reiz- oder Stimuluswort intuitiv mit dem Begriff zu antworten, den sie mit diesem assoziieren.
Als Indikator für die Assoziationsstärke in die unterschiedlichen Richtungen (z.B. (bad → good) vs. (good → bad)) wird der Prozentsatz der Probanden gewertet, die auf den jeweiligen Stimulus die entsprechende Antwort gab. Letztgenanntes Wortpaar (good, bad) wird in [59] als Beispiel für eine symmetrische Beziehung aufgeführt, wie den in [53] veröffentlichten Ergebnissen zu entnehmen ist.

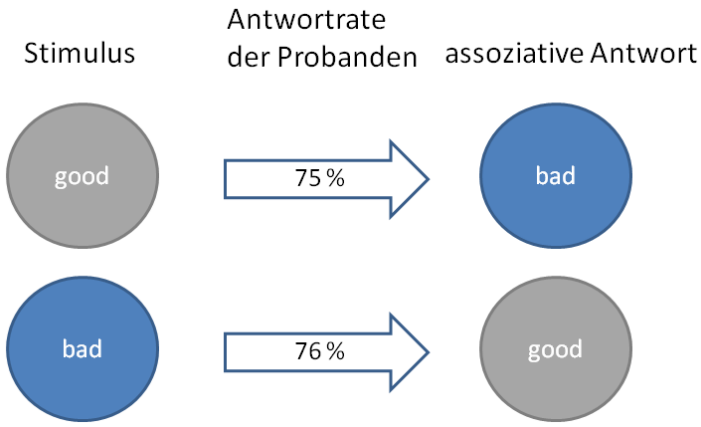


Abbildung 3. Beispiel für eine symmetrische Wortassoziation

Abbildung 3 visualisiert die Ergebnisse dieses Wortpaares und gibt so ein Beispiel für eine symmetrische Wortassoziation. Auf den präsentierten Stimulus ‘good‘ antworteten 75% der Befragten mit ‘bad‘. Betrachtet man die Ergebnisse der umgekehrten Fragestellung, bei der eine assoziative Antwort auf den Stimulus ‘bad‘ gefordert war, stellte sich heraus, dass 76% der Teilnehmer mit ‘good‘ antworteten. Folglich wird die Assoziation für das Wortpaar (good, bad) als symmetrisch bezeichnet, da die Assoziationsstärke in beide Richtungen ungefähr gleich ist.

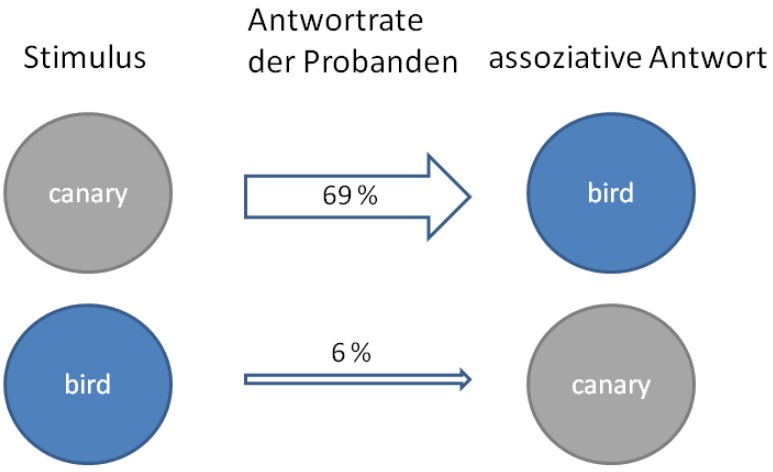


Abbildung 4. Beispiel für eine asymmetrische Wortassoziation

Als Wortpaar mit starken Tendenzen zu einer asymmetrischen Beziehung wurde (bird, canary) identifiziert. Für den Stimulus ‘canary‘ gaben 69% der Versuchspersonen ‘bird‘ als assoziative Antwort. Umgekehrt antworteten allerdings lediglich 6% der Teilnehmer mit ‘canary‘ auf den Stimulus ‘bird‘. Der signifikante Unterschied in den Antwortraten lässt den Schluss zu, dass es sich bei dieser Relation um eine asymmetrische Beziehung handelt. Ohne genaue Raten anzugeben, wie groß der Gesamtanteil an asymmetrischen Wortassoziationen in ihren Tests ist, folgern die Autoren, dass prinzipiell asymmetrische Effekte in der menschlichen Wortassoziation verankert sind [59].
Zu einem ähnlichen Schluss kommen Steyvers et al., indem sie ausführen:

> “[…] *in the norms, the associative strengths* […] *are often highly asymmetric where the associative strength in one direction is strong while it is weak or zero in the other direction.*” [60]

Um die beschriebenen Ergebnisse aus der Literatur und die Hinweise auf Asymmetrie in der menschlichen Wortassoziation zu überprüfen, wurde eine entsprechend ausgelegte Studie durchgeführt. Die Ergebnisse werden im nächten Abschnitt geschildert.

2.4.2.1 Testreihen zur menschlichen Wortassoziation

Die zuvor beschriebenen Ergebnisse der Literaturrecherche lassen bereits erkennen, dass die menschliche Wortassoziation in der Tendenz asymmetrische Züge aufweist. Um diese Resultate zu überprüfen, wurden eigens zu diesem Zweck mehrere Testreihen mit menschlichen Probanden durchgeführt. Der Aufbau und die Durchführung dieser Experimente wurden im Speziellen für die Analyse der Charakteristik der menschlichen Wortassoziation entwickelt, wobei ermittelt werden soll, inwieweit die Assoziationen zwischen Worten symmetrisch oder asymmetrisch sind.
Im Rahmen der Versuche wurden zwei Testreihen mit jeweils 20 Probanden durchgeführt. In Testreihe 1 mussten die Assoziationen zwischen 14 Worten bewertet werden, Testreihe 2 hat einen Umfang von 18 Worten. Die Anzahl der zu bewertenden Wortassoziationen ergibt sich aus der Wortmenge. Je Testreihe wurde den Probanden jedes mögliche Wortpaar genau zweimal vorgelegt. Das doppelte Abfragen der Wortpaare liegt begründet in den unterschiedlichen Assoziationsrichtungen, die zu bewerten waren, denn für die Fokussierung auf asymmetrische Aspekte der menschlichen Wortassoziation ist die Reihenfolge der Worte von entscheidender Bedeutung. Bei der Anordnung ($wort_1 \rightarrow wort_2$) bildet $wort_1$ das Reizwort und die an die Testpersonen gestellte Frage lautet: Wie stark assoziieren Sie $wort_2$ mit $wort_1$. Bei geänderter Reihenfolge ($wort_2 \rightarrow wort_1$) kehrt sich auch die Fragestellung entsprechend um und es ist gefragt, wie stark ist Wort $wort_1$ mit Wort $wort_2$ assoziiert. Interessant für die Untersuchung der (A)Symmetrie sind hier die Unterschiede in der Bewertung für beide Richtungen (($wort_1 \rightarrow wort_2$) vs. ($wort_2 \rightarrow wort_1$)). Ähnliche Werte in beiden Richtungen weisen auf eine tendenziell symmetrische Assoziation hin, wobei signifikante Unterschiede auf eine asymmetrische Assoziation hindeuten. In der Summe wurden jedem Teilnehmer in

Testreihe 1 genau 182 Wortassoziationen (196 minus die 14 Wortpaare bei denen beide Worte identisch gewesen wären) zur Bewertung vorgelegt und in Testreihe 2 exakt 306.
Bei der Durchführung der Tests wurde wie folgt vorgegangen: Den Testpersonen wurden sukzessive sämtliche Wortpaare in unterschiedlicher Reihenfolge und mit zeitlichem Abstand präsentiert. Aufgabe der Probanden war es, die Stärke der Assoziation zwischen den Worten auf einer Skala von 1 (starke Assoziation), über 2 (schwache Assoziation) bis zu 3 (keine Assoziation) zu bewerten.
In den folgenden Tabellen werden die Ergebnisse der Testreihen detailliert aufgeführt. Die Ziffern in der Kopfzeile und der Führungsspalte der Tabellen beinhalten jeweils die Ziffern von 1 – 14 in Testreihe 1 bzw. in Testreihe 2 von 1 – 18. Jede Ziffer pro Testreihe steht für ein Wort. Zu lesen sind die Tabellen wie folgt: Die Ziffern der Führungsspalte stehen für das Stimuluswort, die nachfolgenden Spalten beinhalten die Assoziationen zu den entsprechenden assoziativen Antworten. Die numerischen Werte entsprechen der durchschnittlichen Assoziationsstärke zwischen den Worten, evaluiert durch unsere Probanden.

Tabelle 6. Ergebnisse Testreihe 1 [61]

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1		2.00	2.30	1.20	2.70	2.90	1.05	1.30	2.65	2.05	2.00	2.35	1.70	2.40
2	2.25		2.45	2.40	2.30	2.40	2.60	2.25	2.35	2.65	2.75	2.60	2.25	2.85
3	2.05	2.45		1.75	2.75	2.90	2.25	1.80	2.90	2.70	2.70	2.65	2.75	2.15
4	1.15	2.25	1.90		2.50	2.45	1.35	1.30	2.65	2.75	2.70	2.40	2.30	2.60
5	2.70	2.10	2.90	2.65		1.15	2.95	2.60	1.30	1.20	1.40	2.30	1.90	2.70
6	2.70	2.25	2.85	2.65	1.05		2.75	2.50	1.10	1.25	1.55	2.50	2.05	2.75
7	1.15	2.40	2.40	1.60	2.70	2.85		1.75	2.65	2.20	2.40	2.70	2.70	2.10
8	1.20	2.25	1.90	1.20	2.40	2.50	1.80		2.65	2.55	2.75	2.10	2.65	2.15
9	2.65	2.35	2.80	2.65	1.20	1.20	2.75	2.70		1.40	1.40	2.45	2.15	2.85
10	2.00	2.70	2.60	2.80	1.25	1.20	2.30	2.60	1.50		1.40	2.85	2.80	2.55
11	2.10	2.65	2.70	2.70	1.40	1.25	2.35	2.80	1.45	1.20		2.80	2.80	2.75
12	2.40	2.60	2.80	2.10	2.30	2.40	2.65	2.15	2.25	2.80	2.85		2.75	2.85
13	1.70	2.20	2.90	2.45	1.95	2.20	2.70	2.60	2.05	2.75	2.85	2.75		2.65
14	2.30	2.70	2.25	2.35	2.95	2.65	1.95	2.60	2.95	2.70	2.65	3.00	2.60	

Ein Beispiel verdeutlicht Vorgehensweise der Experimente und Darstellung der Ergebnisse. Dazu soll exemplarisch die Assoziation zwischen $wort_1$ und $wort_2$ in Testreihe 1 analysiert werden. In Tabelle 6 (Zeile 2, Spalte 3; grau hinterlegt) ist die durchschnittliche Bewertung der Probanden mit 2.00 angegeben. Dieser Wert repräsentiert die gerichtete Assoziation ($wort_1 \rightarrow wort_2$), und beantwortet die Frage, wie stark $wort_2$ assoziiert ist mit $wort_1$. In diesem Fall fungiert $wort_1$ als das Stimuluswort und $wort_2$ als assoziative Antwort. Die umgekehrte Richtung und die Assoziation ($wort_2 \rightarrow wort_1$) in (Tabelle 6; Zeile 3, Spalte 2) wird im Durchschnitt mit 2.25 bewertet. Der Unterschied in der Bewertung der Richtungen beträgt in diesem Fall 0.25 (Tabelle 7; Zeile 3, Spalte 2), was in der Tendenz auf eine asymmetrische Assoziation hindeutet. Bei der Untersuchung der Charakteristik der Wortassoziationen in Bezug auf die Symmetrie sind eben diese Differenzen von großer Bedeutung. Aus diesem Grund werden in Tabelle 7 diese Differenzen der Assoziationsrichtungen berechnet und tabellarisch visualisiert. Der Übersichtlichkeit geschuldet ist die zusätzliche Kennzeichnung der Tabellenfelder in unterschiedliche Graustufen.

Tabelle 7. Unterschiede in der Assoziationsrichtung Testreihe 1 [61]

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1		-0.25	0.25	0.05	0.00	0.20	-0.10	0.10	0.00	0.05	-0.10	-0.05	0.00	0.10
2	0.25		0.00	0.15	0.20	0.15	0.20	0.00	0.00	-0.05	0.10	0.00	0.05	0.15
3	-0.25	0.00		-0.15	-0.15	0.05	-0.15	-0.10	0.10	0.10	0.00	-0.15	-0.15	-0.10
4	-0.05	-0.15	0.15		-0.15	-0.20	-0.25	0.10	0.00	-0.05	0.00	0.30	-0.15	0.25
5	0.00	-0.20	0.15	0.15		0.10	0.25	0.20	0.10	-0.05	0.00	0.00	-0.05	-0.25
6	-0.20	-0.15	-0.05	0.20	-0.10		-0.10	0.00	-0.10	0.05	0.30	0.10	-0.15	0.10
7	0.10	-0.20	0.15	0.25	-0.25	0.10		-0.05	-0.10	-0.10	0.05	0.05	0.00	0.15
8	-0.10	0.00	0.10	-0.10	-0.20	0.00	0.05		-0.05	-0.05	-0.05	-0.05	0.05	-0.45
9	0.00	0.00	-0.10	0.00	-0.10	0.10	0.10	0.05		-0.10	-0.05	0.20	0.10	-0.10
10	-0.05	0.05	-0.10	0.05	0.05	-0.05	0.10	0.05	0.10		0.20	0.05	0.05	-0.15
11	0.10	-0.10	0.00	0.00	0.00	-0.30	-0.05	0.05	0.05	-0.20		-0.05	-0.05	0.10
12	0.05	0.00	0.15	-0.30	0.00	-0.10	-0.05	0.05	-0.20	-0.05	0.05		0.00	-0.15
13	0.00	-0.05	0.15	0.15	0.05	0.15	0.00	-0.05	-0.10	-0.05	0.05	0.00		0.05
14	-0.10	-0.15	0.10	-0.25	0.25	-0.10	-0.15	0.45	0.10	0.15	-0.10	0.15	-0.05	

Den Schattierungen in Tabelle 7 liegen die folgenden Einteilungen zugrunde: Keine Unterschiede in der Bewertung der Richtungen bleiben weiß hinterlegt. Die Intervalle [0.01, 0.1]; (0.1, 0.2]; (0.2, 0.3] und (0.3, ∞] sind in immer dunkler werdenden Graustufen hinterlegt, so dass an der Stärke der Schattierung der Grad der Asymmetrie abgelesen werden kann. Je dunkler die Hinterlegung, desto stärker die Asymmetrie in der vorliegenden Assoziation. Die Vorzeichen in Tabelle 7 zeigen im Detail, in welcher Richtung die Assoziation stärker eingeschätzt wurde. Zum jetzigen Zeitpunkt und zur Feststellung der Charakteristik der menschlichen Wortassoziation genügt die Betrachtung der Absolutbeträge zur Erkennung von asymmetrischen Tendenzen.
Die Ergebnisse der Testreihe 2 bezüglich der durchschnittlichen Bewertung der Probanden sind in Tabelle 8 dargestellt.

Tabelle 8. Ergebnisse Testreihe 2

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1		1.50	2.20	2.45	2.25	2.60	2.65	2.55	2.40	2.65	2.55	2.10	2.85	2.10	2.25	2.60	2.60	2.70
2	1.35		2.00	2.20	2.15	2.50	2.45	2.50	2.20	2.55	2.55	2.05	2.80	2.25	2.15	2.75	2.80	2.80
3	2.25	2.45		1.60	1.55	1.70	1.75	1.55	2.45	2.65	2.55	2.80	2.30	2.80	2.60	1.55	1.85	1.85
4	2.15	2.35	1.70		1.00	1.00	1.00	1.00	2.95	2.50	2.75	2.90	2.25	2.70	2.90	2.30	2.40	2.40
5	2.35	2.30	1.75	1.00		1.00	1.00	1.00	2.95	2.40	2.60	3.00	2.45	2.85	2.85	2.40	2.45	2.30
6	2.45	2.70	1.25	1.05	1.00		1.05	1.00	2.90	2.45	2.60	3.00	2.45	2.85	2.95	1.90	2.20	1.90
7	2.55	2.55	1.55	1.00	1.10	1.10		1.05	3.00	2.45	2.60	3.00	2.35	2.75	3.00	2.25	2.25	2.40
8	2.25	2.35	1.35	1.05	1.05	1.00	1.00		2.85	2.15	2.70	2.90	2.30	2.60	2.90	2.25	2.40	2.40
9	2.35	2.40	2.80	2.95	2.85	2.95	2.95	2.95		2.55	2.70	1.10	2.80	2.80	1.00	1.80	2.10	2.20
10	2.75	2.65	2.45	2.60	2.40	2.65	2.15	2.15	2.75		2.65	2.55	2.00	2.65	2.60	2.15	2.20	2.40
11	2.50	2.50	2.50	2.75	2.45	2.70	2.35	2.65	2.55	2.45		2.55	2.85	2.55	2.60	2.35	2.35	2.80
12	2.15	2.20	2.90	2.90	2.95	2.95	2.95	2.95	1.05	2.45	2.60		2.85	2.75	1.00	1.70	2.10	2.10
13	2.85	2.85	2.15	2.20	2.15	2.35	2.10	2.25	2.60	2.05	2.90	2.80		2.80	2.85	2.25	2.30	2.25
14	2.35	2.35	2.70	2.70	2.75	2.80	2.75	2.70	2.90	2.50	2.65	2.75	2.85		2.85	2.65	2.65	2.75
15	2.10	2.40	2.80	2.95	2.95	2.90	2.95	2.95	1.05	2.50	2.50	1.00	2.80	2.85		1.50	1.95	2.05
16	2.65	2.65	1.70	2.35	2.35	2.15	2.55	2.30	1.95	2.10	2.70	1.60	2.20	2.45	1.30		1.00	1.20
17	2.65	2.80	1.75	2.70	2.50	2.50	2.50	2.40	1.95	2.40	2.60	2.20	2.35	2.75	2.00	1.15		1.20
18	2.70	2.75	1.95	2.30	2.45	2.25	2.60	2.40	2.35	2.40	2.60	2.05	2.40	2.70	1.90	1.20	1.45	

Wie den Werten zu entnehmen ist, handelt es sich bei dieser Testreihe nicht um eine Erweiterung von Testreihe 1, sondern um einen eigenständiges Experiment. Die verwendeten und zu bewertenden Wortpaare differieren von denen in der vorangegangenen Testreihe.

Bezüglich der Darstellungsweise bedient sich Tabelle 9 der identischen Einteilung und Schattierung des vorangegangenen Beispiels.

Tabelle 9. Unterschiede in der Assoziationsrichtung Testreihe 2

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1		0.15	-0.05	0.30	-0.10	0.15	0.10	0.30	0.05	-0.10	0.05	-0.05	0.00	-0.25	0.15	-0.05	-0.05	0.00
2	-0.15		-0.45	-0.15	-0.15	-0.20	-0.10	0.15	-0.20	-0.10	0.05	-0.15	-0.05	-0.10	-0.25	0.10	0.00	0.05
3	0.05	0.45		-0.10	-0.20	0.45	0.20	0.20	-0.35	0.20	0.05	-0.10	0.15	0.10	-0.20	-0.15	0.10	-0.10
4	-0.30	0.15	0.10		0.00	-0.05	0.00	-0.05	0.00	-0.10	0.00	0.00	0.05	0.00	-0.05	-0.05	-0.30	0.10
5	0.10	0.15	0.20	0.00		0.00	-0.10	-0.05	0.10	0.00	0.15	0.05	0.30	0.10	-0.10	0.05	-0.05	-0.15
6	-0.15	0.20	-0.45	0.05	0.00		-0.05	0.00	-0.05	-0.20	-0.10	0.05	0.10	0.05	0.05	-0.25	-0.30	-0.35
7	-0.10	0.10	-0.20	0.00	0.10	0.05		0.05	0.05	0.30	0.25	0.05	0.25	0.00	0.05	-0.30	-0.25	-0.20
8	-0.30	-0.15	-0.20	0.05	0.05	0.00	-0.05		-0.10	0.00	0.05	-0.05	0.05	-0.10	-0.05	-0.05	0.00	0.00
9	-0.05	0.20	0.35	0.00	-0.10	0.05	-0.05	0.10		-0.20	0.15	0.05	0.20	-0.10	-0.05	-0.15	0.15	-0.15
10	0.10	0.10	-0.20	0.10	0.00	0.20	-0.30	0.00	0.20		0.20	0.10	-0.05	0.15	0.10	0.05	-0.20	0.00
11	-0.05	-0.05	-0.05	0.00	-0.15	0.10	-0.25	-0.05	-0.15	-0.20		-0.05	-0.05	-0.10	0.10	-0.35	-0.25	0.20
12	0.05	0.15	0.10	0.00	-0.05	-0.05	-0.05	0.05	-0.05	-0.10	0.05		0.05	0.00	0.00	0.10	-0.10	0.05
13	0.00	0.05	-0.15	-0.05	-0.30	-0.10	-0.25	-0.05	-0.20	0.05	0.05	-0.05		-0.05	0.05	0.05	-0.05	-0.15
14	0.25	0.10	-0.10	0.00	-0.10	-0.05	0.00	0.10	0.10	-0.15	0.10	0.00	0.05		0.00	0.20	-0.10	0.05
15	-0.15	0.25	0.20	0.05	0.10	-0.05	-0.05	0.05	0.05	-0.10	-0.10	0.00	-0.05	0.00		0.20	-0.05	0.15
16	0.05	-0.10	0.15	0.05	-0.05	0.25	0.30	0.05	0.15	-0.05	0.35	-0.10	-0.05	-0.20	-0.20		-0.15	0.00
17	0.05	0.00	-0.10	0.30	0.05	0.30	0.25	0.00	-0.15	0.20	0.25	0.10	0.05	0.10	0.05	0.15		-0.25
18	0.00	-0.05	0.10	-0.10	0.15	0.35	0.20	0.00	0.15	0.00	-0.20	-0.05	0.15	-0.05	-0.15	0.00	0.25	

Als Ergebnis der durchgeführten Testreihen kann geschlussfolgert werden, dass die Schlüsse aus der Literatur bestätigt werden konnten. Im Speziellen zeigen Tabelle 7 und Tabelle 9, dass ein wesentlicher Teil der menschlichen Wortassoziationen asymmetrische Tendenzen aufweist. Für eine technische Realisierung der menschlichen Wortassoziation bedeutet das wiederum, dass dieses charakteristische Merkmal in einer Implementierung zu berücksichtigen ist.

3 Simulation von Wortassoziationen

Die vorliegende Arbeit und das entwickelte Assoziationsberechnungsverfahren ist nicht das erste Verfahren seiner Art. In der Literatur existieren andere Ansätze, die ebenfalls Wortassoziationen auf Basis von Textsammlungen zu bestimmen versuchen. Um die Unterschiede zwischen diesen und dem hier vorgestellten Verfahren herauszuarbeiten, werden in diesem Kapitel die bekanntesten Verfahren definiert und erläutert. Des Weiteren werden die hier vorgestellten Verfahren in den vergleichenden Fallstudien (siehe Kapitel 5) miteinander in Konkurrenz treten, um die Leistungsfähigkeit bei verschiedenen Umgebungsparametern zu ermitteln. Wichtigste Umgebungsparameter in diesem Zusammenhang sind die Werte für die Frequenz der Worte in der Textsammlung, sowie das gemeinsame Vorkommen von Worten innerhalb eines definierten Textfensters (Kookkurrenz).
Wie die folgenden Abschnitte zeigen, bestehen Ähnlichkeiten zwischen den beschriebenen Verfahren. Dies ist auch darin begründet, dass diese Verfahren auf den gleichen statistischen Werten fundieren. In erster Linie sind das Informationen über das gemeinsame Vorkommen von Worten, ihrer Frequenz in der Textsammlung, sowie das erwartete gemeinsame Vorkommen aufgrund der Worthäufigkeiten. Ergebnis der Assoziationsberechnungsverfahren ist ein numerischer Wert, der die Stärke der Assoziation zwischen zwei Worten ausdrückt. Unterschiede zwischen den Verfahren bilden lediglich die Interpretation der berechneten Assoziationswerte, denn bei manchen Verfahren bedeutet ein hoher Wert eine starke Assoziation, bei anderen Verfahren bedeutet ein niedriger Wert eine starke Assoziation. Evert definiert die Assoziationsberechnungsverfahren allgemein in dem er ausführt:

"*An association measure is a formula that computes an association score from the frequency information in a pair type´s contingency table.*" [40]

Analog zu vorausgegangenen Erläuterungen dient der beschriebene 'association score' für Evert als Indikator für die Stärke der Assoziation [40]. Mit der Berechnung der Assoziationsstärke stellt sich die Frage der Interpretation sowie nach dem Nutzen dieser Ergebnisse. Aufgrund der unterschiedlichen Berechnungsweisen der verschiedenen Assoziationsmaße können die Ergebnisse unterschiedlicher Verfahren nicht unmittelbar miteinander verglichen werden. Aus diesem Grund hat sich in der Literatur ein Ranking Prinzip durchgesetzt. Hierbei werden die assoziierten Worte für das jeweilige Stimuluswort absteigend nach ihrer Stärke sortiert. Das nach dem jeweiligen Verfahren am stärksten mit dem Stimuluswort assoziierte Wort bildet den Anfang einer absteigend nach Assoziationsstärke sortierten Liste. Die absoluten Ergebnisse der Berechnungen der einzelnen Verfahren, sowie unter Umständen vorhandene Unterschiede in den Bewertungskriterien können so ausgeglichen werden. Ergebnis bei dieser Vorgehensweise sind Assoziationslisten, welche sich unabhängig vom verwendeten Verfahren vergleichen lassen.
Sieht man von der Vergleichbarkeit zwischen Verfahren ab, können die numerischen Resultate der Einzelverfahren zusätzlich dazu verwendet werden, um die Assoziationsstärke

zwischen ausgewählten Worten direkt zu berechnen. Ersteres Verfahren, basierend auf dem Ranking der Assoziationsworte, kommt auch in den vergleichenden Fallstudien der vorliegenden Arbeit zur Anwendung und dient der Evaluation der Ergebnisse unterschiedlicher Verfahren. Ein Anwendungsbeispiel für die Verwendung der Assoziationsstärke zwischen Worten kommt unter anderem in dem in Kapitel 6 vorgestellten Verfahren zur Erkennung von Themenstrukturen in Texten (siehe Kapitel 6.1) sowie bei der assoziativen Suche auf unternehmensinternen Textdaten (siehe Kapitel 6.4) zum Einsatz.

3.1 Definition der Simulationsparameter

Im Folgenden werden zunächst die Einflussparameter der Verfahren einheitlich definiert und eine Einteilung der Verfahren in Hauptkategorien wird vorgegeben.
Die überwiegende Anzahl an Assoziationsberechnungsverfahren nimmt einen Vergleich zwischen den beobachteten Frequenzen in der Textsammlung (‘observed frequencies‘) und den erwarteten Frequenzen (‘expected frequencies‘) vor. Unter beobachteten Frequenzen wird in diesem Zusammenhang die tatsächliche Anzahl der Worte in der Textsammlung verstanden. Die erwarteten Frequenzen beschreiben das auf Basis der beobachteten Frequenzen zur erwartende gemeinsame Vorkommen zweier Worte in einer Textsammlung. Da die Wahrscheinlichkeit eines gemeinsamen Vorkommens zweier hochfrequenter Worte in einer Textsammlung deutlich höher ist als die zweier Worte, die nur selten in der Textsammlung verwendet werden, stellen die erwarteten Frequenzen einen aussagekräftigen Parameter für die Assoziationsberechnungsverfahren dar. Evert führt in [40] eine Darstellung ein, die auch in dieser Arbeit Anwendung findet. Folgende Tabelle 10 beschreibt die beobachteten Frequenzen.

Tabelle 10. Beobachtete Frequenzen [40]

	$V = v$	$V \neq v$	
$U = u$	O_{11}	O_{12}	$O_{11} + O_{12} = \mathbf{Z_1}$
$U \neq u$	O_{21}	O_{22}	$O_{21} + O_{22} = \mathbf{Z_2}$
	$O_{11} + O_{21} = \mathbf{S_1}$	$O_{12} + O_{22} = \mathbf{S_2}$	$O_{11} + O_{12} + O_{21} + O_{22} = \mathbf{N}$

Die Parameter ‘U‘ und ‘V‘ stehen in obiger Tabelle 10 stellvertretend für beliebige Teilmengen von Worten eines Textkorpus. Beispielsweise könnte U die Menge der Adjektive und V die Menge der Substantive darstellen. Die Kleinbuchstaben ‘u‘ bzw. ‘v‘ repräsentieren eine Einzelinstanz aus dieser Menge, sprich in obigem Beispiel ein Adjektiv respektive Substantiv aus der Textsammlung. ‘O_{11}‘ beschreibt die Anzahl des gemeinsamen Vorkommens von u und v, innerhalb eines definierten Textfensters. ‘O_{12}‘ zeigt an, wie häufig u vorkommt, ohne dass in dem umgebenden Textfenster v vorkommt, wobei ‘O_{21}‘ den umgekehrten Fall darstellt. ‘O_{22}‘ steht für die Anzahl an Wortpaaren in denen weder u noch v zu beobachten sind. ‘S_1‘ und ‘S_2‘ sind die Spaltensummen, und ‘Z_1‘ und ‘Z_2‘ stehen für die Zeilensummen der berechneten Parameter. Die Gesamtanzahl der beobachteten Wortpaarungen wird in ‘N‘ ausgedrückt.

Ein Beispiel aus [58] soll zeigen, wie eine solche Tabelle in der Praxis aussieht. Das Adjektiv / Nomen Konstrukt ‘new companies‘ dient dabei als das zu analysierende Wortpaar. Tabelle 11 zeigt, dass die Kombination ‘new companies‘ genau 8 mal in der Textsammlung beobachtet wurde, wohingegen ‘new‘ mit einem anderen Substantiv (z.B. ‘machines‘) 15820 vorkam. Der umgekehrte Fall, in dem ‘companies‘ unmittelbar nach einem anderen Adjektiv als ‘new‘ benutzt wurde (z.B. ‘old companies‘), wurde in 4667 Fällen beobachtet.

Tabelle 11. Beispieltabelle für beobachtete Frequenzen [58]

	v = companies	v ≠ companies	
u = new	8 (new companies)	15820 (e.g., new machines)	8 + 15820 = 15828 **(Z_1)**
u ≠ new	4667 (e.g., old companies)	14287173 (e.g., old machines)	4667 + 14287173 = 14291840 **(Z_2)**
	8 + 4667 = 4675 **(S_1)**	15820 + 14287173 = 14302993 **(S_2)**	8 + 15820 + 4667 + 14287173 = 14307668 **(N)**

Tabelle 11 zeigt, dass im gesamten von [58] verwendeten Textkorpus 14287173 Adjektiv / Substantiv Kombinationen gezählt wurden, in denen weder ‘new‘ noch ‘companies‘ vorkamen. Die Zeilen bzw. Spaltensummen (Z_1, Z_2 bzw. S_1, S_2 in Tabelle 11) sowie die Gesamtanzahl aller Adjektiv / Substantiv Paare (N) wurden ebenfalls berechnet. Diese werden insbesondere benötigt, wenn im Folgenden die zu erwartenden Frequenzen berechnet werden. Tabelle 12 enthält die Berechnungsvorschriften für die erwarteten Frequenzen. Die zugehörigen Parameter entstammen Tabelle 10.

Tabelle 12. Erwartete Frequenzen [40]

	V = v	V ≠ v
U = u	$E_{11} = \frac{Z_1 S_1}{N}$	$E_{12} = \frac{Z_1 S_2}{N}$
U ≠ u	$E_{21} = \frac{Z_2 S_1}{N}$	$E_{22} = \frac{Z_2 S_2}{N}$

Um das Beispiel aus Tabelle 11 und das Wortpaar ‘new companies‘ wieder aufzunehmen, wird exemplarisch die erwartete Frequenz des gemeinsamen Vorkommens dieser beiden Worte (E_{11} in Tabelle 12) berechnet in :

$$E_{11} = \frac{Z_1 S_1}{N} = \frac{15828 * 4675}{14307668} \approx 5{,}17$$

Das bedeuted, dass basierend auf den Werten aus Tabelle 11 und einem Text von der Größe des verwendeten Korpus, sowie der Annahme, dass die beiden Begriffe vollständig unabhängig voneinander darin vorkommen, die erwartete Frequenz des gemeinsamen Vorkommens von ‘new companies‘ bei 5,17 liegt.
Die in Tabelle 10 und Tabelle 12 definierten Parameter bilden die Grundlage für die im Folgenden definierten Assoziationsberechnungsverfahren, die in den vergleichenden Fallstudien in Kapitel 5 implementiert wurden.

3.2 Kategorisierung der Simulationsverfahren

Wie bereits in Kapitel 2 herausgearbeitet, enthält die menschliche Wortassoziation sowohl symmetrische als auch asymmetrische Züge. Die im Folgenden vorgestellten Assoziationsberechnungsverfahren orientieren sich an dieser Kategorisierung. Die Verfahren werden eingeteilt in symmetrische und asymmetrische Verfahren. Diese grenzen sich bezüglich der Unterscheidung zwischen den Assoziationsrichtungen voneinander ab.
Die symmetrischen Verfahren berechnen für jedes Wortpaar einen numerischen Wert, der die Stärke der Assoziation ausdrückt. Diese Assoziationsstärke repräsentiert die Beziehung der Worte ohne Einbeziehung der Richtung. Asymmetrische Verfahren hingegen berechnen die Assoziationsstärke in Abhängigkeit der Assoziationsrichtung, das bedeutet für jede Richtung separat. Exemplarisch soll an dieser Stelle auf Abbildung 4 verwiesen werden. Das dargestellte Wortpaar (canary, bird) bildet ein Beispiel für ein stark asymmetrisches Worttupel. Da die asymmetrischen Verfahren die Werte für jede Richtung einzeln berechnen, besteht die Möglichkeit einer präzisen Abbildung der tatsächlichen Wortassoziation. Symmetrische Verfahren sind hier aufgrund ihrer Konzeption auf einen Einheitswert beschränkt, was eine Einschränkung der Genauigkeit zur Folge haben kann. Trotz dieser Beschränkungen ist die Mehrheit der in der Vergangenheit entwickelten Verfahren symmetrisch konzipiert. Dieser starken Tendenz zu symmetrischen Verfahren wird auch in dieser Arbeit Rechnung getragen, was sich in der Auswahl der getesteten Verfahren niederschlägt. Die meisten getesteten Verfahren sind symmetrisch konzipiert, wobei diese durch ein rein asymmetrisches Verfahren und die hybride Eigenentwicklung ergänzt werden.
Bei der Vorstellung der Verfahren werden zunächst die symmetrischen und im Anschluß die übrigen Verfahren vorgestellt.

3.2.1 Verfahren zur symmetrischen Assoziationsberechnung

Abbildung 5 deutet die Konzeption der symmetrischen Assoziationsberechnungsverfahren an. Deutlich wird die bereits angedeutete Herangehensweise, in der für jedes Wortpaar ($wort_1$, $wort_2$) ein numerischer Wert berechnet wird, der beide Assoziationsrichtungen repräsentiert.

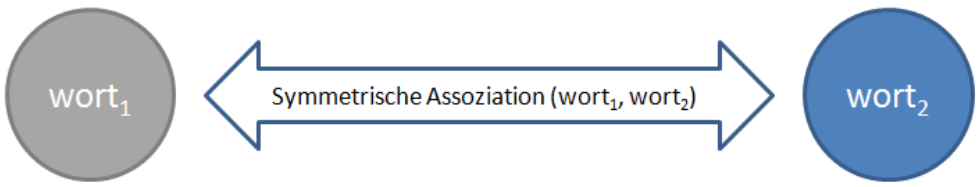


Abbildung 5. Symmetrische Assoziationsberechnung

Bei der Kategorisierung der Simulationsverfahren orientiert sich die vorliegende Arbeit an den in [40] erstellten Kategorien. Der Autor teilt die Assoziationsverfahren in vier übergeordnete Kategorien ein, die im Folgenden näher beschrieben werden.

I. ‘Significance of association group‘ [40]: Die Verfahren dieser Kategorie stellen zunächst die Hypothese auf, dass die extrahierten Wortpaare keine Assoziation besitzen. Ziel ist es im Anschluss, diese These zu widerlegen. Eine weitere Unterteilung dieser Gruppe in ‘likelihood measures‘, ‘exact hypothesis tests‘ und ‘asymptotic hypothesis tests‘ wird von [40] vorgeschlagen. Die in dieser Arbeit implementierten und getesteten Verfahren dieser (Unter-)gruppen sind die folgenden:

- poison-stirling,
- z-score,
- chi-squared,
- t-score,
- log-likelihood

II. ‘Degree of Association group‘ [40]: Die Verfahren dieser Kategorie arbeiten auf den in Tabelle 10 beschriebenen beobachteten Frequenzen und lassen sich weiter unterteilen in die ‘point estimates‘ und ‘conservative estimates‘. Implementiert wurden die folgenden Verfahren zur Assoziationsberechnung:

- liddel,
- dice,
- jaccard,
- g-mean,
- minimum sensivity (MS),
- odds-ratio

III. ‘Information Theory group‘ [40]: Der Unterschied zu den Verfahren der anderen Kategorien liegt in der vergleichenden Analyse von beobachteten (Tabelle 10) und erwarteten Frequenzen (Tabelle 12). Aus dieser Kategorie wurden die folgenden Verfahren getestet:

- pointwise mutual information,
- average mutual information,
- local mutual information

IV. ‘Heuristic Measures‘ [40]: Diese Kategorie fasst diejenigen Methoden zusammen, die aus einer Kombination verschiedener anderer Verfahren bestehen. Unter Umständen ist eine eindeutige Einordnung der Verfahren nicht möglich. Umgesetzt wurden die heuristischen Varianten der ‘mutual information‘:

- MI^2,
- MI^3

Die folgende Tabelle 13 fasst die Formeln zur Berechnung der Assoziationsstärken der oben genannten Verfahren zusammen.

Tabelle 13. Formelsammlung zur symmetrischen Berechnung von Assoziationen

Gruppe	Bezeichnung	Formel	Quelle(n)
I.	poisson-stirling	$O_{11} * (log\ O_{11} - \log\ E_{11} - 1)$	[62]
	z-score	$\frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$	[63], [64]
	chi-squared	$\frac{N(O_{11} - E_{11})^2}{E_{11}E_{22}}$	[65]
	t-score	$\frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$	[66]
	log-likelihood	$2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}}$	[67]
II.	Liddel	$\frac{O_{11}O_{22} - O_{12}O_{21}}{S_1S_2}$	[68]
	Dice	$\frac{2O_{11}}{Z_1 + S_1}$	[69]
	Jaccard	$\frac{O_{11}}{O_{11} + O_{12} + O_{21}}$	[70]
	g-mean	$\frac{O_{11}}{\sqrt{NE_{11}}}$	[40]
	minimum sensivity (MS)	$min\left\{\frac{O_{11}}{Z_1}, \frac{O_{11}}{S_1}\right\}$	[71]
	odds-ratio	$\log \frac{O_{11}O_{22}}{O_{12}O_{21}}$	[72]
III.	pointwise mutual information (PMI)	$\log \frac{O_{11}}{E_{11}}$	[73]
	average mutual information (average MI)	$\sum_{ij} O_{ij} * \log \frac{O_{ij}}{E_{ij}}$	[73]
	local mutual information (local MI)	$O_{11} * \log \frac{O_{11}}{E_{11}}$	[73]
IV.	MI^2	$\log \frac{(O_{11})^2}{E_{11}}$	[73], [74]
	MI^3	$\log \frac{(O_{11})^3}{E_{11}}$	[73], [74]

Die Kategorisierung entspricht der zuvor beschriebenen Gruppeneinteilung. Die vereinheitlichte Grundlage bezogen auf die in den Formeln verwendeten Variablen ist in Tabelle 10 und Tabelle 12 beschrieben. Die Variablen beschränken sich auf die beobachteten (O_{11}, O_{12}, O_{21}, O_{22}) sowie die erwarteten Frequenzen (E_{11}, E_{12}, E_{21}, E_{22}).
Die in obiger Tabelle beschriebenen Assoziationsberechnungsverfahren und deren Formeln liegen der vergleichenden Implementierung in Kapitel 5 zugrunde.

3.2.2 Verfahren zur asymmetrischen Assoziationsberechnung

Neben den zuvor beschriebenen Verfahren zur symmetrischen Assoziationsberechnung wird in dieser Arbeit zusätzlich ein konzeptionell anders ausgerichtetes Verfahren getestet. Abbildung 6 verdeutlicht die grundlegenden Unterschiede zur symmetrischen Assoziationsberechnung. Die Assoziationen zwischen zwei Worten werden in Abhängigkeit von der Assoziationsrichtung bestimmt. Visualisiert sind diese Richtungen in Abbildung 6 durch die Verbindungspfeile zwischen den Worten. Das bedeutet, dass für jedes Wortpaar zwei Assoziationsberechnungen durchgeführt werden müssen, um die Stärke der Assoziation für jede Richtung zu bestimmen.

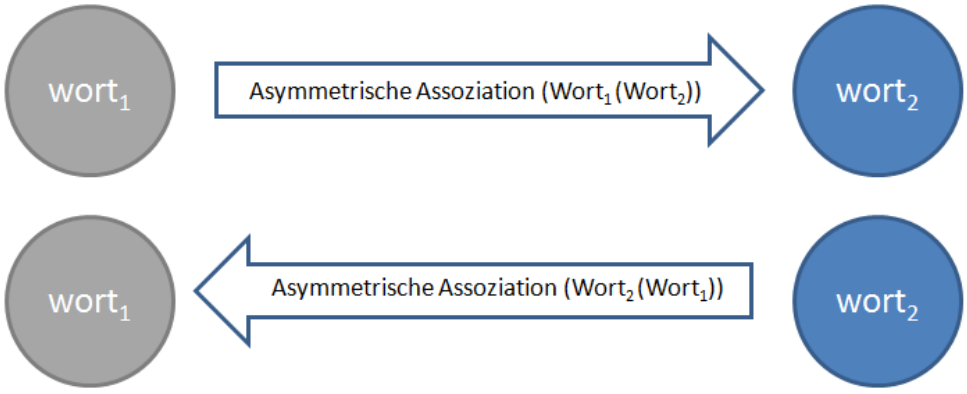


Abbildung 6. Asymmetrische Assoziationsberechnung

Ein Vergleich zwischen Abbildung 5 und Abbildung 6 unterstreicht den Hauptunterschied zwischen symmetrischen und asymmetrischen Assoziationsberechnungsverfahren. Symmetrische Verfahren berechnen für jedes Wortpaar einen Wert, der die Assoziationsstärke ausdrückt. Asymmetrische Verfahren berechnen zwei Werte pro Wortpaar, was einem für jede Richtung entspricht.
Die folgende Formel 2 lässt sich laut Wettler, Rapp und Ferber auf psychologische Lerngesetze zurückführen [47].

Formel 2. Bestimmung der Assoziationsstärke nach Wettler, Rapp und Ferber [75]

$$\tilde{A}(x,y) = \frac{H(x\&y)}{H(y)^{\alpha}}$$

Ã(x,y) steht hier für die Assoziationsstärke zwischen den Worten ‘x‘ und ‘y‘. H(x&y) drückt aus, wie häufig die beiden zu untersuchenden Worte x und y gemeinsam in einem bestimmten Textfenster in der Textsammlung vorkommen. H(y) steht für die Anzahl des Wortes y in der Textsammlung. Rapp führt zu dieser Formel in [47] aus, dass Schätzfehler bei seltenen Wörtern starke Auswirkung auf die berechneten Assoziationsstärken haben, da H(y) im Nenner steht. Aus diesem Grund wurde die in Formel 3 beschriebene Fallunterscheidung eingeführt, die diese negativen Auswirkungen unterdrückt.

Formel 3. Fallunterscheidung für Formel 2 nach Wettler und Rapp **[**76**]**

$$\tilde{A}(x,y) = \begin{cases} \frac{H(x\&y)}{H(y)^{\alpha}} \; for\; H(y) > \beta * Q \\ \frac{H(x\&y)}{(\gamma * Q)} \; for\; H(y) \leq \beta * Q \end{cases}$$

Laut Rapp [47] wurden die besten Ergebnisse mit Formel 2 und der zugehörigen Fallunterscheidung mit den folgenden Parameterwerten erreicht: $\alpha = 0{,}68$, $\beta = \gamma = 0{,}000005$. Q hängt vom verwendeten Textkorpus ab und steht für die Gesamtanzahl der Worte in diesem.

4 CIMAWA – Entwicklung einer textbasierten Assoziationsberechnungs - Methode

Ein unter Linguisten sehr bekannter und häufig zitierter Ausspruch des englischen Sprachwissenschaftlers John Rupert Firth lautet:

"You shall know a word by the company it keeps" [77]

Mit dieser These fasst Firth prägnant zusammen, was bei näherer Betrachtung des Sachverhalts intuitiv erscheint. Sätze einer Sprache, ob gesprochen oder geschrieben, sind keine willkürlich zusammengesetzten Konstrukte, gebildet lediglich nach den zugrunde liegenden Regeln der Syntax. Vielmehr haben Worte, gebraucht in unmittelbarer Nähe, eine Beziehung zueinander oder anders, ein Wort gewinnt an Bedeutung, wenn man die umgebenden Wörter kennt. Aufbauend auf dieser These sollte es möglich sein, Rückschlüsse auf Assoziationen zwischen Worten zu ziehen, wenn man deren unmittelbare Nachbarschaft in Texten untersucht.
Wie bereits zuvor beschrieben, wird diese Untersuchung des unmittelbaren Kontextes eines Wortes im Text durch die Kookkurrenzberechnung realisiert. In Kapitel 3 wurden zahlreiche Verfahren vorgestellt, die in ihrer Gesamtheit auf den Kookkurrenzstatistiken großer Textsammlungen beruhen. Das vorliegende Kapitel widmet sich dem im Laufe meiner Arbeit am Institut für Wissensbasierte Systeme und Wissensmanagement entwickelten Assoziationsberechnungsverfahren, dem '*Concept for the Imitation of the Mental Ability of Word Association*' oder kurz CIMAWA. Gezeigt werden zunächst die konzeptuellen Unterschiede zu bereits existenten Verfahren in Unterkapitel 4.1, sowie eine detailierte Beschreibung des entwickelten Verfahrens in Unterkapitel 4.2.

4.1 Einordnung CIMAWA – Ein Vergleich mit bekannten Verfahren

Ausgehend von der in Kapitel 3.2 getroffenen Kategorisierung der Assoziationsmaße lässt sich CIMAWA als eine Neuentwicklung in den Bereich der mathematisch-statistischen Berechnungsverfahren einordnen. Basis dieser Verfahren sind ausschließlich statistische Kennzahlen, die aus der Analyse großer Textsammlungen stammen. Als wichtigste Größen können zum einen die Kookkurrenzdaten, also das gemeinsame Vorkommen von Begriffen innerhalb eines definierten Textfensters und Werte bezüglich der Anzahl der zu analysierenden Begriffe definiert werden.
Neben den genannten Gemeinsamkeiten ist es jedoch unverzichtbar, an dieser Stelle auf die konzeptuellen Unterschiede der verschiedenen Verfahren einzugehen. Dabei ist im Kern zu untersuchen, wie eine Assoziation zwischen einem Wortpaar berechnet wird. Die überwiegende Zahl der bis dato entwickelten lexikalischen Assoziationsberechnungsverfahren

berechnet pro Wortpaar einen Wert, der die Assoziation dieses Paares numerisch bewertet. Damit ist die Stärke der Assoziation zwischen diesen Worten definiert. Vergleicht man das Konzept dieser Ansätze mit den in Kapitel 2 erarbeiteten Grundlagen über die menschliche Wortassoziation, so kommt man zu dem Schluss, dass in den beschriebenen Fällen eine symmetrische Assoziation zwischen allen Wortpaaren vorausgesetzt wird. Basierend auf den veröffentlichten Forschungsergebnissen führender Linguisten sowie der im Rahmen dieser Arbeit durchgeführten Studien, konnte jedoch herausgearbeitet werden, dass ein relevanter Teil der von Probanden bewerteten Wortassoziationen stark asymmetrische Tendenzen zeigt (siehe Kapitel 2.4.2.1).

Zwischen den praktischen Realisierungen und den Erkenntnissen aus Literatur und Fallstudien besteht eine Diskrepanz. Die Grundidee, die zur Entwicklung von CIMAWA führte, bestand in der Nutzbarmachung der Potentiale, die in der Beseitigung dieser Diskrepanz zu vermuten war. Ziel war die Entwicklung eines Assoziationsberechnungsverfahrens, mit dessen Hilfe nicht nur symmetrische Beziehungen zwischen Wortpaaren dargestellt werden können. In der Folge ist das entwickelte Verfahren darauf ausgelegt, die menschliche Wortassoziation mit ihren charakteristischen Eigenschaften möglichst genau nachzubilden. Dies bedeutet insbesondere, dass es möglich sein muss, auch Wortrelationen mit nachgewiesen asymmetrischer Beziehung korrekt nachzubilden. Dieser Prämisse Rechnung tragend, wurde CIMAWA weder als ein strikt symmetrisches noch als ein strikt asymmetrisches Assoziationsberechnungsverfahren entwickelt. Die folgende Abbildung 7 zeigt die konzeptuellen Unterschiede zwischen den klassischen symmetrischen Verfahren, den weniger gebräuchlichen asymmetrischen Verfahren und der Eigenentwicklung CIMAWA als einen hybriden Ansatz.

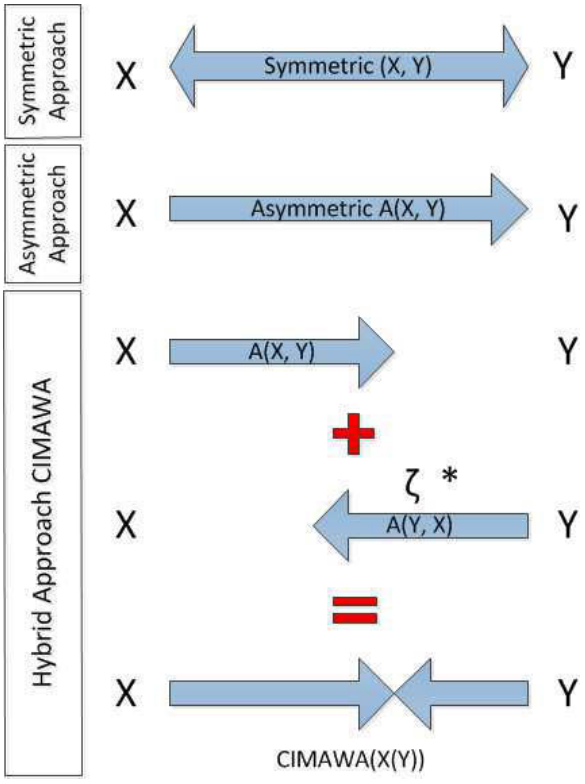


Abbildung 7. Konzeptuelle Unterschiede zwischen Assoziationsberechnungsverfahren (in Anlehnung an [61], [78])

Um die Eigenheiten der dargestellten Ansätze zu spezifizieren, sind insbesondere die Unterschiede herauszuarbeiten. Die Platzhalter ‘X‘ und ‘Y‘ in Abbildung 7 repräsentieren zwei Worte, deren Assoziation zueinander bestimmt werden soll. Das obere Beispiel in Abbildung 7 zeigt die generelle Funktionsweise der symmetrischen Ansätze. Diese berechnen auf Basis statistischer Daten über das gemeinsame Vorkommen der Worte ‘X‘ und ‘Y‘ in einem bestimmten Textfenster sowie deren Vorkommen in der verwendeten Textsammlung einen numerischen Wert. Dieser Wert beschreibt die Stärke der Assoziation zwischen den der Analyse zugrunde liegenden Worte. Die Assoziationsrichtung bleibt bei diesen symmetrischen Verfahren unberücksichtigt. Der Hinweis auf einige in Tabelle 5 eingeführte Beispiele zeigt Stärken und Schwächen der symmetrischen Verfahren. Die Assoziation zwischen dem Wortpaar (bad, good) wurde in den Tests von Nelson et al. in **[53]** als symmetrisch kategorisiert. Das bedeutet, dass etwa gleich viele Probanden auf den Stimulus ‘bad‘ die assoziative Antwort ‘good‘ gaben und umgekehrt. Folglich ist die Assoziation zwischen diesen beiden Worten in beiden Richtungen etwa gleich stark und kann durch einen einzigen numerischen Wert abgebildet werden. Symmetrische Assoziationsberechnungsverfahren sind in der Lage, einen solchen Assoziationswert zu bestimmen und zeigen dementsprechend gute Ergebnisse bei symmetrischen Wortassoziationen. Anders verhält es sich bei der Betrachtung des Wortpaares (bird, canary) aus Tabelle 5. Die hier vorliegende Assoziation wird als asymmetrisch kategorisiert. Deutlich weniger Probanden antworteten mit ‘canary‘ auf den Stimulus ‘bird‘ als umgekehrt. Überträgt man das Ergebnis auf die lexikalischen Assoziationsberechnungsverfahren, so ist festzustellen, dass ein numerischer Wert bestimmt durch ein symmetrisches Verfahren für die Beschreibung der Assoziation dieses Wortpaares keine zufriedenstellenden Ergebnisse in Aussicht stellt.

Bei näherer Betrachtung der symmetrischen Konzepte zur Assoziationsberechnung wird deutlich, dass jeder Versuch, die Stärke der Assoziation zwischen zwei Worten mit nachgewiesen asymmetrischer Assoziation durch einen einzigen numerischen Wert darzustellen, immer einen Kompromiss darstellen muss. Denn ohne Unterscheidung der Assoziationsrichtungen kann der errechnete Wert die Assoziation zwischen den Worten nur undifferenziert beschreiben.

Eine andere Herangehensweise wird durch die strikt asymmetrischen Ansätze verfolgt. Diese berücksichtigen die Assoziationsrichtung und berechnen die Assoziationsstärke für jede Richtung separat. Übertragen auf die menschlichen Assoziationstests sind die asymmetrischen Verfahren in der Lage, zwischen Stimulus und assoziativer Antwort zu unterscheiden. Bleiben wir bei den zuvor gewählten Beispielen und dem Wortpaar (bird, canary). Wie beschrieben ist die Assoziation zwischen diesen Worten asymmetrisch und die Assoziationsrichtungen unterscheiden sich bezüglich der Assoziationsstärke signifikant. Berechnet man die Assoziationsstärke für (bird → canary), so kann dies gleichgesetzt werden mit dem Versuch am Probanden, in dem er zum Stimulus ‘bird‘ eine assoziative Antwort geben soll. In der Folge ist die Assoziationsberechnung für (canary → bird) gleichzusetzen mit dem vorgegebenen Stimulus ‘canary‘ und der assoziativen Antwort ‘bird‘. Unter Zuhilfenahme eines asymmetrischen Assoziationsberechnungsverfahrens können im Besonderen diese asymmetrischen Wortassoziationen präzise abgebildet werden, was mit symmetrischen Verfahren aufgrund ihrer Konzeption nicht möglich ist. Dies bedeutet jedoch nicht, dass beide

Assoziationsrichtungen notwendigerweise unterschiedlich stark bewertet werden müssen, denn es sind auch Fälle denkbar, in denen beide Richtungen ähnlich stark bewertet werden und so eine symmetrische Beziehung nachgebildet werden kann.
Die Konzeption von CIMAWA als eine Art hybridem Ansatz, wird im unteren Teil der Abbildung 7 dargestellt. Die Assoziationsstärke CIMAWA (X(Y)) repräsentiert im menschlichen Assoziationsversuch ‘X‘ als Stimulus und ‘Y‘ als assoziative Antwort. Für die Berechnung dieser Assoziationsstärke werden jedoch im Unterschied zu den symmetrischen und asymmetrischen Konzeptionen beide Assoziationsrichtungen in die Berechnung einbezogen. Hier liegt der Hauptunterschied zu anderen Verfahren und gleichzeitig der Grund für die Kategorisierung von CIMAWA als hybriden Ansatz. Anstatt die Assoziationsrichtungen völlig zu vernachlässigen, wie die symmetrischen Verfahren, oder sich auf jeweils eine Assoziationsrichtung zu beschränken, wie die asymmetrischen Verfahren, fließen in die Assoziationsberechnung mit CIMAWA beide Assoziationsrichtungen ein. Zunächst wird die asymmetrische Assoziation (X → Y) bestimmt. Um den gemachten Beobachtungen zum Charakter der menschlichen Wortassoziation gerecht zu werden, fließt in die Berechnung von CIMAWA (X(Y)) auch die umgekehrte Richtung (Y → X) mit ein. Das bloße aufaddieren der beiden Assoziationsrichtungen würde dazu führen, dass die berechnete Assoziation symmetrisch wird, da sich die Werte von CIMAWA(X(Y)) und CIMAWA(Y(X)) nicht unterscheiden. Deshalb wird die asymmetrische Berechnung der Distanz der Gegenrichtung (Y → X) durch einen Dämpfungsfaktor angepasst und erst dann zu dem Wert für (X → Y) addiert. Durch den Dämpfungsfaktor ζ, der definiert ist im Intervall [0, 1], kann CIMAWA anwendungsbezogen angepasst werden. Er steuert den Grad der Einflussnahme der rückwärtigen Assoziationsrichtung (Y → X). Genauere Untersuchungen mit verschiedenen Ausprägungen des Dämpfungsfaktors folgen im nächsten Kapitel. Hierzu wurde eine umfangreiche Fallstudie durchgeführt, deren Resultate Aufschluss über die Güte der Ergebnisse bei verschiedenen Dämpfungsfaktoren geben.
In der hier beschriebenen Weise wird CIMAWA zu einem Berechnungsverfahren, welches sich als ein hybrides Verfahren zwischen den rein symmetrischen und rein asymmetrischen Verfahren einordnen lässt.

