

**Data Mining und Big Data Analytics:
Semantische Suche, Prognose und
Entscheidungsunterstützung mit Künstlichen Neuronalen Netzen**

Von der Wirtschaftswissenschaftlichen Fakultät der
Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des akademischen Grades

Doktor der Wirtschaftswissenschaften
- Dr. rer. pol. -

genehmigte Dissertation

von

Dipl.-Ök. Christoph Gleue

2019

Betreuer und Gutachter:	Prof. Dr. Michael H. Breitner
Weiterer Gutachter:	Prof. Dr. J.-Matthias Graf von der Schulenburg
Vorsitzender der Prüfungskommission:	Prof. Dr. Stefan Wielenberg
Weiteres Mitglied (beratend):	Dr. Michelle Muraz
Tag der Promotion	15.05.2019

Danksagungen

Die Zeit während meines Promotionsstudiums war für mich eine der prägendsten in meinem bisherigen Privat- und Berufsleben. Der Weg von einem „frisch gebackenen“ Diplom-Ökonom bis zum heutigen Tage war mit vielen Entscheidungen, beruflichen und privaten Scheidewegen, Bekanntschaften und Freundschaften gepflastert, daher möchte ich auf diesem Wege allen besonderen Menschen, die mich auf diesem Weg begleitet haben, von Herzen danken.

Besondere Erwähnung verdienen dabei:

Meine Eltern, dafür, dass sie mich jederzeit mit allen erdenklichen Mitteln und Ratschlägen unterstützten und mir stets ein Wegweiser auch in schweren Zeiten waren, sind und bleiben.

Meine Lebensgefährtin Sandra Pöllmann, ohne deren stetige und liebevolle Unterstützung diese Zeit für mich um ein Vielfaches schwieriger zu meistern gewesen wäre – danke Sandra, dafür, dass Du mir nicht nur eine große moralische Unterstützung, sondern auch ein intellektueller Sparringspartner und so eine tolle Begleiterin warst und bist.

Berthold Neumann, der mich stets gefördert und mir als Mentor, Kollege, guter Freund und Berater vieles davon beibrachte, was mich in meiner beruflichen Karriere letztlich erfolgreich machen würde. Auch nach seinem plötzlichen, tragischen Tod kurz vor Abgabe dieser Doktorarbeit, zu deren Verwirklichung er einen maßgeblichen Anteil beigetragen hat, werde ich ihn immer als einen der einflussreichsten Menschen in meinem Leben in bester Erinnerung behalten und ihm für alles, was er für mich getan hat auf ewig dankbar sein.

Dennis Eilers und André Koukal für die hervorragende Teamarbeit, die schöne und inspirierende gemeinsame Zeit am Institut, den Konferenzbesuchen in Israel, Neuseeland und der Schweiz und natürlich bei der wissenschaftlichen Arbeit, die mir stets großen Spaß gemacht und letztlich zu unseren gemeinsamen Publikationserfolgen geführt hat.

Prof. Dr. Michael H. Breitner für seine vielseitige Unterstützung in dieser Zeit sowie seine fachliche Expertise und Beratung bei der Erarbeitung der in dieser Zeit entstandenen, wissenschaftlichen Arbeiten.

Abschließend möchte ich mich bei allen Freunden, Weggefährten, Kollegen und „Mitstreitern“ am Institut sowie meinem Projektteam der letzten Jahre für die schöne gemeinsame Zeit, die Inspiration, die vielen tollen Gespräche und gemeinsamen Abende bedanken, die wir zusammen verbringen durften.

I. Abstract

This cumulative dissertation summarizes and critically reflects on a total of six scientific publications in which I was involved as an author. The overall objective of this thesis is the analysis and discussion of the question of how traditional, historically grown workflows and business processes can be optimized using methods and technologies from the broad fields of data mining, big data analytics, and data science. It is of particular interest how this can be accomplished successfully, beyond the large "lighthouse projects" in industry and research but in small and medium-sized projects in science and in the private sector. All publications are based on two independent research projects:

- (1) **Semantic Similarity Analysis in the Literature Research Process:** This section presents the development, demonstration and benchmarking of a novel tool for semantic literature research based on Latent Semantic Indexing (LSI). The "Tool for Semantic Indexing and Similarity Queries" (TSISQ) enables scientists, regardless of their field of research, to quickly and easily identify semantically similar research papers. In addition, the structure and architecture of the system are presented, hints for its practical application are given, and the efficiency and quality improvements that can be achieved through the use of TSISQ in the literature research process are pointed out.

- (2) **Decision Support for the automotive industry:** This section describes the results and findings of a cooperation project with a large German car manufacturer, in which a comprehensive decision support system was designed and implemented. The most important core functionality of this system is the forecast of market and residual values for used vehicles using Artificial Neural Networks. In addition to the presentation of the findings in the field of long-term forecasts, our results regarding internal (influenceable by the manufacturer) and external (macroeconomic) influencing factors on vehicle residual values are shown and critically discussed.

Keywords: Semantic Search, Latent Semantic Indexing, Information Retrieval, Text Mining, Literature Research, Decision Support, Data Mining, Big Data Analytics, Data Warehouse, Artificial Neural Networks, Leasing, Residual Value Management, Used Vehicles

Kurzdarstellung

Diese kumulative Dissertation ist eine Zusammenfassung und kritische Reflexion von insgesamt sechs wissenschaftlichen Publikationen, an denen ich als Autor beteiligt war. Das übergeordnete Ziel dieser Arbeit ist die Analyse und Diskussion der Frage, wie abseits der großen „Leuchtturmprojekte“ in Industrie und Forschung auch in kleineren und mittleren Projekten in Wissenschaft und Privatwirtschaft mit Methoden und Technologien aus dem weiten Feld des Data Mining, Big Data Analytics und Data Science althergebrachte Arbeitsabläufe und Geschäftsprozesse erfolgreich optimiert werden können. Alle Veröffentlichungen basieren auf zwei voneinander unabhängigen Forschungsprojekten:

- (1) Semantische Ähnlichkeitsanalyse im Literaturrechercheprozess: In diesem Abschnitt werden Entwicklung, Demonstration und Benchmark eines neuartigen Werkzeugs zur semantischen Literaturrecherche basierend auf Latent Semantic Indexing (LSI) vorgestellt. Das „Tool for Semantic Indexing and Similarity Queries“ (TSISQ) ermöglicht Wissenschaftlern, unabhängig von ihrem Forschungsgebiet, die schnelle und einfache Identifikation von semantisch ähnlichen Forschungsarbeiten. Es werden zudem Aufbau und Architektur des Systems präsentiert, Hinweise für dessen praktische Anwendung gegeben und die Effizienz- und Qualitätsverbesserungen aufgezeigt, die durch den Einsatz von TSISQ im Literaturrechercheprozess erzielt werden können.
- (2) Entscheidungsunterstützung für die Automobilindustrie: Dieser Abschnitt beschreibt die Ergebnisse und Erkenntnisse eines Kooperationsprojektes mit einem großen deutschen Automobilhersteller, wobei ein umfangreiches Entscheidungsunterstützungssystem (EUS) konzipiert und implementiert wurde. Die wichtigste Kernfunktionalität dieses Systems stellt dabei die bedarfsgerechte Prognose von Markt- und Restwerten für Gebrauchtfahrzeuge mit Künstlichen Neuronalen Netzen (KNN) dar. Neben der Vorstellung der Erkenntnisse im Bereich der Langfristprognosen werden auch unsere Ergebnisse bzgl. interner (vom Hersteller beeinflussbarer) und externer (makroökonomischer) Einflussfaktoren auf Fahrzeugrestwerte präsentiert.

Schlagworte: Semantische Suche, Latent Semantic Indexing, Information Retrieval, Text Mining, Literaturrecherche, Entscheidungsunterstützung, Data Mining, Big Data Analytics, Data Warehouse, Künstliche Neuronale Netze, Leasing, Restwertmanagement, Gebrauchtfahrzeuge

II. Management Summary

Problemstellung und Forschungsziele

Data Mining, Business Intelligence, Big Data Analytics und Data Science gehören seit einigen Jahren in nahezu allen Branchen zu den wichtigsten Zukunftsthemen. Die Nachfrage nach Mitarbeitern mit entsprechenden Fähigkeiten ist inzwischen deutlich höher als das Angebot (vgl. Davenport und Patil, 2012). Auch in der Wirtschaftsinformatik als wissenschaftlicher Disziplin steigt seit Jahren stetig die Anzahl an Publikationen in diesem überaus umfangreichen Themengebiet. Moderne Technologien, benutzerfreundlichere Analysesoftware, ein hoher Vernetzungsgrad und vergleichsweise günstige Rechen- und Speicherkapazitäten machen die Verarbeitung und Analyse von „Big Data“ inzwischen einem breiteren Publikum zugänglich (vgl. Agarwal und Dhar, 2014).

Ziel dieser Dissertation ist es, zu untersuchen, wie abseits der großen „Leuchtturmprojekte“ in Industrie und Forschung mit Methoden und Technologien aus dem weiten Feld des Data Mining, (Big) Data Analytics und Data Science althergebrachte Arbeitsabläufe automatisiert und optimiert werden können. Dazu werden im Folgenden zwei Projekte mit jeweils drei Veröffentlichungen vorgestellt, die exemplarisch zeigen sollen, wie auch kleinere und mittlere Projekte in der Welt der Wissenschaft und der Privatwirtschaft erfolgreich durchgeführt werden können und welche Herausforderungen sich dabei typischerweise ergeben. Diese kumulative Doktorarbeit besteht aus zwei unabhängigen Teilen. Beide Teile bewegen sich im Themenkomplex des Data Mining und Big Data Analytics, basieren jedoch auf zwei voneinander unabhängigen Projekten.

Semantische Ähnlichkeitsanalyse im Literaturrechercheprozess:

In Kapitel 2 werden die Erkenntnisse aus drei Forschungsarbeiten zusammengefasst, die einen Beitrag zum Themengebiet des Text Mining bzw. der semantischen Textanalyse zur Unterstützung der Literaturarbeit in der wissenschaftlichen Praxis leisten. Als Ergebnis ist dabei ein Werkzeug zur automatisierten Literaturrecherche auf Basis der Erkennung semantischer Ähnlichkeiten zwischen Texten entstanden, das „Tool for Semantic Indexing and Similarity Queries“ (TSISQ). Die folgenden Themen und Publikationen bauen aufeinander auf und werden im ersten Hauptteil der vorliegenden Doktorarbeit adressiert:

- Koukal A., Gleue, C. und Breitner, M. H. (2013). „Enhancing Literature Research Processes: A Glance at an Approach Based on Latent Semantic Indexing“, veröffentlicht in den Proceedings der Gesellschaft für Informatik (GI) 2013, „Lecture Notes in Informatics (LNI)“,
- Koukal A., Gleue, C. und Breitner, M. H. (2014). „Enhancing Literature Review Methods - towards more efficient Literature Research with Latent Semantic Indexing“, veröffentlicht in den Proceedings der 22. European Conference on Information Systems (ECIS 2014),
- Koukal A., Gleue, C. und Breitner, M. H. (2014). „Enhancing Literature Review Methods: Evaluation of a Literature Search Approach based on Latent Semantic Indexing“, veröffentlicht in den Proceedings der International Conference on Information Systems (ICIS 2014).

Die übergeordnete Motivation des in Kapitel 2 vorgestellten Forschungsprojekts ist es, den Literaturrechercheprozess als eine der zeitintensivsten, wiederkehrenden Aufgaben für Wissenschaftler jeden Fachgebiets mithilfe von Methoden aus dem Gebiet des Text Mining und der semantischen Suche zu optimieren (siehe Abbildung 1). Die dabei entstehende Zeitersparnis kann somit für andere Forschungstätigkeiten aufgewendet werden und, neben einer Vervollständigung des theoretischen Fundaments, insgesamt eine Qualitätsverbesserung der Forschungsarbeiten bei gleichbleibendem Zeitaufwand ermöglichen.

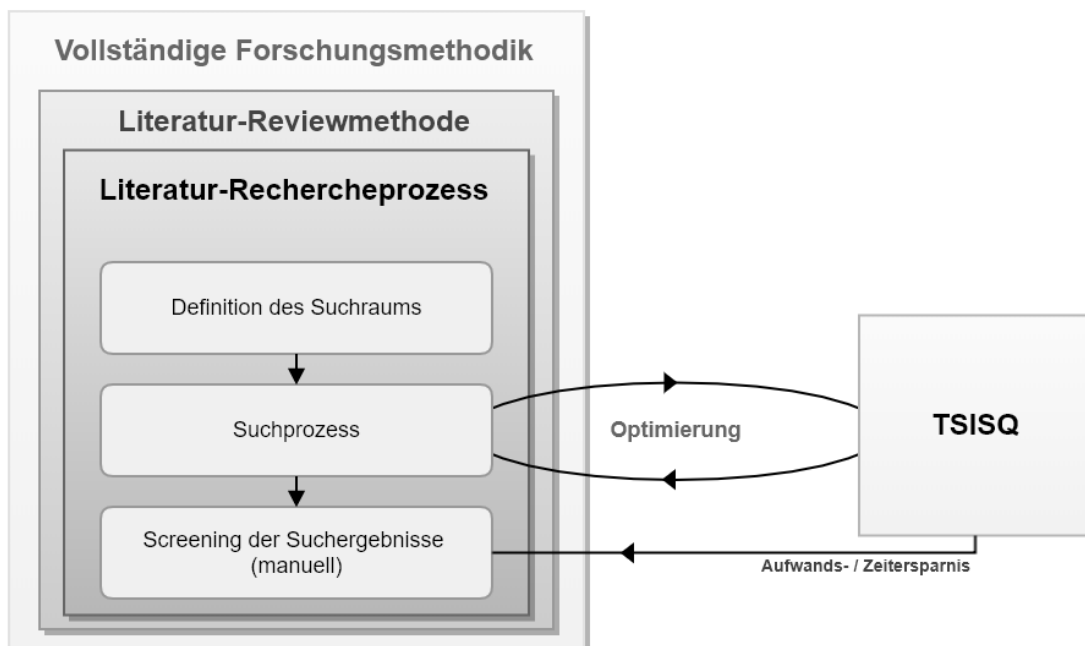


Abbildung 1: Optimierung des Literaturrechercheprozesses mit TSISQ

Die fehlende Automatisierung von Volltextsuchen in Kombination mit einer mangelnden Berücksichtigung semantischer Konzepte bzw. semantischer Ähnlichkeiten trägt wesentlich dazu bei, dass die Literaturrecherche im Rahmen des wissenschaftlichen Arbeitens ein langsamer, zeitintensiver (weil größtenteils manueller) Prozess ist. Um hier eine optimale technologische Unterstützung zu gewährleisten, bedarf es eines effizienteren, toolgestützten Ansatzes, der den manuellen Anteil an der Literaturrecherche, beispielsweise beim Screening der initialen Suchergebnisse, auf ein Minimum reduziert. Diese Idee führte letztlich zu der Entwicklung von TSISQ.

Stichwortbasierte Suchverfahren bringen trotz ihrer einfachen Realisierbarkeit Nachteile mit sich (vgl. Blair und Maron, 1985; Homayouni et al., 2004; LaBrie und St. Louis, 2003) und sind daher alles andere als optimal (vgl. Dumais et al., 1988): Ambiguität von Suchbegriffen, Synonymie, Polysemie, die unsachgemäße Anwendung von Stoppwörtern wie „die“, „in“ oder „und“, Bindestriche, Klammern, Pluralformen sowie letztlich die auftretende Unschärfe und Inkonsistenz bei der Indizierung des zu durchsuchenden Textmaterials führen schnell zu einer Verzerrung der Suchergebnisse. Folglich führen stichwortbasierte Ansätze sehr wahrscheinlich zu falsch-positiven oder falsch-negativen Suchergebnissen. Anders ausgedrückt werden potenzielle Treffer nicht gefunden oder schlicht unzutreffende Suchergebnisse geliefert (vgl. Blair und Maron, 1985; Dumais et al., 1988; Hofmann, 1999; LaBrie und St. Louis, 2003; Salton und McGill, 1986; Yandell und Majoros, 2002).

Trotz dieser Defizite sind rein stichwortbasierte Suchverfahren auch heute noch bei einem Großteil der wissenschaftlichen Datenbanken üblich (vgl. Cui et al., 2003). Daher wurden alle für die Wirtschaftsinformatik relevanten wissenschaftlichen Datenbanken wie AISel, IEEE, JStor, ScienceDirect oder Wiley untersucht und dabei festgestellt, dass sich das beschriebene Fehlverhalten in den Suchergebnissen widerspiegelt, was darauf schließen lässt, dass eine Reduzierung oder gar Beseitigung der zuvor genannten Defizite stichwortbasierter Suchverfahren einen wichtigen Beitrag darstellt. Um einer Lösung für die hier umrissene Problemstellung näherzukommen ist es das Ziel unserer Forschung zu diesem Thema, einen alternativen, modernen und intelligenteren Ansatz zur automatisierten Ähnlichkeitsanalyse von Texten zu entwickeln, der die folgenden Kriterien erfüllt:

- Verlässliche Identifikation wissenschaftlicher Arbeiten, die einem bestimmten Forschungsfeld angehören (Vorhandensein der initialen Suchbegriffe).

- Identifikation semantisch ähnlicher Arbeiten aus demselben oder sogar aus anderen Forschungsgebieten (Vorhandensein von Synonymen und/oder mit den initialen Suchbegriffen „sinnhaft verwandter“, also semantisch ähnlicher Begriffe).
- Die Überwindung des Problems mit Synonymie und Polysemie von Suchbegriffen und somit die größtmögliche Vermeidung falsch-positiver und falsch-negativer Suchergebnisse.
- Die Möglichkeit, Suchanfragen in natürlicher Sprache zu formulieren (Verwendung von Stichwörtern, vollständigen Sätzen, Abstracts und ganzen Forschungsarbeiten als Suchanfrage).

Das zu diesem Zweck entwickelte System ist schematisch in Abbildung 2 dargestellt. Zunächst wird dazu eine beliebig große Zahl wissenschaftlicher Arbeiten im PDF- oder Textformat eingelesen. Stark vereinfacht ausgedrückt wird anschließend mithilfe eines Verfahrens aus dem Fachgebiet des Information Retrieval, „Latent Semantic Indexing“ (LSI), eine numerische Repräsentation der Volltexte als Vektoren in einem Vektorraum erzeugt und einem Gesamtindex hinzugefügt, der als Basis für zukünftige Suchanfragen dient. Nach erfolgtem Training des Gesamtindex können einzelne Suchanfragen auf die gleiche Weise transformiert und die resultierenden Vektoren mit denen innerhalb des Index verglichen werden.

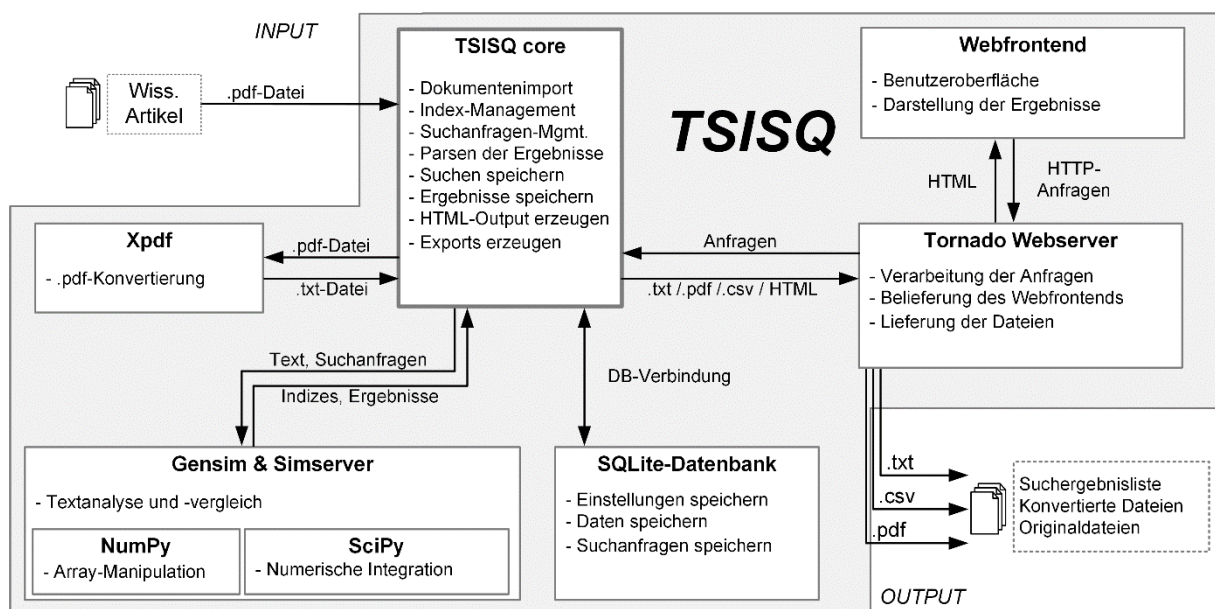


Abbildung 2: Tool for Semantic Indexing and Similary Queries

Abhängig von dem Winkel, in dem die Vektoren letztlich zueinander stehen, kann eine Aussage über die semantische Ähnlichkeit der durch sie repräsentierten Texte getroffen werden: Je größer der Winkel ist, desto größer ist auch die semantische Abweichung. Suchanfragen an TSISQ können in natürlicher Sprache über das dazugehörige Webfrontend gestellt werden,

sowohl als Upload im PDF- oder TXT-Format als auch per Volltexteingabe in ein Textfeld. Sobald der Gesamtindex einmal trainiert ist erscheint bei Klick auf den „Start Query“-Button (vgl. Abbildung 13) ohne nennenswerte Zeitverzögerung die Suchergebnisliste mit den als semantisch ähnlich identifizierten Artikeln, in absteigender Reihenfolge bzgl. ihrer Ähnlichkeit zur Suchanfrage (vgl. Abbildung 14). Mit dem Ziel, Eignung und Nutzen von TSISQ für den Literaturrechercheprozess zu bewerten und messbar zu machen, wurden in mehreren Anwendungsfällen zweierlei Tests durchgeführt:

1. Ein Benchmark zu den Ergebnissen einer stichwortbasierten Suchmaschine (mit Unterstützung von Fachexperten) auf dem 12.332 Paper umfassenden Index aus den führenden Zeitschriften in der Wirtschaftsinformatik, des sog. „AIS Basket of Eight“ und den renommiertesten Konferenzen. Auf diesem Wege wurden ein Vergleich und eine Bewertung der falsch-positiven Fehlerraten der verschiedenen Suchverfahren durchgeführt.
2. Ein Laborexperiment in einer kontrollierten Umgebung, das auf eine Bemessung der falsch-positiven und (insbesondere) der falsch-negativen Fehlerrate abzielt.

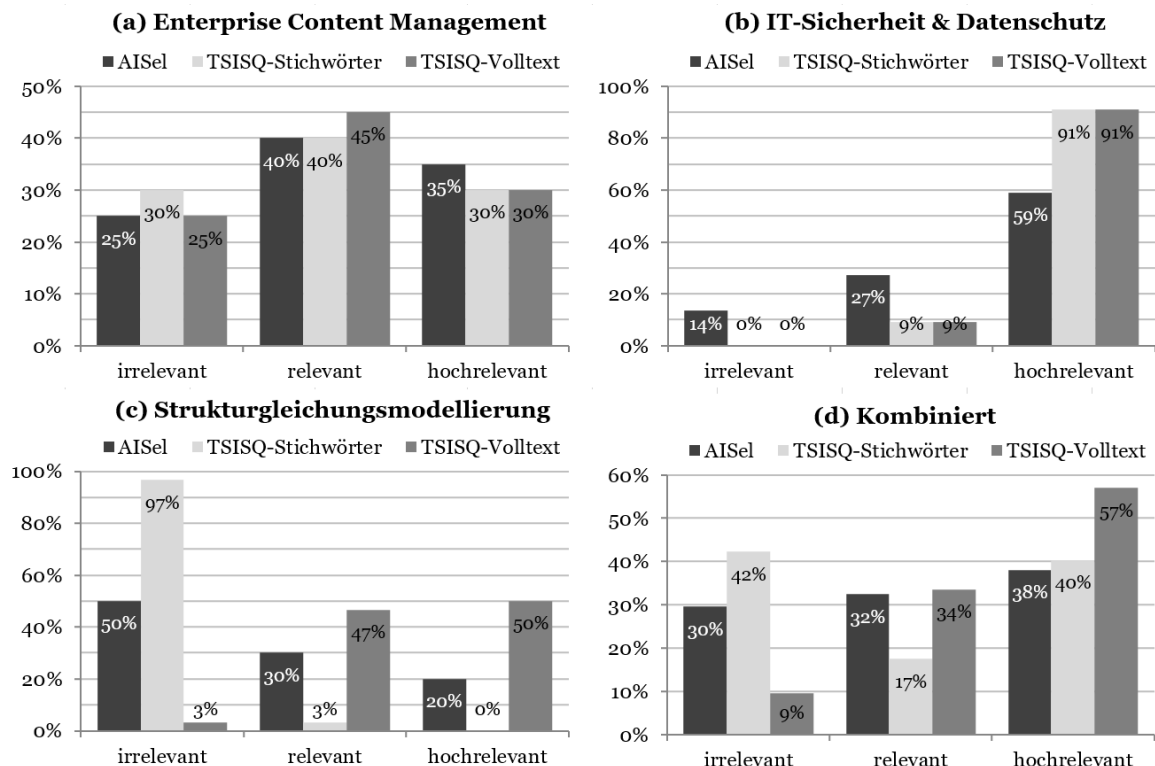


Abbildung 3: Ergebnisse des Benchmarks von TSISQ gegenüber Schlagwortsuchen

Zu 1.: Der Benchmark bezieht sich auf eine durch Fachexperten in den jeweiligen Gebieten „Enterprise Content Management“ (ECM), „IT-Sicherheit und Datenschutz“ und „Strukturgleichungsmodellierung“ (SEM) durchgeführte Bewertung der Relevanz der

Suchergebnisse auf dem großen Gesamtindex in „irrelevant“, „relevant“ und „hoch relevant“. Es wurden zudem geeignete Suchanfragen formuliert, um die den entsprechenden Fachgebieten zugehörigen Artikel im Gesamtindex zu identifizieren.

Zu 2.: Das Laborexperiment zielt darauf ab, die falsch-positive und insbesondere die falsch-negative Fehlerrate von TSISQ zu bemessen. Es soll demonstriert werden, wie viele den o.g. Themengebieten zugehörige Publikationen durch TSISQ „übersehen“ werden und wie viele Suchergebnisse fälschlicherweise als dazu passend identifiziert wurden. Dazu wurde ein deutlich kleinerer, kontrollierbarer Index aus 100 zufällig ausgewählten Veröffentlichungen manuell erstellt und anschließend deren Inhalte gemäß der unter (i) vorgestellten Kategorisierung für das Thema „IT-Sicherheit und Datenschutz“ bewertet. TSISQ lieferte ein Suchergebnis von insgesamt 23 Ergebnissen, obwohl im Gesamtindex nur 18 Artikel von 100 zur Zieldomäne gehörten. Die 23 gefundenen Ergebnisse enthielten dabei 15 (mindestens) relevante und 8 irrelevante Artikel. Folglich wurden 3 der 18 insgesamt relevanten Artikel nicht identifiziert, was einer falsch-negativen Fehlerquote von 17 Prozent entspricht. Dabei waren die ersten fünf Suchergebnisse mit einer Trefferquote von 100 Prozent am genauesten. Während im Bereich der Top 17-Ergebnisse die durchschnittliche falsch-positive Fehlerrate mit weniger als 20 Prozent immer noch auf einem akzeptablen Niveau liegt, erhöht sich diese Rate auf 35 Prozent, wenn alle 23 Suchergebnisse berücksichtigt werden (vgl. Abbildung 4).

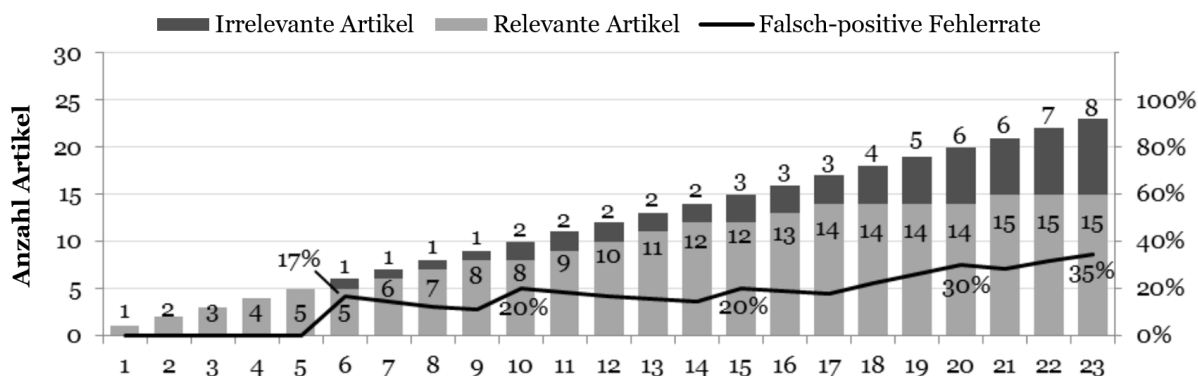


Abbildung 4: Ergebnisse des Laborexperiments

Das Laborexperiment erlaubt eine zusätzliche Quantifizierung der falsch-positiven und falsch-negativen Fehlerrate, deren Reduktion eine weitere der zu Beginn geforderten Eigenschaften darstellt. In dieser kontrollierten Umgebung beträgt die durchschnittliche Suchgenauigkeit mindestens 65% innerhalb der oberen 23 Ergebnisse. Innerhalb der ersten 17 Ergebnisse liegt die durchschnittliche Suchgenauigkeit jedoch nie unter 80%. Folglich schneidet TSISQ etwas schlechter ab als in der Pilotstudie (vgl. Koukal et al., 2013) (75%), bestätigt aber dennoch die insgesamt bemerkenswerten Ergebnisse.

Die Ergebnisse dieser Studien zeigen insgesamt, dass TSISQ sehr gut dazu geeignet ist, wissenschaftliche Literaturrechercheprozesse zu verbessern, insbesondere bei der Verwendung von Volltext in natürlicher Sprache mit inhaltlichem Bezug zur Zieldomäne als Input. In allen Suchszenarien generiert die TSISQ-Volltextsuche qualitativ mindestens gleichwertige, häufig auch höherwertige Ergebnisse hinsichtlich der Relevanz der gefundenen Artikel. TSISQ eignet sich zudem, wie zu Beginn gefordert, am besten für die Formulierung von Suchanfragen in natürlicher Sprache. Zum Beispiel liefert ein vollständiger Abstract als Suchanfrage bei AISel, unserem Benchmark für eine stichwortbasierte Suchmaschine, entweder keine Ergebnisse (bei der Anwendung von „AND“-Bedingungen), oder die Ergebnisse bestehen (bei der Anwendung von „OR“-Bedingungen) aus einer völlig zufälligen Auswahl an Dokumenten aus den verschiedensten Forschungsgebieten, während TSISQ (mit den o.g. Einschränkungen) sinnvolle Ergebnisse liefert.

Bei LSI, dem Verfahren auf dem TSISQ basiert, handelt es sich um eine Methode aus dem Gebiet des Information Retrieval zur Identifikation semantischer Ähnlichkeiten, es ist hauptsächlich für die Suche nach semantischen Konzepten gedacht. Daher kann die klare Empfehlung gegeben werden, statt Stichwörtern fachgebietsspezifische Volltexte in natürlicher Sprache für die Suchanfragen zu verwenden. Aus der Erkenntnis, dass die Ergebnisse bei z.B. Abstracts als Suchanfrage thematisch spezifischer sind, während sie bei Verwendung von vollständigen Forschungsarbeiten diffuser bzw. weniger fokussiert ausfallen, folgt für ein optimales Ergebnis die Empfehlung, einen iterativen, zyklischen Prozess bei der Literaturrecherche einzuhalten. Die erste Suchanfrage sollte aus einem vollständigen Artikel des gewünschten Themengebiets bestehen, um eine semantisch „unschärfere“ Auswahl ähnlicher Paper als Ergebnis zu erhalten. Anschließend folgt eine manuelle Überprüfung der Ergebnisse zur Identifikation der relevantesten Artikel für eine spezifischere thematische Eingrenzung. Der nächste Schritt umfasst das Erstellen einer Sammlung prägnanter Textbausteine, Sätze und/oder Abschnitte, die in der vorab identifizierten Literaturlauswahl enthalten sind. Im letzten Schritt wird die daraus resultierende Sammlung als Suchanfrage verwendet. Der Prozess sollte zyklisch wiederholt werden, bis keine neue, relevante Literatur mehr gefunden wird.

In Bezug auf das übergeordnete Forschungsziel, mit TSISQ die Effizienz wissenschaftlicher Literaturrechercheprozesse mithilfe von Methoden des Data Mining und Big Data Analytics zu erhöhen, können folgende Schlüsse gezogen werden:

- Die Anwendung des hier dargestellten Ansatzes hilft dabei, wertvolle Zeit beim Auffinden der relevanten Literatur zu einem beliebigen Forschungsfeld einzusparen.
- Auch wenn genügend Zeit vorhanden ist kann TSISQ dennoch dazu beitragen, die Vollständigkeit und die Effizienz von Literaturrecherchen zu steigern: Zum einen durch die im Vergleich zu menschlichen Fähigkeiten deutlich höhere Arbeitsgeschwindigkeit und zum anderen durch die Identifikation von Quellen, die andernfalls möglicherweise nicht gefunden oder berücksichtigt worden wären.
- Die im Rahmen der Literaturrecherche eingesparte Zeit kann für eine gründlichere Durchführung anderer Forschungsaufgaben aufgewendet und damit die Qualität der Forschung insgesamt erhöht werden.

Abschließend kann festgehalten werden, dass der im Rahmen dieses Projektes entwickelte Ansatz zur Optimierung der wissenschaftlichen Literaturrecherche eine nützliche Ergänzung zu den etablierten Suchmaschinen und -methoden in der Wissenschaft darstellt. Er kann im Rahmen jeder umfassenderen, etablierten Forschungsmethodik zum Einsatz kommen und dabei deutliche Effizienzgewinne bei der Literaturrecherche erzielen.

Operative Entscheidungsunterstützung in der Automobilindustrie mit Künstlicher Intelligenz:

In zweiten Hauptteil der vorliegenden Dissertation (Kapitel 3) werden die Ergebnisse und Erkenntnisse aus insgesamt drei Forschungsarbeiten zusammengefasst, die dem oben genannten Themenkomplex der Entscheidungsunterstützung in der Automobilindustrie durch Data Mining, Big Data Analytics, Business Intelligence und KI bzw. KNN angehören. Diese Veröffentlichungen sind das theoretische Fundament eines 4-jährigen Kooperationsprojektes mit einem großen deutschen Automobilhersteller. Dabei ist eine Sammlung an (Software-) Tools zur Entscheidungsunterstützung im Restwert- und Risikomanagement entstanden. Auch diese Veröffentlichungen bauen aufeinander auf. Die folgenden Themen und Publikationen werden in Teil B der vorliegenden Doktorarbeit adressiert:

- Gleue, C., Eilers, D., von Mettenheim, H.-J. und Breitner, M. H. (2017). “Decision Support for the Automotive Industry: Forecasting Residual Values Using Artificial Neural Networks”, veröffentlicht in den Proceedings der Wirtschaftsinformatik-Jahrestagung (WI 2017).
- Gleue, C., Eilers, D., von Mettenheim, H.-J. und Breitner, M. H. (2018). “Decision Support for the Automotive Industry: Forecasting Residual Values Using Artificial Neural Networks”, Veröffentlichung voraussichtlich in der Business & Information Systems Engineering in 2019, Online-Vorveröffentlichung über Springer (2018).

- Eilers, D., Köpp, C., Gleue, C. und Breitner, M. H. (2017). "It's Not a Bug, It's a Feature: How Visual Model Evaluation Can Help to Incorporate Human Domain Knowledge in Data Science", veröffentlicht in den Proceedings der International Conference on Information Systems (ICIS 2017).

Die Herausforderung eines großen deutschen Automobilherstellers bestand vor einigen Jahren darin, bereits bei Abschluss von Leasingverträgen eine möglichst genaue Vorstellung über den aktuellen Marktwert und den zukünftigen Restwert seiner Fahrzeuge zu entwickeln. Diese Ausgangslage motivierte uns, das im Folgenden dargestellte Projekt gemeinsam mit dem Hersteller durchzuführen, mit dem Ziel, Risiken, Rückstellungen, Leasingraten und -angebote genauer berechnen und dementsprechend besser steuern zu können.

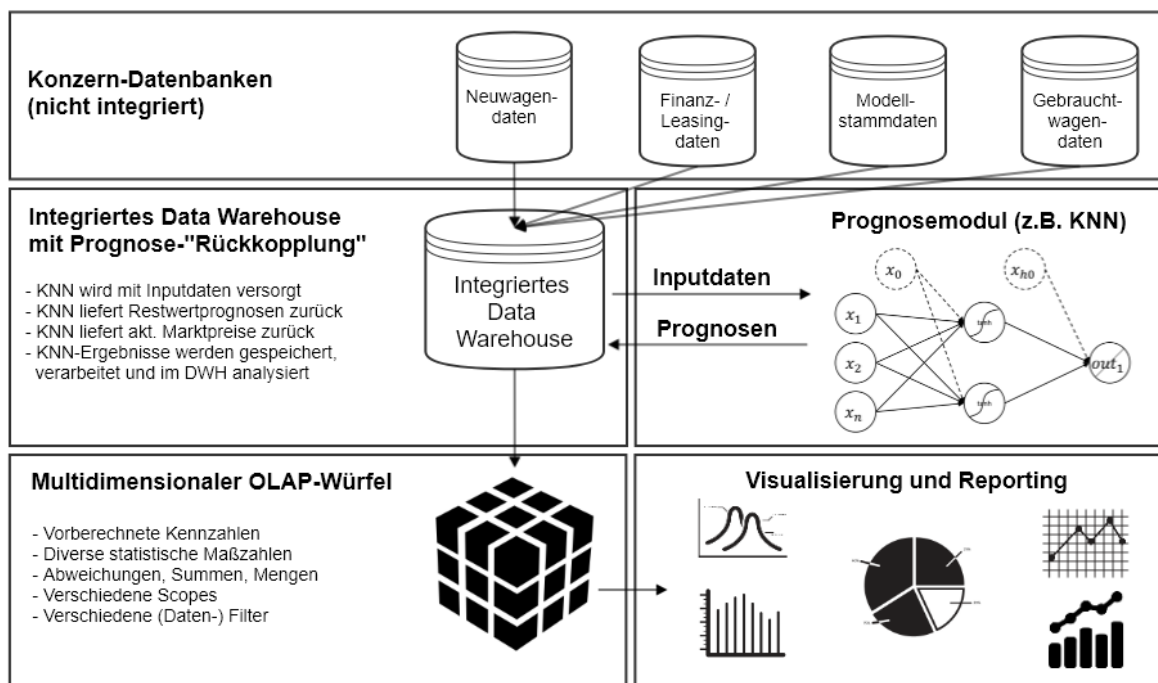


Abbildung 5: EUS für Markt- und Restwertprognosen im Leasinggeschäft

Bereits Jahre zuvor hatte der Konzern sich bereits darum bemüht, die dafür notwendige Datenbasis zu schaffen, indem allen Vertragshändlern starke finanzielle Anreize geboten wurden, alle ihre Gebrauchtfahrzeugverkäufe inkl. Fahrzeuginformationen zu erfassen und die entsprechenden Transaktionsdaten zu historisieren. Diese Daten galt es nun mithilfe der neuesten Erkenntnisse und Methoden aus den Gebieten der Entscheidungsunterstützung, Business Intelligence, Data Mining, Künstlicher Intelligenz und Big Data Analytics zielorientiert mit Daten aus zahlreichen weiteren Konzernsystemen zu integrieren, sie zu bereinigen, zu analysieren und entsprechend aufzubereiten. Nach einer ersten „Proof of Concept“-Phase und einer knapp 4-jährigen Projektlaufzeit unter Beteiligung diverser

Fachabteilungen, dem konzerneigenen BI-Kompetenzcenter und dem Institut für Wirtschaftsinformatik ist so ein intelligentes EUS entstanden, das in Abbildung 5 schematisch dargestellt ist.

Die Ergebnisse dieses Projekts lassen sich zusammenfassen wie folgt:

1. Wir stellen den exemplarischen Aufbau des im Projekt entstandenen EUS mit dem zuvor beschriebenen Zweck vor (vgl. Abbildung 5).
2. Der Einfluss der uns zur Verfügung stehenden Inputfaktoren auf die Markt- und Restwerte wurde untersucht und zahlreiche Erkenntnisse daraus abgeleitet.
3. Wir beschreiben unseren Ansatz zur Verwendung von KNN für eine Prognoseanwendung in diesem Anwendungsgebiet.

Darüber hinaus mussten verschiedenste, charakteristische Herausforderungen bewältigt werden: Der Umgang mit kurz- und insbesondere langfristigen Prognosehorizonten, die für genaue Ergebnisse notwendige, mehrstufige Methodik zur Bereinigung verrauschter Daten „aus der realen Welt“ sowie die Behandlung von zeit- und konjunkturabhängigen Variablen.

Eines der zunächst insbesondere von Fachexperten eher unerwarteten, empirischen Ergebnisse ist der Nachweis, dass potenzielle externe Einflussfaktoren wie Öl- bzw. Kraftstoffpreise, diverse makroökonomische Indizes wie DAX, der Ifo-Geschäftsklimaindex oder auch das BIP, der EZB-Leitzins oder die Arbeitslosenrate keinen signifikanten Einfluss auf die Markt- und Restwerte ausüben. Im Gegensatz dazu wurden die internen, vom Hersteller selbst beeinflussbaren Faktoren wie Fahrzeugalter und Kilometerleistung zum Zeitpunkt der Wiedervermarktung, Ausstattung des Fahrzeugs, Motorisierung oder Getriebeart als maßgebliche Einflussfaktoren identifiziert, was abermals den Wert eines solchen internen „Datenschatzes“ unterstreicht. Die letztlich aus unserer Forschung ableitbaren Erkenntnisse sind für den Hersteller besonders wertvoll, da auf deren Basis bereits frühzeitig, durch eine entsprechende Steuerung der Leasingangebote und damit des späteren Leasingportfolios¹, Fahrzeuge so konfiguriert werden können, dass sie sich zu einem späteren Zeitpunkt am Markt möglichst restwertstabil verhalten.

Weitere Erkenntnisse gab es auch bzgl. des Einflusses zeitbezogener Faktoren auf die Prognosegenauigkeit. Anhand des unseren Analysen zugrundeliegenden Datensatzes, der in dieser Form in der Forschung einzigartig ist, konnten wir nachweisen, dass die Marktpreise von Gebrauchtfahrzeugen sowohl saisonalen Effekten als auch einem über die Zeit sinkenden Trend

¹ Anm. des Verfassers: Nach Möglichkeit wäre sogar eine auf Restwertstabilität gesteuerte Produktion denkbar.

unterliegen, der, nachdem der preissenkende Effekt eines mit der Zeit stetig steigenden Modellalters bereits in der Modellbildung berücksichtigt wurde, auf die gesamtwirtschaftliche Situation im Betrachtungszeitraum zurückzuführen ist.

Bzgl. der Saisonalitäten konnten zum Ende eines Jahres generell niedrigere Marktwerte als zu Beginn beobachtet werden (vgl. Abbildung 6).

Es wurde zudem in einem Benchmark gezeigt, dass die letztlich auch operativ implementierte, auf KNN basierende Prognoseanwendung leistungsfähiger ist und deutlich genauere Ergebnisse liefert als lineare Modelle (Ridge Regression). Eine der großen Herausforderungen bei der technischen Umsetzung war dabei die Berücksichtigung des Faktors „Zeit“, bzw. der Variablen mit einem zeitlichen Bezug. Die letztlich von uns favorisierte Lösung besteht aus einer Kombination von KNN, um die nicht-linearen Beziehungen innerhalb des Basisdatensatzes abbilden zu können und einer linearen, abwärts gerichteten Adjustierung der Ergebnisse abhängig von dem gewünschten Prognosehorizont. Unsere Tests haben dabei gezeigt, dass diese lineare Anpassung der KNN-Ergebnisse (bis zu 1,5 Jahre in die Zukunft) keine nennenswerte Verzerrung verursacht und im Anschluss zumindest im Vergleich zu anderen von uns getesteten Methoden eine „Best Practice“ darstellt.

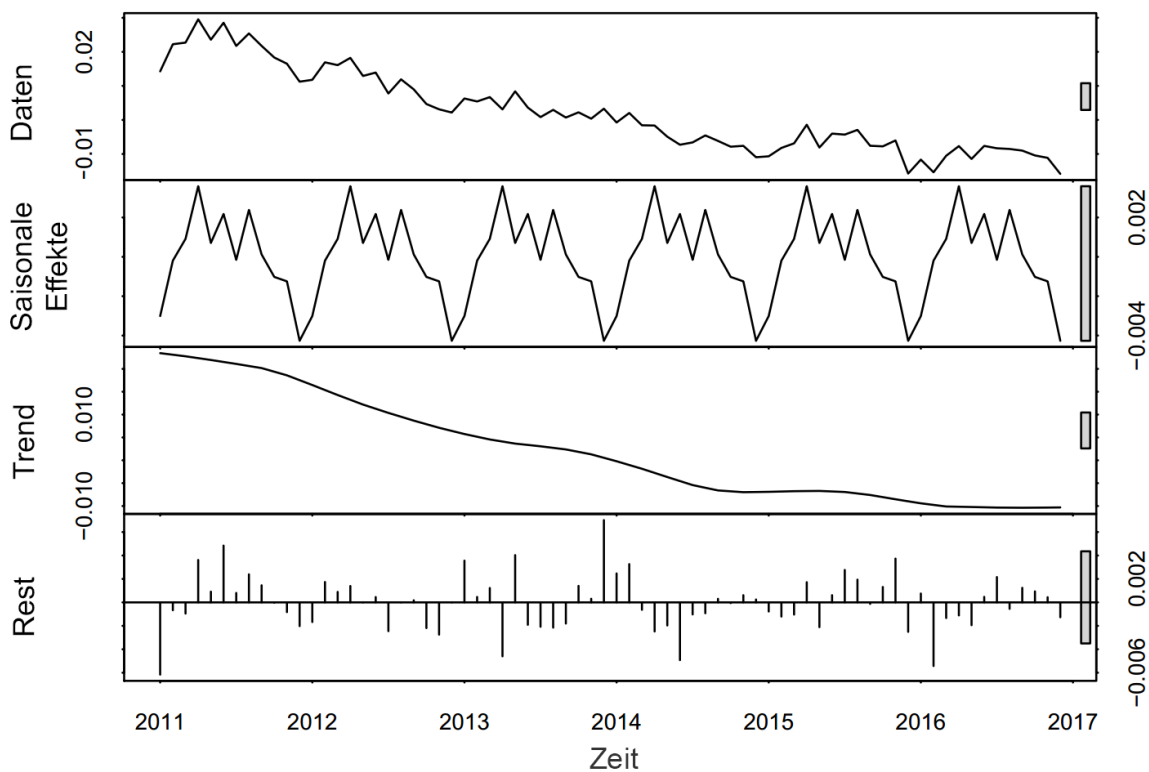


Abbildung 6: Zeitreihenzerlegung: Saisonale Effekte und Trends

Es wurde zudem untersucht, wie ein auf dieser Technologie und Methodik basierendes EUS in der wirtschaftlichen Praxis operationalisiert werden kann. Die gewonnenen Erkenntnisse und

Prozessoptimierungen wurden inzwischen bei dem Automobilhersteller mit Erfolg implementiert. Neben der Schaffung von tagesaktueller Transparenz am Gebrauchtwagenmarkt stand dabei die möglichst genaue Vorhersage der Zukunftsrestwerte im Vordergrund, wobei das Prognosemodul letztlich den entscheidenden Bestandteil des insgesamt deutlich umfangreicheren EUS ausmacht. Die mithilfe dieses EUS geschaffene Markttransparenz und die nachweislich deutlich verbesserte Prognosequalität sind ausschlaggebend für das Risikomanagement, denn eine systematische Verzerrung der kalkulierten Markt- und Restwerte in beiden Richtungen (über- oder unterschätzte Restwerte) bringt kurz- mittel- und langfristig negative Konsequenzen für Händler und Hersteller mit sich, beeinträchtigt die Wettbewerbsfähigkeit bzw. die Höhe der notwendigen Verkaufshilfen für den Vertrieb und/oder hat einen maßgeblichen Einfluss auf die Handelsmargen bei Wiederverkauf der Fahrzeuge. Somit ist eine bedarfsgerechte Rückstellungsbildung und marktgerechte Preisgestaltung im Leasinggeschäft von gleichermaßen hoher (wirtschaftlicher und strategischer) Bedeutung für Automobilhersteller und -händler.

Das Hauptziel dieses Projektes, die Etablierung eines rein datengetriebenen Ansatzes zur Restwertprognose von Leasingfahrzeugen, wurde erreicht und führte durch enge Zusammenarbeit zwischen dem Institut für Wirtschaftsinformatik, dem Fachbereich Finanz und dem Entwicklungsteam des Business Intelligence Competence Centers (BICC) des Automobilherstellers zu gesteigerter (Technologie-) Akzeptanz und Vertrauen innerhalb des Konzerns. Die Ergebnisse zeigen, dass KNN für die in der Praxis häufig verrauschten und unstrukturierten Daten gut geeignet sind. Dennoch ist eine gründliche Datenbereinigung und -aufbereitung unabdingbar, um gute und verlässliche Ergebnisse zu erzielen. Es wurde gezeigt, wie KNN zur Vorverarbeitung und Filterung der Datenbasis verwendet werden und wie ein Prognosemodell auf Basis eines KNN-Ansatzes als „Herzstück“ eines auf die Bedürfnisse der Finanzabteilung des Automobilherstellers maßgeschneiderten EUS entworfen werden kann.

III. Inhaltsverzeichnis

I. Abstract	II
II. Management Summary	IV
III. Inhaltsverzeichnis	XVI
IV. Abbildungsverzeichnis	XVIII
V. Tabellenverzeichnis	XVIII
VI. Abkürzungsverzeichnis	XIX
VII Überblick über Publikationen und Aufgabenverteilung	1
1. Einleitung	6
1.1 Problemstellung, Relevanz und Motivation	6
1.2 Aufbau der Arbeit und wissenschaftlicher Beitrag	7
2 Semantische Suche im Literaturrechercheprozess	9
2.1 Motivation und Zielsetzung	10
2.2 Literaturrecherche und Methoden des Information Retrieval	13
2.2.1 Stand der Forschung	13
2.2.2 Forschungsmethodik	19
2.3 Optimierung des Literaturrechercheprozesses mit TSISQ	21
2.3.1 Theoretischer Hintergrund und Funktionsweise	22
2.3.2 Implementierung, Architektur und Benutzeroberfläche	25
2.4 Demonstration, Evaluation und Benchmark von TSISQ	30
2.4.1 Benchmark mit einer stichwortbasierten Suchmaschine	30
2.4.2 Identifikation der falsch-positiven und falsch-negativen Fehlerrate	33
2.5 Kritische Würdigung und Diskussion der Ergebnisse	35
2.6 Fazit	41
3 Operative Entscheidungsunterstützung mit KI	44
3.1 Gesamtwirtschaftlicher Kontext, Motivation und Projektziele	45
3.1.1 Wirtschaftliche Bedeutung des Leasinggeschäfts für Automobilhersteller	46
3.1.2 Markt- und Restwertprognosen in der wissenschaftlichen Literatur	47

3.1.3 Projektziele	49
3.2 Gesamtsystem zur Entscheidungsunterstützung im Leasinggeschäft	51
3.2.1 Systemarchitektur und Datenquellen	52
3.2.2 Inputfaktoren für die Markt- und Restwertprognose	56
3.2.3 Markttrends und saisonale Effekte.....	57
3.2.4 Der Einfluss externer Faktoren auf Fahrzeugrestwerte.....	59
3.3 Das Prognosemodul als Kern des EUS.....	61
3.3.1 KNN für die Restwertprognose: Eignung und Funktionsweise.....	63
3.3.2 Erkennung von Ausreißern und Datenbereinigung.....	65
3.3.3 Vorhersagekraft der Inputfaktoren und Prognosequalität.....	66
3.3.4 Prognoseverfahren: Anwendungsszenarien und Benchmark.....	68
3.4 Kritische Würdigung und Diskussion der Ergebnisse	72
3.5 Fazit	76
4 Kritische Reflexion und Ausblick	78
5 Fazit	85
6 Literaturverzeichnis.....	86
7 Appendix	A
A: Enhancing Literature Research Processes: A Glance at an Approach Based on Latent Semantic Indexing	A
B: Enhancing Literature Review Methods - towards more efficient Literature Research with Latent Semantic Indexing	B
C: Enhancing Literature Review Methods - Evaluation of a Literature Search Approach based on Latent Semantic Indexing.....	C
D: Decision Support for the Automotive Industry: Forecasting Residual Values Using Artificial Neural Networks (WI)	D
E: It's Not a Bug, It's a Feature: How Visual Model Evaluation Can Help to Incorporate Human Domain Knowledge in Data Science	E
F: Decision Support for the Automotive Industry: Forecasting Residual Values Using Artificial Neural Networks (BISE).....	F

IV. Abbildungsverzeichnis

Abbildung 1: Optimierung des Literaturrechercheprozesses mit TSISQ.....	V
Abbildung 2: Tool for Semantic Indexing and Similary Queries	VII
Abbildung 3: Ergebnisse des Benchmarks von TSISQ gegenüber Schlagwortsuchen.....	VIII
Abbildung 4: Ergebnisse des Laborexperiments.....	IX
Abbildung 5: EUS für Markt- und Restwertprognosen im Leasinggeschäft	XII
Abbildung 6: Zeitreihenzerlegung: Saisonale Effekte und Trends	XIV
Abbildung 7: Aufbau der vorliegenden Dissertation	8
Abbildung 8: Forschungsdesign des DSR-Prozesses nach Peffers et al. (2007)	19
Abbildung 9: Optimierung des Literaturrechercheprozesses mit TSISQ.....	22
Abbildung 10: Reduzierte Singular Value Decomposition der Term-by-Document-Matrix...	24
Abbildung 11: TSISQ-Systemarchitektur	26
Abbildung 12: TSISQ Index-Management	27
Abbildung 13: Eingabemaske für neue Suchanfrage in TSISQ.....	28
Abbildung 14: TSISQ-Suchergebnisliste, absteigend sortiert nach Ähnlichkeits-Score	29
Abbildung 15: TSISQ-Suchhistorie	29
Abbildung 16: Benchmark-Ergebnisse zu Suchmethoden.....	32
Abbildung 17: Ergebnisse des Laborexperiments.....	35
Abbildung 18: EUS für Markt- und Restwertprognosen im Leasinggeschäft	51
Abbildung 19: Der Einfluss des Faktors "Zeit" auf die Restwerte.....	58
Abbildung 20: Zeitreihenzerlegung: Saisonale Effekte und Trends	59
Abbildung 21: Schematische Darstellung eines (3-schichtigen "feed-forward"-) KNN	65

V. Tabellenverzeichnis

Tabelle 1: Publikationsübersicht	5
Tabelle 2: Wiss. Beiträge zu praktischen und theoretischen Anwendungsfällen für LSI.....	16
Tabelle 3: Datenbasis aus wiss. Journal- und Konferenzbeiträgen (2007-2014).....	30
Tabelle 4: Abfragen zum Suchmaschinenvergleich in ausgewählten Forschungsgebieten	31
Tabelle 5: Input- und Output-Variablen des Prognosemodells	56
Tabelle 6: Einfluss externer (makroökonomischer) Inputfaktoren	61
Tabelle 7: Inputfaktoren/Features nach Relevanz	67
Tabelle 8: Prognosegenauigkeit für akt. Marktwerte: RMSE v. KNN u. Ridge-Regression... ..	70
Tabelle 9: Prognosegenauigkeit für zuk. Marktwerte: RMSE v. KNN u. Ridge Regression ..	70
Tabelle 10: Prognosen für einzelne Fahrzeuge: RMSE des KNN u. d. Ridge Regression	71

VI. Abkürzungsverzeichnis

AIS	Association for Information Systems
ANN	Artificial Neural Network (Künstliches Neuronales Netz)
AMCIS	Americas Conference on Information Systems
BI	Business Intelligence
BISE	Business & Information Systems Engineering
DB	Datenbank
DWH	Data Warehouse
DSR	Design Science Research
ECIS	European Conference on Information Systems
EJIS	European Journal of Information Systems
ECM	Enterprise Content Management
EUS	Entscheidungsunterstützungssystem (engl.: Decision Support System (DSS))
FF	Forschungsfrage
GI	Gesellschaft für Informatik
HICSS	Hawaii International Conference on System Sciences
HTML	Hypertext Markup Language, Sprache zur Strukturierung digitaler Dokumente
ICIS	International Conference on Information Systems
IR	Information Retrieval
IS	Information Systems
ISR	Information Systems Research
IT	Informationstechnologie
JAIS	Journal of the Association for Information Systems
JIS	Journal of Information Systems
KDD	Wissensentdeckung in DBen (engl.: „Knowledge Discovery in Databases“)
KNN	Künstliches Neuronales Netz
KI	Künstliche Intelligenz
kW	Kilowatt (physikalische Einheit für Leistung; SI: Watt)
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
MDM	Master Data Management System (dt.: Stammdatenmanagementsystem)
MISQ	Management Information Systems Quarterly
MLP	Multi-layer Perceptron
NLP	Natural Language Processing

OLAP	Online Analytical Processing
PDF	Portable Document Format
PS	Pferdestärken (veraltete physikalische Maßeinheit für Leistung, insbes. Verwendung im automobilen Kontext; SI: Watt)
QE	Query Expansion
RMSE	Root Mean Squared Error (Mittlerer quadratischer Fehler)
SEM	Structural Equation Modeling (Strukturgleichungsmodellierung)
SVD	Singular Value Decomposition
TFIDF	Term frequency-inverse document frequency
TSISQ	Tool for Semantic Indexing and Similarity Queries (Werkzeug für semantische Indizierung und Ähnlichkeitssuchen)
UPE	Unverbindliche Preisempfehlung (des Herstellers)
WI	Internationale Tagung Wirtschaftsinformatik
WKWI	Wissenschaftliche Kommission Wirtschaftsinformatik im Verband der Hochschullehrer für Betriebswirtschaft
VHB	Verband der Hochschullehrer für Betriebswirtschaft
VIN	Fahrzeug-Identifikationsnummer (engl. „Vehicle Identification Number“)
VSM	Vector Space Model (Vektorraummodell)

VII Überblick über Publikationen und Aufgabenverteilung

Diese Dissertation basiert auf 6 Publikationen in „peer-reviewed journals“ und Konferenzen. Die Veröffentlichungen wurden in den verschiedenen Tagungsbänden der ICIS 2014, der ECIS 2014, der Jahrestagung der GI 2013, der WI 2017, der ICIS 2017 sowie der BISE 2018 veröffentlicht. Insgesamt waren an diesen Veröffentlichungen fünf Co-Autoren beteiligt: Michael H. Breitner, Dennis Eilers, Christoph Gleue, André Koukal und Hans-Jörg von Mettenheim (in alphabetischer Reihenfolge).

Alle im Rahmen dieser Publikationen vorgestellten Themen zielen im Grunde darauf ab, mithilfe von Methoden des Data Mining, Big Data Analytics und KI neue Erkenntnisse aus großen Datenmengen zu gewinnen und dabei einen messbaren, praktischen Nutzen in der täglichen Arbeit zu entfalten; sowohl in einem wissenschaftlichen als auch (und insbesondere) in einem rein wirtschaftlich geprägten Umfeld.

Tabelle 1 liefert im Anschluss an dieses Kapitel eine Übersicht über diese Veröffentlichungen. Zur deren Beurteilung wurden drei verschiedene Bewertungsmaßstäbe angelegt: Die WI-Orientierungslisten 2008 der Wissenschaftlichen Kommission für Wirtschaftsinformatik (WKWI), das JOURQUAL3-Ranking des Verbandes der Hochschullehrer für Betriebswirtschaft (VHB) von 2015 und der „Impact Factor“ (IF) von Thomson Reuters.

Alle Publikationen wurden von zwei verschiedenen Autorentams verfasst. Das Paper „Enhancing Literature Research Processes: A Glance at an Approach Based on Latent Semantic Indexing“ (Koukal et al., 2013) ist die erste in einer Reihe von drei Veröffentlichungen und stellt den ersten Prototyp unseres „Tools for Semantic Indexing and Similarity Queries“ (TSISQ) vor, der die Grundlage für die gesamte Reihe an Veröffentlichungen zu diesem Thema bildet. Dieser Prototyp wurde von mir in Python entwickelt und stellt den Kern aller weiteren Forschungsschritte und Erweiterungen zu diesem Thema dar. Die Idee, ihn für die „teilautomatisierte Literaturrecherche“ zu verwenden stammt von André Koukal und mir zu gleichen Teilen, ebenso wie die in dieser ersten Veröffentlichung vorgestellten Anwendungsfälle zur Evaluation meines Prototyps.

Das Paper „Enhancing Literature Review Methods – towards more efficient Literature Research with Latent Semantic Indexing“ (Koukal et al., 2014) baut auf den Erkenntnissen der letztgenannten Arbeit auf. Der von mir entwickelte TSISQ-Prototyp (als Kernelement) wurde von Herrn Koukal und mir gemeinsam um zahlreiche Komponenten erweitert, um den Automatisierungsgrad zu erhöhen und ihn für Kollegen und Leser unserer Veröffentlichung

einfacher nutzbar zu machen. Das in Abschnitt 2.3.2 dargestellte Webfrontend wurde von Herrn Koukal implementiert. Der Forschungshintergrund, verwandte Arbeiten und die Darstellung unserer Methodik wurden von uns gemeinsam gründlich überarbeitet, ebenso wie die neu hinzugekommenen Anwendungsfälle und Evaluationen, die wir gemeinsam konzipiert, durchgeführt und ausgewertet haben.

Bei dem Titel „Enhancing Literature Review Methods - Evaluation of a Literature Search Approach based on Latent Semantic Indexing“ handelt es sich um die letzte Ausbaustufe unserer Forschung zu diesem Thema, der Fokus liegt hier zum einen auf einer Erweiterung der Datenbasis und zum anderen auf der Evaluation der von uns entwickelten Vorgehensweise. Die Idee, einen handverlesenen Index als „kontrollierte Umgebung“ zu verwenden, um nicht nur die falsch-positive, sondern auch die falsch-negative Fehlerrate bewerten zu können, entstand in der stetigen Diskussion zwischen Herrn Koukal und mir während der gemeinsamen Arbeit an der Veröffentlichung, wie auch der Benchmark zur stichwortbasierten Suche.

Das Paper "Decision Support for the Automotive Industry: Forecasting Residual Values Using Artificial Neural Networks" (Gleue et al., 2017), das im Tagungsband der Internationalen Tagung Wirtschaftsinformatik 2017 veröffentlicht wurde, basiert auf der Implementierung eines Entscheidungsunterstützungssystems, dessen Kern eine KNN-Prognoseanwendung für die Vorhersage von Markt- und Restwerten von Gebrauchtfahrzeugen darstellt. Die Grundidee bzgl. Architektur, Vorgehensweise, Projektstruktur, Aufbau des Gesamtsystems, Integration der einzelnen Komponenten und Auswahl der Tools stammt von mir. Die gesamte Fachlichkeit (Leasinggeschäft, Finanzen, Vertrieb, Geschäftsprozesse), die Analyse und Definition von Rahmenbedingungen für die praktische Anwendbarkeit des Systems, die dazu notwendige Kommunikation und Abstimmung mit diversen Fachbereichen (Finanz, Vertrieb, IT, Business Intelligence, Herstellerbank etc.) des Automobilherstellers sowie die Einbettung des Systems in die bestehenden Geschäftsprozesse wurde ebenfalls von mir übernommen. Zu meinen Aufgaben im Projekt gehörte zudem die gesamte Anforderungsdefinition und Spezifikation des Zielszenarios, Konzeption und Implementierung eines dedizierten DWH inkl. der Beschaffung und Integration aller im Projekt verwendeten Daten inklusive deren Aufbereitung, Bereinigung, Strukturierung (nach technischen und fachlichen Vorgaben) und die Konzeption und Umsetzung „bedarfsgerechter“ Visualisierungen. Als Schnittstelle zwischen dem Institut für Wirtschaftsinformatik, dem Finanzbereich und dem Projektteam in der IT des Automobilherstellers war ich verantwortlich für das Projektmanagement, die Kommunikation unserer Ergebnisse, die Aufnahme und „Übersetzung“ von Anforderungen sowie für die Gesamtkonzeption des Projekts. Die Idee, KNN für die angeforderten Prognosen zu verwenden

entstand gemeinsam mit Dennis Eilers, Prof. von Mettenheim und Prof. Breitner, wobei diese mit Ihrem umfangreichen Wissen und Erfahrungsschatz bei der praktischen Anwendung von KNN einen wesentlichen Anteil zum Projekt- und Publikationserfolg beigetragen haben. Dennis Eilers hat dabei die Prognoseanwendung entwickelt, inkl. des Benchmarks mit dem linearen Modell und der empirischen Studie bzgl. des Einflusses des Zeitfaktors. Prof. Breitner und Prof. von Mettenheim haben uns maßgeblich bei der Konzeption der nachgelagerten „Ausreißerererkennung“ und -bereinigung, der Aufteilung des Datensatzes in Trainings- und Validierungsmengen und der Optimierung der Künstlichen Neuronalen Netze unterstützt. Die technische bzw. mathematisch/statistische Interpretation der Ergebnisse wurde von Dennis Eilers durchgeführt, während die fachliche Übersetzung bzw. die Beurteilung der Implikationen dieser Ergebnisse für den Automobilhersteller und sein Leasinggeschäft meine Aufgabe war.

Die anschließende Erweiterung der WI-Publikation, die in der wissenschaftlichen Zeitschrift *Business & Information Systems Engineering* (Gleue et al., 2018) veröffentlicht wurde, wurde von Herrn Eilers und mir methodisch grundlegend überarbeitet. Wir haben in diesem Rahmen eine Fallstudie mit drei verschiedenen Prognose-Szenarien entwickelt und vorgestellt, die umfangreich getestet und mit Benchmarks belegt wurde. Außerdem wurde eine Studie bzgl. des Einflusses externer Faktoren auf die Markt- und Restwerte durchgeführt und analysiert, welche Inputfaktoren den größten Einfluss auf die Prognosegenauigkeit besitzen. Während die „fachliche Idee“ für diese Analysen von mir stammt, wurde die technische Umsetzung und mathematische Interpretation der Ergebnisse von Herrn Eilers durchgeführt, was auch die Auswahl der dafür angewandten, mathematisch/statistischen Methoden umfasst. Die fachliche Interpretation und Evaluation der Ergebnisse dieser Verfahren wie auch die Ableitung von Implikationen für das Risikomanagement des Automobilherstellers waren dabei meine Aufgabe. Prof. Breitner und Prof. von Mettenheim haben uns dabei grundsätzlich mit ihrem umfangreichen Erfahrungsschatz zu KNN, lineare Mixed-Effects-Modelle und bzgl. unseres methodischen Ansatzes beraten und unterstützt.

Das Paper "It's Not a Bug, It's a Feature: How Visual Model Evaluation Can Help to Incorporate Human Domain Knowledge in Data Science" (Eilers et al., 2017) ist die zweite in einer Reihe von Veröffentlichungen, die auf den Daten, Erkenntnissen und identifizierten Forschungslücken im Rahmen des zuvor beschriebenen Kooperationsprojektes des Instituts für Wirtschaftsinformatik mit einem Automobilhersteller basieren. Im Kern geht es hier um die Auflösung von Kommunikationsbarrieren, die in derartigen Projekten häufig aufgrund der verschiedenen Ausbildungen und Erfahrungshorizonte der einzelnen Projektteilnehmer entstehen. Gemeinsam haben Herr Eilers und ich während unserer Projektarbeit erlebt, dass

sich die Kommunikation unserer Ergebnisse gegenüber Mitarbeitern eines Fachbereiches oder sogar Mitgliedern des Managements, die über ein zu geringes technisches Hintergrundwissen verfügen, häufig komplizierter gestaltet als von uns angenommen: Es kam dabei häufig zu Verständnis- und Abstimmungsschwierigkeiten. Auf der anderen Seite fehlt einem (insbesondere externen) Projektmitarbeiter in der Rolle des „Datenwissenschaftlers“ häufig das fachliche Hintergrund- und Erfahrungswissen, das es ihm ermöglicht, bedarfsgerechte Analysen zu erstellen und fachliche Fehler im Design seiner Implementierung frühzeitig zu erkennen. Vereinfacht ausgedrückt geht es darum, diese „Kommunikationslücke“ mithilfe passender Visualisierungen zu schließen. Herr Eilers hatte die Idee, die Ergebnisse seiner statistischen Analysen mithilfe von Heatmaps so darzustellen, dass Unregelmäßigkeiten in den Inputdaten oder auch den Ergebnissen leicht visuell erkennbar werden. Warum diese Unregelmäßigkeiten entstehen bzw. vorhanden sind ist meist eine fachliche Frage, die nur von jemandem mit umfangreichem fachlichem Hintergrundwissen beantwortet werden kann. Diese Erkenntnisse auf Basis von Analysen in großen Datenmengen zunächst auf wissenschaftlich valide Art herzuleiten und visuell darzustellen bedarf eines entsprechenden technischen, mathematischen und statistischen Hintergrundwissens. Während Herr Eilers und ich also die Erfahrungen und Erkenntnisse, die letztlich zu diesen Veröffentlichungen geführt haben, gemeinsam machen durften, wurde die gesamte Konzeption und Implementierung von Herrn Eilers durchgeführt. Die Idee für die Verwendung von Heatmaps für Restwertmatrizen kam ursprünglich von mir, diese jedoch zu dem in diesem Paper verfolgten Zweck einzusetzen war gänzlich Herrn Eilers Idee. Ich habe die gesamte, für seine Analysen notwendige Datenbasis geliefert, ihn bei der Konzeption geeigneter Szenarien unterstützt und seine Ergebnisse als „fachlicher Gegenpart“ mit ihm gemeinsam überprüft. Prof. Breitner hat Herrn Eilers dabei mit seinen umfangreichen mathematischen Kenntnissen und seiner langjährigen Erfahrung in der Anwendung von KNN unterstützt.

Tabelle 1: Publikationsübersicht

Jahr	Titel	Co-Autoren	Konferenz / Journal	WKWI	JQ3	IF	Appendix
2013	Enhancing Literature Research Processes: A Glance at an Approach Based on Latent Semantic Indexing	Koukal A., Breitner M.H.	Lecture Notes in Informatics (LNI) Proceedings (GI 2013), Koblenz, Deutschland	B	C	-	A
2014	Enhancing Literature Review Methods – towards more efficient Literature Research with Latent Semantic Indexing	Koukal A., Breitner M.H.	European Conference on Information Systems 2014 Proceedings, Tel Aviv, Israel	A	B	-	B
2014	Enhancing Literature Review Methods - Evaluation of a Literature Search Approach based on Latent Semantic Indexing	Koukal A., Breitner M.H.	International Conference on Information Systems 2014 Proceedings, Auckland, Neuseeland	A	A	-	C
2017	Decision Support for the Automotive Industry: Forecasting Residual Values Using Artificial Neural Networks	Eilers D., von Mettenheim H.-J., Breitner M. H.	Wirtschaftsinformatik 2017 Proceedings, St. Gallen, Schweiz	A	C	-	D
2017	It's Not a Bug, It's a Feature: How Visual Model Evaluation Can Help to Incorporate Human Domain Knowledge in Data Science	Eilers D., Köpp C., Breitner M. H.	International Conference on Information Systems 2017 Proceedings, Seoul, Südkorea	A	A	-	E
2018	Decision Support for the Automotive Industry: Forecasting Residual Values Using Artificial Neural Networks	Eilers D., von Mettenheim H.-J., Breitner M. H.	Business & Information Systems Engineering	A	B	3.392	F

1. Einleitung

*„Gute Informationen sind schwer zu bekommen.
Noch schwerer ist es, mit ihnen etwas anzufangen.“*

– Sir Arthur Conan Doyle

1.1 Problemstellung, Relevanz und Motivation

In der weltweiten IT-Szene herrscht Goldgräberstimmung. Auch wenn die allgemein bekannte Floskel „Daten sind das Gold des 21. Jahrhunderts“ inzwischen etwas abgedroschen wirken mag, so könnte sie nach wie vor zutreffender nicht sein: Der Hype um Begriffe wie Data Mining, Business Intelligence, Big Data Analytics und Data Science hält seit Jahren in nahezu allen Branchen unaufhaltsam Einzug, die Nachfrage nach Mitarbeitern mit Fähigkeiten in diesem Gebiet übersteigt deutlich das Angebot (vgl. z.B. Davenport und Patil, 2012). Die Anzahl wissenschaftlicher Publikationen aus diesem überaus umfangreichen Themengebiet steigt auch in den renommiertesten Zeitschriften und wichtigsten Konferenzen der Wirtschaftsinformatik seit Jahren stetig (vgl. Agarwal und Dhar, 2014).

Das „Schürfen“ nach Informationen begann gegen Ende des 20. Jahrhunderts, Konzerne verstärkten ihre Bemühungen, die bereits vorhandenen, verteilten Datenbanken in Data Warehouses zu integrieren, die anschließend für (Daten-)Analysen verwendet werden konnten, um geschäftskritische Entscheidungen auf Basis besserer und detaillierterer Informationen treffen zu können. Der Begriff „Business Analytics“ wurde in den späten 2000er Jahren eingeführt, um die analytische Komponente der Business Intelligence hervorzuheben (vgl. Davenport, 2006). Die vergleichsweise jungen Begriffe „Big Data“ und damit „Big Data Analytics“ entstanden, um die stetig größer werdenden Datenmengen und die dafür notwendigen, höher entwickelten Speicherungs-, Management-, Analyse- und Visualisierungstechniken zu beschreiben (vgl. Chen et al., 2012). Während der Gedanke, Analysetechniken zum Erkenntnisgewinn aus Daten zu verwenden, so alt ist wie die Statistik selbst, so machen heute die höhere Geschwindigkeit und Menge ökonomischer und sozialer Interaktionen bei gleichzeitig deutlich höherem Vernetzungsgrad, geringere Kosten für Rechenleistung und Datenhaltung sowie benutzerfreundlichere Analysesoftware den entscheidenden Unterschied, der das Entstehen, Erfassen und Verarbeiten von „Big Data“ letztlich einem breiteren Publikum zugänglich macht. (vgl. Agarwal und Dhar, 2014).

Ziel dieser kumulativen Dissertation ist es, zu untersuchen, wie abseits von den großen „Leuchtturmprojekten“ in Industrie und Forschung auch in kleineren und mittleren Projekten

mit Methoden und Technologien aus dem weiten Feld des Data Mining, Big Data Analytics und Data Science althergebrachte Arbeitsabläufe und Geschäftsprozesse verbessert, automatisiert und optimiert werden können. Dazu werden im Folgenden zwei im Rahmen meiner Promotionszeit durchgeführte Projekte vorgestellt: Zum einen die Entwicklung eines neuartigen Werkzeugs zur semantischen Literaturrecherche basierend auf Algorithmen aus dem Gebiet des Text Mining und der semantischen (Ähnlichkeits-)Suche, zum anderen ein Kooperationsprojekt des Instituts für Wirtschaftsinformatik der Leibniz Universität Hannover mit einem großen, deutschen Automobilhersteller zur Entwicklung eines Entscheidungsunterstützungssystems für das Leasinggeschäft, das im Kern Markt- und Restwertprognosen für Gebrauchtfahrzeuge berechnet und letztlich mit dynamischen Reports geschäftskritische Fragen des Risikomanagements beantworten und besser informierte Entscheidungen zu treffen hilft.

1.2 Aufbau der Arbeit und wissenschaftlicher Beitrag

Insgesamt besteht diese kumulative Dissertation aus zwei voneinander unabhängigen Teilen, die jedoch beide den Gebieten Big Data Analytics und Data Mining, also der systematischen Anwendung statistischer Methoden auf große Datenbestände, auch „Knowledge Discovery in Databases“ (KDD) zugerechnet werden können. Zunächst sollen im Folgenden unsere Beiträge zum Gebiet der semantischen Ähnlichkeitsanalyse von Texten vorgestellt werden (vgl. Kapitel 2). Dieses Forschungsvorhaben basiert auf der Idee, den zeitaufwendigen und auch heute immer noch überwiegend manuellen Teil jeder etablierten wissenschaftlichen Vorgehensweise – der Literaturrecherche – unter Einsatz moderner Methoden aus dem Feld der Datenanalyse und des Text Mining weitestgehend zu automatisieren, um so die Effizienz des Forschungsprozesses insgesamt zu erhöhen. Zu diesem Zweck wurde eigens eine „Literatur-Suchmaschine“ von uns entwickelt, die anhand geeigneter Algorithmen die semantische Ähnlichkeit zwischen Texten mit hoher Genauigkeit erkennen kann und so eine deutlich genauere und effizientere Suche in größeren Datenbanken (bestehend aus wissenschaftlichen Volltexten eines Fachgebiets) ermöglicht. In Kapitel 3 wird das oben bereits beschriebene Kooperationsprojekt mit der Automobilindustrie zu Design, Konzeption und Implementierung eines EUS vorgestellt und kritisch diskutiert. Die Kernfunktionalität und gleichzeitig den Fokus unserer Arbeit in diesem Projekt stellt die Fähigkeit des Systems dar, mithilfe von KNN automatisiert feingranulare Kurz- und Langfristprognosen bzgl. zu erwartender Markt- und Fahrzeugrestwerte zu berechnen. Dabei werden große, kontinuierlich wachsende Datenmengen sowohl des Leasing- und Gebrauchtwagenmarktes als auch aus zahlreichen Konzernsystemen automatisiert bereinigt, integriert, aufbereitet, gespeichert und verarbeitet, um letztlich diverse, sich laufend

aktualisierende Entscheidungsvorlagen für das Management des Automobilherstellers zu liefern. Die im Rahmen dieser beiden Forschungsprojekte in der Welt der Wissenschaft und Privatwirtschaft gesammelten Erfahrungen und Erkenntnisse werden in Kapitel 4 kritisch beleuchtet und diskutiert. Der Fokus liegt dabei auf den Gemeinsamkeiten, Unterschieden und Herausforderungen bei der Durchführung von Projekten mit wissenschaftlichem Anspruch und wissenschaftlicher Vorgehensweise in diesen grundlegend verschiedenen Umfeldern. Im Vordergrund soll hierbei insbesondere der Aspekt des Profils des Wirtschaftsinformatikers als „Schnittstelle“ zwischen (im vorliegenden Falle) den Fachbereichen eines Wirtschaftsbetriebs, Software-Entwicklungsteams und Wissenschaftlern bei der kommerziell getriebenen Durchführung dieser Art von Projekten stehen. Kapitel 5 rundet die vorliegende Dissertation mit einem kurzen Fazit ab. Abbildung 1 stellt die Gesamtstruktur dieser Arbeit schematisch dar und zeigt die Konferenzen und Journals, in denen die Arbeiten jeweils publiziert wurden.

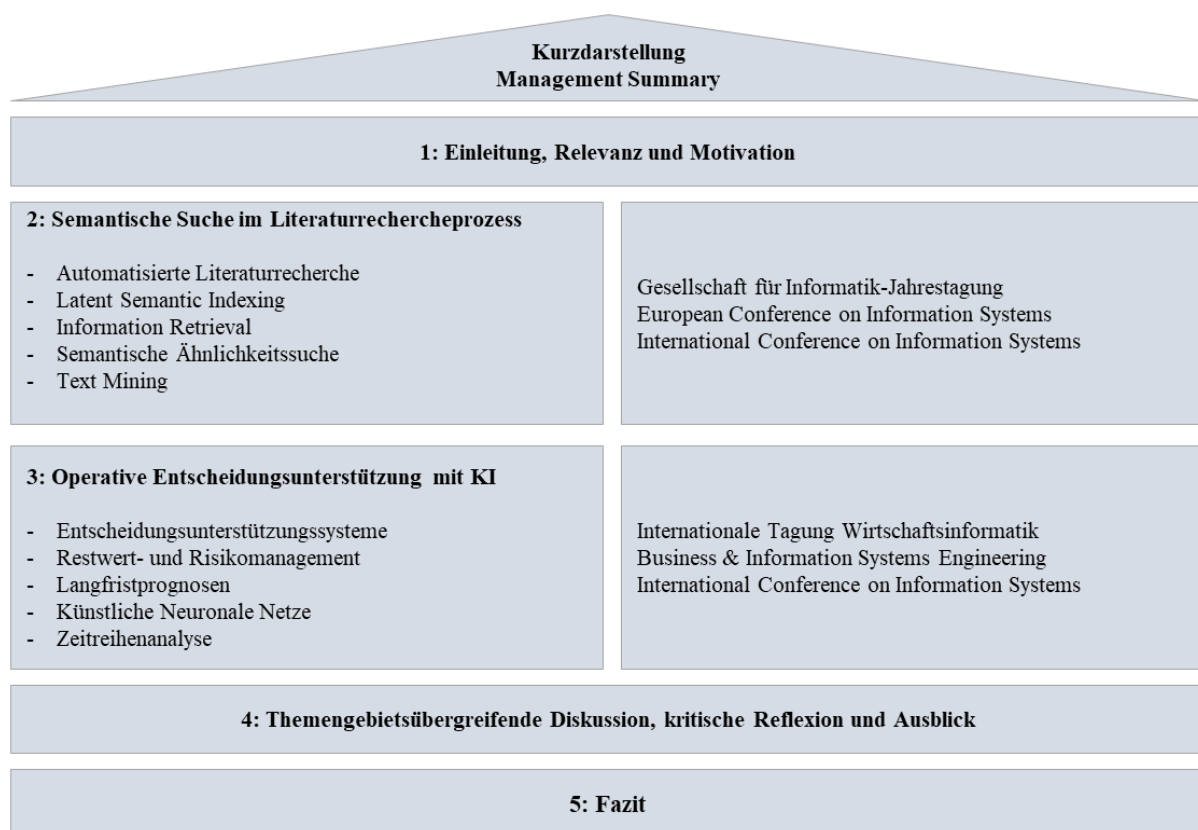


Abbildung 7: Aufbau der vorliegenden Dissertation

2 Semantische Suche im Literaturrechercheprozess

*„Oft finden wir etwas ganz anderes, ja, besseres, als wir suchten;
oft auch das Gesuchte selbst auf einem ganz anderen Wege,
als den wir zuerst vergeblich danach eingeschlagen hatten.“*

– Arthur Schopenhauer

In diesem Abschnitt werden drei Veröffentlichungen aus dem Themengebiet des Text Mining und der semantischen Suche zur Ähnlichkeitsanalyse von unstrukturierten Texten vorgestellt. Es handelt sich dabei um aufeinander aufbauende Publikationen, beginnend mit dem Titel „Enhancing Literature Research Processes: A Glance at an Approach Based on Latent Semantic Indexing“. Es handelt sich dabei um eine erste, prototypische Annäherung an das Thema, die auf der Jahrestagung der Gesellschaft für Informatik 2013 in Koblenz erstmalig vorgestellt wurde. Durch die überaus positive Resonanz und den wissenschaftlichen Diskurs in der (zunächst deutschsprachigen) IS-Community wurden wir ermutigt, das Thema weiter zu verfolgen. Daraufhin wurden sowohl das Tool als auch die wissenschaftliche Methodik maßgeblich weiterentwickelt. Die Ergebnisse dieses Prozesses wurden in der Publikation „Enhancing Literature Review Methods - towards more efficient Literature Research with Latent Semantic Indexing“ verarbeitet und der europäischen IS-Community im Rahmen der ECIS 2014 in Tel Aviv vorgestellt. Auch hier wurden der Forschungsansatz, die Bedeutung der Thematik für die Wissenschaftscommunity und die Relevanz der Ergebnisse durchweg sehr positiv bewertet. Darüber hinaus wurde sowohl durch die Einschätzungen der Gutachter als auch durch regen Meinungs- und Erfahrungsaustausch mit Kollegen während der Konferenz deutlich, dass die weitere Forschung in diesem Themengebiet auch international von wissenschaftlichem Interesse ist. Folglich wurde das Forschungsprojekt aufbauend auf diesen Erkenntnissen weitergeführt, die bekannten Kritikpunkte adressiert und weitere Evaluationen und Benchmarks auf einem deutlich vergrößerten Datensatz durchgeführt. Die neuen Ergebnisse dieses Prozesses wurden letztlich im Rahmen der ICIS 2014 in Auckland, Neuseeland mit dem Titel „Enhancing Literature Review Methods - Evaluation of a Literature Search Approach based on Latent Semantic Indexing“ vorgestellt.

Die an den o.g. Publikationen beteiligten Autoren waren André Koukal, Christoph Gleue und Michael H. Breitner.

2.1 Motivation und Zielsetzung

Die Literaturrecherche als komplexer und sehr wichtiger erster Schritt eines jeden Forschungsprojektes (vgl. Wolfswinkel et al., 2013) bildet in der Wissenschaft das theoretische Fundament desselben (vgl. Baker, 2000). Unabhängig von der verfolgten Forschungsmethodik ist eine umfassende Kenntnis der bereits existierenden Arbeiten im jeweiligen Forschungsgebiet unabdingbar und letztlich entscheidend für die korrekte Identifikation der Forschungslücken, die es im weiteren Projektverlauf zu schließen gilt. Nur so kann sichergestellt werden, dass die eigene Forschung einen wertvollen Beitrag zum designierten Forschungsfeld leistet und auch auf internationalem Niveau konkurrenzfähig ist (vgl. Hart, 1998; Levy und Ellis, 2006). Insbesondere im noch relativ jungen Forschungsgebiet der Wirtschaftsinformatik ist laut Webster und Watson (2002) häufig ein Mangel an qualitativ hochwertiger Literaturrecherche zu beobachten. Die Autoren stellen fest, dass, um die Wirtschaftsinformatik als Forschungsgebiet weiter zu etablieren, effiziente Literaturrecherchemethoden einen wertvollen Beitrag darstellen. Gleichzeitig erhöhen fundierte und umfassende Literaturrecherchen die Chancen auf hochklassige Publikationen von Forschungsarbeiten.