4.2 Assoziationsberechnung mit CIMAWA

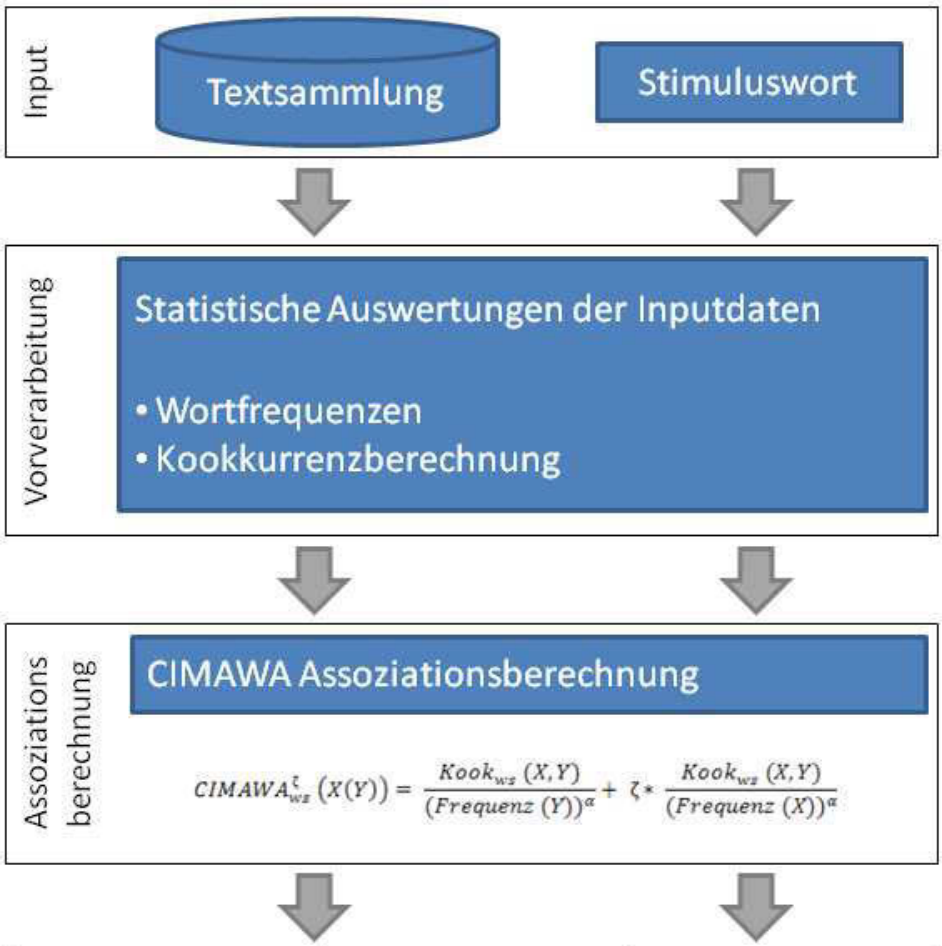
Nachdem die konzeptuellen Unterschiede herausgearbeitet wurden, soll im Folgenden die Assoziationsberechnung mit CIMAWA vorgestellt werden.
Unabhängig von der verwendeten Methode und deren Konzeption als symmetrisches, asymmetrisches oder hybrides Verfahren, verfolgen alle das gemeinsame Ziel einer möglichst realitätsnahen und präzisen Nachbildung der menschlichen Wortassoziation. Konsequenterweise erfolgt die Evaluation der Präzision der Assoziationsberechnungsverfahren über eine vergleichende Analyse der Berechnungsergebnisse mit den an menschlichen Probanden durchgeführten

Assoziationsexperimenten (siehe Kapitel 2.4.2.1). Die im deutschen Sprachraum meist verwendete und einzige dem Autor bekannte Assoziationsstudie von ausreichendem Umfang ist die Studie von Russell und Meseck [52]. Daher dient diese auch für sämtliche in dieser Arbeit gemachten Tests als Referenz.
Neben der gemeinsamen Zielsetzung hat die Aufarbeitung der mathematisch-statistischen Verfahren ergeben, dass die Eingabeparameter für sämtliche Verfahren aus statistischen Auswertungen über gemeinsames Vorkommen (Kookkurrenz) innerhalb eines definierten Textfensters und des Vorkommens der Worte in der Textsammlung bestehen. Alle Verfahren, inklusive CIMAWA, arbeiten folglich auf identischen statistischen Werten.
Diese statistischen Auswertungen hängen direkt von Art und Größe der gewählten Textsammlung oder dem Korpus ab. Wie sich unterschiedliche Textsammlungen und insbesondere verschiedene Größen dieser Textsammlungen auf die Ergebnisse der Assoziationsberechnung auswirken, wird ausführlich in Kapitel 5 getestet. Ein weiterer Einflussfaktor sind die Kookkurrenzen. Diese unterscheiden sich bei unterschiedlich gewählten Fenstergrößen signifikant. Wird die Fenstergröße für die Kookkurrenzanalyse sehr groß gewählt, kommt man zu dem Ergebnis, dass nahezu jedes Wort mit jedem anderen in Beziehung steht. Wählt man die Fenstergröße sehr klein, werden nur wenige Kookkurrenzen und in der Folge auch nur wenige Assoziationen erkannt. Diese Illustration zeigt die Bedeutung der Wahl einer angemessenen Fenstergröße für die Kookkurrenzanalyse. In Anbetracht dieser Tatsache wurden speziell in der frühen Entwicklungsphase von CIMAWA umfangreiche Testreihen durchgeführt, die dieser Problematik geschuldet sind. In welcher Weise die Fenstergröße die Ergebnisse der Assoziationsanalyse beeinflusst und welche Fenstergröße die besten Ergebnisse liefert, wird in Kapitel 5 im Detail analysiert.

4.2.1 Aufbau der CIMAWA-Assoziationsberechnung

Analog zu den freien Assoziationstests an Probanden ist der Input für CIMAWA ein Stimuluswort für das die am stärksten assoziierten Worte gefunden werden sollen.
Basierend auf einer Sammlung von Texten, aus der die benötigten statistischen Auswertungen über Kookkurrenzen und Frequenzen extrahiert werden, kann die Berechnung der meist assoziierten Begriffe erfolgen.
Die Bestimmung der Assoziationsstärke erfolgt dabei in aufeinanderfolgenden Schritten, die im Folgenden erläutert werden. Abbildung 8 illustriert die Prozessschritte.
Zunächst werden die der Analyse zugrunde liegenden Daten eingelesen. Als Inputdaten gegeben, ist zum einen die Textsammlung und zum anderen das zu analysierende Stimuluswort, zu dem die meist assoziierten Worte gefunden werden sollen. In der Vorverarbeitung werden die Kookkurrenzen des Stimuluswortes in der Textsammlung bestimmt. Ergebnis ist eine Liste mit Worten, die in einem bestimmten Textfenster gemeinsam mit dem gegebenen Stimuluswort vorkommen. Worte in dieser Liste sind potentiell assoziierte Begriffe. Welchen Einfluss die gewählte Fenstergröße auf die Ergebnisse der Assoziationsanalyse hat, wird zu einem späteren Zeitpunkt ausführlich diskutiert. Im Zusatz werden in der Vorverarbeitung die Frequenzen der Worte in der Kookkurrenzliste sowie die Frequenz des Stimuluswortes bestimmt und gespeichert. Im

Anschluss erfolgt die eigentliche Assoziationsberechnung durch CIMAWA. Für jedes Wort in der Kookkurrenzliste wird der CIMAWA-Wert bestimmt. Abschließend werden die Worte aus der Kookkurrenzliste absteigend nach CIMAWA-Wert sortiert und ausgegeben. Der Begriff mit dem höchsten CIMAWA-Wert in dieser Liste besitzt die stärkste Assoziation zum Stimuluswort und repräsentiert das rechnerische Pendant zur Primärantwort aus den FATs.



Stimuluswort	Assoziierte Antworten	CIMAWA Wert
Butter	Käse	0,834
	weich	0,801
	Brot	0,768
	...	...
	essen	0,0045

Abbildung 8. Ablauf CIMAWA-Assoziationsberechnung

Wie die Berechnung der Assoziationsstärke rechnerisch umgesetzt ist, wird im Anschluss thematisiert. Die folgende Formel 4 beschreibt die CIMAWA-Berechnungsformel.

Formel 4. Berechnung CIMAWA [61], [78]

$$CIMAWA_{ws}^{\zeta}(X(Y)) = \frac{Kook_{ws}(X,Y)}{(Frequenz\ (Y))^{\alpha}} + \zeta * \frac{Kook_{ws}(X,Y)}{(Frequenz\ (X))^{\alpha}}$$

In obiger Formel steht ‘X‘ für den gegebenen Stimulus, also für das Wort, zu dem die assoziierten Begriffe gefunden werden sollen. Parameter ‘Y‘ repräsentiert die potentiell mit ‘X‘ assoziierten Begriffe, also die Worte aus der Kookurrenzliste. Parameter ‘ws‘ steht für die Textfenstergröße (window-size) der Kookkurrenzberechnung. Dementsprechend steht ‘$Kook_{ws}$ (X, Y)‘ für die Häufigkeit des gemeinsamen Vorkommens von Wort ‘X‘ und ‘Y‘ in der gegebenen Fenstergröße ‘ws‘. ‘Frequenz(Y)‘ bzw. ‘Frequenz(X)‘ repräsentiert die Anzahl der entsprechenden Worte in der Textsammlung. Der Parameter ζ definiert den Dämpfungsfaktor der gegengerichteten Assoziation. Der Einfluss dieses Parameters auf die Assoziationsergebnisse wird in Kapitel 4.2.2 im Detail untersucht.
Die CIMAWA Assoziationsformel setzt sich zusammen aus zwei Summanden, deren konzeptuelle Herkunft bereits in Abbildung 7 diskutiert wurde. Der erste Summand repräsentiert die rein asymmetrische Beziehung (X → Y). Berechnet wird diese, indem die Kookkurrenz von (X, Y) durch die Frequenz des Wortes ‘Y‘ im Textkorpus dividiert wird. Dieser Teil der Berechnung entspricht dem von Wettler, Rapp und Ferber in [75] entwickelten Assoziationsberechnungsverfahren. Der zweite Summand berechnet die asymmetrische Assoziation der gegenläufigen Richtung (Y → X). Dieser zweite Summand fließt allerdings nicht unmittelbar in die Berechnung ein, sondern wird durch den sogenannten Dämpfungsfaktor ζ, der im Intervall [0, 1] definiert ist, angepasst. Dieser Dämpfungsfaktor ermöglicht die Steuerung des Grades der Symmetrie von CIMAWA. Zur Verdeutlichung werden an dieser Stelle die beiden Extremwerte 0 und 1 untersucht. Wählt man den Dämpfungsfaktor 0, so hat die Stärke der rückwärtsgerichteten Assoziation keinen Einfluss auf den berechneten CIMAWA-Wert. In diesem Fall ist CIMAWA ein rein asymmetrisches Assoziationsmaß. Wählt man hingegen den Wert 1 als Dämpfungsfaktor, so wird CIMAWA zu einem symmetrischen Assoziationsmaß. Der CIMAWA-Wert für (X(Y)) gleicht dem Wert für (Y(X)).
Aus der Arbeit von Rapp [47] ist bekannt, dass aufgrund der Tatsache, dass die Frequenzen von X (bzw. Y) im Nenner stehen, Wörter mit sehr niedriger Frequenz starke Auswirkungen auf die berechnete Assoziationsstärke haben. Aus diesem Grund schlagen Wettler und Rapp in [76] eine Fallunterscheidung vor, die auch für die beiden Summanden in der dargestellten CIMAWA-Berechnung verwendet wurde. Formel 5 und Formel 6 beschreiben die für CIMAWA geltenden Fallunterscheidungen genau.

Formel 5. Fallunterscheidung für den ersten CIMAWA Summanden

$$\frac{Kook_{ws}\,(X,Y)}{(Frequenz\,(Y))^{\alpha}} = \begin{cases} \dfrac{Kook\,(X,Y)}{(Frequenz(Y))^{\alpha}} \; falls \;\; Frequenz\,(Y) > \beta * Q \\ \dfrac{Kook\,(X,Y)}{(\gamma * Q)} \; falls \;\; Frequenz(Y) \leq \beta * Q \end{cases}$$

Formel 6. Fallunterscheidung für den zweiten CIMAWA Summanden

$$\frac{Kook_{ws}\,(X,Y)}{(Frequenz\,(X))^{\alpha}} = \begin{cases} \dfrac{Kook\,(X,Y)}{(Frequenz(X))^{\alpha}}\ falls\ \ Frequenz\,(X) > \ \beta * Q \\ \dfrac{Kook\,(X,Y)}{(\gamma * Q)}\ falls\ \ Frequenz(X) \leq \ \beta * Q \end{cases}$$

Rapp erzielt die besten Ergebnisse in [47] mit den folgenden Parameterwerten: α = 0,68; β = 0,000005 und γ = 0,000005. Q ist ein korpusabhängiger Parameter und steht für die Anzahl der Worte in der zugrunde liegenden Textsammlung.

4.2.2 Untersuchungen zum CIMAWA Dämpfungsfaktor

Um herauszuarbeiten, welchen Einfluss der Dämpfungsfaktor ζ auf die Qualität der Assoziationsberechnung hat und welcher Wert für diesen Parameter die besten Ergebnisse liefert, werden im folgenden Abschnitt die durchgeführte Fallstudie und deren Resultate präsentiert.
Als Referenz für die Güte der Prognosen von CIMAWA mit den unterschiedlichen Dämpfungsfaktoren dient der Assoziationstest von Russell und Meseck [52]. Hierbei wurde auf die korrekte Berechnung der Primärantworten aus dem freien Assoziationstest abgezielt. Das bedeutet, die Primärantworten aus dem FAT wurden verglichen mit den von CIMAWA berechneten Assoziationslisten. Wurde die Primärantwort als das am stärksten assoziierte Wort an erster Stelle der Assoziationsliste platziert, so gilt dies als eine korrekt vorhergesagte Primärantwort. Steht ein beliebiges anderes Wort an erster Stelle der Assoziationsliste, wurde die Primärantwort nicht korrekt vorhergesagt. Das den Untersuchungen zugrunde liegende Korpus ist online öffentlich verfügbar und wurde zusammengestellt vom Institut für Deutsche Sprache (IDS) in Mannheim [79]. Zum Zeitpunkt der Analyse hatte das Korpus eine Größe von ca. 2,8 Mrd. Worten.
Abbildung 9 visualisiert die Ergebnisse der Fallstudie für ausgewählte Werte im Intervall [0, 1].

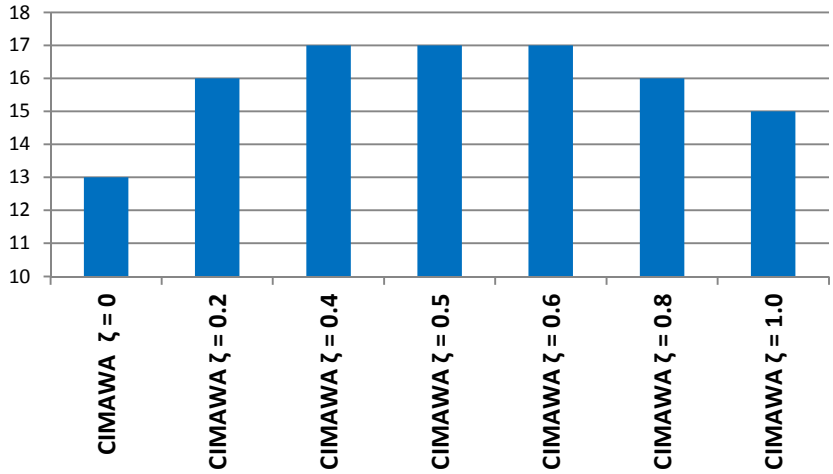


Abbildung 9. Prognose der Primärantworten bei verschiedenen Dämpfungsfaktoren [78]

Die schwächsten Ergebnisse lieferte CIMAWA mit einem Dämpfungsfaktor von 0. Wie bereits zuvor angedeutet und aus der Konzeption von CIMAWA abzuleiteten, implementiert CIMAWA mit einem Dämpfungsfaktor 0 ein rein asymmetrisches Berechnungsverfahren. In dieser Testreihe konnten mit diesem Dämpfungsfaktor lediglich 13 Primärantworten korrekt vorhergesagt werden. Das andere Extrem bildet der Dämpfungsfaktor 1, der bei dieser Belegung im eigentlichen Sinne keine Dämpfung bedeutet. Dieser Wert lässt CIMAWA zu einem rein symmetrischen Assoziationsmaß werden, da beide Assoziationsrichtungen paritätischen Einfluss auf den CIMAWA-Wert haben. Wählt man diesen Faktor, so ist aus den Auswertungen zu entnehmen, dass 16 Primärantworten richtig vorhergesagt werden konnten. Die besten Ergebnisse wurden im Wertebereich zwischen 0,4 und 0,6 erzielt. Alle drei Dämpfungsfaktoren kamen zu dem gleichen Ergebnis und verhalfen CIMAWA zu der korrekten Prognose von 17 Primärantworten. Da sich der Grad der Dämpfung in diesem Bereich als unkritisch in Bezug auf die erzielten Ergebnisse erwies, wird im Folgenden der Mittelwert von 0,5 als Dämpfungsfaktor für die CIMAWA Methode verwendet.

5 Vergleichende Fallstudien statistischer Assoziationsberechnungsverfahren

Nachdem in Kapitel 3 der vorliegenden Arbeit die gängigsten mathematisch-statistischen Verfahren zur Assoziationsberechnung vorgestellt wurden und in Kapitel 4 die Eigenentwicklung CIMAWA im Detail beschrieben wurde, soll an dieser Stelle die vergleichende Analyse aller präsentierten Assoziationsberechnungsverfahren im Fokus stehen.

Tabelle 14. Übersicht Fallstudien Kapitel 5

Fallstudie	Testreihe	Verglichene Assoziationsberechnungsverfahren			Textkorpus	Textfenster-größe zur Kookkurrenz berechnung	Referenz-FAT
1	A	PMI, WR, CIMAWA standard, WR adjusted, CIMAWA adjusted			Selbst erstelltes Textkorpus aus redaktionellen Texten (57.993 Texte; ca. 4.6 Mio. Worte)	± 5, ±12	Russell und Meseck [52]
1	B	PMI, WR, CIMAWA standard, WR adjusted, CIMAWA adjusted			Öffentliches Korpus des Instituts für Deutsche Sprache (IDS) der Universität Mannheim [79]	± 5, ±12	Russell und Meseck [52]
2	A	CIMAWA, WR, poisson-stirling, z-score, t-score, chi-squared,	log-likelihood, dice, jaccard, g-mean, MS, odds-ratio,	liddel, PMI, local MI, average MI, MI^2, MI^3,	Selbst erstelltes Textkorpus aus redaktionellen Texten (57.993 Texte; ca. 4,6 Mio. Worte)	± 5	Russell und Meseck [52]
2	B	CIMAWA, WR, poisson-stirling, z-score, t-score, chi-squared,	log-likelihood, dice, jaccard, g-mean, MS, odds-ratio,	liddel, PMI, local MI, average MI, MI^2, MI^3,	Selbst erstelltes Textkorpus aus redaktionellen Texten (57.993 Texte; ca. 4,6 Mio. Worte)	±12	Russell und Meseck [52]
3	A	CIMAWA, WR, poisson-stirling, z-score, t-score, chi-squared,	log-likelihood, dice, jaccard, g-mean, MS, odds-ratio,	liddel, PMI, local MI, average MI, MI^2, MI^3,	Öffentliches Korpus des Instituts für Deutsche Sprache (IDS) der Universität Mannheim [79] (3% mit ca. 84 Mio. Worten)	± 5, ±12	Russell und Meseck [52]
3	B	CIMAWA, WR, poisson-stirling, z-score, t-score, chi-squared,	log-likelihood, dice, jaccard, g-mean, MS, odds-ratio,	liddel, PMI, local MI, average MI, MI^2, MI^3,	Öffentliches Korpus des Instituts für Deutsche Sprache (IDS) der Universität Mannheim [79] (10% mit ca. 280 Mio. Worten)	± 5, ±12	Russell und Meseck [52]
3	C	CIMAWA, WR, poisson-stirling, z-score, t-score, chi-squared,	log-likelihood, dice, jaccard, g-mean, MS, odds-ratio,	liddel, PMI, local MI, average MI, MI^2, MI^3,	Öffentliches Korpus des Instituts für Deutsche Sprache (IDS) der Universität Mannheim [79] (50% mit ca. 1,4 Mrd. Worten)	± 5, ±12	Russell und Meseck [52]
3	D	CIMAWA, WR, poisson-stirling, z-score, t-score, chi-squared,	log-likelihood, dice, jaccard, g-mean, MS, odds-ratio,	liddel, PMI, local MI, average MI, MI^2, MI^3,	Öffentliches Korpus des Instituts für Deutsche Sprache (IDS) der Universität Mannheim [79] (100% mit ca. 2,8 Mrd. Worten)	± 5, ±12	Russell und Meseck [52]

Zu diesem Zweck wurden vom Autor während der Tätigkeit am Institut für Wissenbasierte Systeme zahlreiche Fallstudien und Testreihen durchgeführt, die in diesem Kapitel der Arbeit zum Thema gemacht werden. Insgesamt lassen sich die durchgeführten Experimente in drei Fallstudien mit zusammengenommen acht Testreihen beziffern. Um die Übersicht dieses Abschnittes zu vereinfachen, sind in Tabelle 14 alle Fallstudien und Testreihen mit den wichtigsten Rahmenbedingungen und Unterschieden dargestellt. Festgehalten sind neben Fallstudie und Testreihe die verglichenen Assoziationsberechnungsverfahren, das verwendete Korpus, die Textfenstergröße zur Kookkurrenzberechnung sowie der als Vergleichsreferenz verwendete FAT. In den hier vorgestellten Tests wurde aus den benannten Gründen stets der FAT von Russel und Meseck [52] verwendet. In jeder Fallstudie wurden in den verschiedenen Testreihen unterschiedliche Einflussparameter variiert. Um diese für den Leser leichter kenntlich zu machen, sind diese in Tabelle 14 jeweils grau hinterlegt.
Bei der Konzeption der Fallstudie stand jeweils eine zentrale Fragestellung zur Beantwortung. Auf Zweck und Unterschiede der Fallstudien untereinander sowie auf die Einflussparameter soll im Folgenden kurz eingegangen werden.
In der ersten Fallstudie wurden die Verfahren ‘PMI‘, ‘WR‘, ‘CIMAWA standard‘, ‘WR adjusted‘ und ‘CIMAWA adjusted‘ gegeneinander getestet. Der Grund für die Auswahl dieser begrenzten Anzahl an Berechnungsverfahren ist in der unterschiedlichen Konzeption der Verfahren begründet. PMI ist ausgelegt als rein symmetrischen Verfahren, WR als rein asymmetrisches und CIMAWA als neuartige hybride Konzeption. Ziel der ersten Fallstudie ist, die Unterschiede in der Konzeption der Verfahren (siehe Abbildung 7) miteinander zu vergleichen. Als Textfenstergröße wurde sowohl ±5 als auch ±12 in beiden Testreihen verwendet. Der Unterschied zwischen den beiden Testreihen dieser Fallstudie liegt in dem zugrunde liegenden Korpus. In Testreihe A wurde ein selbst erstelltes Korpus benutzt, wobei die zweite Testreihe ein öffentliches Korpus des IDS Mannheim [79] zur Basis hat.
Fallstudie 2 vergrößert die Menge der in Fallstudie 1 getesteten Assoziationsberechnungsverfahren um die folgenden: ‘poisson-stirling‘, ‘z-score‘, ‘t-score‘, ‘chi-squared‘, ‘log-likelihood‘, ‘dice‘, ‘jaccard‘, ‘g-mean‘, ‘MS‘, ‘odds-ratio‘, ‘liddel‘, ‘local MI‘, ‘average MI‘, ‘MI^2‘ und ‘MI^3‘. Als Korpus wurde die selbst erstellte Textsammlung aus Fallstudie 1 Testreihe A verwendet. Die zentrale Frage dieser Fallstudie ist die nach der vielversprechendsten Textfenstergröße für die Kookkurrenzberechnung. Testreihe A verwendet Fenstergröße ±5, wobei die Testreihe B die Experimente mit einer Textfenstergröße von ±12 wiederholt.
Die bezüglich der Anzahl der Testreihen sowie des Umfanges der durchgeführten Tests herausragende Studie ist die nun zu beschreibende Dritte. Dieser Fallstudie sind insgesamt 4 Testreihen anhängig. Getestet wurden die in Fallstudie 2 benannten Verfahren. Jede Testreihe wurde mit den Textfenstergrößen ±5 und ±12 durchgeführt. Im Fokus der Fallstudie 3 stand die Frage nach der am besten geeigneten Größe der zugrunde liegenden Textsammlung. Zu diesem Zweck wurde die Textsammlung des IDS Mannheim [79] in verschieden große Fragmente unterteilt: In Testreihe A wurden dafür rund 3% der Gesamttextsammlung als Basis verwendet. Das entspricht ca. 84 Mio. Worten. Sukzessive wird der Umfang des Korpus in den folgenden Testreihen vergrößert. Testreihe B operiert auf 10% der Gesamtsammlung und verwendet dabei ca. 280 Mio. Worte. In Testreihe C werden 50% benutzt, was ca. 1,4 Mrd. Worten entspricht. Die letzte Testreihe macht keine Einschränkungen bezüglich der

Größe der Textsammlung und verwendet den vollen Umfang, was ca. 2,8 Mrd. Worten gleichkommt.
Nachdem der Aufbau der Fallstudien strukturell beschrieben ist, werden in den folgenden Abschnitten die einzelnen Fallstudien und Testreihen im Detail dargestellt und die Ergebnisse auswertend präsentiert.

5.1 Fallstudie 1: Konzeptueller Vergleich statistischer Assoziationsberechnungsverfahren

In dieser ersten Fallstudie werden die konzeptuellen Unterschiede aus Abbildung 7 fokussiert. Wie bereits zuvor umrissen, wurde je ein symmetrisches und ein asymmetrisches sowie das hybride CIMAWA Verfahren gegeneinander getestet.
Repräsentant der symmetrischen Verfahren in dieser Fallstudie ist die ‘Pointwise Mutual Information‘ (PMI). Im Ursprung stammt der PMI Ansatz von Fano [80] und wurde unter anderem von Church [81], [73], [82] auf das Forschungsgebiet des Natural Language Processing (NLP) adaptiert. Als eine der wenigen asymmetrischen Ansätze wurde das Assoziationsmaß von Wettler und Rapp (WR) [76] herangezogen. Die Berechnungsvorschriften der genannten Verfahren sind Kapitel 3 zu entnehmen. Dabei wurden WR und CIMAWA zusätzlich mit mehreren Parameterwerten getestet. ‘WR‘ sowie ‘CIMAWA standard‘ sind mit den von Rapp in [47] erarbeiteten Parameterwerten implementiert. Die Parameterbelegung nach Rapp ist ebenfalls Kapitel 3 zu entnehmen. Nach zahlreichen Tests mit den vorgeschlagenen Werten wurden im Rahmen der Entwicklung von CIMAWA Veränderungen der Parameter vorgenommen. Dabei stellten sich die folgenden Werte als vielversprechende Alternativen heraus: Die Fenstergröße für die Kookkurrenzberechnung wurde auf ±5 herabgesetzt und die Werte für β und γ auf 0,000011 erhöht [61]. Verfahren, die diese justierten Parameterwerte verwenden, werden im Folgenden als ‘WR adjusted‘ bzw. ‘CIMAWA adjusted‘ kenntlich gemacht.
Als Referenz für die erzielten Resultate der einzelnen Berechnungsverfahren wird der bereits beschriebene FAT von Russell und Meseck [52] verwendet.
Fallstudie 1 beinhaltet zwei Testreihen. Diese unterscheiden sich bezüglich des verwendeten Textkorpus. Referenz-FAT und die implementierten Assoziationsberechnungsverfahren sind identisch. Die Evaluation der Verfahren stellt auf die folgenden Kriterien ab:

(a) Anzahl der korrekt vorhergesagten Primärantworten des FAT
(b) Durchschnittliche Position der Primärantworten in der Ergebnisliste

Diese Kriterien (a) und (b) werden in den Vergleichstests der Fallstudie 1 angelegt, Fallstudien 2 und 3 beschränken sich auf die Auswertung von Kriterium (a). Die Ergebnisse werden an entsprechender Stelle mittels Ergebnistabellen und zusammenfassenden Diagrammen präsentiert.

5.1.1 Fallstudie 1: Testreihe A

Testreihe A dieser Fallstudie operiert auf einem selbst erstellten Korpus bestehend, aus redaktionalen Texten. Diese Texte entstammen wöchentlich erscheinenden deutschen Zeitungen. Zum Zeitpunkt der Testreihe waren 57.993 Texte in der Datenbank gespeichert. Die durchschnittliche Länge der Texte betrug 1.733 Zeichen. Zum Vergleich stehen die Verfahren PMI, WR / adjusted sowie CIMAWA standard / adjusted.
Eine ganzheitliche Darstellung der erzielten Ergebnisse ist Tabelle 16 zu entnehmen. Die Tabelle umfasst alle Stimulusworte, bei denen wenigstens eines der getesteten Verfahren ein Ergebnis erzielen konnte. Die erste Spalte listet die Stimulusworte, die als Parameter für die Berechnung an die Verfahren übergeben wurden. In der zweiten Spalte werden die beobachteten Primärantworten aus [52] gelistet. In den folgenden Spalten sind Resultate der getesteten Verfahren dargestellt.
Durch explizieren einiger Beispiele aus Tabelle 16 soll zunächst verdeutlicht werden, wie solche Tabellen zu verstehen sind und welche Interpretationen sie zulassen. Als erstes Beispiel dient die zweite Zeile der Tabelle 16. Das zu analysierende Stimuluswort steht in der ersten Spalte, in diesem Fall handelt es sich um den Begriff ‘Butter‘. In dem Assoziationstest von Russel und Meseck war die am häufigsten gegebene Antwort auf den Stimulus ‘Butter‘ der Begriff ‘Brot‘. Dieser Begriff stellt die Primärantwort des Referenz-FAT dar, welche jeweils in der zweiten Spalte aufgeführt ist. Ziel eines jeden Assoziationsberechnungsverfahrens ist es, genau diesen Begriff an erster Stelle einer Liste von assoziierten Worten auszugeben. Um darzustellen, wie gut die einzelnen Verfahren die Assoziationen zu dem Stimuluswort berechnen, sind in den folgenden Spalten die Positionen der Primärantwort des Referenz-FAT in den Assoziationslisten der Verfahren angegeben. Im konkreten Fall von Stimulus ‘Butter‘ bedeutet dies, das PMI den Begriff ‘Brot‘ an Position 188 der berechneten Assoziationen führt. Anders ausgedrückt, ermittelt PMI 187 andere Begriffe zum Stimulus ‘Butter‘, die eine stärkere Assoziation aufweisen als die tatsächlich beobachtete Primärantwort. Betrachtet man die Ergebnisse der anderen Verfahren, so wird deutlich, dass diese die Primärantwort ‘Brot‘ deutlich weiter vorne in den berechneten Listen führen. Bei WR steht Brot an Position 27, bei ‘CIMAWA standard‘ an Position 19, bei ‘WR adjusted‘ an Position 12. Die Position von ‘Brot‘ in der Assoziationsliste von ‘CIMAWA adjusted‘ kann Tabelle 15 entnommen werden. Der Begriff ‘Brot‘ wird auf Rang 6 geführt (grau hinterlegt in Tabelle 15). Diese Platzierung auf dem sechsten Rang findet sich entsprechend in Tabelle 16 (Zeile 2, Spalte 7, graue Hinterlegung) wieder.

Tabelle 15. Assoziationsliste ‘CIMAWA adjusted' für das Stimuluswort ‘Butter'

Stimuluswort	Rang	Berechnete Antwort
Butter	1	Milch
	2	Sahne
	3	Käse
	4	…
	5	…
	6	Brot
	7	…
	8	…
	9	…
	…	…

Bei der in Tabelle 15 dargestellten Assoziationsliste handelt es sich demzufolge um eine Wortliste, die absteigend nach dem CIMAWA-Wert sortiert ist. Das Wort mit dem höchsten CIMAWA-Wert stellt die berechnete Primärantwort dar. Eine vollkommen korrekte Berechnung des, aus Sicht des FAT, richtigen Begriffs, kommt einer Platzierung der Primärantwort an Position 1 der Assoziationsliste gleich. Keines der getesteten Verfahren hat in diesem Test die Primärantwort der Probanden zum Stimulus ‘Butter‘ korrekt vorhergesagt. Nichtsdestotrotz kann postuliert werden, dass je niedriger die berechnete Position der Primärantwort in der Assoziationsliste, desto besser das Ergebnis.
In Zeile 8 der Tabelle 16 zeigt ein zweites Beispiel eine korrekte Vorhersage der Primärantwort durch zwei der getesteten Verfahren. Gegeben ist das Stimuluswort ‘Berg‘. Die zugehörige Primärantwort dazu lautet ‘Tal‘. In der von PMI berechneten Assoziationsliste erscheint die Primärantwort auf Position 157, WR platziert ‘Tal‘ auf Position 10 und ‘CIMAWA standard‘ auf Position 7. ‘WR adjusted‘ und ‘CIMAWA adjusted‘ hingegen geben ‘Tal‘ jeweils als das am stärksten assoziierte Wort auf Platz 1 der Assoziationslisten aus. In diesen Fällen wurde demzufolge die Primärantwort aus dem FAT richtig vorhergesagt. Das heißt, der berechnete Begriff entspricht dem, der von den menschlichen Probanden im Text am meisten genannt wurde. Diese menschliche Wortassoziation wurde korrekt nachgebildet.

Tabelle 16. Detailergebnisse Fallstudie 1, Testreihe A [61]

Stimulus	Primärantwort des Referenz-FAT	PMI	WR	CIMAWA Standard	WR adjusted	CIMAWA adjusted
Butter	Brot	188	27	19	12	6
rot	grün	117	1	1	1	1
dunkel	hell	48	1	1	1	1
Musik	Ton	185	3	2	1	1
weich	hart	174	419	435	59	65
essen	trinken	92	1	1	1	1
Berg	Tal	157	10	7	1	1
Haus	Hof	1676	319	302	269	266
Obst	Gemüse	59	1	1	1	1
süß	sauer	77	2	1	1	1
kalt	warm	61	2	1	1	2
langsam	schnell	663	540	67	242	37
wünschen	Weihnachten	451	324	315	208	237
Fluss	Wasser	111	195	9	112	113
Fenster	Glas	568	128	106	48	44
Bürger	Staat	716	78	43	42	31
sauer	süß	108	5	9	17	9
Erde	Himmel	112	6	4	2	4
hart	weich	276	258	276	30	42
Magen	Darm	28	1	1	1	1
gelb	rot	154	3	1	2	1
Brot	essen	529	1171	596	206	175
Licht	dunkel	216	4	4	2	3
schnell	langsam	2590	306	270	122	124
Kopf	Haar	253	48	65	206	187
bitter	süß	-	17	18	-	-
Hammer	Amboss	14	1	1	3	1
laut	leise	3557	1679	1842	244	279
ruhig	laut	278	684	445	278	278
Salz	Zucker	119	3	3	2	2
Käse	Butter	269	15	16	7	8
Spinne	Netz	10	4	3	2	1
Ozean	Meer	16	7	7	2	3

Die Ergebnisse von Testreihe A bezüglich Kriterium (a) und der Anzahl der korrekt vorhergesagten Primärantworten ist Abbildung 10 zu entnehmen. Diese fasst die erzielten Ergebnisse aus Tabelle 16 insoweit zusammen, dass für jedes Verfahren die an Platz 1 der Assoziationsliste platzierten Primärantworten aufaddiert.

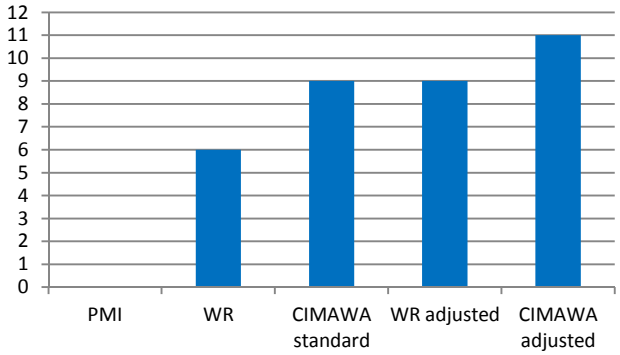


Abbildung 10. Fallstudie 1, Testreihe A, Kriterium (a) [61]

Der symmetrische PMI Ansatz erzielte in diesem Test die schlechtesten Ergebnisse und konnte keine Primärantwort vorhersagen. WR sagte 6 Primärantworten richtig voraus. Die gleichen Parameterwerte führten mit ‘CIMAWA standard‘ zur richtigen Vorhersage von 9 Primärantworten. Ähnliche Ergebnisse wurden nach der Anpassung der Parameterwerte beobachtet. ‘WR adjusted‘ erreicht den Wert von ‘CIMAWA standard‘ und sagt ebenfalls 9 Primärantworten voraus. Die Justierung der Parameter für CIMAWA verbessert die Voraussage auf den Bestwert für diese Testreihe auf 11.

Wie bereits mehrfach beschrieben, ist das Ziel einer jeden Assoziationsberechnungsmethode die Platzierung der beobachteten Primärantwort an Platz 1 bzw. eine möglichst niedrige Platzierung dieser in der berechneten Assoziationsliste. Bezüglich der vollkommen korrekten Vorhersagen wurden die Ergebnisse bereits in Abbildung 10 zusammenfassend dargestellt. An dieser Stelle wird diese Auswertung um ein zusätzliches Kriterium erweitert. In Kriterium (b) wird für jedes Verfahren in Tabelle 16 die Spaltensumme gebildet. Anschließend wird diese durch die Anzahl der Stimulusworte dividiert, um die durchschnittliche Platzierung der Primärantworten zu erhalten. Bezüglich dieses Kriteriums gilt folglich: Je niedriger der berechnete Durchschnittswert, desto besser ist das Ergebnis einzustufen.

Abbildung 11 zeigt die Ergebnisse der Testreihe A bezüglich Kriterium (b) und der durchschnittlichen Position der Primärantworten in den Ergebnislisten der verschiedenen Verfahren. Die Resultate bezüglich der Güte der Vorhersagen bestätigen dabei die bereits beschriebenen Beobachtungen.

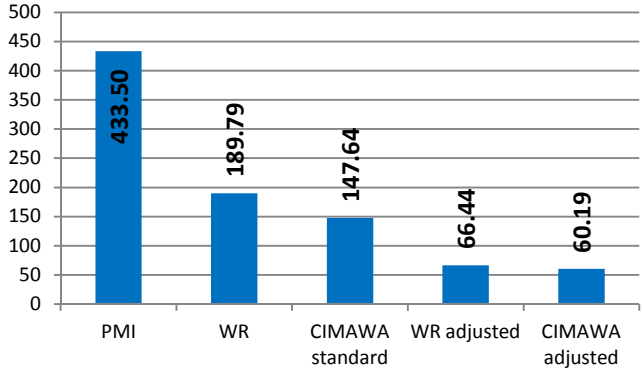


Abbildung 11. Fallstudie 1, Testreihe A, Kriterium (b) [61]

Die schwächsten Ergebnisse realisiert auch bei Kriterium (b) der PMI Ansatz. Die durchschnittliche Position liegt hier bei 433,5. Alle anderen getesteten Verfahren verbessern diesen Wert deutlich. Ein Vergleich zwischen WR und CIMAWA bei gleichen Parameterwerten zeigt gleiche Tendenzen wie bei Kriterium (a). ‘CIMAWA standard‘ liegt mit Position 147,64 im Durchschnitt mehr als 40 Plätze vor WR. Darüber hinaus ist abermals zu erkennen, dass ‘CIMAWA adjusted‘ mit einem dem Wert 60,19 das beste Resultat erzielt und ca. 6 Positionen besser abschneidet als ‘WR adjusted‘.

5.1.2 Fallstudie 1: Testreihe B

In Testreihe B wurden die aus der ersten Testreihe bekannten Verfahren erneut miteinander verglichen. Auch der Referenz-FAT bleibt der gleiche und dient in beschriebener Weise der Evaluation der Ergebnisse. Der Unterschied zur ersten Testreihe ist das den Berechnungen zugrunde liegende Korpus. In der vorliegenden Testreihe wird das vom Institut für Deutsche Sprache (IDS) in Mannheim erstellte und online zugängliche Korpus [79] verwendet. Zum Abfragezeitpunkt bestand diese Textsammlung aus ca. 2,8 Mrd. Worten.
Bezüglich des Umfangs sämtlicher in dieser Arbeit benutzten Textsammlungen, wird in dieser Testreihe (gemeinsam mit Fallstudie 3, Testreihe D) das größte Korpus verwendet. Vergleicht man diesen Umstand mit der Tatsache, dass in der ersten Testreihe dieser Fallstudie das kleinste Korpus dieser Arbeit Verwendung fand, so lässt Fallstudie 1 einen ersten Vergleich der Ergebnisse des größten und des kleinsten Korpus zu.
Tabelle 17 zeigt alle erzielten Ergebnisse mit Stimulusworten und Primärantworten, bei denen mindestens eines der getesteten Verfahren ein Ergebnis vorzuweisen hat. Unter Ergebnis versteht man an dieser Stelle, dass die Primärantwort zum Stimuluswort in der berechneten Assoziationsliste eines getesteten Verfahrens vorkommt. Das bedeutet, dass diejenigen Stimulusworte, deren Primärantwort des FAT nicht in der Liste der assoziierten Worte eines der Verfahren vorkommt, auch nicht in Tabelle 17 aufgeführt sind. Bei dieser Testreihe konnten Ergebnisse für 81 der insgesamt 100 Stimulusworte erzielt werden. Bereits beim ersten Blick auf Tabelle 17 fällt auf, dass in dieser Testreihe im Vergleich zu Testreihe A eine deutlich vergrößerte Anzahl an Ergebnissen erzielt werden konnte. Dies ist auf das deutlich vergrößerte Korpus zurückzuführen. Jedes der getesteten Verfahren basiert unter anderem auf dem gemeinsamen Vorkommen der zu berechnenden Primärantwort und dem gegebenen Stimulus. Deshalb ist es mit den beschriebenen Verfahren nicht möglich, eine Assoziation zwischen zwei Worten zu berechnen, die in der zugrunde liegenden Textsammlung nicht gemeinsam vorkommen. Das quantitative Vergrößern der Textsammlung vergrößert zugleich die Wahrscheinlichkeit, dass die Primärantwort des Referenz-FAT gemeinsam im definierten Textfenster mit dem gegebenen Stimuluswort vorkommt. Mit dieser erhöhten Wahrscheinlichkeit des gemeinsamen Vorkommens bei Vergrößerung des Korpus erklärt sich die deutliche Steigerung in der Anzahl der Ergebnisse zwischen Testreihe A und Testreihe B der hier präsentierten Fallstudie.

Tabelle 17. Detailergebnisse Fallstudie 1, Testreihe B

Stimulus	Primärantwort FAT	PMI	WR	CIMAWA standard	WR adjusted	CIMAWA adjusted
Adler	Vogel	97	65	50	157	35
ängstlich	Kind	60	6	6	13	6
Arzt	Krankheit	109	48	44	114	21
Baby	Kind	20	1	1	1	1
Berg	Tal	21	2	1	2	1
Bett	schlaf	6	6	3	13	40
Bibel	Buch	36	2	2	1	2
blau	Himmel	89	23	18	24	13
Brot	essen	66	9	9	9	5
Bürger	Staat	27	6	5	10	4
Butter	Brot	12	2	2	1	2
Dieb	stehlen	1	1	3	1	1
dunkel	hell	4	2	13	2	26
durstig	hungrig	2	1	1	1	1
Erde	Himmel	35	2	2	1	1
essen	trinken	4	1	1	1	1
Fenster	Glas	70	29	18	41	14
Fluß	Wasser	43	1	1	1	1
Frau	Mann	110	9	25	6	19
Freude	leid	99	108	67	79	45
Fuß	Schuh	27	31	14	51	86
Gedächtnis	Gehirn	16	10	50	26	10
gelb	rot	100	39	26	36	16
Gerechtigkeit	Gericht	129	50	52	61	44
Gesundheit	Krankheit	95	54	45	108	26
glatt	Eis	79	27	15	28	10
grün	Wiese	83	60	32	61	113
Hammelfleisch	essen	34	4	4	43	4
hart	weich	7	13	51	15	70
Haus	hof	60	15	18	15	11
Häuschen	Garten	70	20	18	29	20
hoch	tief	98	38	39	35	30
hungrig	durstig	196	198	198	200	198
Junge	Mädchen	47	10	14	11	13
kalt	warm	7	2	1	5	13
Käse	Butter	23	8	33	9	25
Kind	klein	36	3	1	4	1
Kohl	Gemüse	42	38	23	89	129
König	Kaiser	54	8	7	14	5
Kopf	Haar	30	26	8	45	93
Krankheit	Gesundheit	125	73	46	153	29
Lampe	Licht	78	9	9	16	9
lang	kurz	56	13	13	13	12
langsam	schnell	84	11	11	13	9
laut	leise	-	-	-	67	-
Licht	dunkel	32	17	8	18	56
Mädchen	Junge	9	1	1	1	1
Magen	Darm	2	2	9	1	2
Mann	Frau	27	2	2	3	2
Mond	Stern	19	3	1	2	1
Musik	Ton	44	27	23	38	18
Nadel	spitz	63	72	72	74	72
Obst	Gemüse	1	1	1	1	2
Ofen	wärme	114	61	33	118	61
Ozean	Meer	10	1	1	4	1
pfeifen	singen	83	23	14	27	23
Priester	Kirche	71	2	1	2	1
Quadrat	Viereck	20	70	70	81	70
rauh	glatt	14	13	13	16	13
Religion	Glaube	67	23	13	27	7
rot	grün	134	84	53	76	33
ruhig	laut	181	103	116	-	87
Salz	Zucker	-	-	-	169	-
Schere	schneiden	16	6	6	8	6
schlafen	Bett	24	15	9	21	7
Schmetterling	Falter	197	197	197	-	197
schwer	leicht	56	9	12	22	9
Soldat	Krieg	30	1	1	3	1
Sorge	kummer	193	193	193	190	193
Spinne	Netz	19	1	1	1	1
Stiel	Besen	30	19	19	28	19
Stuhl	Bein	48	29	12	34	29
süß	sauer	35	6	1	4	6
Tabak	rauchen	13	4	3	6	4
tief	hoch	96	9	14	13	9
Tisch	Stuhl	10	12	45	13	58
träumen	schlafen	34	34	19	45	121
weich	hart	23	1	1	1	1
Whisky	Schnaps	51	44	44	81	44
wünschen	Weihnachten	75	49	30	50	22
Zorn	Wut	19	4	1	5	4

In der bekannten Form stellen die Zahlen in Tabelle 17 die Platzierungen der Primärantworten des Referenz-FAT in den Assoziationslisten der jeweiligen Verfahren dar. Im Folgenden werden die präsentierten Detailergebnisse, in der im Rahmen von Testreihe A beschriebenen Art, analysiert. Dabei wird zunächst Kriterium (a) ausgewertet, was gleichbedeutend ist mit der korrekten Vorhersage der im Referenz-FAT beobachteten Primärantwort. Diese Fälle sind in Tabelle 17 genau diejenigen, die in der Spalte für das jeweilige Verfahren mit einer ‘1‘ bewertet sind. Einen Gesamtüberblick mit der Anzahl der an ‘1‘ platzierten Primärantworten für jedes Verfahren gibt Abbildung 12.

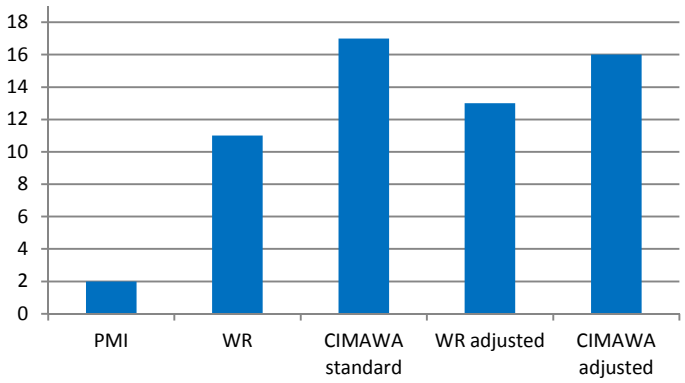


Abbildung 12. Fallstudie 1, Testreihe B, Kriterium (a) [61]

Die Tendenzen aus Testreihe A werden in dieser Testreihe bekräftigt. Der symmetrische PMI-Ansatz stellt sich auch in Testreihe B als der schwächste heraus. PMI prognostiziert lediglich zwei Primärantworten korrekt. WR sagt 11 Primärantworten richtig voraus, wobei sich die Anzahl bei ‘WR adjusted‘ auf 13 erhöht. Die besten Ergebnisse bezüglich Kriterium (a) erzielt ‘CIMAWA standard‘, indem 17 Primärantworten richtig erkannt werden. Ähnliche Ergebnisse erzielt ‘CIMAWA adjusted‘ mit 16, was den beiden Bestwerten dieser Testreihe entspricht.

Vergleicht man diese Ergebnisse mit den Resultaten der ersten Testreihe, so können die Auswirkungen des vergrößerten Korpus herausgearbeitet werden. Bezüglich Kriterium (a) kann zusammengefasst werden, dass sich alle Verfahren verbessert haben. In Zahlen ausgedrückt bedeutet dies folgendes: PMI sagt 2 Primärantworten mehr voraus, WR erreicht 5- und ‘CIMAWA standard‘ sogar 8 Primärantworten mehr. ‘WR adjusted‘ verzeichnet ein Plus von 4 und ‘CIMAWA adjusted‘ verbessert die Ergebnisse um 5 zusätzliche gefundene Primärantworten.

In Abbildung 13 sind die Ergebnisse bezüglich Kriterium (b), was der durchschnittlichen Platzierung der Primärantworten in den Assoziationslisten entspricht, dargestellt. PMI erreicht in Testreihe B eine durchschnittliche Platzierung von 55,03, was einerseits eine deutliche Steigerung zu den in Testreihe A zu verzeichnenden 433,5 ist, andererseits jedoch im Vergleich mit den anderen Verfahren das schwächste Ergebnis bedeutet. Mit einem Durchschnittswert von 35,54 erzielt ‘WR adjusted‘ das zweitschwächste Ergebnis, gefolgt von dem etwas besseren ‘CIMAWA adjusted‘ mit 30,33. Den Bestwert für Kriterium (b) kann ‘CIMAWA standard‘ mit 25,81 verbuchen, gefolgt von dem zweitbesten Wert 28,13 von WR.

Alle getesteten Verfahren erreichen in Testreihe B deutlich verbesserte Ergebnisse bezüglich Kriterium (b). Auch die abgeleiteten Tendenzen bezogen auf Kriterium (a) können nach Auswertung der Ergebnisse der Abbildung 13 bekräftigt werden. ‘CIMAWA standard‘ erweist sich sowohl bei Kriterium (a) als auch bei Kriterium (b) als das beste Verfahren dieser Testreihe.

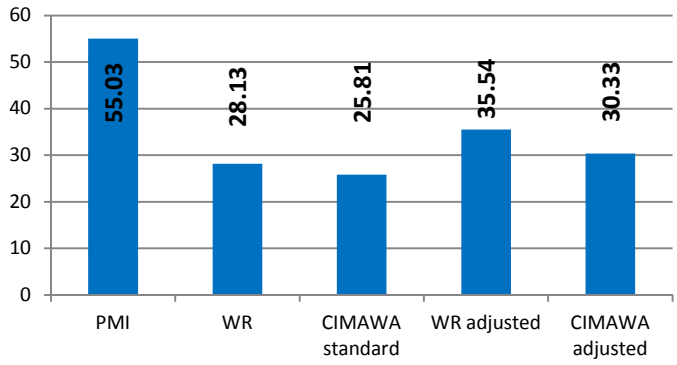


Abbildung 13. Fallstudie 1, Testreihe B, Kriterium (b) [61]

Vergleicht man die Ergebnisse beider Testreihen aus Fallstudie 1, so kann festgehalten werden, dass der symmetrische Ansatz PMI in beiden Testreihen und dort jeweils bezüglich beider Kriterien die schwächsten Ergebnisse erzielt. Etwas differenzierter fällt der Vergleich der übrigen Verfahren aus. In der ersten Testreihe und dem institutseigenen Korpus liefert ‘CIMAWA adjusted‘ jeweils die Bestwerte. Auf dem deutlich vergrößerten Korpus in Testreihe B wird ‘CIMAWA standard‘ bezüglich beider Kriterien zum Spitzenreiter unter den getesteten Verfahren.
Der parameterabhängige Vergleich von CIMAWA und WR zeigt in beiden Testreihen Analogien, denn bei gleichen Parameterwerten liegt CIMAWA in der Tendenz vor WR. Lediglich die Reihenfolge der Parameterkategorien hat sich geändert, denn auf dem Korpus von Testreihe B führen die standard Parameter zu besseren Ergebnissen als die justierten.

5.2 Fallstudie 2: Fenstergrößenabhängiger Vergleich statistischer Assoziationsberechnungsverfahren

Ähnlich der Fallstudie 1 ist auch die vorliegende in zwei Testreihen unterteilt, wobei sich der Fokus von den unterschiedlichen Korpora in Fallstudie 1 hin zur Fenstergröße der Kookkurrenzberechnung verschiebt. Testreihe A dieser Fallstudie umfasst die Simulationen bei einer Textfenstergröße von ±5 und Testreihe B ermittelt die Vergleichswerte bei der Fenstergröße ±12. Beide Testreihen benutzen für die Berechnungen den bereits aus Fallstudie 1 bekannten selbst erstellten Korpus. Der Umfang, bezogen auf die implementierten Verfahren, wurde im Vergleich zur ersten Fallstudie deutlich vergrößert. Neben den aus Fallstudie 1 bekannten Verfahren wurden zusätzlich die statistischen

Verfahren aus Kapitel 3 implementiert und getestet, so dass insgesamt 18 Verfahren verglichen werden.

5.2.1 Fallstudie 2: Testreihe A

Tabelle 18 und Tabelle 19 zeigen die Detailergebnisse der getesteten Verfahren in Testreihe A. Allen Berechnungsmethoden liegen die identischen Eingabeparameter zugrunde. Für die Kookkurrenzstatistik wurde mit einer Textfenstergröße von ±5 gearbeitet. Die Einordnung der Verfahren ist jeweils der ersten Zeile zu entnehmen und entspricht der in Kapitel 3 vorgenommenen Einteilung der Assoziationsberechnungsverfahren. Die dargestellten Tabellen führen sämtliche Stimulusworte auf, bei denen zumindest eines der implementierten Verfahren ein Ergebnis liefert.
Als eines der Ergebnisse aus Fallstudie 1 konnte herausgearbeitet werden, dass auf dem selbst erstellten Korpus sowohl CIMAWA als auch WR die besten Resultate mit den justierten Parametern liefern konnte. Deshalb werden in dieser Fallstudie aufgrund des verwendeten Korpus lediglich ‘CIMAWA adjusted‘ und ‘WR adjusted‘ aufgeführt. Die Ergebnisse dieser Verfahren mit standard Parametern sind den Auswertungen in Fallstudie 1 Testreihe A und im Speziellen Tabelle 16 zu entnehmen.

Tabelle 18. Detailergebnisse Fallstudie 2, Testreihe A, Teil 1

ws ±5			Likelihood Measure	Asymptotic Hypothesis Tests			
Stimulus	CIMAWA adjusted	WR adjusted	poisson-stirling	z-score	t-score	chi-squared	log-likelihood
Butter	6	12	6	97	4	97	6
rot	1	1	1	1	1	1	1
Tisch	597	555	419	401	474	401	433
dunkel	1	1	1	1	1	1	1
Musik	1	1	1	72	1	72	1
weich	65	59	180	177	181	177	180
essen	1	1	1	1	1	1	1
Berg	1	1	4	102	2	102	4
Haus	266	269	1410	1399	335	1395	1260
Obst	1	1	1	1	1	1	1
süß	1	1	1	46	1	46	1
kalt	2	1	2	36	2	36	2
langsam	37	242	373	495	9	495	265
wünschen	237	208	495	490	520	490	503
Fluss	113	112	113	112	113	112	113
Fenster	44	48	73	408	30	408	71
Bürger	31	42	52	378	12	378	47
Spinne	1	2	1	7	1	7	1
sauer	9	17	10	57	11	57	10
Erde	4	2	6	85	5	102	6
hart	42	30	307	290	326	290	308
Magen	1	1	1	3	1	3	1
gelb	1	2	1	2	1	2	1
Brot	175	206	431	478	124	478	420
Licht	3	2	6	15	6	15	6
schnell	124	122	1241	1753	110	1762	895
Ozean	3	2	16	16	16	16	16
Kopf	187	206	56	212	79	212	58
Hammer	1	3	1	1	1	1	1
laut	279	244	2947	3170	764	3218	2985
ruhig	278	278	278	278	278	279	279
Salz	2	2	3	4	3	4	3
Straße	1890	1888	1853	1885	1886	1537	1376
Käse	8	7	13	95	10	95	13

Die Ergebnisse von CIMAWA, WR, der ‘likelihood measure‘ ‘poisson-stirling‘ sowie die ‘asymptotic hypothesis tests‘ ‘z-score‘, ‘t-score‘, ‘chi-squared‘ und ‘log-likelihood‘ sind in Tabelle 18 aufgeführt. Die Platzierungen der Primärantworten zu den einzelnen Stimulusworten sind in den entsprechenden Spalten für die einzelnen Verfahren abgetragen. Der Übersichtlichkeit geschuldet erfolgt die Aufteilung der Detailergebnisse in zwei Tabellen (Tabelle 18 und Tabelle 19).
Die Ergebnisse der Kategorien ‘point estimates of association strength‘ mit den Verfahren ‘dice‘, ‘jaccard‘, ‘g-mean‘, ‘MS‘, ‘odds-ratio‘ und ‘liddel‘, die ‘measures from information theory‘ ‘PMI‘, ‘local MI‘ und ‘average MI‘ sowie die ‘heuristic measures‘ ‘MI^2‘ und ‘MI^3‘ werden in Tabelle 19 dargestellt.