Beachtet man dabei die weltweit stetig steigende Anzahl wissenschaftlicher Veröffentlichungen und den durch neue Technologien massiv erleichterten Zugriff darauf (vgl. Mabe und Amin, 2001; Park und Lee, 2011), so wird klar, dass in der daraus resultierenden, komplexen Informationslandschaft (vgl. Bawden und Robinson, 2009; Manwani et al., 2011) manuell durchgeführte, umfassende Literaturrecherchen zu einer zunehmend zeitaufwendigen Aufgabe werden. Auch wenn einfache, stichwortbasierte Suchverfahren im Vergleich zu komplett manuell durchgeführten Analysen großer, wissenschaftlicher Literaturdatenbanken bereits einen hohen Mehrwert bieten, bringen sie doch gewisse Nachteile mit sich (vgl. Blair und Maron, 1985; Homayouni et al., 2004; LaBrie und St. Louis, 2003) und sind daher alles andere als optimal (vgl. Dumais et al., 1988): Ambiguität von Suchbegriffen, Synonymie, Polysemie, die unsachgemäße Anwendung von Stoppwörtern wie „die“, „in“ oder „und“, Bindestriche, Klammern, Pluralformen sowie letztlich die auftretende Unschärfe und Inkonsistenz bei der Indizierung des zu durchsuchenden Textmaterials führen schnell zu einer Verzerrung der Suchergebnisse. Folglich führen stichwortbasierte Ansätze sehr wahrscheinlich zu falsch-positiven oder falsch-negativen Suchergebnissen. Anders ausgedrückt werden potenzielle Treffer nicht gefunden oder schlicht unzutreffende Suchergebnisse geliefert (vgl. Blair und Maron, 1985; Dumais et al., 1988; Hofmann, 1999; LaBrie und St. Louis, 2003; Salton und McGill, 1986; Yandell und Majoros, 2002).

Trotz der oben beschriebenen Defizite dieser Art von Ansätzen sind rein stichwortbasierte Suchverfahren auch heute noch bei einem Großteil der wissenschaftlichen Datenbanken üblich (vgl. Cui et al., 2003). Zu Beginn des hier vorgestellten Forschungsprojektes wurden alle für die Wirtschaftsinformatik relevanten, wissenschaftlichen Datenbanken wie AISel, IEEE, JStor, ScienceDirect oder Wiley untersucht. Dabei konnte festgestellt werden, dass sich das beschriebene Fehlerverhalten in den Suchergebnissen widerspiegelt, was darauf schließen lässt, dass die zugrundeliegenden Suchmechanismen nach wie vor rein stichwortbasiert sind.

Folglich stellt eine Reduzierung oder gar Beseitigung der zuvor genannten Defizite stichwortbasierter Suchverfahren einen wichtigen Beitrag, nicht nur im Forschungsgebiet der Wirtschaftsinformatik im Speziellen, sondern für die Wissenschaft und Forschungsgemeinschaft im Allgemeinen, dar. Um einer Lösung für die hier umrissene Problemstellung näherzukommen ist das Ziel unserer Forschung zu diesem Thema, einen alternativen, modernen und intelligenteren Ansatz zur automatisierten Ähnlichkeitsanalyse von Texten zu entwickeln, der die folgenden Kriterien erfüllt:

- 1) Verlässliche Identifikation wissenschaftlicher Arbeiten, die einem bestimmten Forschungsfeld angehören (Vorhandensein der initialen Suchbegriffe)
- 2) Identifikation semantisch ähnlicher Arbeiten aus demselben oder sogar aus anderen Forschungsgebieten (Vorhandensein von Synonymen und/oder mit den initialen Suchbegriffen „sinnhaft verwandter“, also semantisch ähnlicher Begriffe)
- 3) Die Überwindung des Problems der Synonymie und Polysemie und somit die größtmögliche Vermeidung falsch-positiver und falsch-negativer Suchergebnisse
- 4) Die Möglichkeit, Suchanfragen in natürlicher Sprache zu formulieren (Verwendung von Stichwörtern, vollständigen Sätzen, Abstracts und ganzen Forschungsarbeiten als Suchanfrage)

Im Bereich des Text Mining existieren in der Literatur bereits einige vielversprechende Ansätze. Die Arbeiten von Corley und Mihalcea (2005), Kontostathis und Pottenger (2006), Kuechler (2007), Rehurek und Sojka (2010) und Zhang et al. (2011) deuten darauf hin, dass die Anwendung von Latent Semantic Indexing (LSI) möglicherweise eine Lösung für die oben aufgeführten Herausforderungen darstellt. Die Eignung und Anwendbarkeit dieses Verfahrens wurde bereits anhand eines breiten Spektrums an Anwendungsfällen nachgewiesen (vgl. Deerwester et al., 1990; Dumais et al., 1988; Gordon und Dumais, 1998; Kontostathis und Pottenger, 2002; Zelikovitz und Hirsh, 2001), wurde jedoch bisher noch nicht im Hinblick auf den hier dargestellten Zweck der automatisierten Literaturrecherche untersucht. Wir vertreten

die grundlegende Auffassung, dass der Einsatz einer semantischen Suchmethode einen deutlichen Effizienzgewinn mit sich bringt, somit wertvolle Zeit bei der Literaturrecherche eingespart werden kann und ein solches Verfahren gleichzeitig dabei hilft, Herausforderungen wie z.B. die in der Wirtschaftsinformatik ausufernde Verwendung unterschiedlicher Begriffe für dasselbe gedankliche Konzept (vgl. Lebek et al. 2013) zu überwinden.

Das erklärte Ziel dieser Forschungsarbeit ist daher die Entwicklung und kritische Diskussion eines alternativen Verfahrens, das Wissenschaftlern den Literaturrechercheprozess mithilfe eines LSI-basierten Suchverfahrens erleichtert und somit dabei hilft, in der wissenschaftlichen Praxis wertvolle Zeit einzusparen. Gleichzeitig soll die Anwendbarkeit von LSI für die semantische Ähnlichkeitserkennung zwischen Texten auf einer großen Datenbasis aus wissenschaftlicher Literatur getestet und evaluiert werden.

Den Kern der vorliegenden Arbeit stellt das „Tool for Semantic Indexing and Similarity Queries“ (TSISQ) dar. Es wurde im Rahmen dieses Vorhabens entwickelt, um unstrukturierte Texte (wie vollständige Forschungsarbeiten oder jede Form von natürlicher Sprache in Textform) als Suchanfrage zu verarbeiten und dazu inhaltlich passende, semantisch ähnliche Texte innerhalb eines großen Indexes wissenschaftlicher Publikationen zu identifizieren und anhand ihrer Ähnlichkeit (verglichen mit der Suchanfrage) zu bewerten. TSISQ stellt somit eine nützliche und sinnvolle Ergänzung diverser Schritte der in der Forschung etablierten Literaturrechercheverfahren dar (vgl. Okoli und Schabram, 2010; Levy und Ellis, 2006; Webster und Watson, 2002). Das Tool hilft dabei, die o.g. Schwierigkeiten bei der Anwendung „klassischer“, stichwortbasierter Suchverfahren zu überwinden und unterstützt, erleichtert und optimiert dabei den gesamten Forschungsprozess. Darüber hinaus wurde TSISQ entwickelt, um jede Forschungsmethodik zu unterstützen, die explizit oder implizit eine umfassende Literaturrecherche verlangt, wie beispielsweise Design Science Research im Hinblick auf die von Hevner geforderte wissenschaftliche Rigorosität und Relevanz (vgl. Hevner et al., 2004) oder den „Rigor Cycle“ (vgl. Hevner, 2007). Der Prototyp des TSISQ-Tools wird im Folgenden verwendet, um die Leistungsfähigkeit und das Potenzial von LSI in diesem Zusammenhang zu demonstrieren und zu evaluieren. Auf die folgende Forschungsfrage soll eine Antwort gefunden werden:

FF: *Wie kann ein LSI-basierter Ansatz die Effizienz von wissenschaftlichen Literaturrechercheprozessen steigern?*

Nachdem in der vorangegangenen Einleitung die Motivation und Relevanz des Themas sowie das Forschungsziel und der wissenschaftliche Beitrag zum Forschungsfeld erörtert wurden,

umfasst der nächste Abschnitt (Kapitel 2.2.1) einen Überblick über die verwandten Arbeiten aus dem Gebiet und beschreibt die theoretischen Grundlagen, auf der die im Folgenden vorgestellte Forschung basiert. Anschließend wird näher auf die dabei angewandte Forschungsmethodik eingegangen (Kapitel 2.2.2) sowie die theoretischen Konzepte hinter LSI und dessen Funktionsweise im Detail beschrieben (Kapitel 2.3.1). Die Architektur, Implementierung und Benutzeroberfläche des Tools werden in Kapitel 2.3.2 illustriert, gefolgt von einigen konkreten Anwendungsfällen, an denen die Leistungsfähigkeit von TSISQ demonstriert wird (Kapitel 2.4). Abschließend werden Verfahren und Tool kritisch diskutiert und mögliche Einschränkungen aufgezeigt (Kapitel 2.5), gefolgt von einem kurzen Zwischenfazit zu diesem Themenkomplex der vorliegenden Dissertation (Kapitel 2.6).

2.2 Literaturrecherche und Methoden des Information Retrieval

2.2.1 Stand der Forschung

Die Literaturrecherche ist die fundamentalste und gleichzeitig eine der wichtigsten Methoden zur Schaffung einer theoretischen Grundlage für jede Art von wissenschaftlicher Forschung. Ihre Qualität und letztlich auch ihr Nutzen hängen dabei maßgeblich von dem dabei verfolgten Literaturrechercheprozess ab (vgl. vom Brocke et al., 2009). In der Wirtschaftsinformatik gibt es bereits zahlreiche etablierte Methoden für die Durchführung qualitativ hochwertiger Literaturrecherchen. Auch wenn die jeweiligen Autoren verschiedenste Richtlinien anführen, kann es als allgemeingültig angesehen werden, dass ein tiefgreifendes Verständnis des eigenen Forschungsgebiets für jeden Wissenschaftler unabdingbar ist. Demzufolge ist die Identifikation aller relevanten wissenschaftlichen Arbeiten im jeweiligen Forschungsgebiet eine wichtige Teilaufgabe im Literaturrechercheprozess (vgl. Wolfswinkel et al., 2013). Webster und Watson (2002) stellen bezugnehmend auf die Forschung in der Wirtschaftsinformatik fest, dass die umfassende Kenntnis der bestehenden Literatur sowohl aus der eigenen als auch aus verwandten Disziplinen erforderlich ist, um einen idealen Artikel zu verfassen. Levy und Ellis (2006) schlagen einen systematischen, dreistufigen Ansatz zur Datenverarbeitung vor, wobei die erste Stufe bereits die Sammlung und das Screening der „Inputs“, also die Identifikation und Analyse qualitativ hochwertiger Literatur mit Blick auf die Verlässlichkeit und Validität der resultierenden Studie und ihrer Ergebnissen umfasst. Okoli und Schabram (2010) weisen darauf hin, dass das Vorgehen bei der Literaturrecherche systematisch (nach einer definierten Methodik) erfolgen und, was noch wichtiger ist, umfassend („comprehensive in its scope“; siehe auch Fink, 2010; Rousseau et al., 2008) sein muss. Das Kriterium der Vollständigkeit wird in der dritten Stufe des Frameworks, „searching for the literature“ im dort vorgestellten

„eight-step-guide to conducting a [scientifically rigorous] systematic literature review“ nochmals unterstrichen.

Vor dem Hintergrund der zuvor genannten etablierten Richtlinien und Vorgehensweisen muss an dieser Stelle darauf hingewiesen werden, dass das Ziel der vorliegenden Arbeit *nicht* ist, eine gänzlich neue Methodik zur Durchführung von Literaturrecherchen zu entwickeln, sondern die entsprechenden Schritte der bereits bestehenden Methodiken bestmöglich mit einer toolgestützten, semantischen Ähnlichkeitssuche zu unterstützen. Um den Rahmen dieser Studie weiter einzugrenzen folgt ein grober Überblick über die dabei in Betracht gezogenen Information Retrieval (IR)-Ansätze im Hinblick auf deren Potential zur Identifikation semantischer Ähnlichkeiten zwischen Texten und unter Berücksichtigung der in Kapitel 2.1 aufgestellten Kriterien.

Query Expansion (QE) ist eine IR-Methode, bei der eine gegebene Suchanfrage um Synonyme oder verwandte Begriffe ergänzt wird, um so die Suchergebnisse zu verbessern und die Effizienz der Suche zu erhöhen. Dabei werden zwar insbesondere die Probleme stichwortbasierter Suchverfahren mit Synonymie und der Verzerrung von Suchergebnissen durch unpassende, unklare oder unscharfe Suchbegriffe adressiert (vgl. Cui et al., 2003; Liu et al., 2011; Mitra et al., 1998; Qiu und Frei, 1993; Xu und Croft, 1996), QE erfüllt dabei jedoch bei weitem nicht alle der in Abschnitt 2.1 aufgeführten Kriterien, auch nicht das der Beseitigung des Problems der Polysemie (vgl. Liu et al., 2011). Darüber hinaus führt die Anwendung von QE durch die Ergänzung von Suchbegriffen zu einer Erhöhung der *Quantität* (vgl. Liu et al., 2011; Mitra et al., 1998; Santos und Riveiro, 2011; Xu und Croft, 1996), nicht aber zu einer höheren *Qualität* der Suchergebnisse. Folglich ist die Methode für den hier verfolgten Zweck ungeeignet, da der Fokus ausdrücklich auf einer höheren *Qualität* der Suchergebnisse und gleichzeitig auf der Eignung des Verfahrens liegt, auch natürliche Sprache wie z.B. vollständige Abstracts oder Forschungsarbeiten als Input verarbeiten zu können. Auch der zuvor erwähnte Effizienzgewinn durch den Einsatz von QE wird in der wissenschaftlichen Literatur kontrovers diskutiert und von einigen Autoren angezweifelt (vgl. Vorhees, 1994; Xu und Croft, 1996).

Das Konzept der semantischen Ähnlichkeit ist „[...]a concept by which a metric is given to groups of terms or documents based on the similitude of their meanings“ (Furlan et al., 2013). Das Fachgebiet, das sich im Kern mit der Verarbeitung natürlicher Sprache befasst, nennt sich Natural Language Processing (NLP) und spielt eine wichtige Rolle bei der Bewertung semantischer Ähnlichkeiten zwischen Dokumenten. Dabei wird zwischen drei grundlegenden Aspekten unterschieden: Informationsextraktion, Semantik und Information Retrieval (IR)

(vgl. Yandell und Majoros, 2002). Latent Semantic Indexing (LSI), auch als Latent Semantic Analysis (LSA) bezeichnet, ist ein Verfahren aus dem Gebiet der NLP-Methoden. Es handelt sich dabei um einen mathematischen Ansatz, der bei der Bewertung semantischer Ähnlichkeiten von Texten/Dokumenten ohne explizites, von Menschen vorstrukturiertes Wissen auskommt. Im Wesentlichen wird eine Dimensionsreduktion vorgenommen, um die wichtigsten (semantischen) Dimensionen in einem Dokument zu ermitteln. Diese Dimensionen stimmen mit den „latenten Konzepten“ überein, die die Bedeutung von Wörtern und Dokumenten innerhalb des durch sie definierten semantischen Raums repräsentieren (vgl. Gabrilovich und Markovitch, 2009). Einfach ausgedrückt bildet LSI die Bedeutung von Begriffen in einem semantischen Raum ab (vgl. Kintsch, 2010). Eine detailliertere Darstellung der Funktionsweise von LSI folgt in Abschnitt 2.3.

In der Praxis wird LSI für die Lösung einer Vielzahl von Problemstellungen eingesetzt, in der größere Mengen an textuellen Daten eine Rolle spielen, z.B. bzgl. Textsuche (vgl. Dumais, 1992; Dumais, 1994), Textklassifikation (vgl. Zelikovitz and Hirsh, 2001) und Filterung (vgl. Zha und Simon, 1998). Die möglichen Anwendungsgebiete für LSI sind vielfältig (z.B. Deerwester et al., 1990; Foltz and Dumais, 1992; Hofmann, 1999; Landauer and Dumais, 1997; Wolfe et al., 1998) und es wurde bereits wissenschaftlich nachgewiesen, dass die Anwendung von LSI die durchschnittliche Suchgenauigkeit erhöht (vgl. Ding, 1999). Kontostathis (2007) beschreibt die Eignung des LSI-Verfahrens um Suchanfragen mit Dokumenten zu vergleichen bzw. semantische Beziehungen zwischen Dokumenten und den darin beinhalteten Begriffen zu analysieren. Um dies zu erreichen, werden die den Suchbegriffen zugrundeliegenden Bedeutungen und Konzepte mit denen aus der Suchanfrage verglichen, anstatt dass auf das Vorhandensein der Suchbegriffe selbst geprüft wird (vgl. Farrús und Costa-jussá, 2013). Folglich stellt die Anwendung von LSI im Vergleich zu rein stichwortbasierten Suchverfahren einen grundlegend andersartigen Ansatz dar, um natürliche Sprache ungefiltert als Input für eine (Literatur-) Suche zu verwenden (vgl. Gordon und Dumais, 1998). Diese Erkenntnis erlaubt die Schlussfolgerung, dass LSI ein hilfreiches und vielversprechendes Verfahren sein kann, um den Literaturrechercheprozess in der wissenschaftlichen Praxis effizienter zu gestalten.²

² Genauer: Bei mindestens gleichbleibender Qualität Zeit einzusparen oder mindestens mit dem gleichen Zeitaufwand eine höhere Qualität der Literaturrecherche zu erreichen.

Tabelle 2: Wissenschaftliche Beiträge zu praktischen und theoretischen Anwendungsfällen für LSI

Nr.	Autor(en)	Forschungsthema
1	Gong und Liu (2001)	Automatische Textzusammenfassung, Benchmark mit Standardmethoden des Information Retrieval
2	Gee (2003)	Evaluation von LSI zur E-Mail-Spamfilterung
3	Wolfe & Goldman (2003)	Vorhersage psychologischer Phänomene
4	Homayouni et al. (2004)	Identifikation funktionaler Zusammenhänge zwischen Genen aus der biomedizinischen Literatur
5	Shen et al. (2004)	Website-Klassifikation durch Zusammenfassung
6	Steinberger und Ježek (2004)	Zusammenfassung von Text, Messung der inhaltlichen Ähnlichkeit zwischen automatisch generierter Zusammenfassung und zugehöriger Originalquelle
7	Yeh et al. (2005)	Automatische Zusammenfassung politischer Artikel
8	Kuechler (2007)	Analyse unstrukturierter Texte, Natural Language Processing, praktische Anwendungsfälle von LSI
9	Bhandari et al. (2008)	Automatische Textzusammenfassung, Extraktion von Themen aus Dokumenten
10	Gansterer et al. (2008)	E-Mail-Spamfilterung
11	Sidorova et al. (2008)	Ermittlung des „intellektuellen Kerns der Disziplin Wirtschaftsinformatik“ mittels Untersuchung von Abstracts aus Artikeln der IS-Forschung
12	Hovorka et al. (2009)	Analyse semantischer Beziehungen in Abstracts der wichtigsten Business Journals, Untersuchung von Konvergenzen zwischen der Wirtschaftsinformatik und anderen Geschäftsfeldern/-disziplinen
13	Blake (2010)	Identifikation von Kernthemen der Daten- und Informationsqualitätsforschung
14	Go et al. (2010)	Sentiment-Analyse, automatische Klassifikation von Twitter-Nachrichten
15	Lee et al. (2010)	Vergleich LSI-verwandter Textmining-Methoden
16	Maas et al. (2011)	Automatische Sentiment-Analyse und Klassifikation
17	Abate et al. (2013)	Semantische Analyse biomedizinischer Literatur; Integration und Extraktion biologischer Information aus dem Internet
18	Arijit (2013)	SMS-basierte Beantwortung von Fragen und Information Retrieval; Meinungsextraktion aus SMS für automatische Beantwortung
19	Koukal et al. (2013, 2014a, 2014b)	Verbesserung/Optimierung wissenschaftlicher Literaturrechercheprozesse
20	Nugumanova und Bessmertny (2013)	Automatische Extraktion von Wortpaar-Kollokationen aus Fachliteratur
21	Shao et al. (2013)	Wiederherstellung von Rückverfolgbarkeitsinformationen zwischen Anforderungsdokumenten und Quellcodes basierend auf LSI und besonderen Merkmalen von Quellcodes

Innerhalb der letzten dreißig Jahre sind zahlreiche Publikationen über LSI und dessen Arbeitsweise, Performancemessungen, theoretische Ansätze zur Erlangung eines detaillierteren Verständnisses sowie der Optimierung des Algorithmus und/oder von Teilen des LSI-Prozesses erschienen (vgl. z.B. Brand, 2006; Cao und Ngo, 2012; Deerwester et al., 1990; Dumais, 1992; Hofmann, 1999; Kontostathis und Pottenger, 2006; Rehurek und Sojka, 2010). Darüber hinaus sind viele Artikel zu praktischen und theoretischen Anwendungsfällen für LSI in zahlreichen Forschungsdisziplinen erschienen. Diejenigen davon aus der jüngeren Vergangenheit sind in Tabelle 2 aufgelistet. Neben der manuellen Literaturrecherche für diese Dissertation wurden die Ergebnisse konsequenterweise mit dem Prototyp von „TSISQ“ überprüft und verbessert. Neben einigen wenigen Publikationen zu LSI, die in der ersten Iteration nicht in Betracht gezogen wurden, wurde ein Artikel aus der ECIS 2011 mit dem vielversprechenden Titel “A systematic, tool-supported method for conducting literature reviews in information systems” (Bandara et al. 2011) identifiziert. Bei näherer Betrachtung stellte sich jedoch heraus, dass die Autoren weder LSI direkt angewandt noch ihr eigenes Tool oder IT-Artefakt entwickelt haben, sondern die Verwendung einer Kombination bereits existierender, kommerzieller Tools vorschlagen, um Literaturrecherchen nach Levy und Ellis (2006) oder Webster und Watson (2002) zu unterstützen. Die mit der vorliegenden Arbeit adressierte Forschungslücke ist der Abgleich einer Suchanfrage in natürlicher Sprache mit einem großen Textkorpus, bestehend aus wissenschaftlichen Veröffentlichungen aus dem Gebiet der Wirtschaftsinformatik. Besonderes Augenmerk wird dabei (unter Verwendung von LSI oder ähnlichen Methoden) auf die Identifikation semantischer Ähnlichkeiten zwischen dem verwendeten Such-Input und den Suchergebnissen gelegt.

Neben unserer eigenen Forschung in diesem Gebiet (Koukal et al. 2013, 2014a, 2014b) ist uns in der wissenschaftlichen Literatur keine Veröffentlichung bekannt, die diese Lücke zu schließen vermag. Inhaltlich ist die Arbeit von Sidorova et al. (2008) unseren Studien am nächsten: Die Autoren stellen einen LSI-basierten Ansatz vor, um den intellektuellen Kern der Wirtschaftsinformatik zu erfassen („[...]to determine the “intellectual core of the information systems discipline”). Dazu wurden die Abstracts von veröffentlichten Wirtschaftsinformatik-Artikeln im Zeitraum von 1985 bis 2006 in drei der renommiertesten wissenschaftlichen Zeitschriften der Disziplin untersucht (MIS Quarterly, Information Systems Research, Journal of Management Information Systems). Hovorka et al. (2009) beziehen sich auf die Arbeit von Sidorova und ihren Kollegen und analysieren die semantischen Beziehungen von 24.841 Abstracts aus Zeitschriften der wichtigsten Wirtschaftsbereiche, um daraus mögliche Konvergenzen zwischen der Wirtschaftsinformatik und anderen wirtschaftswissenschaftlichen

Disziplinen abzuleiten. Blake (2010) verfolgt einen vergleichbaren Ansatz, um die Kernthemen des Gebiets „Data and Information Quality Research“ zu identifizieren. Ein technologisch ähnliches Vorgehen wurde von Homayouni et al. (2004) im Fachgebiet der Bioinformatik gewählt.

Jedoch ist der im Mittelpunkt unserer Betrachtung stehende Textkorpus deutlich umfangreicher als derjenige von Sidorova et al. (2008). Auch bei der Dokumentensammlung von Homayouni et al. (2004) handelt es sich um einen vergleichsweise kleinen Datensatz, unter anderem, da es sich dabei, wie auch bei den vergleichbaren Arbeiten aller zuvor genannten Autoren, ausschließlich um eine Sammlung von Abstracts handelt. Der von uns untersuchte Textkorpus umfasst hingegen Volltexte aller renommierten, wissenschaftlichen Zeitschriften in der Wirtschaftsinformatik, dem sog. „AIS Basket of Eight“ und zusätzlich aller Veröffentlichungen der vier wichtigsten internationalen Konferenzen dieses Gebiets (AMCIS, ECIS, HICSS, ICIS) in einem Zeitraum von Januar 2007 bis März 2014. Hinzu kommt, dass keine der genannten Forschungsarbeiten das Ziel verfolgt, ihre LSI-Implementierung zur Identifikation relevanter Literatur innerhalb eines Literaturrechercheprozesses anzuwenden oder ihre geschaffenen Artefakte in dieser Hinsicht zu evaluieren.

Eine der großen Herausforderungen unseres Ansatzes besteht darin, dass die meisten gängigen Algorithmen sehr umfangreiche Matrizen direkt „in-memory“ berechnen, was auf die hohe Anzahl an Dokumenten und der darin enthaltenen, einzigartigen Begriffe und semantischen Konzepte zurückzuführen ist. Da es sich folglich um eine sehr speicherintensive Anwendung handelt und gleichzeitig hohe Rechenleistung gefordert ist, sind möglichst geringe technologische Anforderungen des Tools ohne massive Leistungseinbußen wünschenswert. Um dies zu gewährleisten, wurde als Basis für TSISQ ein Open-Source-Framework namens gensim verwendet, welches einen Algorithmus für einen speichereffizienten, inkrementellen Prozess nach Brand (2006) anwendet (vgl. Rehurek und Sojka, 2010). Rehurek and Sojka führen an, dass ihre auf Python basierende Implementierung von Brands‘ Algorithmus die einzige, öffentlich verfügbare Implementierung von LSI darstellt, die vollkommen unabhängig von der Größe des Index arbeitet. Dieser Aspekt ermöglicht die Verwendung von TSISQ auf einem durchschnittlichen aktuellen Desktopcomputer oder Notebook.

Die hier dargestellten Erkenntnisse der zuvor erwähnten Autoren, die praktischen Anwendungsfälle im Gebiet des NLP sowie insbesondere im Bereich LSI wie auch die oben genannten technologischen Fortschritte erlauben es uns, die von uns identifizierte Forschungslücke zu adressieren.

2.2.2 Forschungsmethodik

Die Verbesserung des wissenschaftlichen Literaturrechercheprozesses als Teil jeder etablierten, wissenschaftlich stringenten („rigorosen“) Forschungsmethodik kann nur dann erfolgreich sein, wenn das hier vorgestellte Vorgehen selbst einer solchen Methodik folgt. Aus diesem Grund haben wir uns an den Prinzipien des Design Science Research nach Offermann et al. (2009) und insbesondere Peffers et al. (2007) orientiert; unter Beachtung der Empfehlungen von Hevner et al. (2004, 2007) sowie March und Smith (1995). Bei dem gewählten Forschungsdesign handelt es sich um den sechsstufigen, problemzentrierten Ansatz nach Peffers et al. (vgl. Abbildung 8).

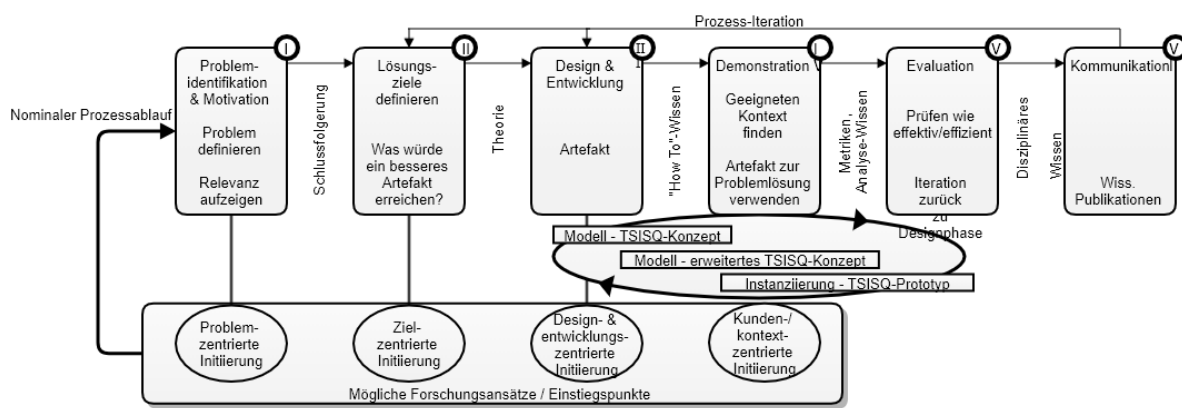


Abbildung 8: Forschungsdesign des DSR-Prozesses nach Peffers et al. (2007)

Die fehlende Automatisierung von Volltextsuchen in Kombination mit einer mangelnden Berücksichtigung semantischer Konzepte bzw. semantischer Ähnlichkeiten trägt wesentlich dazu bei, dass die Literaturrecherche im Rahmen des wissenschaftlichen Arbeitens ein langsamer, zeitintensiver (weil größtenteils manueller) Prozess ist. Um hier eine optimale technologische Unterstützung zu gewährleisten bedarf es eines effizienteren, toolgestützten Ansatzes, der den manuellen Anteil an der Literaturrecherche, beispielsweise beim Screening der initialen Suchergebnisse, auf ein Minimum reduziert. Diese Idee führte letztlich zu der Entwicklung von TSISQ. Unser Forschungsprozess beginnt folglich mit der Identifikation des o.g. Problems (I).

Es wurde zunächst eine umfassende Literaturrecherche zu etablierten Literaturrecherche-methoden in den Gebieten der Wirtschaftsinformatik, des IR, des NLP mit Fokus auf LSI durchgeführt (vgl. Abschnitt 2.2.1), um methodisch ein streng wissenschaftliches Vorgehen zu gewährleisten. Zusätzlich wurde ein gezieltes Literatur-Review im Bereich DSR durchgeführt. Vor dem Hintergrund der gestellten Forschungsfrage ist das Hauptziel (II), die in der Wissenschaft bereits etablierten Literaturrecherche-methoden durch einen toolgestützten, auf

LSI basierenden Ansatz zu optimieren. Im Hinblick auf diese Zielsetzung standen anschließend Design (III), Demonstration (IV) und Evaluation (V) des dafür vorgesehenen IT-Artefakts im Vordergrund. In Anlehnung an den Design-Zyklus nach Hevner (2007) wurden zudem auf Basis der o.g. theoretischen, wissenschaftlichen sowie praktischen Inputs mehrere Iterationen durchgeführt, um eine Verbesserung des Literaturrechercheprozesses zu erreichen und diesen wiederum direkt an den daraus entstandenen Veröffentlichungen (Koukal et al. 2013, 2014a, 2014b) zu verproben.

Im Detail wurden die einzelnen Schritte wie folgt durchgeführt: Nach einer weiteren Ausdifferenzierung der zugrundeliegenden Problemstellung sowie der Konkretisierung der durch das IT-Artefakt zu erfüllenden Kriterien (vgl. Abschnitt 2.1) wurde der erste Prototyp von TSISQ entwickelt (III), zunächst noch beschränkt auf die beiden zentralen Funktionalitäten: Die semantische Indexierung der zu durchsuchenden Volltexte und der Ähnlichkeitssuche basierend auf einer Suchanfrage in natürlicher Sprache auf diesem Index. Anschließend wurden mehrere Iterationen entsprechend der sechsten Richtlinie nach Hevner (2004), „Design as a Search Process“, durchgeführt, um das IT-Artefakt zyklisch weiterzuentwickeln. Auf diese Weise wurde der erste Prototyp von TSISQ um zusätzliche Parameter, eine Schicht für die automatisierte Aufbereitung der Content-Datenbank sowie ein Web-Frontend zur erleichterten Bedienung des Tools ergänzt. March und Smith (1995) klassifizieren die Ergebnisse designorientierter Forschung in Konstrukte, Modelle, Methoden und Instanzierungen. Der Prototyp von TSISQ kann neben dem zugrundeliegenden, formalen Modell als Instanzierung betrachtet werden. Die DSR-Prozesszyklen (vgl. Abbildung 8) wurden durch die Demonstration der Leistungsfähigkeit des Prototyps in Form von umfangreichen Tests (IV), deren Evaluation (V) und Dokumentation sowie der Kommunikation der Ergebnisse (VI) mittels der drei diesem Kapitel zugrundeliegenden, wissenschaftlichen Publikationen vollständig durchgeführt.

Durch weitere, umfassende Tests des Tools wurden die DSR-Prozesszyklen (III, IV und V) mehrfach vollständig durchlaufen, um eine Dokumentation der Forschungsergebnisse zu ermöglichen und das formale Modell wie auch den Prototyp iterativ zu verbessern. Gemäß der Klassifikation von Forschungsmethoden nach Palvia et al. (2006) wurden zwei Fallstudien in Form einer Literaturrecherche im Gebiet der Wirtschaftsinformatik durchgeführt, um die Leistungsfähigkeit des Prototyps und des formalen Modells zu demonstrieren (IV), auf dem TSISQ basiert (vgl. Koukal et al. 2013, 2014a). Die Evaluation von TSISQ (V) wurde mithilfe einer weiteren Fallstudie und einem Laborexperiment durchgeführt (vgl. Kapitel 2.4). Im Rahmen der zuletzt genannten Fallstudie wurden die Suchergebnisse von TSISQ innerhalb von

drei verschiedenen Suchräumen in einem direkten Benchmark mit den Ergebnissen eines gängigen, stichwortbasierten Suchverfahrens im Hinblick auf ihre Qualität verglichen.

Mit dem Laborexperiment wurde hingegen das Ziel verfolgt, innerhalb einer kontrollierten Umgebung bestehend aus zufällig (manuell) ausgewählten Artikeln nicht nur die falsch-positive, sondern auch die falsch-negative Fehlerrate des Ansatzes zu bewerten. Als letzter Schritt wurde die Veröffentlichung der gesammelten Forschungsergebnisse (unter Verwendung von TSISQ für die Verfeinerung der Literaturrecherche) im Rahmen der drei hier zusammengefassten, wissenschaftlichen Publikationen durchgeführt (VI).

2.3 Optimierung des Literaturrechercheprozesses mit TSISQ

Es ist eine grundlegende Voraussetzung für jedes Forschungsprojekt, möglichst vor dessen Beginn das bereits vorhandene Wissen in diesem Gebiet zu kennen (vgl. Hart, 1999). Wie in Abbildung 9 dargestellt, gilt dies nicht nur für das Literatur-Review als eigenständige Forschungsmethodik, sondern auch für Literaturrecherchen als Teil umfassenderer Forschungsmethoden, z.B. innerhalb des Design Science-Forschungsprozesses nach Hevner (2004). In den verschiedenen Forschungsmethoden der Wirtschaftsinformatik existieren zahlreiche Richtlinien für Literaturrecherchen mit diversen Teilschritten zu deren Durchführung. Bevor also der Versuch unternommen werden kann, den Literaturrechercheprozess „in der Forschung“ grundlegend zu optimieren ist es wichtig, zunächst eine konsolidierte Vorstellung dieses Prozesses aus den verschiedenen Forschungsmethoden zur Literaturrecherche abzuleiten. In Abbildung 9 werden die drei Kernpunkte als „gemeinsamer Nenner“ aller Literaturrecherchemethoden dargestellt.

Zunächst begreifen wir den Literaturrechercheprozess als Bestandteil eines Literatur-Reviews, das wiederum Bestandteil einer umfassenderen Forschungsmethodik sein kann. Innerhalb des Literaturrechercheprozesses sind drei Teilschritte zu beachten: Zunächst die Definition des Suchraums, wie z.B. die Festlegung eines oder mehrerer Fachgebiete und/oder der zu durchsuchenden Datenbasis. Der zweite Schritt ist der eigentliche Suchprozess, worin die potenziell vom Autor gesuchte Literatur innerhalb der gewählten Datenbasis identifiziert wird. Der dritte und letzte Schritt ist das (manuelle) Screening der zuvor identifizierten Literatur, um die finale Auswahl der Literatur zu treffen, die letztlich als Forschungsgrundlage verwendet und inhaltlich näher betrachtet werden soll. Da der Aufwand für das manuelle Screening im letzten Schritt von der „Treffsicherheit“ des vorangegangenen Suchprozesses abhängt, sehen wir das größte Potenzial für einen Effizienzgewinn innerhalb dieses Konstrukts in einer

Optimierung des Suchprozesses im Hinblick auf die Qualität und Relevanz der identifizierten Ergebnisse im Zusammenhang mit der Suchanfrage.

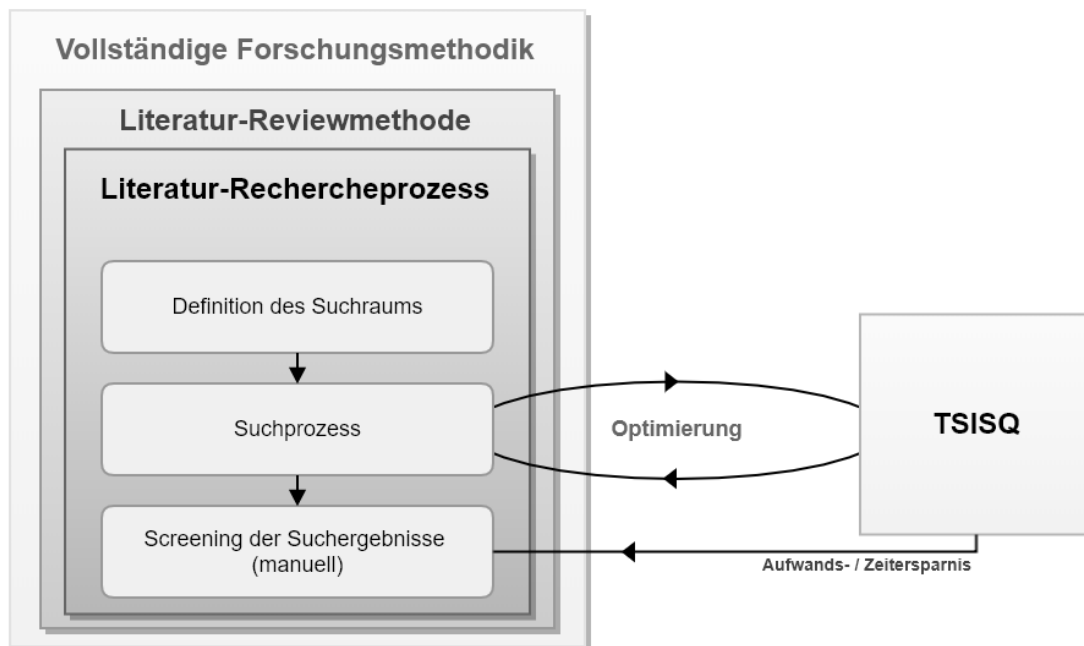


Abbildung 9: Optimierung des Literaturrechercheprozesses mit TSISQ

Vor dem Hintergrund der in Kapitel 2.1 genannten vier Kriterien führt dies zu einer Reduktion der irrelevanten Ergebnisse und spart somit Zeit bei der finalen Auswahl der im Anschluss manuell zu bewertenden Literatur. TSISQ setzt genau an dieser Stelle an: Durch die Anwendung einer Suchmethode, die im Gegensatz zu stichwortbasierten Suchverfahren semantische Ähnlichkeiten berücksichtigt und gleichzeitig, u.a. durch die Reduktion der Anzahl falsch-positiver Treffer, die Qualität der Suchergebnisse steigert, kann wertvolle Zeit bei der manuellen Sichtung eingespart werden.

2.3.1 Theoretischer Hintergrund und Funktionsweise

Der Kerngedanke bei der computergestützten Verarbeitung textueller Inhalte ist die Umwandlung von Dokumenten in einen Vektorraum, das sog. Vector Space Model (VSM) nach Salton et al. (1975). Dabei handelt es sich um den ersten Verarbeitungsschritt eines jeden Textdokumentes mit TSISQ. Der Inhalt jedes in Betracht kommenden Dokumentes D_i in dem zu durchsuchenden Textkorpus besteht aus einem oder (üblicherweise) mehreren Begriffen T_j . Innerhalb des VSM wird jedes dieser Dokumente (i) als ein t-dimensionaler Vektor im euklidischen Raum definiert, wobei t der Anzahl verschiedenartiger Begriffe in einem Dokument entspricht. Jeder dieser Begriffe erhält eine Gewichtung, die seine Relevanz widerspiegelt.

Folglich wird ein Dokument in TSISQ mathematisch beschrieben als

$$D_i = (d_{i1}, \dots, d_{ij}, \dots, d_{it}) \quad (1)$$

wobei d_{ij} die Gewichtung des j -ten Begriffs repräsentiert. Ein einfaches Kriterium zur Festlegung einer Gewichtung ist die Häufigkeit eines Begriffs in einem Dokument, stattdessen wenden wir hier jedoch das „Term Frequency-Inverse Document Frequency“ (TFIDF)-Konzept nach Salton und McGill (1986) an, um die statistische Bedeutung der Begriffe in einem Dokument zu messen. Die Anwendung der Worthäufigkeits- (Term Frequency, TF) und der inversen Dokumentenhäufigkeitsgewichtung (Inverse Document Frequency, IDF) steigert die Performance von Abfrage- und Kategorisierungssystemen (vgl. Maas et al., 2011) und damit auch die von TSISQ mittels der Abwertung üblicherer Stoppwörter und gleichzeitiger Aufwertung seltenerer Begriffe. Daraus ergeben sich zwei weitere Schritte, die anschließend durch Multiplikation kombiniert werden. Zuerst wird die normalisierte Häufigkeit eines Begriffs innerhalb eines Dokuments berechnet:

$$TF_{ij} = n_{ij} / \sum n_i \quad (2)$$

Dabei ist n_j die auftretende Anzahl eines Begriffs und $\sum n_i$ die Gesamtzahl aller Begriffe im betrachteten Dokument. Anschließend wird die Seltenheit eines Begriffs innerhalb des gesamten Textkorpus (der zu durchsuchenden Datenbasis) mittels der inversen Dokumentenhäufigkeit gemessen:

$$IDF_j = \log(N/df_j) \quad (3)$$

Wobei N die Gesamtanzahl der Dokumente im Textkorpus und df_j die Anzahl der Dokumente darstellt, die den Begriff j enthalten. Die Multiplikation beider Maßzahlen ergibt den statistischen Gewichtungsfaktor für jeden Begriff innerhalb eines Dokumentes:

$$TFIDF_{ij} = TF_{ij} \times IDF_j = \left(\frac{n_{ij}}{\sum n_i}\right) \times \log(N/df_j) \quad (4)$$

Als finales Ergebnis dieser Umwandlung der Dokumente im Textkorpus in das VSM und der anschließenden Anwendung des TFIDF-Konzeptes erhalten wir eine sogenannte „Term-by-Document Matrix“ X , deren Spalten die Gewichtungen der Begriffe für jedes einzelne Dokument im Textkorpus enthalten. Diese Umwandlung von Dokumenten beliebiger Länge hin zu Listen aus Zahlen mit fester Länge führt zu keiner Dimensionsreduktion bzgl. der zu verarbeitenden Inhaltmenge. Dies kann zu Problemen bei der Verarbeitung führen, insbesondere bei Suchen in einer Datenbasis großen Umfangs. Zudem werden durch die

Anwendung des VSM und des TFIDF-Konzeptes die Probleme mit Synonymie und Polysemie zunächst nicht adressiert.

Um eine Dimensionsreduktion zu erreichen und gleichzeitig die weiteren Defizite des TFIDF-Konzeptes zu beseitigen, wird die LSI-Methode auf die bereits konvertierte „Term-by-Document Matrix“ X angewandt. LSI kann somit als Erweiterung des VSM betrachtet werden: Häufigeres, gemeinsames Auftreten von Begriffen in einem Dokument wird erkannt und verwendet, um eine implizite semantische Struktur durch die gemeinsame Verwendung dieser Begriffe in Dokumenten zu erkennen (vgl. Zhang et al., 2011).