Tabelle 19. Detailergebnisse Fallstudie 2, Testreihe A, Teil 2

ws ±5	Point Estimates of Association Strength						Measures from Information Theory			Heuristic Measures	
Stimulus	dice	jaccard	g-mean	MS	odds-ratio	liddel	PMI	local MI	average MI	MI^2	MI^3
Butter	20	20	97	20	20	188	188	5	6	97	15
rot	1	1	1	1	8	117	117	1	1	1	1
Tisch	456	456	404	551	825	364	364	446	433	404	438
dunkel	1	1	1	1	45	48	48	1	1	1	1
Musik	2	2	72	1	61	185	185	1	1	72	6
weich	175	175	178	174	76	174	174	180	180	178	180
essen	1	1	1	1	141	92	92	1	1	1	1
Berg	15	15	102	2	82	157	157	4	4	102	59
Haus	245	245	1303	282	560	1676	1676	843	1260	1303	994
Obst	1	1	1	1	2	59	59	1	1	1	1
süß	7	7	46	9	29	77	77	1	1	46	5
kalt	2	2	36	1	39	61	61	2	2	36	3
langsam	564	564	443	663	10	663	663	50	265	443	255
wünschen	505	505	497	226	618	451	451	518	503	497	536
Fluss	111	111	113	111	11	111	111	113	113	113	113
Fenster	57	57	401	27	85	568	568	45	71	401	244
Bürger	18	18	349	8	87	716	716	18	47	349	138
Spinne	10	10	7	10	4	10	10	1	1	7	4
sauer	4	4	57	3	158	108	108	9	10	57	13
Erde	10	10	85	3	83	112	112	5	6	85	59
hart	299	299	292	193	470	276	276	314	308	292	308
Magen	1	1	3	1	22	28	28	1	1	3	1
gelb	2	2	2	4	1	154	154	1	1	2	1
Brot	444	444	472	122	127	529	529	344	420	472	416
Licht	6	6	15	7	238	216	216	6	6	15	9
schnell	120	120	1574	147	554	2590	2590	212	895	1574	828
Ozean	16	16	16	16	6	16	16	16	16	16	16
Kopf	60	60	212	88	624	253	253	65	58	212	173
Hammer	1	1	1	1	104	14	14	1	1	1	1
laut	1504	1504	3136	1686	3095	3557	3557	2298	2985	3136	2831
ruhig	278	278	278	278	14	278	278	278	279	278	278
Salz	3	3	4	3	38	119	119	3	3	4	3
Straße	1887	1887	1889	1887	135	1887	1887	1887	1376	1889	1890
Käse	10	10	95	7	117	269	269	12	13	85	18

Abermals auf den ersten Blick zu erkennen ist der Zusammenhang zwischen Größe des zugrunde liegenden Korpus und der Anzahl der Ergebnisse. Der im Vergleich kleine Korpus von ca. 4,6 Mio. Worten führt auch in dieser Testreihe zu einer vergleichsweise verringerten Ergebnismenge.
Eine zusammenfassende Darstellung der Ergebnisse von Testreihe A liefert das Balkendiagramm in Abbildung 14. In diesem ist für jedes der hier getesteten Verfahren die Anzahl der auf Position 1 der Assoziationsliste platzierten Primärantworten angegeben.

Mit 11 korrekt vorhergesagten Primärantworten liefert CIMAWA das beste Ergebnis dieser Testreihe. Es folgen vier Verfahren (‘poisson-stirling‘, ‘t-score‘, ‘log-likelihood‘, ‘local MI‘, ‘average MI‘), die jeweils 10 richtige Primärantworten prognostizieren. Neun Mal richtig liegt WR, gefolgt von 8 Verfahren, die zwischen 5 und 8 Primärantworten vorhersagen (‘z-score‘, ‘chi-squared‘, ‘dice‘, ‘jaccard‘, ‘g-mean‘, ‘MS‘, ‘MI^2‘, ‘MI^3‘). Die schwächsten Ergebnisse liefern ‘odds-ratio‘ mit einer bzw. ‘liddel‘ und ‘PMI‘ mit keiner korrekt vorhergesagten Primärantwort.

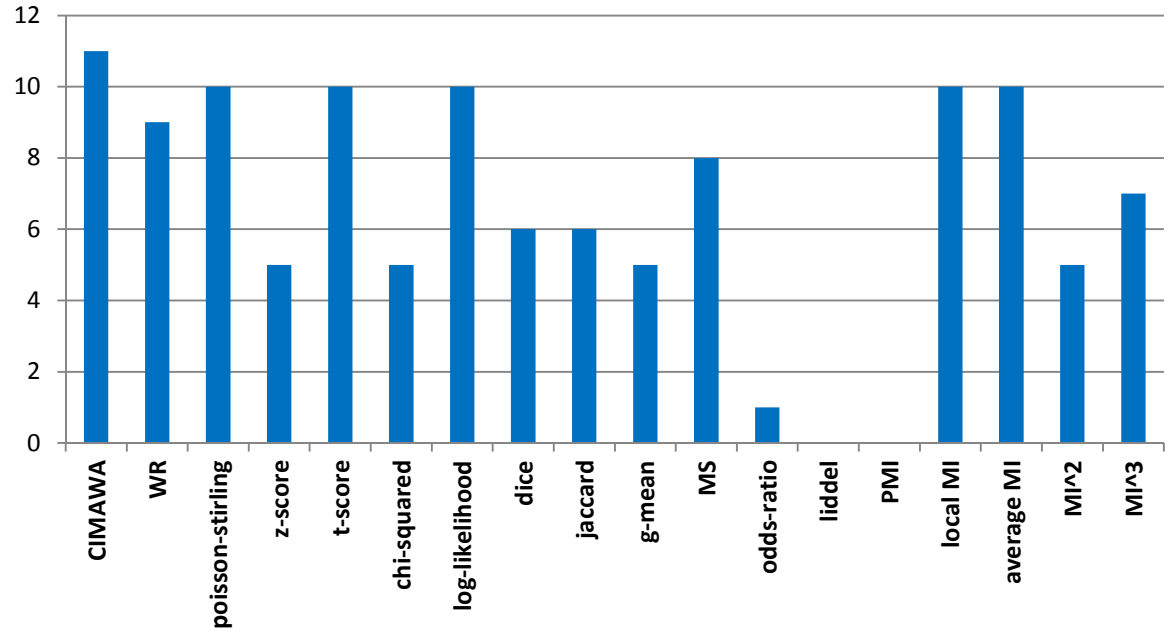


Abbildung 14. Zusammenfassung Fallstudie 2, Testreihe A

Eine abschließende und vergleichende Interpretation der hier erzielten Ergebnisse folgt im Anschluss an die Präsentation der Ergebnisse der Testreihe B.

5.2.2 Fallstudie 2: Testreihe B

Die zweite Testreihe dieser Fallstudie bedient sich des bekannten Testaufbaus aus Testreihe A. Der einzige Unterschied besteht in der Vergrößerung des Textfensters zur Kookkurrenzberechnung von ±5 auf ±12. Nach Auswertung der Ergebnisse dieser Testreihe sind erste vergleichende Schlussfolgerungen über die Qualität der Resultate der verschiedenen Textfenstergrößen möglich.

Tabelle 20. Detailergebnisse Fallstudie 2, Testreihe B, Teil 1

ws ±12			Likelihood Measures	Asymptotic Hypothesis Tests			
Stimulus	CIMAWA adjusted	WR adjusted	poisson-stirling	z-score	t-score	chi-squared	log-likelihood
Butter	14	20	14	198	10	198	14
rot	1	1	1	1	1	1	1
Tisch	592	621	313	536	205	536	324
dunkel	1	1	1	1	1	1	1
Musik	2	2	4	182	1	182	4
weich	154	138	426	422	431	422	427
essen	1	1	1	1	2	1	1
Berg	1	2	14	227	7	227	14
Haus	174	170	1787	2024	155	2033	530
Obst	1	1	1	1	1	1	1
süß	1	1	3	90	2	90	3
kalt	1	1	1	73	1	73	1
langsam	31	276	1117	954	11	956	479
wünschen	121	93	700	840	207	840	691
Fluss	18	70	31	174	641	174	30
schön	1041	976	171	140	264	140	176
Fenster	53	61	88	750	27	750	87
Bürger	28	34	27	557	12	557	25
Fuss	112	72	592	573	641	573	603
Spinne	2	2	1	13	1	13	1
sauer	18	30	13	141	16	141	13
Erde	3	2	8	153	6	153	8
hart	85	62	740	717	799	718	758
Magen	1	1	1	4	1	4	1
Lampe	105	104	105	104	105	104	105
gelb	1	3	1	2	1	2	1
Brot	445	507	1166	1148	329	1153	1117
Licht	4	3	7	33	7	33	7
schnell	143	142	3980	3186	132	3245	747
blau	200	123	521	510	536	510	524
Ozean	5	4	30	30	30	30	30
Kopf	482	529	290	515	210	515	302
Ofen	179	238	430	507	162	507	417
Religion	76	74	143	142	150	142	144
bitter	68	66	73	67	76	67	73
Hammer	1	5	1	3	1	3	1
laut	1185	1013	8691	8027	7372	9427	9689
ruhig	163	684	725	693	686	697	694
Salz	2	2	3	5	3	5	3
Straße	305	780	3825	2944	2803	3190	3195
König	64	50	426	598	226	598	396
Käse	11	10	14	64	15	64	14

Für die Darstellung der Ergebnisse werden abermals die bekannten Tabellen verwendet. Aufgrund der großen Anzahl an getesteten Verfahren sind auch hier die Ergebnistabellen zweigeteilt. Tabelle 20 und Tabelle 21 geben den Gesamtüberblick aller erzielten Ergebnisse.

Tabelle 21. Detailergebnisse Fallstudie 2, Testreihe B, Teil 2

ws ±12	Point Estimates of Association Strength						Measures from Information Theory			Heuristic Measures	
Stimulus	dice	jaccard	g-mean	MS	odds-ratio	liddel	PMI	local MI	average MI	MI^2	MI^3
Butter	32	32	198	41	37	375	375	13	14	198	38
rot	1	1	1	1	13	268	268	1	1	1	1
Tisch	199	199	541	346	1681	664	664	172	324	541	437
dunkel	1	1	1	1	95	115	115	1	1	1	1
Musik	4	4	182	1	154	424	424	3	4	182	15
weich	418	418	423	416	169	416	416	430	427	423	430
essen	1	1	1	1	271	237	237	1	1	1	1
Berg	32	32	227	5	207	347	347	9	14	227	131
Haus	157	157	1717	215	637	3311	3311	284	530	1717	406
Obst	1	1	1	1	10	139	139	1	1	1	1
süß	11	11	90	15	70	155	155	3	3	90	8
kalt	1	1	49	1	94	129	129	1	1	49	2
langsam	757	757	773	251	11	1424	1424	62	479	773	301
wünschen	199	199	809	157	1072	998	998	528	691	809	697
Fluss	218	218	173	220	5	220	220	24	30	173	135
schön	197	197	142	809	1282	135	135	191	176	142	154
Fenster	65	65	715	22	139	1177	1177	63	87	715	166
Bürger	10	10	493	3	111	1497	1497	14	25	493	58
Fuss	581	581	579	253	761	542	542	625	603	579	625
Spinne	16	16	13	17	10	17	17	1	1	13	5
sauer	7	7	141	4	333	243	243	12	13	141	23
Erde	11	11	153	3	178	247	247	5	8	153	76
hart	754	754	728	463	1098	685	685	776	758	728	782
Magen	1	1	4	1	58	72	72	1	1	4	1
Lampe	101	101	104	101	38	101	101	105	105	104	105
gelb	3	3	2	4	2	361	361	1	1	2	1
Brot	1049	1049	1104	316	325	1227	1227	1005	1117	1104	1001
Licht	7	7	33	8	496	522	522	7	7	33	9
schnell	169	169	2247	256	769	5398	5398	259	747	2247	330
blau	501	501		141	359	487	487	531	524	514	530
Ozean	30	30	30	30	19	30	30	30	30	30	30
Kopf	196	196	521	253	1409	636	636	191	302	521	445
Ofen	525	525	500	562	163	562	562	357	417	500	441
Religion	137	137	142	136	130	136	136	146	144	142	144
bitter	67	67	67	67	119	67	67	73	73	67	69
Hammer	1	1	3	1	256	43	43	1	1	1	1
laut	3940	3940	8134	3644	4922	8321	8321	7375	9689	8134	7477
ruhig	704	704	646	704	9	704	704	672	694	646	565
Salz	3	3	5	3	75	303	303	3	3	5	3
Straße	436	436	2075	496	44	3167	3167	2260	3195	2075	1146
König	182	182	595	120	628	673	673	357	396	595	537
Käse	12	12	62	10	193	532	532	13	14	62	20

Die zusammengefasste Form der Ergebnisse dieser Testreihe wird im Balkendiagramm der Abbildung 15 visualisiert. Diese fokussiert die Anzahl der korrekt vorhergesagten Primärantworten.

Die besten Ergebnisse in dieser Testreihe in Fenstergröße ±12 lieferte CIMAWA mit 10 richtig vorhergesagten Primärantworten. Danach folgen insgesamt fünf weitere Verfahren (‘poisson-stirling‘, ‘t-score‘, ‘log-likelihood‘, ‘local MI‘, ‘average MI‘), die unter den gegebenen Umgebungsparametern 9 Primärantworten richtig prognostizierten. Weitere fünf Verfahren bilden das breite Mittelfeld (‘WR‘, ‘dice‘, ‘jaccard‘, ‘MS‘, ‘MI^2‘, ‘MI^3‘) und sagen 5 – 8 Primärantworten richtig voraus. Jeweils 4 mal richtig lagen drei Berechnungsverfahren (‘z-score‘, ‘chi-squared‘, ‘g-mean‘), wobei drei Verfahren keine richtigen Ergebnisse liefern konnten (‘odds-ratio‘, ‘liddel‘, ‘PMI‘).

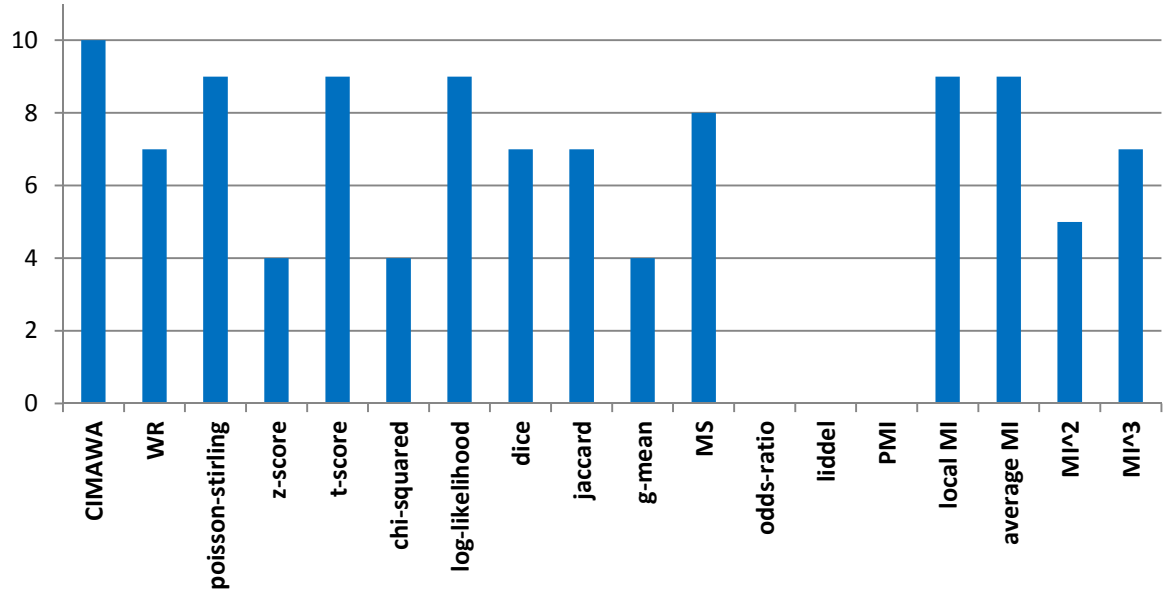


Abbildung 15. Zusammenfassung Fallstudie 2, Testreihe B

Eine vergleichende Analyse der beiden Testreihen bringt zwei Hauptunterschiede zum Vorschein. Auf der einen Seite werden die Ergebnisse offenbar quantitativ verbessert, wenn das Textfenster vergrößert wird. Wie den entsprechenden Tabellen zu entnehmen ist, schaffen es mehr Primärantworten in die Ergebnislisten der getesteten Verfahren. Diese Entwicklung ist bei Vergrößerung des Textfensters zu erwarten, da die reine Anzahl an Kookkurrenzkandidaten bei Vergrößerung des Textfensters ebenfalls vergrößert wird. Die Wahrscheinlichkeit, dass Stimulus und Primärantwort gemeinsam in einem Textfenster vorkommen, steigt analog zur Vergrößerung des Textfensters. Die Argumentation ist folglich die gleiche wie bei der Vergrößerung des zugrunde liegenden Korpus. Auf der anderen Seite ist zu beobachten, dass die quantitative Verbesserung der Ergebnisse eine qualitative Verschlechterung mit sich bringt. Das bedeutet wiederum, dass bei kleinerer Fenstergröße zwar weniger Primärantworten in den Ergebnislisten auftauchen, diese Ergebnisse jedoch bezüglich der tatsächlich gefundenen Primärantworten (Platzierung der Primärantwort auf dem ersten Platz der Assoziationsliste) als besser einzustufen sind. Diese Entwicklung zeigt sich bei allen in dieser Fallstudie getesteten Verfahren. Keines der getesteten Verfahren war in der Lage, die Anzahl an korrekt vorhergesagten Primärantworten zu erhöhen. Gleichbleibend gute Ergebnisse lieferten dabei die folgenden Verfahren: ‘MS‘, ‘MI^2‘ und ‘MI^3‘. Die Verfahren ‘liddel‘ und ‘PMI‘ konnten auch in dieser Testreihe keine korrekten Primärantworten berechnen. Die übrigen 13 Verfahren verschlechterten ihre Ergebnisse in dieser Testreihe im Vergleich zu Testreihe A.

5.3 Fallstudie 3: Korpusgrößenabhängiger Vergleich statistischer Assoziationsberechnungsverfahren

Im Vergleich zu den ersten beiden Fallstudien, in denen vornehmlich das selbst erstellte Korpus Verwendung fand, werden in der dritten Fallstudie sämtliche Kookkurrenzberechnungen auf einem öffentlichem Korpus [79] des Instituts für Deutsche Sprache (IDS) der Universität Mannheim durchgeführt. Mit Hilfe von 'COSMAS II', einer am IDS konzipierten Volltextdatenbank für die linguistisch motivierte Recherche in den Textsammlungen des IDS [83], wurden die entsprechenden Statistiken abgerufen. Dabei steht COSMAS II für: 'Corpus Search, Management and Analysis System' und ist das Nachfolgesystem von COSMAS I (1991-2003) am IDS [83]. COSMAS II beinhaltet eine Funktion, die es dem Nutzer ermöglicht, Auswertungen auf wählbaren Teilbereichen der zur Verfügung gestellten Texsammlung durchzuführen. Neben der möglichen Auswahl eines oder mehrerer Archive (z.B. geschriebene Korpora, historische Korpora, Süddeutsche Zeitung etc.) kann die Textsammlung auch rein quantitativ oder genauer gesagt, prozentual aufgeteilt werden. Dieser Funktion wird sich in Fallstudie 3 bedient, denn die vier Testreihen dieser Studie unterscheiden sich in der Größe des zugrunde liegenden Korpus. Beginnend mit einem vergleichsweise kleinen Korpus mit 84 Mio. Worten (~ 3% des Gesamtkorpus; Testreihe A), wird das Korpus sukzessive auf 10% und 240 Mio. Worte (Testreihe B), 50% und 1,4 Mrd. Worte (Testreihe C) und schließlich 100% mit insgesamt ca. 2,8 Mrd. Worten (Testreihe D) vergrößert. Dieses Vorgehen macht Aussagen über die mindestens erforderliche Korpusgröße möglich und zeigt, welche Berechnungsverfahren bei der jeweiligen Korpusgröße die besten Ergebnisse liefern. Zusätzlich wurden sämtliche Testreihen mit den Textfenstergrößen ±5 und ±12 durchgeführt. Durch dieses Vorgehen werden in dieser abschließenden Fallstudie sämtliche Einflussparameter variiert. In Abbildung 20 und
Abbildung 21 werden die erzielten Ergebnisse für jede getestete Textfenstergröße separat zusammengefasst.
Die Basis für die präsentierten Simulationsergebnisse in Fallstudie 3 wurden zum Teil im Rahmen einer vom Autoren betreuten Masterarbeit [84] am Institut für Wissensbasierte Systeme erarbeitet.

5.3.1 Fallstudie 3: Testreihe A

Der ersten Testreihe dieser Fallstudie liegt das kleinste Korpus zugrunde. Die Textsammlung umfasst ca. 84 Mio. Worte, was in etwa 3% des Gesamtkorpus entspricht. Es wurden insgesamt 18 Verfahren getestet, wobei jedes Einzelverfahren mit den Textfenstergrößen ±5 und ±12 implementiert wurde.
Die Zusammenfassung der erzielten Ergebnisse findet in Form der bereits mehrfach verwendeten Balkendiagramme statt. Auf der Hochachse ist die Anzahl der richtig vorhergesagten Primärantworten abgetragen. In Abbildung 16 sind die Ergebnisse der Testreihe A dargestellt. Da jedes Verfahren jeweils mit den Textfenstergrößen ±5 und ±12 getestet wurde, sind auch in Abbildung 16 jeweils 2 Werte abzulesen. Der linke Balken

repräsentiert die Anzahl der korrekten Primärantworten für die Textfenstergröße ±12 und der rechte diejenigen der kleineren Fenstergröße ±5.

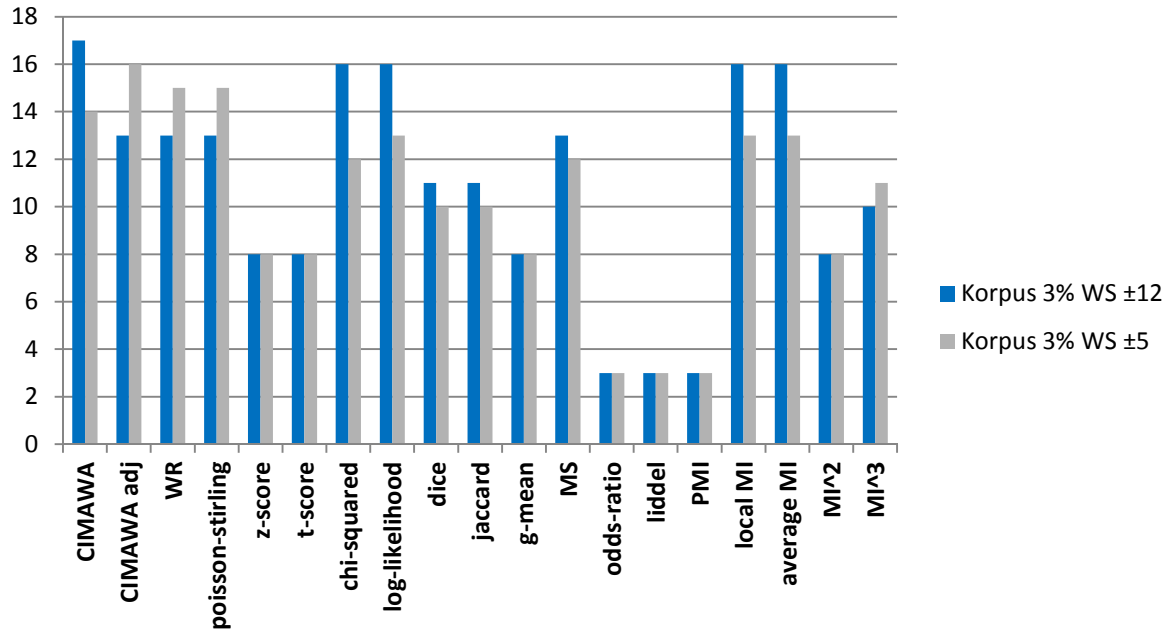


Abbildung 16. Zusammenfassung Fallstudie 3, Testreihe A

Zunächst werden die Resultate der Textfenstergröße ±12 eingeordnet. Die besten Ergebnisse wurden hier von CIMAWA erzielt. Dieses Verfahren konnte 17 der Primärantworten liefern. Der Abstand zu den nächst besseren Verfahren ist bei dieser Korpusgröße gering, denn mit 16 Primärantworten zeigten vier weitere Verfahren (‘chi-squared‘, ‘log-likelihood‘, ‘local MI‘, ‘average MI‘) ähnlich gute Ergebnisse. Diesen folgen vier Verfahren mit jeweils 13 korrekten Primärantworten (‘CIMAWA adjusted‘, ‘WR‘, ‘poisson-stirling‘, ‘MS‘). Die nächsten sieben Verfahren bilden das Mittelfeld mit 8 – 11 richtigen Primärantworten (‘z-score‘, ‘t-score‘, ‘dice‘, ‘jaccard‘, ‘g-mean‘, ‘MI^2‘, ‘MI^3‘). Schlusslichter in dieser Betrachtung sind ‘odds-ratio‘, ‘liddel‘ und ‘PMI‘ mit jeweils 3 Richtigen.

Bei verkleinerter Textfenstergröße stellte sich ‘CIMAWA adjusted‘ als bestes Verfahren heraus. In diesem Fall konnten 16 Primärantworten korrekt vorhergesagt werden. Nur eine richtige Antwort dahinter platzierten sich die zweitbesten Verfahren (‘WR‘, ‘poisson-stirling‘), gefolgt von ‘CIMAWA standard‘ mit 14 Primärantworten. Zwischen 8 und 13 richtige Antworten berechneten die Verfahren ‘z-score‘, ‘t-score‘, ‘chi-squared‘, ‘log-likelihood‘, ‘dice‘, ‘jaccard‘, ‘g-mean‘, ‘MS‘, ‘local MI‘, ‘average MI‘, ‘MI^2‘ und ‘MI^3‘, wobei erneut die schwächsten Ergebnisse von ‘odds-ratio‘, ‘liddel‘ und ‘PMI‘ erreicht wurden.

Vergleicht man die Resultate der beiden Fenstergrößen, so ist kein eindeutiger Trend abzuleiten. Sieben der getesteten Verfahren erzielen mit beiden Fenstergrößen identische Ergebnisse (‘z-score‘, ‘t-score‘, ‘g-mean‘, ‘odds-ratio‘, ‘liddel‘, ‘PMI‘, ‘MI^2‘). Bei acht Verfahren verschlechtern sich die Simulationsergebnisse bei verkleinerter Fenstergröße (‘CIMAWA standard‘, ‘chi-squared‘, ‘log-likelihood‘, ‘dice‘, ‘jaccard‘, ‘MS‘, ‘local MI‘,

‘average MI‘), wobei vier Verfahren (‘CIMAWA adjusted‘, ‘WR‘, ‘poisson-stirling‘, ‘MI^3‘) mit der kleineren Fenstergröße jeweils bessere Ergebnisse liefern.
Summiert man die insgesamt vorhergesagten Primärantworten über alle Verfahren für die beiden Fenstergrößen, so kann ein leichter Trend zur Fenstergröße ±12 ausgemacht werden. Mit dieser Fenstergröße werden über alle Verfahren 206 richtige Antworten vorausgesagt, wobei die verkleinerte Fenstergröße zu 195 richtigen Antworten führt. Zusätzlich für diese Textfenstergröße spricht, dass das absolut beste Ergebnis dieser Testreihe durch ‘CIMAWA standard‘ mit 17 korrekten Primärantworten, ebenfalls mit Textfenstergröße ±12, erreicht wurde.

5.3.2 Fallstudie 3: Testreihe B

Testreihe B vergrößert das Korpus auf 10 % des Gesamtkorpus und umfasst ca. 280 Mio. Worte. Die getesteten Verfahren sowie der Aufbau der Ergebnispräsentation sind in der bereits beschriebenen Weise unverändert. In Abbildung 17 werden die Ergebnisse der Testreihe B zusammengefasst.

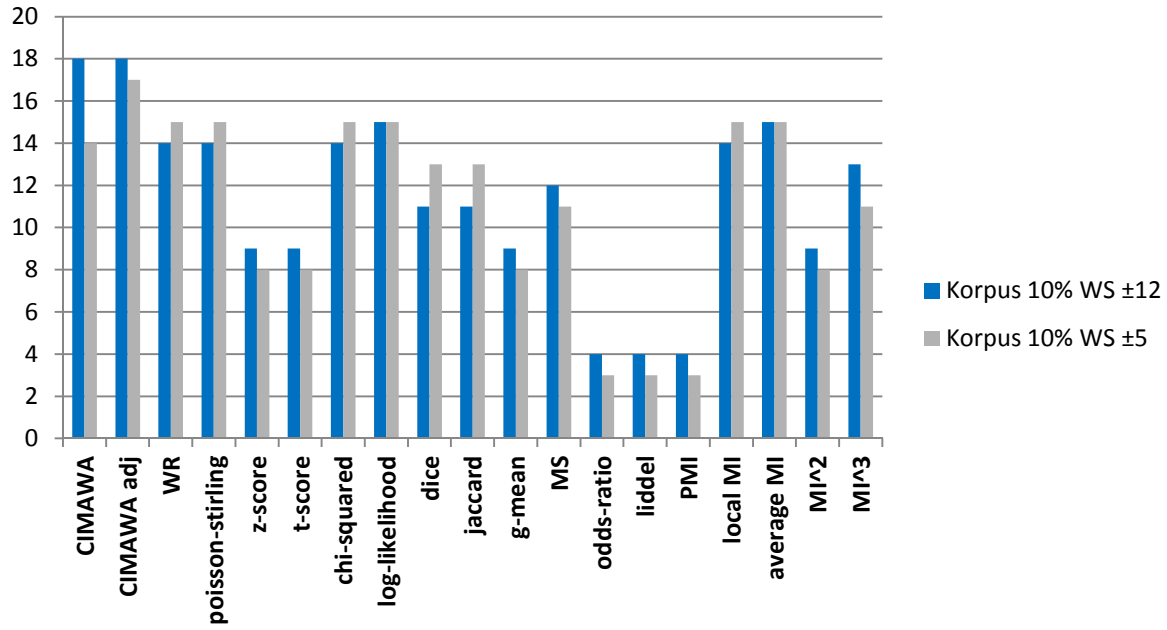


Abbildung 17. Zusammenfassung Fallstudie 3, Testreihe B

Legt man die Fenstergröße ±12 zugrunde, konnten die besten Ergebnisse bei CIMAWA und ‘CIMAWA adjusted‘ beobachtet werden. Mit 18 richtigen Primärantworten konnte in diesen Fällen das beste Resultat erzielt werden. Darauf folgen zwei Verfahren mit jeweils 15 Primärantworten (‘log-likelihood‘, ‘average MI‘) und vier mit 14 (‘WR‘, ‘poisson-stirling‘, ‘chi-squared‘, ‘local MI‘). Zwischen 9 und 13 korrekten Primärantworten werden von den folgenden acht Verfahren berechnet: ‘z-score‘, ‘t-score‘, ‘dice‘, ‘jaccard‘, ‘g-mean‘, ‘MS‘, ‘MI^2‘, ‘MI^3‘. Mit 4 Primärantworten erneut die schwächsten Ergebnisse erzielen die Verfahren ‘odds-ratio‘, ‘liddel‘ und ‘PMI‘.

Ähnlich gute Ergebnisse zeigt ‘CIMAWA adjusted‘ bei Textfenstergröße ±5. Siebzehn richtige Primärantworten sind zu verzeichnen, was zwei Antworten mehr entspricht als das Ergebnis der zweitbesten Berechnungsverfahren. Sechs Verfahren teilen sich diesen zweiten Platz (‘WR‘, ‘poisson-stirling‘, ‘chi-squared‘, ‘log-likelihood‘, ‘local MI‘, ‘average MI‘) bei Fenstergröße ±5. Mit 14 richtigen Antworten folgt CIMAWA. Das Mittelfeld bilden auch hier die Verfahren ‘z-score‘, ‘t-score‘, ‘dice‘, ‘jaccard‘, ‘g-mean‘, ‘MS‘, ‘MI^2‘ und ‘MI^3‘, indem sie zwischen 8 und 13 Primärantworten richtig berechnen. Mit 3 Primärantworten abermals die schwächsten Ergebnisse erreichen die Verfahren ‘odds-ratio‘, ‘liddel‘ und ‘PMI‘.
Elf der getesteten Verfahren (CIMAWA, ‘CIMAWA adjusted‘, ‘z-score‘, ‘t-score‘, ‘g-mean‘, ‘MS‘, ‘odds-ratio‘, ‘liddel‘, ‘PMI‘, ‘MI^2‘, ‘MI^3‘) zeigen die besseren Ergebnisse bei der großen Fenstergröße, wobei sechs Assoziationsberechnungsverfahren (‘WR‘, ‘poisson-stirling‘, ‘chi-squared‘, ‘log-likelihood‘, ‘dice‘, ‘jaccard‘) bessere Ergebnisse in der kleinen Fenstergröße generieren konnten und bei zweien (‘log-likelihood‘, ‘average MI‘) identische Ergebnisse zustande kamen.
Die Gesamtanzahl der korrekt vorhergesagten Antworten liegt bei Fenstergröße ±12 bei 217 und bei ±5 mit 210 etwas darunter. Der Spitzenwert von 18 richtigen Vorhersagen mit ‘CIMAWA adjusted‘ konnte im Vergleich zur kleineren Korpusgröße um das Minimum von einer Antwort gesteigert werden. Bei beiden Fenstergrößen war im Vergleich zu Testreihe A ein leichter Anstieg der insgesamt gefundenen Antworten zu verzeichnen.

5.3.3 Fallstudie 3: Testreihe C

Eine weitere Vergrößerung der zugrunde liegenden Textsammlung wird in Testreihe C vorgenommen. Der Umfang des Korpus wird verglichen mit Testreihe B um das fünffache, auf 50% des Gesamtkorpus erhöht. Das Korpus für diese Testreihe umfasst ca. 1,4 Mrd. Worte. In Abbildung 18 werden die erzielten Ergebnisse zusammengefasst.

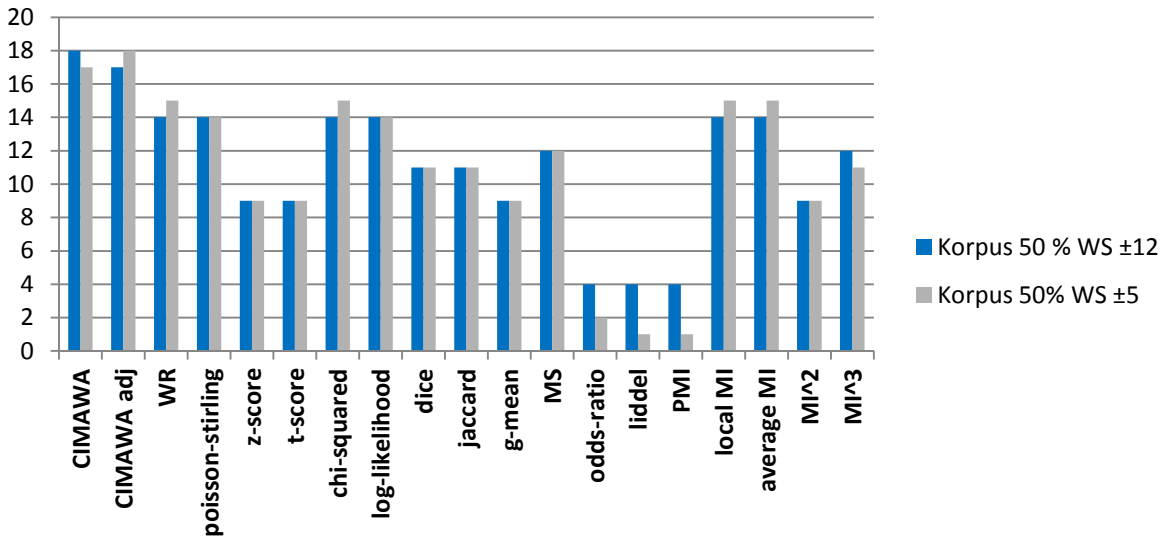


Abbildung 18. Zusammenfassung Fallstudie 3, Testreihe C

Abermals erreicht CIMAWA in der Fenstergröße ±12 das beste Ergebnis und sagt 18 der Primärantworten korrekt voraus, wobei ‘CIMAWA adjusted‘ mit 17 Primärantworten nur knapp dahinter liegt. Es folgen insgesamt sechs Verfahren (‘WR‘, ‘poisson-stirling‘, ‘chi-squared‘, ‘log-likelihood‘, ‘local MI‘, ‘average MI‘) mit jeweils 14 richtigen Voraussagen. Insgesamt acht Verfahren (‘z-score‘, ‘t-score‘, ‘dice‘, ‘jaccard‘, ‘g-mean‘, ‘MS‘, ‘MI^2‘, ‘MI^3‘) liefern zwischen 9 und 12 richtigen Antworten. Abgeschlagen mit 4 Antworten bilden ‘odds-ratio‘, ‘liddel‘ und ‘PMI‘ das Ende der Rangfolge.
Bewertet man die Ergebnisse der verkleinerten Fenstergröße, so ist festzustellen, dass ‘CIMAWA adjusted‘ mit 18 Primärantworten das Topergebnis liefert. CIMAWA folgt auf dem zweiten Platz mit einer Antwort weniger. Die drittbesten Ergebnisse werden von vier Verfahren (‘WR‘, ‘chi-squared‘, ‘local MI‘, ‘average MI‘) mit jeweils 15 Antworten erzielt. Es folgen ‘poisson-stirling‘ und ‘log-likelihood‘ mit einer Anzahl von 14 Primärantworten. Wiederum mittelmäßige Ergebnisse generieren die folgenden Verfahren: ‘z-score‘, ‘t-score‘, ‘dice‘, ‘jaccard‘, ‘g-mean‘, ‘MS‘, ‘MI^2‘, ‘MI^3‘, bei Ergebnissen zwischen 9 und 12 richtigen Vorhersagen. Mit zwei bzw. drei Primärantworten abermals die schwächsten Ergebnisse erreichen die Verfahren ‘odds-ratio‘, ‘liddel‘ und ‘PMI‘.
Fünf der getesteten Verfahren (CIMAWA, ‘odds-ratio‘, ‘liddel‘, ‘PMI‘, ‘MI^3‘) haben bei größerer Fenstergröße bessere Ergebnisse vorzuweisen, wobei weitere fünf Verfahren (‘CIMAWA adjusted‘, ‘WR‘, ‘chi-squared‘, ‘local MI‘, ‘average MI‘) bei der kleineren Fenstergröße besser zu funktionieren scheinen. Bei den restlichen neun Verfahren (‘poisson-stirling‘, ‘z-score‘, ‘t-score‘, ‘log-likelihood‘, ‘dice‘, ‘jaccard‘, ‘g-mean‘, ‘MS‘, ‘MI^2‘) sind die Resultate der beiden Fenstergrößen identisch.
Die Gesamtanzahl der Primärantworten mit Fenstergröße ±12 liegt bei dem 50%-Korpus bei 213. Zum Vergleich beträgt die Summe der Primärantworten mit der verkleinerten Fenstergröße 208.

5.3.4 Fallstudie 3: Testreihe D

In Testreihe D dient der Gesamtkorpus mit ca. 2,8 Mrd. Worten als Basis für die Assoziationsberechnung. Auch in dieser Testreihe werden die bekannten Verfahren getestet und in zwei Textfenstergrößen implementiert. Abbildung 19 fasst die Simulationsergebnisse in bewährter Form zusammen.
Der Trend aus den ersten drei Testreihen bestätigt sich auch in dieser: Bei großer Fenstergröße liegt CIMAWA mit 18 Primärantworten in der Prognose vor ‘CIMAWA adjusted‘ mit 16 richtigen Antworten. Ihnen folgen vier Verfahren (‘chi-squared‘, ‘log-likelihood‘, ‘local MI‘, ‘average MI‘) mit jeweils 13 Richtigen. Die zehn Verfahren (‘WR‘, ‘poisson-stirling‘, ‘z-score‘, ‘t-score‘, ‘dice‘, ‘jaccard‘, ‘g-mean‘, ‘MS‘, ‘MI^2‘, ‘MI^3‘) erreichen 10 - 12 Primärantworten. Mit zwei bzw. drei Primärantworten erneut die schwächsten Ergebnisse erzielen die Verfahren ‘odds-ratio‘, ‘liddel‘ und ‘PMI‘.

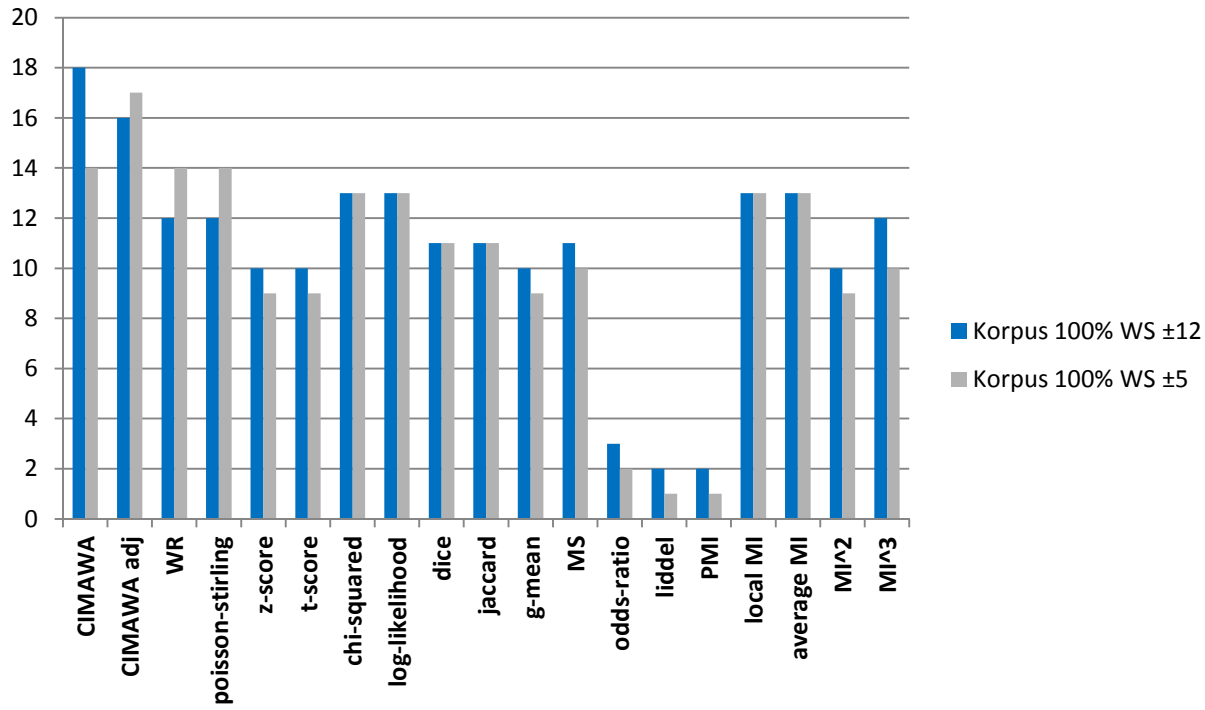


Abbildung 19. Zusammenfassung Fallstudie 3, Testreihe D

Auch die kleine Fenstergröße zeigt ähnliche Ergebnisse wie die ersten Testreihen. Hier legt erneut ‘CIMAWA adjusted‘ das beste Ergebnis mit 17 Antworten vor. Die zweitbesten Resultate mit Fenstergröße ±5 liefern drei Verfahren (CIMAWA, ‘WR‘, ‘poisson-stirling‘) mit jeweils 14 Primärantworten. Zwischen 9 und 13 richtige Antworten prognostizieren 12 Verfahren (‘z-score‘, ‘t-score‘, ‘chi-squared‘, ‘log-likelihood‘, ‘dice‘, ‘jaccard‘, ‘g-mean‘, ‘MS‘, ‘local MI‘, ‘average MI‘, ‘MI^2‘, ‘MI^3‘) korrekt.
Zehn der getesteten Verfahren (CIMAWA, ‘z-score‘, ‘t-score‘, ‘g-mean‘, ‘MS‘, ‘odds-ratio‘, ‘liddel‘, ‘PMI‘, ‘MI^2‘, ‘MI^3‘) erzielen die besten Ergebnisse mit der großen Fenstergröße, während drei Verfahren (‘CIMAWA adjusted‘, ‘WR‘, ‘poisson-stirling‘) mit der kleinen Fenstergröße die besseren Ergebnisse erreichen.
Die Gesamtanzahl der Primärantworten liegt bei 202 mit Fenstergröße ±12 und bei 193 mit der Fenstergröße ±5.

5.3.5 Fallstudie 3: Zusammenfassung der Ergebnisse

Im letzten Abschnitt wurden die Einzelergebnisse der Fallstudie 3 detailliert behandelt und deren Ergebnisse dargestellt. Um eine weitergehende Interpretation zu ermöglichen, ist eine Zusammenfassung der Resultate zielführend. Diese vereinfacht den Vergleich zwischen den einzelnen Testreihen und ermöglicht das Erkennen von Trends und Tendenzen zwischen den gemachten Experimenten.
Die Zusammenfassung erfolgt zweigeteilt. Da sämtliche Testreihen dieser Fallstudie Tests in den Fenstergrößen ±5 und ±12 enthalten, werden die Ergebnisse getrennt nach Fenstergröße dargestellt. Zunächst sind die Resultate der Testreihen A – D für die Textfenstergröße ±5

aufgearbeitet. Abbildung 20 fasst diese Resultate der Assoziationsberechnungsverfahren zusammen, wobei der zweite Teil der Zusammenfassung in
Abbildung 21 einen Überblick der Resultate in Fenstergröße ±12 ermöglicht.

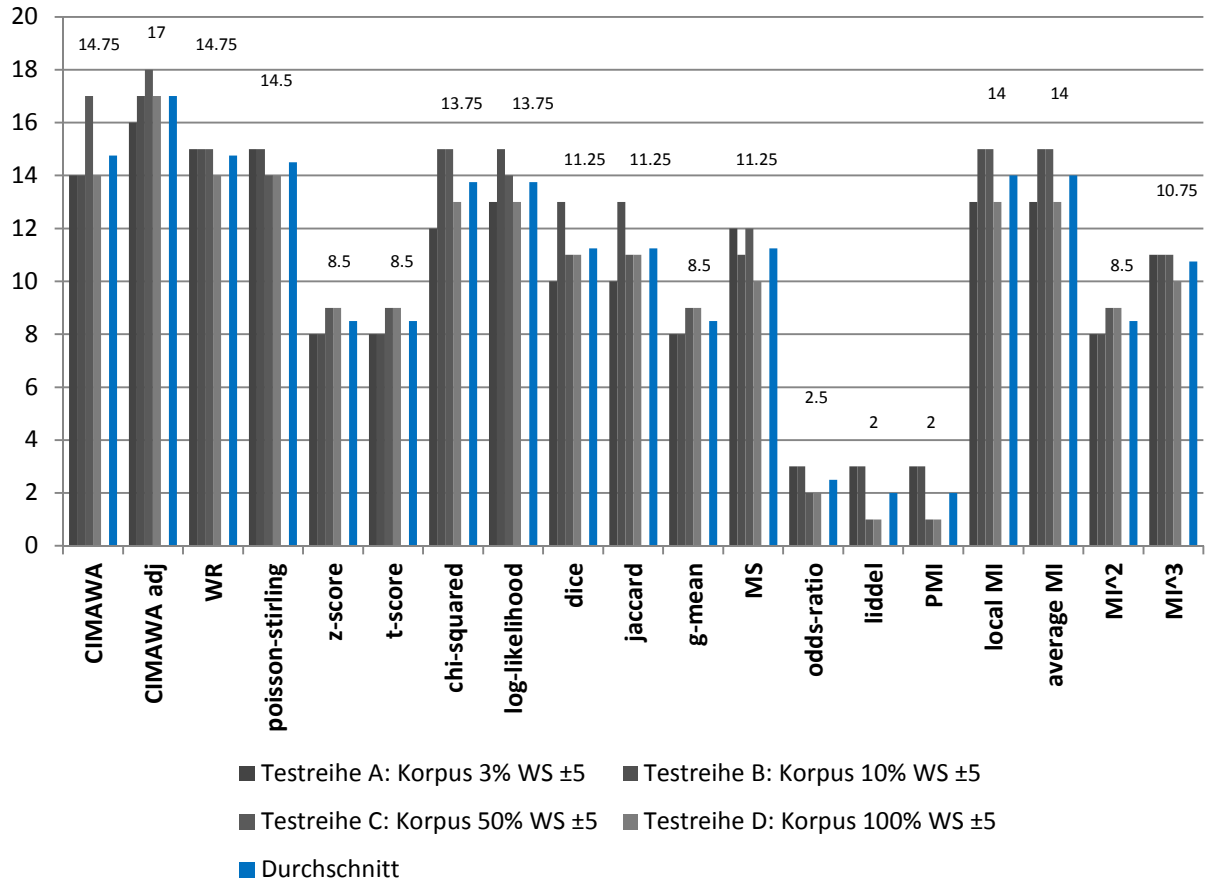


Abbildung 20. Zusammenfassung Fallstudie 3, Textfenstergröße ±5

Für jedes der getesteten Verfahren sind in den Diagrammen jeweils fünf Werte abgetragen. Die ersten vier Balken repräsentieren dabei die Testreihen A – D. Der fünfte Balken visualisiert den Durchschnittswert der Testreihen für das jeweilige Verfahren und macht so Aussagen über die Qualität der Resultate für die entsprechende Fenstergröße möglich. Zusätzlich wird der berechnete Durchschnittswert in den Diagrammen numerisch angegeben. Abschließend werden in Tabelle 24 die Simulationsergebnisse der CIMAWA-Assoziationsberechnung als Ganzes dargestellt und diskutiert.
An einem Beispiel aus Abbildung 20 wird exemplarisch verdeutlicht, wie die zusammenfassenden Balkendiagramme zu interpretieren sind. Wie bereits aus zahlreichen Beispielen in dieser Arbeit bekannt, sind auf der Hochachse die Anzahl der korrekt berechneten Primärantworten des FAT und auf der Längsachse die getesteten Verfahren abgetragen. Die Besonderheit von Abbildung 20 besteht darin, dass für jedes Verfahren alle Ergebnisse aus Fallstudie 3 für die Fenstergröße ±5 enthalten sind. Da zusätzlich je Verfahren ein Durchschnittswert berechnet und dargestellt wurde, ergeben sich für jedes Verfahren 5 Werte. Das ist eines für jede Testreihe dieser Fallstudie, ergänzt durch eben beschriebenen

berechneten Durchschnitt. Am Beispiel von ‘CIMAWA adjusted‘ (zweites Verfahren von links in Abbildung 20) werden die enthaltenen Werte erklärt. Von links angefangen beträgt der erste Wert für ‘CIMAWA adjusted‘ 16. Wie zusätzlich der Legende von Abbildung 20 zu entnehmen, stammt dieser Wert aus Testreihe A und der zugehörigen Korpusgröße von 3% der Gesamttextsammlung. Die nächsten drei Werte zeigen die Ergebnisse bei sukzessiver Vergrößerung der Textbasis. Den Bestwert mit 18 richtigen Primärantworten erreicht ‘CIMAWA adjusted‘ in Testreihe C, d. h. bei einer Korpusgröße von 50%. Dementsprechend kann in Abbildung 20 für jedes Verfahren ein Trend abgelesen werden, welche Korpusgröße zu den besten Ergebnissen führt. Im Falle von ‘CIMAWA adjusted‘ beträgt der Durchschnitt über die vier Testreihen 17, was als numerischer Eintrag über den Verfahren kenntlich gemacht ist.
Bei genauer Analyse der Ergebnisse wird erkennbar, dass die 18 Primärantworten von ‘CIMAWA adjusted‘ den absoluten Bestwert der Fenstergröße ±5 über alle Korpusgrößen bedeuten. Auch beim berechneten Durchschnittswert zeigt sich, dass ‘CIMAWA adjusted‘ bei dieser Fenstergröße die besten Ergebnisse erzielt. Im Schnitt wurden 17 Primärantworten richtig vorhergesagt. Nimmt man die erzielten Durchschnittswerte zum Maßstab für die Güte der Resultate, folgen CIMAWA und ‘WR‘ mit 14,75 Primärantworten als nächstbeste Verfahren. Mit 14,5 Antworten im Durchschnitt liefert ‘poisson-stirling‘ ähnlich gute Resultate und liegt damit im oberen Mittelfeld der getesteten Verfahren.
Betrachtet man jedes Verfahren für sich und vergleicht die Ergebnisse, so sind keine eindeutigen Trends bezüglich der optimalen Korpusgröße für die Assoziationsberechnung zu beobachten. Es können Verfahren ausgemacht werden, deren Ergebnisse bei Vergrößerung des Korpus besser werden oder gleich bleiben (‘z-score‘, ‘t-score‘, ‘g-mean‘, ‘MI^2‘). Genauso können jedoch Verfahren benannt werden, die gegenteilige Entwicklungen zeigen (‘WR‘, ‘poisson-stirling‘, ‘odds-ratio‘, ‘liddel‘, ‘PMI‘, ‘MI^3‘). Bei den übrigen Verfahren sind die jeweils besten Ergebnisse in den beiden mittleren Korpusgrößen zu lokalisieren, wobei in diesen Fällen die Ergebnisse in den Testreihen mit sehr kleinem und sehr großem Korpus abfallen.
Um einen Gesamtüberblick zu gewähren, sind in Tabelle 22 die richtigen Antworten von allen getesteten Verfahren für jede Korpusgröße aufsummiert.
Bei der Analyse der Tabelle 22 scheint sich eine bereits angedeutete Tendenz zu bestätigen, denn die Korpusgröße 10% liefert in der Summe über allen Verfahren 210 Primärantworten und erreicht damit den Bestwert.

Tabelle 22. Gesamtergebnis Fenstergröße ±5

WS ±5	∑ Primärantworten
Korpus 3%	195
Korpus 10%	210
Korpus 50%	208
Korpus 100%	193
Gesamtergebnis WS ±5	806

Die nächstbessere Anzahl liefert Korpusgröße 50% mit 208 Antworten. Das zeigt, dass die erfolgversprechendste Textsammlungsgröße in diesem mittleren Bereich zu liegen scheint. Das Ergebnis für den vollen Korpus (100%) hingegen liegt bei 193 Antworten, wobei auf den

kleinsten Korpus 195 richtige Vorhersagen entfallen. Das Gesamtergebnis für ±5 über alle Korpora und Verfahren liegt bei 806 Primärantworten.