TSISQ zerlegt die „Term-by-Document Matrix“ in drei weitere Matrizen mittels eines Prozesses namens „Singular Value Decomposition“ (SVD) (vgl. Forsythe et al., 1977). Die Matrizen haben ein spezielles Format: T_0 und D_0 besitzen orthonormale Spalten und sind Matrizen linker und rechter Singulärvektoren, S_0 ist eine Diagonalmatrix aus Singulärwerten. Eine Zerlegung der Matrix X in dieses Format reicht jedoch noch nicht aus, denn die Multiplikation dieser drei Matrizen miteinander ergäbe wieder die Matrix X in ihrer ursprünglichen Form und brächte keinen Informationsgewinn mit sich. Im nächsten Schritt wird eine Dimensionsreduktion durchgeführt, um Informationen zu übergeordneten Beziehungen zwischen Begriffen zu berücksichtigen. Abbildung 10 zeigt die Kürzung der Matrizen $T_0 \times S_0 \times D_0'$ von m Dimensionen zu den neuen Matrizen $T \times S \times D'$ mit k Dimensionen. Die Multiplikation von $T \times S \times D'$ ergibt die beste Rang- k -Approximation der ursprünglichen Term-by-Document Matrix X (vgl. Kontostathis, 2007). Da die Anwendung der üblichen SVD-Prozedur sehr speicherintensiv ist, wenden wir einen inkrementellen SVD-Algorithmus nach Brand (2006) an. Eine beispielhafte Darstellung der Anwendung von VSM, TFIDF und LSI kann bei Sidorova et al. (2008) nachgelesen werden.

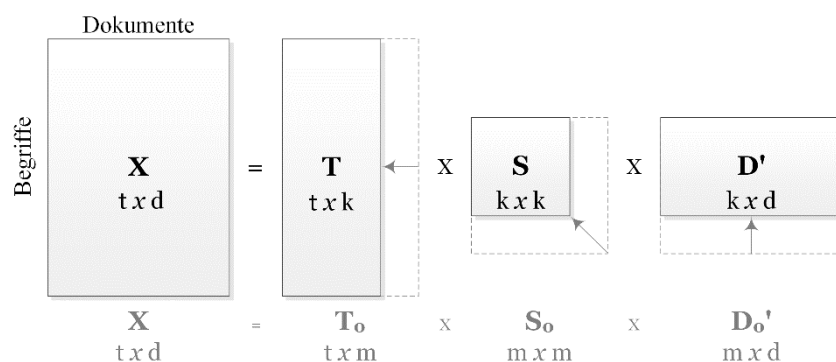


Abbildung 10: Reduzierte Singular Value Decomposition der Term-by-Document-Matrix

Im nächsten Schritt werden die Spalten der Matrix D' für die angestrebten Dokumentvergleiche und Ähnlichkeitssuchen verwendet. Jede Spalte steht für einen Vektor, der das aggregierte semantische Konzept des ursprünglichen Inhaltes eines Dokumentes widerspiegelt.

Um nun eine Suchanfrage mit den Dokumenten innerhalb der gewählten Datenbasis zu vergleichen, oder genauer, sie im Hinblick auf ihre Ähnlichkeit zueinander zu analysieren, konvertiert TSISQ den Inhalt der Anfrage zunächst in die entsprechende Repräsentation im VSM und transformiert diese anschließend in den gleichen Raum, in dem die o.g. Dokumentvektoren liegen:

$$Q = Q^T T S^{-1} \quad (5)$$

Der Vergleich zwischen den Dokumenten im Textkorpus, also der zu durchsuchenden Datenbasis, und der Suchanfragen wird mithilfe des Kosinus-Maßes durchgeführt, welches das am häufigsten verwendete Maß für die Feststellung von Ähnlichkeiten zwischen Dokumenten darstellt (vgl. Korenius et al., 2007). Anstatt den Winkel zwischen den Vektoren der Suchanfrage und der Dokumente miteinander zu vergleichen wird der Kosinus dieser Winkel berechnet. Da die TFIDF-Werte niemals negativ sein können, kann der Winkel zwischen zwei Vektoren nicht mehr als 90° betragen. Folglich liefert TSISQ letztlich Kosinuswerte in einem Intervall von $[0, 1]$, die die Ähnlichkeit zwischen der Suchanfrage und den Dokumenten innerhalb des Textkorpus ausdrücken: Je höher der Wert ist, desto höher ist die Ähnlichkeit.

2.3.2 Implementierung, Architektur und Benutzeroberfläche

Der TSISQ-Prototyp wurde in Python implementiert, um eine plattformübergreifende Nutzung zu ermöglichen. Das System (abgesehen von dessen Webfrontend) wurde jedoch bisher ausschließlich auf Linux-Systemen oder in virtuellen Linux-Maschinen unter Windows getestet. Die Systemkomponenten, ihre Funktionen und entsprechenden Datenströme werden in Abbildung 11 dargestellt und im Folgenden detailliert beschrieben.

Als Input werden ausschließlich wissenschaftliche Artikel im üblichen PDF-Format verwendet. Diese Dateien werden mit Xpdf in einfache Textdateien umgewandelt. Für die Anwendung der zuvor beschriebenen Methoden (VSM, TFIDF, SVD) wird das Softwareframework „gensim“ in Kombination mit „simserver“, einer übergeordneten Kontrollschicht, verwendet. Gensim ist ein NLP-Softwareframework, das auf der Idee des „Document Streaming“ basiert (Rehurek and Sojka 2010). Es benötigt die Open Source-Bibliotheken NumPy und SciPy. NumPy erlaubt die Manipulation n-dimensionaler Arrays, SciPy liefert die Routinen für numerische Integration

und Optimierung. Gemäß der Gensim-Dokumentation liegen die Vorteile des Frameworks in der schnellen und speicherunabhängigen Verarbeitung großer Datensätze, da die Term-by-Document Matrix nicht „in-memory“ vorgehalten werden muss.

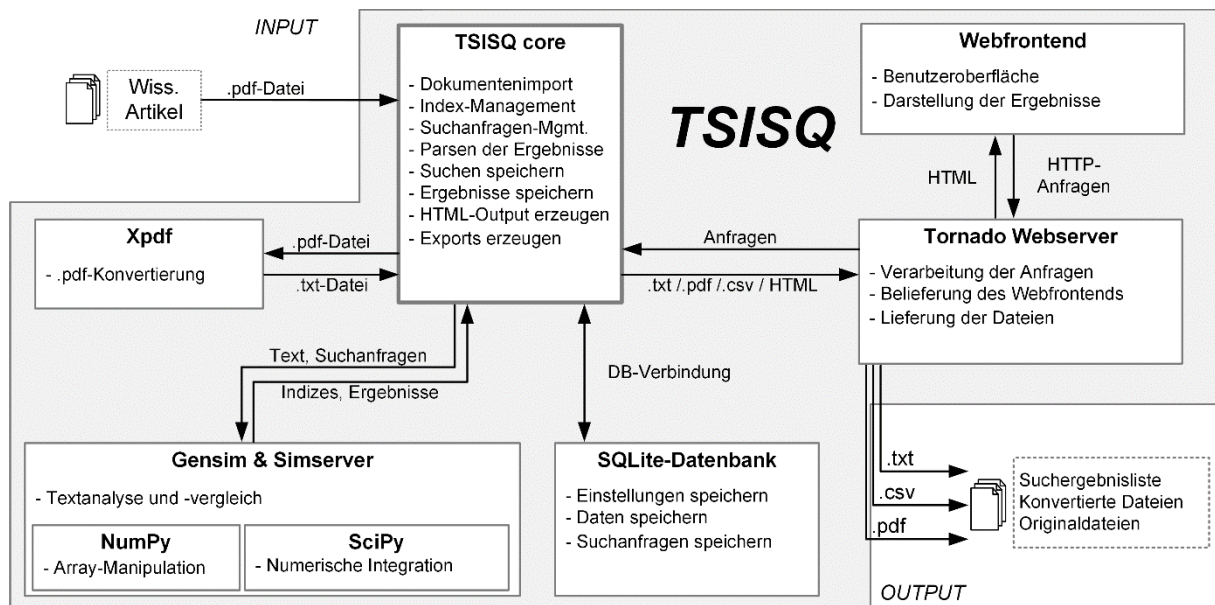


Abbildung 11: TSISQ-Systemarchitektur

Gensim kann den Input in Kleinbuchstaben umwandeln, Tokenisieren, Stemming (Stammform-/Normalformreduktion) durchführen sowie den Textinput in Unicode transformieren. Darüber hinaus erlaubt es die direkte Anwendung des SVD-Konzeptes auf eine Term-by-Document Matrix mit Worthäufigkeitsgewichtungen oder, wie im Falle von TSISQ, einer vorherigen Anwendung des TFIDF-Gewichtungsschemas. Obwohl alle Berechnungen in TSISQ speicherunabhängig sind, ist In-Memory-Processing deutlich schneller als die Verwendung einer Auslagerungsdatei, d.h. letztlich bestimmt die Größe des Arbeitsspeichers auch die Größe der bei optimaler Geschwindigkeit berechenbaren Matrizen. TSISQ ist somit frei skalierbar, genügend Arbeits- und Festplattenspeicher vorausgesetzt. Der Indexierungsprozess ist sehr rechen- und speicherintensiv, daher kann das initiale Training bei größeren Dokumentensammlungen länger dauern. Auf einem Intel Core i7-2640M-Prozessor mit 2.80 GHz und 8GB RAM benötigte dieser Prozess für 12.332 Dokumente etwa eine halbe Stunde. Als Ergebnis des Indexierungsprozesses erstellt TSISQ eine Index-Korpusdatei für die weitere Verarbeitung und Abfragen bereit. Wurde der Index bereits vorab berechnet, liefert die eigentliche Suchprozedur ohne nennenswerte Zeitverzögerung direkt die Ergebnisse.

TSISQ speichert die Informationen über Indizes, Inhalte, Dateigrößen der PDF- und Textdateien sowie die Suchhistorie in einer SQLite-Datenbank. Um nach Fertigstellung des rein funktionalen TSISQ-Prototyps die Benutzerinteraktion („request and response“) mit dem Tool

zu erleichtern und die Usability zu verbessern, wurde zusätzlich ein HTML-Webfrontend entwickelt. Dazu wurde das pythonbasierte Web-Framework Tornado verwendet, um die Benutzereingaben an die TSISQ-Kernkomponente weiterzuleiten und die Ergebnisse im HTML-Webfrontend übersichtlich darzustellen (siehe Abbildungen 12-15). Die grafische Darstellung des Webfrontends basiert auf dem Twitter Bootstrap-Framework.

#	Index name	Folder with pdf files	# of files	Add new files	Delete index
3	journals only	ais_basket_of_eight_2007_2013	1660	+ Add new files	🗑 Delete index
2	conference articles only	is_conferences_2007_2013	8626	+ Add new files	🗑 Delete index
1	main	journals_and_conferences_2007_2013	10286	+ Add new files	🗑 Delete index

Abbildung 12: TSISQ Index-Management

Abbildung 12 stellt die Eingabemaske für die Verwaltung der verschiedenen Indizes dar. Dies erlaubt eine komfortable Auswahl des gewünschten Suchraums, z.B. können, wie in der Abbildung dargestellt, die Journals des „AIS Basket of Eight“ und die Konferenzbeiträge der wichtigsten IS-Konferenzen getrennt voneinander oder auch gemeinsam durchsucht werden. Auch die Ablage von Indizes aus verschiedenen Forschungsgebieten ist somit möglich, ohne dass bei einem Wechsel des Index die für eine Berechnung der Indizes notwendige Zeit und Rechenleistung aufgebracht werden muss. Darüber hinaus können die verschiedenen, bereits vorberechneten Indizes inkrementell erweitert werden, was den zusätzlichen Zeit- und Rechenaufwand für diese Prozedur minimiert.

Abbildung 13 zeigt die Eingabemaske für die semantische Volltextsuche mit TSISQ. In der aktuellen Ausbaustufe werden sowohl frei formulierte Suchanfragen per Eingabe in das Textfeld als auch die Eingabe einer lokalen Datei im PDF- oder TXT-Format unterstützt. Die

letztere Möglichkeit dient beispielsweise dazu, zu vorhandener Basisliteratur semantisch ähnliche Artikel im gewählten Index zu identifizieren.

The screenshot shows the 'Performing a New Query' interface in TSISQ v0.1. The interface is divided into several sections:

- Navigation Bar:** Contains 'TSISQ v0.1', 'New Query' (selected), 'Query history', and 'Manage Indexes'.
- Select .txt or .pdf file to upload:** A file selection area with a 'File' input field and a 'select file' button.
- Insert query text:** A text input area containing the following text: "Recent academic investigations of computer security policy violations have largely focused on non-malicious noncompliance due to poor training, low employee motivation, weak affective commitment, or individual oversight. Established theoretical foundations applied to this domain have related to protection motivation, deterrence, planned behavior, self-efficacy, individual adoption factors, organizational commitment, and other individual cognitive factors. But another class of violation demands greater research emphasis: the intentional commission of computer security policy violation, or insider computer abuse. Whether motivated by greed, disgruntlement, or other psychological processes, this act has the greatest potential for loss and damage to".
- choose index to search in:** A dropdown menu with 'main' selected and a 'select' button.
- Start Query:** A button with a magnifying glass icon and the text 'Start Query'.

Abbildung 13: Eingabemaske für neue Suchanfrage in TSISQ

Diese Funktion kann insbesondere zu Beginn eines Forschungsvorhabens dazu dienen, sich einen ersten Überblick über die in einem Forschungsgebiet bereits existierende Literatur zu verschaffen und/oder bestehende Forschungslücken zu identifizieren, bevor in weiteren Iterationen des (teil-)automatisierten Literaturrechercheprozesses immer spezifischere Suchanfragen formuliert werden, die letztlich zu einem entsprechend (semantisch) spezifischeren „Ergebnisraum“ führen. In dieser Maske können zudem die im oben beschriebenen Index-Management entsprechend organisierten Indizes ausgewählt werden, in denen die semantische Suche durchgeführt werden soll.

Nachdem der zu durchsuchende Index ausgewählt, die Suchanfrage formuliert und die Suche durch Klick auf „Start Query“ ausgeführt wurde, erscheint ohne nennenswerte Verzögerung bereits die Suchergebnisliste (vgl. Abbildung 14).

Query Results

Query input

Recent academic investigations of computer security policy violations have largely focused on non-malicious noncompliance expand ▾

Result list Download Results as txt-file

Rank	Title of document	Similarity score
1	MISQ_2013_An_Expanded_View_of_Employee_Computer_Abuse	92.3%
2	EJIS_2011_A_review_and_analysis_of_deterrence_theory_in_the_IS_security_literature_making_sense_of_the_disparate_findings	86.7%
3	MISQ_2010_Neutralization_New_Insights_into_the_Problem_of_Employee_Systems_Security_Policy_Violations	85.5%
4	ICIS_2010_UNDERSTANDING_THE_EFFECT_OF_DETERRENCE_MECHANISMS_ON_CYBERLOAFING_EXPLORING_A_GENERAL_DETERRENCE_MODEL_WITH_A_SOCIAL_PERSPECTIVE	85.4%
5	AMCIS_2007_Managing_the_Dark_Side_of_Computer_Use_at_Work_A_Typology_of_Inf	80.2%
6	AMCIS_2009_Measuring_Severity_of_Internet_Abuse_in_the_Workplace_Creation_o	80.0%

Abbildung 14: TSISQ-Suchergebnisliste, absteigend sortiert nach Ähnlichkeits-Score

Die Suchergebnisse sind nach ihrer semantischen Ähnlichkeit zur Suchanfrage absteigend sortiert, wobei ein Ähnlichkeits-Score von 100% darauf hindeutet, dass im Index ein Text identifiziert wurde, der identisch zur Suchanfrage ist. Die zugehörigen Volltexte zu den einzelnen Ergebnissen können zudem per Klick auf das Suchergebnis direkt geöffnet werden.

Query History

Search #	Date / time	Query text	Show results
253	2013-11-29 13:27	Recent academic investigations of computer security expand ▾	Show results
252	2013-11-29 13:12	With China emerging as a new frontier of global IT ou expand ▾	Show results
251	2013-11-29 12:58	Abstract. Virtual worlds or three-dimensional compute expand ▾	Show results
250	2013-11-27 15:44	Online communities are increasingly important to organizations and the general public, but there is little theoretically based research on what makes some online communities more successful than others. In this article, we apply theory from the field of social psychology to understand how online communities develop member attachment. collapse ▾	Show results
249	2013-11-27 15:41	Firms invest in a variety of information technologies a expand ▾	Show results
248	2013-11-27 14:58	Taking a control theory view of software process inno expand ▾	Show results
247	2013-11-27 14:51	Abstract While many corporations and Information Sy expand ▾	Show results
246	2013-11-27 13:17	This study develops a research model of how the tec expand ▾	Show results
245	2013-11-27 12:59	Despite the tremendous commercial success of gene expand ▾	Show results

Abbildung 15: TSISQ-Suchhistorie

Alle bisher durchgeführten Suchanfragen werden letztlich von TSISQ in einer Datenbank gespeichert und können über die in Abbildung 15 dargestellte Suchhistorien-Maske im TSISQ-Webfrontend nachvollzogen und bei Bedarf erneut aufgerufen werden.

2.4 Demonstration, Evaluation und Benchmark von TSISQ

Mit dem Ziel, Eignung und Nutzen von TSISQ für den Literaturrechercheprozess zu bewerten und messbar zu machen, wurden in mehreren Anwendungsfällen verschiedene Tests durchgeführt: Ein direkter Vergleich mit den Ergebnissen einer stichwortbasierten Suchmaschine (mit Unterstützung von Fachexperten) und ein Laborexperiment in einer kontrollierten Umgebung. Mit einer manuellen Literaturrecherche würde sinnvollerweise in den führenden, wissenschaftlichen Zeitschriften des jeweiligen Fachgebiets begonnen, da es wahrscheinlich ist, dass die wichtigsten Beiträge dort veröffentlicht wurden (vgl. Webster und Watson, 2002). Folglich wurden als Suchraum für das Gebiet der Wirtschaftsinformatik die acht führenden Zeitschriften im AIS Senior Scholars‘ Basket, dem „AIS Basket of Eight“, ausgewählt. Zusätzlich wurden Veröffentlichungen der vier wichtigsten Konferenzen in der Wirtschaftsinformatik dem Index hinzugefügt: Die International Conference on Information Systems (ICIS), die European Conference on Information Systems (ECIS), die Americas Conference on Information Systems (AMCIS) und die Hawaiian Conference on System Sciences (HICSS). Detailliertere Informationen bzgl. der Gesamtzahl der Artikel und deren Anteil pro Zeitschrift oder Konferenz am Gesamtindex können Tabelle 3 entnommen werden.

Tabelle 3: Datenbasis aus wiss. Journal- und Konferenzbeiträgen (2007-2014)

Journal	Summe	EJIS	ISJ	ISR	JAIS	JIT	JMIS	JSIS	MISQ
Artikel	1.685	372	166	323	219	223	280	147	326
Anteil (%)	13,66%	3,02%	1,35%	2,62%	1,78%	1,81%	2,27%	1,19%	2,64%
Konferenz	Summe	AMCIS	ECIS	HICSS	ICIS				Summe
Artikel	10.647	3.604	1.599	3.681	1.763				12.332
Anteil (%)	86,34%	29,22%	12,97%	29,85%	14,30%				100%

2.4.1 Benchmark mit einer stichwortbasierten Suchmaschine

Um TSISQ für eine Verwendung in der wissenschaftlichen Praxis empfehlen zu können reicht ein einfacher Beweis dafür, dass das Tool zutreffende Suchergebnisse liefert, bei weitem nicht aus. Vielmehr muss eine höhere Qualität der Suchergebnisse im direkten Vergleich zu etablierten, wissenschaftlichen Suchmaschinen erreicht werden. In dieser ersten Fallstudie wurde zu diesem Zweck AISel, eine der führenden wissenschaftlichen Datenbanken in der Wirtschaftsinformatik, als Benchmark betrachtet. Es wurden drei verschiedene Forschungsgebiete für einen Direktvergleich herangezogen: (1) „Enterprise Content

Management“ (ECM), da es sich dabei um eine klar abgegrenzte Domäne im Gebiet der Wirtschaftsinformatik handelt, (2) „IT-Sicherheit und Datenschutz“ (engl. „IS Security and Privacy“), da es sich dabei um ein vergleichsweise breites und interdisziplinäres Forschungsfeld handelt, und (3) „Strukturgleichungsmodellierung“ (engl. „Structural Equation Modeling“, kurz SEM), um nicht nur einzelne Fachgebiete, sondern auch eine Forschungsmethode zu betrachten, die in den verschiedensten Fachgebieten innerhalb der Wirtschaftsinformatik zum Einsatz kommt.

Für jede der o.g. Kategorien wurde eine Sammlung von Stichwörtern ausgewählt, die anschließend als Input für die stichwortbasierte Suche (a) bei AISel und die stichwortbasierte Suche mit TSISQ (b) verwendet wurden. Zusätzlich wurden pro Kategorie passende Suchanfragen, bestehend aus Volltexten in natürlicher Sprache, als Input für die TSISQ-Volltextsuche (c) ausgewählt. Tabelle 4 zeigt die Stichwörter und Texte, die als Suchanfragen für TSISQ und AISel verwendet wurden. Für die AISel-Stichwortsuche wurden die Suchbegriffe in Anführungszeichen gesetzt, während bei TSISQ die Stichwörter als Volltext ohne Anführungszeichen eingegeben wurden.

Tabelle 4: Abfragen zum Suchmaschinenvergleich in ausgewählten Forschungsgebieten

Forschungsgebiet	AISel und TSISQ Stichwortsuche	TSISQ-Volltextsuche
Enterprise Content Management (ECM)	- Enterprise Content Management - ECM	Abstract eines Literatur-Reviews von Rickenberg et al. (2012).
IT-Sicherheit und Datenschutz	- IT security - IS security - IT privacy - IS privacy	Beschreibung des „IS Security and Privacy“-Tracks der ICIS 2014.
Strukturgleichungsmodellierung	- Structural equation model - Structural equation modeling - Structural equation modelling - SEM	Abstract eines Artikels bzgl. der Evaluation von Indikatoren in der Strukturgleichungsmodellierung von Bollen (2011).

Um die Vergleichbarkeit der Ergebnisse der beiden stichwortbasierten Suchmethoden (a) und (b) sicherzustellen, wurde die Analyse auf diejenigen Zeitschriften und Konferenzen beschränkt, die in beiden Indizes gleichermaßen enthalten sind (AMCIS, ECIS, ICIS, ISJ, JAIS, MISQ). Für die Suchmethode (c) mit natürlicher Sprache/Volltext als Input wurden die Suchanfragen wie folgt ausgewählt: Für den Bereich ECM wurde ein umfassender Abstract eines Literatur-Reviews verwendet, weil darin ein Großteil der relevanten, ECM-bezogenen Begriffe enthalten ist. Im Bereich IT-Sicherheit wurde die komplette Beschreibung des Tracks „IT Security and Privacy“ der ICIS 2014 unter der Annahme verwendet, dass auch dieser die wichtigsten Begriffe und semantischen Konzepte des Fachgebiets enthält. Analog wurde für

das Forschungsgebiet des SEM der Abstract eines MISQ-Artikels ausgewählt, der sich nicht nur mit einem einzelnen Anwendungsfall von SEM, sondern mit verschiedenen SEM-Ansätzen befasst.

Der Fokus dieser ersten Fallstudie liegt auf der Relevanz der Suchergebnisse für das jeweilige Forschungsgebiet, da es sich im Rahmen des Literaturrechercheprozesses dabei um das wichtigste Qualitätskriterium handelt. Um einen vergleichbaren, gemeinsamen Qualitätsmaßstab für die Suchergebnisse der drei Suchmethoden herzustellen, wurde eine dreistufige Likert-Skala definiert: Eine Bewertung von „0“ klassifiziert einen Artikel (bezogen auf die Suchanfrage) als *irrelevant*, während „1“ als *relevant* und „2“ als *hochrelevant* eingestuft wird. Um die Unabhängigkeit dieses Klassifikationsvorgangs zu gewährleisten wurde die Bewertung von Experten in den jeweiligen Fachgebieten vorgenommen. Sie bekamen dazu die Top 50-Suchergebnisse der drei Suchmethoden geliefert und wurden gebeten, jedes davon einer der drei o.g. Relevanzkategorien zuzuordnen. Um dabei Objektivität zu gewährleisten, wurde den Experten während der Bewertung das ursprüngliche Ranking der Suchergebnisse vorenthalten.

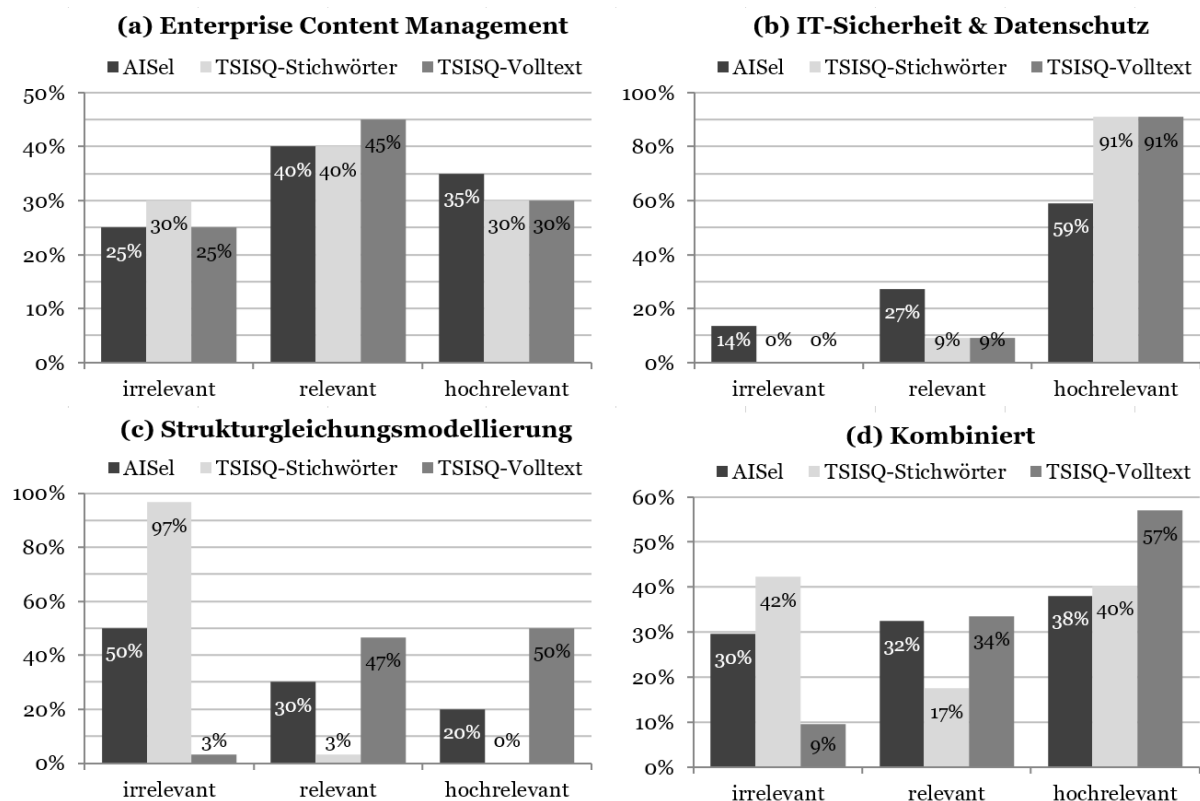


Abbildung 16: Benchmark-Ergebnisse zu Suchmethoden

Die einzelnen Ergebnisse je Forschungsgebiet sind in Abbildung 16 (a)-(c) grafisch aufbereitet. Während im Bereich ECM (Abbildung 16a) kaum ein Unterschied der Ergebnisqualität von TSISQ und AISel zu beobachten ist, zeigt Abbildung 16b insbesondere in der Kategorie der

hochrelevanten Literatur eine deutliche Qualitätsverbesserung bei der Suche mit TSISQ. Zudem fällt hier auf, dass TSISQ im Bereich IT-Sicherheit und Datenschutz keine als irrelevant klassifizierte Literatur identifiziert. Im Bereich SEM (Abbildung 16c) müssen die beiden verschiedenen TSISQ-Varianten getrennt voneinander betrachtet werden: Während die stichwortbasierte Suche fast ausschließlich irrelevante Artikel liefert, ist die Variante mit Volltext-Input in natürlicher Sprache wiederum leistungsfähiger als AISel in Bezug auf relevante und hochrelevante Literatur. Abbildung 16d zeigt schließlich eine Aggregation der Ergebnisse aus den drei Suchszenarien. Die AISel-Suchergebnisse verteilen sich zu etwa gleichen Teilen auf die drei Relevanzkategorien, während die TSISQ-Stichwortsuche sowohl einen hohen Anteil irrelevanter als auch hochrelevanter Artikel identifiziert. Besonderes Augenmerk sollte jedoch an dieser Stelle auf die verbliebene Suchvariante, TSISQ mit Volltext-Input, gelegt werden: Der Anteil irrelevanter Literatur in den Suchergebnissen ist deutlich geringer als bei den beiden stichwortbasierten Varianten, während der Anteil relevanter und hochrelevanter Literatur in den Suchergebnissen die höchsten Werte erreicht. Eine ausführlichere Diskussion der Ergebnisse folgt in Kapitel 2.5.

2.4.2 Identifikation der falsch-positiven und falsch-negativen Fehlerrate

Das im Folgenden vorgestellte Laborexperiment dient einer tiefergehenden Evaluation von TSISQ und soll die zu Beginn dieses Forschungsprojektes getroffenen Annahmen bzgl. der Eignung und der Vorteile einer Verwendung von LSI für die semantische Ähnlichkeitserkennung zwischen Texten nochmals unterstreichen. Es wurde bereits hervorgehoben, dass eine LSI-basierte, semantische Ähnlichkeitssuche dabei helfen kann, die Herausforderungen, die linguistische Phänomene wie Synonymie und Polysemie mit sich bringen, zu überwinden. Damit einher geht die Annahme, dass eine LSI-basierte Suchmethode die Anzahl falsch-positiver und falsch-negativer Suchergebnisse messbar reduziert, sprich weniger irrelevante Suchtreffer findet und weniger relevante Artikel „übersieht“.

Während falsch-positive Ergebnisse (wie z.B. thematisch unpassende Artikel in den oberen Rängen der Suchergebnisliste) vergleichsweise leicht zu identifizieren sind, ist die Identifikation falsch-negativer Ergebnisse deutlich komplexer, da sie sich per Definition nicht in der Suchergebnisliste befinden. Um die falsch-negative Fehlerrate einer Suchmethode dennoch analysieren zu können ist es eine Grundvoraussetzung, dass alle Dokumente im zu durchsuchenden Index inhaltlich bekannt sind.

Ausgehend davon und von den Ergebnissen unserer Pilotstudie (Koukal et al., 2013), dass die überwiegende Mehrheit der für die Abfrage relevantesten Beiträge in den Top 25-Ergebnissen zu finden ist, wurde ein gezielt manipulierter Index erstellt. Dieser setzt sich zusammen aus 100 zufällig ausgewählten, wissenschaftlichen Artikeln aus Konferenzen und Zeitschriften in unserer Gesamt-Datenbank (vgl. Tabelle 3). Der Hauptzweck dieses Verfahrens besteht darin, den Index klein genug zu halten, um gleichzeitig signifikante und dennoch kontrollierbare Ergebnisse in einer kontrollierten Umgebung zu erhalten.

Um eine Selektionsverzerrung durch die manuelle Auswahl besonders relevanter oder eindeutig irrelevanter Artikel zu einem bestimmten Thema zu vermeiden, wurden im ersten Schritt wie zuvor beschrieben 100 Artikel zufällig ausgewählt. Anschließend wurde jeder in diesem kleinen Index enthaltene Artikel manuell von einem Fachexperten überprüft und in zwei Kategorien eingeteilt. Die erste Kategorie enthielt nur wissenschaftliche Arbeiten, die sich auf „IT-Sicherheit und Datenschutz“ beziehen und daher mit hoher Wahrscheinlichkeit semantisch ähnlich sind. Die zweite Kategorie enthielt alle Artikel, die nicht zu diesem Bereich gehören und sich mit verschiedensten Themengebieten der Wirtschaftsinformatik befassen.

Um die Qualität unserer Datenbank für dieses Experiment sicherzustellen, wurde die Kategorisierung des Fachexperten mit den Klassifikationen der Experten unserer oben dargestellten Fallstudie zum Thema „IT-Sicherheit und Datenschutz“ abgeglichen. Für den Fall, dass ein Artikel von TSISQ innerhalb der ersten Fallstudie gefunden wurde und derselbe Artikel unter den 100 Artikeln für dieses Experiment enthalten ist, mussten die Klassifikationen in "relevant" und "irrelevant" identisch sein. Die Einteilung in diese beiden Kategorien führte zu einem Satz von 18 für das Themengebiet „IT-Sicherheit und Datenschutz“ relevanten Artikeln, die restlichen 82 Artikel im Index wurden demnach als irrelevant eingestuft. Als Input für die Suche wurde erneut die vollständige Beschreibung des ICIS-2014-Tracks zu „IT-Sicherheit und Datenschutz“ verwendet.

Aufgrund des sehr kleinen Index und um zu vermeiden, dass alle 100 darin enthaltenen Artikel als Suchergebnisse ausgegeben werden, wurde die Ergebnisliste bei einem (vergleichsweise niedrigen) Kosinuswert von 0,01 abgeschnitten. TSISQ lieferte eine Ausgabe von 23 Ergebnissen, obwohl im Gesamtindex nur 18 Artikel von 100 zur Zieldomäne „IT Sicherheit und Datenschutz“ gehörten. Die 23 gefundenen Ergebnisse enthielten dabei 15 relevante und 8 irrelevante Artikel.

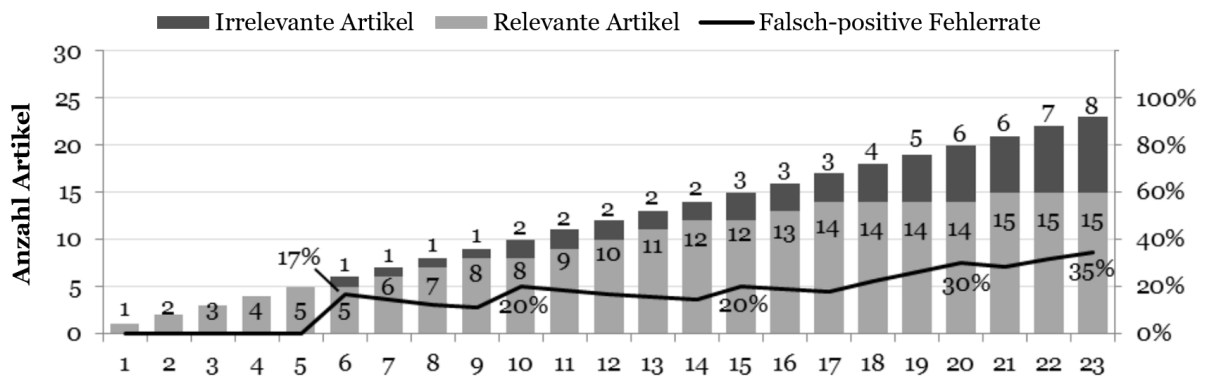


Abbildung 17: Ergebnisse des Laborexperiments

Folglich wurden drei der 18 insgesamt relevanten Artikel nicht identifiziert, was einer falsch-negativen Fehlerquote von 17 Prozent entspricht. Dabei waren die ersten fünf Suchergebnisse mit einer Trefferquote von 100 Prozent am genauesten (siehe Abbildung 17). Während im Bereich der Top 17-Ergebnisse die durchschnittliche falsch-positive Fehlerrate mit weniger als 20 Prozent immer noch auf einem akzeptablen Niveau liegt, erhöht sich diese Rate auf 35 Prozent, wenn alle 23 Suchergebnisse berücksichtigt werden. Die Ergebnisse des zweiten Kapitels der vorliegenden Dissertation werden im folgenden Abschnitt aufgegriffen und kritisch diskutiert.

2.5 Kritische Würdigung und Diskussion der Ergebnisse

Die Fallstudie in Abschnitt 2.4.1 wurde durchgeführt, um zu demonstrieren und zu bewerten, ob ein LSI-basierter Ansatz im direkten Vergleich mit einer etablierten, stichwortbasierten Suchmaschine wie AISel qualitativ hochwertigere Suchtreffer liefert, sprich eine geringere falsch-positive Fehlerrate erreicht. Die Ergebnisse zeigen, dass TSISQ sehr gut geeignet ist, wissenschaftliche Literaturrechercheprozesse zu verbessern, insbesondere bei der Verwendung von Volltext in natürlicher Sprache mit Bezug zur Zieldomäne als Input. In allen drei Suchszenarien (a-c) generiert die TSISQ-Volltextsuche qualitativ mindestens gleichwertige Ergebnisse hinsichtlich der Relevanz der gefundenen Artikel. In den Suchfällen (b) und (c) übertrifft TSISQ-Volltext die stichwortbasierten Ansätze (AISel- und TSISQ-Stichwörter) deutlich. In Szenario (a), innerhalb des ECM-Fachgebiets, ist die Qualität der Ergebnisse aller getesteten Verfahren nahezu identisch, die TSISQ-Textsuche schneidet im direkten Vergleich minimal schlechter ab. Um die teils großen Unterschiede zwischen den einzelnen Ergebnissen zu verstehen bedarf es einer detaillierteren Analyse. Dabei ist der Inhalt der Suchanfrage der offensichtlichste und wichtigste variable Faktor, da das Setup der Suchszenarien fix und die Datenbasis bei allen Suchen identisch ist.

Nach Erhalt der Evaluationsergebnisse wurden alle Suchanfragen rückblickend betrachtet. Folgende Begründungen für die abweichende Performance in den Suchszenarien (a)-(c) sind denkbar: Zunächst ist der Bereich ECM im Vergleich zu SEM und „IT-Sicherheit und Datenschutz“ stärker abgegrenzt, während die beiden letztgenannten Forschungsfelder einen eher universellen Charakter aufweisen und stärker fachgebietsübergreifend Anwendung finden. Zweitens enthält die Suchanfrage des Suchszenarios (c) eindeutig spezifischere Begriffe und semantische Konzepte mit Bezug zu SEM, was sich in den qualitativ deutlich besseren Ergebnissen der TSISQ-Volltextvariante zeigt. In Anbetracht der Tatsache, dass es sich bei SEM um ein statistisches Verfahren und nicht notwendigerweise um ein eigenes Forschungsgebiet handelt und auch, weil SEM in so gut wie jedem Fachgebiet Anwendung finden kann, muss die gute Performance von TSISQ-Volltext an dieser Stelle noch einmal näher betrachtet werden: Es wäre davon auszugehen gewesen, dass bei Artikeln, in denen die Anwendung von SEM in einem beliebigen Fachgebiet dargestellt wird, dieses Fachgebiet wiederum die mittels TSISQ-Volltext extrahierten, semantischen Konzepte innerhalb des Artikels dominiert. Die Ergebnisse von Suchszenario (c) widersprechen jedoch dieser Annahme: Die in diesem Fall verwendete Suchanfrage beschreibt einen „meta-research“-Artikel, der sich explizit mit SEM als Verfahren und nicht mit dessen Anwendung in einem konkreten Anwendungsfall befasst. Anscheinend ist TSISQ zu der Identifikation von Artikeln, die einer bestimmten Forschungsmethodik folgen, nicht gut geeignet. Es bestätigt sich jedoch die Annahme, dass TSISQ besonders nützlich bei der Identifikation *thematisch verwandter* Texte ist.

Anders als die AISEL-Suche behandelt TSISQ Ausdrücke wie "structural equation modeling" nicht als einen einzigen Suchbegriff, sondern behandelt jedes einzelne Wort gemäß dem im Kapitel 2.3.1 beschriebenen Verfahren. Die Begriffe werden zu ihrer Repräsentation im VSM konvertiert, ihre jeweiligen Gewichtungen werden durch die Anwendung des TFIDF-Konzepts bestimmt und anschließend mittels LSI transformiert. Da nicht-alphanumerische Zeichen während dieses Prozesses entfernt werden, macht es keinen Unterschied für eine Suche mit TSISQ, ob die Begriffe in Anführungszeichen gesetzt wurden oder nicht. Obwohl im Suchszenario (c) der TSISQ-Stichwortsuche nach SEM die Begriffe "model", "modeling" und "modelling" einzeln relativ niedrige TFIDF-Gewichte erhalten, so haben sie gemeinsam nur einen geringen Einfluss auf den Vektor eines Gesamtdokuments. Diese Begriffe können zusammen in vielen Dokumenten in der Datenbank gefunden werden, da das Ergebnis eines "modeling" -Prozesses normalerweise ein "model" ist. Selbst wenn der Begriff "modeling" konsequent in der gleichen Schreibweise verwendet wird (was eher selten der Fall ist), so

kommen beide Schreibweisen ("modeling" und "modelling") spätestens im Literaturverzeichnis eines verwandten Artikels vor. Folglich identifiziert TSISQ, der Suchanfrage entsprechend, Dokumente, die jeden dieser Begriffe enthalten. Das bedeutet, dass beispielsweise Literatur über Geschäftsprozessmodellierung oder beliebige statistische „Modelle“ zwar unerwünschte, aber durchaus nachvollziehbare Suchergebnisse darstellen. Da die Begriffe "model" und "modeling" in fast jedem Forschungsbereich der Wirtschaftsinformatik in vielen verschiedenen Kontexten auftreten, führt eine Stichwortsuche mit TSISQ daher möglicherweise zu diffusen Ergebnissen. Dasselbe gilt im Übrigen auch für den Begriff „equation“, da Gleichungen, auch im Zusammenhang mit Modellen („model“), eine häufig vorkommende Wortkombination in der Literatur verschiedenster Teilgebiete der Wirtschaftsinformatik darstellen.

Im Gegensatz dazu findet die TSISQ-Stichwortsuche bei der Suche in einer ausschließlich aus Wirtschaftsinformatik-Literatur bestehenden Datenbank nur relevante Artikel für das Suchszenario (b), „IS security and privacy“, da erstens die Schreibweise der Begriffe „security“ und „privacy“ eindeutig ist, zweitens die Begriffe „IS“ und „IT“ im Englischen als Stoppwörter identifiziert und somit herausgefiltert werden, da TSISQ nicht zwischen Groß- und Kleinschreibung unterscheidet und drittens die Aggregation der Begriffe "Sicherheit" und "Datenschutz" zu einem semantischen Konzept voraussichtlich genau die erwarteten Suchergebnisse liefert. Im Falle einer erweiterten, nicht auf fachgebietsspezifische („IS“- oder „IT“-) Literatur beschränkten Datenbank sind weitaus weniger genaue Suchergebnisse zu erwarten, zumindest wenn zusammenhanglose Stichwörter als Suchanfrage für TSISQ verwendet werden.

Die Ergebnisse der verschiedenen Suchszenarien deuten letztlich darauf hin, dass die Ergebnisse der TSISQ-Volltextsuche die der stichwortbasierten Suchverfahren übertreffen werden, da der Benutzer keine bestimmten Konventionen bei der Formulierung der Suchanfrage befolgen muss. Wenn in AISel und vergleichbaren, stichwortbasierten Suchmaschinen die notwendigen Konventionen vernachlässigt werden, sind die Suchergebnisse mit hoher Wahrscheinlichkeit nutzlos. Zum Beispiel liefert ein vollständiger Abstract als Suchanfrage bei AISel entweder keine Ergebnisse (bei der Anwendung von „AND“-Bedingungen), oder die Ergebnisse bestehen (bei der Anwendung von „OR“-Bedingungen) aus einer völlig zufälligen Auswahl an Dokumenten aus den verschiedensten Forschungsgebieten.