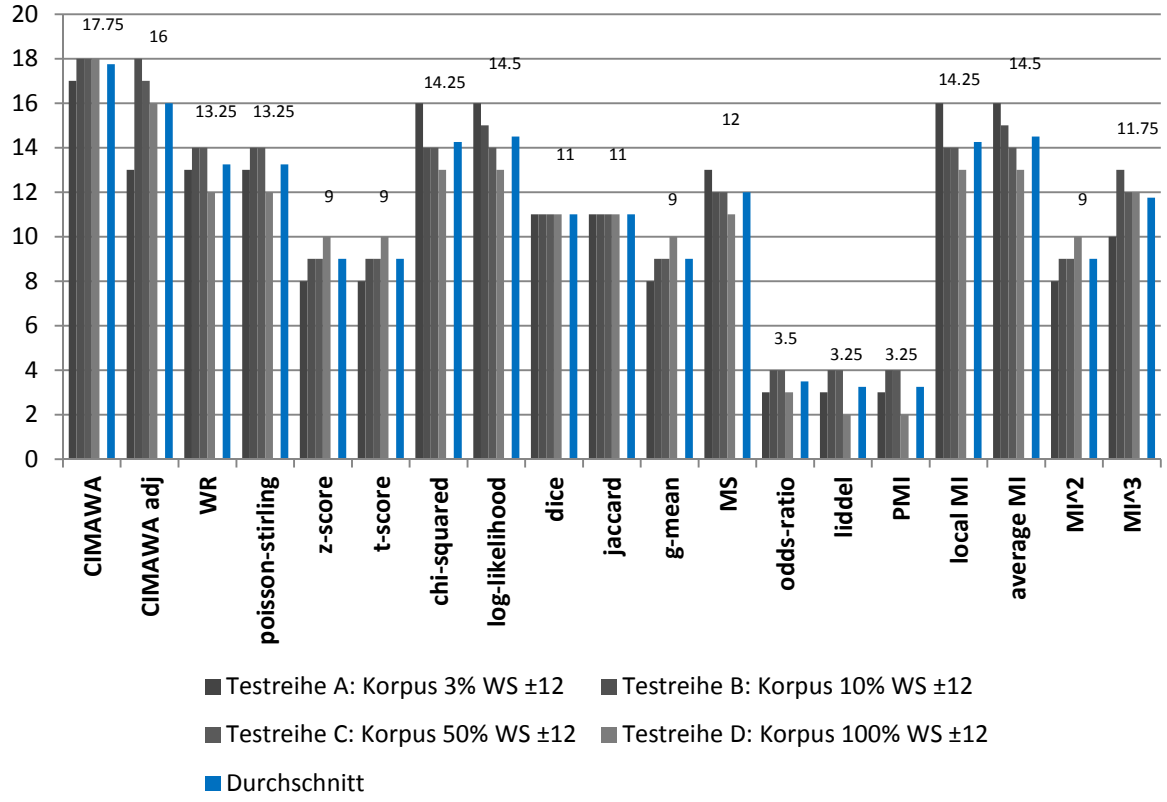


Abbildung 21. Zusammenfassung Fallstudie 3, Textfenstergröße ±12

Der zweite Teil der zusammenfassenden Auswertung ist
Abbildung 21 zu entnehmen. Hier werden die Ergebnisse in der zuvor beschriebenen Weise zusammengestellt und präsentiert. Einziger Unterschied ist die Grundlage der Auswertung, denn die folgenden Analysen beinhalten die Ergebnisse der Testreihen A – D, die mit Textfenstergröße ±12 berechnet wurden.

Den Bestwert für die Fenstergröße ±12 erzielte CIMAWA und ‘CIMAWA adjusted‘ mit 18 richtigen Primärantworten. CIMAWA konnte diese Anzahl an Antworten auf dem 10%-, 50%- und dem 100%-Korpus bestätigen. ‘CIMAWA adjusted‘ kam lediglich auf dem 10%-Korpus zu diesem Bestergebnis. Bestätigt werden die vielversprechenden Ergebnisse der beiden Verfahren nach der Betrachtung der Durchschnittswerte. Mit 17,75 richtigen Antworten im Schnitt legt CIMAWA den Bestwert für diese Textfenstergröße vor. ‘CIMAWA adjusted‘ liegt mit 16 Antworten im Durchschnitt auf dem ersten Verfolgerplatz. ‘Log-likelihood‘ und ‘average MI‘ simulieren 14,5 Primärantworten und bilden die Spitze des breiten Mittelfelds.

Ein klarer Trend, welche Korpusgröße für diese Fenstergröße die besten Ergebnisse liefert, kann auch an dieser Stelle nicht abgeleitet werden. Fünf implementierte Verfahren (CIMAWA, ‘z-score‘, ‘t-score‘, ‘g-mean‘, ‘MI^2‘) zeigen bei sukzessiver Vergrößerung der Textbasis eine stetige Verbesserung oder teils stagnierende Ergebnisse. Bei zwei Verfahren (‘dice‘, ‘jaccard‘) scheint die Korpusgröße keine entscheidende Rolle zu spielen, denn die

Resultate bleiben konstant. Weitere fünf Methoden (‘chi-squared‘, ‘log-likelihood‘, ‘MS‘, ‘local MI‘, ‘average MI‘) berechnen bei steigender Korpusgröße weniger oder zum Teil gleich viele Primärantworten korrekt. Bei den übrigen Assoziationsberechnungsverfahren werden die Bestwerte in den mittleren beiden Korpusgrößen erreicht.

Tabelle 23. Gesamtergebnis Fenstergröße ±12

WS ±12	∑ Primärantworten
Korpus 3%	206
Korpus 10%	217
Korpus 50%	213
Korpus 100%	202
Gesamtergebnis WS ±12	838

Die Gesamtergebnisse für die Fenstergröße ±12 sind Tabelle 23 zu entnehmen. Die meisten Antworten (217) wurden von den Verfahren auf dem 10%-Korpus errechnet. Nur wenig schwächer sind die Ergebnisse mit 213 Antworten bei Korpusgröße 50%. Die Resultate der kleinen und der kompletten Textsammlung fallen mit 206 bei 3% und 202 bei 100% ähnlich den zuvor gemachten Beobachtungen bei Fenstergröße ±5 leicht ab.
Insgesamt liegt die Fenstergröße ±12 mit 838 korrekt vorhergesagten Primärantworten nur leicht vor der Textfenstergröße ±5, die alles in allem 806 richtige Antworten generierte.
Analysiert man die erzielten Simulationsergebnisse, so kann zusammengefasst werden, dass CIMAWA bzw. ‘CIMAWA adjusted‘ die besten Ergebnisse erzielen. Aus diesem Grund werden in Tabelle 24 und Tabelle 25 die von CIMAWA berechneten Primärantworten für alle durchgeführten Testreihen der Fallstudie 3 aufgeführt. Der Tabelle zu entnehmen sind jeweils die am stärksten mit dem Stimuluswort assoziierten Begriffe, was den Begriffen mit den höchsten berechneten CIMAWA-Werten entspricht.

Tabelle 24. Detailergebnisse Fallstudie 3, CIMAWA, Teil 1

Stimulus	Primärantwort	Testreihe A		Testreihe B		Testreihe C		Testreihe D	
		WS ±5	WS ±12	WS ±5	WS ±12	WS ±5	WS ±12	WS ±5	WS ±12
Adler	Vogel	mannheimer	mannheimer	mannheimer	mannheimer	mannheimer	mannheimer	mannheimer	Mannheimer
arbeiten	faulenzen	beginnen	beginnen	erledigen	erledigen	beginnen	beginnen	erledigen	Erledigen
Arzt	Krankheit	Patient	Patient	Patient	Patient	Patient	Patient	Patient	Patient
ängstlich	Kind(er)	viel	viel	agieren	wirken	wirken	agieren	agieren	Wirken
Bad	Wasser	Kreuznach	Kreuznach	Kreuznach	Kreuznach	Kreuznach	Münster	Münster	Münster
Bequemlichkeit	Sessel	sicher	gewiss	sicher	sicher	mehr	mehr	sicher	Mehr
Berg	Tal	Tal	Tal	Tal	Tal	Tal	Tal	Tal	Tal
Bett	Schlaf	liegen	schlafen	liegen	schlafen	liegen	schlafen	liegen	Schlafen
Bibel	Buch	lesen	Gott	lesen	Gott	lesen	Gott	lesen	Gott
bitter	süß	nötig	nötig	nötig	nötig	nötig	nötig	nötig	Nötig
blau	Himmel	rot	rot	rot	rot	rot	rot	rot	Rot
Blüte	Blume	voll	voll	voll	voll	voll	voll	voll	Voll
Brot	essen	täglich	Wein	täglich	täglich	täglich	täglich	täglich	Täglich
Butter	Brot	Brot	Brot	Brot	Milch	Brot	Milch	Brot	Milch
Bürger	Staat	interessieren	interessieren	interessieren	interessieren	interessieren	interessieren	interessieren	Interessieren
Dieb	stehlen	unbekannt	unbekannt	Polizei	Polizei	stehlen	Polizei	unbekannt	Polizei
dunkel	hell(e)	licht	licht	licht	licht	licht	licht	licht	Licht
durstig	hungrig	hungrig	hungrig	hungrig	hungrig	hungrig	hungrig	hungrig	Hungrig
Erde	Himmel	Himmel	Himmel	Himmel	Mond	Himmel	Mond	Himmel	Mond
essen	trinken	trinken	trinken	trinken	trinken	trinken	trinken	trinken	Trinken
Fenster	Glas	öffnen	öffnen	öffnen	öffnen	öffnen	öffnen	öffnen	Öffnen
Fluß	Wasser	Wasser	Wasser	Wasser	Wasser	Wasser	Wasser	Wasser	Wasser
Frau	Mann	jung	jung	jung	jung	jung	jung	jung	Jung
Freude	Leid	gross	gross	gross	gross	gross	gross	gross	Gross
Fuß	Schuh(e)	fassen	fassen	fassen	fassen	fassen	fassen	fassen	Fassen
Gedächtnis	Gehirn	bleiben	rufen	bleiben	bleiben	bleiben	bleiben	bleiben	Bleiben
gelb	rot	rot	rot	rot	rot	rot	rot	grün	Farbe
Gerechtigkeit	Gericht	sozial	sozial	sozial	sozial	sozial	sozial	sozial	Sozial
Gesundheit	Krankheit	sozial	sozial	sozial	sozial	sozial	sozial	sozial	Sozial
glatt	Eis	laufen	laufen	laufen	laufen	laufen	laufen	laufen	Laufen
grün	Wiese	blau	blau	gelb	blau	blau	blau	blau	Blau
Hammelfleisch	essen	Bohne	Lamm	Bohne	Lamm	Bohne	Lamm	Lamm	Lamm
Hammer	Amboß	kommen	kommen	kommen	kommen	kommen	kommen	kommen	Kommen
Hand	Fuß	öffentlich	öffentlich	öffentlich	öffentlich	öffentlich	öffentlich	öffentlich	Öffentlich

Tabelle 25. Detailergebnisse Fallstudie 3, CIMAWA, Teil 2

Stimulus	Primärantwort	Testreihe A		Testreihe B		Testreihe C		Testreihe D	
		WS ±5	WS ±12	WS ±5	WS ±12	WS ±5	WS ±12	WS ±5	WS ±12
hart	weich	arbeiten	arbeiten	arbeiten	arbeiten	arbeiten	arbeiten	arbeiten	Arbeiten
Haus	Hof	weiss	wohnen	weiss	wohnen	weiss	wohnen	weiss	Wohnen
hoch	tief	Meter	Meter	Meter	Meter	Meter	Meter	Meter	Meter
hungrig	durstig	machen	machen	machen	machen	machen	machen	machen	Machen
Häuschen	Garten	klein	klein	klein	klein	klein	klein	klein	Klein
Junge	Mädchen	Mensch	Mensch	Mensch	Mensch	Mensch	Mensch	Mensch	Mensch
kalt	warm	lassen	warm	lassen	warm	lassen	warm	lassen	Warm
Kind	klein	Mutter	Mutter	Mutter	Eltern	Mutter	Eltern	klein	Klein
Kohl	Gemüse	Helmut	Helmut	Helmut	Helmut	Helmut	Helmut	Helmut	Helmut
kommandieren	befehlen	lass	Soldat	lass	lass	Mann	kontrollieren	lass	Lass
Kopf	Haar(e)	Hals	schlagen	Hals	schlagen	Hals	schlagen	Hals	Schlagen
Krankheit	Gesundheit	schwer	schwer	schwer	schwer	schwer	schwer	schwer	Schwer
kurz	lang	Mitternacht	Mitternacht	Mitternacht	Mitternacht	Mitternacht	Mitternacht	Mitternacht	Mitternacht
Käse	Butter	Milch	Milch	Milch	Milch	Milch	Milch	Milch	Milch
König	Kaiser	Ewald	Ewald	kardinal	kardinal	kardinal	kardinal	kardinal	Kardinal
Lampe	Licht	meist	meist	meist	meist	meist	meist	meist	Meist
lang	kurz	Jahr	Jahr	Jahr	Jahr	Jahr	Jahr	Jahr	Jahr
langsam	schnell	sicher	gehen	sicher	gehen	gehen	gehen	gehen	Gehen
Laut	leise	Polizei	Polizei	Polizei	Polizei	Polizei	Polizei	Polizei	Polizei
Licht	dunkel	grün	grün	grün	grün	grün	grün	grün	Grün
Löwe	Tiger	Stern	Stern	Stern	Stern	golden	Stern	Stern	Stern
Magen	Darm	schlagen	liegen	schwer	liegen	Darm	liegen	schwer	Liegen
Mann	Frau	jung	jung	jung	jung	jung	jung	jung	Jung
Mond	Stern(e)	Stern	Stern	Stern	Stern	Stern	Stern	Stern	Stern
Musik	Ton(Töne)	Text	Konzert	Tanz	Tanz	Text	Konzert	Text	Komponist
Mädchen	Junge	Junge	Junge	Junge	Junge	Junge	Junge	Junge	Junge
Nadel	spitz	golden	golden	golden	golden	golden	golden	golden	Golden
Obst	Gemüse	Gemüse	Gemüse	Gemüse	Gemüse	Gemüse	Gemüse	Gemüse	Gemüse
Ofen	Wärme	Holz	Holz	Schuss	Holz	backen	Holz	warm	Holz
Ozean	Meer	still	Wasser	still	Meer	tief	Meer	tief	Meer
pfeifen	singen	Spatz	Spatz	Orgel	Orgel	Spatz	Schiedsrichter	Loch	Schiedsrichter
Priester	Kirche	katholisch	Kirche	Orden	Kirche	Orden	Kirche	Orden	Kirche
Quadrat	Viereck	Cinema	Cinema	Cinema	Cinema	Cinema	Cinema	Cinema	Cinema
rauh	glatt	Inge	Inge	Inge	Inge	Inge	Inge	Inge	Inge
Religion	Glaube(n)	Kultur	Kultur	Islam	Islam	Islam	Islam	Islam	Islam
rot	grün	blau	blau	blau	blau	blau	blau	blau	Blau
ruhig	laut	bleiben	bleiben	bleiben	bleiben	bleiben	bleiben	bleiben	Bleiben
Salz	Zucker	Pfeffer	Pfeffer	Pfeffer	Pfeffer	Pfeffer	Pfeffer	Pfeffer	Pfeffer
sauer	süß	reagieren	Fussball	reagieren	Trainer	Rolf	reagieren	reagieren	Reagieren
Schaf	Wolle	schwarz	schwarz	schwarz	schwarz	schwarz	schwarz	schwarz	Schwarz
Schere	schneiden	kleben	kleben	kleben	kleben	arm	kleben	arm	Kleben
schlafen	Bett	gehen	Nacht	nachts	Nacht	legen	Nacht	legen	Nacht
Schmetterling	Falter	Meter	raupen	Meter	Meter	Meter	Meter	Meter	Meter
schnell	langsam	gehen	fahren	mögen	gehen	mögen	gehen	mögen	Gehen
schwarz	weiß	weiss	weiss	weiss	weiss	weiss	weiss	weiss	Weiss
schwer	leicht	krank	tun	krank	tun	tun	tun	fällen	Tun
schön	häßlich	ganz	ganz	ganz	ganz	ganz	ganz	ganz	Ganz
Soldat	Krieg	töten	Krieg	töten	Krieg	Ryan	Krieg	töten	Krieg
Sorge	Kummer	gross	gross	gross	gross	gross	gross	gross	Gross
Spinne	Netz	schwarz	schwarz	Netz	Netz	Netz	Netz	Netz	Netz
Stadt	Land	Salz	Salz	Salz	Salz	Salz	Salz	Salz	Salz
Stiel	Besen	Jörg	Jörg	Jörg	Jörg	Jörg	Jörg	Jörg	Jörg
Straße	Weg	richten	richten	richten	Strasse	richten	Uhr	richten	Richten
Stuhl	Bein(e)	Heilige	sitzen	sitzen	sitzen	sitzen	sitzen	sitzen	Sitzen
süß	sauer	Jud	Jud	sauer	sauer	Christian	sauer	Christian	Sauer
Tabak	rauchen	Austria	Austria	Austria	Austria	Austria	Austria	Austria	Austria
Teppich	weich	rot	rot	rot	rot	rot	rot	rot	Rot
tief	hoch	greifen	greifen	greifen	greifen	greifen	greifen	greifen	Greifen
Tisch	Stuhl	legen	legen	legen	legen	legen	legen	legen	Legen
träumen	schlafen	dürfen	Traum	dürfen	dürfen	dürfen	dürfen	dürfen	Dürfen
weich	hart	hart	hart	kochen	hart	hart	hart	kochen	Hart
weiß	schwarz	schwarz	genau	genau	genau	schwarz	genau	genau	Genau
Whisky	Schnaps	Flasche	Flasche	Flasche	Flasche	Flasche	Flasche	Flasche	Flasche
wünschen	Weihnachten	übrig	übrig	übrig	übrig	übrig	übrig	übrig	Übrig
Zorn	Wut	Stephan	Wut	Stephan	Wut	Stephan	Wut	Wut	Wut

Die dargestellten Beispiele aus Tabelle 24 zeigen, dass auch wenn die Primärantwort nicht exakt vorhergesagt wurde, die errechneten Ergebnisse zumindest subjektiv als sinnhaft bezeichnet werden können. So ist die beobachtete Primärantwort zu ‘Musik‘ zwar ‘Ton‘ bzw. ‘Töne‘, allerdings sind die berechneten Antworten von CIMAWA mit ‘Text‘, ‘Konzert‘, ‘Tanz‘ und ‘Komponist‘ ebenfalls als sinnhaft zu bewerten.

5.4 Schlußbetrachtung zu den Fallstudien der menschlichen Wortassoziation

Das vorliegende Kapitel beschreibt die während meiner am Institut für Wissensbasierte Systeme und Wissensmanagement durchgeführten Fallstudien zur menschlichen Wortassoziation. Die auf drei Fallstudien aufgeteilten acht Testreihen unterschieden sich jeweils in einem zentralen Parameter. Um die Ergebnisse der implementierten Assoziationsberechnungsverfahren in jeder Testreihe vergleichbar zu machen, wurde sichergestellt, dass stets identische Parameterwerte an die Verfahren übergeben wurden. Die Güte der errechneten Ergebnisse wurde jeweils am Referenz-FAT von Russel und Meseck [52] gemessen. Aufgrund der großen Anzahl der getesteten Verfahren und der Fülle der Inputparameter werden im folgenden Abschnitt die wichtigsten Resultate der einzelnen Fallstudien bezogen auf die jeweiligen Zielsetzungen abschließend zusammengefasst.
Das Ziel der Fallstudie 1 war ein erster konzeptueller Vergleich mittels ausgewählter Verfahren. Dafür wurden je ein symmetrisches, ein asymmetrisches und die hybride Eigenentwicklung CIMAWA gegeneinander getestet. Um diese Untersuchungen unabhängig von der verwendeten Textbasis zu machen, wurden zwei Testreihen durchgeführt, die sich in nichts, außer dem zugrunde liegenden Textkorpus, unterschieden. Die Analyse der Ergebnisse ergab, dass CIMAWA die besten Resultate erzielte. Das getestete asymmetrische Verfahren WR lag deutlich vor dem getesteten symmetrischen Verfahren PMI und ordnete sich so im Mittelfeld der Auswertungen ein.
Fallstudie 2 erweiterte das Spektrum der getesteten Assoziationsberechnungsverfahren auf 18 unterschiedliche Methoden. Hierbei wurden zwei Testreihen mit unterschiedlichen Textfenstergrößen für die Berechnung der Kookkurrenzen durchgeführt. Wie in Fallstudie 1 wurden auch in dieser Testserie im Durchschnitt die besten Ergebnisse von ‘CIMAWA adjusted‘ erzielt. Wie Tabelle 26 zu entnehmen ist, folgen auf dem zweiten Rang fünf Verfahren, die gemittelt über die beiden Testreihen identische Resultate erzielen.
In Fallstudie 3 werden im Vergleich zu Fallstudie 3 die gleichen Verfahren erneut miteinander verglichen. Der Unterschied bildet die verwendete Textgrundlage. Fallstudie 3 verwendet ein öffentliches Korpus und unterteilt diesen in insgesamt vier Testreihen in unterschiedliche Größen. Ziel dieser Fallstudie war es, den Zusammenhang zwischen Güte der Ergebnisse und Größe der zugrunde liegenden Textbasis herauszufinden. Im Mittel wurden die besten Resultate von ‘CIMAWA adjusted‘ und CIMAWA geliefert. An dritter Position des Ranking der Fallstudie 3 platzierte sich das symmetrische Assoziationsberechnungsverfahren ‘average MI‘.
Tabelle 26 zeigt die jeweils besten Verfahren der drei Fallstudien in einer zusammenfassenden Übersicht. Als Maß für die Einteilung ist die gemittelte Anzahl an korrekt berechneten Primärantworten über alle Testreihen gewählt.

Tabelle 26. Zusammenfassung der besten Ergebnisse aller Fallstudien

Fallstudie	Top 3 Assoziationsberechnungsverfahren	Durchschnittliche Anzahl korrekt berechneter Primärantworten
1	1. CIMAWA adjusted	13,5
	2. CIMAWA	13,0
	3. WR adjusted	10,5
2	1. CIMAWA adjusted	10,5
	2. poisson-stirling	9,5
	t-score	9,5
	log-likelihood	9,5
	local MI	9,5
	average MI	9,5
3	1. CIMAWA adjusted	16,5
	2. CIMAWA	16,25
	3. average MI	14,25

Obige Ergebnisübersicht zeigt die Konkurrenzfähigkeit der Eigenentwicklung CIMAWA in den durchgeführten Fallstudien. ‘CIMAWA adjusted‘ belegt in jeder der drei Fallstudien den ersten Platz und beweist so, unabhängig von den von Testreihe zu Testreihe wechselnden Randparametern, die Einsetzbarkeit als Berechnungsmethode für der menschlichen Wortassoziation.

Die in absoluten Werten besten Ergebnisse dieser Fallstudie lagen bei 18 Primärantworten (erzielt von CIMAWA in mehreren Testreihen der Fallstudie 3), was verglichen mit dem theoretisch möglichen Wert von 100 zunächst relativ wenig erscheint. Zu beachten gilt jedoch, dass auch die Antworten der menschlichen Probanden im zugrunde liegenden FAT zum Teil stark variieren. So stellt **[47]** über den Assoziationstest von Russell und Meseck fest, dass die durchschnittliche Versuchsperson lediglich 22,5 Primärantworten produzierte. Legt man diesen Wert der oben gemachten Bewertung zugrunde, erscheinen die Simulationen zur menschlichen Wortassoziation auf Textbasis vielversprechend.

Zusätzlich muss der zeitliche Abstand zwischen dem Assoziationstest und den Simulationen in die abschließende Bewertung aufgenommen werden, denn die Texte, die den Simulationen zugrunde liegen, stammen nicht aus der Zeit, in der die Assoziationstests durchgeführt wurden. Zu vermuten ist daher, dass die Ergebnisse der Verfahren in der Gesamtheit auf einem zeitlich aktuelleren Assoziationstest bessere Ergebnisse liefern werden. Da allerdings kein aktueller Assoziationstest in angemessener Größe bekannt ist, bleibt dies die subjektive Einschätzung des Autors. Ein Beispiel aus Tabelle 24 steht exemplarisch für diese Vermutung: Die Primärantwort für ‘Kohl‘ im Test von Russell und Meseck **[52]** aus den 50er Jahren ist ‘Gemüse‘, die berechnete Antwort der meisten Verfahren lautet jedoch ‘Helmut‘. Diese berechnete Assoziation dürfte bei den Menschen erst in den 80er und 90er Jahren entstanden sein und schied somit in den frühen Assoziationstests als mögliche Antwort aus.

6 CIMAWA Anwendungen

6.1 CIMAWA zur Erkennung von Multi-Themenstrukturen in Textdokumenten

Die in diesem Unterkapitel vorgestellte CIMAWA-Anwendung basiert auf dem vom Autor mitverfassten Artikel ‘A Framework to utilize the Human Ability of Word Association for Detecting Multi Topic Structures in Text Documents‘. Der Artikel wurde am 27.10.2013 zur Veröffentlichung in der Printversion der IEEE *Intelligent Systems* akzeptiert und am 06.11.2013 vorab online veröffentlicht.
Die Menge digital verfügbarer Daten nimmt stetig zu [85]. Ein großer Teil dieser Daten liegt in Form unstrukturierter Texte vor, was zu einem großen Interesse an Text Mining Techniken führt [86], die es Nutzern ermöglichen, Wissen aus dieser großen Menge an Informationen zu gewinnen [87]. Etablierte Verfahren, basierend auf automatischer Schlüsselwort Extraktion [88], können Nutzern helfen, die für sie im jeweiligen Kontext relevanten Texte zu finden. Dabei werden statistische Auswertungen bezüglich Vorkommen und Position der Worte im Text benutzt, um die wichtigsten Worte eines Textes herauszufiltern. In der Literatur sind verschiedene Ansätze zur Extraktion der Schlüsselworte zu finden: Ein graph-basiertes ranking Modell namens ‘TextRank‘ wird von Mihalcea und Tarau in [89] vorgestellt. Dieses Verfahren extrahiert die wichtigsten Worte oder Sätze eines Textes. Dabei wird der Text an sich durch einen Graphen repräsentiert, der Worte und andere textuelle Einheiten miteinander verbindet [89]. Einen anderen Ansatz verfolgen Goyal et. al in [88], indem sie ein Ähnlichkeitsmaß auf Satzbasis definieren und dafür die Kookkurrenzen in einem großen Textkorpus verwenden. Ergebnis ist ein kontextbasiertes Modell zur Zusammenfassung von Dokumenten. Die meisten der entwickelten Verfahren zur Schlüsselwortextraktion basieren auf den sogenannten Term Frequency-Inverse Document Frequency (TF-IDF) Ansätzen [90], [91], [92], [93]. Bei einem solchen handelt es sich laut Klahold

„[…] *um ein einfaches Verfahren zur Bestimmung von Wort-Gewichtungen, welches das häufige Auftreten eines Wortes im konkreten Text „belohnt“ und ein häufiges Vorkommen über alle Texte „bestraft“.“* [94].

Ein Algorithmus auf Basis von TF-IDF und Naïve Bayes Machine Learning zur Extraktion der wichtigsten Begriffe wird von Wu und Salton in [95] vorgestellt. Diese Methoden erzielen gute Ergebnisse bezüglich einer grobgranularen Darstellung der wichtigsten Worte eines Textes.
Die klassische Schlüsselwortextraktion stößt jedoch an Grenzen, wenn es darum geht Texte zu analysieren, die mehrere Themen gleichzeitig behandeln. Die Schlüsselworte werden zwar auch in diesen Fällen in gewohnter Weise extrahiert, jedoch sind in diesem

Schlüsselwortkonglomerat sämtliche Themen gleichermaßen vertreten. Eine präzise Definition der im Text enthaltenen Themen ist nicht möglich. Aus dieser fehlenden Funktionalität traditioneller Schlüsselwortextraktionsverfahren ergab sich die Motivation zur Entwicklung einer eigenen Methode zur Themenbestimmung in Texten. Mit dem Ziel, die vorhandenen Potentiale, die sich aus der bislang fehlenden Funktionalität ergeben, zu nutzen, begann die Konzeption der ‘Associative Gravity‘.
Die Verbindung zwischen Statistiken der Kookkurrenz und Themenstrukturen in Texten wurde empirisch belegt von Deerwester, Furnas und Landauer [96]. Das Grundkonzept des hier präsentierten Ansatzes zur Erkennung von Themenstrukturen in Texten besteht darin, den zuvor entwickelten CIMAWA-Ansatz einzusetzen, um die Assoziationen zwischen den Schlüsselworten zu bestimmen und für das Themenclustering einzusetzen. So ist es das übergeordnete Ziel, aus den Schlüsselworten Cluster zu bilden, die die Themenstruktur des Dokumentes abbilden. Auf der Basis von CIMAWA wurde die ‘Associative Gravity‘ [78] entwickelt, der ihrerseits drei Text Mining-Methoden in einem integrierten Ansatz kombiniert:

(1) Schlüsselwortextraktion für die Bestimmung der wichtigsten Worte im Text
(2) Berechnung der Assoziationen zwischen den Schlüsselworten mit CIMAWA
(3) Clustern der Schlüsselworte zu Themenclustern

Wie die verwendeten Komponenten aufeinander abgestimmt und aufgebaut sind, wird in Kapitel 6.1.3 dargestellt.
Grundidee der Associative Gravity ist es, die Wortassoziation zwischen den Schlüsselworten eines Textes als eine Art Anziehungskraft zu interpretieren und diese zu nutzen, um die Worte in Themenclustern abzulegen. In der frühen Entwicklungsphase der Associative Gravity wurden Parallelen zum Gravitationsgesetz nach Newton gezogen, welches besagt, dass Objekte einander anziehen, basierend auf ihrer Masse und deren Abstand zueinander [97]. Die Anziehungskraft ist direkt proportional zum Produkt der Massen und umgekehrt proportional zum Quadrat des Abstandes der Objekte.
Associative Gravity adaptiert diese Prinzipien auf den vorliegenden Anwendungsfall des Themenclustering von Begriffen. Die sich gegenseitig anziehenden Objekte aus Newtons Gravitationstheorie sind die Schlüsselworte aus den Texten. Das Gegenstück zur physikalischen Masse dieser Objekte ist die Wichtigkeit der Worte im Text. Diese Wichtigkeit wird ausgedrückt durch den sogenannten ‘importance value‘ (iv). Der iv wird im Zuge der Schlüsselwortextraktion ermittelt. Dafür benutzt Associative Gravity ein Derivat der bewährten TF-IDF-Ansätze.
Das Gegenstück zum Abstand der Objekte aus Newtons Theorie ist die durch CIMAWA berechnete Assoziation. Je stärker die Assoziation, desto geringer der Abstand zwischen den Objekten. Die eigentliche Anziehungskraft der Worte untereinander berechnet sich in der Associative Gravity aus einer Kombination aus Wichtigkeit der Worte (iv) und deren Assoziation (CIMAWA). Diese Anziehung wird bezeichnet als ‘Associative Gravitational Force‘ oder kurz AGF. Die Berechnungsvorschrift für die AGF ist Formel 7 zu entnehmen.
An dieser Stelle muss auf die Bedeutung der AGF-Berechnung eingegangen werden, denn auf konzeptueller Ebene muss die Frage beantwortet werden, warum eine zusätzliche Berechnung

des AGF vonnöten ist, anstatt für diesen Parameter den bereits bekannten CIMAWA-Wert selbst zu verwenden. Die Notwendigkeit einer angepassten Anziehungsberechnung in Form der AGF entwickelte sich im Verlauf der zahlreichen Tests zur Optimierung des Verfahrens. Dabei zeigte sich, dass CIMAWA zwar akkurate Ergebnisse im Bereich der Assoziationsberechnung erzielte, diese Ergebnisse jedoch vielmehr auf das Gesamtkorpus als auf den einzelnen Text anwendbar waren. Die CIMAWA-Resultate waren in gewisser Hinsicht global betrachtet aussagekräftig, allerdings für den einzelnen Text zu unspezifisch. Die Überwindung dieses Problems gelang durch die Integration der textspezifischen Bestimmung des iv. Kombiniert ergänzen sich die Assoziationen, errechnet durch CIMAWA, als übergeordnete Variable und der Wichtigkeit der Worte im Einzeltext durch den iv zu einem aussagekräftigen Maß für die Anziehungskraft zwischen den Worten.

6.1.1 Konzeptuelle Visualisierung der Associative Gravity

Der folgende Abschnitt beschreibt die Zusammenhänge zwischen Worten und deren Anziehungskraft auf konzeptueller Ebene. Dabei werden insbesondere die Paralellen zu Newtons Gravitationsgesetz deutlich.
Um die wichtigsten Worte eines Textes zu extrahieren, wurde in den Associative Gravity-Ansatz eine Eigenentwicklung des Instituts für Wissensbasierte Systeme zur Extraktion der Schlüsselworte integriert [98]. Nach der Definition der Schlüsselworte liegt zunächst ein Konglomerat der wichtigsten Worte vor. Diese beschreiben den Inhalt des Textes, machen jedoch noch keine Themen innerhalb des Textes unterscheidbar. Zu diesem Zweck wird die Anziehung zwischen den Worten mittels AGF berechnet und mit einem Clusteralgorithmus entsprechende Themencluster gebildet.
In Abbildung 22 und Abbildung 23 werden die extrahierten Schlüsselworte durch Kugeln (K1, K2,…, K5) visualisiert. Jede Kugel steht dabei für eines der extrahierten Schlüsselworte. Die Größe der Kugeln repräsentiert in der Visualisierung die Wichtigkeit der einzelnen Worte im Text. Dabei gilt: Je größer die Kugel, desto wichtiger das Schlüsselwort.

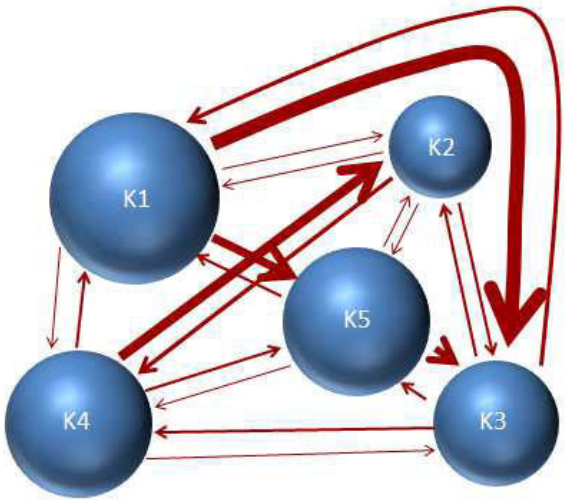


Abbildung 22. Visualisierung der Schlüsselworte und Anziehungskräfte [78]

Verbunden sind alle Worte mit gegenläufigen Pfeilen, welche die Anziehung darstellen. Die Stärke der Pfeile steht für den Grad der Anziehung, je stärker die Pfeile, desto stärker die Anziehung, die die Worte aufeinander ausüben.
Beachtenswert sind jedoch nicht ausschließlich die Stärken der Pfeile bzw. die auftretenden Anziehungskräfte, sondern auch die Pfeilrichtung. Die Verbindung zwischen zwei Worten in Abbildung 22 wird jeweils durch zwei gegenläufige Pfeile visualisiert. Diese Gegenläufigkeit ist auf die Integration des CIMAWA-Ansatzes als hybrides Assoziationsberechnungsverfahren zurückzuführen. Die Stärke der Anziehung zwischen zwei Worten ist nicht auf einen Wert beschränkt, sondern kann richtungsabhängig exakt dargestellt werden. Die Einbeziehung der Assoziationsrichtung in die Berechnung ermöglicht die genaue Abbildung der Assoziation und trägt so dazu bei, das Themenclustering zu präzisieren.
Auf Basis der in Abbildung 22 dargestellten Grundlage kommt ein eigens im Rahmen von Associative Gravity entwickelter Clusteralgorithmus zum Einsatz. Dieser prozessiert die errechneten Ergebnisse und separiert die Schlüsselworte in einzelne Cluster, die im Idealfall jeweils ein Themengebiet des Textes zusammenfassen. Ein beispielhaftes Resultat zeigt Abbildung 23.

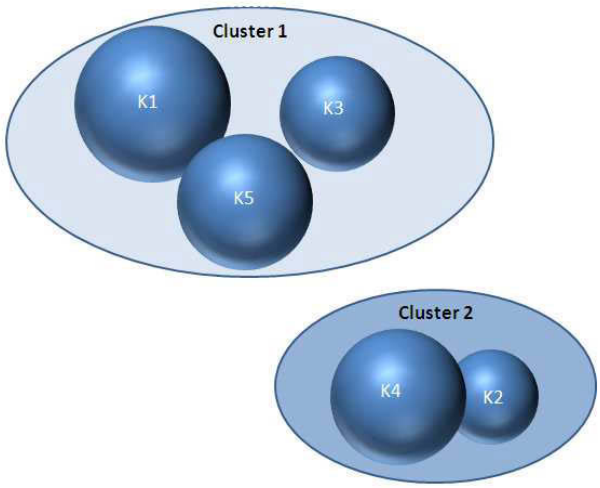


Abbildung 23. Beispielhaft gebildete Themencluster durch Associative Gravity

Das Clustering-Resultat in Abbildung 23 ist wie folgt zu interpretieren: Ein Themenstrang des analysierten Textes wird durch die Schlüsselworte K1, K3 und K5 beschrieben, ein zweiter Schwerpunkt des Textes wurde durch die Worte K4 und K2 identifiziert.
Der mathematische Hintergrund der verwendeten Formel für AGF und der entwickelte Clusteralgorithmus werden im Folgenden detailiert beschrieben.

6.1.2 Associative Gravity im Detail

Die Bestimmung von Themenclustern durch Associative Gravity ist in vier aufeinanderfolgende Schritte unterteilt. Diese Prozessschritte werden im Folgenden

dargestellt. Die selbst entwickelte Architektur der Associative Gravity ist in Abbildung 25 visualisiert.

- Prozessschritt 1: Schlüsselwortextraktion

Im ersten Vorverarbeitungsschritt der Associative Gravity werden die Schlüsselworte des Textes extrahiert. Dafür wird eine Eigenentwicklung, basierend auf dem TF-IDF-Ansatz, verwendet [98], wobei zusätzlich zu dem Auftreten der Worte in Korpus und Analysetext die Position der Worte im Text Berücksichtigung findet. Frühe Untersuchungen zu diesem Thema weisen auf Zusammenhänge zwischen Position eines Wortes im Text und dessen Bedeutung hin. In einer experimentellen Studie von Baxendale wurden in [99] menschliche Probanden gebeten, denjenigen Satz aus einem ihnen vorgelegten Text zu identifizieren, der ihnen am repräsentativsten erscheint. Im Ergebnis zeigte sich, dass in 85% der Fälle der erste Satz des Textes als solcher identifiziert wurde und in 7% der Letzte. Demnach sind Worte bzw. Sätze, die den Anfang (Titel, Einleitung) und das Ende (Zusammenfassung) eines Textes bilden, als tendenziell wichtiger einzustufen als solche im Hauptteil. Klahold erarbeitet in [94] die folgende dreistufige Gewichtung, welche auch in Associative Gravity inkludiert ist:

(1) Position innerhalb der ersten 20% des Textes: Gewichtung 2,25
(2) Position innerhalb der letzten 10% des Textes: Gewichtung 1,5
(3) Position sonst: Gewichtung 1,0

Demzufolge werden nicht nur das reine Vorkommen der Worte im Text berücksichtigt, sondern auch deren Positionen im Text einbezogen.

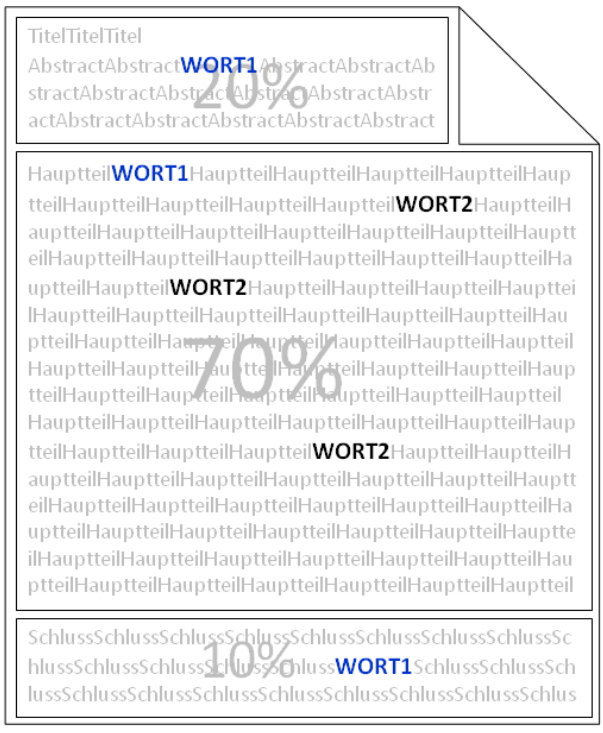


Abbildung 24. Wortpositionen im Textverlauf

Die beschriebene Einteilung wirkt sich in der Praxis wie folgt aus: Jedes Wort innerhalb der ersten 20% eines Textes geht nicht mit dessen gezählter Anzahl in die TF-IDF Berechnung ein, sondern mit dem 2,25-fachen. Kommt ein Wort im Schlussteil, also innerhalb der letzten 10% des Textes vor, so wird dieses mit 1,5 gewichtet. Im Hauptteil des Textes erfolgt die Gewichtung mit der tatsächlich beobachteten Häufigkeit. Abbildung 24 zeigt beispielhaft ein Textdokument, eingeteilt in der beschriebenen Weise und zwei Beispielworten anhand derer eine Berechnung durchgeführt wird.
Beispielsweise wird ein ‘WORT1‘, das je einmal pro Stufe vorkommt, nicht mit seiner gezählten Frequenz ‘3‘ in die Berechnung aufgenommen, sondern mit ‘4,75‘. Dieser Wert kommt wie folgt zustande: Das einmalige Vorkommen in den ersten 20% des Textes wird mit 2,25 gewichtet (1 * 2,25). Das einmalige Vorkommen in den letzten 10% des Textes wird mit 1,5 facher Gewichtung in die Berechnung aufgenommen (1 * 1,5). Addiert man das einmalige Vorkommen im Hauptteil dazu, so ergibt sich folgende Rechnung für ‘WORT1‘: (1 * 2,25) + (1 * 1,5) + (1 * 1) = 4,75. Demnach wird dieses Wort höher bewertet als ‘WORT2‘, das zwar ebenfalls insgesamt 3 mal vorkommt, jedoch ausschließlich im Mittelteil, woraus sich die folgende Bewertung ergibt: (1 * 1) + (1 * 1) + (1 * 1) = 3.
Aufgebrochen wird der beschriebene Standard bei sehr kurzen und sehr langen Texten. In diesen Fällen schlägt Klahold in [94] folgende Differenzierung vor: In Texten die kleiner sind als 100 Zeichen bekommen alle Worte die höchste Gewichtung und bei Texten, die länger sind als 10.000 Zeichen, gelten die ersten 2.000 Zeichen als Anfang und die letzten 1.000 Zeichen als Ende des Textes.
Das Ergebnis dieses ersten Schrittes ist eine Liste mit den wichtigsten Worten des zu analysierenden Textes, absteigend sortiert nach Wichtigkeit (iv).

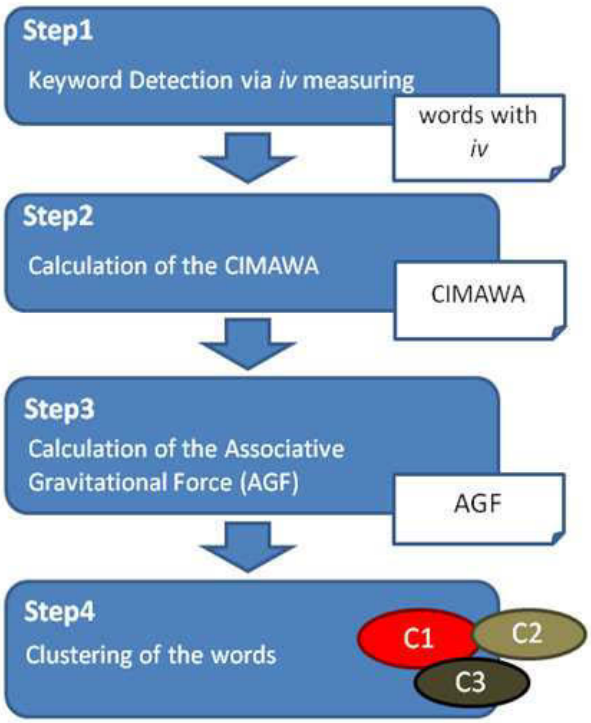


Abbildung 25. Associative Gravity Architektur

- Prozessschritt 2: CIMAWA-Berechnung

Im zweiten Schritt werden mit Hilfe von CIMAWA die Assoziationen zwischen den im ersten Schritt extrahierten Schlüsselworten berechnet. Dafür müssen zunächst die erforderlichen statistischen Daten aus dem Korpus gewonnen werden. Benötigt werden Daten über gemeinsames Vorkommen (Kookkurrenz) der Schlüsselworte und deren Häufigkeit (Frequenz) in der Textsammlung. Im Anschluss kann die CIMAWA-Assoziationsberechnung gemäß Formel 4 durchgeführt werden.
Das Ergebnis dieses Schrittes sind die CIMAWA-Werte zwischen sämtlichen Schlüsselwortpaaren.

- Prozessschritt 3: Berechnung der Associative Gravity Force (AGF)

Im dritten Schritt der Associative Gravity werden die in den ersten beiden Stufen berechneten Werte herangezogen und für die Berechnung der Anziehungskraft zwischen den Worten verwendet. Die selbst entwickelte Berechnungsvorschrift für die AGF ist in Formel 7 definiert.

Formel 7. Berechnung der Associative Gravity Force [78]

$$AGF\big(wort_1(wort_2)\big) = \left(\frac{CIMAWA_{ws}^{\zeta}\big(wort_1(wort_2)\big) * iv(wort_1)}{iv(wort_2)}\right)$$

Da für die Berechnung der hybride Ansatz CIMAWA benutzt wird, sind auch bei der AGF die unterschiedlichen Assoziationsrichtungen zu unterscheiden. Wir differenzieren für jedes Wortpaar zwischen AGF($wort_1(wort_2)$) und der Gegenrichtung AGF($wort_2(wort_1)$). Beispiele für AGF-Werte aus den durchgeführten Tests sind Tabelle 27 zu entnehmen.
Der iv-Wert als Maß für den Informationsgehalt beeinflusst die Gravitationskraft dabei wie folgt: Ein höherer Wert für $wort_1$ (verglichen mit $wort_2$) erhöht die Anziehung von $wort_1$ auf $wort_2$. Im umgekehrten Fall wird die Anziehungskraft entsprechend vermindert. Damit orientiert sich die Berechnung der Anziehung zwischen den beiden Worten wiederum am Newtonschen Gravitationsgesetz, laut dem größere Objekte eine stärkere Anziehung auf kleinere ausüben als umgekehrt.

- Prozessschritt 4: Clustern der Wörter

Auf Basis der berechneten Anziehungskräfte zwischen den extrahierten Schlüsselworten wurde ein Algorithmus entwickelt, der diese Kennzahlen benutzt um Themencluster zu bilden. Ziel ist die Separierung der Schlüsselworte, so dass pro Cluster genau ein Thema des Textes abgebildet wird.
Für jedes extrahierte Schlüsselwort wird zunächst das Schlüsselwort gesucht, auf das die stärkste Anziehung ausgeübt wird. Die gefundenen Wortpaare werden mitsamt des AGF-Wertes in der sogenannten AGF-Tabelle zwischengespeichert. Ein Beispiel für eine solche

Aufstellung ist in Tabelle 27 zu finden. Diese enthält für jedes extrahierte Schlüsselwort (Spalte 1) das Schlüsselwort mit dem höchsten AGF-Wert.

Tabelle 27. Beispiel einer AGF-Tabelle

Schlüsselwort	Schlüsselwort mit höchstem AGF-Wert	AGF-Wert
roy	makaay	0,154
gogh	amsterdam	0,085
feyenoord	rotterdam	0,778
amsterdam	gogh	0,024
zabavnik	club	0,2
rückrunde	roy	0,009
club	saisonende	0,006
saisonende	nürnberg	0,014
nürnberg	saisonende	0,008
fortuyn	gogh	0,286
makaay	roy	0,667
wechsel	nürnberg	0,008
aziz	amsterdam	0,083
rotterdam	feyenoord	0,28
spektakulären	nürnberg	0,012

Auf der Basis dieser AGF-Tabellen wurde der folgende Algorithmus zur Bildung von Themenclustern implementiert:

I. *Bilde neues Cluster:*
Bilde neues Cluster mit dem Wortpaar aus der AGF-Tabelle mit dem größten AGF-Wert; lösche diese Zeile aus der AGF-Tabelle;

II. *Füge den Clustern Worte zu:*

a) Speichere die in (*I.*) geclusterten Worte in einer temporären Liste
b) Durchsuche die AGF-Tabelle nach dem ersten Wort in der temporären Liste
c) Wird das Wort nicht gefunden: Lösche es aus der temporären Liste
sonst:
Füge den (rechten oder linken) Nachbarn aus der AGF-Tabellen Zeile der temporären Liste und dem Cluster hinzu; lösche die Zeile in der AGF-Tabelle und entferne das Suchwort aus der temporären Liste;

So lange die temporäre Liste nicht leer: GOTO (b).

III. *Wenn AGF-Tabelle nicht leer:* GOTO (I.)

IV. *Optimierung der Cluster:*

a) Für jedes Wort in jedem Cluster: Berechne die innere Cluster Anziehung als Summe der AGF-Werte zu den Worten im Cluster;
b) Für jedes Wort in jedem Cluster: Berechne die äußere Cluster Anziehung als Summe der AGF-Werte zu den Worten in anderen Clustern;

c) Vergleiche die innere und äußere Cluster-Anziehung für jedes Wort in jedem Cluster; Übersteigt die äußere Cluster Anziehung die innere, merke dieses Wort vor für Clusterwechsel;
d) Falls Worte für Clusterwechsel vorgemerkt: Alle Wechsel vollziehen und GOTO (a)
sonst
STOP Clustering;

Um Associative Gravity mitsamt dem entwickelten Clusterverfahren zu testen, wurden zweierlei Fallstudien durchgeführt. Zum einen wurde die Associative Gravity als Einzelverfahren getestet und herausgearbeitet, mit welcher Präzision der Ansatz beim Themenclustering arbeitet (siehe Kapitel 6.1.4). Zum anderen wurden verschiedene State of the Art-Clusteralgorithmen implementiert und deren Ergebnisse mit den Associative Gravity-Resultaten verglichen (siehe Kapitel 6.1.7). Für beide Fallstudien wurde zunächst ein Testsaufbau entwickelt, welcher im Folgenden erläutert wird.

6.1.3 Testaufbau zu den Fallstudien der Associative Gravity

Nach der Entwicklung der Associative Gravity muss ein Testaufbau gefunden werden, der es erlaubt, die erzielten Ergebnisse objektiv zu evaluieren. Als wichtige Voraussetzung werden Textdokumente benötigt, die nachweislich mehrere Themen enthalten, um diese zu separieren. Zusätzlich muss es im Rahmen einer Evaluation möglich sein, die Schlüsselworte den einzelnen Themen exakt zuzuordnen. Da dem Autor kein entsprechendes Testszenario bekannt ist, wurde ein eigener Testaufbau entwickelt.
Um sicherzustellen, dass die zu analysierenden Textdokumente mehrere Themen beinhalten, wurden zwei unabhängige Texte in einem Textkonglomerat zusammengefasst und als neues Textdokument abgelegt. Dieser Prozessschritt ist visualisiert im oberen Teil von Abbildung 26. In dieser wird ‘$text_a$‘ kombiniert mit ‘$text_b$‘ zu Compound ‘$text_c$‘. Davon ausgehend, dass jeder der Ursprungstexte mindestens ein Thema behandelt, enthält das neu entstandene Dokument mindestens zwei Themen. Die zusammengesetzten Texte werden jetzt gemäß der beschriebenen Schritte der Associative Gravity prozessiert: Schlüsselwortextraktion, CIMAWA-Assoziationsberechnung und AGF Berechnung. Anschließend werden die Schlüsselworte sowohl mit dem eigens entwickelten Associative Gravity Clusterverfahren (siehe Kapitel 6.1.2), als auch mit anderen State of the Art Algorithmen (siehe Kapitel 6.1.7) in Themencluster eingeteilt. An dieser Stelle zeigt sich die Anwendbarkeit des entwickelten Testaufbaus, denn aufgrund der Textkonglomerate ist es möglich, die gebildeten Cluster auf Basis des Ursprungs der Schlüsselworte zu evaluieren: Enthält ein Cluster ausschließlich Worte aus einem der Ursprungstexte, kann daraus gefolgert werden, dass dieses Cluster keine falschen Zuordnungen enthält. Sind jedoch in einem Cluster Worte aus beiden Ursprungstexten beinhaltet, kann dies auf eine falsche Zuordnung hindeuten.

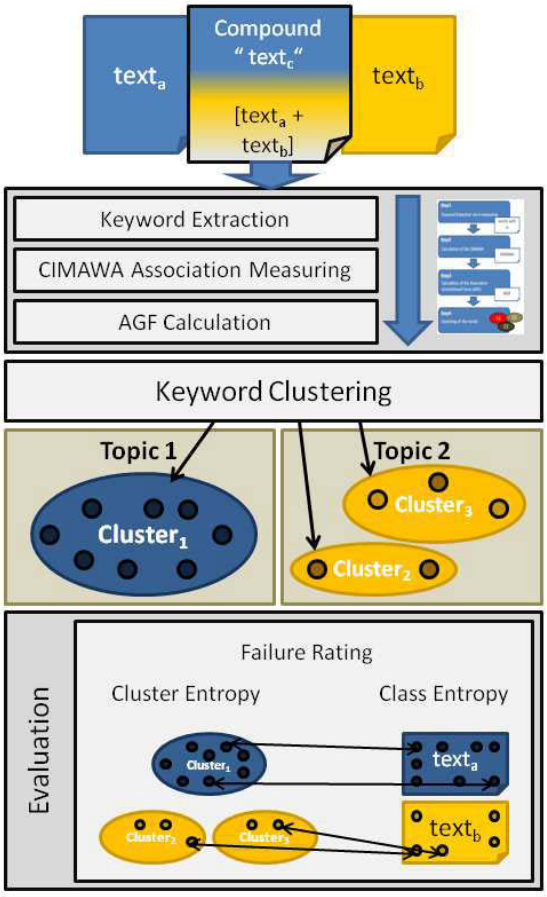


Abbildung 26. Testaufbau zur Associative Gravity (in Anlehnung an [78])

Für die Evaluation sind im Grundsatz zwei Ansätze zu unterscheiden. Zum einem, das sogenannte ‘Failure Raiting‘, was der Berechnung einer Fehlerquote entspricht, und zum anderen die Hinzunahme von qualitativen Messverfahren zur Evaluation von Clusteralgorithmen (siehe Kapitel 6.1.6). Die Fehlerquote wird im Folgenden verwendet um die Präzision der Associative Gravity als Einzelverfahren zu beschreiben. Cluster Entropy und Class Entropy hingegen, die in Kapitel 6.1.7 ausführlich beschrieben werden, sind aus der Literatur bekannte Methoden zur Clusterevaluation und machen einen Vergleich der Resultate mehrerer Verfahren möglich.

6.1.4 Fallstudie 1: Themenclustering mit Associative Gravity

Gemäß des zuvor geschilderten Testaufbaus wird im Folgenden die erste Themenclustering Fallstudie durchgeführt. Die Eigenentwicklung Associative Gravity wird auf insgesamt 20.000 zusammengesetzten Texten ausgeführt und die Ergebnisse im Detail diskutiert. Alle Ursprungstexte der Textkonglomerate wurden der Nachrichtenwebseite ‘Welt Online‘ entnommen und sind frei verfügbar. Die durchschnittliche Länge der zusammengesetzten Texte beträgt 1150,74 Worte und es wurden im Schnitt 14,982 Schlüsselworte pro zusammengesetztem Text extrahiert.

Die Evaluation der Ergebnisse dieser Fallstudie erfolgt mittels des sogenannten Failure Rating, bei dem für jedes berechnete Cluster eine Fehlerrate bestimmt wird. Zum besseren Verständnis der Fehlerrate dient das in Abbildung 27 dargestellte Beispiel.

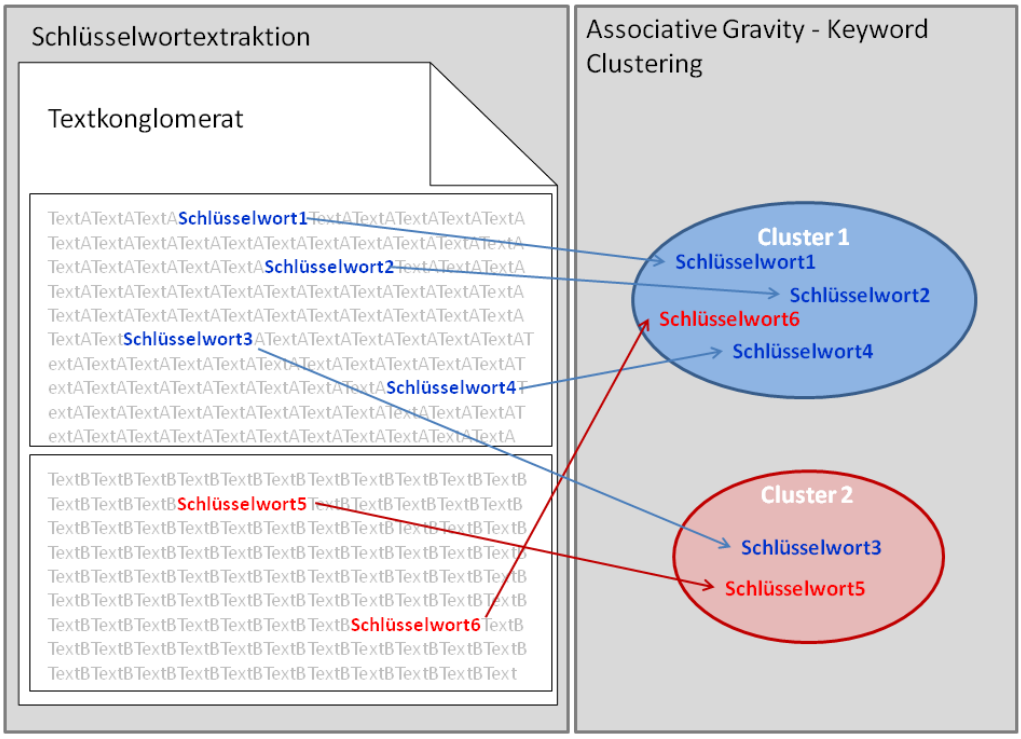


Abbildung 27. Beispiel Failure Rating

Auf der linken Seite von Abbildung 27 ist das aus zwei Texten zusammengesetzte Textkonglomerat zu erkennen. Dieses entspricht ‘$text_c$‘ aus Abbildung 26. Zunächst erfolgt die Schlüsselwortextraktion auf dem Gesamttext. In unserem Beispiel wurden 6 Schlüsselworte extrahiert und exemplarisch hervorgehoben. Schlüsselworte 1 – 4 entstammen ‘$text_a$‘ und Schlüsselworte 5 und 6 sind ‘$text_b$‘ entnommen. Auf der rechten Seite der Abbildung ist ein beispielhaftes Ergebnis des Associative Gravity Keyword Clustering wiedergegeben. An den beiden gebildeten Clustern wird nun veranschaulicht, wie die Fehlerrate zu berechnen ist. ‘Cluster 1‘ in Abbildung 27 enthält insgesamt vier Schlüsselwörter, von denen drei aus ‘$text_a$‘ und eines aus ‘$text_b$‘ stammt. Dies entspricht einer Fehlerrate von 25%, da eines von vier Schlüsselwörtern nicht korrekt geclustert wurde. In ‘Cluster 2‘ befinden sich jeweils ein Schlüsselwort aus jedem Ursprungstext, was eine Fehlerrate von 50% bedeutet. Die maximale Fehlerrate eines Clusters liegt bei dieser Art der Berechnung bei 50%, was einer Gleichverteilung der Schlüsselworte in einem Cluster gleichkommt.