Die hier dargestellte Forschung zielt explizit *nicht* darauf ab, bestehende Suchansätze vollständig zu ersetzen. Stattdessen soll der Literaturrechercheprozess durch die Identifizierung

semantisch ähnlicher Literatur erleichtert und so optimiert werden. Es wird vorausgesetzt, dass Wissenschaftlern, die TSISQ einsetzen, mindestens ein wissenschaftlicher Artikel oder Text zu ihrem gewählten Forschungsgebiet bekannt ist. Unter dieser Annahme würde die Definition einer Reihe geeigneter Stichwörter, die nicht nur ein gesamtes Forschungsfeld, sondern auch die spezifische Subdomäne einschließlich aller relevanten Synonyme und/oder verwandter Begriffe abdecken, zusätzlichen Aufwand erfordern. Bei einer zumindest identischen oder sogar überwiegend besseren Qualität der Suchergebnisse (siehe Abbildung 16a-d) führt insbesondere der TSISQ-Volltext-Suchansatz, der keine Transformation bestehender Texte in ein spezielles Format für die Abfrage erfordert, zu einer nachweisbaren Effizienzsteigerung des Literaturrechercheprozesses.

Die Pilotstudie zu diesem Themenkomplex (Koukal et al. 2013) zeigt, dass die Chance, gute Ergebnisse in den Top 25-Suchtreffern zu erzielen, potenziell hoch ist. Wenn jedoch die Anzahl der Artikel, die sich mit dem gesuchten Inhalt befassen, aufgrund der Größe des Index unbekannt ist, kann über die Performance der Suchmaschine keine Aussage getroffen werden. Im Falle einer vergleichsweise großen Anzahl passender Artikel im Index, z.B. 250 Stück, wäre somit die Tatsache, dass die Suchgenauigkeit nach 25 Ergebnissen stark abnimmt, kein befriedigendes Ergebnis. Wenn im Gegensatz dazu die Menge an geeigneten Artikeln in dem gleichen, großen Index relativ niedrig ist, z.B. 20 Stück, ergäbe die Identifikation von 16 Artikeln innerhalb der ersten 25 Ergebnisse eine bemerkenswert hohe Suchgenauigkeit.

Während die Suchgenauigkeit und die Ergebnisqualität von TSISQ im Vergleich zu der stichwortbasierten Suchmaschine von AISel hier bereits diskutiert wurde, erlaubt das Laborexperiment zusätzlich eine Quantifizierung der falsch-positiven und falsch-negativen Fehlerrate. Dabei geht es insbesondere um die Frage, ob die zuvor dargestellten, vielversprechenden Ergebnisse letztlich aus einer unerwartet großen Anzahl ECM-, IT-Sicherheit und Datenschutz- oder SEM-bezogener Publikationen im Index oder aus der guten Performance von TSISQ resultieren. In dieser kontrollierten Umgebung beträgt die durchschnittliche Suchgenauigkeit mindestens 65% innerhalb der oberen 23 Ergebnisse. Innerhalb der ersten 17 Ergebnisse liegt die durchschnittliche Suchgenauigkeit jedoch nie unter 80%. Folglich schneidet TSISQ etwas schlechter ab als in der Pilotstudie (75%) (vgl. Koukal et al. 2013), bestätigt aber dennoch die insgesamt bemerkenswerten Ergebnisse der Fallstudie. Die geringe Menge an relevanten Artikeln in der Datenbank des Experiments führt zu großen Abweichungen bei der durchschnittlichen Suchgenauigkeit, wenn ein relevanter Artikel in der Ergebnisliste einen höheren Rang aufweist. Anders ausgedrückt kann ein zusätzlicher Treffer oder ein geringfügig anderer Kosinuswert eines Artikels zu sehr unterschiedlichen Ergebnissen

führen, besonders in den oberen Rängen der Ergebnisliste. Obwohl die Ergebnisse des Laborexperiments aufgrund ihrer Empfindlichkeit gegenüber leichten Veränderungen nicht verallgemeinert werden können, so bestätigen sie doch in Verbindung mit den Ergebnissen der Fallstudie und der Pilotstudie die Eignung unseres LSI-basierten Ansatzes zur Unterstützung des Literaturrechercheprozesses.

Ein weiterer allgemeiner, aber dennoch erwähnenswerter Aspekt ist die Beobachtung, dass die Textmenge in der Suchanfrage einen signifikanten Einfluss auf die Ähnlichkeitswerte hat, nach denen die Treffer in der Ergebnisliste sortiert werden. Je umfangreicher der Text in der Suchanfrage ist, desto geringer sind die Differenzen der Ähnlichkeitswerte eines Ranges in der Ergebnisliste zu seinen „Nachbarn“. Dies kann der Tatsache zugeschrieben werden, dass z. B. Beschreibungen von Konferenz-Tracks oder vollständige Forschungsarbeiten zu einem bestimmten Thema eine umfangreichere und somit vor allem allgemeinere Auswahl an Begriffen und semantischen Konzepten enthält. Da in Forschungsarbeiten üblicherweise Abschnitte zur Darstellung verwandter Forschung und/oder zur thematischen Abgrenzung enthalten sind, wird semantisch mehr als nur ein bestimmtes Thema behandelt, was in der Folge zu ungenaueren Ergebnissen führt. Im Gegensatz dazu sind Abstracts vergleichsweise kurz und konzentrieren sich auf das Wesentliche, sprich den Hauptforschungsbereich des zugehörigen Artikels. Somit ist die darin enthaltene Auswahl an Begriffen und semantischen Konzepten deutlich spezifischer. Eine Sammlung geeigneter Stichwörter stellt die höchste Ebene einer solchen Aggregation dar, was sich insbesondere in den Ergebnissen des Suchszenarios (b) mit dem TSISQ-Stichwort-Ansatz widerspiegelt. Damit jedoch semantische Konzepte durch den LSI-Algorithmus abgeleitet werden können, muss die Suchanfrage umfangreich und kohärent genug formuliert sein. Ist sie dies nicht, können schlechte Ergebnisse wie in Suchszenario (c) unter Verwendung des TSISQ-Stichwort-Ansatzes entstehen. Zusammenfassend lässt sich feststellen: Je gezielter eine Suchanfrage nach den hier dargestellten Konventionen formuliert wird, desto gezielter und „relevanter“ werden auch die Suchergebnisse.

Aufgrund der hohen Anzahl an Veröffentlichungen, die sich mit LSI in verschiedenen Anwendungsfällen sowie den entsprechenden Algorithmen im Detail befassen, gibt es in der Literatur zu LSI viel Positives zu finden, es gibt jedoch auch einige Einschränkungen. Lee et al. (2010) und Bhandari et al. (2008) stellen fest, dass LSI zwar geeignet ist, die bekannten Schwierigkeiten durch Synonymie zu beseitigen, Probleme mit Polysemie jedoch lassen sich aufgrund der orthogonalen Faktoreigenschaften in der Term-by-Document-Matrix nur begrenzt lösen. Kintsch (2010) kritisiert die vereinfachende Annahme, dass der berechnete Vektor zur Textrepräsentation der Zentroid seiner Wortvektoren ist. Er führt ein Beispiel an, das zeigt, dass

die Summe der Vektoren für die beiden Sätze "the lion killed the deer" und "the deer killed the lion" identisch ist. Der signifikante Unterschied in der Bedeutung dieser beiden Sätze würde von LSI übersehen werden. Dennoch wird letztlich betont, dass LSI für Texte ab einer Mindestlänge von 100 bis 150 Wörtern (z. B. Abstracts) eine erstaunlich gute Annäherung und nützliche Resultate liefert. Dieser Aspekt wird auch durch die Ergebnisse unserer Fallstudie bestätigt (vgl. Abbildung 16), wobei umfassendere Suchanfragen in natürlicher Sprache (Abstracts oder Track-Beschreibungen, vgl. Tabelle 4) im Gegensatz zu Suchanfragen mit nur einigen Stichwörtern zu vergleichsweise guten und nützlichen Ergebnissen führen.

Wie in vielen Forschungsprojekten ist die verwendete Datenbasis ein kritischer Aspekt. Unsere Volltext-Datenbank besteht aus 12.332 Artikeln. Auch wenn damit eine breite Palette von Top-Wirtschaftsinformatikartikeln in diversen Forschungsgebieten abgedeckt wird, entspricht diese Teilmenge nur einem Bruchteil der existierenden Literatur im Fachgebiet der Wirtschaftsinformatik und nochmals einem deutlich geringeren Anteil der insgesamt existierenden wissenschaftlichen Literatur (in „allen“ Fachgebieten). Ebenfalls könnte der betrachtete Zeitraum von sieben Jahren unzureichend sein. Um jedoch die Eignung von LSI im Allgemeinen und TSISQ im Speziellen zur Verbesserung des Literaturrechercheprozesses zu evaluieren, entspricht der festgelegte Zeitrahmen unseren Anforderungen.

Um die bisherigen Ergebnisse zu bestätigen, sind umfassendere Tests und Suchszenarien in weiteren Forschungsgebieten nötig. Aktuell ist das TSISQ-Tool und somit der gesamte Ansatz zur Unterstützung des Literaturrechercheprozesses nur für die Verwendung englischsprachiger Inhalte/Texte geeignet. Dies ergibt sich aus der Tatsache, dass ein „Stopp-Wörterbuch“ in der semantischen Indexierungsschicht von TSISQ benötigt wird. Die Erstellung solcher NLP-Ressourcen ist aufwendig und kostet viel Zeit, daher sind diese Wörterbücher für viele Sprachen nicht (oder nicht frei) verfügbar (vgl. Furlan et al., 2013). Da der Großteil der Top-IS-Literatur jedoch noch immer in englischer Sprache veröffentlicht wird, hat dieser Aspekt keinen negativen Einfluss auf die Ergebnisse.

Nach sorgfältiger Prüfung der Resultate aller Suchszenarien und des Laborexperiments kann abschließend eine Empfehlung für die Vorgehensweise bei der Verwendung von TSISQ gegeben werden: Bei LSI handelt es sich um ein Prinzip zur Identifikation semantischer Ähnlichkeiten, es ist hauptsächlich für die Suche nach semantischen Konzepten gedacht. Daher kann die klare Empfehlung gegeben werden, statt Stichwörtern fachgebietsspezifische Volltexte in natürlicher Sprache als Suchanfrage zu wählen. Aus der Erkenntnis, dass die Ergebnisse bei z.B. Abstracts als Suchanfrage thematisch spezifischer sind, während sie bei

Verwendung von umfassenderen Track-Beschreibungen oder vollständigen Forschungsarbeiten diffuser bzw. weniger fokussiert ausfallen, kann folgende Empfehlung abgeleitet werden: Es sollte ein iterativer, zyklischer Prozess eingehalten werden, um bei der Verwendung von TSISQ für die Literatursuche optimale Ergebnisse zu erzielen. Die erste Suchanfrage sollte möglichst aus einem vollständigen Artikel bestehen, der thematisch dem designierten Forschungsgebiet möglichst nah ist. Das Ziel dabei ist es, eine erste, thematisch weiter gefasste Auswahl semantisch ähnlicher Paper als Ergebnis zu erhalten. Anschließend sollten die Ergebnisse manuell überprüft werden, um die relevantesten Artikel für eine spezifischere Eingrenzung des Themas zu identifizieren. Der nächste Schritt umfasst das Erstellen einer Sammlung prägnanter Textbausteine, Sätze und/oder Abschnitte, die in der vorab identifizierten Literatúrauswahl mit Blick auf die zuvor dargestellten Konventionen enthalten sind. Im letzten Schritt wird die daraus resultierende Sammlung als Suchanfrage verwendet. Der Prozess sollte zyklisch wiederholt werden, bis keine neue, relevante Literatur mehr gefunden wird.

2.6 Fazit

Das im ersten Hauptteil (Kapitel 2) dieser Dissertation vorgestellte Forschungsprojekt fasst alle unsere Ergebnisse und Erkenntnisse im Gebiet der semantischen (Ähnlichkeits-) Suche zusammen. Das Projekt liefert einen neuen, toolgestützten Ansatz zur Verbesserung und Weiterentwicklung der wissenschaftlichen Literaturrecherche mittels der Erkennung semantischer Ähnlichkeit zwischen Texten. Dieser kann im Rahmen jeder umfassenderen, etablierten Forschungsmethodik zum Einsatz kommen und dabei deutliche Effizienzgewinne bei der Literaturrecherche erzielen: Die komplexe und sehr wichtige Aufgabe der Literaturrecherche als grundlegender Schritt jeder Forschungsmethodik ist zeitaufwendig und erfordert viel Aufwand. Je umfassender und gründlicher sie durchgeführt wird, desto wahrscheinlicher ist es, dass bestehende Forschungslücken und -fragen präzise identifiziert und entsprechend adressiert werden können. Auch wenn die Eignung und der Nutzen von TSISQ in diesem Forschungsprojekt ausschließlich an Literatur aus dem Gebiet der Wirtschaftsinformatik getestet und nachgewiesen wurde, so kann jedoch davon ausgegangen werden, dass das Vorgehen ohne weitere Anpassungen auf beliebige wissenschaftliche Disziplinen übertragbar ist.

Im Hinblick auf die in Abschnitt 2.1 gestellte Forschungsfrage, wie ein LSI-basierter Ansatz die Effizienz von wissenschaftlichen Literaturrechercheprozessen steigern kann, wurde der Versuch unternommen, etablierte Forschungsmethoden durch den Einsatz eines theoretisch

fundierten Ansatzes zu erweitern. Zu diesem Zweck wurde das „Tool for Semantic Indexing and Similarity Queries“ (TSISQ) implementiert, getestet und evaluiert. Es ermöglicht Wissenschaftlern, abhängig von der gewählten Vorgehensweise bei der Anwendung des Tools und der Auswahl der Suchanfragen, sich schnell einen Überblick über ein Forschungsgebiet zu verschaffen, ihr Wissen darin beliebig zu vertiefen und/oder die theoretischen Grundlagen ihrer Forschung zu vervollständigen oder zu verfeinern. Die Anwendbarkeit von TSISQ wurde in einer Fallstudie, bestehend aus drei verschiedenen Suchfällen und einem Laborexperiment, evaluiert. Die Ergebnisse wurden kritisch diskutiert, mögliche Grenzen und Schwächen des Tools identifiziert und erläutert sowie Empfehlungen für dessen Anwendung gegeben. In Bezug auf das übergeordnete Forschungsziel, mit TSISQ die Effizienz wissenschaftlicher Literaturrechercheprozesse mithilfe von Methoden des Data Mining³ zu erhöhen, können folgende Schlüsse gezogen werden:

- Die Anwendung des hier dargestellten Ansatzes hilft dabei, wertvolle Zeit beim Auffinden der relevanten Literatur zu einem beliebigen Forschungsfeld einzusparen.
- Auch wenn genügend Zeit vorhanden ist kann TSISQ dennoch dazu beitragen, die Vollständigkeit und die Effizienz von Literatur-Reviews zu steigern: Zum einen durch die im Vergleich zu menschlichen Fähigkeiten deutlich höhere Arbeitsgeschwindigkeit und zum anderen durch die Identifikation von Quellen, die andernfalls möglicherweise nicht gefunden oder berücksichtigt worden wären.
- Die im Rahmen der Literaturrecherche eingesparte Zeit kann für eine gründlichere Durchführung anderer Forschungsaufgaben aufgewendet und damit die Qualität der Forschung insgesamt erhöht werden.
- Folglich stellt TSISQ eine nützliche Ergänzung zu den etablierten Suchmaschinen und -methoden in der Wissenschaftswelt dar.

Dennoch wurden im Rahmen dieses Projektes einige Grenzen des Ansatzes offensichtlich, denen nachzugehen weitere Forschungsschritte erfordert:

- Um die Ergebnisse weiter zu optimieren, sollte die für den Index verwendete Datenbank um zusätzliche Literatur von Konferenzen und Zeitschriften erweitert werden.
- Auch sollte dabei ein längerer Zeitraum betrachtet werden, um sicherzustellen, dass auch ältere Basisliteratur in der zu durchsuchenden Datenbasis enthalten ist.

³ Anm. des Verfassers: In diesem Falle aus den Gebieten „Text Mining“, „Information Retrieval“ und NLP.

- Um die Validität der Evaluierungsergebnisse zu erhöhen, sollten bei einer Fortführung der Forschung in diesem Gebiet umfassendere und anspruchsvollere Tests durchgeführt sowie Bewertungen und Meinungen einer größeren Gruppe von Domänenexperten eingeholt werden.
- Eine detailliertere Analyse der Auswirkungen auf die Suchergebnisse, wenn Suchanfragen im selben Fachgebiet (ähnlich einer Sensitivitätsanalyse) minimal verändert werden, könnte ebenfalls wertvolle Erkenntnisse liefern.
- Eine derartige, kontrollierte Anpassung einzelner Suchbegriffe erlaubt und/oder erleichtert die Formulierung stärker fokussierter Suchanfragen.
- Zukünftige Arbeiten sollten sich auch mit der Festlegung klarer Leitlinien für die Zusammensetzung von Suchanfragen befassen. Die Messung der Auswirkungen auf die Suchergebnisse für SEM im Suchfall (c) beispielsweise, wenn die Begriffe "model", "modeling" und "modelling" einzeln ausgelassen werden, kann dabei helfen, allgemeine Richtlinien für die Formulierung von Suchanfragen zu erstellen.

Basierend auf den zuvor genannten Schritten und gemäß unseren Empfehlungen kann die Einbettung des hier vorgestellten Ansatzes in einen strukturierten und gut konzipierten, iterativen Prozesszyklus eine vielversprechende zusätzliche Erweiterung sein. Schließlich könnte das TSISQ-System zu einer rein webbasierten Lösung mit Onlineanbindung an wissenschaftliche Datenbanken weiterentwickelt werden. Mit der Kooperation und Unterstützung von deren Betreibern könnte ein besserer, umfangreicherer Service für Wissenschaftler angeboten werden, der das vorhandene Angebot komfortabel, verlässlich und zielgerichtet nach semantisch ähnlichen Themen durchsucht. Da wissenschaftliche Literatur zumeist kostenpflichtig ist, könnte die Suchergebnisliste Links zu den gefundenen Quellen liefern, während die vollständigen Artikel hinter den Paywalls der wissenschaftlichen Datenbanken verbleiben.

3 Operative Entscheidungsunterstützung mit KI

*„Information technology and business are becoming inextricably interwoven.
I don't think anybody can talk meaningfully about one
without the talking about the other.“*

– Bill Gates

In diesem Abschnitt werden zwei von drei in diesem Kontext entstandene Veröffentlichungen detailliert vorgestellt. Die dritte Veröffentlichung (Eilers et al. 2017) baut direkt auf den hier in Kapitel 3 präsentierten Arbeiten auf. Sie basiert zudem zu 100% auf den im Rahmen des hier vorgestellten Forschungsprojektes gewonnenen Daten und beschreibt eine logische Konsequenz der von den Autoren während dieser Zeit gewonnenen Erfahrungen und Erkenntnisse. Aufgrund der Aufgabenteilung der beiden federführenden Autoren (Christoph Gleue und Dennis Eilers) werden diese Erkenntnisse zwar auch in dieser Dissertation behandelt, sollen jedoch im Rahmen einer den einzelnen Themenkomplexen in Kapitel 2 und 3 übergeordneten Diskussion behandelt und kritisch beleuchtet werden (vgl. Kapitel 4). Die Ergebnisse und Erkenntnisse aus den beiden verbleibenden Publikationen stellen hingegen das Fundament des zweiten Hauptteils der vorliegenden Dissertation dar und tragen beide den Titel „Decision Support for the Automotive Industry: Forecasting Residual Values Using Artificial Neural Networks“ (Gleue et al., 2017; Gleue et al., 2018). Die an diesen Publikationen beteiligten Autoren waren Christoph Gleue, Dennis Eilers, Hans-Jörg von Mettenheim und Michael H. Breitner. Es handelt sich bei beiden Publikationen um die Veröffentlichung der Forschungsergebnisse aus dem o.g. Kooperationsprojekt der Autoren mit einem großen, deutschen Automobilhersteller, sie wurden nach einer Revision bei der Internationalen Tagung Wirtschaftsinformatik (WI) 2017 in St. Gallen (Gleue et al., 2017) angenommen und präsentiert. Im VHB-JOURQUAL3-Ranking werden die Proceedings dieser „peer-reviewed“-Konferenz mit „C“ bewertet, in der WI-Orientierungsliste der WKWI ist die WI mit der höchstmöglichen Bewertung „A“ gelistet. Die „WI“ ist die führende Konferenz der Wirtschaftsinformatik-Community im deutschsprachigen Raum. Das Motto der WI 2017 war "Towards Thought Leadership in Digital Transformation". Das auf dieser Konferenz veröffentlichte Paper wurde als eine von wenigen Arbeiten von der „Business & Information Systems Engineering“ und von der „Electronic Markets“ für eine „Fast-Track-Publikation“ in weiterentwickelter Form in den genannten wissenschaftlichen Zeitschriften nominiert. Die Autoren haben sich für eine Einreichung in der renommierten Zeitschrift, der BISE,

entschieden. Das inhaltlich zu diesem Zweck deutlich erweiterte Paper wurde konsequent eingereicht und nach nur zwei „minor revisions“ zur Publikation freigegeben (Gleue et al., 2018).

Der Impact Factor von Thomson Reuters in 2016 für BISE lag bei 3,392, im VHB-JOURQUAL3-Ranking für Wirtschaftsinformatik (Information Systems Research) wird die Zeitschrift mit einem „B“, in der WI-Orientierungsliste der WKWI mit einem "A" bewertet (höchstmögliche Bewertung). Im genannten VHB-JOURQUAL3-Ranking für Wirtschaftsinformatik bekam die BISE die zweithöchste Anzahl Bewertungen (168), direkt nach der renommierten Zeitschrift „Management Information Systems Quarterly“ (MISQ).

Die BISE beschreibt die Voraussetzungen für eine Veröffentlichung in ihrer Zeitschrift wie folgt: „BISE (Business & Information Systems Engineering) is an international scholarly and double-blind peer reviewed journal that publishes scientific research on the effective and efficient design and utilization of information systems by individuals, groups, enterprises, and society for the improvement of social welfare. Information systems are understood as socio-technical systems comprising tasks, people, and information technology. Research published in the journal examines relevant problems in the analysis, design, implementation, and management of information systems.

BISE has been the flagship journal of the German-language Information Systems community for almost 60 years. It is now one of the leading European journals in the field. BISE is sponsored by the Section “Information Systems” (Wirtschaftsinformatik, WKWI) of the German Association for Business Research (VHB) and the special interest group “Business Informatics” (GI-FB WI) of the Gesellschaft für Informatik e. V. (GI) with more than 1200 members. BISE is also an affiliated journal of the Association for Information Systems (AIS).“⁴

3.1 Gesamtwirtschaftlicher Kontext, Motivation und Projektziele

Die stetig fortschreitende Digitalisierung der Gesellschaft führt in der heutigen Zeit auch in der Wirtschaft zu einem tiefgreifenden Wandel: Unternehmen erkennen zunehmend das enorme Potenzial neuer Technologien im Bereich des maschinellen Lernens. Daten sind ubiquitär, die Generierung und Speicherung immer größer werdender Datenmengen wird immer günstiger. Da es sich bei (Daten-)Analysen sozusagen um das dazugehörige Komplement handelt, steigt gleichzeitig entsprechend der Bedarf an ihnen – und an „Data Scientists“, die in der Lage sind, diese Analysen durchzuführen. Die Anwendung von Data Science-Verfahren, z.B. zur

⁴ <http://www.bise-journal.com/>, zuletzt aufgerufen am 28.09.2018.

datengetriebenen Entscheidungsunterstützung, hält in nahezu jeder Branche unaufhaltsam Einzug (vgl. Chen et al., 2012).

Insbesondere historisch gewachsene IT-Landschaften in Großkonzernen halten bereits große Mengen an Daten vor, welche jedoch häufig in über diverse Abteilungen verteilter, unstrukturierter Form und in variierender Qualität, Konsistenz sowie unterschiedlichem Format vorliegen. Werden diese Daten integriert, zielgerichtet analysiert, verarbeitet und visualisiert, so können wertvolle und exklusive Informationen für das jeweilige Unternehmen gewonnen und so enorme Wettbewerbsvorteile generiert werden (vgl. Chaudhuri et al., 2011). Dabei kann die Kooperation zwischen Forschung und Wirtschaft für beide Seiten von Vorteil sein: Konzerne sind im Besitz von einzigartigen Daten bzgl. ihres Geschäftsfeldes, ihnen fehlt jedoch häufig das Know-How, aus diesen Daten die gewünschten Informationen zu extrahieren. Auf der anderen Seite sind in Wissenschaft und Forschung zwar die Technologien und mathematisch/statistische Verfahren bekannt, doch fehlen die meist unter strenger Geheimhaltung stehenden, realen Daten aus der Wirtschaft, um aus dem konkreten Anwendungsfall einen generalisierbaren Beitrag für das jeweilige Forschungsgebiet leisten zu können.

Die in diesem Abschnitt zusammengefassten Ergebnisse aus einem mehrjährigen Kooperationsprojekt mit einem großen, deutschen Automobilhersteller bewegen sich in genau diesem Spannungsfeld. Motivation und Forschungsziel hierbei war es, sowohl einen nachhaltigen Mehrwert im Bereich der Entscheidungsunterstützung für das Unternehmen zu schaffen, als auch mithilfe der durch das Unternehmen in großem Umfang bereitgestellten Daten „aus der realen Welt“ einen wertvollen Beitrag zum Forschungsfeld der künstlichen Intelligenz im Anwendungsbereich der Prognose realer Marktpreise (in diesem Falle von Leasingfahrzeugen) zu leisten.

3.1.1 Wirtschaftliche Bedeutung des Leasinggeschäfts für Automobilhersteller

Das Leasinggeschäft ist eines der wichtigsten Geschäftsfelder für Automobilhersteller. Allein im Jahre 2016 machten die neu abgeschlossenen Leasingverträge für Pkw und Nutzfahrzeuge 67% (223,8 Milliarden Euro) des gesamten europäischen Leasingvolumens aus; gegenüber 2014 entspricht dies einer Steigerung von 45,6 Milliarden Euro. Es handelt sich dabei folglich um das größte Geschäftsfeld des europäischen Leasingmarktes (vgl. Leaseurope, 2016). Bei knapp einem Drittel aller von deutschen Premium-Automobilherstellern in 2015 veräußerten Neufahrzeuge handelte es sich um Leasingfahrzeuge (DAT-Report, 2016); laut dem Verband der Banken der Automobilwirtschaft (BDA), dem die herstellereigenen Banken und

Leasinginstitute aller namhaften Automobilhersteller angehören, wurden in 2017 bereits rund 75 Prozent aller Pkw-Neuzulassungen in Deutschland über Leasing- und Finanzierungsmodelle⁵ auf die Straße gebracht, was einem Kreditvolumen allein in Deutschland von 60 Milliarden Euro⁶ entspricht (BDA, 2017).

An den zuvor genannten Zahlen und dem massiven Wachstum der letzten Jahre ist bereits zu erkennen, dass der Markt für automobiler Finanzdienstleistungen zu einem volkswirtschaftlich bedeutenden Wachstumsmarkt gereift ist. Dieses Geschäftsfeld bietet daher enorme Marktchancen, insbesondere für Fahrzeughersteller und Leasinggesellschaften. Gleichzeitig bringt das Leasinggeschäft jedoch schwer quantifizierbare Risiken mit sich, die sogenannten Restwerttrisiken (vgl. Prado und Ananth, 2012), auf deren Minimierung (durch eine genauere Prognose von Zukunftsrestwerten der Fahrzeuge als wesentliche Berechnungsgrundlage für marktgerechte Leasingraten) sich die in diesem Abschnitt vorgestellte Forschung konzentriert.

3.1.2 Markt- und Restwertprognosen in der wissenschaftlichen Literatur

Aufgrund des bereits erwähnten intensiven Wettbewerbs am Leasingmarkt und der Tatsache, dass entsprechende Marktdaten üblicherweise unter strenger Geheimhaltung der Automobilindustrie stehen, handelt es sich bei den in diesem Forschungsprojekt verwendeten Daten um einen einzigartigen, exklusiven und nicht öffentlich zugänglichen Datensatz. In diesem Forschungsgebiet existieren daher nur sehr wenige Studien, wobei jede der im Folgenden aufgeführten Forschungsarbeiten auf einem eigenen, ähnlich exklusiven Datensatz oder lediglich theoretischen Annahmen basiert. Die relative Exklusivität dieser Daten begründet zum einen die Knappheit an Forschungsarbeiten zu diesem Thema, zum anderen aber auch das enorme Forschungspotenzial in diesem Gebiet. Die für dieses Projekt wichtigste, vergleichbare Studie wurde von Lessmann et al. (2010) veröffentlicht. Aufbauend auf 124.386 Gebrauchtfahrzeugs-Transaktionsdaten von Fahrzeugen der Oberklasse eines anderen großen Automobilherstellers entwickeln die Autoren ein Entscheidungsunterstützungssystem unter Verwendung von Support Vector Regression (SVR). Diese Methode stellt eine Erweiterung der klassischen linearen Regression dar, um eine nicht-lineare Transformation der unabhängigen Variablen zu ermöglichen. Jedes Fahrzeug wird durch 176 Attribute beschrieben, wobei der Großteil der Attribute aus Platzhaltervariablen besteht, um Eigenschaften wie z.B. verschiedene

⁵ Anmerkung des Verfassers: In der vorliegenden Forschungsarbeit steht jedoch lediglich das „klassische“ Leasing im Mittelpunkt der Betrachtung. Andersartige Finanzierungsangebote der Hersteller, bei denen *nicht* die Herstellerbank Eigentümer des Fahrzeugs ist, sind aus Sicht des Risikomanagements eines Herstellers uninteressant, da u.a. die Restwerttrisiken vollständig vom Kunden getragen werden.

⁶ In 2016 belief sich das Kreditvolumen noch auf 47,7 Milliarden € (vgl. AKA, 2016).

Sonderausstattungen abzubilden. Der Kern der in dieser Studie entwickelten Methode besteht aus der Verwendung transaktionsbezogener Charakteristika wie beispielsweise typischer Eigenschaften der Kunden.

Die Autoren zeigen Vorteile und Nutzen der Verwendung solcher interner Informationen in einem Prognosemodell auf und empfehlen, die Restwertprognosen des Unternehmens um interne Faktoren zu erweitern, um die Prognosegenauigkeit zu verbessern. Diese kann folglich auf Basis der ausschließlich dem Unternehmen selbst vorliegenden Daten gesteigert werden, insbesondere im Vergleich zu Prognosen externer Dienstleister oder Restwertinstitute, die ihre Daten in meist deutlich kleinerem Umfang und zu höheren Kosten selbst erheben müssen. Aufbauend auf den Erkenntnissen von Lessmann et al. (2010) liegt der Fokus der hier vorgestellten Forschung insbesondere auf internen sowie externen Faktoren und deren Einfluss auf die Wiederverkaufspreise von Leasingfahrzeugen nach Ablauf der Leasingdauer. Seit der Veröffentlichung der dieser Dissertation zugrundeliegenden Publikationen haben die letztgenannten Autoren ihre Forschung fortgesetzt, um die Prognosegenauigkeit weiter zu erhöhen (vgl. Dress et al., 2017). Sie verwenden dabei asymmetrische Kostenfunktionen, da Über- oder Unterschätzungen des Restwertes nicht zwangsläufig dieselben Auswirkungen auf die unterschiedlichen Geschäftssituationen haben.

Eine weitere Forschungsarbeit in diesem Forschungsgebiet wurde von Wu et al. (2009) veröffentlicht. Die Autoren prognostizieren Gebrauchtfahrzeugpreise auf dem taiwanesischen Markt. Ihre Inputfaktoren bestehen aus der Marke des Fahrzeugs, dem Herstellungsjahr, dem Motortyp und einem Ausstattungsindex. Sie verwenden eine neue Kombination aus künstlichen neuronalen Netzen (KNN) und adaptiven Neuro-Fuzzy-Inferenzsystemen (ANFIS) zur Erhöhung der Prognosegenauigkeit. Eine frühere Studie von Lian et al. (2003) beschreibt Restwertprognosen als Zeitreihenproblem. Evolutionäre künstliche neuronale Netze (EKNN) werden für die Modellierung der Restwerte 24 Monate alter Fahrzeuge in einem Zeitraum von 1993 bis 1997 verwendet. Die Autoren stellen zyklische, saisonale Schwankungen fest, wobei das Restwertniveau am Anfang eines Jahres auf einem hohen Niveau startet und im weiteren Verlauf des Jahres auf ein niedrigeres Niveau fällt. Weitere Studien befassen sich häufig mit dem Einfluss makroökonomischer Indikatoren in Kombination mit internen Faktoren, um die Verteilung der Restwerte zu erklären. Prado (2009) verwendet zusätzlich zu fahrzeugspezifischen Variablen wie Alter und km-Laufleistung den Dieselpreis sowie den Index der Industrieproduktion als erklärende Variablen. Fan et al. (2008) vergleichen verschiedene Verfahren wie AutoRegressiveTrees (ART), KNN und lineare Regression, konzentrieren sich dabei jedoch auf Leasing von schweren Baumaschinen, wobei das CART-

Modell die besten Ergebnisse für diesen Anwendungsfall liefert. Letztlich bleibt festzuhalten, dass sich zum aktuellen Zeitpunkt in der wissenschaftlichen Literatur noch keine Standardverfahren für die Prognose von Fahrzeugrestwerten etabliert haben.

Neben den rein datengetriebenen Prognoseverfahren existieren weitere, theoretische Ansätze um Preisentwicklungen zu erklären und daraus Implikationen für das Risikomanagement abzuleiten. Da es sich bei diesen Ansätzen jedoch nicht um den Fokus dieser Arbeit handelt, soll an dieser Stelle für einen tieferen Einblick in die Thematik auf Rode et al. (2002), Storchmann (2004) und Smith und Jin (2007) verwiesen werden.

3.1.3 Projektziele

Das o.g. Kooperationsprojekt wurde mit dem Ziel begonnen, die Restwerte von Leasingfahrzeugen möglichst genau bereits vor Abschluss eines Leasingvertrages vorherzusagen. Der (zukünftige) Restwert des betreffenden Leasingfahrzeuges stellt einen der wichtigsten Faktoren für die Berechnung von konzernweit gültigen Leasingraten dar. Da insbesondere im deutschen Neu- und Gebrauchtwagenmarkt ein sehr intensiver Wettbewerb herrscht müssen Leasingraten jedoch zwingend wettbewerbsfähig sein, was die Notwendigkeit zu einem individuellen, möglichst genauen und tagesaktuellen Monitoring der Entwicklungen am Gebrauchtwagenmarkt mit sich bringt. Zusammengefasst war die Vision hinter dem hier vorgestellten Projekt, auf der Basis von historischen Verkaufs- und Marktdaten sowie Daten aus diversen weiteren Konzernsystemen des Automobilherstellers ein selbstlernendes Entscheidungsunterstützungssystem (EUS) zu entwickeln, das mithilfe moderner Methoden aus dem Gebiet des Data Mining und der Künstlichen Intelligenz (KI) nicht nur die Prognose von Gebrauchtwagen-Restwerten deutlich präziser als vor Beginn des Projekts ermöglicht, sondern darüber hinaus auch bei wichtigen Entscheidungen und der Beantwortung einer Vielzahl fachlicher Fragen des Fachbereichs Finanz des Automobilherstellers unterstützen kann.

Dazu wurden zunächst die Daten aus den Gebrauchtwagenverkäufen (nahezu) aller Vertragshändler des Automobilherstellers von 2011-2017 verwendet; die Verkaufspreise entsprechen dem gesuchten Restwert der Fahrzeuge am Ende der Leasingdauer. Diese Daten enthalten beispielweise den Herstellungszeitpunkt, den Kilometerstand und modellspezifische wie auch fahrzeugspezifische Informationen (z.B. Motorisierung, Getriebeart, Antriebsart, Ausstattung). Für Hersteller und Leasinggeber ist es besonders wichtig, eine möglichst genaue Vorstellung des zukünftigen Verkaufspreises eines Fahrzeuges bereits vor Abschluss des Leasingvertrages zu haben, da der Wertverlust über die gesamte Nutzungszeit des Leasinggutes der Hauptfaktor zur Berechnung der passenden, monatlichen Leasingrate ist. Während die

monatliche Rate über die gesamte Leasingdauer ein unveränderlicher Wert ist, ist der zukünftige Wiederverkaufswert des Fahrzeugs unsicher. Das führt letztlich zu einem kalkulierten Risiko bei der Wiedervermarktung für den Hersteller.

Folglich ist es besonders wichtig, den Wiederverkaufspreis so früh und so genau wie möglich zu bestimmen, denn sowohl eine Über- als auch eine Unterschätzung führen zu negativen Konsequenzen für den Hersteller:

Eine *Überschätzung* des Leasing-Restwertes führt zu einer Fehlkalkulation der Leasingraten derart, dass deren Summe kumuliert über die gesamte Leasingdauer nicht den eingetretenen Wertverlust am Fahrzeug kompensiert; die Raten sind zu gering. Der so entstehende Fehlbetrag muss i.d.R. vom Händler ausgeglichen werden, d.h. er nimmt bei einem Gebrauchtwagenverkauf (zu wettbewerbsfähigen Preisen) weniger ein, als er für das Fahrzeug (üblicherweise bei der hauseigenen Bank des Herstellers, dem Eigner des Fahrzeugs während der Leasingdauer) selbst bezahlen musste⁷. Dieser Wirkungszusammenhang führt global betrachtet (bei teilweise hunderten oder tausenden Fahrzeugen pro Jahr pro Händler) zwar kurzfristig zu einem Wettbewerbsvorteil im Neuwagengeschäft durch günstige Leasingraten, früher oder später jedoch unweigerlich zur Insolvenz der Händler. Weil die Hersteller jedoch ein funktionierendes, gesundes Händlernetz gewährleisten müssen, werden derartige Verluste üblicherweise zumindest anteilig oder vollständig vom Hersteller kompensiert.

Eine *Unterschätzung* des Leasing-Restwertes zum Zeitpunkt des Abschlusses eines Leasingvertrags führt zu einem gänzlich anderen Effekt: Der zu niedrig prognostizierte Zukunftsrestwert unterstellt einen höheren Wertverlust des Leasingguts über die Leasingdauer als tatsächlich eintritt, was den Hersteller dazu veranlasst, überhöhte Leasingraten zu kalkulieren. Da in Deutschland sowohl der Leasing- als auch der Gebrauchtwagenmarkt einem hohen Konkurrenzdruck unterliegt, müssen diese kalkulatorischen Leasingraten folglich gesenkt werden, um wettbewerbsfähige Leasingkonditionen anbieten zu können. Die dafür notwendigen Verkaufshilfen stammen ebenfalls direkt vom Hersteller und können dort zu ausgesprochen hohen Verlusten führen. Die Minimierung des Risikos einer Unter- sowie Überschätzung von Leasingraten über das gesamte Produktportfolio hinweg ist folglich eines der Hauptziele des hier beschriebenen Projektes. Um dies zu erreichen, wurde ein EUS entwickelt, um auf Basis historischer Daten aus dem Leasing- und Gebrauchtwagengeschäft mithilfe von KI einen möglichst genauen Wiederverkaufspreis zu prognostizieren und so die

⁷ Anm. des Verfassers: Die Händler werden i.d.R. vom Hersteller vertraglich dazu verpflichtet, Fahrzeuge zu einem zu Beginn des Leasingvertrages mit einem Endkunden festgelegten Rückkaufwert zum Ende des Leasingvertrages von der Herstellerbank abzunehmen; das so entstehende Risiko trägt der Händler selbst.

oben beschriebenen Risiken zu minimieren. Dies erlaubt dem Hersteller nach erfolgreichem Abschluss des Projekts eine deutlich feingranularere Steuerung seines Leasing- und Gebrauchtwagengeschäfts. Darüber hinaus ermöglichen die für diesen Zweck entwickelten Softwaretools ein tagesaktuelles Markt-Monitoring und somit eine gezieltere Steuerung von Managementaktivitäten und -entscheidungen in den Bereichen Finanzen, Controlling, Marketing und Vertrieb.

Im folgenden Abschnitt werden die Ergebnisse unserer Forschung auf dem Gebiet der Restwertprognosen mit KI bzw. KNN vorgestellt und dabei ein besonderes Augenmerk auf die gewonnenen Erkenntnisse gelegt, die, insbesondere im Vergleich zu anderen Studien, dabei entstanden sind.

3.2 Gesamtsystem zur Entscheidungsunterstützung im Leasinggeschäft

In diesem Abschnitt werden der Aufbau des entstandenen Gesamtsystems, die angewandte Methodik sowie die verwendeten Inputdaten im Hinblick auf das übergeordnete Projektziel, die Entwicklung eines EUS für restwert- und leasingbezogene Entscheidungen, zusammenfassend dargestellt. Abschnitt 3.2.1 gibt einen Überblick über die Architektur des EUS und das Zusammenwirken der einzelnen IT-Komponenten, deren Kombination letztlich die mit diesem Projekt verfolgten Forschungsziele zu erreichen ermöglicht.

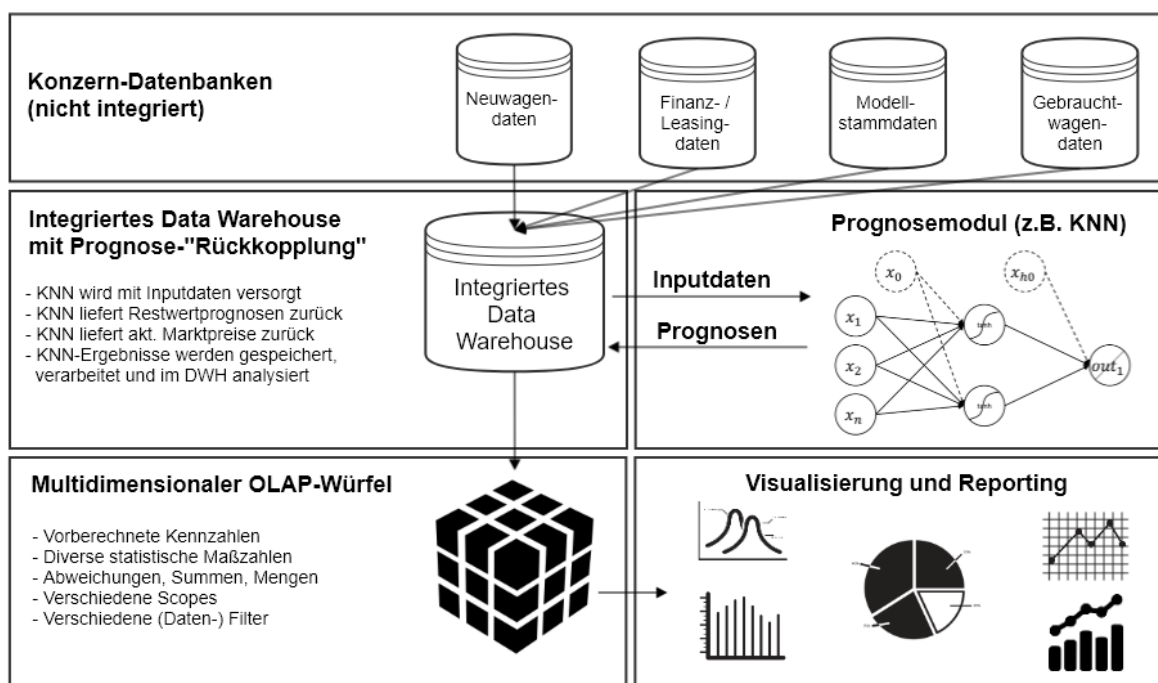


Abbildung 18: EUS für Markt- und Restwertprognosen im Leasinggeschäft

Anschließend wird detaillierter auf die in Abbildung 12 dargestellten Datenquellen eingegangen, gefolgt von einer genauen Untersuchung der in dieser Studie verwendeten

externen und internen Inputfaktoren, von Markttrends und saisonalen Effekten sowie deren Einfluss auf die Fahrzeugrestwerte (Kapitel 3.2.2 bis 3.2.4).

3.2.1 Systemarchitektur und Datenquellen

Um ein EUS zu schaffen, das automatisiert Daten aus diversen Konzernsystemen integriert, normiert, die später benötigten Kennzahlen berechnet und diese letztlich ansprechend, informativ, zielgerichtet und für den Fachbereich verständlich und komfortabel nutzbar aufbereitet, sind eine Vielzahl an Systemkomponenten notwendig. Jede dieser Komponenten übernimmt eine für das Endergebnis wichtige Funktion in einem Workflow, an dessen Ende die vielfältigen Fragen des Fachbereiches Finanz mithilfe maßgeschneiderter Analysen, Berichte, Dashboards etc. beantwortet werden können. Die einzelnen Systemkomponenten und deren Zusammenwirkung sind exemplarisch in Abbildung 18 dargestellt.

Die KNN-Prognosemodelle werden mit historischen Daten trainiert, daher bestand der erste Entwicklungsschritt zunächst daraus, die bereits im Konzern in verteilter Form vorliegenden Daten zu finden, anschließend die entsprechenden Quellsysteme anzubinden, deren Daten zu importieren, zu integrieren und aufzubereiten. Die mit Abstand wichtigste aller Datenquellen ist die Gebrauchtwagenmarkt-Datenbank, in der alle von den Vertragshändlern in Deutschland gesammelten Gebrauchtwagendaten gespeichert werden. Sie beinhaltet jeden einzelnen in Deutschland abgewickelten Gebrauchtwagenverkauf seit 2011, einschließlich der Fahrzeugidentifikationsnummer (VIN) als eindeutigen Schlüssel für das Fahrzeug, fahrzeugbezogene Informationen (z.B. Kilometerstand, Fahrzeugalter, Modelltyp, Motor, Kraftstoffart), das Verkaufsdatum und letztlich die geplante Output-Variable des Prognosemodells, den Netto-Verkaufspreis. Alle Daten aus dieser Quelle werden von den Händlern manuell eingegeben und anschließend über eine gemeinsame Schnittstelle zu allen am Markt im Einsatz befindlichen Händlermanagementsystemen übertragen. Aufgrund dieses „menschlichen Faktors“ sind die Daten nicht immer korrekt und müssen daher, um spätere Ungenauigkeiten und Verzerrungen zu vermeiden, zur Qualitätssicherung mit Daten aus anderen internen Quellen verknüpft und (kreuz-)validiert werden. Erst nach einer gründlichen, automatisierten Prüfung und einer entsprechenden Bereinigung werden sie als Trainingsdaten für die Prognosemodelle verwendet.

Eines dieser qualitätssichernden Systeme ist die Neufahrzeug-Datenbank, welche die Daten aus dem Fakturierungssystem des Herstellers direkt nach Abschluss des Produktionsprozesses bezieht. Neben dem Primärschlüssel, der VIN, ist darin auch die zum Zeitpunkt der Produktion eines Fahrzeuges gültige, unverbindliche Preisempfehlung (UPE) historisiert abgelegt, die als

Summe der (separat ausgewiesenen) Fahrzeug-Basispreise, der Aufpreise für die jeweilige Lackierung und der kumulierten Aufpreise für die gewählte Sonderausstattung eines Fahrzeugs vorliegt. Darüber hinaus enthält die Neufahrzeug-Datenbank pro Fahrzeug eine vollständige Liste aller (Sonder-)Ausstattungsmerkmale sowie einen Indikator, um zu erkennen, ob es sich bei dem Fahrzeug um ein Aktionsmodell⁸ handelt (und wenn ja, welches).

Diese Informationen sind aus verschiedensten Gründen sehr wertvoll: Zum einen ist es in der Leasingbranche üblich, Restwerte von Leasingfahrzeugen und alle damit zusammenhängenden Berechnungen in Prozent der ursprünglichen UPE auszudrücken. Da die Gebrauchtwagenmarkt-Datenbank jedoch nur Gebrauchtwagen-Verkaufspreise enthält und die zum Herstellungszeitpunkt gültige UPE i.d.R. unbekannt ist, kann der benötigte Prozentwert nur durch die Zusatzinformation aus der Neuwagen-Datenbank berechnet werden; die Zuordnung erfolgt über die (eindeutige) VIN als Primärschlüssel. Zum anderen sind die zusätzlich gewonnenen Attribute und Preisbestandteile ein unabdingbarer Bestandteil aller weiteren Schritte: Die getrennte Darstellung von Basispreis, Lackierungsaufpreis und insbesondere des kumulierten Aufpreises für etwaige Sonderausstattungen erlaubt eine spätere, feingranulare Differenzierung der Restwertentwicklung von gut und weniger gut ausgestatteten Fahrzeugen und ein genaueres Clustering bei den nachgelagerten Analysen. Die genaue Auflistung aller einzelnen Ausstattungsmerkmale ermöglicht zudem beispielsweise die Beantwortung von Fragen wie:

- Welchen Einfluss hat das Fehlen eines Schiebedachs, einer Klimaanlage, eines Navigationssystems auf die Wertentwicklung eines einzelnen Fahrzeugs?
- Wie muss ein Fahrzeug konfiguriert sein, um eine möglichst stabile Wertentwicklung und damit vergleichsweise niedrige Leasingraten zu gewährleisten?
- Wie sollten Dienstwagen und Fahrzeuge im Werksangehörigen-Leasing konfiguriert werden, um später als junge Gebrauchtwagen einen möglichst hohen Wiederverkaufspreis zu erzielen?
- Welche Ausstattungsmerkmale haben welchen Einfluss auf den Wiederverkaufswert im Zeitverlauf? Beispiel: Ein Doppelkupplungsgetriebe erhöht den Restwert nur bei

⁸ Anm. des Verfassers: Aktionsmodelle sind (wie Sondermodelle) vom Hersteller „sinnvoll vorkonfigurierte“ Angebotsmodelle, bei denen die Summe der Aufpreise für die paketierte Sonderausstattungen mit einem signifikanten Preisvorteil gegenüber einem identisch konfigurierten Serienfahrzeug angeboten wird. Aktionsmodelle bekommen im Gegensatz zu Sondermodellen kein spezielles „Branding“ und können daher optisch letztlich nicht von identisch konfigurierten Serienmodellen unterschieden werden.

niedrigen km-Laufleistungen, bei hohen km-Laufleistungen werden robustere Schaltgetriebe vom Kunden bevorzugt.

Letztlich dient der Aktionsmodell-Indikator in Kombination mit zusätzlichen Stammdaten einer Unterscheidung zwischen Serien- und Aktionsmodellen. Da zwei identisch ausgestattete Fahrzeuge am Gebrauchtwagenmarkt per Definition auch denselben Verkaufspreis erzielen, ein Aktionsmodell jedoch eine aus Herstellersicht finanziell „sinnvolle“ Paketierung von Sonderausstattungen enthält, werden Aktionsmodelle mit einem teilweise wesentlichen Rabatt angeboten. Konfiguriert ein Kunde ein Serienfahrzeug identisch, so sind auch die Fahrzeuge identisch. Ihre Listenpreise unterscheiden sich jedoch üblicherweise maßgeblich, was bei Auswertungen auf Basis der prozentualen Restwerte (bei einem Aktionsmodell-Anteil an der Gesamtabsatzmenge von bis zu 50%) zu massiven Verzerrungen führen kann.

Die aus den zuvor beschriebenen Systemen bezogenen Informationen liegen größtenteils in codierter Form vor, sodass diese mithilfe der Daten aus einem eigens für das Projekt entwickelten MDM-System (engl.: „Master Data Management“) vor der Weiterverarbeitung entschlüsselt werden müssen. Dies ist auch wichtig für das spätere Reporting, um die intern im System verwendeten Codierungen für die Berichtsempfänger verständlich darzustellen. Das MDM enthält allgemeine Informationen über Modelltypen, Aktionen (die Höhe der Rabatte auf Aktionsmodelle), Motorleistung, Hubraum, Getriebetyp (manuell, Automatik, Doppelkupplung), Anzahl Gänge und Modellalter (Anzahl Monate seit Markteinführung) sowie das „Mapping“ zwischen den für einzelne Eigenschaften verwendeten Codierungen und deren „für Menschen verständliche“ Beschreibung in natürlicher Sprache. Letztlich besteht auch eine Verbindung zur Leasing-Datenbank des Herstellers bzw. der Herstellerbank, die die konzernweit offiziell gültigen, konventionellen Restwertprognosen und Marktpreis-Schätzungen der konzerneigenen Leasinggesellschaft bereitstellt.

Im nächsten Schritt werden für jedes einzelne am Gebrauchtwagenmarkt verkaufte Fahrzeug alle zuvor beschriebenen Informationen automatisiert in einem zentralen Data Warehouse (DWH) gesammelt, integriert, bereinigt und historisiert. Das DWH ist wiederum direkt mit dem Prognosemodul verbunden, worin das bereits trainierte Prognosemodell hinterlegt ist. Es wird mit Inputdaten direkt aus dem DWH gespeist, prognostiziert auf deren Basis Marktwerte (Istwerte) und erwartete Zukunftsrestwerte⁹ für alle Fahrzeuge in der Datenbank und liefert diese zurück an das Data Warehouse, wo sie zur weiteren Verarbeitung gespeichert werden.

⁹ im Folgenden der Einfachheit halber als „Restwerte“ bezeichnet

Die nächste und vorletzte Schicht des EUS ist ein multidimensionaler Online Analytical Processing-Würfel (OLAP), der als Grundlage für Visualisierungen und das gesamte Reporting dient, z. B. für Benchmarks der verschiedenen Prognosemodelle, Marktanalysen und vor allem Auswertungen und Entscheidungshilfen zur Leasing- und Preisgestaltung. Der OLAP-Würfel enthält zahlreiche (aus den Basisdaten) vorberechnete Indikatoren, Abweichungskennzahlen, Summen, Beträge, arithmetische Mittel und weitere statistische Kennzahlen. Es gibt darin vorkonfigurierte „Scopes“ (Abgrenzungen) und Filtermöglichkeiten, um z.B. Auswertungen auf Aktions- oder Serienmodelle zu begrenzen (z.B. zur Vermeidung von Verzerrungen) oder in Auswertungen komfortabel zwischen jungen Gebrauchtwagen (jünger als 12 Monate) und Fahrzeugen unterscheiden zu können, die das klassische Leasingsegment mit einer Leasingdauer von mindestens 12 bis 72 Monaten repräsentieren.

Der OLAP-Würfel kann folglich als umfangreicher „Werkzeugkasten“ für die Anfertigung von statischen oder dynamischen ad-hoc Berichten, Entscheidungshilfen und Visualisierungen für Fachbereich und Management angesehen werden; außerdem ist er die direkte Quelle für das Standard-Reporting. Der hohe Automatisierungsgrad bei der Datenintegration, -aufbereitung und Qualitätssicherung, die Integration des Prognosemodells sowie vordefinierte Filter und Selektionskriterien garantieren maximale Vergleichbarkeit der Ergebnisse, auch zu verschiedenen Zeitpunkten.

Korrekt angewandt kann der OLAP-Würfel Antworten auf zahlreiche Fragestellungen rund um das Gebrauchtwagen- bzw. Leasinggeschäft liefern, zum Beispiel:

- Händler-Rankings zu deren individueller Verkaufperformance
- Monitoring und Analyse des Preisniveaus am Fahrzeugmarkt nach externen Schocks (wie der Finanzkrise, der „Dieselkrise“ oder politischen Entscheidungen)
- Früherkennung von Abweichungen der Markt- und Restwerte vom Erwartungswert
- Monitoring und Controlling der aktuell gültigen Restwerte
- Tagesaktuelle Bewertung des Leasingportfolios
- Messung des Einflusses bestimmter Ausstattungsmerkmale auf den Wiederverkaufspreis der Leasingfahrzeuge nach Vertragsende

Solche und ähnliche Berichte können einmalig auf Anforderung oder in Zusammenarbeit mit Fachbereichen und/oder Management auf OLAP-Ebene „ad-hoc“ konfiguriert werden und anschließend mit zusätzlicher Software standardisiert und visualisiert werden. Auch automatisch aktualisierte, dynamische Darstellungen und interaktive Dashboards können so problemlos und in gleichbleibender Qualität bereitgestellt werden.

Die zuvor beschriebenen Module, von der Datenerfassung und -integration bis hin zu Berichterstellung und Visualisierung, spielen eine wichtige Rolle bei dem in diesem Abschnitt beschriebenen, operativen Entscheidungsunterstützungssystem. Dennoch sind alle Ergebnisse und Berichte in dem hier verfolgten, speziellen Anwendungszweck von möglichst genauen Prognosen der aktuellen Markt- und zukünftigen Restwerte abhängig. Folglich stellen das Prognosemodul und seine interne Funktionsweise den Kern unseres EUS und damit auch den Fokus der hier vorgestellten Forschung dar. Abschnitt 3.2.2 beschreibt die für das Prognosemodell verwendeten Inputfaktoren.

3.2.2 Inputfaktoren für die Markt- und Restwertprognose

Die Fahrzeuge werden in die Datenbank aufgenommen, sobald nach Ende des Leasingvertrags der Wiederverkauf auf dem Gebrauchtwagenmarkt erfolgt ist. Auf diesem Wege werden die am Markt realisierten Restwerte in der Datenbank abgebildet, wobei zwischen *fahrzeugspezifischen* und *modellspezifischen* Variablen unterschieden wird. Die *fahrzeugspezifischen* Variablen beziehen sich auf die Vertragsvereinbarungen in den jeweiligen Leasingverträgen, wie km-Laufleistung, Leasingdauer (Fahrzeugalter) und die Aufpreise für Lackierung und Sonderausstattungen. Die *modellspezifischen* Variablen beschreiben die grundlegenden Eigenschaften des Fahrzeugs und sind innerhalb jeder Modellgruppe identisch. Die dieser Forschungsarbeit zugrundeliegenden Daten enthalten 928 verschiedene Modelltypen. Für weiterführende Analysen werden kategoriale Variablen wie „Kraftstoffart“ als Binärvariablen (pro vorkommender Kategorie) kodiert. Die resultierenden Merkmale sind in Tabelle 5 dargestellt.

Tabelle 5: Input- und Output-Variablen des Prognosemodells

Variable	Kodierung	Beschreibung
Unfallfrei	binär	Unfallfrei (ja/nein)
Farbe/Lackierung	kontinuierlich	Aufpreis für Sonderlackierung (in % UPE)
Kunde	binär	Kundenart (z.B. Endkunde oder Händler)
Vertriebszentrum	binär	Ort des Handelsbetriebs (Region, Bundesland)
Sonderausstattung	kontinuierlich	Aufpreis für Sonderausstattung (in % UPE)
Finanzierungsart	binär	Leasing, Finanzierung, Kauf
Initialer Listenpreis (UPE)	kontinuierlich	„Historischer“ Listenpreis des Fahrzeugs
Laufleistung	kontinuierlich	Kilometerstand des Fahrzeugs in km
Fahrzeugalter	kontinuierlich	Datum EZ bis Wiederverkauf in Tagen
Hubraum	kontinuierlich	Hubraum des Motors in Kubikzentimeter
Leistung	kontinuierlich	Leistung des Motors in kW
Antriebsart	binär	Allrad-Indikator (ja/nein)
Kraftstoffart	binär	Benzin, Diesel, CNG, LPG
Anzahl Gänge	binär	Anzahl Gänge {4, 5, 6, 7, 8}
Getriebeart	binär	Manuell, Wandlerautomatik, Doppelkupplung
Modellalter	kontinuierlich	Modellalter seit Markteinführung in Monaten
Modellspezifikation	binär	Feingranularere Spezifikation des Modelltyps
Restwert	kontinuierlich	Restwert in Prozent: $\left(\frac{\text{Wiederverkaufspreis}}{\text{Init. Listenpreis}}\right)$

Die dargestellten Variablen beziehen sich direkt auf die untersuchten Fahrzeuge. Für die explorative Analyse wurde ein einfacher linearer Regressionsansatz angewendet, um die Restwerte auf Basis der Faktoren aus Tabelle 5 (in-sample) zu erklären. Die Faktoren Fahrzeugalter und km-Laufleistung üben den stärksten negativen Einfluss auf den Restwert aus. Auch das Modellalter, sprich die Zeitspanne zwischen Markteinführung eines Modelltyps und dem individuellen Verkaufszeitpunkt eines Fahrzeugs dieses Typs, stellt einen entscheidenden Inputfaktor für die Prognose dar. Hier besteht die Herausforderung, dass das Alter eines bestimmten Modelltyps (kurz: „Modellalter“) mit der Zeit kontinuierlich steigt. Verglichen mit dem Alter eines spezifischen Fahrzeugs, wozu in den Daten zahlreiche Trainingsbeispiele vorhanden sind, ist das Modellalter (und dessen Auswirkungen) über den gegenwärtigen Zeitpunkt hinaus nicht bekannt.

Die Ergebnisse dieser Regressionsanalyse zeigen, dass die Auswirkung des Modellalters in der Stichprobe signifikant negativ ist, was bedeutet, dass Fahrzeuge bei steigendem Modellalter über die Zeit an Wert verlieren. Die Verwendung dieses Faktors als Input für Prognosen erfordert daher eine Extrapolation. Eine weitere, interessante Erkenntnis ist die, dass sich der Aufpreis für optionale Ausstattung negativ auf den Restwert auswirkt, was zeigt: Je mehr optionale Sonderausstattung ein Fahrzeug hat, desto größer ist sein prozentualer Wertverlust über die Zeit (im Verhältnis zum ursprünglichen Listenpreis). Anders ausgedrückt: Sonderausstattungen verlieren schneller an Wert als das Basisfahrzeug selbst.

3.2.3 Markttrends und saisonale Effekte

Nach abgeschlossener Durchführung der zuvor beschriebenen Regressionsanalyse werden deren Modellfehler verwendet, um die Saisonabhängigkeit und Preistrends auf dem Gebrauchtwagenmarkt zu untersuchen. Dieses Vorgehen liefert wichtige Erkenntnisse zur nachgelagerten Durchführung einer unverzerrten Prognose. Durch die Verwendung des Regressionsansatzes werden Einflüsse der oben genannten, unabhängigen Variablen eliminiert. Anschließend wird der folgende Ansatz angewendet: Erstens werden die Modellfehler im Hinblick auf den Faktor „Zeit“ sortiert. Zweitens werden die Modellfehler für jeden Monat des Beobachtungszeitraums gemittelt, was letztlich eine nach (Verkaufs-)Monaten gegliederte Zeitreihe gemittelter Modellfehler ergibt. In Abbildung 19 ist die resultierende Zeitreihe dargestellt.

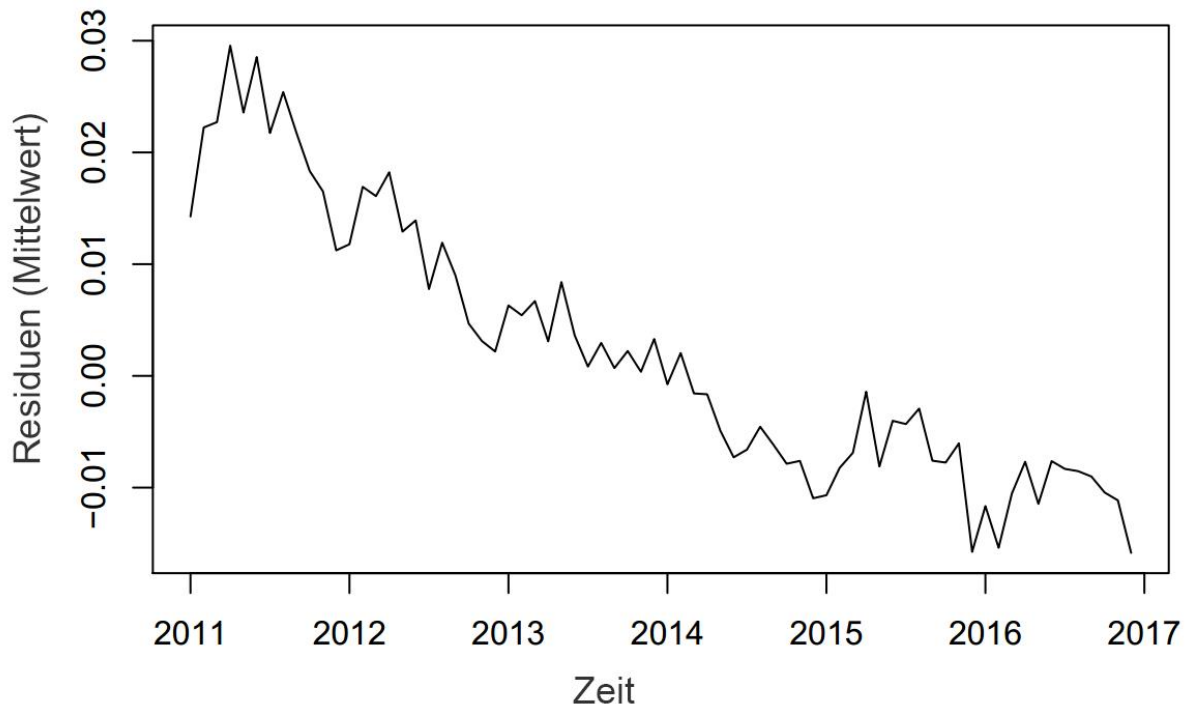


Abbildung 19: Der Einfluss des Faktors "Zeit" auf die Restwerte

Die Ergebnisse weisen während des 6-jährigen Beobachtungszeitraums einen klar erkennbaren Negativtrend auf, der in den letzten beiden Jahren (von 2015 bis 2017) leicht abflacht. Nachdem für alle vertragsspezifischen Variablen kontrolliert wurde, deuten die Ergebnisse auf einen Preisrückgang hin, der auf die allgemeinen Marktbedingungen zurückzuführen ist (das im Beobachtungszeitraum stetig steigende Modellalter wurde in der Regression bereits berücksichtigt).

Da saisonale Schwankungen ebenfalls von Interesse sind, wird als nächster Schritt eine Zerlegung der Zeitreihe durch die „Seasonal and Trend Decomposition“-Methode nach Loess (STL) (vgl. Cleveland et al., 1990) durchgeführt, wobei eine saisonale Periode von 12 (Monaten) angenommen wurde. Abbildung 20 zeigt die Zeitreihenzerlegung zur Bestimmung von Saison- und Trendeffekten. Die Ordinate zeigt die mittleren Residuen insgesamt sowie ihre drei Komponenten. Es ist zu beobachten, dass die Residuen sowohl einem Trend als auch saisonalen Effekten unterliegen. Die trendbereinigte Saisonkomponente zeigt, dass gegen Ende des Jahres niedrigere Restwerte erzielt werden als im Rest des Jahres, wonach die Restwerte im Frühjahr tendenziell höher sind. Mit einem Einfluss im Bereich von 0,8 Prozentpunkten des Listenpreises im Jahresverlauf sind diese Effekte jedoch von eher geringer Bedeutung.

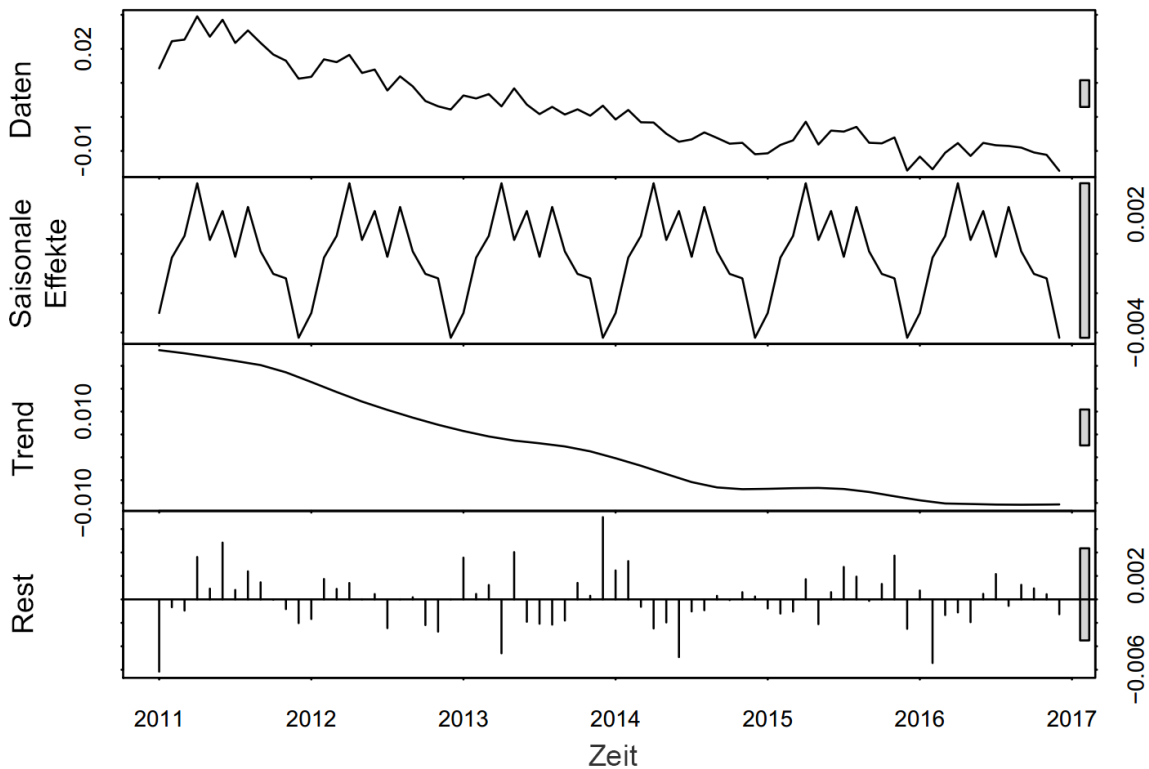


Abbildung 20: Zeitreihenzerlegung: Saisonale Effekte und Trends

Die vorangegangenen Beobachtungen veranschaulichen eine „zeitliche“ Abhängigkeit der Preise auf dem Gebrauchtwagenmarkt. Für weiterführende Analysen empfiehlt es sich folglich, zusätzlich entsprechende Dummy-Variablen (z.B. „Verkaufsmonat“) zu berücksichtigen, um mögliche Verzerrungen durch saisonale Einflüsse und Trends zu minimieren. Diese können zwar durch das Hinzufügen eines kontinuierlich zunehmenden Zeitfaktors in den Regressionsmodellen simuliert werden. Gleichwohl bedeutet die zuvor beschriebene Abflachung der preislichen Entwicklung in den letzten beiden Jahren, dass dieser Effekt schwer messbar ist und daher zu irreführenden (Prognose-)Ergebnissen führen kann.

3.2.4 Der Einfluss externer Faktoren auf Fahrzeugrestwerte

Um die Entwicklung von Restwerten auf dem Gebrauchtwagenmarkt zu verstehen und zu erklären, können auch externe Einflüsse wie der Ölpreis oder die Anzahl Neuzulassungen in einem bestimmten Monat wichtige Faktoren darstellen. In den Studien von Prado (2009) beispielsweise werden derartige Faktoren in den Prognosemodellen verwendet. Die von uns auf ihren Einfluss untersuchten Variablen sind in Tabelle 6 dargestellt.

Die Eignung externer Faktoren für die Verwendung in einem Prognosemodell hängt grundsätzlich davon ab, welches Ziel mit dem Modell verfolgt wird. Die Einbeziehung von Faktoren, die beispielsweise die aktuelle makroökonomische Situation widerspiegeln, ist nur dann sinnvoll, wenn das Ziel darin besteht, den *aktuellen* Marktwert eines bestimmten

Gebrauchtwagens zu schätzen. Für eine Anwendung im Bereich des Restwert- bzw. Risikomanagements muss hingegen eine tatsächliche Prognose des *zukünftigen* Fahrzeugrestwertes bereits bei Abschluss eines Leasingvertrags erstellt werden. Um dabei einen „Look-Ahead Bias“ zu vermeiden, sollten nur die zu diesem Zeitpunkt bekannten Faktoren in das Prognosemodell miteinbezogen werden. Dies bedeutet, dass nur die zum Zeitpunkt der Erstellung des Prognosemodells bekannten Werte für externe Inputfaktoren einbezogen werden dürfen, es sei denn, die Faktoren selbst werden ebenfalls vorhergesagt, was wiederum andere Verzerrungseffekte verursachen kann. Dennoch ist eine gründliche Untersuchung externer Einflüsse auf die Restwerte wichtig, um deren grundsätzliche Wirkungsweise und die Charakteristika des Gebrauchtwagenmarktes besser zu verstehen, denn eine möglichst exakte Schätzung aktueller Marktwerte ist gleichzeitig die Grundlage für entsprechend präzise Zukunftsprognosen.

In unserer Studie wurde eine Stichprobe von 49.297 Fahrzeugen, bestehend aus den zehn am häufigsten vorkommenden Modelltypen in der Datenbasis ausgewählt, um ein lineares Mixed-Effects-Modell zu schätzen. Fahrzeuge derselben Modelltyp-Klasse sind dabei identisch in Bezug auf ihre Leistung, Kraftstoffart, Getriebeart, Karosserietyp usw.; so können die jeweiligen Einflüsse aller individuellen, fahrzeugspezifischen Eigenschaften auf den Restwert effektiv aus dem Regressionsmodell eliminiert werden. Zusätzlich werden der Einfluss vertragsspezifischer Charakteristika wie Fahrzeugalter und Kilometerstand (zum Ende des Leasingvertrags) aus dem Modell entfernt und Dummy-Variablen für den Verkaufsmonat und ein Zeitfaktor berücksichtigt. Für jeden untersuchten externen Faktor ergeben sich verschiedene Achsenabschnitte sowie modelltypabhängige Steigungen. Die p-Werte in Tabelle 6 spiegeln die Ergebnisse des Likelihood-Ratio-Tests wider.

Bei Betrachtung dieser Ergebnisse fällt auf, dass keiner der untersuchten Effekte eine signifikante Erklärungskraft für den realisierten Restwert besitzt (d.h. keiner der p-Werte liegt unter 0,1). An dieser Stelle ist es jedoch wichtig, darauf hinzuweisen, dass die hier untersuchte Stichprobe rückblickend einen Zeitraum ohne größere wirtschaftliche Krisen oder anderweitige Störungen repräsentiert. Folglich kann aus den Ergebnissen nicht direkt geschlossen werden, dass Fahrzeugrestwerte von äußeren Einflüssen wie Marktschocks oder Wirtschaftskrisen völlig unabhängig sind. Unter normalen Marktbedingungen jedoch, die für eine allgemeine Prognose i.d.R. angenommen werden, ist ein systematischer Einfluss dieser Faktoren nicht nachweisbar.

Tabelle 6: Einfluss externer (makroökonomischer) Inputfaktoren

Inputfaktor	t-Wert	X²	p-Wert
DAX	1,41	1,9043	0,1676
Ölpreis	-1,39	1,8376	0,1752
Ifo-Geschäftsklimaindex	-0,87	0,7638	0,3821
Ifo-Teilindex zur Beurteilung der Geschäftslage	0,88	0,7822	0,3765
Ifo-Teilindex zu den Geschäftserwartungen	0,89	0,7929	0,3732
GfK-Konjunkturaussichten	0,08	0,0058	0,9390
GfK-Einkommenserwartung	-0,08	0,0056	0,9401
GfK-Anschaffungsneigung	-0,21	0,0459	0,8304
GfK-Konsumklima	-1,18	1,3827	0,2396
Arbeitslosenrate	0,5	0,2511	0,6163
Dieselpreis	-0,19	0,0368	0,8479
Benzinpreis (E5)	0,64	0,4113	0,5213
Verbraucherpreisindex (VPI)	0,73	0,5249	0,4688
EZB-Zinssatz	0,14	0,0190	0,8903
BIP	0,01	0,0002	0,9891
BIP-Wachstumsrate	-0,43	0,1832	0,6687
Kfz-Zulassungen im Wiederverkaufsmonat	1,1	1,1780	0,2778
Durchschnittliche Kfz-Zulassungen (vor 1 Jahr)	1,44	2,0017	0,1571
Durchschnittliche Kfz-Zulassungen (vor 2 Jahren)	1,17	1,3676	0,2422
Durchschnittliche Kfz-Zulassungen (vor 3 Jahren)	0,74	0,5523	0,4574
Durchschnittliche Kfz-Zulassungen (vor 4 Jahren)	-0,15	0,0226	0,8806
Durchschnittliche Kfz-Zulassungen (vor 5 Jahren)	0,35	0,1206	0,7284

Auch wenn lineare Modelle sich leicht interpretieren lassen und sich daher gut dafür eignen, um Rückschlüsse aus ihnen zu ziehen, so bringen sie doch gewisse Nachteile bei deren Anwendung für Prognosen mit sich. Obwohl KNN im Gegensatz dazu nur sehr schwer zu interpretieren sind, so sind sie dennoch in der Lage, komplexe Wechselbeziehungen innerhalb der Datenbasis zu „erlernen“, was sie letztlich zu einer geeigneten Methode macht, wenn die Prognosegenauigkeit im Mittelpunkt der Betrachtung steht.

3.3 Das Prognosemodul als Kern des EUS

Das Prognosemodul repräsentiert den Kern des hier vorgestellten Gesamtsystems zur Entscheidungsunterstützung. Während die der Prognose vorgelagerten Systemkomponenten für die automatisierte Beschaffung, Bereinigung, Aufbereitung und Bereitstellung der Inputdaten zuständig sind und die nachgelagerten Systeme die Prognoseergebnisse clustern, Mittelwerte bilden, auswerten und in bedarfsgerechtem Format in dynamische Reports umwandeln, so ist die eigentliche Prognose auf Einzelfahrzeugebene das Herzstück des Systems. Wenn im Folgenden von „Prognosen“ gesprochen wird, so sind damit jedoch nicht nur *zukünftige* Marktwerte gemeint. Auch bei den *aktuellen* Marktwerten handelt es sich um Prognosen, wenn auch nicht in Bezug auf die Zeit, sondern z.B. auf die Kilometerleistung. Aufgrund der Tatsache, dass keines der Gebrauchtfahrzeuge in der Datenbasis dem anderen zu 100% gleicht, jedoch allein für die Berechnung der aktuellen Marktwerte pro Modelltyp und Alters-/Kilometerleistungscluster eine Mindestmenge an Fahrzeugen mit nahezu identischen

Eigenschaften benötigt wird, haben wir uns dazu entschlossen, für jedes einzelne Fahrzeug in der Datenbasis eine individuelle Prognose (u.a. bzgl. seines Alters, der Kilometerleistung und des Verkaufszeitpunktes) zu erstellen. Diese umfasst insbesondere auch dessen *prognostizierten* Wert bei den aus fachlicher Sicht wichtigsten Kombinationen aus Kilometerleistung und Alter, die sich je nach Fahrzeugkategorie im Bereich des größten Leasingvertragsvolumens bewegen.

Die übliche Leasingdauer liegt dabei im Bereich von 12 bis 72 Monaten, jüngere Fahrzeuge werden als „junge Gebrauchte“ vermarktet und unterliegen einem anderen Geschäftsmodell. Ältere Fahrzeuge werden üblicherweise nicht von den Vertragshändlern des Herstellers sondern von Drittanbietern weitervermarktet und sind daher monetär für den Automobilhersteller nicht von Interesse. Die jährliche Kilometerleistung der Fahrzeuge liegt i.d.R. zwischen 10.000 und 70.000km (in 5000km-Schritten). Der überwiegende Anteil des Leasingvolumens liegt im mittleren Bereich, zwischen 24 und 48 Monaten und 10.000 bis 40.000km jährlicher Kilometerleistung. Die Festlegung der für das Halbjahr konzernweit gültigen Markt- und Restwerte erfolgt in genau diesen „Clustern“, also den entsprechenden Kombinationen pro Modelltyp aus Kilometerleistung pro Jahr und Alter in Monaten, z.B. „36/20“ für Fahrzeuge mit einem Alter von 36 Monaten bei 20.000km Laufleistung pro Jahr. Da es sich bei Gebrauchtwagendaten des Handels, die die wichtigste Grundlage für die Auswertungen des hier vorgestellten Gesamtsystems sind, um *echte* Fahrzeuge handelt, weichen die tatsächlichen Kilometerstände jedoch mitunter weit von den im Leasingvertrag vereinbarten Gesamtkilometerleistungen ab.

Sind beispielsweise in einem Cluster eines bestimmten Modelltyps, der für das Reporting ausgewertet werden soll, insgesamt 3.000 Fahrzeuge bei 36/20 enthalten, 2.500 dieser Fahrzeuge haben jedoch einen *realen* Kilometerstand von etwa 16.000km pro Jahr und die restlichen 500 liegen bei z.B. etwa 24.000km pro Jahr, so können sich Verzerrungen ergeben. Diese können letztlich die Ergebnisse so stark verfälschen, dass auf Basis dieser Informationen Fehlentscheidungen in der Restwertkommission getroffen werden.

Aus diesem Grund werden für jedes einzelne Fahrzeug in der Datenbasis diverse Prognosen erstellt, sowohl für verschiedene Zeithorizonte als auch für verschiedene Kilometerstände, um eine genauere Berechnung auch der *aktuellen* Marktwerte zu erreichen und Verzerrungen bei der Durchschnittsbildung im nachgelagerten, dynamischen Reporting zu verhindern. Für die Berechnung dieser Prognosewerte wurden in unserem Ansatz KNN verwendet. Deren Eignung

für diesen konkreten Anwendungsfall und die genaue Vorgehensweise im Gesamtkontext des Leasinggeschäfts werden im Folgenden detailliert vorgestellt und kritisch diskutiert.

3.3.1 KNN für die Restwertprognose: Eignung und Funktionsweise

KNN eignen sich besonders gut für die Analyse nicht-linearer Zusammenhänge (vgl. Bishop, 1995; von Mettenheim et al., 2010). Entsprechende Anwendungsfälle erstrecken sich von Vorhersagen im Finanzbereich (vgl. Sermpinis et al., 2014; Zimmermann et al., 2001) über das Risikomanagement (vgl. von Spreckelsen et al., 2014) bis hin zu EUS (vgl. Eilers et al., 2014; Gleue et al., 2017; Gleue et al., 2018; Kuo et al., 2011). Sie sind robust gegenüber stark verrauschten, unstrukturierten und/oder lückenhaften Daten (vgl. Schocken und Ariav, 1994), sind leistungsfähig bei der Mustererkennung (vgl. Zhang et al., 1998) und eignen sich folglich sehr gut für das hier vorgestellte Praxisproblem. Dennoch handelt es sich bei KNN lediglich um einen von vielen denkbaren „Supervised Machine Learning“-Ansätzen. Es existieren diverse weitere Ansätze (wie z.B. Support Vector Machines oder Gradient Boosting), doch der entscheidende Erfolgsfaktor für eine solche Anwendung ist nicht die Wahl des Algorithmus, sondern vielmehr die korrekte Auswahl und Verwendung der richtigen Variablen für die vorliegende Problemstellung (vgl. Domingos, 2012). Dieser Erkenntnis wird in der Folgenden Untersuchung durch eine besonders umfangreiche Analyse der möglichen Inputvariablen Rechnung getragen.

Insbesondere wenn es darum geht, verlässliche und vor allem reproduzierbare Ergebnisse zu erzielen, ist über eine korrekte Auswahl der Inputfaktoren hinausgehend zudem Erfahrung bei der Anwendung der ausgewählten Methode entscheidend. Dies ist eines der Hauptargumente für die Anwendung von KNN in diesem Anwendungsfall. Obwohl das Potenzial dieses Verfahrens dafür in vielen Studien bereits nachgewiesen werden konnte, ist eine sachgemäße Implementierung und Validierung wie auch eine gründliche Vorverarbeitung der Datenbasis eine Grundvoraussetzung für verlässliche Ergebnisse (vgl. Adya und Collopy, 1998).

KNN können als eine Methode zur nicht-linearen Funktionsapproximation verstanden werden. In der vorliegenden Arbeit werden dafür sog. „Feed-Forward-Networks“ verwendet, die aus mehreren Schichten an Neuronen („Multi-Layer“) bestehen. Die erste Schicht („Input Layer“) beschreibt dabei die unabhängigen Variablen, die zur Beschreibung des untersuchten Phänomens verwendet werden. Die Neuronen darin sind über die Gewichte θ_1 mit einer weiteren, versteckten Schicht („Hidden Layer“) verbunden, die wiederum über die Gewichte θ_2 mit der Outputschicht $h_\theta(X)$ oder einem weiteren „Hidden Layer“ verbunden ist. Der „Hidden Layer“ ist damit für die jeweils folgende Schicht die Inputschicht. Dieses Muster kann

beliebig oft wiederholt werden (beliebig hohe Anzahl an versteckten Schichten), wobei gleichzeitig jede Schicht auch eine beliebige Anzahl Neuronen beinhalten kann. Ein „Three-Layer-Feed-Forward“-KNN lässt sich demnach wie folgt definieren:

$$h_{\theta}(X) = \theta_2 \tanh(\theta_1 X) \quad (6)$$

Die versteckten Neuronen (engl. „hidden neurons“) und (optional) die Output-Neuronen wandeln die gewichtete Summe ihrer Inputs mittels einer Aktivierungsfunktion um, üblicherweise wird dafür der Tangens Hyperbolicus verwendet.

$$f(X) = \tanh(X) = 1 - \frac{2}{e^{2X} + 1} \quad (7)$$

Der Aufbau eines KNN ist in Abbildung 21 dargestellt. Um eine Funktionsapproximation durchführen zu können werden die KNN mit Trainingsdaten bzw. Trainingsmustern trainiert (Input x_k , Output y_k repräsentieren das Trainingsmuster k). Die zufällig initialisierten Parameter (Gewichte) werden mittels eines iterativen „Error Backpropagation“-Prozesses bestimmt, um den jeweiligen Inputvektor mit dem zugehörigen Output zu verbinden. Bei KNN handelt es sich um nicht-konvexe Optimierungen, was bedeutet, dass jeder Trainingsprozess auf derselben Datenbasis zu unterschiedlichen Ergebnissen führen kann (lokale Minima). Um nicht nur eine, ggf. zufällig schlechte Lösung auszuwählen, werden viele KNN mit zufällig ausgewählten Initialisierungen der Gewichte simultan trainiert und die jeweilige Performance auf Basis von Validierungsdaten, welche nicht in den Trainingsdaten enthalten sind, gemessen. Aus diesem Grund werden die Trainingsdaten aufgeteilt, zum einen in einen Datensatz, mit dem die KNN-Modelle tatsächlich trainiert werden, I_t , und zum anderen in einen Validierungsdatensatz, I_v , der der Bewertung der Out-of-Sample-Performance dient. Demzufolge werden 80% der verfügbaren Trainingsdaten für das Training der Modelle und die restlichen 20% für deren Validierung verwendet.