Abbildung 28 zeigt, welche Anzahl von Texten in wie viele Cluster zerlegt wurde. Auf der Hochachse sind die Anzahl der Texte aufgeführt und die Längsachse zeigt die Clustereinteilung. Aus 695 der analysierten Texte wurde jeweils ein einziges Themencluster gebildet. 4.359 Texte wurden in zwei, 8.257 in drei, 5.109 in vier, 155 in sechs und ein Text wurde in sieben Cluster eingeteilt. Die meisten Texte dieser Fallstudie wurden demzufolge in drei Themencluster eingeteilt. Associative Gravity bildete insgesamt 62.677 Themencluster,

was bei 20.000 analysierten Textkonglomeraten zu einem Durchschnittswert von ca. 3,134 Themenclustern pro Text führte.

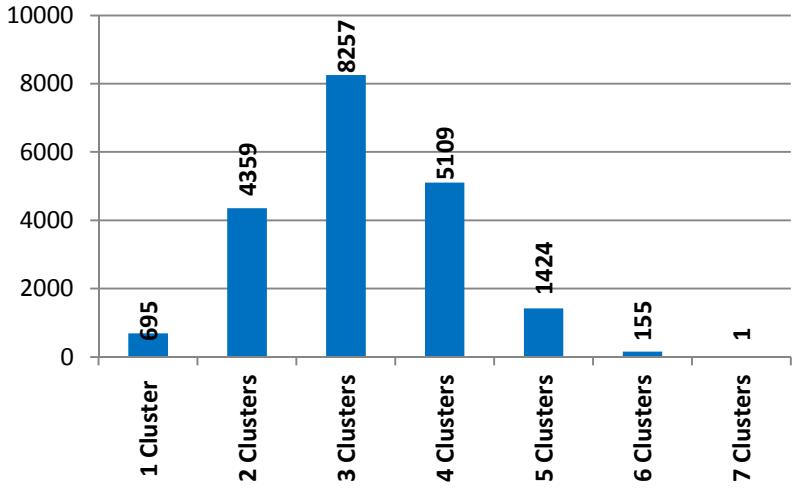


Abbildung 28. Ergebnisse Fallstudie 1: Anzahl der Texte unterteilt in Anzahl der Cluster

Diese Beobachtungen bezüglich der Anzahl der Cluster macht noch keine Aussagen über die Qualität der Clusterergebnisse möglich. Um dies bewerten zu können, sind in Abbildung 29 die berechneten durchschnittlichen Fehlerraten, aufgeteilt nach Clustern, aufgeführt.
Kombiniert man die dargestellten Ergebnisse aus Abbildung 28 und Abbildung 29, kann geschlussfolgert werden, dass die 695 Texte, deren Schlüsselworte in einem einzigen Themencluster zusammengeführt wurden, mit einer durchschnittlichen Fehlerrate von 6,39% die schwächsten Resultate erzielt. Die 4.359 Texte die in 2 Themencluster unterteilt wurden, verbessern diese Quote bereits auf 3,29%. Die entsprechenden Ergebnisse bei steigender Clusterzahl verbessern sich stetig und enden bei 7 Clustern pro Text und einer Fehlerrate von null.

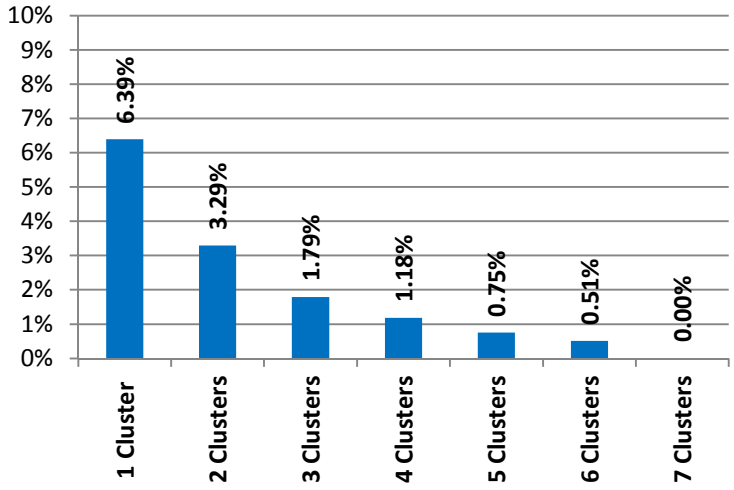


Abbildung 29. Ergebnisse Fallstudie 1: Fehlerquote Clustering

Offensichtlich ist der Zusammenhang zwischen zunehmender Anzahl der Cluster und sinkender Fehlerrate. Die schwächsten Ergebnisse dieser Fallstudie wurden bei den Texten erzielt, deren Schlüsselworte in ein einziges Cluster eingeteilt wurden. Dieses Ergebnis ist aufgrund des Testszenarios zu erwarten, da die Schlüsselworte im zusammengesetzten Text aus zwei Einzeltexten bestehen und die Einteilung der Schlüsselworte in ein gemeinsames Cluster daher tendenziell fehlerbehaftet ist. Eine weitere Erklärung für diesen Verlauf kann nach Betrachtung des anderen Extremfalles gegeben werden. Angenommen es wird nicht ein Cluster mit allen Schlüsselworten gebildet, sondern jedes Cluster besteht aus einem Schlüsselwort, so gleicht die Fehlerrate eines jeden Clusters null. Die alleinige Ausrichtung der Associative Gravity auf eine möglichst niedrige Fehlerrate, birgt die Gefahr der Bildung sehr vieler und gleichzeitig sehr kleiner Cluster. Der Schnitt von etwas über drei Themenclustern pro Text zeigt jedoch, dass bei fast 15 durchschnittlich extrahierten Schlüsselworten, keine Optimierung des Verfahrens auf Kosten der Clustergröße durchgeführt wurde.
Ein zu diskutierender Aspekt ist die relativ niedrige Fehlerrate für die Fälle, in denen nur ein Cluster gebildet wurde, denn es wirft die Frage auf, wie Texte mit tendenziell mehr als einem Thema mit einer solch geringen Fehlerquote geclustert werden können. Die Antwort darauf ergibt sich bei einem erneuten Blick auf den Testaufbau. Denn aus dem zusammengesetzten Text werden die Schlüsselworte extrahiert, ohne jedwede Restriktion, dass diese Schlüsselworte aus beiden Ursprungstexten oder gar paritätisch aus diesen entstammen müssen. Daraus resultiert wiederum die Möglichkeit, dass alle extrahierten Schlüsselworte aus nur einem der beiden Ursprungstexte stammen und folglich das Einteilen dieser Wörter in nur ein einziges Cluster vollkommen korrekt ist.

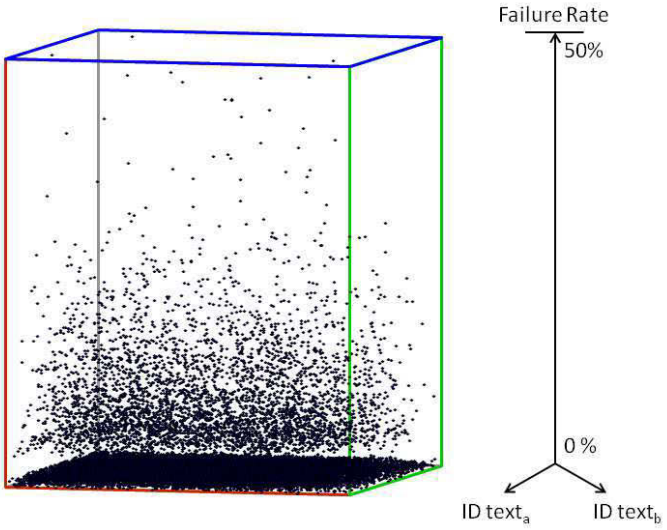


Abbildung 30. Ergebnisse Fallstudie 1: Gesamtübersicht Fehlerrate

Eine Visualisierung der Gesamtergebnisse dieser Fallstudie wird in Abbildung 30 vorgenommen. Jeder Datenpunkt in dem 3-dimensionalen Diagramm steht für einen der 20.000 analysierten Texte. Auf der vertikalen Achse ist die Fehlerrate abgetragen. Die Skala

erstreckt sich von 0, was einem vollkommen korrekten Clusterergebnis gleichkommt und 50%, dem schlechtesten Ergebnis mit Gleichverteilung der Schlüsselworte in einem Cluster. Die horizontale Position der Datenpunkte ist definiert durch eine intern vergebene ID der Ursprungstexte. Die vertikale Position der Datenpunkte hat daher für Auswertungszwecke keine Bedeutung.
Zusammenfassend zeigt diese durchgeführte Fallstudie sehr gute Ergebnisse bezüglich des Themenclusterings des Associative Gravity-Ansatzes. Wie Abbildung 30 zeigt, werden 15.570 der 20.000 Texte mit einer Fehlerrate von 0 geclustert und als Datenpunkte auf der untersten Ebene des Würfels dargestellt. Die höchte Fehlerrate von 50%, welche einer Gleichverteilung der Schlüsselworte in einem Cluster gleichkommt, war in lediglich 3 Fällen zu beobachten. Die Gesamtfehlerquote, gemessen über alle Texte und Cluster in dieser Fallstudie, beläuft sich auf 2,04% und erfüllte die in das Verfahren gesetzten Erwartungen vollends.
Basierend auf den erzielten Ergebnissen stellt sich die Frage nach der Vergleichbarkeit mit anderen etablierten Clusteringverfahren. Zu diesem Zweck schließt sich eine Analyse von Clusterverfahren an, welche im Bereich Themenclustering zur Anwendung gebracht werden können.

6.1.5 Analyse geeigneter Clusterverfahren

Generell werden unter dem Begriff ‘Clustering‘ Methoden und Algorithmen verstanden, die Objekte aus einem Datenbestand in Gruppen einteilen, dessen Elemente einander ähnlich sind [101], [102], [103], [104].
Hauptanforderung für die Auswahl geeigneter Verfahren zum Themenclustering in dem beschriebenen Anwendungsfall ist, dass die Verfahren im Vorfeld der Analyse keine Informationen bezüglich der Anzahl der zu bildenden Cluster benötigen. Die in Frage kommenden Algorithmen sollten keinen Einschränkungen bezüglich der Anzahl und dem Inhalt der Cluster unterliegen. Diese Voraussetzung ergibt sich aus dem beschriebenen Anwendungsszenario. Im Idealfall sollte jedes Cluster ein Thema eines Textes abbilden und da die Anzahl der Themen eines Textes im Vorfeld nicht bekannt sind, müssen Algorithmen für den beschriebenen Anwendungsfall ohne Informationen über Anzahl der zu bildenden Cluster auskommen. Aus diesem Grund können bereits einige Verfahren von vornherein ausgeschlossen werden. Beim sogenannten ‘graph partitioning‘, ‘partitional clustering‘ [105], [106] und ‘spectral clustering‘ [107], [108] wird die Anzahl der Cluster als bekannt vorausgesetzt und kommen aus diesem Grund nicht in Betracht.
Hierarchische Clusterverfahren hingegen benötigen im Voraus weder Informationen über Anzahl noch über die Größe der Cluster [101], [102], [103], [104] und erfüllen die Mindestanforderung. Fortunato beschreibt diesen Sachverhalt wie folgt:

“Hierarchical Clustering has the advantage that it does not require a preliminary knowledge on the number and size of the clusters.” [101]

Unter den hierarchischen Clusterverfahren wird unterschieden zwischen ‘agglomerativen‘ und ‘divisiven‘ Ansätzen [101], [102], [103]. Die agglomerativen Verfahren verfolgen einen ‘bottom-up‘ Ansatz und werden von Cimiano, Hotho und Staab folgendermaßen beschrieben:

“Hierarchical Agglomerative Clustering is a similarity-based bottom-up clustering technique in which at the beginning every term forms a cluster of its own. Then the algorithm iterates over the step that merges the two most similar clusters [...].” [109]

Die agglomerativen Verfahren beginnen demzufolge mit ebenso vielen Clustern wie zu clusternden Elementen. Jedes Element bildet initial ein separates Cluster. In der Folge werden diese Cluster zu immer größeren Clustern verbunden. Die divisiven Ansätze hingegen verfolgen einen ‘top-down‘-Ansatz und beginnen den Clustervorgang mit einem Cluster. Dieses beinhaltet sämtliche Elemente und wird sukzessive aufgeteilt. Hierbei spielen die Verbindungen zwischen den Elementen eine entscheidende Rolle, indem solche mit geringer Gewichtung entfernt werden und somit neue Cluster entstehen.
Im Bereich des Text-, Dokumenten-, oder Termclustering gibt es ebenfalls Ansätze, die auf der hierarchischen Clustering-Strategie aufbauen. In diesem Zusammenhang wurden hierarchische Termclustering-Verfahren identifiziert, die das definierte Anforderungsprofil erfüllen. Namentlich sind dies ‘Single-Link‘, ‘Complete-Link‘ und ‘Average-Link‘ [110], [111].
Die folgende Definition des Single-Link-Verfahrens stammt von Manning, Raghavan und Schütze:

“In single-link clustering [...] *the similarity of two clusters is the similarity of their most similar members.* [...] *Other, more distant parts of the cluster and the clusters’ overall structure are not taken into account.”* [112]

Complete-Link wird von den Autoren wie folgt beschrieben:

“In complete-link clustering or complete-linkage clustering, the similarity of two clusters is the similarity of their most dissimilar members. This is equivalent to choosing the cluster pair whose merge has the smallest diameter.” [112]

Die Ähnlichkeit zweier Cluster wird im Average-Link-Verfahren durch den Durchschnittswert der paarweisen Ähnlichkeiten der Datenpunkte je Cluster ausgedrückt [113]. Je höher die Ähnlichkeit der Elemente zweier Cluster zueinander, desto ähnlicher sind sich die Cluster. Für den in Kapitel 6.1.7 durchgeführten Vergleichstest wurden alle drei Verfahren implementiert und deren Ergebnisse mit der Associative Gravity verglichen.
Bevor diese Vergleichstests im Detail dargestellet werden, muss zunächst eine objektive Methode zur vergleichenden Clusterevaluation gefunden werden. Zu diesem Zweck werden im nächsten Abschnitt verschiedene Maße zur qualitativen Clusterbewertung eingeführt.

6.1.6 Cluster Evaluation

Die intuitive Methode Clusterresultate zu evaluieren, ist die Kontrolle durch den Menschen, indem der Output mit den erwarteten Ergebnissen verglichen wird [114]. Die Probleme mit einer solchen Vorgehensweise zur objektiven Clusterevaluation beschreiben He et al. treffend:

> *"[...] human inspection lacks the scalability to high dimensional, large, and complicated problem domains. In addition, manual inspection is always not desirable and not feasible in real-life applications."* [114]

Aus diesem Grund sollten nach Kashef und Kamel die Ergebnisse von Clusteralgorithmen durch objektive Messverfahren bewertet werden, die die Güte der gebildeten Cluster bewertet [115]. Im beschriebenen Anwendungsszenario wird ein Evaluationsverfahren benötigt, das auf korrekter Klassifikation [116] der Schlüsselworte beruht. Eine Voraussetzung für den von Boley in [117] entwickelten Ansatz ist, dass jedes Objekt ein zuvor definiertes Label für die Klassenzugehörigkeit erhält.

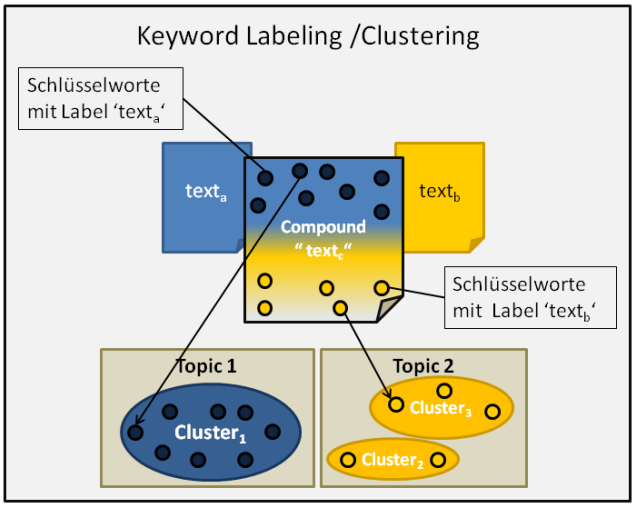


Abbildung 31. Keyword Labeling und Clustering

Im betrachteten Fall bedeutet die Klasse die Zugehörigkeit zum Ursprungstext ($text_a$ oder $text_b$). Folglich sind zwei Arten von Label für die Schlüsselworte (siehe Abbildung 31) zu unterscheiden: Entweder Label '$text_a$' bei Ursprung des Schlüsselwortes in '$text_a$', oder Label '$text_b$' bei Zugehörigkeit zu '$text_b$'. Dieser Ansatz wird als 'Cluster Entropy' bezeichnet und wird für jedes Cluster *j* berechnet:

Formel 8. Cluster Entropy [117]

$$e_j = -\sum_i \left(\frac{c\,(i,j)}{\sum_i c\,(i,j)}\right) * log\left(\frac{c\,(i,j)}{\sum_i c\,(i,j)}\right)$$

Mit *c(i, j)* als Anzahl von Label *i* in Cluster *j*. Die Cluster Entropy e_j beträgt 0, für den Fall, dass alle Objekte in Cluster *j* das gleiche Label *i* aufweisen, anderenfalls ist die Cluster

Entropy positiv [117]. Je niedriger die Cluster Entropy, desto besser das Ergebnis. Alle Beispielcluster in Abbildung 31 besitzen demnach eine Cluster Entropy von 0, da sie jeweils ausschließlich Elemente des gleichen Labels enthalten. Die durchschnittliche Cluster Entropy e_j^{total} berechnet sich als gewichteter Durchschnitt über alle Cluster. Wie diese berechnet wird, ist Formel 9 zu entnehmen.

Formel 9. Durchschnittliche Cluster Entropy [117]

$$e_j^{total} = \frac{1}{m} \sum_j e_j * \left(\sum_i c\ (i,j) \right)$$

Variable *m* steht hier für die Anzahl der geclusterten Objekte. Der berechnete Wert repräsentiert die Qualität der gebildeten Cluster bezüglich Homogenität der Objekte in den Clustern. Je kleiner der berechnete Wert für die durchschnittliche Cluster Entropy, desto homogener die Clusterstruktur. Übertragen auf das Anwendungsszenario steht die Cluster Entropy stellvertretend für die Fähigkeit der Clusterverfahren, die Schlüsselworte korrekt voneinander zu trennen und Cluster zu bilden, die, wenn möglich, ausschließlich einheitlich gelabelte Elemente enthalten. Da sämtliche Beispielcluster eine Cluster Entropy von 0 aufweisen, gleicht auch die durchschnittliche Cluster Entropy dem Bestwert.
Als weiteres Verfahren zur Bewertung von Clusterresultaten kann die Kompaktheit der Lösung gemessen werden:

Formel 10. Class Entropy [114]

$$e_i = -\sum_j \left(\frac{c\ (i,j)}{\sum_j c\ (i,j)} \right) * log \left(\frac{c\ (i,j)}{\sum_j c\ (i,j)} \right)$$

Die sogenannte ‘Class Entropy‘ e_j stammt von He et. al und drückt aus, wie stark Objekte einer Klasse in den Clusterergebnissen fragmentiert sind [114]. Das geschilderte Anwendungsszenario beinhaltet entsprechend der Ursprungstexte zwei Klassen: ‘$text_a$‘ und ‘$text_b$‘.
Die durchschnittliche Class Entropy e_i^{total} ergibt sich als gewichtete Summe der Einzelwerte:

Formel 11. Durchschnittliche Class Entropy [114]

$$e_i^{total} = \frac{1}{m} \sum_i e_i * \left(\sum_j c\ (i,j) \right)$$

Das in Abbildung 31 illustrierte Beispiel ist bezüglich der Class Entropy wie folgt zu bewerten: Aus den Schlüsselworten der Klasse ‘$text_a$‘ wurde ein einziges Cluster ($Cluster_1$) gebildet, daher ist e_i für diese Klasse mit dem bestmöglichen Wert 0 zu bewerten. Dies entspricht der maximal erreichbaren Kompaktheit eines Clusterergebnisses. Diese wurde im Beispiel für ‘$text_b$‘ nicht erreicht, denn die Schlüsselworte aus diesem Text wurden in 2 Cluster ($Cluster_2$, $Cluster_3$) unterteilt. Die Class Entropy für ‘$text_b$‘ ist daher positiv.

Folglich werden durch dieses Maß diejenigen Clusterresultate weniger gut bewertet, die die Schlüsselworte aus einem Ursprungstext in mehr als ein Cluster einteilen. Dieses Qualitätskriterium könnte jedoch nur in dem Fall angelegt und als aussagekräftig bezeichnet werden, wenn mit Gewissheit davon ausgegangen werden kann, dass jeder Ursprungstext exakt ein Thema behandelt. Es ist jedoch denkbar, dass ‘$text_b$‘ in Abbildung 31 mehr als nur ein Thema behandelt, und demzufolge die Einteilung in zwei Cluster korrekt ist. Da keine gesicherten Informationen über die Anzahl der behandelten Themen vorliegen, muss die Class Entropy für die in dieser Arbeit durchgeführten Fallstudien als weniger zielführend bezeichnet werden, als die zuvor eingeführte Cluster Entropy. Entsprechend aussagekräftige Ergebnisse der Class Entropy wären nur dann möglich, wenn die Unterthemen der Ursprungstexte bekannt und exakt im Text zu lokalisieren wären.

6.1.7 Fallstudie 2: Vergleichstest Themenclustering

Wie in Kapitel 6.1.5 beschrieben, wurden für die Vergleichstests die hierarchischen Clusterverfahren Single-Link, Complete-Link und Average-Link ausgewählt. Basiseigenschaften der hierarchischen Clusterverfahren beschreibt Fortunato prägnant:

> *“The starting point of any hierarchical clustering method is the definition of a similarity measure between vertices. After a measure is chosen, one computes the similarity for each pair of vertices* […]. *At the end of this process, one is left with a new n x n matrix X, the similarity matrix.”* [101]

Die beschriebenen ‘vertices‘ sind in dem hier vorliegenden Fall gleichzusetzen mit den extrahierten Schlüsselworten aus den Texten, wobei das ‘similarity measure‘ durch die Berechnung des AGF-Wertes (Formel 7) implementiert wird. Für jeden der 20.000 getesteten Texte dieser Fallstudie wurde eine Matrix erstellt, die sämtliche AGF-Werte zwischen allen, aus dem jeweiligen Text extrahierten, Schlüsselworten enthält. Um die Ergebnisse der verschiedenen Clusteralgorithmen, inklusive der Eigenentwicklung vergleichbar zu machen, operieren alle Verfahren ausschließlich auf diesen Eingabeparametern in Form einer solchen ‘similarity matrix‘.
Zum Zweck der späteren Evaluation der Clusterergebnisse werden alle Schlüsselworte bezüglich ihrer Zugehörigkeit zum Ursprungstext gelabelt (siehe Abbildung 31). Der Umfang dieser Fallstudie gleicht dem der ersten, denn es wurden abermals 20.000 zusammengesetzte Texte analysiert. Diese Texte wurden von jedem der Clusterverfahren (Single-Link, Complete-Link, Average-Link und Associative Gravity) analysiert.
Die Resultate bezüglich der Anzahl der im Durchschnitt gebildeten Cluster sind Abbildung 32 zu entnehmen.

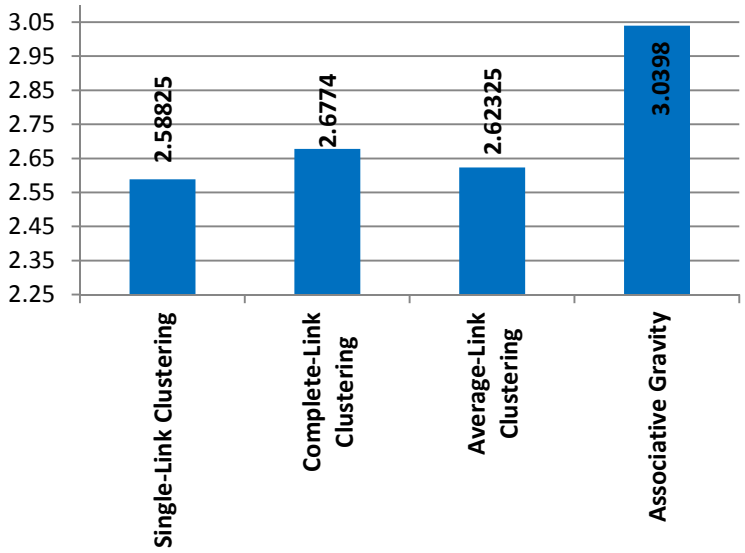


Abbildung 32. Durchschnittliche Anzahl Cluster

Associative Gravity bildet im Durchschnitt 3,0398 Cluster pro zusammengesetztem Text. Im Vergleich dazu bilden die anderen Verfahren weniger, dafür größere Cluster.
Die Cluster Entropy der getesteten Verfahren ist Abbildung 33 zu entnehmen. Die besten Ergebnisse und die niedrigste Cluster Entropy verzeichnet Associative Gravity mit 0,0380488. Die zweitbesten Resultate liefert der Average-Link Algorithmus mit 0,0415596, an dritter Stelle ist der Complete-Link mit 0,04191965 platziert und die schwächsten Ergebnisse lieferte das Single-Link Verfahren mit 0,04248115.

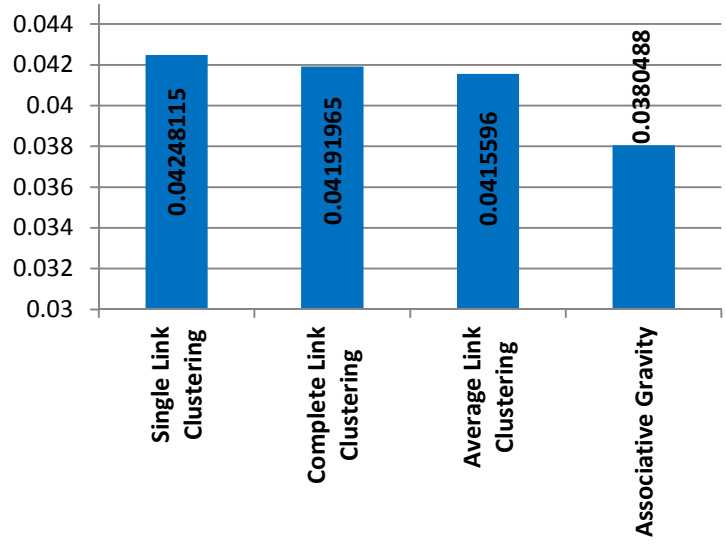


Abbildung 33. Durchschnittliche Cluster Entropy Fallstudie 2 [78]

Aus zuvor erörterten Gründen ist die Cluster Entropy das aussagekräftigere Evaluationsmaß für dieses Anwendungsszenario. Es misst die Durchmischung der gebildeten Cluster in Bezug auf deren Herkunft. Enthält ein Cluster Schlüsselworte aus beiden Ursprungstexten, so kann gefolgert werden, dass dieses Cluster ein bestimmtes Thema nicht scharf von anderen trennt.

Im Gegensatz dazu misst die Class Entropy, in wie viele Cluster die Schlüsselworte eines Ursprungstexts unterteilt wurden und bestraft das Bilden mehrerer Cluster pro Klasse.
Im folgenden, aus der vorliegenden Fallstudie stammenden, Beispiel wird erläutert, aus welchen Gründen die Class Entropy zur Bewertung der Ergebnisse weniger geeignet ist als die Cluster Entropy.
Der zusammengesetzte ‘$text_c$‘ besteht zum einen aus einem politischen Nachrichtentext ($text_a$) und zum anderen aus einem internationalen Sport-Nachrichtentext ($text_b$). Associative Gravity bildet aus 12 Schlüsselworten die folgenden fünf Cluster:

(a) $Cluster_1$: {Terrorist; Anführer; Mehsud}
(b) $Cluster_2$: {van Bommel; Ribéry; Kroos}
(c) $Cluster_3$: {Ballack; Chelsea}
(d) $Cluster_4$: {Inter; Mailand}
(e) $Cluster_5$: {SC Freiburg; 18:30}

$Cluster_1$ beinhaltet ausschließlich Begriffe aus $text_a$, was zu einer Cluster Entropy von 0 führt. Zusätzlich beinhaltet $Cluster_1$ auch sämtliche Schlüsselworte aus $text_a$, woraus eine Class Entropy von 0 resultiert. Die übrigen Cluster beinhalten ebenfalls jeweils ausschließlich Schlüsselworte aus einem Ursprungstext ($text_b$) und besitzen daher ebenfalls eine Cluster Entropy von 0. Eine Detailanalyse dieser Cluster erklärt deren Zusammensetzung: $Cluster_2$ enthält ausschließlich Spieler des FC Bayern München; $Cluster_3$ enthält den Spielernamen ‘Ballack‘ und dessen (damaligen) Verein FC Chelsea; $Cluster_4$ enthält den Namen des italienischen Fussballvereins Inter Mailand; $Cluster_5$ beinhaltet den Namen des Bundesligisten SC Freiburg und den Anstoßzeitpunkt der nächsten Partie (18:30). Sieht man sich die gebildeten Cluster aus der Perspektive der Themenclusterung an, kann gefolgert werden, dass die Clustereinteilung korrekt getroffen wurde und $text_b$ mehrere Unterthemen behandelt. Diese subjektive Einschätzung wird jedoch in keiner Weise von der berechneten Class Entropy gestützt, denn diese liegt bei 0,427, was eines der schlechtesten Resultate in der gesamten Fallstudie ist.

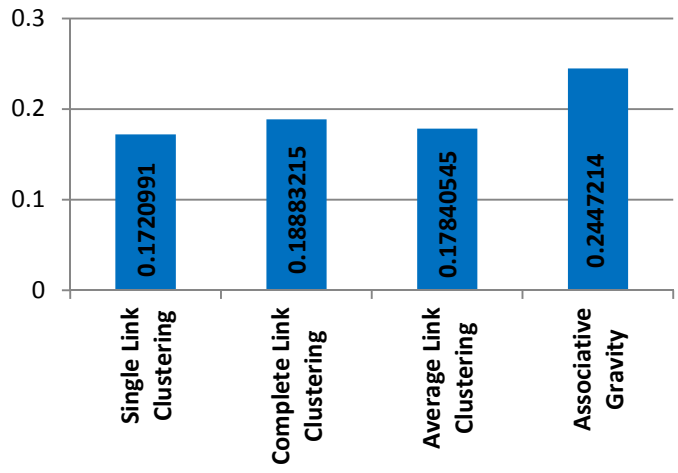


Abbildung 34. Durchschnittliche Class Entropy Fallstudie 2 [78]

Trotz der zuvor beschriebenen eingeschränkten Aussagekraft der Class Entropy für dieses Anwendungsszenario, werden die berechneten Ergebnisse in Abbildung 34 zusammengefasst.
Die durchgeführten Fallstudien zur Erkennung von Themenstrukturen in Texten haben gezeigt, dass die Associative Gravity in der Lage ist, die Themenstruktur eines Textdokumentes mit hoher Präzision abzubilden. Auch ein Vergleich mit bekannten Clusterverfahren konnte die Anwendbarkeit des entwickelten Verfahrens bestätigen.
Es kann postuliert werden, dass es sich bei der Associative Gravity um ein selbst entwickeltes und in seiner Anwendbarkeit geprüftes Verfahren handelt, welches auf Basis der CIMAWA-Assoziationsberechnung innovative Elemente und etablierte Text Mining Komponenten integriert.
Die hier beschriebene Implementierung ‘Associative Gravity‘ zur Erkennung von Multi-Themenstrukturen in Texten, besitzt als eigenständige Lösung Potential für den Praxiseinsatz. Beispielsweise können durch Kenntnis der Themenstruktur eines Textes die Suchergebnisse von Volltextsuchen verbessert werden. Anstatt Texte nach den eingegebenen Begriffen zu durchsuchen und diejenigen Dokumente anzubieten, die diesen Begriff enthalten, sind verfeinerte Suchverfahren denkbar, die die Themenstruktur der Texte einbeziehen. Diese können, anstelle von kompletten Dokumenten, nur die themenspezifisch relevanten Abschnitte herausstellen und dem Suchenden die zeitintensive Recherchearbeit erleichtern.
Im Bereich der Textkategorisierung sind ebenfalls mögliche Einsatzgebiete zu definieren. Bei der Kategorisierung von großen Textmengen, die in Behörden und Unternehmen häufig über Jahre gewachsen sind und so einen Teil der organisationalen Wissensbasis bilden, kann eine Zuordnung von Texten oder Textteilen in Themenkategorien bei der Wiederauffindung und Wiederverwendung von abgelegten digitalen Dokumenten hilfreich sein. Schnittstellen sind hier die bereits etablierten Informations-, oder Dokumentenmanagement-Systeme.
Ein weiteres, derzeit unter Domänenexperten diskutiertes Anwendungsgebiet der Erkennung von Themenstrukturen in Texten, ist in einem anderen Forschungszweig des Text Mining beheimatet. So thematisieren Cambria et al. in [118] den Einsatz von Methoden wie der Associative Gravity im ‘Opinion Mining‘ oder der ‘Sentiment Analysis‘. Eine der Aufgaben im Rahmen der Sentiment Analysis ist die sogenannte ‘polarity classification‘, die von Cambria et al. wie folgt beschrieben wird:

“Polarity classification occurs when a piece of text stating an opinion on a single issue is classified as one of two opposing sentiments.” [118]

Die Zuordnung einer Polarität zu einem Textteil oder Text bildet für sich genommen bereits eine große Herausforderung. Diese wird zusätzlich erschwert, wenn es sich bei dem zu analysierenden Text um einen solchen mit mehr als einem behandelten Thema handelt. In diesen Fällen muss die Themenstruktur erkannt und die identifizierten Meinungen den Themen zugeordnet werden. Cambria et al. erklären dazu:

“If the text […] *covers more than one issue or item, new challenges arise, such as* […] *opinion target identification.”* [118]

Des Weiteren führen die Autoren aus:

”In such instances, it’s important to identify topics and separate the opinions associated with each topic.” [118]

Bei der Identifikation der oben genannten ‘topics‘ kann die in diesem Kapitel erarbeitete Associative Gravity zum Einsatz gebracht werden. Im beschriebenen Fall wird es für die Vorverarbeitung der im Rahmen der Sentiment Analysis durchgeführten Analysen eingesetzt, um die Themen eines Textes vor der eigentlichen Einordnung scharf voneinander zu trennen.

6.2 CIMAWA zur textuellen Metaanalyse im Instandhaltungsmanagement

Die nächste CIMAWA-Anwendung beschäftigt sich mit der automatischen Analyse von Textdokumenten im Instandhaltungsmanagement. Die in diesem Kapitel der Arbeit vorgestellten Ergebnisse basieren auf einem vom Autor mitverfassten Artikel im *International Journal of Services, Economics and Management* mit dem Titel 'Textual Meta-analysis of Maintenance Management's Knowledge Assets' [119].
Es wird eine virtuelle Applikation vorgestellt, die aus den vom Unternehmen erhobenen und in Datenbanken abgelegten Dokumenten der Instandhaltung Assoziationen zwischen relevanten Entitäten berechnet und darstellt. Unter Entitäten oder Instanzen einer Klasse versteht man beispielsweise handelnde Personen, Maschinen oder Aktivitäten. Die drei Klassen `practitioners` als handelnde Personen in der Instandhaltung, `machines`, was den zu wartenden Maschinen entspricht und `maintenance operations`, was die Instandhaltungsaktivitäten zusammenfasst, dienen als Beispielinstanzen in diesem Kapitel. Diese Liste erhebt keinen Anspruch auf Vollständigkeit, sondern kann in einer Anwendungsentwicklung zielorientiert erweitert werden. Eine detailierte Darstellung der Instanzen der derzeit verwendeten Klassen ist Abbildung 37, Abbildung 38 und Abbildung 39 zu entnehmen. Für die Bestimmung der Assoziationen zwischen den Klassen oder zwischen einzelnen Instanzen findet der CIMAWA-Ansatz Anwendung. Die so berechneten Assoziationen werden im Stile einer Assoziationskarte dargestellt. Aus diesem Grund wird der Prozess der Assoziationsberechnung und Visualisierung im Folgenden als 'Association Mapping' bezeichnet.
Um das Verständnis des Lesers für die folgenden Ausführungen zu erleichtern, sollen zunächst einige wichtige Begriffe und Zusammenhänge in Bezug auf das Instandhaltungsmanagement erläutert werden.
Aktivitäten im Rahmen des Instandhaltungsmanagement beinhalten die Identifikation, die Erstellung und die Speicherung sogenannter 'knowledge assets'. Unter knowledge assets verstehen Nonaka et al.:

> *"*[…] *inputs, outputs and moderating factors of the organization's knowledge creating activities* […]." [120]

Dawson identifiziert derweil drei Gruppen von knowledge assets und benennt diese als 'human capitals', 'structural capital' und 'relationship capital' [121]. Im Kontext des Instandhaltungsmanagement kann Wissen während der Instandhaltungsaktivitäten gesammelt und gespeichert werden. Dies kann beispielsweise in Form von manuell verfassten Fehlerberichten oder Reparaturprotokollen geschehen, welche als zuvor definierte knowledge assets in geeigneter Form abgelegt werden. Die so gewonnenen Informationen können zu einem späteren Zeitpunkt bei Reparaturen oder Inspektionen von Nutzen sein, da sie in dokumentierter Form vorliegen und unter Umständen bereits vorverarbeitet in Datenbanken gespeichert sind. Zum nicht dokumentierten Wissen gehört in diesem Zusammenhang das implizite Wissen der handelnden Personen. Dazu zählt Erfahrungswissen, spezielle

Fähigkeiten oder anwendungsbezogene Kompetenzen. Implizites Wissen muss extrahiert, dokumentiert, geteilt und transferiert werden [122], [123]. Die sogenannten ‘Computerized Maintenance Management Information Systems‘ (CMMIS) [124], [125] dienen dabei der systematischen Identifizierung, Aquirierung und Speicherung der knowledge assets in der Instandhaltung. Laut Bagadia beinhalten die CMMIS eine strukturierte Datenbank, in der alle Arten von dokumentiertem Wissen abgelegt werden können [124]. Anders ausgedrückt handelt es sich bei den CMMIS um speziell für das Instandhaltungsmanagement entwickelte Informationssysteme.
Um bislang brach liegende Potentiale in den vorhandenen Datenbeständen zu erkennen und zu nutzen, kann eine Metaanalyse zielführend sein. Wie Abbildung 35 zeigt, unterscheidet man bei der Metaanalyse zwischen drei Arten von Methoden:

(1) statistische,
(2) mathematische und
(3) textuelle.

Den ersten beiden Methoden bzw. Methodengruppen liegen sogenannte hard-facts und numerische Werte zugrunde, die in dieser Arbeit weniger im Fokus stehen als die textbasierten Methoden der textuellen Metaanalyse. Aus diesem Grund werden erstere lediglich in ihren Grundzügen vorgestellt und auf eine tiefere Analyse verzichtet.

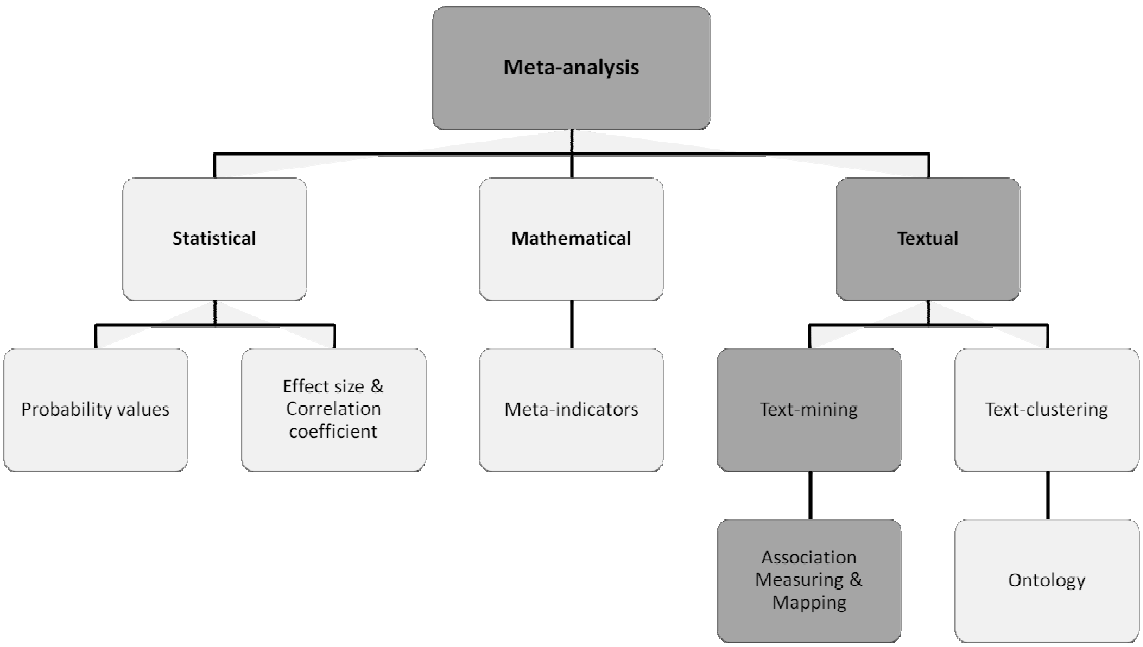


Abbildung 35. Methoden der Metaanalyse [119]

Der Oberbegriff der Metaanalyse geht zurück auf Glass, der in [126] eine erste Definition verfasst und dessen Arbeit die Basis weiterer Forschung und Publikationen [127] darstellt. Eine detaillierte Studie der wichtigsten statistischen Methoden der Metaanalyse liefert Lyons in [128]. Auch die in Abbildung 35 dargestellte Gliederung der statistischen Methoden in ‘Probability values‘ sowie ‘Effect size & Correlation coefficient‘ geht auf diese Publikation

zurück. Die mathematischen Methoden sind dargestellt im mittleren Teil von Abbildung 35 und werden als zweite Gruppe der Metaanalyse zusammengefasst. Im Fokus stehen hier Meta-Indikatoren, die Instanzen von Datensätzen aggregieren und interpretieren. Ein konkretes Beispiel hierfür aus dem Bereich der Instandhaltung ist die Verfügbarkeit einer bestimmten Maschine. Die Feststellung der Verfügbarkeit basiert auf der Zusammenfassung entsprechender Daten, wie der mittleren Zeitspanne bis zum nächsten Ausfall und der mittleren Ausfallzeit in Folge einer Reparatur [129]. Durch die Zusammenfassung dieser Daten ist es möglich, eine erste Prognose zur Verfügbarkeit der entsprechenden Maschine zu erstellen.
Die dritte und für die vorliegende Arbeit bedeutsamste Gruppe, stellen die textuellen Methoden der Metaanalyse dar. Diese lässt sich laut Heyer, Quasthoff und Wittig in Text Mining und Text-clustering unterteilen [16], wobei beide Begriffe Methoden zur Analyse von unstrukturierten textuellen Inhalten beherbergen. Die textuelle Metaanalyse beeinhaltet in erster Linie die Assoziationsberechnung (siehe Kapitel 3 und Kapitel 4) sowie die Ontologien, wobei erstere dem Text Mining und letztere dem Text-Clustering zugeordnet werden kann. Eine präzise Definition des Text Mining-Begriffs ist Kapitel 2 zu entnehmen und in Kapitel 4 wird der Text-Clustering-Begriff genauer erläutert.
Die für die vorliegende Arbeit wichtigsten Bereiche, die die Basis für den hier entwickelten Lösungsansatz darstellen, wurden in Abbildung 35 dunkelgrau hinterlegt. Beim vorgestellten Ansatz des Association Mapping handelt es sich um eine textuelle Metaanalyse mittels Assoziationsberechnung unter Verwendung von CIMAWA.
Die mathematischen Methoden der Metaanalyse bilden ein aktuelles Forschungsgebiet unseres Instituts für Wissensbasierte Systeme und Wissensmanagement. Ansari et al. thematisieren in [130] die mathematische Metaanalyse von Instandhaltungskosten am Beispiel von Maschinen aus der Produktion. Ein Ergebnis dieser Arbeit ist die prototypische Implementierung einer Instandhaltungsmanagement-Software. Eine weitere institutsinterne Entwicklung wurde publiziert in [131]. Diese beschreibt eine Dashboard-Lösung zur Unterstützung des ‘Chief Maintenance Officer‘ (CMO) bei Planung und Kontrolle von Instandhaltungstätigkeiten. Die in [130] und [131] entwickelten Lösungen basieren ausschließlich auf hard facts und numerischen Werten und klammern die in Textform vorliegenden Daten gänzlich aus. Um diese Lücke zu schließen und die teils offen liegenden, teils verborgenen Wissenspotentiale der abgelegten Texte zu nutzen, war eine neue Herangehensweise im Sinne einer Neukonzeption erforderlich.
Zu diesem Zweck wurden zunächst die auf dem Markt befindlichen CMMIS-Systeme auf ihre Möglichkeiten zur automatischen Analyse von Texten untersucht. Diesbezüglich war in erster Linie die Frage nach Text Mining-Lösungen, die über die Grundfunktionalitäten hinaus gehen, zu beantworten. Ein zusammenfassender Überblick der Resultate wird im Folgenden wiedergegeben.

6.2.1 Textuelle Metaanalyse von Textdokumenten in der Instandhaltung

Wie bereits zuvor herausgearbeitet, können Textdokumente in der Instandhaltung wie auch in anderen Bereichen einer Unternehmung kodifiziertes Wissen von Experten enthalten. Solche Texte werden in unterschiedlichen Formaten und an verschiedene Orten im Unternehmen gespeichert und geraten oftmals bereits unmittelbar nach ihrer Erstellung in Vergessenheit. Mitarbeiter haben keinen Einblick in die gewachsene organisationale Wissensbasis und verfügen nicht über die nötigen Werkzeuge um abgelegte Dokumente zu finden und die gespeicherten Wissenspotentiale zu nutzen.
Dieses oder ähnliche Szenarien wurden durch Kontakte zu unseren Industriepartnern wiederholt geschildert und bestätigen die Notwendigkeit der Entwicklung von nutzerorientierten Text Mining-Anwendungen zur automatischen Analyse großer Textmengen. Speziell das Instandhaltungsmanagement mit den angepassten CMMIS erscheint vor diesem Hintergrund als vielversprechendes Anwendungsgebiet von automatischer Analysesoftware.
Nach umfangreicher Studie der gängigen kommerziellen CMMIS muss postuliert werden, dass die meisten dieser Systeme zwar die standard Dokumentenmanagement-Funktionalitäten wie Suche oder Schlüsselwortextraktion besitzen, eine echte textuelle Metaanalyse jedoch nicht Teil des angebotenen Funktionsumfanges ist. Mobley, Higgins und Wikoff stellen diese Funktionslücke in [125] ebenfalls fest und erkennen zugleich die Notwendigkeit solcher Lösungen um Qualität und Kosteneffizienz von Instandhaltungsprozessen und Servicleistungen zu gewährleisten [125]. Ein aus [119] entnommenes Beispiel für die von CMMIS bereitgestellten Funktionalitäten zeigt Abbildung 36. Diese beinhaltet eine erstellte Zusammenfassung eines Textdokumentes aus der CMMIS-Datenbank und enthält neben Autor, Erstellungsdatum, Identifikationsnummer und Dringlichkeit auch die wichtigsten Schlüsselworte des Textes.

Author	Thomas Mueller
Date	20.04.2012
ID No.	ID2X54F
Title	Inspection of drilling machine
Priority	One week
Keywords	Inspection, Repair

Abbildung 36. Beispielhafte Repräsentation eines Instandhaltungsdokuments im CMMIS [114]

Eine darüber hinausgehende textuelle Metaanalyse der gespeicherten Textdokumente hingegen wird von den auf dem Markt befindlichen CMMIS nicht unterstützt. Aus diesem Grund setzt sich der Ansatz des Association Mappings zum Ziel, eine Erweiterung der Text Mining Funktionalitäten im Sinne einer echten textuellen Metaanalyse zu realisieren und die Herstellung einer Verbindung zwischen der mathematischen Metaanalyse auf numerischen Werten und den Ergebnissen der Textanalyse zu gewährleisten. Denn auch wenn die detailierte Analyse von numerischen Daten für den CMO unabdingbar ist, sind die knowledge assets in der Instandhaltung nicht auf diese limitiert. In der Praxis liegen eine Vielzahl verschiedener Typen von Textdokumenten vor, die von Monteuren, Technikern oder Ingenieuren verfasst und abgelegt wurden. Bei diesen Textdokumenten kann es sich beispielsweise um Reparaturberichte, Handlungsbeschreibungen, Wartungsprotokolle oder

Fehlerdiagnosen handeln. Der größte Teil dieser Texte beinhaltet Beschreibungen, verfasst von handelnden Personen, in denen expliziert wird, welche Handlungen durchgeführt wurden und was der initiale Grund für diese war.
Ein leicht nachzuvollziehendes Beispiel für einen solchen Bericht ist eine Handlungsbeschreibung für den Wechsel einer Öldichtung an einer Produktionsmaschine aufgrund einer Leckage. Ein derartiges Dokument enthält explizites Wissen, denn der Autor kodifiziert seine Erfahrungen, indem er beschreibt, welche Handlungen für den Dichtungswechsel vollzogen werden müssen. Im richtigen Kontext bereitgestellt, in diesem Beispiel einem Monteur nach der Feststellung einer Leckage der entsprechenden Maschine, kann ein solcher Bericht einen echten Mehrwert darstellen. Das Auffinden eines solchen Dokumentes, zur richtigen Zeit am richtigen Platz, kann unter Umständen, bei entsprechendem Aufwand, durch die Grundfunktionalitäten Schlüsselwortextraktion und Suche realisiert werden. Eine echte textuelle Metaanalyse der Texte bedeutet dieses Vorgehen noch nicht. Für eine solche werden in dem hier vorgeschlagenen Ansatz ganze Textsammlungen, bestehend aus Textdokumenten der Instandhaltung, automatisiert analysiert. Indem die Assoziationen zwischen Klassen oder deren Instanzen allein auf Basis der zugrunde liegenden Textdokumente berechnet werden, können Beziehungen aufgedeckt und nutzbar gemacht werden, die nach den Auswertungen der numerischen Werte noch verborgen waren.
Ein detailiertes Konzept, welches verdeutlicht, wie der CIMAWA-Ansatz zum Association Mapping in der Instandhaltung funktioniert, wird in Kapitel 6.2.2.1 diskutiert. Zusätzlich wird ein Überblick über die technischen Voraussetzungen der CMMIS gegeben, auf die das Association Mapping aufbaut, sowie die virtuellen Applikation erläutert.

6.2.2 Textuelle Metaanalyse via Association Mapping mit CIMAWA

Bevor das Konzept des Association Mapping und die virtuelle Applikation veranschaulicht werden, ist es unverzichtbar, die technischen Voraussetzungen im Rahmen der CMMIS und die zugrunde liegenden Daten zu untersuchen.
In erster Linie dienen die CMMIS dem CMO und weiteren, in die Instandhaltung involvierte Personen, bei der Ausführung und dem Management der Instandhaltungstätigkeiten [125]. Dafür halten die CMMIS eine strukturierte Datenbank zur Speicherung operationaler, taktischer und strategischer Instandhaltungsdaten vor. Diese beinhaltet beispielsweise Zeitpläne, verschiedene Arten von Berichten, Rollenzuweisungen, statistische Fehleranalysen oder Budgetinformationen [119]. Wichtige Inputdaten für das Association Mapping sind die abgelegten Textdokumente, da auf diesen die Assoziationen berechnet werden können.
Darüber hinaus sind Informationen über die Entitätsstruktur der zu untersuchenden Klassen `practitioners`, `machines` und `maintenance operations` unabkömmlich. Diese Strukturen geben Aufschluss über die Zugehörigkeit der verschiedenen Elemente und weisen im Falle der `practitioners` Rollen zu. Die im Folgenden dargestellten Klassen wurden, basierend auf den in DIN 13306 [132] beschriebenen Hauptkategorien, identifiziert. Jede Klasse besitzt Unterklassen, die in den Abbildung 37, Abbildung 38 und Abbildung 39 exemplarisch und ohne Anspruch auf Vollständigkeit dargestellt sind.

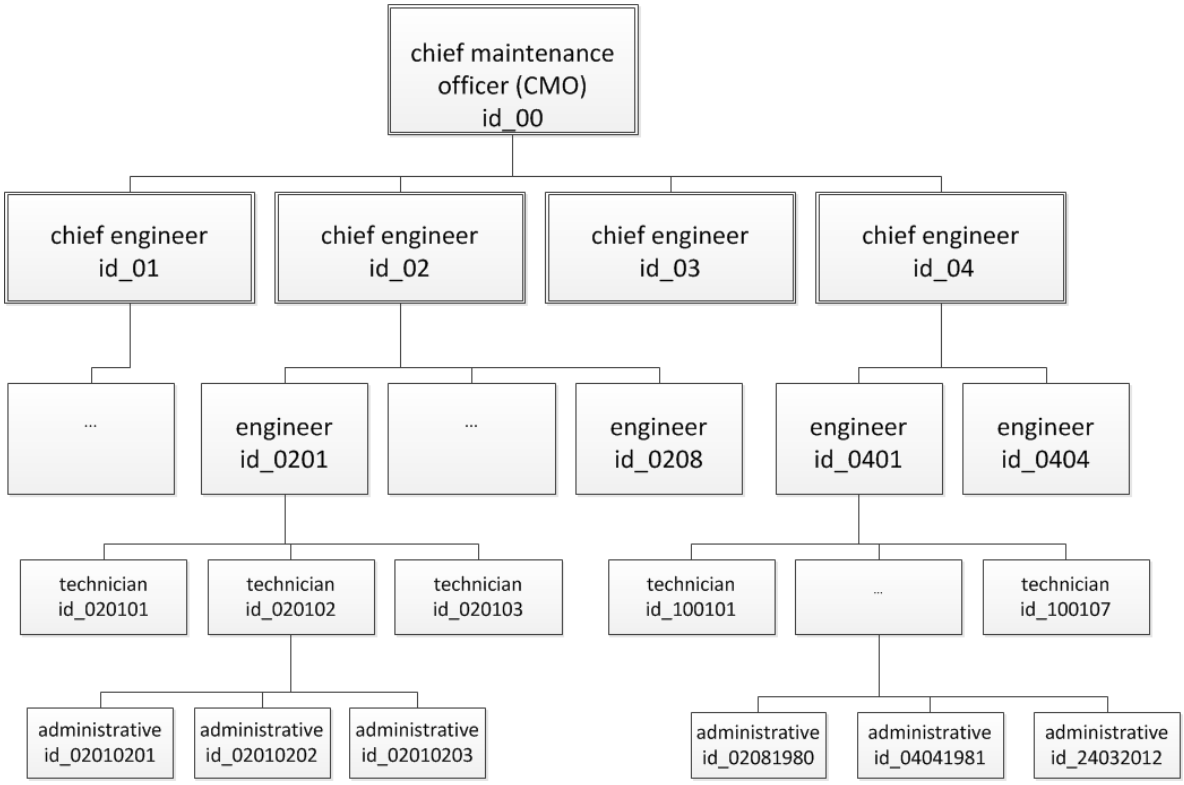


Abbildung 37. Zusammensetzung der Klasse `practitioners` [119]

In Abbildung 37 sind die Unterklassen bzw. Instanzen der Klasse `practitioners` in einer Art Organigramm dargestellt. Zu erkennen ist, dass einem `CMO` im abgebildeten Fall vier `chief engineer` unterstellt sind. Jeder dieser trägt die Verantwortung für mehrere `engineer`, wobei die `technician` eine Hierarchiestufe darunter angesiedelt sind. Das im Beispiel als `administrative` bezeichneten Instandhaltungspersonal ist im untersten Level eingeordnet und untersteht direkt den entsprechenden `technician`. Jeder Mitarbeiter in dieser Struktur wird identifiziert durch eine einmalig vergebene ID. Die Struktur eines Aufbaus, wie ihn Abbildung 37 zeigt, hängt von der Organisation der verantwortlichen Abteilung ab. Der hier dargestellte Aufbau steht exemplarisch für eine solche Organisationsstruktur.

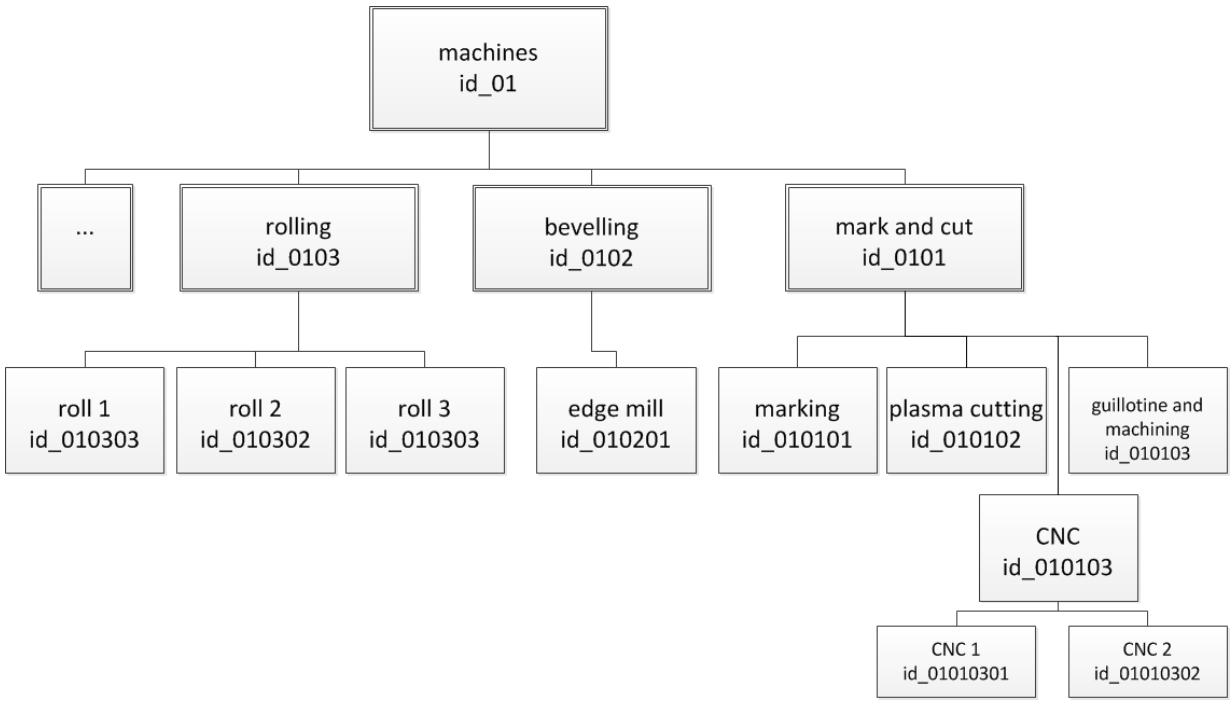


Abbildung 38. Zusammensetzung der Klasse `machines` [119]

Analog zur Klasse der `practitioners` werden in Abbildung 38 die Unterklassen und Instanzen der Klasse `machines` visualisiert. Die Unterklassen sind definiert als `rolling` (Walzen), `bevelling` (Schleifmaschinen) und `mark and cut` (Markierungs- und Schneidewerkzeuge). Mit Ausnahme der `CNC` Unterklasse, die ihrerseits zwei Instanzen beherbergt, sind alle Instanzen der genannten Unterklassen als eigenständige Maschinen identifiziert.
Grundlage der in Abbildung 39 gezeigten Einteilung der Instandhaltungstätigkeiten ist Unterscheidung zwischen den Unterklassen `preventive maintenance` (vorbeugende oder präventive Instandhaltung) und `corrective maintenance` (Bedarfsinstandhaltung). Alle weiteren Tätigkeiten lassen sich einer dieser Unterklassen zuteilen. Zusätzlich unterscheidet man bei der `preventive maintenance` zwischen den sogenannten `clock based`, `age based` und `condition based` Kategorien. Die Unterklasse der `corrective maintenance` wird aufgeteilt in die beiden Unterklassen `repair` und `compensating`. Im gegebenen Beispiel in Abbildung 39 lassen sich diesen Unterklassen noch jeweils eine bis vier konkrete Maßnahmen zuordnen, die in beschriebener Weise mit eindeutiger id gekennzeichnet sind.

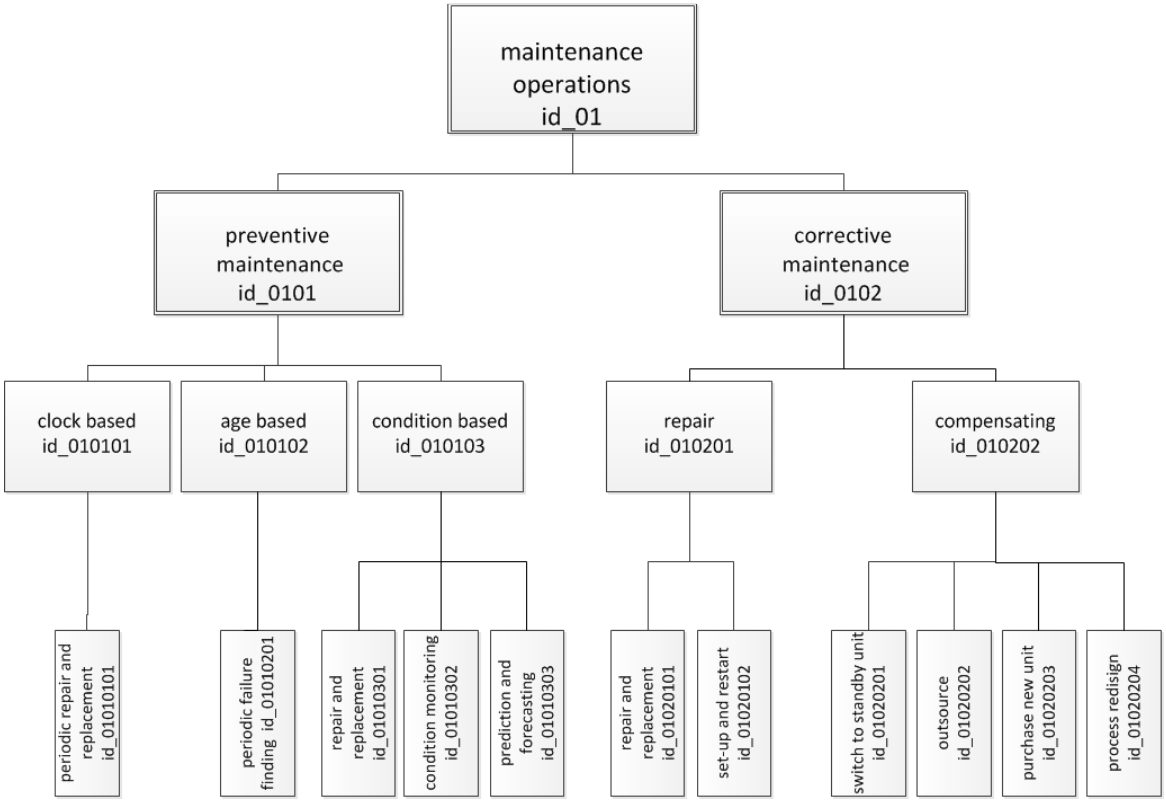


Abbildung 39. Zusammensetzung der Klasse `maintenance operations` **[119]**

Für den hier verfolgten Ansatz des Association Mapping sind vorgegebene Strukturen der zu analysierenden Klassen von großer Wichtigkeit. Die vorgestellten Klassen sind nur ein Teil derer, die für die Instandhaltung von Bedeutung sind. Beispielhaft seien an dieser Stelle die Betriebsstoffe für Maschinen und Anlagen genannt, deren unmittelbarer Bezug zu Instandhaltungstätigkeiten oder Maschinen hergestellt werden kann. Aus Gründen der Übersichtlichkeit beschränkt sich die Klassendarstellung sowie die virtuelle Applikation in Kapitel 6.2.2.2 auf die drei Klassen `practitioners`, `machines` und `maintenance operations`.

6.2.2.1 Konzeptionierung des Association Mapping mit CIMAWA

Nachdem die Grundlagen für das Association Mapping erläutert wurden, geht es im nächsten Schritt darum, das Konzept und den Ablauf näher zu erläutern. Der Ansatz des Association Mapping ist dabei nicht als eigenständige Anwendung, sondern als eine Art Add-on für bestehende CMMIS konzipiert. Hauptgrund dafür ist zum einen die benötigte Datengrundlage in Form der abgelegten Textdokumente und Klassenstrukturen, zum anderen ist davon auszugehen, dass durch die Integration einer solchen Lösung in ein bestehendes, etabliertes System die Akzeptanz der Nutzer höher ist, als bei einer entsprechenden Insellösung.
Abbildung 40 zeigt das, dem Association Mapping zugrunde liegende, Konzept. Die Datenbasis in Form der in der CMMIS-Datenbank gespeicherten Texte muss im ersten Schritt für den weiteren Ablauf vorverarbeitet werden. Hierzu gehören Text Mining Funktionalitäten ‘tokenization‘ und ‘part-of-speech tagging‘.

Erstere unterteilt den Text in sogenannte ‘tokens‘. Bei diesen handelt es sich um textuelle Einheiten wie Worte, Ziffern, Satzzeichen oder ähnliches [58]. Beim part-of-speech tagging werden die einzelnen Worte gelabelt, je nachdem, ob es sich bei dem entsprechenden Wort um ein Nomen, Verb, Adjektiv oder sonstiges handelt [58].
Aufbauend auf der Vorverarbeitung der Dokumente folgt der Prozessschritt der Entitätenidentifizierung. An dieser Stelle wird die Bedeutsamkeit der im letzten Abschnitt erarbeiteten Strukturen von `practitioners`, `machines` und `maintenance operations` deutlich. Die dargestellten Strukturvorgaben stellen die Hierarchien innerhalb der Klassen dar und definieren die Zugehörigkeiten der einzelnen Instanzen. Um die Assoziationen zwischen ausgewählten Instanzen berechnen und darstellen zu können, müssen diese zunächst eindeutig in den Texten identifiziert und zugeordnet werden. Als Grundlage für diese Funktionalität kann eine von Uhr et al. in [133] beschriebene Methode zur Anwendung gebracht werden, die Produkte und Produktkomponenten in Texten erkennt und entsprechend labelt (siehe Kapitel 6.3.1). Umfunktioniert und angepasst auf das hier beschriebene Anwendungsszenario, muss der Algorithmus, die in Abbildung 37, Abbildung 38 und Abbildung 39 beschriebenen Instanzen in den Texten erkennen und kennzeichnen. Erschwert wird dieser Prozessschritt durch den Umstand, dass es sich bei den in den CMMIS gespeicherten Texten größtenteils um digital verfasste Freitexte handelt. Das bedeutet wiederum, dass sowohl Tipp- als auch Rechtschreibfehler weitestgehend abgefangen werden sollten. Darüber hinaus muss bei der Entitätenerkennung berücksichtigt werden, dass jede Klasse, Unterklasse oder Instanz zwar mit einer eindeutigen ID gekennzeichnet ist, diese aber in den seltensten Fällen auch so im Text Verwendung findet. Beispielsweise kann ein und derselbe `practitioner` mit seiner ID, seinem Vor- und oder Nachnamen oder mit einer anderen Bezeichnung (z.B. „der Monteur“, „dieser Techniker“ oder per Personalpronomen) beschrieben werden. Gleiches gilt für die anderen beschriebenen Klassen, bei denen ähnliches auftreten kann. Unter Umständen ist es nötig, zur Auflösung dieser Problematik die Metadaten zum Textdokument genauer zu untersuchen und Strukturen zu erarbeiten, die synonym verwendete Begriffe abbilden. Wie im Mittelteil von Abbildung 40 dargestellt, sollen nach der Entitätenerkennung die Instanzen der Klassen in den Textdokumenten identifiziert und gekennzeichnet sein. Die Zuordnung der Wörter im Text zu den entsprechenden Instanzen wird in Abbildung 40 durch die entsprechenden Pfeile visualisiert. In der Praxis bedeutet diese Zuordnung ein eindeutiges Label zur exakten Identifizierung dieser Instanz.
Im Anschluss daran erfolgt die Auswahl der Instanzen, Klassen oder Unterklassen, mit denen eine Assoziationsanalyse durchgeführt werden soll. Diese erfolgt manuell durch den Nutzer und kann so zielorientiert angepasst werden. Die Ausgestaltung des Auswahlverfahrens und die Möglichkeiten, die dem Systemnutzer offen stehen, werden im nächsten Abschnitt erläutert, wenn der virtuelle Prototyp vorgestellt wird.

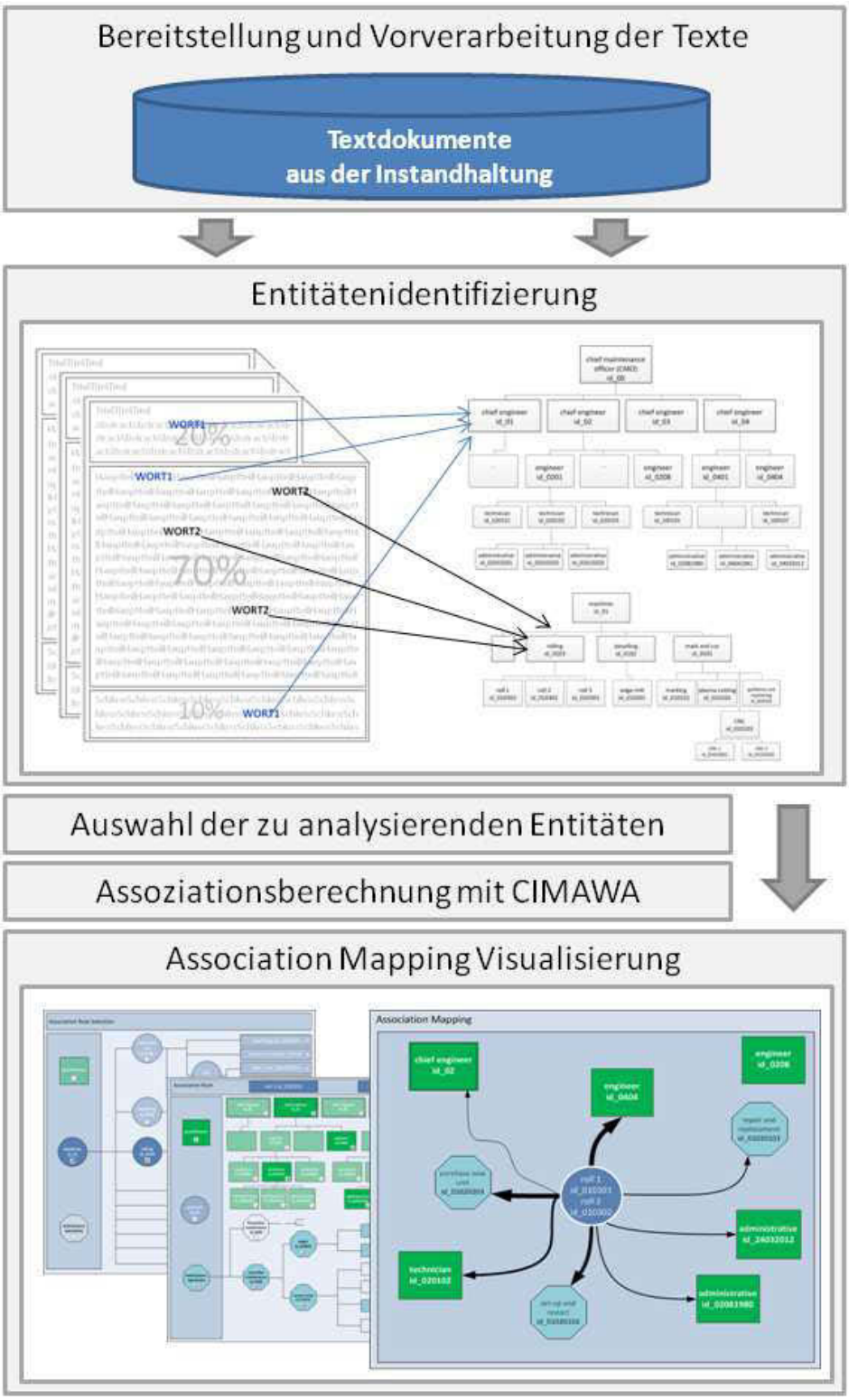


Abbildung 40. Konzept Association Mapping

Basierend auf der Auswahl des Nutzers müssen im nächsten Schritt die Assoziationen, die analysiert werden sollen, berechnet werden. Hierzu wird die Eigenentwicklung CIMAWA eingesetzt, deren Berechnung und Konzeption in Kapitel 4 im Detail dargestellt sind.
An dieser Stelle muss noch einmal auf die Entitätenerkennung hingewiesen werden. Ohne deren Ergebnisse sind im vorgestellten Szenario bei der Assoziationsberechnung keine

zufriedenstellenden Resultate zu erwarten. Wird eine Instanz einer Klasse ausgewählt, so erfolgt die Assoziationsberechnung auf allen, durch die Entitätenerkennung gefundenen, synonym verwendeten Begriffen. Die Assoziationsberechnung erfolgt dann nicht mehr auf den von handelnden Personen verwendeten Begriffen in den Texten, sondern auf den von der Entitätenerkennung vergebenen Label. Durch dieses Vorgehen wird sichergestellt, dass möglichst viele in der Textbasis vorhandenen Beziehungen erkannt und dargestellt werden.
Abschließend werden die Ergebnisse der Assoziationsanalyse visualisiert. Die Assoziationen zwischen den Instanzen sind mit Hilfe von Pfeilen unterschiedlicher Stärke sowie dem Abstand der Elemente zueinander abgebildet. Ein Beispiel für eine solche Darstellung ist im unteren Drittel der Abbildung 40 dargestellt.
Nachdem die Konzeption des Association Mapping erläutert wurde, soll im nächsten Abschnitt der Aufbau der zugehörigen Applikation erläutert werden. Hierbei liegt der Fokus auf der Nutzerinteraktion, sprich der Auswahl der Instanzen für die Assoziationsanalyse sowie der Präsentation der Ergebnisse in einer ‘Association Map‘.

6.2.2.2 Virtuelle Applikation zum Association Mapping als Add-on für CMMIS

Um die textuelle Metaanalyse im Instandhaltungsmanagement zu etablieren, wird im vorliegenden Abschnitt eine virtuelle Applikation entwickelt, die auf Basis der Assoziationsberechnung als ein Add-on für CMMIS-Lösungen konzipiert ist. Es werden verschiedene Screenshots dargestellt, anhand derer beispielhafte Nutzerinteraktionen nachvollzogen und erläutert werden. Abschließend sind exemplarisch einige Analysen eines solchen Systems beschrieben, mit Hilfe derer der Nutzen einer solchen Lösung begründet wird.
Vor der Analyse des hier erarbeiteten Ansatzes muss auf Zielsetzung und Potential einer spezialisierten textuellen Metaanalyse eingegangen werden. Eine wie auch immer umgesetzte Metaanalyse auf Textdokumenten setzt sich nicht zum Ziel, die Analyse der numerischen Werte oder die Ergebnisse anderer traditioneller Auswertungsmethoden zu ersetzen. Vielmehr sollen die Textanalysen diese ergänzen und unter Umständen Informationen zu Tage fördern, die auf anderem Wege nicht zu gewinnen sind. Dabei sehen wir am Institut für Wissensbasierte Systeme durch die Zusammenarbeit mit unseren Partnern aus der Industrie ein sich stetig vergrößerndes Potential in der zunehmenden Menge an digital verfügbaren und abgelegten Textdokumenten. Diese Textdokumente enthalten mitunter wertvolle Informationen, die im richtigen Kontext zu Wissen veredelt werden können. Zur Nutzbarmachung dieser textuellen Ressourcen sind innovative und soweit möglich automatisierte Analyseverfahren von großem Interesse. Einem solchen Ansatz in Form einer virtuellen Applikation ist dieses Kapitel gewidmet.
Die erste Interaktion mit dem Nutzer ist in Abbildung 41 angedeutet. Hierbei muss sich der Nutzer für die sogenannte ‘Association Root‘ entscheiden. Unter der Association Root versteht man diejenigen Instanzen oder Klassen, die im Fokus der anstehenden Assoziationsanalyse stehen sollen. Von den ausgewählten Instanzen oder Klassen ausgehend, werden in späteren Prozessschrittten die Assoziationen zu anderen Instanzen aus dem Bereich der Instandhaltung berechnet. Bezogen auf die Verwendung findende CIMAWA-

Berechnungsformel, bilden die hier ausgewählten Elemente den Stimulus, also denjenigen Begriff, zu dem die assoziierten Begriffe gefunden werden sollen. Abbildung 41 zeigt einen Screenshot, der als ‘Association Root Selection‘ bezeichneten Auswahl.

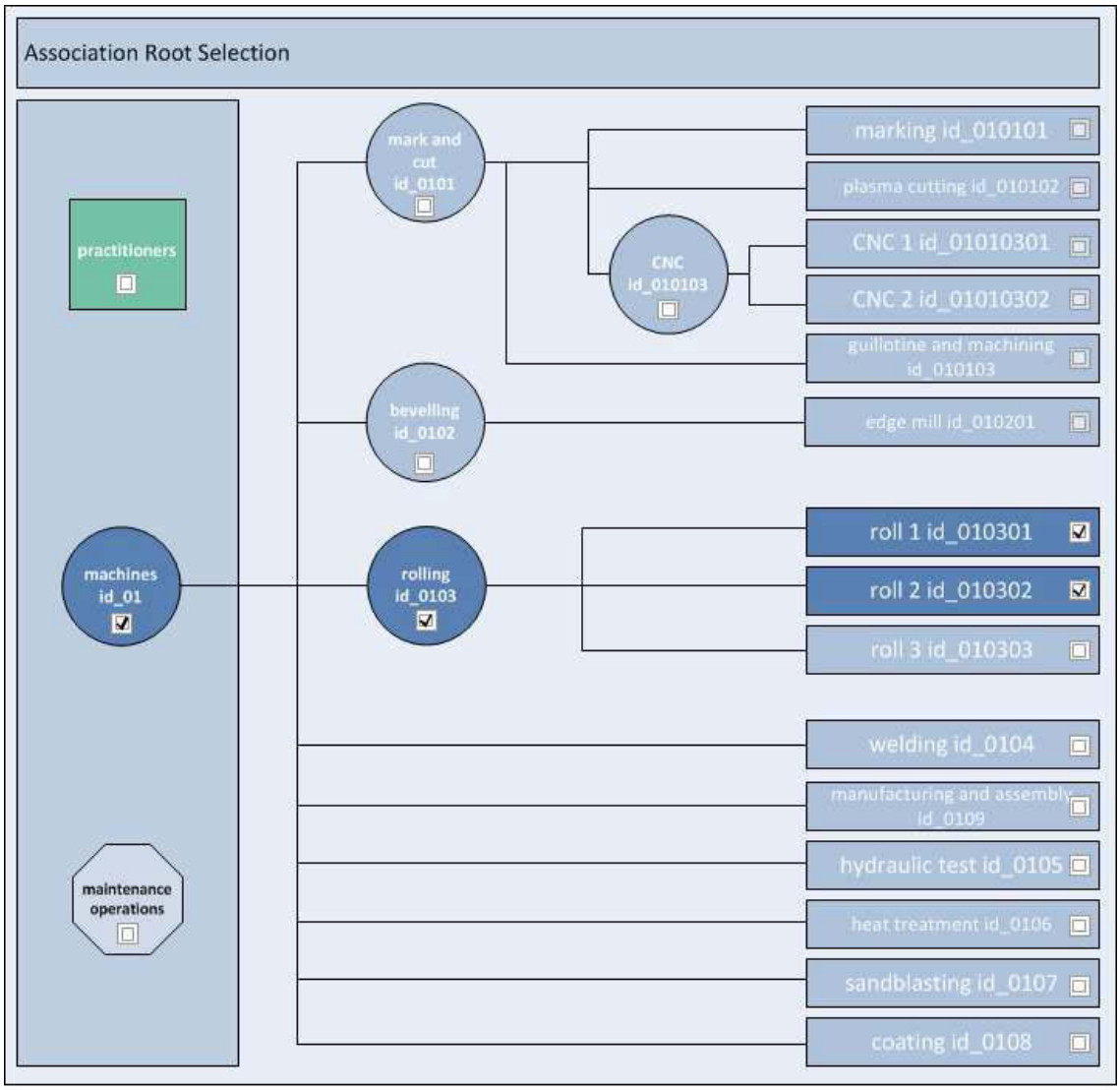


Abbildung 41. Auswahl Association Root [119]

Dabei befinden sich auf der linken Seite drei verschiedene geometrische Formen. Diese stehen symbolisch für die beschriebenen Klassen `practitioners`, `machines` und `maintenance operations`. Durch die Aktivierung der Checkbox im jeweils unteren Bereich der Elemente, eröffnet sich dem Nutzer die entsprechende Klassenstruktur auf der rechten Seite von Abbildung 41. Diese entstammt den in Abbildung 37, Abbildung 38 und Abbildung 39 erarbeiteten Diagrammen.
Im dargestellten Beispiel von Abbildung 41 wurde durch den Nutzer die Klasse `machines` durch die Checkboxaktivierung ausgewählt. Demzufolge wird die Klassenstruktur von `machines` auf der rechten Seite dargestellt. Alle enthaltenen Unterklassen und Instanzen sind für den Nutzer ebenfalls durch Aktivierung der jeweiligen Checkbox auswählbar. Die markierten Elemente werden farblich hervorgehoben, wobei nicht ausgewählte Instanzen oder Klassen durch erhöhte Transparenz im Hintergrund bleiben. Für den Nutzer besteht hier die

Möglichkeit, beliebige Kombinationen aus Klassen, Unterklassen oder Instanzen auszuwählen, je nachdem, auf welcher Basis die folgende Assoziationsanalyse durchgeführt werden soll.
Im Beispiel wurde zunächst die Unterklasse `rolling` ausgewählt, wobei hier nicht alle Maschinen, sondern nur `roll 1 id_010301` und `roll 2 id_010302` markiert wurde. In der späteren Assoziationsanalyse werden diese beiden Instanzen und sämtliche synonym verwendeten Begriffe für beide Maschinen den Ausgangspunkt der Berechnung bilden. Ohne die Markierung dieser beiden speziellen Instanzen wären implizit sämtliche Instanzen der Unterklasse `rolling` als Association Root ausgewählt. Gleiches gilt für die Klasse `machines`. Ohne Auswahl einer Unterklasse oder Instanz gehen alle Maschinen in die Assoziationsberechnung ein. Einzige Grundregel bei der Wahl der Association Root ist, dass jeweils nur Untermengen einer der drei Klassen `practitioners`, `machines` und `maintenance operations` gewählt werden können. Das Mischen von Elementen aus zwei oder mehreren dieser Klassen ist nicht möglich.
Nachdem die Auswahl der Association Root getroffen wurde, muss im nächsten Schritt durch Interaktion mit dem Nutzer die Auswahl der potentiell assoziierten Elemente stattfinden. Abbildung 42 zeigt den entspechenden Auswahlbildschirm. Das in Abbildung 41 begonnene Anwendungsbeispiel wird in Abbildung 42 fortgeführt. Zu erkennen ist dies an den im oberen Bereich als Association Root gekennzeichneten Elementen `roll 1 id_010301` und `roll 2 id_010302`, welche der zuvor getroffenen Auswahl der Association Root entsprechen. Der linke Teil des Auswahlmenüs in Abbildung 42 gleicht dem der Auswahl der Association Root in Abbildung 41. Unterschieden wird auch hier zunächst zwischen den drei Klassen `practitioners`, `machines` und `maintenance operations`. Im Unterschied zur vorherigen Auswahl besteht hier jedoch die Möglichkeit mehrere Klassen gleichzeitig auszuwählen, deren Instanzen zu markieren und so eine gemischte Auswahl aus mehreren Klassen zu treffen. In der Beispielauswahl zeigt sich, dass neben der Klasse `practitioners` auch die Klasse `maintenance operations` ausgewählt wurde. Die Klasse der `machines` findet in diesem Beispiel keine Berücksichtigung und wurde nicht zur Analyse ausgewählt. Das grün gehaltene Diagramm mit der Klassenstruktur der `practitioners` ist dabei im oberen Teil von Abbildung 42 dargestellt und das türkise `maintenance operations` Strukturdiagramm ist im unteren Teil zu erkennen.
Aus dem Bereich der `practitioners` wurden der `chief engineer id_02` der `engineer id_0208` und `engineer id_0404`, der `technician id_020102` sowie `administrative id_02081980` und `administrative id_24032010` ausgewählt. Die zusätzlich markierten Instanzen der `maintenance operations` fließen ebenfalls in die folgende Analyse ein. Diese sind in Gänze lokalisiert im Bereich der `corrective maintenance`. Hier wurden beide Teilbereiche `repair` und `compensating` ausgewählt, wobei die Instanzen `repair and replacement id_01020101`, `set-up and restart id_01020102` und `purchase new unit id_01020204` markiert sind. Alle in Abbildung 42 markierten Instanzen sind Teil der anschließenden Auswertung. In dieser werden zunächst die CIMAWA-Werte zwischen diesen und der ausgewählten Association Root berechnet.

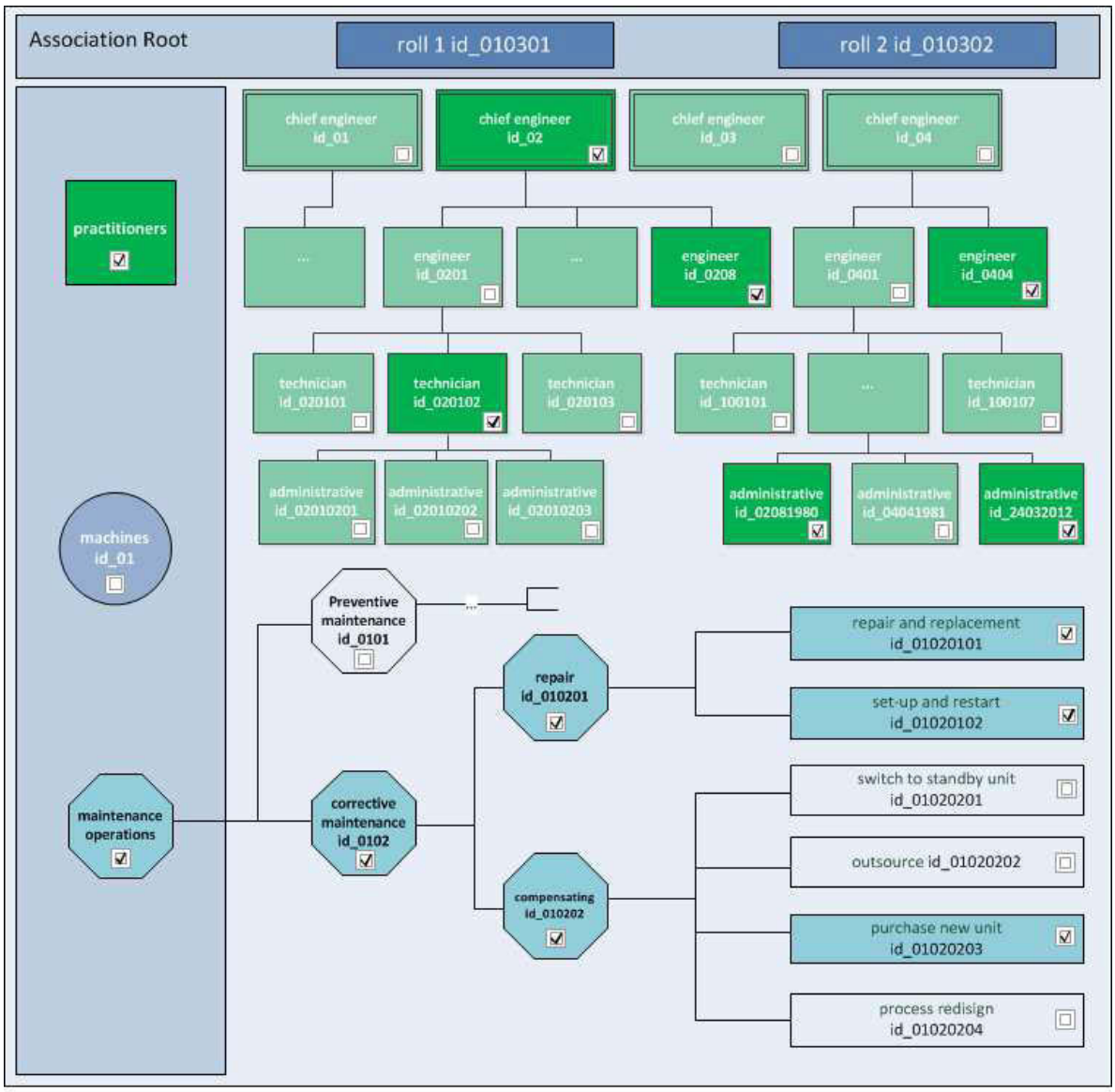


Abbildung 42. Auswahl der zu analysierenden Elemente [119]

Wie die abschließende Visualisierung des Association Mapping in Abbildung 43 zeigt, werden ausschließlich die Assoziationen zwischen Association Root und den zu analysierenden Elementen berechnet und dargestellt. Die Assoziationen zwischen den in Abbildung 42 ausgewählten Elementen ist nicht Gegenstand der Analyse.

Nachdem die Auswahl aller benötigten Elemente vollzogen wurde, kann die Assoziationsberechnung angestoßen werden. Hierzu kommt die in Formel 4 beschriebene CIMAWA-Assoziationsberechnung zur Anwendung. Abbildung 43 komplettiert den in den vorangestellten Abbildungen begonnenen Anwendungsfall und visualisiert zugleich exemplarisch die Ergebnisse der Assoziationsberechnung.

Im Fokus der Analyse und zugleich im Zentrum der Visualisierung in Abbildung 43 steht der Association Root. Dieser wurde zuvor durch den Nutzer ausgewählt (Abbildung 41) und besteht im dargestellten Fall aus den beiden Maschinen `roll 1 id_010301` und `roll 2 id_010302`. Im Rahmen der Entitätenerkennung wurden alle Begriffe in den zugrunde liegenden Texten, die diese beiden konkreten Maschineninstanzen beschreiben, identifiziert

und entsprechend gelabelt. Gleiches gilt für die Elemente der Klassen `practitioners` und `maintenance operations`.
Die Stärke der Assoziation zwischen den Elementen ist in Abbildung 43 zum einen durch die Verbindungspfeile und deren Stärke und zum anderen durch den Abstand der assoziierten Elemente zur Association Root visualisiert. Hierbei gilt, je stärker die Pfeile und geringer der Abstand der Elemente zum Zentrum, desto stärker die berechnete Assoziation.

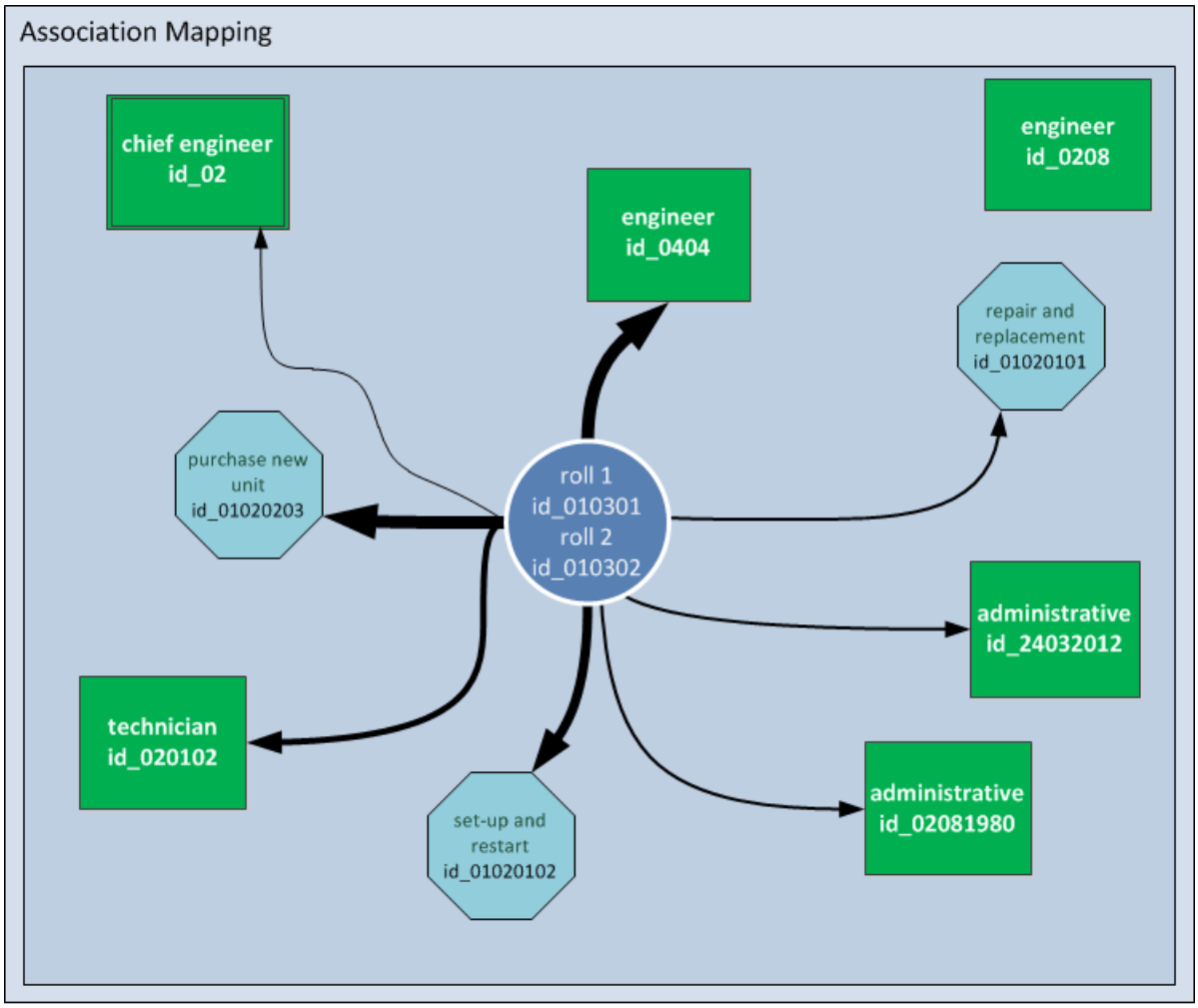


Abbildung 43. Visualisierung der Ergebnisse des Association Mapping [119]

Die beispielhaft dargestellte Association Map in Abbildung 43 enthält alle zuvor ausgewählten Elemente, auch solche, die auf Basis der Textdokumente keine Assoziation zum Association Root aufweisen. Ein Beispiel für eine Instanz ohne Assoziation ist der `engineer id_0208` im oberen rechten Bereich der Association Map in Abbildung 43. Basierend auf der Analyse der Textdokumente besitzt dieser Ingenieur keine Verbindung zu den ausgewählten Instanzen der Klasse `rolling`. Die anderen Instanzen der Klasse `practitioners`, jeweils visualisiert durch grüne Rechtecke, weisen hingegen mehr oder weniger starke Assoziationen auf. Die stärkste Assoziation besteht zwischen der Association Root und dem `engineer id_0404`. Die übrigen vier Instanzen der `practitioners` Klasse (`chief engineer id 02`, `technician id_020102`, `administrative`

id_02081980, administrative id_24032012) weisen weniger starke Assoziationen auf, wobei die schwächste zum chief engineer id_02 im linken oberen Bereich der Abbildung 43 besteht.
Die drei aus der Klasse der maintenance operations ausgewählten Instanzen sind in Abbildung 43 dargestellt durch türkis gehaltene Achtecke. Die stärkste Assoziation, ausgehend von der Association Root, besteht zu purchase new unit id_01020203. Eine im Vergleich dazu leicht abgeschwächte Assoziation ist zu set-up and restart id_01020102 im unteren Bereich der Association Map zu erkennen. Die schwächste berechnete Verbindung besteht zu der Instanz repair and replacement id_01020101 am oberen rechten Rand der Association Map.

6.2.2.3 Anwendung des Association Mapping im Instandhaltungsmanagement

Nachdem der Aufbau einer Applikation zur textuellen Metaanalyse beschrieben und ein Beispielergebnis in Abbildung 43 dargestellt wurde, soll die Frage der Anwendbarkeit eines solchen Ansatzes beantwortet werden.
Wie bereits zuvor angeführt, bieten kommerzielle CMMIS-Lösungen keine Funktionalitäten, die eine textuelle Metaanalyse von Textdokumenten ermöglicht. Die hier vorgeschlagene Methode des Association Mapping bietet die Möglichkeit einer solchen Metaanalyse von großen Textmengen und so das Einbeziehen von abgelegtem Wissen aus der Instandhaltung in den Entscheidungsprozeß der handelnden Personen. Die dabei erzielten Ergebnisse können als zusätzliche Quelle zu den üblichen statistischen und mathematischen Auswertungen der CMMIS angesehen werden [119]. Als alleinige Basis zur Entscheidungsfindung sind die Resultate der textuellen Analyse in den meisten Fällen unzureichend und sollten vielmehr als Ergänzung etablierter Verfahren angesehen werden.
Als ein Beispielszenario kann die Assoziation zwischen den Instanzen der Klasse machines (roll 1 id_010301, roll 2 id_010302) als Association Root im Zentrum der Abbildung 43 und dem engineer id_0404 herangezogen werden. Den handelnden Personen im Instandhaltungsmanagement ist der Zusammenhang zwischen diesen Maschinen und dem Ingenieur unter Umständen bereits bekannt. Denkbar ist jedoch, dass sich der CMO durch die Betrachtung der Association Map dazu veranlasst sieht, sich vor einer Entscheidung bezüglich dieser beiden Maschinen, die Wartungsberichte des entsprechenden Ingenieurs anzuschauen oder diesen Mitarbeiter gezielt zu kontaktieren.
Zusammenfassend liegt der Nutzen eines textuellen Ansatzes in der automatisierten Analyse von großen Textmengen, durch die bislang nicht oder unzureichend genutztes Wissen aufgearbeitet und nutzbar gemacht werden kann. Herauszustellen gilt in diesem Zusammenhang im Besonderen, dass die Metaanalyse auf Bestandsdaten aufsetzt und keine zusätzlichen Methoden zur Wissensextraktion von den Mitarbeitern erfordert. In diesem Sinne ist die textuelle Metaanalyse dazu geeignet, das im Unternehmen entwickelte Wissen zu bewahren und durch stetige Nutzung darüber hinaus neues Wissen zu entwickeln. Dabei ist zu beachten, dass die vorgeschlagene Klassenstruktur nicht auf die drei Klassen practitioners, machines und maintenance operations beschränkt sein muss, sondern, dass die Erweiterung dieser Liste weitere Anwendungsbereiche eröffnen kann.

Beispielsweise bergen die potentiellen Erweiterungsklassen `resources`, `external practictioners`, `vendors` oder `customer` Anwendungsfelder, die über das Instandhaltungsmanagement hinausreichen und für andere Geschäftsbereiche eines Unternehmens von Interesse sind.

6.3 CIMAWA zur kontextbasierten Bereitstellung von Textdokumenten im Produktverbesserungsprozess

Die in diesem Kapitel vorgestellten Ergebnisse basieren im Wesentlichen auf den vom Autor mitverfassten Artikeln ‘Kontextbasierte Bereitstellung von Textdokumenten im Produktverbesserungsprozess‘ [133], veröffentlicht im Fachmagazin *wt Werkstattstechnik online* sowie dem Konferenzbeitrag ‘Concept for Improving Industrial Goods via Contextual Knowledge Provision‘ [134], der auf der Wissensmanagement Konferenz *i-Know* ’12 präsentiert und veröffentlicht wurde.
Der technologische Fortschritt und der verstärkte Konkurrenzdruck durch die fortschreitende Globalisierung haben dazu geführt, dass sich die Produktentwicklungszeiten stetig verkürzen [133]. Nichtsdestotrotz sind die Qualitätsanforderungen an Produkte, insbesondere im Maschinen- und Anlagenbau, bezüglich Laufzeit und Fehleranfälligkeit unvermindert hoch. Dieser Zwiespalt, der immer kürzer werdenden Produktentwicklungszeiten und der nach wie vor hohen Qualitätsanforderungen, stellt hohe Anforderungen an die Arbeit der Produktentwickler. Diese müssen Schwachstellen von Produkten erkennen und diese unter Anwendung alternativer Lösungen beseitigen.
Zur Unterstützung des Produktentwicklers im Produktverbesserungsprozess werden Daten in der Produktnutzungsphase gesammelt und dem Produktentwickler zur Verfügung gestellt. Hierbei unterscheidet man zwischen Daten, die am Produkt selbst, beispielsweise über Sensoren an einer Maschine oder im Umfeld, zum Beispiel im Kundenservice oder der Instandhaltung gewonnen werden. Im Rahmen des von unserem Institut für Wissensbasierte Systeme und Wissensmanagement in Kooperation mit dem Lehrstuhl für Maschinenbauinformatik der Ruhr-Universität Bochum erfolgreich abgeschlossenen Forschungsprojektes der DFG namens „WiRPro – *Erweiterung des Product Lifecycle Managements durch wissensbasierte Rückführung von Produktnutzungsinformationen in die Produktentwicklung*“, wurde ein Feedback Assistenz System (FAS) zur Verwaltung und Analyse der Daten aus der Produktnutzung umgesetzt [135]. Dieses FAS wurde mit der Zielsetzung entwickelt, ausschließlich objektives Feedback zu prozessieren. Das objektive Feedback besteht aus strukturierten Daten wie numerischen oder booleschen Werten, die analysiert werden, um Aussagen zu Verbesserungspotentialen an dem in der Nutzung befindlichen Produkt abzuleiten [136]. Die Notwendigkeit neben den strukturierten auch unstrukturierte Daten und insbesondere Texte in den Produktverbesserungsprozess einfließen zu lassen, wurde zur Laufzeit des WiRPro Forschungsprojektes diagnostiziert.
Es wurde festgestellt, dass große Teile der für das Unternehmen relevanten Informationen, beispielsweise in Montage- Wartungs- oder Störungsprotokollen, Reklamationen, Garantie- und Störungsmeldungen, Kundenkorrespondenzen oder Diagnoseberichten, lokalisiert sind [133]. Mit der wachsenden Menge an archivierten, textbasierten Dokumenten stellen sich neue Herausforderungen, da mit der Anzahl der Texte auch die Menge der verfügbaren Informationen steigt. Auf der einen Seite bedeutet die Verbreiterung der unternehmensinternen Datenbasis Chancen bezüglich der Know-how Sicherung einer Unternehmung. Auf der anderen Seite müssen die Probleme berücksichtigt werden, die den Bereich der Auffindbarkeit und somit zugleich die Wiederverwendbarkeit von Dokumenten

betreffen [133]. Ohne entsprechende Methoden zur automatischen Analyse von großen Textsammlungen, können die negativen Auswirkungen der beschriebenen Entwicklung überwiegen. Mit dem Ziel, diese Informationen im Produktverbesserungsprozess nutzbar zu machen und die Potentiale der sich vergrößernden Datenbasis im Produktverbesserungsprozess zu nutzen, wurde die in dem vorliegenden Kapitel konzipierte Text Mining-Lösung entwickelt. Der Ansatz sieht vor, das FAS um eine Empfehlungskomponente zur Integration von Texten zu erweitern, um so den Feedbackfluss zu komplettieren [133]. Zur Realisierung werden mehrere eigenentwickelte Text Mining-Methoden in einem integrierten Ansatz kombiniert und auf das Anwendungsszenario abgestimmt.
Der entwickelte Ansatz basiert auf den im Unternehmen gespeicherten Textdokumenten. Wie bereits ausgeführt, ist aufgrund der Quantität dieser Dokumente eine Lösung erstrebenswert, die den Nutzer beim Auffinden von Dokumenten im jeweiligen Aufgabenkontext unterstützt. Daher wurde als Zieldefinition die kontextbasierte Bereitstellung von Textdokumenten im Produktverbesserungsprozess gewählt. Produktentwickler stellen die Nutzergruppe des Textempfehlungssystems, wobei der Kontext durch das zu verbessernde oder neu zu entwickelnde Produkt definiert ist.
Im folgenden Abschnitt wird die Konzeption des Ansatzes im Detail dargestellt sowie die Funktionsweise der Einzelkomponenten erläutert.

6.3.1 Konzeption der kontextbasierten Textempfehlung auf Basis von CIMAWA

Abbildung 44 zeigt den Aufbau des konzipierten Textempfehlungssystems. Wie die Einzelkomponenten aufgebaut sind und wie diese im integrierten Konzept interagieren, soll im Folgenden thematisiert werden.
Als Eingangsdaten werden die in Datenbanken abgelegten Textdokumente der Unternehmung benötigt. Da das Anwendungsszenario im Kontext der Produktentwicklung / Produktweiterentwicklung angesiedelt ist, kann sich auf die Textdokumente aus diesem Bereich beschränkt werden. Wie in Art und Weise in Kapitel 6.2.2.1 beschrieben, wird die Textbasis für die nächsten Schritte vorverarbeitet. Wichtige Schritte sind abermals ‘tokenization‘ und ‘part-of-speech tagging‘. Auf der vorbereiteten Textbasis können sich die nächsten Prozessschritte anschließen.
Die Produkterkennung ähnelt der Entitätenidentifizierung im Anwendungsfall der textuellen Metaanalyse von Instandhaltungsdokumenten aus dem letzten Kapitel. Genau genommen handelt es sich hierbei um einen Spezialfall dieser Methode. Ausgehend von einer definierten Produktstruktur, die Zusammenschlüsse von Produkten zu Produktgruppen, aber auch Produktkomponenten zu Produkten enthält, werden sämtliche Texte nach Instanzen von Produkten durchsucht und entsprechend gelabelt. Das Verfahren arbeitet auf einer zentral vorgehaltenen Produktdatenbank, die neben den genauen Produktbezeichnungen zusätzliche Informationen zum Produkt selbst enthält [133]. Die dort hinterlegte/n Produktbezeichnung/en dienen als ‘pattern‘ für die Suche in der Textsammlung, wobei keine klassische Volltextsuche, sondern ein spezialisiertes Verfahren zur Anwendung gebracht wird, das Sonderzeichen wie Bindestrichen oder Leerzeichen eine andere Gewichtung zuweist

als Buchstaben oder Ziffern [133]. Die korrekte Erkennung der Produktinstanzen in den Textdokumenten bildet eine wichtige Voraussetzung für die sich anschließende Assoziationsanalyse mit CIMAWA. Ohne genaue Zuordnung, welcher Begriff welche Produktinstanz im Text darstellt, wird die Assoziationsberechnung erschwert und die Berechnungsergebnisse können negativ beeinflusst werden.

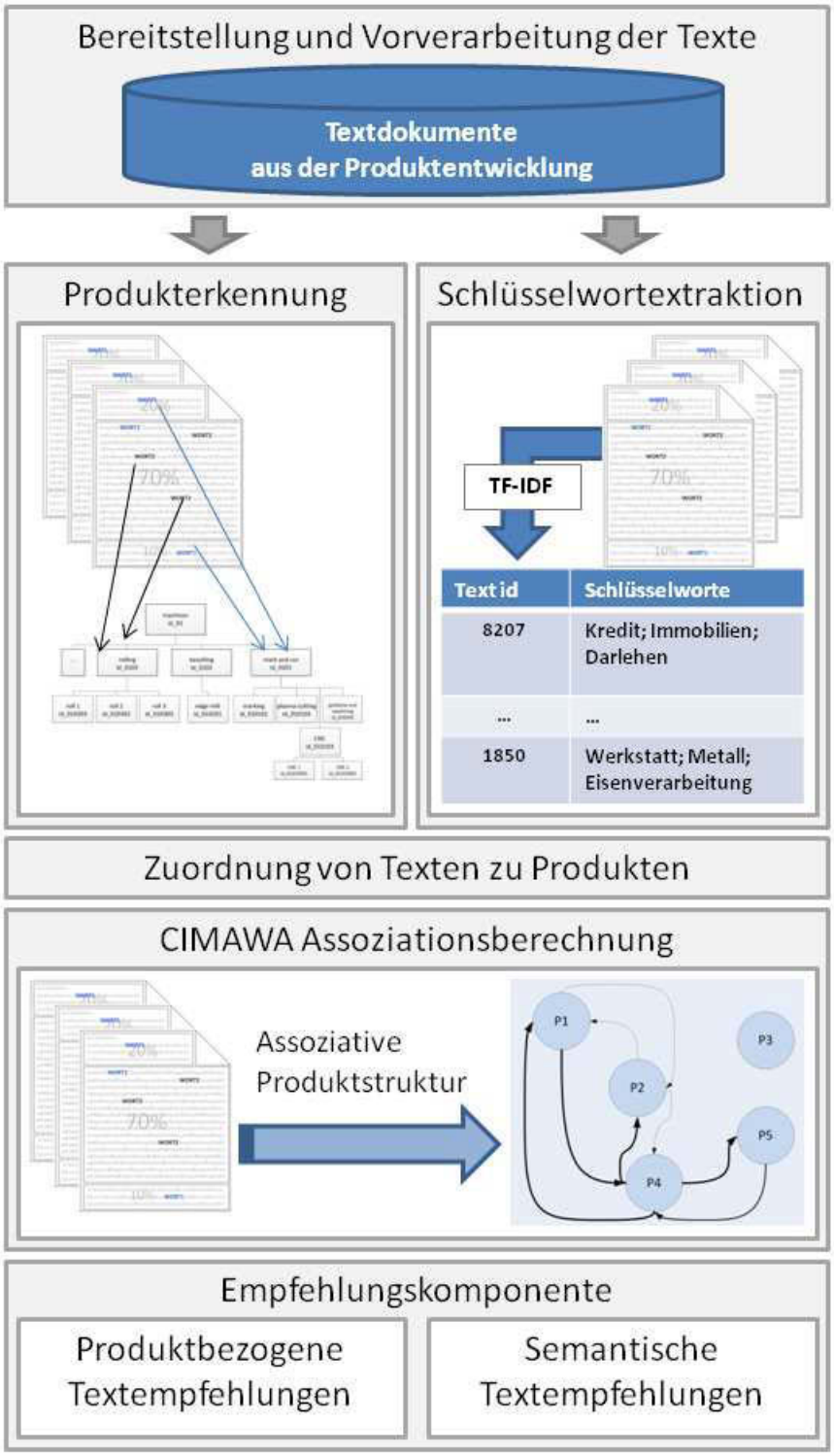


Abbildung 44. Konzept kontextbasierte Bereitstellung von Textdokumenten im Produktverbesserungsprozess

Die Schlüsselwortextraktion bildet einen weiteren für die spätere Textempfehlung essentiellen Verarbeitungsschritt. Die Texte der Datenbasis werden in diesem Prozessschritt analysiert, um für jeden Text die wichtigsten Worte zu identifizieren. Das Grundprinzip der angewendeten Methode kann heruntergebrochen werden auf eine einfache These: Worte, die in vielen Texten vorkommen, sind tendenziell für den einzelnen Text weniger bedeutend als solche, die nur vereinzelt, dafür aber gehäuft in Einzeltexten vorkommen. Aus technischer Sicht kommt zur Realisierung ein TF-IDF-Derivat zum Einsatz, das am Institut für Wissensbasierte Systeme entwickelt wurde [78]. Im Rahmen von Kapitel 6.1.2 wurde bereits im Detail auf die Schlüsselwortextraktion eingegangen. Aus diesem Grund soll diese hier nicht erneut thematisiert werden. Wichtig bleibt festzuhalten, dass nach der Extraktion der Schlüsselworte zu jedem Text in der Textbasis eine Liste von Schlüsselworten verfügbar ist. Beispielhafte Ergebnisse sind in Abbildung 44 durch eine Tabelle mit ‘Text id‘ und identifizierten Schlüsselworten visualisiert.

Auf Basis der Produkterkennung auf der einen und der Schlüsselwortextraktion auf der anderen Seite, kann die Zuordnung von Texten zu Produkten erfolgen. In diesem Schritt werden die jeweiligen Schlüsselworte der Texte daraufhin überprüft, ob diese einen Begriff enthalten, der in der Produkterkennung als Produkt gelabelt wurde. Ist dies der Fall, so kann davon ausgegangen werden, dass in diesem speziellen Text, das dem Label entsprechende Produkt eine große Bedeutung zuzumessen ist. Dementsprechend wird der Text diesem Produkt zugeordnet. Hierbei ist eine Mehrfachzuordnung möglich. Diese tritt ein, wenn die Schlüsselworte eines Textes mehrere Produktbezeichnungen enthalten. Auf Grundlage dieser Zuordnung von Texten zu Produkten können im weiteren Verlauf Textempfehlungen für den Produktentwickler abgegeben werden (siehe Abbildung 47).