Die aus diesem Prozess resultierenden Ergebnisse werden letztlich auf einem unabhängigen Testdatensatz, der weder zum Anpassen der Gewichte (Modelltraining) noch zur Modellauswahl verwendet wird, im Hinblick auf ihre Performance bewertet. Während des Modelltrainings wird die Approximationsgenauigkeit der KNN-Modelle auf Basis ihrer Trainings- und Validierungsfehlerfunktionen

$$\varepsilon_t = \frac{1}{2} \sum_{k \in I_t} (h_{\theta}(x_k) - y_k)^2 \text{ und } \varepsilon_v = \frac{1}{2} \sum_{k \in I_v} (h_{\theta}(x_k) - y_k)^2 \quad (8)$$

ausgewertet. Um das sog. „Overfitting“ zu vermeiden, wird eine Kreuzvalidierung durchgeführt. Overfitting bezeichnet in der Statistik eine Überanpassung des Modells an die

Trainingsdaten derart, dass das Rauschen in den Trainingsdaten mittrainiert wird und so die tatsächlich geringere Anpassungsgüte verschleiert wird. Das Modell ist folglich zwar besser auf die Daten der Stichprobe angepasst, allerdings besteht aufgrund fehlender Generalität keine Übertragbarkeit auf die Grundgesamtheit. Um dieses Phänomen zu vermeiden wird ein „Early Stopping“-Ansatz angewendet. Mit dem Ziel, KNN-Modelle mit guten Generalisierungseigenschaften zu erhalten, wird das Training beendet, sobald der Fehler auf dem Validierungsdatensatz wieder ansteigt. Die beiden in diesem Prozess festzulegenden, wichtigsten Hyperparameter sind die Anzahl der versteckten Schichten und die jeweils darin enthaltene Anzahl Neuronen, wofür jedoch keine klar definierten Regeln genannt werden können. Deren optimale Anzahl hängt dabei sowohl von der Problemstellung und den verfügbaren Inputfaktoren als auch dem Stichprobenumfang ab, was bedeutet, dass das Experimentieren mit verschiedenen Topologien gute Ergebnisse hervorbringen kann.

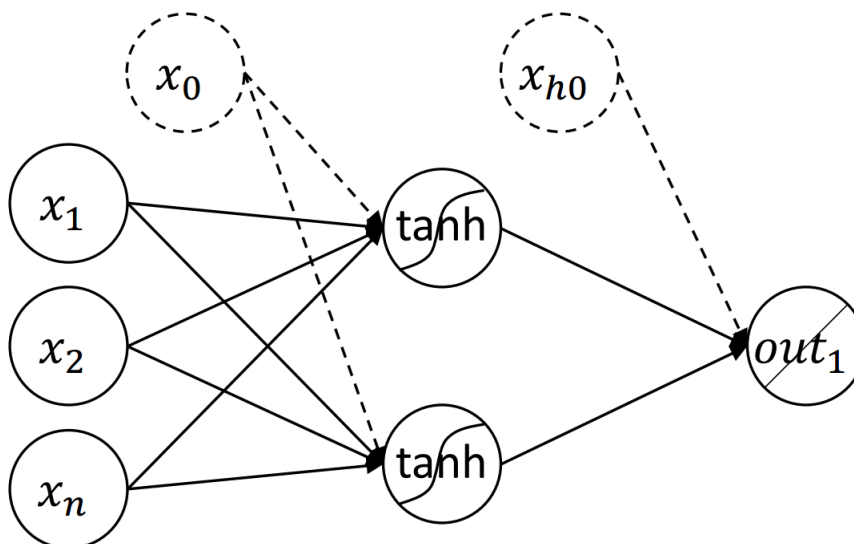


Abbildung 21: Schematische Darstellung eines (3-schichtigen "feed-forward"-) KNN

Die Anwendung von Verfahren wie beispielsweise das o.g. „Early Stopping“ gegen Overfitting erlaubt die Verwendung größerer Netzwerke ohne erhebliche Einbußen in der Modell-Performance, kann jedoch zu einer Erhöhung der Rechenzeit führen. In dem in dieser Studie vorgestellten Beispiel konnten mit zwei versteckten Schichten mit jeweils 50 und 25 Neuronen gute Ergebnisse erzielt werden.

3.3.2 Erkennung von Ausreißern und Datenbereinigung

Da KNN sensibel auf Ausreißer in den Trainingsdaten reagieren, muss vor dem eigentlichen Trainingsprozess eine gründliche Datenbereinigung durchgeführt werden. Zu diesem Zweck wird zur Vorverarbeitung ein kleines, dreischichtiges KNN mit allen in Tabelle 5 dargestellten Variablen und allen verfügbaren Trainingsdaten auf ein lokales Minimum trainiert.

Anschließend wird jeder Datensatz entfernt, für den $|h_{\theta}(x_k) - y_k|$ größer als ein Schwellwert c ist. Im vorliegenden Beispiel wurde dieser Wert als das 95% -Quantil der Fehler definiert, was bedeutet, dass 5% der Daten aus der Datenbasis entfernt werden. Eine nachgelagerte, manuelle Analyse der durch dieses Verfahren identifizierten Ausreißer zeigt, dass im Falle eines deutlich *höheren* realen Restwertes als durch das KNN geschätzt bei den betroffenen Fahrzeugen nachträglich stark restwerterhöhende Aufbauten (Kühlung, Werkzeugschränke, Kräne etc.) installiert wurden. Diese Aufbauten finden typischerweise nach Abschluss des Leasingvertrages (Zeitpunkt der Datenerhebung) durch externe Spezialdienstleister statt und führen daher ohne „messbare“ Indikation in der Datenbasis zu einer nachträglichen, massiven Werterhöhung des Neufahrzeugs und damit folglich zu einem erhöhten Restwert bei dessen Wiederverkauf als Gebrauchtwagen bzw. zum Ende des Leasingvertrags. Im Falle eines deutlich *niedrigeren* realen Restwertes des jeweiligen Fahrzeugs als durch das KNN geschätzt wurde der allgemeine Zustand des Fahrzeugs (z.B. „unfallfrei“) im manuellen Datenerhebungsprozess seitens des Händlers bei Eingabe in das Händlermanagementsystem nicht korrekt angegeben, was den erhöhten Wertverlust erklärt. Da es das erklärte Ziel dieses Forschungsprojektes ist, möglichst genaue Restwertprognosen für *unfallfreie* Fahrzeuge *ohne nachträgliche Aufbauten* zu ermitteln, wurden diese Ausreißer konsequent aus der Datenbasis entfernt.

3.3.3 Vorhersagekraft der Inputfaktoren und Prognosequalität

Neben der Aussagekraft der verschiedenen Faktoren bzgl. ihres Einflusses auf aktuelle Marktwerte ist insbesondere auch deren Vorhersagekraft im Rahmen einer KNN-Prognose von Interesse. Die Bewertung des individuellen Einflusses einer Variablen innerhalb eines KNN ist im Vergleich zu leichter interpretierbaren, linearen Modellen komplizierter, es besteht ein Zielkonflikt zwischen Nachvollziehbarkeit und Prognosequalität. Außerdem ist zu bedenken, dass je nach konkretem Anwendungsfall oder sogar Marktsegment sich die verfügbaren Daten, Faktoren sowie deren Abhängigkeiten stark unterscheiden können. Fundierte, fachliche Kenntnisse (in diesem Falle des Leasinggeschäfts) und zumindest eine grobe Vorstellung bzgl. des Einflusses einzelner (Input-) Faktoren auf die Restwerte, sprich des mit KNN zu prognostizierenden (Output-) Faktors, sind hier von ausschlaggebender Bedeutung. Diese fachliche Expertise wird letztlich auch benötigt, um später die Ergebnisse des komplexen Optimierungsmodells besser interpretieren zu können.

Um den jeweiligen Nutzen der einzelnen Variablen für das KNN-Prognosemodell zu bewerten, wird das „Input Perturbation Ranking“ nach Breiman (2001) angewendet. Dazu werden erneut alle Datensätze der zehn am häufigsten vorkommenden Modelltypen aus der Datenbasis

ausgewählt und auf deren Grundlage ein KNN trainiert, bis ein lokales Minimum erreicht ist. Dabei werden alle verfügbaren Inputfaktoren aus Tabelle 5 miteinbezogen, wie auch die externen Faktoren aus Tabelle 6, ein Zeitfaktor und die monatlichen Dummy-Variablen. Das trainierte KNN wird anschließend verwendet, um den Output für dieselben Daten n-fach zu berechnen, wobei n die Anzahl der betrachteten Inputfaktoren darstellt. Bei jedem Durchgang wird einer der Inputfaktoren durch zufälliges Mischen seiner Werte eliminiert. Auf diese Weise werden mögliche Beziehungen zwischen dem betreffenden Faktor und dem daraus resultierenden Output eliminiert, wobei jedoch statistische Eigenschaften wie der Mittelwert und die Standardabweichung beibehalten werden.

Das Ranking der Faktoren (nach ihrem Einfluss auf den Output-Faktor „Restwert“) erfolgt schließlich auf Basis des resultierenden In-Sample-Fehlers. Zur Beurteilung der Prognosegüte wird in diesem Fall dafür der mittlere quadratische Fehler (engl.: „Root Mean Square Error“, RMSE) verwendet (siehe Gleichung 9), den die KNN-Modelle mit den jeweils eliminierten Inputfaktoren erzeugen.

$$RMSE_n = \sqrt{\frac{1}{m} \sum_{k=1}^m (h_{\theta}(x_{n,k}) - y_k)^2} \quad (9)$$

Je höher der Fehler, desto wichtiger ist er für das Prognosemodell, bzw. desto höher ist sein Einfluss auf den Output des Modells. Seine resultierende Relevanz entspricht folglich dem Modellfehler des jeweiligen Modells mit einem eliminierten Faktor, geteilt durch den maximal auftretenden Fehlerwert über alle Faktoren. Der wichtigste Faktor erhält daher einen Wert von 1. Die zehn wichtigsten Inputfaktoren sind in Tabelle 7 in absteigender Reihenfolge dargestellt.

Tabelle 7: Inputfaktoren/Features nach Relevanz

Inputfaktor	RMSE	Relevanz
Fahrzeugalter	0,18670350	1
Km-Laufleistung	0,10002972	0,5357678
Anzahl Gänge	0,08053335	0,4313436
Listenpreis (zum Zeitpunkt der Erstzulassung)	0,07457451	0,3994275
Finanzierungstyp	0,07097613	0,3801543
Vertriebszentrum	0,06793025	0,3638403
Modellalter	0,06675006	0,3575191
Hubraum	0,06654379	0,3564143
Summe Aufpreis für Sonderausstattung	0,06652828	0,3563312
Leistung (kW/PS)	0,06614927	0,3543012

Die Baseline mit allen Inputfaktoren im „Urzustand“ liegt bei einem RMSE von 0,06114059. Wie erwartet haben die Faktoren „Alter“ und „Laufleistung“ den größten Einfluss auf den Output des Modells und damit auf den Restwert der Fahrzeuge. Eine ausgesprochen wichtige Erkenntnis aus dieser Analyse ist die Beobachtung, dass anscheinend interne, nur dem Hersteller bekannte Faktoren wie die UPE, die Art der Finanzierung oder das Vertriebszentrum

ebenfalls einen massiven Einfluss auf die Prognosequalität zeigen, während alle externen Faktoren am Ende der Ergebnisliste stehen und keinen besonderen Einfluss auf den resultierenden RMSE des Modells ausüben. Dieses Ergebnis unterstreicht die zuvor getroffene Aussage, dass externe Faktoren wie makroökonomische Indizes (DAX, Ölpreise, Arbeitslosenquote etc.) innerhalb unserer Stichprobe von 2011 bis 2017 weder für die Erklärung der Entwicklung aktueller Marktwerte noch zur Erhöhung der Prognosequalität eine Rolle spielen, sobald fahrzeugspezifische Faktoren mitberücksichtigt werden.

Anhand dieser Erkenntnisse soll an dieser Stelle nochmals ausdrücklich betont werden wie wichtig es ist, die exklusiv bei den Automobilherstellern vorhandenen, internen Daten zu erheben, zu sammeln, zu integrieren und letztlich dafür zu verwenden, die bestehenden Entscheidungsunterstützungssysteme (u.a. im Leasinggeschäft) zu verbessern.

3.3.4 Prognoseverfahren: Anwendungsszenarien und Benchmark

Im hier vorgestellten Unternehmenskontext können Prognosen zu verschiedenen Zwecken eingesetzt werden:

(1) Eine genaue Schätzung aktueller Marktwerte von Gebrauchtfahrzeugen beispielsweise kann zum einen die Händler bei marktgerechten Einpreisungen und/oder Verhandlungen auf dem Gebrauchtwagenmarkt unterstützen. Zudem dienen die gemittelten, aktuellen Marktwerte des Modellportfolios eines Automobilherstellers immer auch als Basis für die Berechnung der aktuell konzernweit gültigen Leasingkonditionen. Obwohl externe Faktoren, insbesondere im Falle von Krisen, die am Markt erzielbaren Preise durchaus beeinflussen könnten, wurde gezeigt, dass innerhalb des hier betrachteten Zeitraums von 2011 bis 2017 bei keinem der getesteten Faktoren ein signifikanter Einfluss auf die erzielbaren Gebrauchtwagenpreise nachweisbar ist. Auch bei der Schätzung aktueller Marktwerte handelt es sich im weiteren Sinne um Prognosen, da die Datenbasis aus vielen heterogenen Fahrzeugen mit individuellen Eigenschaften besteht. Eine einfache Durchschnittsbildung der Verkaufspreise würde daher zu vielfältigen Verzerrungen und Ungenauigkeiten führen. Um für einen bestimmten Modelltyp eines bestimmten Alters und mit einer bestimmten Laufleistung belastbare Durchschnittsmarktwerte bilden zu können, müssen Fahrzeuge eines Clusters zusammengefasst und vor der Durchschnittsbildung auf „Einzelfahrzeugebene“ homogenisiert werden. Beispielsweise würden zwei ansonsten identische Fahrzeuge, eines mit einer Laufleistung von 17.500km, eines mit einer Laufleistung von 22.500km, mithilfe des KNN-Modells auf eine einheitliche Laufleistung von 20.000km „vor- bzw. zurückgerechnet“.

(2) Ein weiterer Anwendungsfall ist die Prognose zukünftiger Marktwerte für Gebrauchtfahrzeuge, um Fragen zu beantworten wie z.B.: „Welchen Preis können wir für ein bestimmtes Fahrzeug oder einen bestimmten Modelltyp in einem Jahr am Markt erzielen?“ Die möglichst genaue Beantwortung dieser Frage ist besonders wichtig für das Risikomanagement eines Autoherstellers, denn realistische Erwartungen über erzielbare Preise am Gebrauchtwagenmarkt zu verschiedenen Zeitpunkten (aktuell und in der Zukunft) beeinflussen die Höhe der bilanziellen Rückstellungen des Herstellers und/oder der Leasinggesellschaft für das aktuelle Leasingportfolio. Diese sind notwendig, da die möglichen Verluste der Händler bei der Rücknahme der Leasingfahrzeuge entsprechend dem einleitend dargestellten Zusammenhang durch die Hersteller anteilig oder vollständig getragen werden, um ein funktionierendes und intaktes Händlernetz zu gewährleisten.

(3) Der dritte Anwendungsfall ist letztlich die Prognose von Fahrzeugrestwerten nach Ende eines Leasingvertrags, jedoch bereits zum Zeitpunkt des Vertragsbeginns. Diese Information für das komplette Modellportfolio vorab möglichst genau vorhersagen zu können unterstützt in hohem Maße die vielfältigen Managemententscheidungen, die bei der Festlegung bzw. der regelmäßigen Anpassung konzernweit gültiger Leasingraten getroffen werden müssen.

Um den ersten Anwendungsfall zu simulieren wurde ein initiales KNN-Modell mit allen Transaktionen der Jahre 2011 bis 2013 konstruiert. Wieder wurden alle Faktoren aus Tabelle 5, Pseudovariablen für die 12 Verkaufsmonate und ein allgemeiner Zeitfaktor verwendet. Die Daten aus dem darauffolgenden Halbjahr wurden für Testzwecke verwendet. Die Modelle wurden anschließend erneut mit den initialen Daten und zusätzlich den Daten der zuvor als Testmenge verwendeten Transaktionen trainiert. Für wiederholte Tests wurden wiederum die Daten der darauffolgenden 6 Monate herangezogen.

Die Ergebnisse in Tabelle 8 zeigen die Verbesserung der Prognosegenauigkeit des KNN-Modells im Vergleich zu einem regularisierten linearen Modell. Ein Ensemble-Durchschnitt von 30 verschiedenen KNN mit zufälligen Gewichtsinitialisierungen (was zu unterschiedlichen lokalen Optimierungsminima führt) wurde verwendet, um mögliche Prognosefehler einzelner Modelle zu minimieren. Für den linearen Benchmark verwenden wir einen Ridge-Regression-Ansatz. Der Regularisierungsparameter wird durch eine naive Grid-Suche (engl.: „naïve grid search“) und das leistungsstärkste Modell anhand eines Validierungsdatensatzes ausgewählt.

Das KNN übertrifft den linearen Ansatz in jedem der getesteten Zeiträume, da es in der Lage ist, nichtlineare Zusammenhänge in den Daten (auf Kosten der leichten Interpretierbarkeit) zu

erlernen. Insbesondere bei KNN ist es wichtig, den Einfluss der Inputfaktoren auf die Prognosegenauigkeit zu verstehen, wobei zeitabhängige Inputfaktoren eine wichtige Rolle spielen.

Tabelle 8: Prognosegenauigkeit für akt. Marktwerte: RMSE von KNN-Modell und Ridge-Regression

<i>Halbjahr</i>	<i>KNN</i>	<i>Linear</i>
2014/1	0,07403902	0,08272359
2014/2	0,07478254	0,08319453
2015/1	0,07331877	0,08253960
2015/2	0,07562336	0,08428427
2016/1	0,07326041	0,08350813
2016/2	0,07443271	0,08617191

KNN eignen sich besonders gut für *Interpolationsaufgaben*, bei denen die Informationen, von denen das KNN lernt, zum Zeitpunkt des Trainings tatsächlich zur Verfügung stehen. In der hier betrachteten praktischen Anwendung stehen wir vor der Herausforderung, zwei unabhängige Variablen miteinbeziehen zu müssen, die im Laufe der Zeit stetig zunehmen: Das Modellalter und der allgemeine Zeitfaktor. In einem realen Anwendungsfall müssen daher Modelle, die diese Faktoren berücksichtigen, in der Lage sein, zu *extrapolieren*.

Der zweite Anwendungsfall beschreibt die Problemstellung, die Modelle zur Prognose zukünftiger Marktwerte zu verwenden, ohne die Möglichkeit zu haben, sie für jede neue Entscheidung neu zu trainieren. Tabelle 9 zeigt die Ergebnisse für die Testperioden von 2014 bis 2016, jedoch auf Basis von Prognosemodellen, die lediglich auf Transaktionsdaten aus dem Zeitraum von 2011 bis 2013 trainiert wurden. Um die Extrapolationsfähigkeiten zu analysieren wurden Modelle verwendet, die jeweils mit und ohne zeitabhängige Faktoren trainiert wurden. Die Ergebnisse zeigen insgesamt eine Zunahme der Unsicherheit für spätere Zeithorizonte. Das beste Ergebnis für das erste Halbjahr erzielt das KNN-Modell, das den Zeitfaktor und das Modellalter berücksichtigt.

Tabelle 9: Prognosegenauigkeit für zuk. Marktwerte: RMSE von KNN-Modell und Ridge Regression

<i>Halbjahr</i>	<i>KNN (mit t)</i>	<i>Linear (mit t)</i>	<i>KNN (ohne t)</i>	<i>Linear (ohne t)</i>
2014/1	0,07403902	0,08272359	0,0751566	0,08405203
2014/2	0,07738575	0,08410964	0,07653743	0,08534023
2015/1	0,0770185	0,0839343	0,07636575	0,08659457
2015/2	0,0816638	0,08615005	0,07740517	0,08799502
2016/1	0,08367484	0,08551692	0,07874524	0,0903501
2016/2	0,09278381	0,08765334	0,0796895	0,08937356

Spätere Prognosen sind genauer, wenn keine zeitabhängigen Faktoren miteinbezogen wurden. Die Schwierigkeit von KNNs mit Extrapolationen kann durch einen Vergleich mit dem linearen Modell beobachtet werden, und zwar wenn beide Modelle zeitabhängige Faktoren enthalten. Die Unsicherheit des KNN-Modells steigt schneller in den Regionen außerhalb der Verteilung

der Trainingsdaten; in der zweiten Jahreshälfte 2016 übertrifft sogar das lineare Modell den KNN-Ansatz. Während das lineare Modell in allen analysierten Perioden von der Berücksichtigung der Zeitfaktoren profitiert, tut das KNN-Modell dies nur im ersten Halbjahr der Betrachtung. Dies gilt, auch wenn aus Tabelle 7 ersichtlich ist, dass zumindest das Modellalter einer der Faktoren mit der höchsten Vorhersagekraft in der Stichprobe ist. Der generelle Umgang mit zeitabhängigen Faktoren im Falle von Prognosen im Leasinggeschäft hängt daher im Wesentlichen vom betrachteten Zeithorizont und dem Marktkontext ab. In diesem Anwendungsfall können folglich zeitabhängige Faktoren bei der Prognose nur für einen kleinen Zeitrahmen von Vorteil sein („Kurzfrist-Zukunftsprognose“). Wenn diese Faktoren einen klaren (linearen) Trend aufweisen, kann es dennoch sinnvoll sein, die nichtlinearen Modelle nachträglich anzupassen, z.B. durch eine lineare Adjustierung der Interpolationsergebnisse. Eine Extrapolation im Rahmen von KNN kann jedoch zu irreführenden Ergebnissen führen.

Der dritte Anwendungsfall ist die Prognose von Fahrzeugrestwerten am Ende des Leasingvertrags, jedoch bereits zum Zeitpunkt des Vertragsbeginns. Auch hier werden alle Transaktionsdaten der ersten drei Jahre (2011 bis 2013) für das Training der Modelle verwendet (ohne zeitabhängige Faktoren). Alle Leasingverträge, die in den darauffolgenden sechs Monaten abgeschlossen werden, werden anhand dieser Modelle bewertet und die Prognosen bei der Rückgabe dieser Fahrzeuge (ab 2015) evaluiert, sprich mit den realen Transaktionspreisen in 2015 abgeglichen, die dem Modell jedoch nicht „bekannt“ sind. Die Ergebnisse sind in Tabelle 10 dargestellt. Es ist zu beobachten, dass KNN in allen betrachteten Zeiträumen das lineare Modell (bzgl. ihrer Prognosegenauigkeit) übertreffen. Je älter die Fahrzeuge werden, desto schneller steigt dabei auch die Unsicherheit.

Tabelle 10: Prognosen für einzelne Fahrzeuge: RMSE des KNN-Modells und der Ridge Regression

<i>Halbjahr</i>	<i>KNN</i>	<i>Linear</i>
2015/1	0,06396709	0,07228487
2015/2	0,06782139	0,07600988
2016/1	0,07978839	0,08955229
2016/2	0,0960669	0,1029367

Neben den oben dargestellten Benchmarks sind die mit dem hier vorgestellten Ansatz erzielten Performance-Verbesserungen in der Praxis durchaus wirtschaftlich relevant. Im Vergleich zu früheren, vergleichsweise simplen Prognosen auf Basis von Daten externer Restwertinstitute konnte unter Einbeziehung der internen Daten des Herstellers eine durchschnittliche Verbesserung in der Größenordnung von mindestens drei Prozentpunkten des ursprünglichen Listenpreises näher am tatsächlichen Restwert erzielt werden.

Dieses Ergebnis unterstreicht einmal mehr die Bedeutung der Nutzung aller intern verfügbaren Daten im Bereich der Entscheidungsunterstützung im Leasinggeschäft.

3.4 Kritische Würdigung und Diskussion der Ergebnisse

In diesem Teil der vorliegenden Dissertation wurde unter Zuhilfenahme von Methoden aus den Forschungsgebieten Data Mining, Big Data Analytics und der Künstlichen Intelligenz ein datengetriebener Ansatz zur Prognose von Fahrzeugrestwerten entwickelt. Dabei wurde insbesondere untersucht, wie ein auf dieser Technologie und Methodik basierendes EUS in der wirtschaftlichen Praxis operationalisiert werden kann und die gewonnenen Erkenntnisse in einem 3-jährigen Praxisprojekt bei einem großen deutschen Automobilhersteller implementiert. Dabei stand neben der Schaffung von Transparenz am Gebrauchtwagenmarkt die möglichst genaue Vorhersage von Zukunftsrestwerten im Vordergrund, wobei das Prognosemodul letztlich den entscheidenden Bestandteil eines umfangreicheren EUS darstellt.

Die mithilfe dieses EUS geschaffene Markttransparenz und die nachweislich deutlich verbesserte Prognosequalität sind ausschlaggebend für das Risikomanagement und gleichzeitig wichtig für die Rückstellungsbildung sowie die Preisgestaltung im Leasinggeschäft, die eine gleichermaßen hohe (wirtschaftliche und strategische) Bedeutung für Automobilhersteller und -händler besitzen. Eine systematische Verzerrung in beide Richtungen (über- oder unterschätzte Restwerte) bringt kurz- mittel- und langfristig negative Konsequenzen für Händler und Hersteller mit sich, beeinträchtigt die Wettbewerbsfähigkeit bzw. die Höhe der notwendigen Verkaufshilfen und/oder hat einen maßgeblichen Einfluss auf die Handelsmargen bei Wiederverkauf der Fahrzeuge.

Das Hauptziel dieses Projektes, die Etablierung eines rein datengetriebenen Ansatzes zur Restwertprognose von Leasingfahrzeugen, wurde erreicht und führte durch enge Zusammenarbeit zwischen dem Institut für Wirtschaftsinformatik, dem Fachbereich Finanz und dem Entwicklungsteam des Business Intelligence Competence Centers (BICC) des Automobilherstellers zu gesteigerter (Technologie-) Akzeptanz und Vertrauen innerhalb des Konzerns.

Die Ergebnisse zeigen, dass KNN für die in der Praxis häufig verrauchten und unstrukturierten Daten gut geeignet sind. Dennoch ist eine gründliche Datenaufbereitung und Ausreißererkennung unabdingbar, um gute und verlässliche Ergebnisse zu erzielen. Es wurde gezeigt, wie KNN zur Vorverarbeitung und Filterung der Datenbasis verwendet werden und wie ein Prognosemodell auf Basis eines KNN-Ansatzes entworfen werden kann. KNN werden oft wegen ihres Black-Box-Charakters kritisiert. Durch eine transparente Beschreibung der

Topologie, die genaue Untersuchung der Stärke des Einflusses der Inputvariablen auf den Restwert und einen sorgfältigen Datenbereinigungsprozess ist es möglich, dieses Problem zu entschärfen und die Ergebnisse verständlich und reproduzierbar zu machen.

KNN werden häufig nur als alternatives Instrument zum Benchmarking verschiedener Methoden eingesetzt. Anstatt jedoch diverse Arten maschineller Lernansätze miteinander zu vergleichen, sind insbesondere die zur Verfügung stehenden Daten, die KNN-Modellspezifikation und das Training die wichtigsten Erfolgsfaktoren. So ist eines der Hauptziele dieser Studie die Untersuchung der restwertbestimmenden Faktoren, während KNN als eine mögliche Technologie verstanden werden können, um zuverlässige Ergebnisse zu erzielen. Dennoch sind zahlreiche Möglichkeiten denkbar um die Prognoseergebnisse weiter zu verbessern, z.B. durch eine weniger fachlich, stärker datengetriebene Clusterung der einzelnen (Fahrzeug-)Modelltypen, eine weitere Optimierung von Hyperparametern oder den Einsatz anderer Regularisierungstechniken. Asymmetrische Kostenfunktionen für den Lernalgorithmus, die verschiedene Prognosefehler nach ihren jeweiligen wirtschaftlichen Auswirkungen gewichten, könnten ebenfalls von Vorteil sein, um dem jeweiligen Geschäftskontext besser Rechnung zu tragen (vgl. Lessmann, 2013).

In der Stichprobe aus dem Zeitraum 2011 bis 2017 konnte weder ein signifikanter erklärender noch einen prognoserelevanten Einfluss extern verfügbarer (makroökonomischer) Faktoren wie dem Ölpreis nachgewiesen werden. Im Gegensatz dazu sind die für den hier untersuchten Anwendungsfall wichtigsten Faktoren Variablen wie der anfängliche Listenpreis oder der Finanzierungstyp, wobei es sich ausschließlich um interne Informationen handelt, sprich Daten, die nur dem Hersteller zur Verfügung stehen. Demzufolge ist es deutlich wichtiger, die im Unternehmen verfügbaren Daten zu sammeln und ihr Potenzial auszuschöpfen, als sich auf externe Daten und/oder Dienstleister zu verlassen.

Die hier vorgestellten Ergebnisse gelten jedoch ausdrücklich nur für die im Rahmen dieses Forschungsprojektes untersuchte Stichprobe, die (auch implizit) keine Informationen über größere Krisen oder wirtschaftliche Schocks enthält. Diese Einschränkung gilt insbesondere für die Langfristprognosen in dieser Untersuchung. Um valide Aussagen über den Einfluss externer Effekte oder gar disruptiver Veränderungen der Marktbedingungen abzuleiten ist eine größere Datenmenge, mindestens über den Zeitraum eines vollständigen Konjunkturzyklus erforderlich. Bei der zuvor beschriebenen, tiefgehenden Analyse zeitabhängiger Faktoren wurde ein deutlicher Abwärtstrend der Restwerte über die Zeit und bei zunehmendem Modellalter festgestellt. Ein Grund dafür könnte die in den Augen des Kunden nachlassende

Attraktivität eines bestimmten Fahrzeugmodells im Zeitverlauf sein. Anders ausgedrückt neigen Kunden dazu, neuere Produktgenerationen den älteren Modellen vorzuziehen, was sich letztlich im Restwert der entsprechenden Fahrzeugmodelle manifestiert.

Darüber hinaus führt eine saisonale Komponente des Gebrauchtwagenmarktes zu vergleichsweise niedrigen Restwerten im Dezember und weist höhere Restwerte im ersten Quartal eines Jahres auf. Diese empirische Beobachtung wurde durch den Hersteller bestätigt und kann beispielsweise durch Handelsbetriebe erklärt werden, die versuchen, ihre Verkaufsziele gegen Ende des Jahres noch zu erreichen. Ob diese Beobachtungen ein spezielles Phänomen des in unserer Studie untersuchten Marktsegments (leichte Nutzfahrzeuge) darstellen ist ein Aspekt, der in weiterführenden Analysen dieses Musters über eine breitere Palette von Fahrzeugsegmenten untersucht werden muss, um hier allgemeingültigere Aussagen treffen zu können.

Die Prognosesimulation in Abschnitt 3.3.4 veranschaulicht die verschiedenen Anwendungsszenarien des Systems. Die größte Herausforderung bei der hier vorgestellten Aufgabenstellung besteht darin, die zeitabhängigen Faktoren korrekt zu berücksichtigen. Es wurde gezeigt, dass KNN linearen Modellen überlegen sind, auch wenn sie sich nicht explizit für Extrapolationsaufgaben wie die Berücksichtigung des stetig zunehmenden Modellalters eignen. Weiterführende Untersuchungen sollten diesen Teilaspekt näher betrachten und beispielsweise einen Benchmark verschiedener Extrapolationstechniken durchführen. Kombinationen aus mehreren Modellen, wie beispielsweise der Einsatz von KNN für den Interpolationsanteil und lineare „Nachjustierung“ der Ergebnisse basierend auf den jeweils vorkommenden zeitabhängigen Faktoren, könnten zu einer weiteren Verbesserung der Langfristprognosen führen.

Im Gesamtkontext dieser Dissertation und in Bezug auf die übergeordnete Forschungsfrage sollen die in diesem, zweiten Hauptteil vorgestellten Anwendungsfälle als ein allgemeines Beispiel dafür dienen, dass Unternehmen mithilfe von Methoden und Tools aus den Bereichen Data Mining, Business Intelligence, Big Data Analytics und KI aus Daten, die häufig bereits in großer Menge und ausschließlich intern verfügbar sind, wertvolle, geschäftskritische Erkenntnisse gewinnen können.

Diese Informationen, integriert und in verständlicher Form aufbereitet und visualisiert, können eine maßgebliche Unterstützung bei strategischen Managemententscheidungen sein und/oder Transparenz über deren Konsequenzen schaffen. Im speziellen Fall des in Abschnitt 3.2 vorgestellten EUS konnten über die ursprüngliche Zielsetzung hinaus bereits zahlreiche

Anwendungsfälle jenseits der Restwertprognosen realisiert werden. Von Benchmarks der Gebrauchtwagen-Verkaufperformance einzelner Handelsbetriebe bis zur Definition der optimal restwerterhaltenden Ausstattungskonfiguration von Leasingfahrzeugen für Großkundenaufträge sind zahlreiche Aufgaben für ein solches System ohne großen zusätzlichen Zeitaufwand realisierbar. Auch die Simulation verschiedener Marktszenarien, wie z.B. die Risikobewertung des aktuellen Leasingportfolios, wenn aufgrund unerwarteter, externer Schocks wie Finanzkrisen, neuen Gesetzen oder Entscheidungen der Wettbewerber die Marktwerte plötzlich stark sinken, ist problemlos möglich. Die Ergebnisse solcher "Stresstests" unterstützen die Finanz-/Controlling-Abteilung des Herstellers bei der Rückstellungsbildung in geeigneter Höhe zur Abdeckung dieser Risiken.

Weitere Konzepte zur wertorientierten Datenanalyse könnten in diesem Zusammenhang die Entwicklung neuer Geschäftsmodelle im Leasinggeschäft allgemein umfassen. Ein Beispiel dafür wäre die Implementierung von Fahrzeugsensordaten, um eine Kalkulation flexibler Leasingraten in Abhängigkeit von der beabsichtigten Fahrzeugnutzung zu ermöglichen. Kunden und Hersteller könnten so gleichermaßen von niedrigeren Leasingraten bzw. Wettbewerbsvorteilen aufgrund einer geringeren Prognoseunsicherheit profitieren. Live-Sensordaten können auch zur Beurteilung des Verschleißes herangezogen werden; entsprechende Projekte wurden bereits diskutiert.

Einer der größten Nachteile von Systemen, wie sie in unseren Studien vorgestellt werden, ist, dass sie in der Regel sehr schnell sehr komplex werden. Das Verständnis des gesamten Ergebnisermittlungsprozesses erfordert ein tiefgreifendes technisches und mathematisches Know-how und vor allem Zeit, um die einzelnen Prozesse gründlich nachvollziehen und verstehen zu können. An dieser Stelle wird die Komplexität zum Problem, da die dafür notwendige Zeit im Tagesgeschäft, insbesondere im Management, meist ein knapper Faktor ist. Entscheidungsträger müssen sich auf die Ergebnisse verlassen können, vor allem, wenn wichtige Entscheidungen auf deren Basis getroffen werden sollen. Im operativen, praktischen Einsatz des EUS im Rahmen des hier vorgestellten Projektes konnten wir die Erfahrung sammeln, dass sich Technologieakzeptanz und Vertrauen in die Ergebnisse durch bestimmte Maßnahmen steigern lässt, z.B. durch ein übersichtliches Dashboard für das kontinuierliche Monitoring der Datenqualität und allgemein durch Berichte, deren Erscheinungsbild für den Empfänger nicht völlig neu ist.

Neben einem möglichst selbsterklärenden Berichtsdesign und der im Konzern üblichen Farbwahl liegt die besondere Herausforderung im Reporting darin, komplexe Zusammenhänge

möglichst einfach zu visualisieren und einzelne Berichte nicht mit zu viel Information zu überfrachten. Darüber hinaus ist die professionelle, bedarfsorientierte Kommunikation von Projekt- und Analyseergebnissen nach dem Grundsatz „so wenig wie möglich, so viel wie nötig“ und letztlich eine enge Zusammenarbeit zwischen Managern, Analysten und Entwicklern unumgänglich und ist wichtig, um das notwendige Vertrauen zu gewinnen. Da leicht verständliche Berichte Komplexität reduzieren, Vertrauen in die Ergebnisse schaffen und damit auch die Kommunikation erleichtern, sollte sich die zukünftige Forschung insbesondere auf die richtige Visualisierung von Daten und Analysen konzentrieren (vgl. Eilers et al. 2017).

3.5 Fazit

Im zweiten Hauptteil (Kapitel 3) dieser Dissertation wurde untersucht, mit welchen Mitteln sich das Restwertisiko für Automobilhersteller und Leasinggesellschaften in ihrem Tagesgeschäft reduzieren lässt. Da sowohl das Restwertisiko selbst als auch der Zeitpunkt, zu dem es ggf. eintritt, aus der Perspektive eines Herstellers maßgeblich von den in Leasingverträgen festgelegten Zukunftsrestwerten der Fahrzeuge abhängt, stand deren möglichst genaue Prognose im Mittelpunkt der Untersuchung.

Um die preisbestimmenden Faktoren auf dem Gebrauchtwagenmarkt besser zu verstehen wurde die Stärke des Einflusses interner und externer Variablen auf die Restwerte analysiert und nachgewiesen. Während interne, vom Hersteller direkt beeinflussbare Faktoren wie der initiale Listenpreis und die Finanzierungsart einen starken Einfluss auf die Restwerte haben, zeigen die Ergebnisse dieser Fallstudie auch, dass externe, makroökonomische Faktoren wie der Ölpreis, Aktienindizes oder die Arbeitslosenrate unter normalen, beständigen Marktbedingungen keine signifikante Erklärungs- oder Vorhersagekraft besitzen.

Dies unterstreicht nochmals wie wichtig es ist, die exklusiven, in Konzernen häufig bereits in großer Menge vorhandenen, internen Daten zur Realisierung von Wettbewerbsvorteilen zu nutzen. Die Ergebnisse zeigen, dass auch in diesem speziellen Anwendungsgebiet, dem Leasinggeschäft in der Automobilindustrie, insbesondere die internen Variablen den größten Einfluss auf die Restwerte besitzen. Entsprechend sinnvoll und lohnenswert ist es, diese möglichst genau zu analysieren. Der im Rahmen des hier vorgestellten Promotionsprojektes entwickelte Prognoseansatz basiert hauptsächlich auf KNN, deren überlegene Performance im Vergleich zu linearen Modellen deutlich nachgewiesen werden konnte. Anhand dreier typischer Anwendungsfälle für Restwertprognosen wurde jedoch auch gezeigt, dass deren Genauigkeit, insbesondere im Falle der nichtlinearen Modelle, stark vom Prognosehorizont abhängt. Der untersuchte Einfluss zeitabhängiger Faktoren muss als generelle Herausforderung für

Restwertprognosen in allen (Anwendungs-)Bereichen gesehen werden und sollte daher in der zukünftigen Forschung besondere Beachtung finden.

Zusammenfassend kann an dieser Stelle die Aussage getroffen werden, dass Unternehmen in der Regel unbeschränkten Zugang zu großen Mengen exklusiver, unstrukturierter interner Daten haben, die, ordnungsgemäß bereinigt, aufbereitet und korrekt interpretiert, das Potenzial besitzen, im Tagesgeschäft und insbesondere durch die datengetriebene Beantwortung wichtiger, strategischer Fragen entscheidende Wettbewerbsvorteile zu generieren. Dieses Potenzial sollte nicht ungenutzt bleiben, solange die so gewonnenen Informationen und Ergebnisse letztlich von Menschen mit entsprechender Erfahrung und tiefgreifenden Kenntnissen innerhalb des untersuchten Fachgebietes interpretiert und bewertet werden. In der Praxis ist für derartige Projekte eine enge Abstimmung zwischen diesen Experten, Entwicklern und Entscheidungsträgern der Schlüssel zum Erfolg.

4 Kritische Reflexion und Ausblick

*“Mankind's greatest achievements have come about by talking
and its greatest failures by not talking.
It doesn't have to be like this.”*

– *Stephen Hawking*

Diese Dissertation stellt eine Zusammenfassung der Ergebnisse und Erkenntnisse aus zwei voneinander unabhängigen Forschungsprojekten dar, wovon eines in einem rein wissenschaftlich geprägten Umfeld (vgl. Kapitel 2) und eines (insbesondere aus meiner persönlichen Perspektive) in einem stark wirtschaftlich geprägten Umfeld, der Automobilindustrie, durchgeführt wurde (vgl. Kapitel 3). Da jeder der beiden Hauptteile für sich bereits kritisch diskutiert und jeweils einen Ausblick auf zukünftige Forschungsthemen enthält, soll der folgende Abschnitt dafür genutzt werden, die Gemeinsamkeiten, Unterschiede und Herausforderungen bei der Durchführung von IT-Projekten mit Bezug zu datengetriebenen Technologien in beiden Umfeldern mit größerem Abstand zu betrachten.

Ein übergeordnetes Ziel dieser beiden Projekte war es, zu zeigen, dass Methoden aus dem Gebiet der Künstlichen Intelligenz, des Data Mining und Big Data Analytics helfen können, bestehende Prozesse und Arbeitsabläufe zu optimieren, dass es sich lohnen kann, Althergebrachtes kritisch zu hinterfragen, und dass der Einsatz dieser Verfahren sowohl in der Wissenschaft als auch in der Wirtschaft beiderseitigen Nutzen stiften kann. Je nachdem, in welchem dieser beiden, grundlegend verschiedenen Umfeldern sich ein IT-Projekt bewegt, ergeben sich jedoch unterschiedliche Herausforderungen: Während das in Kapitel 2 vorgestellte Projekt zur Optimierung von Literaturrechercheprozessen hauptsächlich auf eine Zeitersparnis in der täglichen Arbeit von Wissenschaftlern abzielt, so liegt der Nutzen des in Kapitel 3 vorgestellten Kooperationsprojektes mit der Automobilindustrie in letzter Konsequenz in den monetären Vorteilen, die sich der Hersteller durch das Treffen „besserer“ bzw. besser informierter Entscheidungen bzgl. seines Leasinggeschäfts und seines Risikomanagements verschaffen kann. Da ein Projekt dieser finanziellen Größenordnung (Integration einer zweistelligen Anzahl von Konzernsystemen, Aufbau eines gesamten DWH, „Einkauf“ externer Softwareentwicklungs- und Data Science-Expertise) und einer mehrjährigen Entwicklungszeit unter Beteiligung ganzer Fachabteilungen sich wirtschaftlich rechnen *muss*, ist der kurzfristige Erfolgsdruck erheblich größer und die Kontrolle der Projektfortschritte deutlich engmaschiger,

als dies (zumindest nach meiner persönlichen Erfahrung) in rein wissenschaftlichen Projekten der Fall ist.

Dabei ist mir besonders die Herausforderung in Erinnerung geblieben, auch nach einer erfolgreichen „Proof of Concept“-Phase regelmäßig neue, wissenschaftlich valide und gleichzeitig praxisrelevante Ergebnisse präsentieren zu können, um einerseits den teilweise enormen Ressourcenaufwand zu rechtfertigen und andererseits bereits frühzeitig Vertrauen in die gewählte Vorgehensweise und Technologie zu vermitteln.

Vertrauen ist einer der primären Faktoren zur Schaffung von Akzeptanz (vgl. Gefen et al., 2003), und Technologieakzeptanz auf breiter Ebene ist dabei ein wichtiger Aspekt: Denn um wirklich praxisrelevante Entscheidungsunterstützungssysteme, z.B. wie in unserem Fall mit Komponenten der Künstlichen Intelligenz, in einem Konzern mit historisch gewachsenen Prozessen erfolgreich einzuführen, sind nicht nur „Leuchtturmprojekte“ gefragt, sondern Systeme, die sowohl das Management als auch die Mitarbeiter einer Fachabteilung in ihrer täglichen Arbeit zuverlässig unterstützen. Nur wenn bereits während der Projektlaufzeit diese Akzeptanz durch zielgerichtete und bedarfsgerechte Kommunikation geschaffen werden kann, können Methoden des maschinellen Lernens sowie aus dem Gebiet der Data Science und Big Data Analytics ihre wirkliche Macht entfalten.