Mit Abschluss der beschriebenen Vorverarbeitungen kann die CIMAWA-Assoziationsberechnung angestoßen werden. Im Gegensatz zu dem im letzten Kapitel vorgestellten Ansatz des Association Mappings im Instandhaltungsmanagement, beschränken sich die zu berechnenden Assoziationen im vorliegenden Anwendungsfall auf Produkte bzw. Produktkomponenten. Eine Auswahl von Instanzen verschiedener Oberklassen ist nicht nötig. Um die Übersichtlichkeit zu verbessern, kann hier für den Fall, dass zu viele Produkte vorhanden sind, eine Einschränkung vorgenommen werden. Die wichtigste Voraussetzung für eine aussagekräftige Assoziationsberechnung zwischen den Produkten sind die Ergebnisse der Produkterkennung in den Texten. Die in diesem Schritt gelabelten und eindeutig zugeordneten Begriffe bilden die Basis der CIMAWA-Assoziationsberechnung. CIMAWA arbeitet auf den Labeln und nicht auf den ursprünglichen Begriffen und stellt so sicher, dass möglichst viele, für die Bezeichnung ein und desselben Produktes verwendeten Begriffe, in die Berechnung einfließen. Ergebnis der Assoziationsberechnung ist die sogenannte assoziative Produktstruktur (siehe Abbildung 48). In dieser werden die berechneten Ergebnisse in der aus dem Association Mapping bekannten Weise visualisiert.

In Abbildung 44 wird eine solche Struktur exemplarisch gezeigt, wobei P1, …, P5 die Produkte darstellen und die verbindenden Pfeile und deren Stärke die CIMAWA-Assoziation visualisieren.

Die Empfehlungskomponente im unteren Bereich von Abbildung 44 baut auf den Ergebnissen der vorgenannten Prozessschritte auf und nutzt diese um produktspezifische

Textempfehlungen zu entwickeln. Dabei können im Grundsatz zwei Arten von Empfehlungen unterschieden werden. Die produktbezogenen und die semantischen Textempfehlungen.
Wie die Bezeichnung vermuten lässt, steht bei den produktbezogenen Textempfehlungen das einzelne Produkt und dessen Beziehungen zu anderen Produkten im Vordergrund. Hierzu wählt der Produktentwickler zunächst das Produkt aus, zu dem er Informationen in Form von Texten benötigt. Durch die zuvor durchgeführte Zuordnung von Texten zu Produkten können unmittelbar die relevanten Texte zu diesem Produkt als Informationsquelle empfohlen werden. Darüber hinaus wird die durch CIMAWA berechnete assoziative Produktstruktur dafür verwendet, um die Empfehlung um relevante Texte zu erweitern. Dies geschieht durch die Analyse der assoziativen Produktstruktur. Wird bei der Analyse festgestellt, dass das aktuell betrachtete Produkt starke Assoziationen zu einem weiteren Produkt oder einer Produktkomponente aufweist, so wird die Menge der empfohlenen Texte um diejenigen Texte erweitert, die diesem Produkt direkt zugeordnet sind. Demzufolge wird der Produktentwickler mit Informationen aus dem direkten Umfeld des ausgewählten Produktes versorgt. Hierbei kommen Textempfehlungen zustande, die mit klassischen Methoden wie der Volltextsuche nicht zu erzielen sind. Ein konkretes Beispiel für eine solche Erweiterung der empfohlenen Texte findet sich im nächsten Abschnitt.
Als Ergänzung der zuvor beschriebenen produktspezifischen Textempfehlung können zusätzlich die sogenannten semantischen Textempfehlungen definiert werden. Hier stehen nicht die Produkte, sondern die mit den Produkten in Relation stehenden Texte im Fokus der Analyse. Die semantischen Textempfehlungen beziehen die Texte selbst und deren Ähnlichkeit untereinander mit ein. Die Textempfehlungen zu einem spezifischen Produkt werden zunächst analysiert und anschließend wird in der vorhandenen Textbasis nach Texten gesucht, die semantische Ähnlichkeit zu den Ausgangstexten besitzen. Dabei ist zu beachten, dass in diesem Fall auch Texte in die Empfehlungsmenge aufgenommen werden, die ansonsten keinen Bezug zu dem aktuell ausgewählten Produkt vorweisen.
Von Nutzen kann eine solche semantische Textempfehlung zum Beispiel im folgenden Fall sein: Eine direkte Textempfehlung enthält eine Problembeschreibung mit Fehlerursache und Schadensbericht zu dem aktuell ausgewählten Produkt, ohne jedoch einen Lösungsansatz zu formulieren. Ein anderer Text in der Datenbasis beschreibt ein ähnliches Problem mit entsprechendem Hinweis auf einen Lösungsansatz, bei einem vollkommen anderen Produkt. Dieser Lösungsansatz kann allein durch die Identifikation der semantischen Ähnlichkeit der beiden Texte in die Textempfehlung aufgenommen werden und so einen Mehrwert für den Produktentwickler darstellen.
Für die Analyse der semantischen Verwandtschaft von Texten kommt ein vereinfachtes Verfahren zum Einsatz, das im Ursprung von Klahold in [94] entwickelt wurde. Die Besonderheit dieser Methode liegt darin, dass die Beziehungen zwischen den Texten nicht notwendigerweise symmetrisch sind. In anderen Worten bedeutet dies, dass Text T_2 zugleich semantisch verwandt zu Text T_1 sein kann, ohne dass gleichzeitig T_1 semantische Verwandtschaft zu T_2 aufweisen muss. Wie diese auf den ersten Blick widersprüchliche Verwandtschaftsbeziehung zu Stande kommt, wird im Folgenden durch die Erläuterung des Verfahrens beschrieben.
Basis der Bestimmung der semantischen Verwandtschaft ist die Schlüsselwortextraktion, die vor der Berechnung durchzuführen ist. Dabei ist die semantische Verwandtschaft zwischen

zwei Texten wie folgt zu verstehen. Enthält ein Text T_2 eine oder mehrere Schlüsselworte eines Textes T_1, so gilt T_2 als semantisch verwandt zu T_1. Da im Umkehrschluss die in T_2 gefundenen Schlüsselworte aus T_1 nicht automatisch Schlüsselworte, also bedeutende Worte in T_2 sein müssen, bedeutet dies nicht notwendigerweise eine semantische Verwandtschaft von T_1 zu T_2. Die folgende Abbildung 45 greift dieses Beispiel auf und versucht anhand von konkreten Texten die semantische Verwandtschaft zwischen diesen exemplarisch zu erklären.

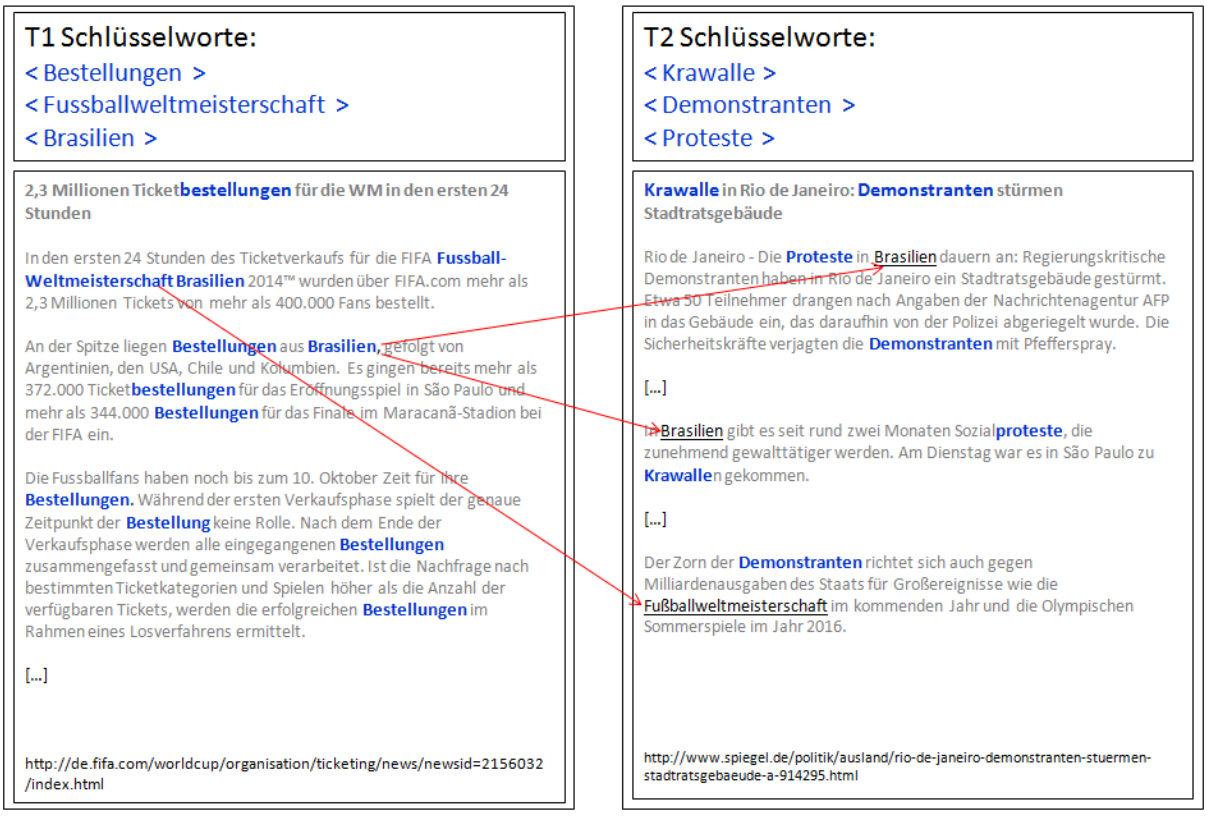


Abbildung 45. Beispiel für die semantische Verwandtschaft von Texten

Die beiden in Abbildung 45 verwendeten Texte sind jeweils online abgerufene Nachrichtentexte. Der Text T_1 auf der linken Seite der Abbildung entstammt [137] und T_2 stellt Auszüge aus [138] auf der rechten Seite dar. Die jeweiligen Schlüsselworte der Texte sind sowohl im oberen Bereich in Listenform als auch in den Texten durch Einfärbung hervorgehoben. Zunächst soll die Frage beantwortet werden, ob der Text T_2 semantische Verwandtschaft zu Text T_1 aufweist. Hierzu wird eine Volltextsuche angestoßen, in der T_2 auf die Schlüsselworte von T_1 untersucht wird. In dem konkreten Beispiel bedeutet dies die Suche nach den Begriffen ‘Bestellungen‘, ‘Fußballweltmeisterschaft‘ und ‘Brasilien‘ in T_2. Die Ergebnisse der Suche sind in Abbildung 45 mittels Unterstreichung der gefundenen Begriffe angedeutet. Die Zuordnung wird visualisiert durch die Verbindungspfeile zwischen den entsprechenden Worten. In T_2 wurden die beiden Begriffe ‘Brasilien‘ und ‘Fußballweltmeisterschaft‘ gefunden. Aufgrund dieser Suchergebnisse ist der Text T_2 semantisch verwandt zu Text T_1, da ein oder mehrere Schlüsselwörter von T_1 in T_2 vorkommen. Als nächstes wird die umgekehrte Verwandtschaftsrichtung beider Texte getestet. Entsprechend wird in T_1 nach den in T_2 identifizierten Schlüsselworten ‘Krawalle‘,

‘Demonstranten‘ und ‘Proteste‘ gesucht. Diese kommen in T_1 nicht vor, woraus zu schließen ist, dass T_1 keine semantische Verwandtschaft zu T_2 besitzt. Dieses Beispiel verdeutlicht, dass der Ansatz zum Ermitteln von semantischen Verwandtschaften zwischen Texten asymmetrische Beziehungen zulässt und solche identifiziert.
Die in der zuvor beschriebenen Art gefundenen Ergebnisse werden in einer Matrix visualisiert und gespeichert, die die binäre Verwandtschaft zwischen sämtlichen in der Textbasis befindlichen Texten darstellt. In Abbildung 46 ist eine solche Beispielmatrix visualisiert. Diese beschreibt die semantische Verwandtschaft zwischen sechs Texten (T_1, T_2, …, T_6). Neben den trivialen semantischen Verwandtschaften von jedem Text zu sich selbst, welche durch die Markierungen auf der Diagonalen gekennzeichnet sind, werden sämtliche anderen Textverwandtschaften repräsentiert.

	T_1	T_2	T_3	T_4	T_5	T_6
T_1	●	●		●		
T_2		●	●	●		
T_3		●	●		●	
T_4	●	●	●	●		
T_5	●	●	●		●	
T_6						●

Abbildung 46. Matrix für die semantische Verwandtschaft zwischen Texten [133]

Auch das in Abbildung 45 im Detail dargestellte Beispiel ist in der Matrix wiederzufinden. Die entsprechenden Zellen der Matrix wurden der Übersichtlichkeit halber farbig hervorgehoben. Die Markierung in Zeile 2, Spalte 3 zeigt eine festgestellte semantische Verwandtschaft von T_2 zu T_1, die durch das Vorkommen der Schlüsselworte aus T_1 in T_2 begründet ist. Im umgekehrten Fall und Zeile 2, Spalte 3, fehlt diese Markierung und zeigt damit, dass Text T_1 keine semantische Verwandtschaft zu T_2 aufweist. Die gesetzten Markierungen in den weiteren Feldern der dargestellten Matrix sind analog zu interpretieren. Im Allgemeinen kann formuliert werden, dass eine Markierung in der entsprechende Zeile / Spalte bedeutet, dass der Text in dieser Spalte semantische Verwandtschaft zu dem in dieser Zeile abgetragenen Text besitzt.
Auf Basis einer in der beschriebenen Art aufgebauten Matrix können semantische Textempfehlungen gegeben werden. Wird einem Nutzer beispielsweise der Text T_5 aufgrund einer produktspezifischen Empfehlung angeboten oder empfindet der Nutzer diesen Text als hilfreich in seinem derzeitigen Arbeitskontext, so werden dem Nutzer laut Abbildung 46 die semantisch verwandten Texte T_1, T_2 und T_3 zusätzlich angeboten.
Dieser Teil der Empfehlungskomponente sowie die produktspezifischen Empfehlungen und die Assoziationsberechnung werden an einem Beispiel im nächsten Abschnitt im Detail erklärt.

6.3.2 Anwendungsbeispiel für die kontextbasierte Bereitstellung von Textdokumenten

Anhand mehrerer aufeinander aufbauender Darstellungen, wird in diesem Abschnitt ein Anwendungsbeispiel für die kontextbasierte Bereitstellung von Textdokumenten gezeigt. Dabei wird das in Abbildung 44 entwickelte Konzept schrittweise abgearbeitet. Da die Vorverarbeitung der Texte, die Produkterkennung sowie die Schlüsselwortextraktion bereits in den vorangegangenen Anwendungsbeispielen im Detail erläutert wurden, kann an dieser Stelle auf eine ausführlichere Darstellung verzichtet werden. Beginnend mit der Zuordnung der Texte zu den Produkten, bauen die im Folgenden durchgeführten Schritte auf den zuvor erzielten Ergebnissen auf.
Die folgende Abbildung 47 zeigt eine beispielhafte Zuordnung der aus Abbildung 46 bekannten Texte zu fünf Produkten (P1, P2,…, P5). Ein Verbindungspfeil zwischen Produkt und Text bedeutet, dass das entsprechende Produkt in dem zugehörigen Text als eines der Schlüsselworte identifiziert wurde. Das wiederum bedeutet, dass der Text mit dem Produkt in enger Verbindung steht. Mehrfachzuordnungen sind in diesem Zusammenhang zulässig. Beispielsweise ist Text 2 in Abbildung 47 gleichzeitig den Produkten P3, P4 und P5 zugeordnet, was in der Tatsache begründet liegt, dass alle drei genannten Produkte als Schlüsselworte dieses Textes erkannt wurden. Der umgekehrte Fall mit keiner Zuordnung zu einem oder mehreren Produkten stellt Text 6 dar. Dieser Text enthält keine Produktbezeichnungen in seiner Schlüsselwortliste und kann dementsprechend keinem Produkt zugeordnet werden.

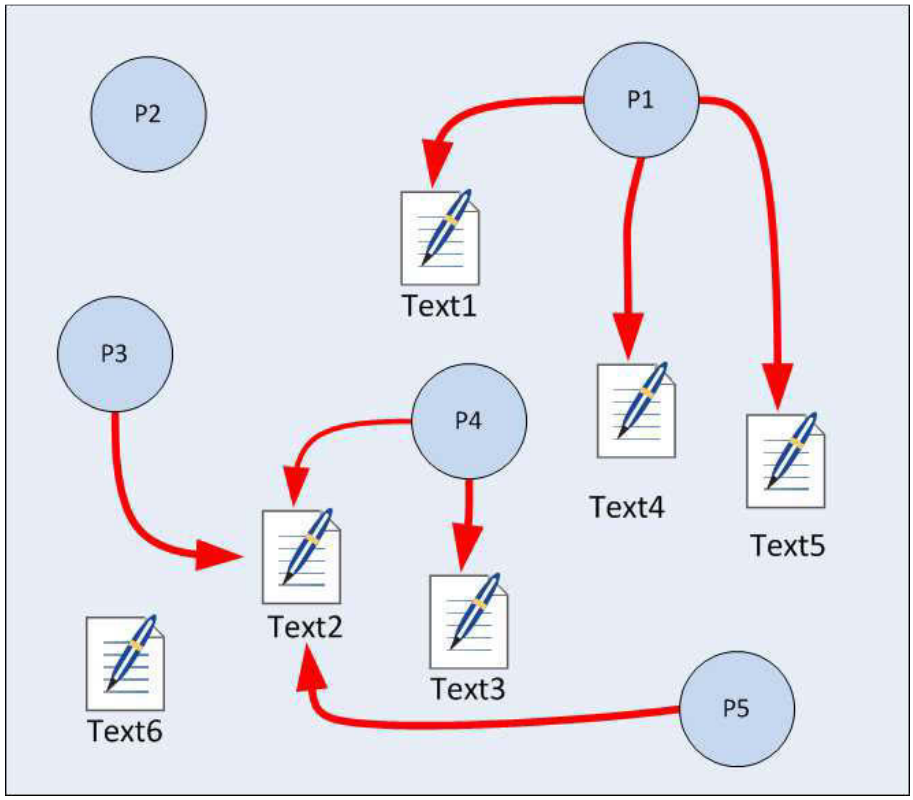


Abbildung 47. Zuordnung von Texten zu Produkten

Ohne Verbindungen zu konkreten Texten ist in obigem Beispiel Produkt P2. In keinem der sechs Beispieltexte wird Produkt P2 eine gesteigerte Bedeutung zugemessen. Mit Abschluss der Zuordnung von Texten zu Produkten ist ein weiterer Prozessschritt abgearbeitet. Bezogen auf die Texte ist das direkte Umfeld der Produkte definiert und die Beziehungen zwischen Texten und Produkten liegt offen. Zur Komplettierung des Kontext eines jeden Produktes fehlen die Verbindungen der Produkte untereinander. Um diese Produkt / Produkt-Beziehungen zu bestimmen, wird auf die Assoziationsberechnung mit CIMAWA zurückgegriffen.
Wie bereits angedeutet, wird die Assoziationsberechnung mittels CIMAWA in diesem Anwendungsfall dazu verwendet, aus den gesammelten Textdokumenten der Produktentwicklung eine sogenannte assoziative Produktstruktur zu generieren. Dieses CIMAWA-Anwendungsbeispiel funktioniert nach dem Prinzip des im letzten Kapitel entwickelten Association Mapping. Wenn auch prinzipiell ähnlich, können zwei Hauptunterschiede identifiziert werden. Zum einen fließen in die hier zu berechnenden Assoziationen ausschließlich Produktinstanzen ein, handelnde Personen, Tätigkeiten oder andere Instanzen finden keine Berücksichtigung. Zum anderen wurde beim Association Mapping im Instandhaltungsmanagement eine Association Map entwickelt, die von der Association Root ausgehend, die Assoziationen zu ausgewählten Elementen berechnet und entsprechend darstellt. Die Auswertungen der Association Map sind auf das definierte Zentrum der Assoziation ausgerichtet, da ausschließlich die vom Zentrum ausgehenden Assoziationen berechnet werden. Die assoziative Produktstruktur in Abbildung 48 hingegen beeinhaltet sämtliche berechenbaren Assoziationen zwischen den verfügbaren bzw. den zur Auswertung ausgewählten Produkten.

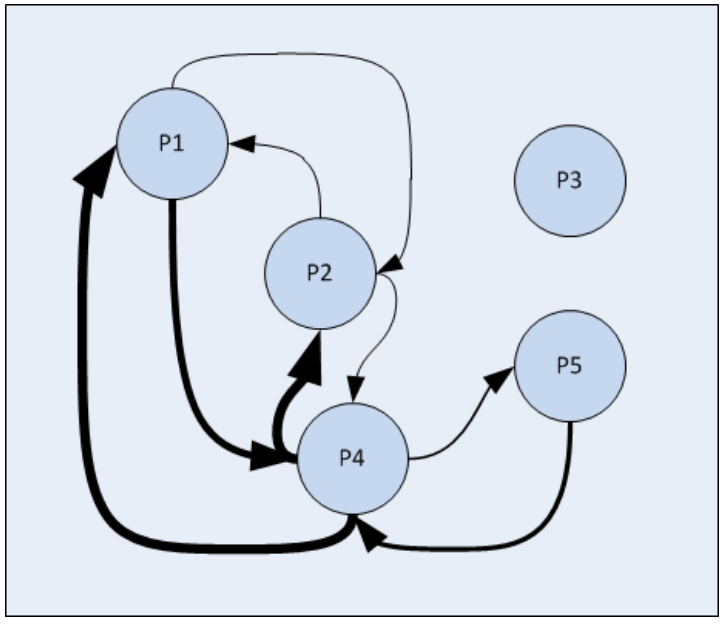


Abbildung 48. Assoziative Produktstruktur [134]

Die Produkte (P1, P2, …, P5) in Abbildung 48 sind in gewohnter Form durch Kreise visualisiert. Die Stärke der Assoziation zwischen den Produkten ist dargestellt durch die Verbindungspfeile. Je stärker die Pfeile, desto stärker die berechnete Assoziation.

Demzufolge sind diejenigen Produkte, die nicht durch Pfeile verbunden sind, auf Textebene nicht miteinander assoziiert. In diesen Fällen konnte mit der CIMAWA-Assoziationsberechnung keine Verbindung hergestellt werden. Dies ist genau dann der Fall, wenn die betreffenden Produkte, zum Beispiel P3 und P5 in Abbildung 48 kein gemeinsames Vorkommen in den Textdokumenten aufweisen.
Der asymmetrische Charakter der CIMAWA-Assoziationsberechnung wird am Beispiel der Produkte P2 und P4 deutlich. Während P2 stark mit P4 assoziiert ist, ist die Gegenrichtung deutlich weniger stark ausgeprägt. Die Assoziation zwischen den beiden Produkten differiert bezogen auf die unterschiedlichen Assoziationsrichtungen.
Wie den folgenden Darstellungen zu den Empfehlungskomponenten zu entnehmen ist, dienen die assoziative Produktstruktur in Verbindung mit der Zuordnung der Texte zu den Produkten als Inputdaten für die kontextbasierten Textempfehlungen für den Produktentwickler.

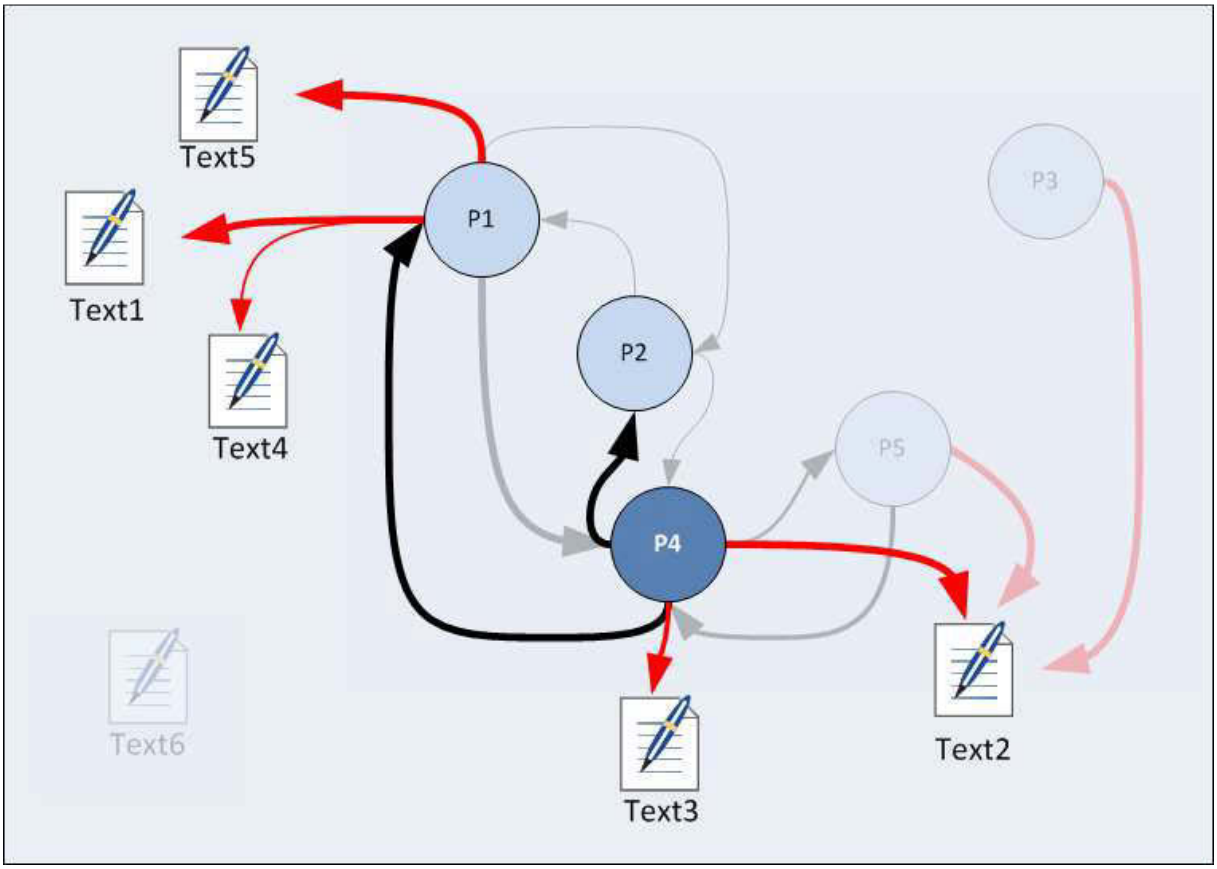


Abbildung 49. Darstellung des Produktkontext durch Zusammenführung der assoziativen Produktstruktur und der Textzuordnung [133]

Der in Abbildung 49 visualisierte Produktkontext verbindet die Produktstruktur mit der Textzuordnung und ermöglicht einen umfassenden Überblick der bis hierher erarbeiteten Beziehungen. Anhand der gegebenen Darstellung kann eine erklärende Beschreibung des ersten Teilbereichs der Empfehlungskomponente, der sogenannten produktbezogenen Textempfehlungen (Abbildung 44) gegeben werden. Im dargestellten Beispiel steht beispielhaft Produkt P4 im Fokus der Analyse, worauf die dunkelblaue Kennzeichnung

hindeutet. Alle Verbindungen, Produkte und Texte, die für dieses Produkt keine Relevanz besitzen, wurden in der Abbildung der Übersichtlichkeit halber ausgegraut.
Als Anwendungsfall wird die Entwicklung der nächsten Produktgeneration von Produkt P4 angenommen. Mit diesem Auftrag bedient der Produktentwickler das Textempfehlungssystem, um relevante textuelle Informationen aus dem direkten Produktumfeld zu erlangen.
Als unmittelbare Textempfehlungen können Text 2 und Text 3 angegeben werden (unterer Bereich Abbildung 49). Diese Texte besitzen durch die direkte Zuordnung den größten direkten Produktbezug. Die Menge der Textempfehlungen kann erweitert werden, indem die assoziative Produktstruktur analysiert wird. Zu diesem Zweck werden zunächst die am stärksten mit Produkt P4 assoziierten Produkte identifiziert. Im gegebenen Beispiel sind dies die Produkte P1 und P2. Auf dieser Basis wird die Menge der empfohlenen Texte um diejenigen erweitert, die direkten Bezug zu diesen Produkten P1 und P2 haben. Konkret bedeutet das die Hinzunahme von Text 1, Text 4 und Text 5 (oberer linker Bereich Abbildung 49), da diese direkte Verbindungen zu Produkt P1 aufweisen. Das Produkt P2 erweitert die Empfehlungsmenge der Texte nicht, da kein Text dieser Produktinstanz zugeordnet ist.
Aus der Text Mining-Perspektive heraus ist zu erwarten, dass die unmittelbar mit Produkt P4 in Verbindung stehenden Texte (Text 2, Text 3) eine höhere Relevanz zu P4 aufweisen als die über die assoziative Produktstruktur gefundenen Texte (Text 1, Text 4, Text 5). Aus diesem Grund ist eine Kategorisierung der Textempfehlungen umzusetzen, der die direkt produktbezogenen Texte höher bewertet als die übrigen. Dennoch können durch die Nutzung der berechneten Produktstruktur Informationen an den Produktentwickler weitergegeben werden, die diesen in seinem Aufgabenkontext unterstützen.
Wie Abbildung 44 zu entnehmen ist, wird in der Empfehlungskomponente zwischen produktbezogenen und semantischen Textempfehlungen unterschieden. Nachdem die produktbezogenen Empfehlungen auf Basis der assoziativen Produktstruktur erläutert wurden, werden im Folgenden die semantischen Textempfehlungen fokussiert.

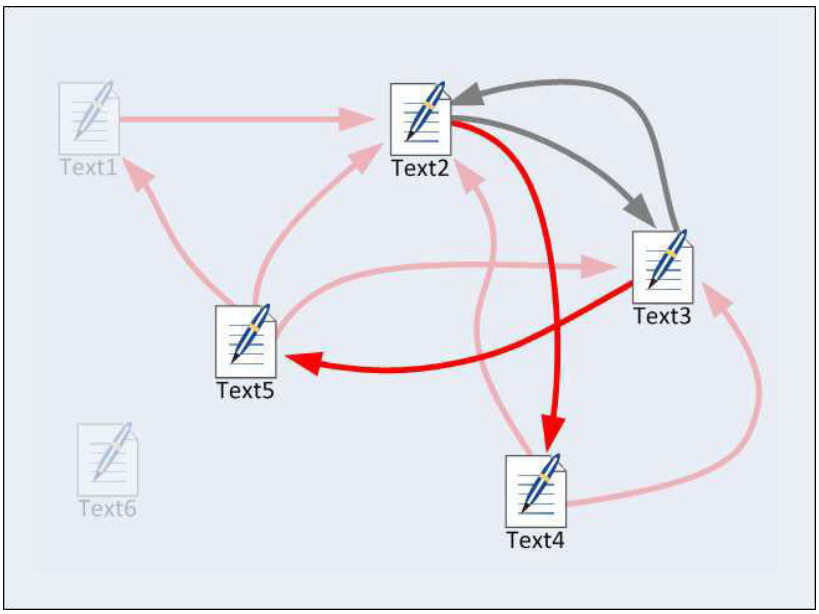


Abbildung 50. Semantische Textempfehlungen

Abbildung 50 visualisiert die semantischen Textverwandtschaften aus Abbildung 46. Die Pfeile stellen dabei die Verwandtschaft zwischen den Texten dar. Um ein zuvor bereits benutztes Beispiel aufzugreifen, soll die Beziehung zwischen Text 1 und Text 2 analysiert werden. Der Pfeil, ausgehend von Text 1 in Richtung Text 2 steht für die semantische Verwandtschaft von Text 2 zu Text 1. In umgekehrte Richtung gibt es keine Verbindung, da, laut Abbildung 46, Text 1 keine semantische Ähnlichkeit zu Text 2 aufweist. Alle in Abbildung 50 benutzten Pfeile sind in dieser Weise zu deuten und repräsentieren jeweils einen Eintrag in der Verwandtschaftsmatrix.
Das Anwendungsbeispiel in dem Produkt P4 der Analyse durch den Produktentwickler unterliegt, wird auch in Bezug auf die semantischen Textempfehlungen fortgesetzt. Die laut Abbildung 49 direkten Produktbezug besitzenden Texte 2 und 3 sind für die semantische Textempfehlung interessant und stehen im Zentrum der Analyse. Aufgrund der produktbezogenen Analyse der Texte kann davon ausgegangen werden, dass Text 2 und Text 3 aus dem direkten Umfeld von Produkt P4 im Aufgabenkontext des Produktentwicklers gesteigerte Relevanz besitzen. Darum wird auf Basis dieser Texte eine Textempfehlung erstellt, die dem Produktentwickler semantisch ähnliche Texte bereitstellt. Die von diesen Texten ausgehenden Verwandtschaften, welche den Pfeilen in Abbildung 50 entsprechen, sind hervorgehoben, während die übrigen, weniger relevanten Verbindungen und Texte, ausgegraut sind. Die Analyse zeigt, dass semantisch verwandte Texte zu Text 2 in Text 4 und Text 3 zu finden sind. Da Text 3 im Anwendungsbeispiel bereits Teil der Empfehlungsmenge ist, ist diese Verbindung zu vernachlässigen. Als semantisch ähnlicher Text bleibt Text 4, der entsprechend in die Empfehlungsmenge aufgenommen werden kann. In Text 5 findet sich ein ähnlicher Text zu Text 3, was diesen ebenfalls zum Element der Textempfehlungsmenge werden lässt.
Auswahlkriterium für das Auffinden von semantisch ähnlichen Texten waren in diesem Beispiel die produktbezogenen Textempfehlungen. Denkbar sind jedoch auch andere Kriterien, wie zum Beispiel die Empfehlung von ähnlichen Texten zu denen, die aktuell vom Produktentwickler gelesen werden. In diesem Fall würden die Empfehlungen in Echtzeit identifiziert und angeboten.

6.4 Verwendung von CIMAWA zur Entwicklung eines assoziativen Suchverfahrens auf Textbasis

Die in diesem Unterkapitel präsentierten Ergebnisse basieren auf der vom Autor mitbetreuten Diplomarbeit mit dem Titel „Konzeptionierung und prototypische Realisierung eines intelligenten Suchverfahrens auf Textbasis als ein intranet basierter Ansatz bei Mubea“ [139]. Oben genannte Diplomarbeit entstand in Kooperation mit dem Unternehmen Muhr und Bender (im Folgenden Mubea genannt). Mubea ist ein Unternehmen der Automotive Branche mit weltweit verteilten Produktions- und Entwicklungsstandorten mit Hauptsitz in Attendorn. Zu den Kunden zählen namhafte nationale und internationale Automobilkonzerne wie Audi, BMW, Daimler oder Toyota [140]. Das Unternehmen ist spezialisiert auf die Entwicklung und den Vertrieb von Produkten in Leichtbauweise, wobei Karosserieteile, Tellerfedern, Stabilisatoren und Achsfedern den wichtigsten Teil des Produktportfolio bilden [139].
Um ein Suchverfahren auf Textbasis realisieren zu können, war zunächst eine Analyse der Ist-Situation im Unternehmen durchzuführen. Hierbei steht anwendungsbedingt der Umgang und die Speicherung der unternehmenseigenen Textdokumente im Vordergrund. Im vorgefundenen Fall werden Dokumente wie Normen, Kundenvorgaben oder Arbeitsanweisungen zentral im firmeneigenen Intranet gespeichert, wobei eine Einteilung der Dokumente in allgemein gehaltene Oberkategorien festgestellt wurde [139]. Um Textdokumente für ihren Aufgabenkontext zu identifizieren, steht den Mitarbeitern eine Volltextsuche auf Basis von Einzeltexten zur Verfügung, welche voraussetzt, dass das relevante Dokument identifiziert und eine Kopie heruntergeladen wurde. Eine Suche nach Schlagworten auf der gesamten Textbasis oder auf definierten Teilbereichen ist derzeit nicht realisiert.
Das Auffinden von kontextspezifischen Dokumenten gestaltet sich Mitarbeiterberichten zufolge mitunter schwierig, da die einfachen ordnerähnlichen Strukturen das Sichten großer Textmengen vom Suchenden verlangt. Darüber hinaus werden die Textdateien im unternehmenseigenen Intranet nicht mit inhaltsorientierten Titeln versehen, sondern bestehen aus kryptischen eher codeähnlichen Bezeichnungen. Diese Art der Dokumentenablage macht es speziell Mitarbeitern mit wenig Erfahrung bei der intranetbasierten Informationssuche schwer, relevante Dokumente für anstehende Aufgaben zu identifizieren. Aber selbst diejenigen Mitarbeiter, die ständig Textdokumente aus dem Intranet verwenden, können aufgrund der vorhandenen und laufend wachsenden Menge an Dokumenten keinen ganzheitlichen Überblick über die gesammelte unternehmensintere Textbasis wahren. Dies erweist sich unter mehreren Gesichtspunkten kritisch.
Zum einen gibt es für die Mitarbeiter bislang keine Alternative zur zeitintensiven manuellen Textrecherche, welche sich im unternehmerischen Alltag zu einem echten Kostentreiber entwickeln kann. Zum anderen sei an dieser Stelle auf den Aspekt der textgebundenen unternehmerischen Wissensbasis hingewiesen. Die gespeicherten Texte bilden einen Teil dieser Wissensbasis, da Wissen von internen oder externen Experten in Form von Handlungsbeschreibungen, Kommunikationsprotokollen, Arbeitsanweisungen oder Normen

kodifiziert ist. Ohne den Mitarbeitern einen unkomplizierten und schnellen Zugang zu diesen Informationsquellen zu ermöglichen, bleiben im Unternehmen vorhandene Wissenspotentiale ungenutzt. Unter Umständen werden Lösungen für Problemstellungen entwickelt, die an anderer Stelle im Betrieb bereits erarbeitet wurden, was unabhängig von Größe und Relevanz der zu erledigenden Aufgabe einen unerwünschten Umstand darstellt. Vor diesem Hintergrund erscheint die Entwicklung eines unternehmensinternen Suchverfahrens auf Textbasis als geeigneter Ansatz zur Zeit- und Kostenminimierung für ein breites Anwenderspektrum.
Die geschilderten Gegebenheiten vor Ort und die identifizierten Verbesserungspotentiale im Bereich der Erkennung und Nutzung relevanter Textdokumente bildeten die Motivation zur Entwicklung eines Prototypen, der auf gewachsenen Textsammlungen eingesetzt werden kann. Das Konzept hinter dem entwickelten Ansatz, wie CIMAWA als Assoziationsberechnungsmethode eingesetzt wird, sowie die Unterschiede zu herkömmlichen Suchverfahren, werden in den nächsten Abschnitten geschildert. Zusätzlich wird der Prototyp selbst und zugehörige Anwendungsbeispiele thematisiert und anhand von Screenshots erläutert.

6.4.1 Konzeption des assoziativen Suchverfahrens auf Basis von CIMAWA

Um die zugrunde liegende Idee hinter dem entwickelten Suchverfahren zu verdeutlichen, wird im Folgenden auf das Konzept der assoziativen Suche eingegangen. Hierbei wird insbesondere die Rolle der Assoziationsberechnung durch CIMAWA und der Nutzen der so berechneten Ergebnisse erklärt. Abbildung 51 visualisiert das Konzept des realisierten Suchverfahrens auf Basis von CIMAWA. Die einzelnen Prozessschritte werden im Folgenden detailiert beschrieben.
Wie die vorher beschriebenen Anwendungsbeispiele von CIMAWA, bildet auch in diesem Fall ein Grundstock an Texten die Basis für den entwickelten Ansatz. Wie bereits zuvor erwähnt, besteht diese spezielle Textsammlung aus Texten der unternehmensinternen Wissensbasis, die im Intranet des Unternehmens abgelegt sind. Diese Texte bilden die Grundlage für die Assoziationsberechnung und bilden zugleich die zu durchsuchende Textsammlung. Die Texte werden in einem ersten Schritt soweit vorverarbeitet, wie es die Anwendung verlangt. Hierzu gehört zunächst die Extraktion der Texte aus abgelegten PDF-Dokumenten und deren Speicherung in der Datenbank. Die darauf aufbauende ‘tokenization‘ stellt sicher, dass im weiteren Verlauf präzise statistische Daten zur Assoziationsberechnung erhoben werden können. Diese Vorverarbeitung geschieht im Hintergrund und verlangt keine Interaktion mit dem Benutzer des Suchsystems. Der Nutzer tritt erst im nächsten Prozessschritt in Erscheinung. Die erste Interaktion mit dem System stellt die Eingabe der Anfrage in Form eines oder mehrerer Suchbegriffe in die Suchmaske dar. Diese Anfrage oder genauer gesagt die eingegebenen Begriffe, werden vom System in zweierlei Hinsicht weiterverarbeitet.
Zum einen wird die gesamte eingelesene Textbasis im Rahmen einer Volltextsuche nach den eingegebenen Begriffen durchsucht. Die so gewonnenen Ergebnisse, in Form, der die Begriffe enthaltenden Textpassagen, werden dem Benutzer als Suchergebnisse der Kategorie 1 sofort

angeboten. Zu erkennen sind diese Ergebnisse im unteren Bereich von Abbildung 51 unter ‘Suchergebnisse Kategorie 1‘. Diese Ergebnisse stellen die direkten Treffer des eingegebenen Begriffs dar und bieten somit, abgesehen von der Einbeziehung der gesammten Textbasis an Stelle nur einzelner Texte, keinen über die Basisfunktionen hinausgehenden Mehrwert.
Zum anderen werden die vom Benutzer eingegebenen Suchbegriffe auf ihre Assoziationen in der eingelesenen Textbasis analysiert. Diese Assoziationen werden mit dem CIMAWA-Ansatz berechnet. Das Suchverfahren verwendet CIMAWA in der in Kapitel 4.2.1 erläuterten Form, wobei zwei unterschiedliche Varianten der CIMAWA-Assoziationsberechnung verwendet wurden. Diese unterscheiden sich bezüglich der Textfenstergröße der Kookkurrenzberechnung.

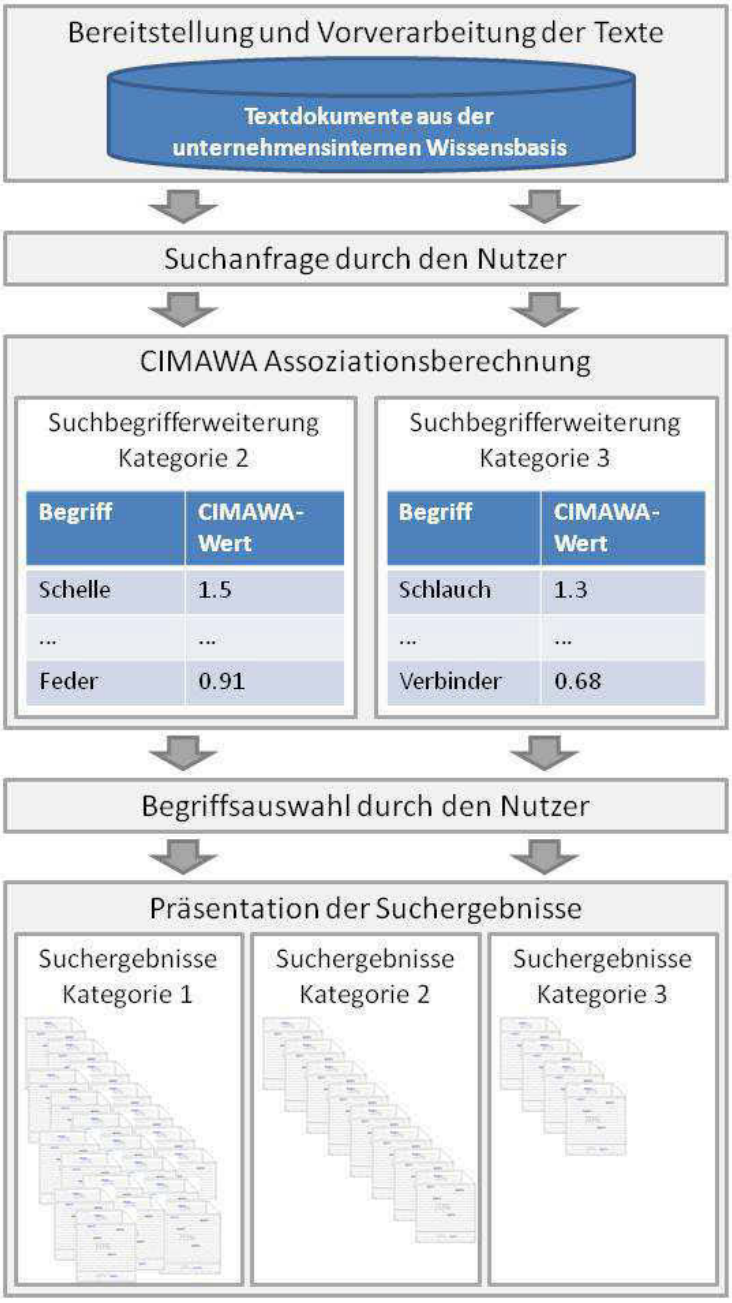


Abbildung 51. Konzept des Suchverfahrens auf Basis von CIMAWA

Das präsentierte Suchverfahren empfiehlt Textpassagen bzw. ganze Absätze aus den eingelesenen Dokumenten. Aus diesem Grund wurden die Kookkurrenzen und demzufolge auch die Assoziationen auf Satzbasis und nicht wie in vorigen Beispielen (siehe Kapitel 6.1) auf Wortbasis bestimmt. Durchgeführte Tests im Rahmen von [139] zeigen, dass aus der Kookkurrenzberechnung mit 7-Satz bzw. 15-Satz Textfenster die besten Ergebnisse resultieren. Aus diesem Grund wurden diese Textfenstergrößen im Prototyp umgesetzt. Die Resultate der CIMAWA-Berechnung im 7-Statz-Radius ist dargestellt in ‘Suchbegrifferweiterung Kategorie 2‘ im mittleren Teil der Abbildung 51. Die Assoziationsergebnisse der größeren Textfenstergröße im 15-Satz-Radius sind in der ‘Suchbegrifferweiterung Kategorie 3‘ visualisiert. Die jeweiligen Tabellen beinhalten die am stärksten mit dem Suchbegriff assoziierten Begriffe im entsprechenden Textfenster. Die Begriffe sind absteigend sortiert nach CIMAWA-Wert. Diese Wortlisten mit den am stärksten mit dem Suchbegriff assoziierten Begriffen stellen Vorschläge zur Suchbegrifferweiterung dar, aus denen der Nutzer auswählen kann. Zu beachten ist an dieser Stelle, dass die angebotenen Begriffe in Kategorie 2 und 3 diejenigen Begriffe sind, die nachgewiesen in der unternehmensinternen Wissensbasis verankert sind und im Kontext des Suchbegriffes vorkommen. Diese Begriffe beinhalten in Teilen solche, die außerhalb des Unternehmens unbekannt sind, da es sich um interne Bezeichnungen, Produktnamen oder Tätigkeitsbeschreibungen handelt, die sich im Laufe der Zeit innerhalb des Unternehmens entwickelt haben. Jeder der angebotenen Begriffe kann vom Nutzer ausgewählt werden, um seine Suchanfrage zu spezifizieren. Da die Assoziationsberechnung auf der Grundlage der unternehmensinternen Wissensbasis durchgeführt wurde, die gleichzeitig die zu durchsuchende Textsammlung darstellt, ist sichergestellt, dass jeder Begriff in den Listen zugleich in ein oder mehreren Textstellen der Wissensbasis vorkommt. Das bedeutet für den Nutzer einen echten Mehrwert, da hinter jeder angebotenen Suchbegrifferweiterung, Substanz in Form von Textpassagen steckt.

Der Nutzer wählt je nach aktuellem Aufgabenkontext einen Begriff aus den angebotenen Listen aus. Dieser Ansatz ist speziell dann vielversprechend, wenn es sich bei der ursprünglichen Suchanfrage um einen hochfrequenten Begriff mit zahlreichen direkten Suchtreffern in der Textbasis handelt. Ohne Spezifikation durch einen den derzeitigen Aufgabenkontext beschreibenden Erweiterungsbegriff, müssten die gesamten direkten Treffer manuell durch den Nutzer gesichtet werden. Folglich verringert die Auswahl die Menge der direkten Suchtreffer aus Kategorie 1, auf die entsprechende Teilmenge an Textpassagen mit dem ausgewählten Erweiterungsbegriff. Die Ergebnisse werden je nach Herkunft des ausgewählten Begriffs in Kategorie 2 oder Kategorie 3 im unteren Bereich der Abbildung 51 angezeigt.

An einem Beispiel aus Abbildung 51 soll der Ablauf verdeutlicht werden. Angenommen der Nutzer wählt zusätzlich zu seinem ursprünglich angegebenen Suchwort das an Position 1 platzierte Wort ‘Schelle‘ der Kategorie 2. In diesem Fall werden bezogen auf die Präsentation der Ergebnisse in Kategorie 1 alle Textpassagen mit dem ursprünglichen Suchwort angezeigt. Die Suchergebnisse der Kategorie 2 reduzieren diese Menge an Textpassagen auf genau diejenigen, in denen das ursprüngliche Suchwort gemeinsam mit der Erweiterung ‘Schelle‘ in einem 7-Satz-Radius vorkommt.

6.4.2 Prototyp der assoziativen Suche in der unternehmensinternen Wissensbasis

Nachdem die Konzeption des Ansatzes vorgestellt wurde, soll in diesem Abschnitt der entwickelte Prototyp im Detail vorgestellt werden. Um eine praxisnahe Darstellung zu gewährleisten, wird das System anhand von Screenshots und entsprechenden Suchanfragen vorgestellt.
Die folgende Abbildung 52 zeigt einen Screenshot der Anwendung im Ausgangszustand. Das bedeutet, es sind noch keine Dokumente eingelesen und dementsprechend auch keine Suchanfragen an das System möglich. Im linken Bereich der Anwendung kann der Benutzer unter ‘Assoziationssuche: Dokumenteneingabe‘ neue Dokumente in das System einlesen. Darunter ist alternativ in ‘Assoziationssuche: Suchbereich‘ die Suchmaske zu öffnen. Abbildung 52 fokusiert die Dokumenteneingabe, weshalb der Suchbereich minimiert und nicht zu sehen ist.

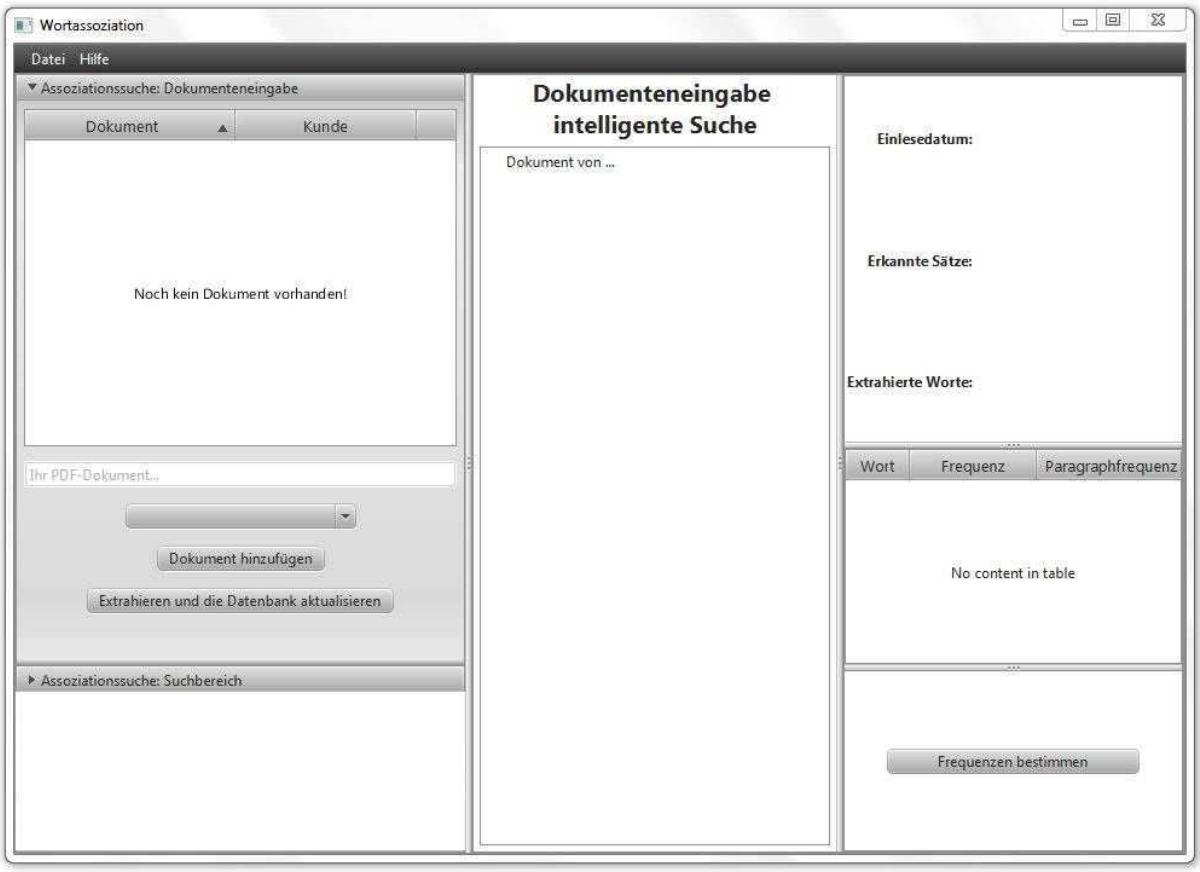


Abbildung 52. Prototyp intelligente Suche: Dokumenteneingabe

Da in obiger Abbildung noch keine Dokumente eingelesen wurden, bleibt auch der gesamte rechte Teil der Anwendung inhaltlos. Die folgenden Detailansichten zeigen, wie die Dokumenteneingabe auf der einen und die Eigenschaften der eingelesenen Dokumente auf der anderen Seite im Prototyp integriert wurden und wie der Nutzer des Systems diese Funktionen zur Anwendung bringt.
Zunächst soll gezeigt werden, wie der Nutzer neue Dokumente ins System einspeist. Dazu wird in Abbildung 53 die Dokumenteneingabe vergrößert dargestellt. Das Beispiel zeigt das

System in einem Zustand, in dem bereits zehn Dokumente ausgewählt wurden. Diese werden mit Dateipfad und dem zugehörigen Kunden im oberen Bereich listenartig aufgeführt. Direkt darunter im blau umrandeten Bereich hat der Nutzer die Möglichkeit, den Pfad eines neuen PDF-Dokumentes anzugeben und zusätzlich in einem drop-down-Menü einen entsprechenden Kunden zuzuordnen. Im dargestellten Fall handelt es sich um eine Norm des Automobilherstellers BMW, die für die Einfügeoperation ausgewählt wurde. Durch die Betätigung der Schaltfläche ‘Dokument hinzufügen‘ wird die angegebene Datei zur obigen Liste hinzugefügt. Sobald der Nutzer keine zusätzlichen Dokumente mehr einfügen möchte, kann er sämtliche gelisteten Dokumente durch die Betätigung der Schaltfläche ‘Extrahieren und die Datenbank aktualisieren‘ in der Datenbank ablegen. Daraufhin extrahiert das System die Texte aus den PDF-Dokumenten, verarbeitet diese vor, und legt sie in der Datenbank ab. Im Anschluss stehen die Inhalte dieser Dokumente für die Suchanwendung zur Verfügung.

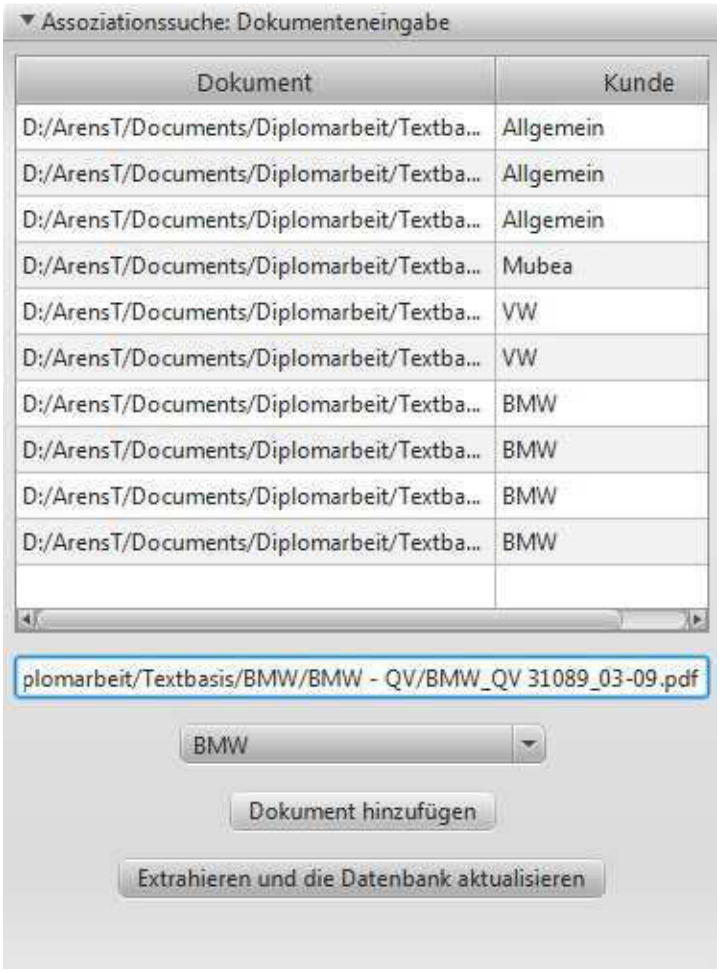


Abbildung 53. Detailansicht der Dokumenteneingabe [139]

Darüber hinaus ist es dem Nutzer zusätzlich möglich, Informationen über die eingelesenen Texte abzurufen. Abbildung 54 zeigt exemplarisch statistische Auswertungen zu einem eingelesenen Textdokument. Zu unterscheiden sind vier Bereiche, die durch die entsprechenden Ziffern gekennzeichnet sind und im Folgenden näher beschrieben werden. Im

Bereich 1 kann der Nutzer alle Daten einsehen, zu denen ein oder mehrere Dokumente eingelesen wurden. Für den Fall, dass mehrere Dokumente an ein und demselben Tag hinzugefügt wurden, ist für jedes Dokument zusätzlich die Uhrzeit der Datenbankablage hinterlegt. Der dargestellte Fall zeigt, dass am 23.05.2013 eine große Anzahl an Dokumenten zur gleichen Zeit eingelesen wurden. Hier steht jede Uhrzeit für ein spezifisches Textdokument. Der Nutzer wählt ein Textdokument aus und betätigt die Schaltfläche ‘Frequenzen bestimmen‘ im Bereich 4 der Abbildung 54. Die daraufhin ermittelten Werte werden sodann in den Bereichen 2 und 3 dargestellt. Bereich 2 gibt an, wann das Dokument eingelesen wurde, wie viele Sätze es enthält und wie viele Wörter es beinhaltet. Im Bereich 3 wird dem Nutzer eine Wortliste angeboten, die absteigend nach der Wortfrequenz im ausgewählten Dokument sortiert ist.

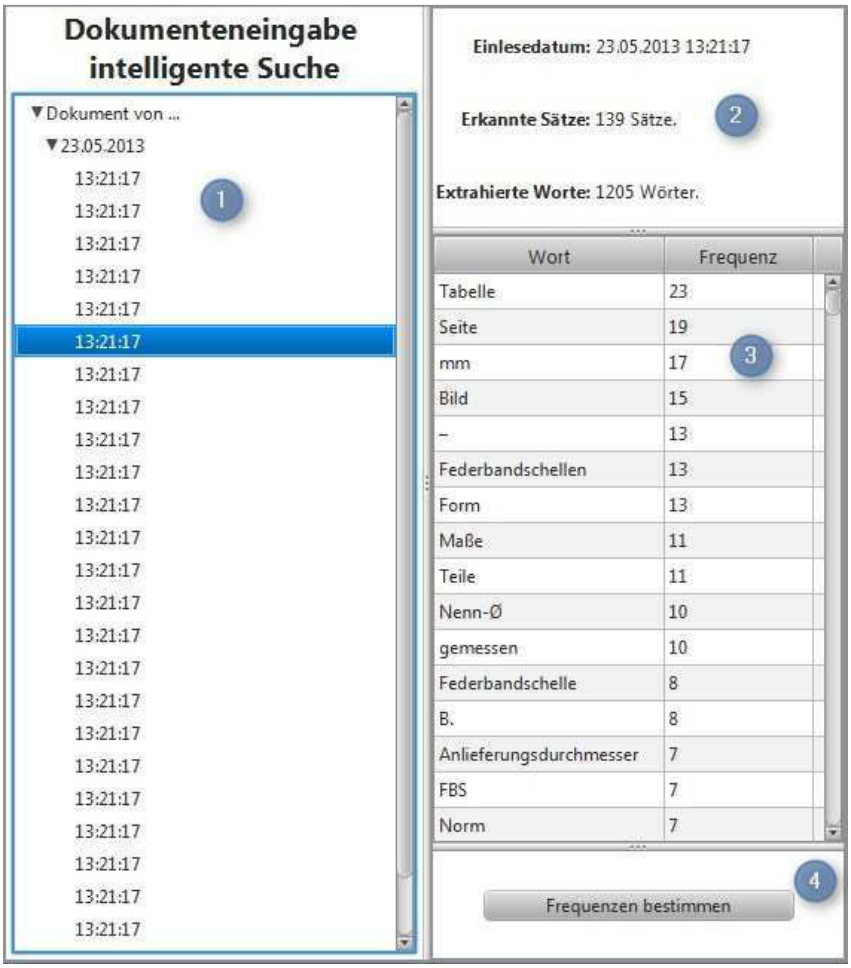


Abbildung 54. Detailinformationen zu den eingelesenen Texten [139]

Nachdem die Funktionsweise der Dokumenteneingabe und die damit verbundenen statistischen Auswertungen aufgezeigt wurden, kann sich der Hauptkomponente gewidmet werden. Hierbei handelt es sich um eine Suchanwendung, die die zuvor eingelesenen Dokumente einbezieht und darauf abzielt, die Recherchezeiten für Informationen zu reduzieren und die im Unternehmen vorhandenen Dokumente kontextbasiert zur Verfügung zu stellen. Damit kann ein Beitrag zur Wiederverwendung vorhandener Informationsobjekte

innerhalb des Unternehmens geleistet werden, da den Mitarbeitern die Möglichkeit zur aufgabenorientierten Suche in großen Textmengen ermöglicht wird.
Abbildung 55 zeigt einen Screenshot des Prototyps mit dem Suchfenster für die Rechercheanfrage im oberen linken Bereich der Abbildung. Zunächst ist der Nutzer aufgefordert eine Suchanfrage an das System zu stellen. Hierfür gibt er seinen Suchbegriff in das Dialogfeld oben links ein. Mit Betätigung der darunter angesiedelten Schaltfläche ‘Suchen‘ werden zugleich zwei Prozesse angestoßen. Zum einen wird eine Volltextsuche mit der Suchanfrage auf den eingelesenen Dokumenten initiiert. Die Ergebnisse dieser Suche werden als sogenannte direkte Treffer in der Ergebnisskategorie 1 im mittleren Bereich der Abbildung 55 angezeigt. Zum anderen wird gleichzeitig mit dem Start der Volltextsuche die Assoziationsberechnung ausgeführt. Wie bereits zuvor diskutiert, wird hierzu der CIMAWA-Ansatz in zwei verschiedenen Textfenstergrößen angewendet. Die Ergebnisse der 7-Satz-Assoziationsberechnung werden im Prototyp als zusätzliche Suchbegriffe der Kategorie 2 bezeichnet und mitsamt dem CIMAWA-Wert in einer absteigend sortierten Liste angeboten. Das gleiche Vorgehen bezüglich der Assoziationsberechnung verbirgt sich hinter den zusätzlichen Suchbegriffen der Kategorie 2, mit dem einzigen Unterschied der größeren Textfenstergröße von 15 Sätzen um das Suchwort. Beide Ergebnisfelder sind auf der linken Seite der Abbildung 55 direkt unterhalb des Dialogfeldes für die Suchanfrage lokalisiert.
In der rechten Hälfte der Abbildung sind die Felder für die Suchergebnisse zu erkennen. Die Laufzeit der Berechnung der Suchergebnisse sowie die Berechnung der Assoziationen zu dieser Suchanfrage hängen vom gewählten Suchwort ab. Diejenigen Suchworte, die relativ häufig im eingelesenen Textkorpus vorkommen, verursachen längere Such- und Assoziationsberechnungszeiten als solche, mit geringerer Frequenz in der Textsammlung.

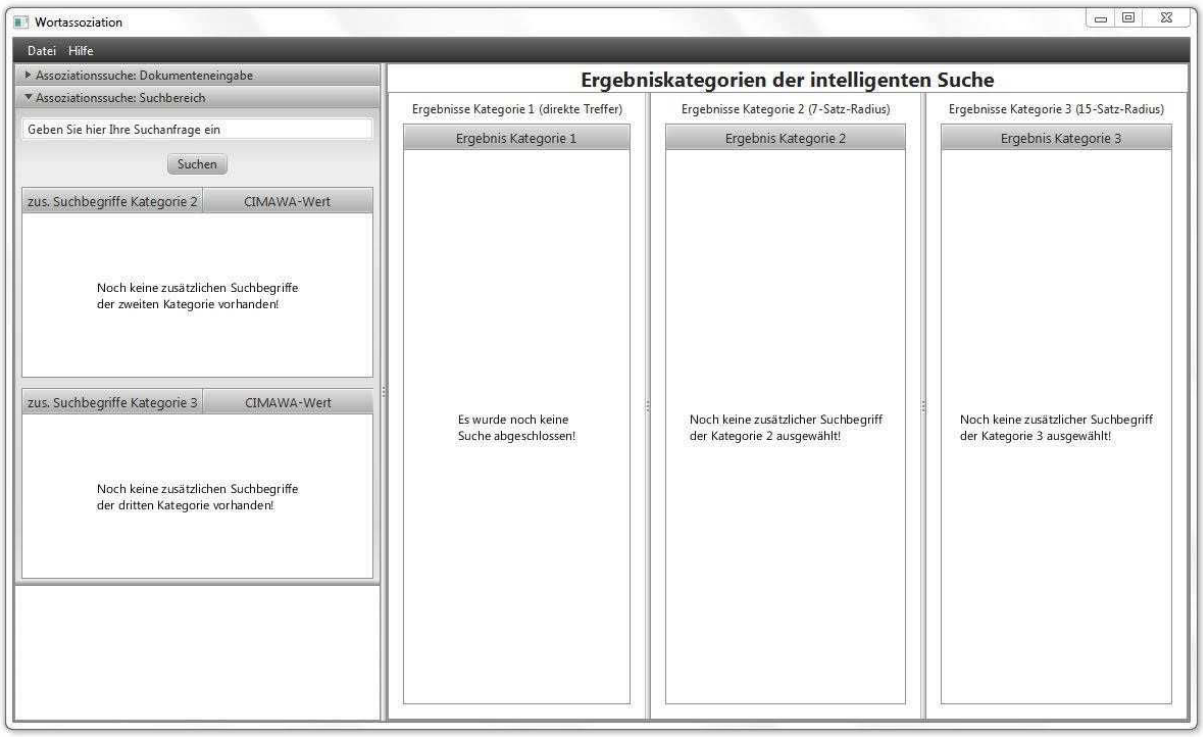


Abbildung 55. Prototyp intelligente Suche: Suchanfrage und Ergebnisdarstellung

Neben dem bereits beschriebenen Ergebnisfeld der Kategorie 1, in der die Suchresultate der Volltextsuche angeboten werden, gibt es zwei weitere Ergebnisfelder, in denen die Ergebnisse der Assoziationssuche dargestellt werden. Diese werden mit Inhalt gefüllt, sobald der Nutzer zusätzliche Begriffe aus den Assoziationslisten ausgewählt hat. Angezeigt werden sodann die Textpassagen, in denen die ursprüngliche Suchanfrage gemeinsam mit den ausgewählten Begriffen aus der Assoziationsliste vorkommen. Der genaue Ablauf einer solchen Suchanfrage wird exemplarisch in Abbildung 56 und Abbildung 57 dargestellt.
Abbildung 56 zeigt in dem mit 1 gekennzeichneten Bereich eine Suchanfrage des Nutzers an das System. Im dargestellten Fall wurde der Begriff ‘Schlauchschelle‘ in das Dialogfeld eingetragen und die Schaltfläche ‘Suchen‘ betätigt. Die in Abbildung 56 im Bereich 2 eingetragenen Ergebnisse entsprechen den von CIMAWA berechneten Assoziationen zum Begriff ‘Schlauchschelle‘. Um diese zu ermitteln, wurden sämtliche eingelesenen Texte nach diesem Begriff durchsucht. Für die Kategorie 2 wurden die Textbereiche in einem 7-Satz-Radius um den Suchbegriff analysiert, die vorkommenden Wörter identifiziert und auf dieser Basis die am meisten assoziierten Worte für diesen Radius ermittelt.

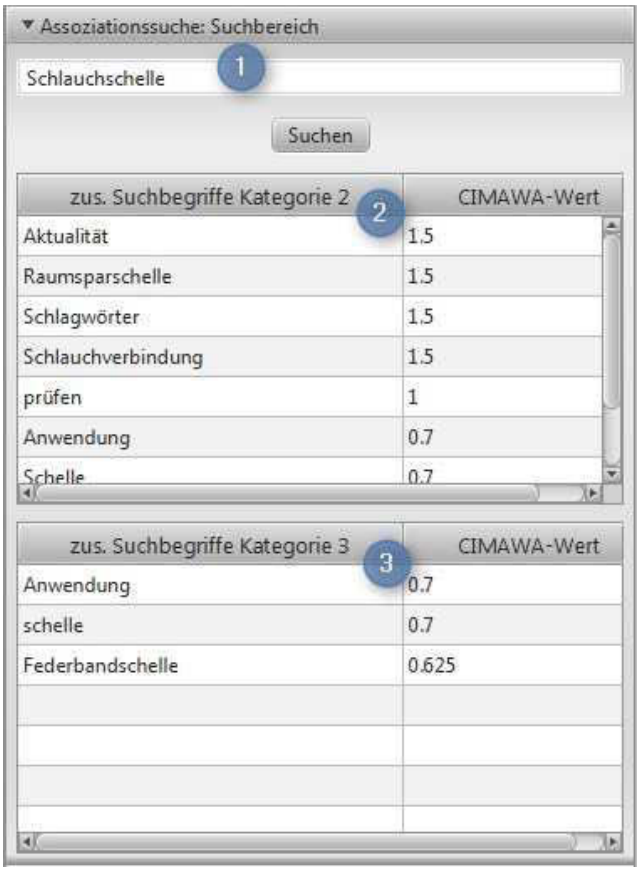


Abbildung 56. Detailansicht der Suchmaske [139]

Ebenso wurden die assoziierten Begriffe für die Kategorie 3 ermittelt. Im Unterschied zu der Kategorie 2 wurde hier allerdings ein Radius von 15 Sätzen um das Suchwort analysiert. Da beide Suchbereiche zum Teil überlappen, was bedeutet, dass Begriffe, die im 7-Satz-Radius vorkommen, naturgemäß zugleich in dem 15-Satz-Radius auftreten, werden die Ergebnisse der zweiten Kategorie nicht erneut in die Kategorie 3 übernommen, es sei denn, diese treten

zusätzlich in dem vergrößerten Suchradius auf. Der Nutzer hat an dieser Stelle die Möglichkeit, mithilfe der angebotenen Begriffe aus beiden Kategorien seine Suchanfrage zu präzisieren. Durch Auswahl eines der Begriffe werden in den entsprechenden Ergebniskategorien die Textstellen angezeigt, in denen der ursprüngliche Suchbegriff gemeinsam mit dem ausgewählten assoziierten Begriff vorkommt.
Diese Textstellenempfehlungen können dem Nutzer zeitintensive Recherchearbeit abnehmen, da er unmittelbar zu der Textstelle geleitet wird, die seinem Aufgabenkontext entspricht. Die manuelle Suche nach entsprechenden Passagen langer Textdokumente kann somit verkürzt und die Effizienz der Informationssuche gesteigert werden. Ein weiterer Vorteil des Prototypen zeigt sich darin, dass der Nutzer nicht wie bisher zunächst einzelne Dokumente aus dem Intranet herunterladen muss, um diese zu durchsuchen, sondern dass ganze Sammlungen von Textdokumenten zeitgleich mit einer Anfrage in die Suche einbezogen werden. Nachfolgende Abbildung 57 zeigt exemplarisch das Ergebnis einer Suchanfrage mit zusätzlich ausgewählten assoziierten Begriffen.

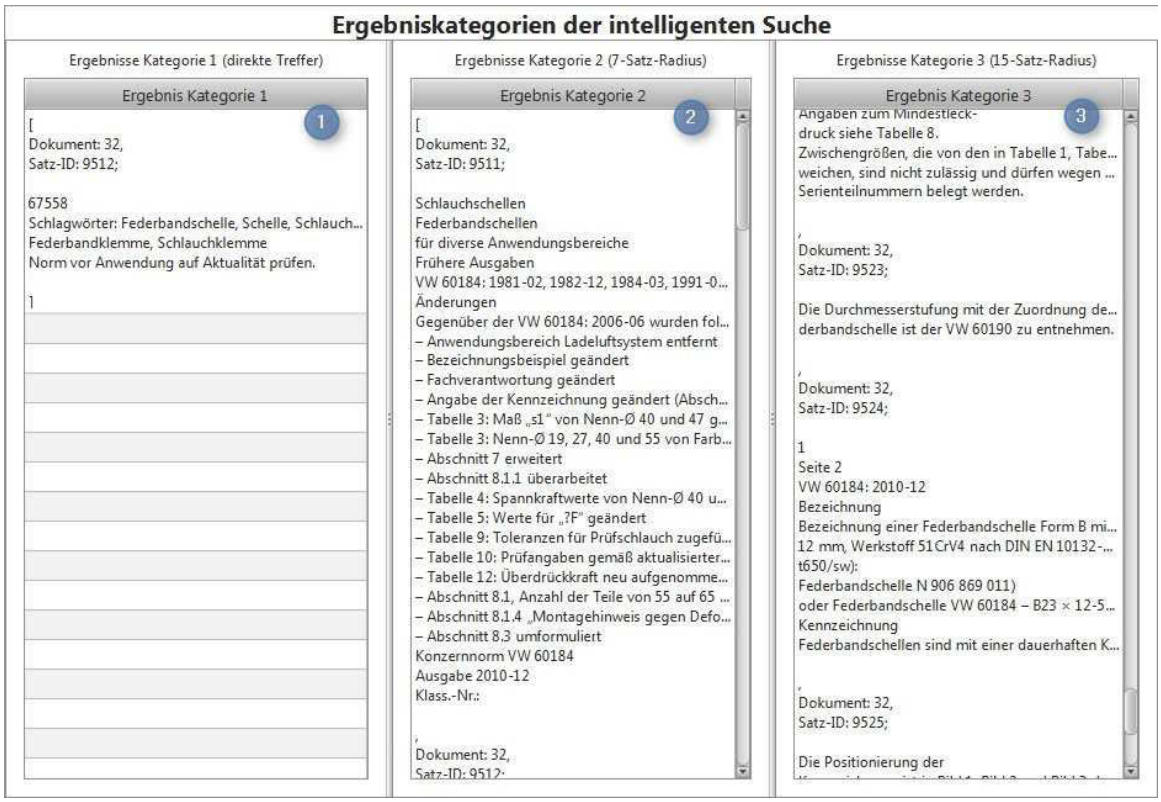


Abbildung 57. Detailansicht der Ergebniskategorien [139]

Die angebotenen Resultate sind nach den beschriebenen Kategorien geordnet. Kategorie 1 im Bereich 1 der Abbildung 57 zeigt die direkten Treffer der Suchanfrage. Im Bereich 2 und 3 sind die Ergebnisse einer präzisierten Recherche dargestellt. Diese umfassen die identifizierten Textstellen in den entsprechenden Radien um den Suchbegriff.

6.4.3 Vergleich des entwickelten Suchverfahrens mit dem Marktführer

Seit Entstehung des Internet, hat sich dieses zu einem Medium entwickelt, welches seinen Nutzern zuvor unerreichbare Ressourcen an Infomationen überall und jederzeit zugänglich macht. Aufgrund der Masse an Informationen und der Anzahl an Webseiten entwickelten sich die bekannten Internetsuchmaschinen, wobei der Marktführer ‘google‘ derzeit den mit Abstand größten Marktanteil einnimmt. Da es sich auch bei dem entwickelten Ansatz zur assoziativen Suche um eine Art Suchmaschine handelt, sollen im Folgenden die konzeptuellen Unterschiede zu google herausgearbeitet werden.
Genau genommen handelt es sich bei der vorgestellten Assoziationssuche um eine Art Erweiterung des Suchbegriffes oder eine spezielle Art der Autovervollständigung, da Suchbegrifferweiterungen zum ursprünglichen Suchbegriff angeboten werden. Aus diesem Grund wird das entwickelte Verfahren mit der von google angebotenen Autovervollständigung verglichen. Dafür wird zunächst dargestellt, wie die google-Autovervollständigung konzeptionell aufgebaut ist und welchen Regeln sie unterliegt. Es folgt ein Vergleich mit der Eigenentwicklung sowie exemplarische Suchanfragen an beide Systeme.
Zunächst soll ein Blick auf den von google angebotenen Dienst geworfen werden. Google selbst beschreibt die Grundzüge der automatischen Vervollständigung in [141]. Nach Eingabe eines Begriffs oder eines Teilbegriffs bietet der Suchmaschinenbetreiber in Echtzeit vier Suchbegrifferweiterungen an, aus denen der Benutzer optional auswählen kann.

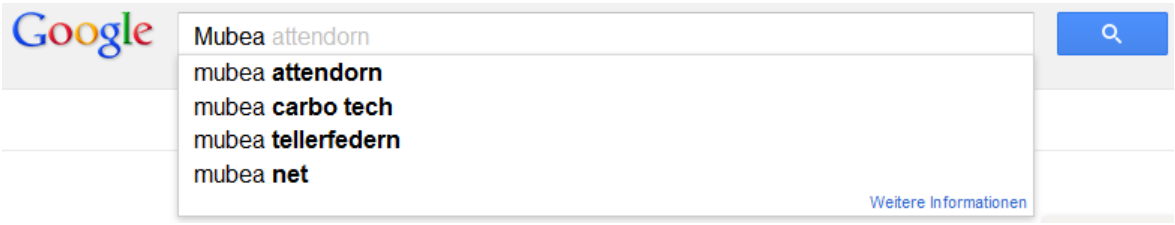


Abbildung 58. Automatische Vervollständigung bei google [142]

Abbildung 58 zeigt exemplarisch eine Suchanfrage des Begriffs ‘Mubea‘. Daraufhin werden die Suchbegrifferweiterungen ‘Attendorn‘, ‘carbo tech‘, ‘tellerfedern‘ und ‘net‘ angeboten. Interressant ist, wie die Auswahl der angebotenen Begriffe funktioniert und warum gerade diese Begriffe in der automatischen Vervollständigung erscheinen. Laut google stammen diese im Rahmen der automatischen Vervollständigung angezeigten Suchanfragen aus den Suchaktivitäten aller Webnutzer [141]. Das bedeutet, dass die angezeigten Erweiterungsbegriffe ein Indiz dafür ist, wie häufig diese Begriffe gemeinsam in einer Suchanfrage anderer google Nutzer vorkamen. Wird ein Suchbegriff häufig genug gemeinsam mit einem anderen in die google Suchmaske eingegeben, so wird dieser Begriff unter Umständen auch anderen Nutzern als Suchbegrifferweiterung angeboten.
Einen grundsätzlich anderen Ansatz verfolgt der hier vorgestellte Prototyp. Die Suchbegrifferweiterungen werden ausschließlich auf den in der Textbasis berechneten Wortassoziationen erstellt. Die eigenen Suchanfragen oder die Suchanfragen anderer Nutzer fließen in keiner Weise mit ein. Es handelt sich um ein strikt inhaltsorientiertes Verfahren,

was sich speziell für eine Suche auf firmeneigenen und somit firmenspezifischen Textdokumenten als vielversprechend erweist.
Die Wortassoziationen werden auf Basis der eingelesenen Dokumente berechnet. Diese Dokumente bilden zugleich die potentielle Ergebnismenge der Suchanfrage. Somit ist sichergestellt, dass jede berechnete Assoziation, also jede vorgeschlagene Vervollständigung aus der Textbasis resultiert und zugleich ein Suchergebnis in Form von Textpassagen garantiert. Darüber hinaus enthalten firmeneigene Dokumente gehäuft Spezialbegriffe und interne Bezeichnungen, die außerhalb des unternehmerischen Umfeldes gänzlich unbekannt sind. Solche Begriffe können im Rahmen der Assoziationssuche erkannt und als Suchbegrifferweiterung angeboten werden, sofern diese Begriffe in den eingelesenen Textdokumenten vorkommen.
Die Vor- und Nachteile beider Ansätze lassen sich am besten an einem konkreten Beispiel verdeutlichen. Hierzu wurden exemplarisch Suchanfragen an beide Systeme gestellt um die angebotenen Begriffe der Autovervollständigung bzw. der Suchbegrifferweiterung miteinander zu vergleichen.
Die folgende Abbildung 59 zeigt auf der linken Seite die Suchbegrifferweiterungen der entwickelten assoziativen Suche und auf der rechten Seite die Ergebnisse der automatischen Vervollständigung von google. In beiden Fällen wurde der Begriff ‘Federbandschellen‘ in das Dialogfeld eingetragen.
Federbandschellen entstammen der Produktpalette der Firma Mubea und dienen der Herstellung einer leckagesicheren Verbindung zwischen einem Schlauch und einem Stutzen.

Abbildung 59. Suchbegrifferweiterung des Prototypen vs. Autovervollständigung von google

Die google-Autovervollständigung schlägt die Begriffe ‘Zange‘, ‘mubea‘, ‘din 3021‘ und ‘englisch‘ vor. Bei genauer Betrachtung dieser Vorschläge ist zu konstatieren, dass zumindest

die ersten drei aus dem Umfeld des Suchbegriffs stammen. ‘Federbandschelle + Zange‘ beschreibt ein Spezialwerkzeug, das für die Installation der Schelle benötigt wird. ‘Mubea‘ ist der Hersteller und ‘din 3021‘ stellt eine Spezifikation für die Federbandschelle dar. Der letzte angebotene Begriff ‘englisch‘ weist auf die Suche nach einer englischen Übersetzung des Begriffs hin. Subjektiv betrachtet erscheinen alle von google angebotenen Begriffe als plausible Grundlage für eine Internetsuche.
Fraglich ist jedoch die Anwendbarkeit auf eine unternehmensinterne Suche auf Textdokumenten. Für diesen Zweck erscheinen die von google angebotenen Begriffe zu allgemein. Für die Suche in unternehmensinternen Textdokumenten sind sehr viel spezifischere Vorschläge vonnöten.
Die vorgeschlagenen Suchbegrifferweiterungen der assoziativen Suche sind auf der linken Seite der Abbildung 59 dargestellt. Unter anderem werden die Begriffe ‘Steckplätze‘, ‘Druckmedium‘, ‘Anschluss‘ oder ‘Typ‘ als assoziierte Worte angezeigt. Auf den ersten Blick fällt auf, dass diese Begriffe im Vergleich zu den von google empfohlenen Begriffen aus dem unmittelbaren Kontext der Suchanfrage stammen. Tests und Befragungen im Kooperationsunternehmen Mubea haben gezeigt, dass solche kontextspezifischen Begriffe für Domänenexperten auf der Suche nach Informationen aus unternehmensinternen Textdokumenten wertvoller sind, als die eher allgemein veranlagten Autovervollständigungen der Internetsuchmaschinen.
Im umgekehrten Fall muss zum Ausdruck gebracht werden, dass die spezialisierten Suchbegrifferweiterungen der assoziativen Suche für den durchschnittlichen Nutzer einer Internetsuchmaschine weniger geeignet sind, da dieser im Regelfall allgemeine Informationen oder Internetseiten sucht. Die assoziative Suchbegrifferweiterung ist konzipiert für Domänenexperten, die gezielt Informationen aus großen Textmengen extrahieren müssen.
Abschließend kann zusammengefasst werden, dass mit der Konzeptionierung und prototypischen Umsetzung der assoziativen Suche auf Textbasis ein vielversprechender Ansatz entwickelt wurde, den es weiterzuverfolgen gilt. Dafür wird derzeit im Kontakt mit unseren Praxispartnern an der Verbesserung der Konzeption und der Weiterentwicklung des Prototypen gearbeitet.

7 Zusammenfassung und Ausblick

In der vorliegenden Arbeit wurde ein Verfahren zur Imitation der menschlichen Wortassoziation entwickelt. Das als ‘Concept for the Imitation of the Human Ability of Word Association‘ bezeichnete Verfahren, wird in der Kurzform ‘CIMAWA‘ genannt. Ausgehend von der Annahme, dass Wortassoziationen bis zu einem gewissen Grad auch in Texten, beispielsweise durch vermehrten Gebrauch miteinander assoziierter Worte im selben Kontext, abgebildet sind, basiert die Eigenentwicklung auf Sammlungen von Texten. Aus genannten Textsammlungen werden zunächst statistische Werte bezüglich Wortfrequenzen und des gemeinsamen Vorkommens von Wortpaaren extrahiert. Aus diesen Werten berechnet CIMAWA für beliebige Wörter die am stärksten assoziierten Begriffe und gibt diese in Form einer absteigend nach CIMAWA-Wert sortierten Liste aus.
Nachdem die einleitenden Kapitel eine Untersuchung der menschlichen Wortassoziation aus psychologischer Sicht sowie Testreihen zur Unterlegung der Ergebnisse der Literaturarbeit enthalten, zeigt ein konzeptueller Vergleich zwischen der Eigenentwicklung und aus der Literatur bekannten Assoziationsberechnungsverfahren die fundamentalen Unterschiede der Ansätze. Darauf aufbauend zeigt ein qualitativer Vergleich, durchgeführt in drei Fallstudien mit insgesamt acht Testreihen, die Differenzen im Leistungsverhalten etablierter Assoziationsberechnungsverfahren.
Im praxisorientierten Schwerpunkt der Arbeit wird gezeigt, wie die mit CIMAWA berechneten Ergebnisse nutzbar zu machen sind. Hierfür werden sowohl umgesetzte Anwendungen und Prototypen, als auch CIMAWA basierte Konzepte vorgestellt.
Im ersten Anwendungsfall wird die CIMAWA-Assoziationsberechnung dazu eingesetzt, Texte auf ihre Themenstruktur hin zu untersuchen. In diesem Fall wird CIMAWA in Verbindung mit bewährten Text Mining-Verfahren, wie der Schlüsselwortextraktion, verwendet, um unterschiedliche Themen in ein und demselben Text zu identifizieren und präzise voneinander zu trennen. Ergebnis der als ‘Associative Gravity‘ benannten Methode sind zusammenfassende Cluster von themenbeschreibenden Schlüsselworten. Eine qualitative Evaluation der erzielten Ergebnisse wies Associative Gravity als leistungsstärkstes der getesteten Verfahren aus.
Der Ansatz des ‘Association Mapping‘ nutzt CIMAWA für die Analyse von Textdokumenten im Bereich der Instandhaltung von Produktionsmaschinen. Hierfür wurde ein Konzept entwickelt, das die Verbindung zwischen den Forschungsgebieten des Text Mining auf der einen und dem Instandhaltungsmanagement auf der anderen Seite herstellt. Das Association Mapping zielt darauf ab, die bisweilen ungenutzen Potentiale der abgelegten Dokumente der Instandhaltung zu nutzen, indem es die Verbindungen zwischen definierten Instanzen (Maschinen, Mitarbeiter, Instandhaltungsaktivitäten) identifiziert und visualisiert. Wie das Association Mapping im Anwendungsfeld der Produktentwicklung und in Verbindung mit kontextspezifischen Textempfehlungen eingesetzt werden kann, wurde in einem eigenen Unterkapitel herausgearbeitet.

Der Versuch, einen Beitrag zur Verbesserung der Wiederauffindung und zur Steigerung der Wiederverwendung unternehmensinterner Textdokumente zu leisten, wird mit der prototypischen Entwicklung der ‘Assoziativen Suche‘ unternommen. In Verbindung und mit der Unterstützung unseres Kooperationspartners Mubea, ist im Rahmen einer Diplomarbeit die erste Version einer unternehmensinternen Suchmaschine implementiert worden, die die CIMAWA Assoziationen zur Erweiterung und Präzisierung von Suchanfragen auf der internen Wissensbasis der Unternehmung verwendet. Erste Testergebnisse und interne Nutzerbefragungen bestätigen dabei die Nutzbarkeit der erzielten Ergebnisse.

Im Hinblick auf die weitere Nutzung des entwickelten CIMAWA-Ansatzes sei zunächst auf die bereits entwickelten bzw. auf die in der Entwicklung befindlichen Anwendungen verwiesen. Allein die in dieser Arbeit vorgestellten CIMAWA-Anwendungen decken ein breites Spektrum an Anwendungsfeldern ab. Es ist davon auszugehen, dass sich dieses Anwendungsfeld im Zuge der weiteren Zusammenarbeit mit unseren Kooperationspartnern sowohl im industriellen als auch im wissenschaftlichen Umfeld noch erweitert.
Außerdem ist die Ausweitung der vorgestellten CIMAWA-Anwendungen an dieser Stelle zu thematisieren. Aus Sicht des Autors besteht die Möglichkeit der Weiterentwicklung bzw. der Spezialisierung der vorgestellten Konzepte, bezüglich konkreter praxisnaher Anwendungsszenarien.
Ferner befinden sich weitere Forschungen für den CIMAWA-Einsatz in der Entwicklungsphase. Exemplarisch zu nennen ist der sich ausweitende Bereich der Sentiment Analyse / Opinion Mining, in dem es darum geht, Texte auf Meinungen oder Polarität hin zu untersuchen. Erste Forschungsergebnisse liegen bereits im Bereich der temporalen Analyse von Assoziationen vor. Dieser Ansatz untersucht regelmäßig aktualisierte Textsammlungen, bestehend aus redaktionellen Artikeln von Presseargenturen oder online Zeitungen. Ziel hierbei ist es, Trendforschung zu betreiben, indem die Assoziationen und insbesondere die sprunghaften Veränderungen dieser Assoziationen mit CIMAWA berechnet und ausgewertet werden.

8 Abbildungsverzeichnis

9 Tabellenverzeichnis

10 Formelverzeichnis

11 Literaturverzeichnis

[1] G. Simmel, The sociology of Georg Simmel, Glencoe, 1950.

[2] J. G. Miller, „Information Input Overload and Psychopathology," *American Journal of Psychiatry 116 (2),* pp. 695 - 704, 1960.

[3] P. Hemp, „Death by Information Overload," *Harvard Business Review,* 12.04.2009.

[4] J. Naisbitt, Megatrends, Warner Books, 1982.

[5] Y. Chen, L. Wang, M. Dong und J. Hua, „Exemplar-based Visualization of Large Document Corpus," *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 15, NO. 6,* pp. 1161 - 1168, 11/12 2009.

[6] A. Rzhetsky, M. Seringhaus und M. Gerstein, „Seeking a New Biology through Text Mining," *Cell, Volume 134, Issue 1,* pp. 9 - 13, 11.07.2008.

[7] Y. Zhou, Y. Zhang, N. Vonortas und J. Williams, „A Text Mining Model for Strategic Alliance Discovery," in *Proceedings of 45th Hawaii International Conference on System Sciences*, Hawaii, USA, 2012.

[8] S. Ananiadou, T. Ohta und M. K. Rutter, „Text Mining Supporting Search for Knowledge Discovery," *Curr Cardiovasc Risk Rep; Springer Science+Business Media New York 2012,* 22.12.2012.

[9] A. Mehler und C. Wolff, „Perspektiven und Positionen des Text Mining," *Zeitschrift für Computerlinguistik und Sprachtechnologie,* pp. 1 - 18, Heft 1, Band 20, 2005.

[10] R. Feldman und J. Sanger, The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data, New York: Cambridge University Press, 2007.

[11] M. Hearst, „What is Text Mining?," 07.02.2003. [Online]. Available: www.sims.berkeley.edu/~hearst/text-. [Zugriff am 11.06.2013].

[12] C. C. Aggarwal und C. Zhai, Mining Text Data, New York, Dordrecht, Heidelberg, London: Springer, 2012.

[13] H. Chen, Knowledge Management Systems: A Text Mining Perspective, Tucson, Arizona: University of Arizona, 2001.

[14] K.-E. Sommerfeldt, G. Starke und W. Hackel, Einführung in die Grammatik der deutschen Gegenwartssprache, Tübingen: Niemeyer, 1998.

[15] C. Shannon und W. Weaver, The Mathematical Theory of Communication, Urbana: University of Illinois Press, 1949.

[16] G. Heyer, U. Quasthoff und T. Wittig, Text Mining: Wissensrohstoff Text, Herdecke: W3L-Verlag, 2012.

[17] M. A. Hearst, „Untangling Text Data Mining," in *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland*, 1999.

[18] I. Nonaka und H. Takeuchi, The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation, Oxford University Press, 1995.

[19] T. Joachims und E. Leopold, „Vorwort der Herausgeber," *Themenheft: Text-Mining, Künstliche Intelligenz,* 2002.

[20] A. -. H. Tan, „Text Mining: The State of the Art and the Challenges," in *Proceedings of the Pacific Asia Conference on Knowledge Discovery and Data Mining PAKDD'99*, 1999.

[21] R. Feldman und I. Dagan, „Knowledge Discovery in Textual Databases (KDT)," in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD'95)*, 1995.

[22] U. Hahn und K. Schnattinger, „Towards Text Knowledge Engineering," in *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98) and of the 10th Conference on Innovative Applications of Artificial Intelligence (IAAI-98)*, Menlo Park, 1998.

[23] Y. Kodratoff, „Knowledge Discovery in Texts: A Definition and Applications," in *Proceedings of the 11th International Symposium on Foundations of Intelligent Systems (ISMIS'99)*, Berlin/Heidelberg/New York, 1999.

[24] D. Merkl, Text Data Mining. In R. Dale, H. Moisl,& H. Somers (EDS), Handbook of Natural Language Processing, New York: Dekker, 2000.

[25] P. Losiewicz, D. W. Oard und R. N. Kosthoff, „Textual Data Mining to support Science and Technology Management," *Journal of Intelligent Information Systems, 15,* pp. 99 - 119, 2000.

[26] J. Franke, G. Nakhaeizadeh und I. Renz, Text Mining: Theoretical Aspects and Applications, Heidelberg/New York: Physika-Verlag/Springer, 2003.

[27] M. W. Berry, Survey of Text Mining, New York: Springer-Verlag, 2003.

[28] M. C. Meier und M. Beckh, „State-of-the-Art: Text Mining," *Wirtschaftsinformatik,* pp. 165 - 167, 04/2000.

[29] D. C. Hoaglin, F. Mosteller und J. W. Tukey, Understanding Robust and Exploratory Data Analysis, John Wiley & Sons, Inc, 1983.

[30] J. W. Tukey, Exploratory Data Analysis, Addison-Wesley Publishing Company, 1977.

[31] B. Thuraisingham, Data Mining: Technologies, Techniques, Tools, and Trends, Boca Raton: CRC Press; 1 edition, 1999.

[32] U. Fayyad, G. Piatetsky-Shapiro und P. Smyth, „The KDD Process for Extracting Useful Knowledge from Volumes os Data," *Communications of the ACM / Vol. 39, No 11,* pp. 27 - 34, 11/1996.

[33] C. Coronel, S. A. Morris und P. Rob, Database Systems - Design, Implementation and Management, Cengage Learning, 2011.

[34] S. Armstrong, Using Large Corpora, A Bradford Book; 1st MIT Press edition, 1994.

[35] J. H. Kroeze, M. C. Mathee und T. J. D. Bothma, „Differentiating between Data Mining and Text Mining Terminology," *South African Journal of Information Management, Vol. 6(4),* 12 2004.

[36] G. K. Zipf, The Psych-Biology of Language. An Introduction to Dynamic Philology, Boston: Houghton-Mifflin, 1935.

[37] A. Klahold, Empfehlungssysteme, Wiesbaden: Vieweg + Teubner / GWV Fachverlag GmbH, 2009.

[38] B. B. Mandelbrot, „An Information Theory of the Statistical Structure of Language," *Communication Theory,* pp. 503 - 512, 1953.

[39] Duden, „Fremdwörterlexicon," [Online]. Available: http://www.duden.de/node/764582/revisions/1069264/view. [Zugriff am 11.06.2013].

[40] S. Evert, The Statistics of Word Cooccurrences Word Pairs and Collocations, Stuttgart: PhD, 2005.

[41] P. Pecina und P. Schlesinger, „Combining Association Measures for Collocation Extraction," in *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, Sydney, 2006.

[42] M. E. Stevens, V. E. Giuliano und L. B. Heilprin, Proceedings of the Symposium on Statistical Association Methods For Mechanized Documentation, Washington: volume 269 of National Bureau of Standards Miscellaneous, 1965.

[43] H. Grimm und E. Johannes, Sprachpsychologie - Handbuch und Lexicon der Psycholinguistik, Berlin: Erich Schmidt Verlag, 1981.

[44] M. Jahn, Psychologie als Grundwissenschaft der Pädagogik, Paderborn: Trapeza, 2012 (Reprint des Originals von 1904).

[45] D. Hartley, Observations on Man, his Frame, his Duty and his Expectations, Reprinted for J. Johnson, by W. Eyres, 1801.

[46] J. Mill, Analysis of the phenomena of the Human Mind, Georg Olms Verlag, 1869.

[47] R. Rapp, Die Berechnung von Assoziationen: Ein Korpuslinguistischer Ansatz, Hildesheim; Zürich; New York: Georg Olms Verlag, 1996.

[48] W. James, The Principles of Psychology, New York: Dover Publications, 1890.

[49] P. Seidensticker, Simulation von Wortassoziationen mit Hilfe von mathematischen Lernmodellen in der Psychologie, Paderborn, 2006.

[50] A. W. Staats, Learning Language and Cognition, Holt, Reinhart and Winston, 1986.

[51] G. H. Kent und A. J. Rosanoff, „A study of association in insanity," *American Journal of Insanity,* pp. 317 - 390, 1910.

[52] W. A. Russell, „The complete german language norms for responses to 100 words from the Kent-Rosanoff word association test," *L. Postmann & G. Keppel (Eds.), Norms of word association,* pp. 53 - 94, 1970.

[53] D. L. Nelson, C. L. McEvoy und T. A. Schreiber, „The University of South Florida word association, rhyme, and word fragment norms," http://www.usf.edu/FreeAssociation/ (Stand 30.11.2012), 1998.

[54] D. L. Nelson, C. L. McEvoy und T. A. Schreiber, „The University of South Florida Word Association, Rhyme and Word Fragment Norms," *Behavior research methods instruments computers a lournal of the Psychonomic Society Inc,* pp. 402 - 407, 2004.

[55] F. d. Saussure, Grundfragen der allgemeinen Sprachwissenschaft, Berlin: de Gruyter, 2001 (reprint).

[56] R. H. Robins, General linguistics: An introductory survey, Longmans, 1968.

[57] W. Ulrich, Wörterbuch Linguistische Grundbegriffe, Borntraeger; Auflage: 5., erneut bearb. u. erw. A., 2002.

[58] C. D. Manning und H. Schütze, Foundations of Statistical Natural Language Processing, Massachusetts Institute of Technology, 1999.

[59] L. Michelbacher, S. Evert und H. Schütze, „Asymmetry in corpus-derived and human word associations," *Corpus Linguistics and Linguistic Theory,* p. 245–276, August 2011.

[60] M. Steyvers, R. M. Shiffrin und D. L. Nelson, „Word Association Spaces for Predicting Semantic Similarity Effects in Episodic Memory," *Experimental Cognitive Psychology and its Applications: Festschrift in Honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer,* 2004.

[61] P. Uhr, A. Klahold und M. Fathi, „Imitation of the Human Ability of Word Association," *The International Journal of Soft Computing and Software Engineering, Vol. 3, No. 3, Doi: 10.7321/jscse. v3n3.37,* pp. 248 - 254, 03/2013.

[62] U. Quasthoff und C. Wolff, „The Poisson collocation measure and its application," in s *Workshop on Computational Approaches to Collocations*, Wien, Österreich, 2002.

[63] S. F. Dennis, „The construction of a thesaurus automatically from a sample of text," in *Proceedings of the Symposium on Statistical Association Methods For Mechanized Documentation*, Washington, DC, National Bureau of Standards Miscellaneous Publication, 1965, pp. 61 - 148.

[64] G. L. Berry-Rogghe, „The computation of collocations and their relevance to lexical studies," in *The Computer and Literary Studies*, Edinburgh, pp. 103 – 112, 1973.

[65] M. H. DeGroot und M. J. Schervish, Probability and Statistics, 3rd edition, Boston: Addison Wesley, 2002.

[66] K. W. Church und W. A. Gale, „Concordances for parallel text," in *Proceedings of the 7th Annual Conference of the UW Center for the New OED and Text Research*, Oxford, UK, 1991.

[67] T. E. Dunning, „Accurate methods for the statistics of surprise and coincidence," *Computational Linguistics, 19(1),* p. 61 – 74, 1993.

[68] D. Liddell, „Practical tests of 2 × 2 contingency tables," *The Statistician 25(4),* p. 295–304, 1976.

[69] F. Smadja, K. R. McKeown und V. Hatzivassiloglou, „Translating collocations for bilingual lexicons: A statistical approach," *Computational Linguistics, 22(1),* pp. 1 – 38, 1996.

[70] T. E. Dunning, Finding Structure in Text, Genome and Other Symbolic Sequences, Sheffield: Ph.D. thesis, Department of Computer Science, University of Sheffield, 1998.

[71] T. Pedersen und R. Bruce, „What to infer from a description," Technical Report 96-CSE-04, Southern Methodist University, Dallas, USA, 1996.

[72] D. Blaheta und M. Johnson, „Unsupervised learning of multi-word verbs," in *Proceedings of the ACL Workshop on Collocations*, Toulouse, France, 2001.

[73] K. W. Church und P. Hanks, „Word Association Norms, Mutual Information and Lexicography," *Computational Linguistics, 16(1),* pp. 22 - 29, 1990.

[74] B. Daille, Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques, Paris: Ph.D. thesis, Université Paris 7, 1994.

[75] M. Wettler, R. Rapp und R. Ferber, „Freie Assoziationen und Kontiguitäten von Wörtern in Texten," *Zeitschrift für Psychologie, 201,* pp. 99 - 108, 1993.

[76] M. Wettler und R. Rapp, „Computation of word associations based on the co-occurences of words in large corpora," in *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus, Ohio, 1993.

[77] J. R. Firth, „A Synopsis of Linguistic Theory 1930 - 1950," *Studies in Linguistic Analysis; Reprinted in F.R. Palmer(ed), Selected Papers of J.R.Firth 1952-1959,* pp. 1 - 32, 1968.

[78] A. Klahold, P. Uhr, F. Ansari und M. Fathi, „A Framework to utilize the Humen Ability of Word Association for detecting Multi Topic Structures in Text Documents," *IEEE Intelligent Systems,* 06.11.2013.

[79] „Institut für Deutsche Sprache, Das Deutsche Referenzkorpus DeReKo," http://www.ids-mannheim.de/kl/projekte/korpora, Mannheim, 2012.

[80] R. M. Fano, „Transmission of Information: A statistical Theorie of Communication," *MIT Press, New York,* 1961.

[81] K. W. Church, W. Gale, P. Hanks und D. Hindle, „Using Statistics in Lexical Analysis," *Lexical Aquisition: Exploiting On-Line Resources to build a Lexicon, Hillsdale, NJ: Lawrence Erlbaum,* pp. 115 - 164, 1991.

[82] D. Hindle, „Noun Classification from Predicate Argument Structures," *ACL28,* pp. 268 - 275, 1990.

[83] IDS, „Kurzinformationen zu COSMAS II," 03 2009. [Online]. Available: http://www.ids-mannheim.de/cosmas2/uebersicht.html. [Zugriff am 07.03.2013].

[84] K. Aliyev, „Masterarbeit: Korpusabhängiger Vergleich und Realisierung statistischer Berechnungsverfahren zur Implementierung der menschlichen Wortassoziation im Deutschen und Englischen," Siegen, 2012.

[85] X. J. Ma, W. X. Wang, Y. C. Lai und Z. Zheng, „Information explosion on complex networks and control," *The European Physical Journal B, vol. 76, no 1,* pp. 179 - 183, 2010.

[86] J. Dörre, P. Gerstl und R. Seiffert, „Text Mining: Finding Nuggets in Mountains of Textual Data," in s *Proceedings International Conference of Knowledge Discovery and Data Mining*, 1999.

[87] T. Jiang, A. H. Tan und K. Wang, „Mining Generalized Associations of Semantic Relations from Textual Web Content," *IEEE Transactions on Knowledge and Data Engineering, VOL. 19, NO. 2,* 02/2007.

[88] P. Goyal, L. Behera und T. M. McGinnity, „A Context based Word Indexing Model for Document Summarization," *IEEE Transactions on Knowledge and Data Engineering, VOL. 25, NO. 8,* 25.08.2013.

[89] R. Mihalcea und P. Tarau, „TextRank: Bringing Order into Texts," in s *Proceedings of EMNLP 2004*, Barcelona, Spain, 2004.

[90] K. Sugiyama, K. Hatano, M. Yoshikawa und S. Uemura, „Improvement in tf-idf scheme for web pages based on the contents of their hyperlinked neighbouring pages," *Syst. Comput. Japan 36,* pp. 56 - 68, 2005.

[91] H. Luhn, „The Automatic Creation of Literature Abstracts," *IBM Journal Pages ,* pp. 159 - 165, 1958.

[92] K. Spärck-Jones, „A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation Vol. 28,* pp. 11 - 21, 1972.

[93] H. Wu und G. Salton, „A comparison of search term weighting: term relevance vs. inverse document frequency," in *Proceedings of the 4th annual international ACM SIGIR conference on Information storage and retrieval*, 1981.

[94] A. Klahold, CRIC: Kontextbasierte Empfehlungunstrukturierter Texte in Echtzeitumgebungen, Haiger, 2006.

[95] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin und C. G. Nevill-Manning, „KEA: Practical Automatic Keyphrase Extraction," *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific,* pp. 129 - 152, 2005.

[96] S. Deerwester, G. Furnas und T. Landauer, „Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science, vol. 41, no. 6,* pp. 391 - 407, 1990.

[97] I. Newton, Philosophiæ naturalis principia mathematica (Mathematical principles of natural philosophy), London, 1687.

[98] A. Klahold, A. Holland und M. Fathi, „Computation of Asymmetric Semantic Document Relations," in *Proceedings of the 13th International Conference on Artificial Intelligence and Soft Computing*, Mallorca, Spain, 2009.

[99] P. B. Baxendale, „Machine-made index for technical literature - an experiment," *IBM journal, vol. 10,* pp. 354 - 361, 1957.

[100] A. Klahold, P. Uhr, F. Ansari und M. Fathi, „A Framework to utilize the Human Ability of Word Association for detecting Multi Topic Structures in Text Documents," *IEEE Intelligent Systems,* 06.11.2013.

[101] S. Fortunato, „Community detection in graphs," *Journal of physics Reports 486,* pp. 75 - 174, 2010.

[102] S. E. Schaeffer, „Graph Clustering," *Journal of Computer Science,* pp. 27 - 64, 2007.

[103] C. Biemann, Structure Discovery in Natural Language, Springer, pp. 73 – 77, 2012.

[104] D. Fasulo, „An analysis of recent works on Clustering Algorithms," Department of Computer Science & Engineering, University of Washington, 1999.

[105] J. B. Mac Queen, „Proceedings of the fifth Berkeley Symposium on Mathemetical Statistics and Probability,edited by L. M. L. Cam and J. Neyman," Berkeley, 1967, pp. 281 - 297.

[106] A. Hlaoui und S. Wang, Neural Networtks and Computational Intelligence, 2004.

[107] W. Donath und A. Hoffman, „IBM Journal of Research and Development," 1973.

[108] U. von Luxburg, „A Tutorial on Spectral Clustering," Max Planck Institute for Biological Cybernetics, 2006.

[109] P. Cimiano, A. Hotho und S. Staab, „Comparing Conceptual, Divisive and Agglomerative Clustering for Learning Taxonomies from Text," in *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, 2004.

[110] X. Li, „Parallel Algorithms for Hierarchical Clustering and Clustering Validity," *IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 12,* pp. 1088 - 1092, 1990.

[111] S. Rajasekaran, „Efficient Parallel Hierarchical Clustering Algorithms," *IEEE Transactions on Parallel and Distributet Systems, VOL. 16,* pp. 497 - 502, 2005.

[112] C. D. Manning, P. Raghavan und H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[113] B. F. William und R. Baeza-Yates, Information Retrieval: Data Structures and Algorithms, 1992.

[114] J. He, A. H. Tan, C. L. Tan und S. Y. Sung, „On Quantitative Evaluation of Clustering Systems," *Information Retrieval and Clustering, Kluwer Academic Publishers,* pp. 105 - 134, 2003.

[115] R. Kashef und M. S. Kamel, „Cooperative Clustering," *Pattern Recognition, vol. 43, issue 6 ,* pp. 2315 - 2329, 2010.

[116] U. Maulik und S. Bandyopadhyay, „Performance Evaluation of some Clustering Algorithms and Validity Indicies," *IEEE Transactions on Pattern Analysis and Machine Learning,* pp. 1650 - 1654, 2002.

[117] D. Boley, „Principal Direction Divisive Partitioning," *Data Mining and Knowledge Discovery,* pp. 325 - 344, 1998.

[118] E. Cambria, B. Schuller, Y. Xia und C. Havasi, „New Avenues in Opinion Mining and Sentiment Analysis," *IEEE Intelligent Systems,* pp. 15 - 21, March/April 2013.

[119] F. Ansari, P. Uhr und M. Fathi, „Textual Meta-analysis of Maintenance Management's Knowledge Assets," *International Journal of Services, Economics and Management, Vol. 6, No. 1,* p. 14 – 37, 2014.

[120] I. Nonaka, R. Toyama und N. Konno, „SECI, Ba and Leadership: A Unified Model of Dynamic Knowledge Creation," *Long Range Planning, Vol. 33, Issue 1,* pp. 05 - 34, 01. 02/2000.

[121] R. Dawson, „Knowledge capabilities as the focus of organizational development and strategy," *Journal of Knowledge Management, Vol. 4, Issue 4,* pp. 320 - 327, 2000.

[122] F. Wijnhoven, „Knowledge Integration," in *Knowledge Management: More than a Buzzword*, Germany, Physica Verlag, 2006, pp. 1 - 16.

[123] R. Maier, Knowledge Management Systems, Information and Communication Technologies for Knowledge Management, 3rd edition, Germany: Springer, 2007.

[124] K. Bagadia, Computerized Maintenance Management Systems made easy, McGraw-Hill, 2006.

[125] R. K. Mobley, L. R. Higgins und D. J. Wikoff, Maintenance Engineering Handbook, 7th edition, McGraw-Hill, 2008.

[126] G. Glass, „Integrating Findings: The Mata-analysis of research," *Review of Research in Education, Vol. 5, No. 1,* pp. 351 - 379, 1977.

[127] J. F. Salgado, N. Anderson, S. Moscoso, C. Bertua, G. College, F. de Fruyt und J. P. Rolland, „A Meta-Analytic Study of General Mental Ability Validity for Different Occupations in the European Community," *Journal of Applied Psychology, Vol. 88, No. 6,* pp. 1068 - 1081, 2003.

[128] L. C. Lyons, „Meta-Analysis: Methods of Accumulating Results across Research Domains," Manassas, USA, 2000, (Last revised: 30.01.2003).

[129] D. Parmenter, Key Performance Indicators (KPI): Developing, Implementing, and Using Winning KPIs. 2nd edition, John Wiley & Sons, Inc., 2010.

[130] F. Ansari, M. Fathi und U. Seidenberg, „Developing an Algebraic Model for Administrating Preventive Maintenance Cost of Production Machines," in *Proceedings of the 4th World Conference Production & Operations Management and the 19th International Annual EurOMA Conference, University of Amsterdam*, The Netherlands, 2012.

[131] F. Ansari, A. Dewa und M. Fathi, „Drive System Assistance Tool for Meta-Analysis of Dimensioning and Maintenance Indicators," in *Proceedings of the 38th Annual Conference of the IEEE Industrial Electronics Society (IEEE-IECON 2012)*, Montreal, Canada, 2012.

[132] DIN, „Maintenance - Maintenance terminology, Trilingual version EN 13306," Deutsches Institut für Normung, Berlin, 2010.

[133] P. Uhr, S. Dienst, A. Klahold und M. Fathi, „Kontextbasierte Bereitstellung von Textdokumenten im Produktverbesserungsprozess," *wt Werkstattstechnik online,* pp. 528 - 534, 7/8 2012.

[134] S. Dienst, P. Uhr, M. Fathi, A. Klahold, M. Abramovici und A. Lindner, „Concept for Improving Industrial Goods via Contextual Knowledge Provision," in *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, Austria, Graz, 2012.

[135] M. Abramovici, A. Lindner, A. Walde, M. Fathi und S. Dienst, „Decision Support for Improving the Design of Hydraulic Systems by Leading Feedback into Product Dvelopment," in *Proceedings of the 18th International Conference on Engineering Design*, Kopenhagen, 2011.

[136] S. Dienst, P. Uhr, F. Ansari und M. Fathi, „Using Data Analysis for Discovering Improvement Potentials in Production Process," in *Joint IEEE International Conference on Industrial Electronics*, Auburn, Alabama, 2011.

[137] FIFA.com, „FIFA Fussball-Weltmeisterschaft Brasilien 2014TM," 22.08.2013. [Online]. Available: http://de.fifa.com/worldcup/organisation/ticketing/news/newsid=2156032/index.html. [Zugriff am 26.08.2013].

[138] C. Hebel, „www.spiegel.de," 01.08.2013. [Online]. Available: http://www.spiegel.de/politik/ausland/rio-de-janeiro-demonstranten-stuermen-stadtratsgebaeude-a-914295.html. [Zugriff am 26.08.2013].

[139] T. Arens, „Konzeptionierung und prototypische Realisierung eines intelligenten Suchverfahrens auf Textbasis als ein intranet basierter Ansatz bei Mubea," Siegen, 2013.

[140] Mubea, „www.mubea.com," 16.09.2013. [Online]. Available: http://www.mubea.com/de/unternehmen/kunden/. [Zugriff am 16.09.2013].

[141] google, „Automatische Vervollständigung," google, 2013. [Online]. Available: https://support.google.com/websearch/answer/106230?hl=de. [Zugriff am 24.09.2013].

[142] google, „www.google.de," 2013. [Online]. Available: https://www.google.de/#q=Mubea%20attendorn. [Zugriff am 24.09.2013].