Bei Projekten dieser Art sind Kooperationen zwischen Wissenschaft und Wirtschaft notwendig: Einerseits werden die neuen Technologien, die den höchsten Informationsgewinn und damit den größten Wettbewerbsvorteil und finanziellen Nutzen in der Wirtschaft versprechen, häufig an Forschungsinstituten und Universitäten entwickelt, was im Übrigen auch die dort vorherrschende Akzeptanz dieser Technologien entsprechend erhöht. Andererseits verfügen Wirtschaftsbetriebe üblicherweise über große Mengen der für die Forschung so wertvollen, realen Daten, mit denen sich neueste Forschungsergebnisse in der Praxis testen, evaluieren, bewerten und ggf. bestätigen lassen. Der beiderseitige Nutzen in solch einem Szenario kann sehr hoch sein, doch müssen dazu einige Zielkonflikte überwunden werden, die mir in dieser Form in einem rein wissenschaftlichen Projekt nicht begegnet sind:

Wissenschaftlich valides Vorgehen vs. schnelle, praxisrelevante Ergebnisse: Aufgrund der hohen Kosten eines mehrjährigen IT-Projektes unter Beteiligung von internen und externen Fachexperten, Softwareentwicklern, Wissenschaftlern und IT-Beratern entsteht nach meiner persönlichen Erfahrung generell ein vergleichsweise hoher Erfolgsdruck. Dies gilt nicht nur für die Projektmitarbeiter, sondern auch für den Auftraggeber, der seinerseits an höhere Managementebenen den finanziellen Aufwand regelmäßig rechtfertigen und entsprechende

Erfolge präsentieren muss. Dies führt über kurz oder lang zu einem Konflikt wissenschaftsethischer Natur, denn „schnelle Zwischenergebnisse“ entsprechen häufig nicht den in der Wissenschaft üblichen Standards bzgl. Validität, Methodik, Reproduzierbarkeit und Überprüfbarkeit. Dennoch muss hier durch zielgerichtete, lösungsorientierte Kommunikation und gegenseitige Kompromissbereitschaft ein Mittelweg gefunden werden, um letztlich gemeinsam erfolgreich zu sein. Auch hier ist gegenseitiges Vertrauen von großer Bedeutung, z.B. dass das Herstellen eines validen Zwischenergebnisses länger dauert als vom Management gewünscht, oder auch ein nicht 100%ig nach wissenschaftlichen Standards valides Ergebnis dennoch bereits vorgestellt werden kann. Für Wirtschaftsinformatiker mit wissenschaftlichem Hintergrund in einem Wirtschaftsprojekt ist genau das der „Spagat“, den erfolgreich zu meistern die größte Herausforderung darstellt. Dazu ist neben der Kenntnis sowohl der Fachlichkeit(en) als auch der notwendigen Technologie, Mathematik und Statistik auch Durchsetzungsfähigkeit, Kompromissbereitschaft, Fingerspitzengefühl sowie die Fähigkeit, sich und seine Arbeit kritisch zu hinterfragen (Was habe ich geliefert? Was wird wirklich benötigt? Sind meine Ergebnisse in der Praxis relevant?) wichtige Eigenschaften.

Technologische Beschränkungen: Großkonzerne haben i.d.R. sehr restriktive IT-Landschaften, insbesondere was die einzusetzende Software betrifft. In einem Forschungsprojekt in einem rein wissenschaftlichen Umfeld würde man z.B. für die Berechnung von KNN dasjenige Softwarewerkzeug verwenden, das nach einem sorgfältig durchgeführten Benchmark die genauesten Ergebnisse erzielt. Allein die Auswahl des am besten für den jeweiligen Anwendungsfall geeigneten Werkzeuges wäre ein kleines Projekt für sich und möglicherweise ein zeitintensiver Prozess, während in einem Konzernprojekt üblicherweise die Software bereits zentral (per Vorgabe der Konzern-IT) ausgewählt wurde. Die Erlaubnis zu erhalten, trotz teurer, bereits angeschaffter Lizenzen andere Tools für einen ähnlichen Einsatzzweck zu verwenden kann hier ein langwieriger Prozess mit ungewissem Ausgang sein, was letztlich dazu führen kann, dass sich das Forschungs- und das Praxisprojekt auseinanderentwickeln.

Rechtliche Beschränkungen: Die oft sensiblen und geschäftskritischen Daten, die im Rahmen der Entwicklung eines EUS mit Komponenten und Methoden aus dem Gebiet des Data Mining, Big Data Analytics etc. verarbeitet werden, stehen üblicherweise unter strenger Geheimhaltung. Die Wissenschaft hingegen zeichnet sich jedoch dadurch aus, ihre Ergebnisse und Erkenntnisse zu publizieren und somit offenzulegen, damit eine „kritische Überprüfung durch Dritte“ stattfinden und ein wissenschaftlicher Diskurs entstehen kann. An dieser Stelle ist ein Konflikt vorprogrammiert, denn die detaillierten Ergebnisse eines Kooperationsprojektes wie dem

unseren (vgl. Kapitel 3) sind letztlich Eigentum des Konzerns und nicht für die Öffentlichkeit bestimmt. Auch hier kann durch eine entsprechende Abstraktion der Forschungsergebnisse (ohne sie zu verfälschen) und die enge Abstimmung aller Projektteilnehmer eine Lösung gefunden werden, die für beide „Parteien“ akzeptabel ist.

Gegenseitige Verständnisschwierigkeiten: In unserer BISE-Veröffentlichung (vgl. Gleue et al., 2018) weisen wir bereits darauf hin, dass einer der großen Nachteile des im Rahmen unseres Projektes entwickelten EUS dessen (zwangsläufig) hohe Komplexität ist. Entscheidungsträger, insbesondere mit geringem technischem, mathematischem bzw. statistischem Vorwissen, haben i.d.R. Schwierigkeiten, im Detail nachzuvollziehen, auf welchem Wege die Zahlen entstehen, auf deren Basis sie teils schwerwiegende und weitreichende Entscheidungen treffen sollen. Den gesamten technischen Prozess nachzuvollziehen und im Detail zu verstehen kostet viel Zeit – ein knapper Faktor im Tagesgeschäft eines Managers in einem Automobilkonzern. Auf der anderen Seite ist das tiefgreifende fachliche Expertenwissen eben dieser Manager oder auch Projektmitarbeiter ein wichtiger Baustein, um letztlich ein System zu entwickeln, das in der Praxis den maximal möglichen Nutzen entfaltet. Das von uns entwickelte EUS musste sich, um letztlich erfolgreich zu sein und von den zukünftigen Nutzern „akzeptiert“ zu werden, nahtlos in die bestehenden Geschäftsprozesse eingliedern lassen und Ergebnisse in einem Format liefern, das im Konzern bereits bekannt ist. Um bereits während der Entwicklung den Nutzen des Projektes monetär bewerten, weiteres Projektbudget akquirieren und damit den weiteren Projektfortschritt sichern zu können, mussten regelmäßig Benchmarks zu den Zahlen aus der „alten Welt“ durchgeführt werden. Damit dies überhaupt möglich ist, müssen die KPIs des neuen EUS zu den altbewährten und konzernweit bekannten Kennzahlen passen, unabhängig davon, ob dies aus wissenschaftlicher Sicht sinnvoll erscheint oder nicht. Das dafür notwendige Fach- und Erfahrungswissen findet sich im Konzern bei Sachbearbeitern und Experten, fehlt jedoch (insbesondere externen) Softwareentwicklern und Beratern. Es gilt also, diese „Wissenslücke“ durch zielgerichtete, lösungsorientierte Kommunikation und professionelles Projekt- und Anforderungsmanagement zu schließen. Diese Aufgabe entspricht genau dem Profil des Wirtschaftsinformatikers in seiner Querschnittsfunktion als Schnittstelle bzw. „Verbindungsperson“ zwischen den technischen und den fachlichen Experten sowie dem Management. Dabei ist sowohl eine bedarfsgerechte Kommunikation von Projektergebnissen von hoher Bedeutung, wie auch die korrekte Aufnahme und „Übersetzung“ der fachlichen Anforderungen in für die Softwareentwickler, Datenanalysten und Forscher verständliche Spezifikationen. Diese Funktion und Fähigkeit des Wirtschaftsinformatikers, die bereits während der Ausbildung durch die Interdisziplinarität dieses Fachgebietes entsteht, ist folglich

oft die Rolle, die in IT-Projekten über Erfolg oder Misserfolg eines solchen Kooperationsprojektes entscheidet.

Hier liegt für mich persönlich der Haupterfolg beider in dieser kumulativen Dissertation vorgestellten Projekte. Während viele Promotionsprojekte und wissenschaftliche Arbeiten eher theoretischer Natur sind, wollte ich schon zu Beginn meiner Zeit als Promotionsstudent an etwas arbeiten, das in der Praxis einen sichtbaren Nutzen entfaltet, etwas Bestehendes verbessert. Das ist einer der Gründe für die Durchführung des in Kapitel 2 vorgestellten Ansatzes zur Optimierung von Literaturrecherchemethoden: Das Endergebnis, „TSISQ“, zielt darauf ab, Wissenschaftler unabhängig von ihrer Forschungsdisziplin bei ihrer täglichen Arbeit zu unterstützen und einen der zeitintensivsten, wiederkehrenden Prozesse in der Forschung mithilfe von datengetriebenen Technologien zu optimieren und so wertvolle Forschungszeit einzusparen. Im Grunde trifft diese Grundhaltung auch für das in Kapitel 3 vorgestellte Kooperationsprojekt mit der Automobilindustrie zu: Wir haben von Grund auf ein neuartiges EUS entwickelt, das sich heute im täglichen, operativen Betrieb befindet und durch seine zahlreichen, tagesaktuellen und dynamischen Berichte der Finanzabteilung des Herstellers die tägliche Arbeit und regelmäßig zu treffende Entscheidungen massiv erleichtert. Dieses Projekt mit einem wissenschaftlichen Anspruch innerhalb des zuvor beschriebenen Spannungsfeldes zwischen Wissenschaft und Wirtschaft durchzuführen war für uns alle ein Spagat, den wir, unter Beachtung der hier gegebenen Ratschläge, erfolgreich meistern konnten.

Davenport und Patil (2012) beschreiben in ihrem Artikel „Data Scientist: The Sexiest Job of the 21st Century“ das Profil desselben als einen Querschnitt aus Kenntnissen in den Bereichen der Mathematik, Statistik, Softwareentwicklung und Wirtschaftsinformatik sowie zusätzlich einem Bezug zur jeweiligen fachlichen Domäne. Diese hoch spezialisierte Mischung ist am Arbeitsmarkt sehr knapp, in der Realität finden sich diese Fähigkeiten und Kenntnisse selten in einer Person vereint wieder. Eine mögliche Lösung dafür ist die Bildung interdisziplinärer Teams mit den oben genannten individuellen Fähigkeiten. Agarwal und Dhar (2014) beschreiben es in einem Editorial der Zeitschrift *Information Systems Research* mit dem Titel „Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research“ als Wettbewerbsvorteil der Wirtschaftsinformatik, dass sie seit Jahrzehnten Forschung in der Schnittmenge zwischen „[...]computing technology and data in business and society“ betreibt. Die Autoren stellen fest, dass Wirtschaftsinformatiker möglicherweise die umfassendste Sichtweise auf Unternehmen als Ganzes besitzen, und dass Forschung in diesem Gebiet wahlweise als interdisziplinär gelobt oder als „zu interdisziplinär“ kritisiert wird. Ich möchte mich ihrer Sichtweise anschließen und genau diese Interdisziplinarität des

Wirtschaftsinformatikers als Stärke und Fähigkeit verstehen, in der heutigen, spannenden Zeit in entsprechend zusammengesetzten Teams das Bindeglied zu sein, das letztlich durch die oben genannten Eigenschaften Projekte zum Erfolg führt.

Die Erkenntnis, dass nur eine Kombination aus menschlichem Erfahrungswissen innerhalb eines Fachgebiets bei gleichzeitiger, optimaler Nutzung und Auswertung aller verfügbaren Daten mit Methoden des Data Mining und Big Data Analytics letztlich maximalen Erfolg verspricht, und dass die Kommunikation zwischen den eher technisch, mathematisch und statistisch ausgebildeten Projektmitarbeitern und den erfahrenen Fachexperten letztlich der Schlüssel dazu ist, wurde in reduzierter Form auch bereits im Ausblick unserer BISE-Publikation (Gleue et al., 2018) genannt. In unserer Veröffentlichung „It's not a Bug, it's a Feature: How Visual Model Evaluation can help to incorporate Human Domain Knowledge in Data Science“ (Eilers et al., 2017) gehen wir noch einen Schritt weiter: Aus den bereits genannten Gründen ist der Einbezug fachlichen Expertenwissens nicht erst bei der Vermittlung von ersten Zwischenergebnissen eines Projektes, sondern bereits bei der Datenmodellierung bzw. dem „Feature Engineering“ für Anwendungen des maschinellen Lernens unabdingbar. Folglich bedarf es nicht erst am Ende, sondern auch bereits *während* des Entwicklungsprozesses geeigneter Visualisierungen, die helfen, ein gemeinsames Verständnis zwischen „Datenwissenschaftlern“ und Fachexperten herzustellen. Die Verwendung von Heatmaps für diese Aufgabe wurde dabei unter anderem gewählt, da wir sie als „gewohnte“ Darstellungsform für das Management in unserem Kooperationsprojekt mit der Automobilindustrie mit großem Erfolg anwenden konnten, um die Zwischen- und Endergebnisse unserer Arbeit verständlich zu kommunizieren. Durch die noch frühere Eingliederung der Auftraggeber aus dem Management und der Fachexperten in den Entwicklungsprozess können Missverständnisse früher aufgedeckt und beseitigt werden. Zudem entsteht so gleichzeitig eine Identifikation mit dem (Software-) Produkt, was zu einer gesteigerten Technologieakzeptanz des Fachbereiches und größerem Vertrauen in die Projektergebnisse führt. Für einen tieferen Einblick in dieses Thema verweise ich auf unsere entsprechende ICIS-Publikation (Eilers et al., 2017).

Abschließend möchte ich an dieser Stelle dennoch selbstkritisch hervorheben, dass blindes Vertrauen in die Technologie auch Risiken mit sich bringt. Aufgrund des aktuellen Hypes um Data Science, Maschinelles Lernen, Künstliche Intelligenz und alle damit verwandten Themengebiete werden auch in der Forschung hauptsächlich Fragen adressiert, die auf die Steigerung der Technologieakzeptanz und die Einführung derartiger Technologien in nahezu allen Branchen abzielen. Hingegen steht die Gefahr, dass moderne EUS wie das von mir in

dieser Arbeit vorgestellte, risikofreudiges Verhalten unterstützen könnten, eher selten im Fokus aktueller Forschung (vgl. Chen and Koufaris, 2015). In der Tat wurde das hier vorgestellte Projekt u.a. auch mit der Absicht durchgeführt, bisherige, umfangreiche „Sicherheitspuffer“ im Risikomanagement zu reduzieren, bei gleichzeitiger Gewinnmaximierung und Kosteneinsparung. Dies kann sicherlich bis zu einem gewissen Grad gelingen, denn der hier vorgestellte, datengetriebene Ansatz bringt einen enormen, tagesaktuellen Informationsgewinn mit sich, der zuvor so schlicht nicht denkbar gewesen wäre. Dennoch ist es bei der Entscheidungsfindung wichtig, die von „der Maschine“ erstellten Kennzahlen und Prognosen nicht stärker zu gewichten, als das über Jahrzehnte aufgebaute Erfahrungswissen in den Fachabteilungen.

Da es sich bei der datengetriebenen Entscheidungsunterstützung um ein stetig an Bedeutung gewinnendes Fachgebiet handelt, wird auch in Zukunft die Anzahl der oben beschriebenen Kooperationsprojekte entsprechend ansteigen. Die steigende Verfügbarkeit und Menge an Daten bringt automatisch einen steigenden Bedarf an Analysen und Analysten mit sich (vgl. Agarwal und Dhar, 2014). Nur die richtige Kombination aus optimaler Nutzung der im Konzern vorhandenen Daten und Informationen durch ein entsprechendes EUS inkl. bedarfsgerechtem Reporting und menschlichem Erfahrungs- und Expertenwissen kann jedoch kurz-, mittel- und langfristig zu maximalem Erfolg führen. Methoden des maschinellen Lernens und Künstliche Intelligenz sollten daher nicht als Substitut für menschliches Wissen und die menschlichen Fähigkeiten betrachtet, sondern als Komplement verstanden werden. Zukünftige Forschung im interdisziplinären Gebiet der Wirtschaftsinformatik sollte sich folglich darauf konzentrieren, Ideen zu entwickeln, wie menschliches Expertenwissen optimal mit den o.g. Methoden kombiniert und integriert werden kann.

5 Fazit

In dieser kumulativen Dissertation wurden zwei Forschungsprojekte aus den Gebieten des Data Mining, Big Data Analytics und der KI vorgestellt und kritisch diskutiert. Der erste Hauptteil der vorliegenden Arbeit beschreibt die Entwicklung und Evaluation eines Werkzeugs zur „teilautomatisierten“ Literaturrecherche auf Basis semantischer Ähnlichkeiten zwischen (Voll-)Texten (vgl. Kapitel 2). Im zweiten Hauptteil wird unsere Forschung im Gebiet der datengetriebenen EUS vorgestellt, wobei im Kern die Fähigkeit, mit KNN Markt- und Restwerte von Leasingfahrzeugen für verschiedene Zeithorizonte zu prognostizieren, im Vordergrund steht (vgl. Kapitel 3). Das übergreifende Ziel über beide Projekte hinweg war es zum einen, Antworten auf die Frage zu finden, ob mit den oben genannten Methoden auch abseits von „Leuchtturmprojekten“, die während meiner Promotionszeit in diesem Gebiet in großer Zahl entstanden sind, althergebrachte Arbeits- und Entscheidungsprozesse durch die Anwendung von Verfahren aus den o.g. Forschungsgebieten weiter optimiert werden können. Diese Frage wurde anhand zweier Projekte, eines in einem rein wissenschaftlichen Umfeld, eines in der Wirtschaft, kritisch diskutiert und in beiden Fällen eine maßgebliche Verbesserung des bisherigen Prozesses erreicht. Zum anderen ist durch die Arbeit an den in dieser Dissertation vorgestellten Projekten die Erkenntnis gereift, dass sowohl die Kooperation zwischen Wissenschaft und Wirtschaft mit ihren unterschiedlichen Denkweisen und Geisteshaltungen als auch die Kombination der Stärken von „Menschen und Maschinen“ insbesondere in IT-Projekten mit komplexer Aufgabenstellung für beide Seiten große Vorteile mit sich bringen kann. Das maschinelle Lernen, Künstliche Intelligenz und die zahlreichen weiteren, hier nicht explizit adressierten Methoden aus dem weiten Feld der Data Science werden auch in Zukunft stetig wachsenden Einfluss auf unser tägliches Leben nehmen und sollten daher, wenn auch stets kritisch hinterfragt, als Komplement, nicht als Substitut verstanden werden. Die Frage, wie diese Kooperation und Kombination optimal ausgestaltet werden kann, und wie die verschiedenen Stärken von Fachexperten mit Erfahrungswissen in Ihrer Domäne, Softwareentwicklern, BI-Experten, „Datenwissenschaftlern“ und der entsprechenden Technologien optimal miteinander verzahnt werden können, sollte Gegenstand zukünftiger Forschung in der Disziplin der Wirtschaftsinformatik sein.

6 Literaturverzeichnis

Abate, F., Ficarra, E., Acquaviva, A., Macii, E., 2011. Improving latent semantic analysis of biomedical literature integrating UMLS Metathesaurus and biomedical pathways databases, in: International Joint Conference on Biomedical Engineering Systems and Technologies. Springer, pp. 173-187.

Adya, M., Collopy, F., 1998. How effective are neural networks at forecasting and prediction? A review and evaluation. *Journal of Forecasting* 17, pp. 481-495.

Agarwal, R., Dhar, V., 2014. Editorial – Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. *Information Systems Research* 25, pp. 443-448.

AKA, 2016. Automobilbanken-Studie 2016. AKA - Arbeitskreis Autobanken. URL: <https://www.autobanken.de/publikationen> (letzter Aufruf 22.09.18).

Arijit, D., 2013. SMS Based FAQ Retrieval Using Latent Semantic Indexing, in: *Multilingual Information Access in South Asian Languages*. Springer, pp. 100-103.

Baker, M.J., 2000. Writing a Literature Review. *Marketing Review* 1, pp. 219-247.

Bandara, W., Miskon, S., Fielt, E., 2011. A systematic, tool-supported method for conducting literature reviews in information systems, in: *Proceedings of The 19th European Conference on Information Systems (ECIS 2011)*.

Bawden, D., Robinson, L., 2009. The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of information science* 35, pp. 180-191.

BDA, 2017. Mobilitätsdienstleistungen - vom reinen Finanz- zum umfassenden Mobilitätsangebot. BDA - Banken der Automobilwirtschaft.

URL: <https://www.autobanken.de/publikationen> (letzter Aufruf 27.09.18).

Bhandari, H., Shimbo, M., Ito, T., Matsumoto, Y., 2008. Generic text summarization using probabilistic latent semantic indexing, in: *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume I*.

Bishop, C.M., 1995. *Neural networks for pattern recognition*. Oxford university press.

Blair, D.C., Maron, M.E., 1985. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM* 28, pp. 289-299.

- Blake, R., 2010. Identifying the core topics and themes of data and information quality research., in: Americas Conference on Information Systems, paper 221.
- Bollen, K.A., 2011. Evaluating effect, composite, and causal indicators in structural equation models. *Management Information Systems Quarterly* 35, pp. 359-372.
- Brand, M., 2006. Fast low-rank modifications of the thin singular value decomposition. *Linear algebra and its applications* 415, pp. 20-30.
- Breiman, L., 2001. Random Forests. *Machine Learning* 45, pp. 5-32.
- Cao, T.H., Ngo, V.M., 2012. Semantic search by latent ontological features. *New Generation Computing* 30, pp. 53-71.
- Chaudhuri, S., Dayal, U., Narasayya, V., 2011. An overview of business intelligence technology. *Communications of the ACM* 54, pp. 88-98.
- Chen, C.-W., Koufaris, M., 2015. The impact of decision support system features on user overconfidence and risky behavior. *European Journal of Information Systems* 24, pp. 607-623.
- Chen, H., Chiang, R.H., Storey, V.C., 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS quarterly* 36, pp. 1165-1188.
- Cleveland, R.B., Cleveland, W.S., McRae, J.E., Terpenning, I., 1990. STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics* 6, pp. 3-73.
- Corley, C., Mihalcea, R., 2005. Measuring the semantic similarity of texts, in: *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*. Association for Computational Linguistics, pp. 13-18.
- Cui, H., Wen, J.-R., Nie, J.-Y., Ma, W.-Y., 2003. Query expansion by mining user logs. *IEEE Transactions on Knowledge & Data Engineering* 15, pp. 829–839.
- DAT, 2016. DAT-Report 2016, Deutsche Automobil Treuhand. URL: <https://www.dat.de/presse/dat-report/> (letzter Aufruf 19.09.18).
- Davenport, T.H., 2006. Competing on analytics. *Harvard Business Review* 84, p. 98.
- Davenport, T.H., Patil, D., 2012. Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review* 90, pp. 70-76.

- Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* 41, pp. 391-407.
- Ding, C.H., 1999. A similarity-based probability model for latent semantic indexing, in: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 58-65.
- Domingos, P., 2012. *A Few Useful Things to Know About Machine Learning*. *Communications of the ACM* 55, pp. 78-87.
- Dress, K., Lessmann, S., von Mettenheim, H.-J., 2017. Residual Value Forecasting Using Asymmetric Cost Functions. arXiv:1707.02736 [q-fin, stat].
- Dumais, S.T., 1994. Latent semantic indexing (LSI) and TREC-2. *Nist Special Publication Sp* p. 105.
- Dumais, S.T., 1993. LSI meets TREC: A status report, in: *Proceedings of the First Text REtrieval Conference, TREC1*. pp. 137-152.
- Dumais, S.T., Furnas, G.W., Landauer, T.K., Deerwester, S., Harshman, R., 1988. Using latent semantic analysis to improve access to textual information, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp. 281-285.
- Eilers, D., Dunis, C.L., von Mettenheim, H.-J., Breitner, M.H., 2014. Intelligent trading of seasonal effects: A decision support algorithm based on reinforcement learning. *Decision support systems* 64, pp. 100-108.
- Eilers, D., Köpp, C., Gleue, C., Breitner, M., 2017. It's not a Bug, it's a Feature: How Visual Model Evaluation can help to incorporate Human Domain Knowledge in Data Science. *ICIS 2017 Proceedings*.
- Fan, H., AbouRizk Simaan, Kim Hyoungkwan, Zaïane Osmar, 2008. Assessing Residual Value of Heavy Construction Equipment Using Predictive Data Mining Model. *Journal of Computing in Civil Engineering* 22, pp. 181-191.
- Farrús, M., Costa-jussà, M.R., 2013. Automatic evaluation for e-learning using latent semantic analysis: A use case. *The International Review of Research in Open and Distributed Learning* 14, pp. 239-254.

- Fink, A., 2013. Conducting research literature reviews: From the Internet to paper. Sage Publications.
- Foltz, P.W., Dumais, S.T., 1992. Personalized information delivery: An analysis of information filtering methods. *Communications of the ACM* 35, pp. 51-60.
- Forsythe, G.E., Malcolm, M.A., Moler, C.B., 1979. *Computer Methods for Mathematical Computations*. Englewood Cliffs, New Jersey 07632. Prentice Hall, Inc., 1977. XI, 259 S.
- ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik 59, pp. 141-142.
- Furlan, B., Batanović, V., Nikolić, B., 2013. Semantic similarity of short texts in languages with a deficient natural language processing support. *Decision Support Systems* 55, pp. 710-719.
- Gabrilovich, E., Markovitch, S., 2009. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research* 34, pp. 443-498.
- Gansterer, W.N., Janecek, A.G., Neumayer, R., 2008. Spam filtering based on latent semantic indexing, in: *Survey of Text Mining II*. Springer, pp. 165-183.
- Gee, K.R., 2003. Using latent semantic indexing to filter spam, in: *Proceedings of the 2003 ACM Symposium on Applied Computing*. ACM, pp. 460-464.
- Gefen, D., Karahanna, E., Straub, D.W., 2003. Trust and TAM in online shopping: An integrated model. *MIS quarterly* 27, pp. 51-90.
- Gleue, C., Eilers, D., Mettenheim, H.-J. von, Breitner, M., 2017. Decision Support for the Automotive Industry: Forecasting Residual Values using Artificial Neural Networks. *Wirtschaftsinformatik 2017 Proceedings*.
- Gleue, C., Eilers, D., Mettenheim, H.-J. von, Breitner, M.H., 2018. Decision Support for the Automotive Industry: Forecasting Residual Values using Artificial Neural Networks. *Business & Information Systems Engineering (online vorveröffentlicht)*, pp. 1-13.
- Go, A., Bhayani, R., Huang, L., 2010. Exploiting the unique characteristics of tweets for sentiment analysis. Technical report, Technical Report, Stanford University.
- Gong, Y., Liu, X., 2001. Generic text summarization using relevance measure and latent semantic analysis, in: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 19-25.

- Gordon, M.D., Dumais, S., 1998. Using latent semantic indexing for literature based discovery. *Journal of the American Society for Information Science* 49, pp. 674-685.
- Hart, C. 1998. *Doing a literature review: Releasing the social science research imagination*. London: Sage
- Hevner, A.R., 2007. A three cycle view of design science research. *Scandinavian journal of information systems* 19, p. 4.
- Hevner, A.R., March, S.T., Park, J., Ram, S., 2004. Design Science in Information Systems Research. *MIS Quarterly* 28, pp. 75-105.
- Hofmann, T., 1999. Probabilistic latent semantic analysis, in: *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., pp. 289-296.
- Homayouni, R., Heinrich, K., Wei, L., Berry, M.W., 2004. Gene clustering by latent semantic indexing of MEDLINE abstracts. *Bioinformatics* 21, pp. 104-115.
- Hovorka, D.S., Larsen, K.R., Monarchi, D.E., 2009. Conceptual convergences: Positioning information systems among the business disciplines., in: *ECIS*. pp. 800-812.
- Ionescu, L., Gwiggner, C., Kliewer, N., 2016. Data Analysis of Delays in Airline Networks. *Business & Information Systems Engineering* 58, pp. 119-133.
- Kintsch, W., 2010. Modeling Semantic Memory in Mobile Ad-hoc NETWORKS. in *Mobile Ad-hoc NETWORKS (MANETS)* S. JheanLarose and G. Denhière (eds.).
- Kontostathis, A., 2007. Essential dimensions of latent semantic indexing (lsi), in: *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference On*. IEEE, pp. 73-73.
- Kontostathis, A., Pottenger, W.M., 2006. A framework for understanding Latent Semantic Indexing (LSI) performance. *Information Processing & Management* 42, pp. 56-73.
- Kontostathis, A., Pottenger, W.M., 2002. A mathematical view of latent semantic indexing: Tracing term co-occurrences. Bethlehem (USA): LeHigh Technical Reports.
- Korenius, T., Laurikkala, J., Juhola, M., 2007. On principal component analysis, cosine and Euclidean measures in information retrieval. *Information Sciences* 177, pp. 4893-4905.

- Koukal, A., Gleue, C., Breitner, M., 2014a. Enhancing Literature Review Methods - towards more efficient Literature Research with Latent Semantic Indexing. ECIS 2014 Proceedings.
- Koukal, A., Gleue, C., Breitner, M., 2014b. Enhancing Literature Review Methods - Evaluation of a Literature Search Approach based on Latent Semantic Indexing. ICIS 2014 Proceedings.
- Koukal, A., Gleue, C., Breitner, M.H., 2013. Enhancing Literature Research Processes: A Glance at an Approach Based on Latent Semantic Indexing., in: GI-Jahrestagung. pp. 1937-1942.
- Kuechler, W.L., 2007. Business applications of unstructured text. Communications of the ACM 50, pp. 86-93.
- Kuo, R.J., Chen, C.H., Hwang, Y.C., 2001. An intelligent stock trading decision support system through integration of genetic algorithm based fuzzy neural network and artificial neural network. Fuzzy Sets and Systems 118, pp. 21-45.
- LaBrie, R., St Louis, R., 2003. Information retrieval from knowledge management systems: Using knowledge hierarchies to overcome keyword limitations. AMCIS 2003 Proceedings pp. 2552-2563.
- Landauer, T.K., Dumais, S.T., 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review 104, pp. 211-240.
- Leaseurope - 2016 Leasing Facts & Figures [WWW Document], 2018. . Leaseurope - 2016 Leasing Facts & Figures. URL <http://www.leaseurope.org/index.php?page=key-facts-figures> (letzter Aufruf 28.09.18).
- Lebek, B., Uffen, J., Breitner, M.H., Neumann, M., Hohler, B., 2013. Employees' Information Security Awareness and Behavior: A Literature Review, in: 2013 46th Hawaii International Conference on System Sciences(HICSS). pp. 2978-2987.
- Lee, S., Baker, J., Song, J., Wetherbe, J.C., 2010. An Empirical Comparison of Four Text Mining Methods. 2010 43rd Hawaii International Conference on System Sciences pp. 1-10.
- Lessmann, S., 2013. Modelling Mismatch In Predictive Analytics: A Case Study Illustration And Possible Remedy. ECIS 2013 Completed Research.
- Lessmann, S., Listiani, M., Voß, S., 2010. Decision support in car leasing: A forecasting model for residual value estimation. ICIS 2010 Proceedings.

- Levy, Y., Ellis, T.J., 2006. A systems approach to conduct an effective literature review in support of information systems research. *Informing Science* 9, pp. 181-212.
- Lian, C., Zhao, D., Cheng, J., 2003. A fuzzy logic based evolutionary neural network for automotive residual value forecast, in: *International Conference on Information Technology: Research and Education*, pp. 545-548.
- Liu, Z., Natarajan, S., Chen, Y., 2011. Query expansion based on clustered results. *Proceedings of the VLDB Endowment* 4, pp. 350-361.
- Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C., 2011. Learning word vectors for sentiment analysis, in: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pp. 142-150.
- Mabe, M., Amin, M., 2001. Growth dynamics of scholarly and scientific journals. *Scientometrics* 51, pp. 147-162.
- Manwani, S., Bech, H., Dahlhoff, J., 2001. Managing Information Overload: Is Technology the Answer? *AMCIS 2001 Proceedings*, pp. 34-38.
- March, S.T., Smith, G.F., 1995. Design and natural science research on information technology. *Decision support systems* 15, pp. 251-266.
- Mettenheim, H.-J., Breitner, M., 2010. Robust Decision Support Systems with Matrix Forecasts and Shared Layer Perceptrons for Finance and other Applications. *ICIS 2010 Proceedings*.
- Mitra, M., Singhal, A., Buckley, C., 1998. Improving automatic query expansion, in: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 206-214.
- Nelson, M., Hill, T., Remus, W., O'Connor, M., 1999. Time series forecasting using neural networks: should the data be deseasonalized first? *Journal of Forecasting* 18, pp. 359-367.
- Nugumanova, A., Bessmertny, I., 2013. Applying the latent semantic analysis to the issue of automatic extraction of collocations from the domain texts, in: *International Conference on Knowledge Engineering and the Semantic Web*. Springer, pp. 92-101.
- Offermann, P., Levina, O., Schönherr, M., Bub, U., 2009. Outline of a design science research process, in: *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology*. ACM, paper 7.

- Okoli, C., Schabram, K., 2010. A guide to conducting a systematic literature review of information systems research.
- Palvia, P., Midha, V., Pinjani, P., 2006. Research models in information systems. *Communications of the Association for Information Systems* 17, pp. 1042-1063.
- Park, J., Lee, J.-N., 2011. The Impact of Information Overload on Decision Quality in the Web 2.0 Environment: A Cognitive-Emotional Dichotomy Perspective, in: *Proceedings of the 2011 International Conference on Information Resources Management (CONF-IRM 2011)*.
- Peppers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S., 2007. A design science research methodology for information systems research. *Journal of management information systems* 24, pp. 45-77.
- Prado, S., 2009. The European used-car market at a glance: Hedonic resale price valuation in automotive leasing industry. *Economics Bulletin* 29, pp. 2086-2099.
- Prado, S., Ananth, R., 2012. Breaking Through Risk Management, a Derivative for the Leasing Industry. *Journal of Financial Transformation* 34, pp. 211-218.
- Qiu, Y., Frei, H.-P., 1993. Concept based query expansion, in: *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 160-169.
- Rehurek, R., Sojka, P., 2010. Software framework for topic modelling with large corpora, in: *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.
- Rickenberg, T., Neumann, M., Hohler, B., Breitner, M., 2012. Enterprise Content Management - A Literature Review. *AMCIS 2012 Proceedings*.
- Rode, D.C., Fischbeck, P.S., Dean, S.R., 2002. Residual Risk and the Valuation of Leases Under Uncertainty and Limited Information. *The Journal of Structured Finance* 7, pp. 37-49.
- Rousseau, D.M., Manning, J., Denyer, D., 2008. Evidence in Management and Organizational Science: Assembling the Field's Full Weight of Scientific Knowledge Through Syntheses. *The Academy of Management Annals* 2, pp. 475-515.
- Salton, G., McGill, M.J., 1986. *Introduction to modern information retrieval*, New York, NY: McGraw-Hill.

Salton, G., Wong, A., Yang, C.-S., 1975. A vector space model for automatic indexing. *Communications of the ACM* 18, pp. 613-620.

Santos, J.C.A., Ribeiro, M.F.D.S.B., 2011. Improving search engine Query Expansion techniques with ILP in, in: *Proceedings of the 21st International Conference on Inductive Logic Programming (ILP)*.

Schocken, S., Ariav, G., 1994. Neural networks for decision support:: Problems and opportunities. *Decision Support Systems* 11, pp. 393-414.

Sermpinis, G., Stasinakis, C., Dunis, C., 2014. Stochastic and genetic neural network combinations in trading and hybrid time-varying leverage effects. *Journal of International Financial Markets, Institutions and Money* 30, pp. 21-54.

Shao, J., Wu, W., Geng, P., 2013. An improved approach to the recovery of traceability links between requirement documents and source codes based on latent semantic indexing, in: *International Conference on Computational Science and Its Applications (ICCSA 2013)*. Springer, pp. 547-557.

Shen, D., Chen, Z., Yang, Q., Zeng, H.-J., Zhang, B., Lu, Y., Ma, W.-Y., 2004. Web-page classification through summarization, in: *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 242-249.

Sidorova, A., Evangelopoulos, N., Valacich, J.S., Ramakrishnan, T., 2008. Uncovering the intellectual core of the information systems discipline. *Management Information Systems Quarterly* 467-482.

Smith, L.D., Jin, B., 2007. Modeling exposure to losses on automobile leases. *Review of Quantitative Finance and Accounting* 29, pp. 241-266.

Steinberger, J., Jezek, K., 2004. Using latent semantic analysis in text summarization and summary evaluation. *Proceedings of ISIM'04* 4, pp. 93-100.

Storchmann, K., 2004. On the Depreciation of Automobiles: An International Comparison. *Transportation* 31, pp. 371-408.

Vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., Cleven, A., 2009. Reconstructing the giant: On the importance of rigour in documenting the literature search process., in: *ECIS 2009 Proceedings*. pp. 2206-2217.

- von Spreckelsen, C., von Mettenheim, H.-J., Breitner, M.H., 2014. Real-Time Pricing and Hedging of Options on Currency Futures with Artificial Neural Networks. *J. Forecast.* 33, pp. 419-432.
- Voorhees, E.M., 1994. Query expansion using lexical-semantic relations, in: *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Springer-Verlag New York, Inc., pp. 61-69.
- Webster, J., Watson, R.T., 2002. Analyzing the past to prepare for the future: Writing a literature review. *Management Information Systems Quarterly* xiii–xxiii.
- Wolfe, M.B., Goldman, S.R., 2003. Use of latent semantic analysis for predicting psychological phenomena: Two issues and proposed solutions. *Behavior Research Methods, Instruments, & Computers* 35, pp. 22-31.
- Wolfe, M.B., Schreiner, M.E., Rehder, B., Laham, D., Foltz, P.W., Kintsch, W., Landauer, T.K., 1998. Learning from text: Matching readers and texts by latent semantic analysis. *Discourse processes* 25, pp. 309-336.
- Wolfswinkel, J.F., Furtmueller, E., Wilderom, C.P., 2013. Using grounded theory as a method for rigorously reviewing literature. *European journal of information systems* 22, pp. 45-55.
- Wu, J.-D., Hsu, C.-C., Chen, H.-C., 2009. An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. *Expert Systems with Applications* 36, pp. 7809-7817.
- Xu, J., Croft, W.B., 1996. Query expansion using local and global document analysis, in: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 4-11.
- Yandell, M.D., Majoros, W.H., 2002. Genomics and natural language processing. *Nature Reviews Genetics* 3, pp. 601-610.
- Yeh, J.-Y., Ke, H.-R., Yang, W.-P., Meng, I.-H., 2005. Text summarization using a trainable summarizer and latent semantic analysis. *Information processing & management* 41, pp. 75-95.
- Zelikovitz, S., Hirsh, H., 2001. Using LSI for text classification in the presence of background text, in: *Proceedings of the Tenth International Conference on Information and Knowledge Management*. ACM, pp. 113-118.
- Zha, H., Simon, H., 1998. A Subspace-based Model for Latent Semantic Indexing in Information Retrieval. *Proceedings of the 13th Symposium on the Interface* pp. 315-320.

Zhang, G., Eddy Patuwo, B., Y. Hu, M., 1998. Forecasting with artificial neural networks:: The state of the art. *International Journal of Forecasting* 14, pp. 35-62.

Zhang, W., Yoshida, T., Tang, X., 2011. A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert Systems with Applications* 38, pp. 2758-2765.

Zimmermann, H.G., Neuneier, R., Grothmann, R., 2001. Multi-agent modeling of multiple FX-markets by neural networks. *IEEE Transactions on Neural Networks* 12, pp. 735-743.

7 Appendix

A: Enhancing Literature Research Processes: A Glance at an Approach Based on Latent Semantic Indexing

Zitat: Koukal, A., Gleue, C., & Breitner, M.H. (2013). Enhancing Literature Research Processes: A Glance at an Approach Based on Latent Semantic Indexing. In: Horbach, Matthias (Ed.): Informatik 2013, 43. Jahrestagung der Gesellschaft für Informatik. Lecture Notes in Informatics, Vol. 220, pp. 1937-1943. ISBN 978-3-88579-614-5

Link: <https://dl.gi.de/bitstream/handle/20.500.12116/20626/1937.pdf>

Abstract: Literature search as a fundamental, complex and time-consuming step in a literature research process is part of many established scientific methods. It is still predominantly supported by search techniques based on conventional termmatching methods. We address the lack of semantic approaches in this context by proposing an enhancement of the literature research process with a prototype of our Tool for Semantic Indexing and Similarity Queries (TSISQ), which is based on latent semantic indexing (LSI). Its applicability is evaluated in two cases. Results indicate that our approach can help to save valuable time discovering relevant literature in a desired research field or to increase the comprehensiveness of a review by identifying sources that otherwise would not have been considered. The target audience for our findings includes researchers who need to efficiently gain an overview of a research field, deepen their knowledge and refine the theoretical foundations of their research.

B: Enhancing Literature Review Methods - towards more efficient Literature Research with Latent Semantic Indexing

Zitat: Koukal, A., Gleue, C., & Breitner, M.H. (2014). Enhancing Literature Review Methods - towards more efficient Literature Research with Latent Semantic Indexing. European Conference on Information Systems (ECIS) 2014 Proceedings.

Link: <https://aisel.aisnet.org/ecis2014/proceedings/track19/1/>

Abstract: Nowadays, the facilitated access to increasing amounts of information and scientific resources means that more and more effort is required to conduct comprehensive literature reviews. Literature search, as a fundamental, complex, and time-consuming step in every literature research process, is part of many established scientific methods. However, it is still predominantly supported by search techniques based on conventional term-matching methods. We address the lack of semantic approaches in this context by proposing an enhancement of established literature review methods. For this purpose, we followed design science research (DSR) principles in order to develop artifacts and implement a prototype of our Tool for Semantic Indexing and Similarity Queries (TSISQ) based on the core concepts of latent semantic indexing (LSI). Its applicability is demonstrated and evaluated in a case study. Results indicate that the presented approach can help save valuable time in finding basic literature in a desired research field or increasing the comprehensiveness of a review by efficiently identifying sources that otherwise would not have been taken into account. The target audience for our findings includes researchers who need to efficiently gain an overview of a specific research field, deepen their knowledge or refine the theoretical foundations of their research.

C: Enhancing Literature Review Methods - Evaluation of a Literature Search Approach based on Latent Semantic Indexing

Zitat: Koukal, A., Gleue, C., & Breitner, M.H. (2014). Enhancing Literature Review Methods - Evaluation of a Literature Search Approach based on Latent Semantic Indexing. International Conference on Information Systems (ICIS) 2014 Proceedings.

Link: <https://aisel.aisnet.org/icis2014/proceedings/ResearchMethods/9/>

Abstract: Literature search, as a fundamental and time-consuming step in a literature research process, is part of many established scientific research methods. The facilitated access to scientific resources requires an increasing effort to conduct comprehensive literature reviews. We address the lack of semantic approaches in this context by proposing and evaluating our Tool for Semantic Indexing and Similarity Queries (TSISQ) for the enhancement of established literature review methods. Its applicability is evaluated in different environments and search cases covering realistic applications. Results indicate that TSISQ can increase efficiency by saving valuable time in finding relevant literature in a desired research field, improve the quality of search results, and enhance the comprehensiveness of a review by identifying sources that otherwise would not have been considered. The target audience includes all researchers who need to efficiently gain an overview of a specific research field and refine the theoretical foundations of their research.

D: Decision Support for the Automotive Industry: Forecasting Residual Values Using Artificial Neural Networks (WI)

Zitat: Gleue, C., Eilers, D., Mettenheim, H.-J. von, Breitner, M., 2017. Decision Support for the Automotive Industry: Forecasting Residual Values using Artificial Neural Networks. Wirtschaftsinformatik (WI) 2017 Proceedings.

Link: <http://aisel.aisnet.org/wi2017/track12/paper/1/>

Abstract: The leasing business is one of the most important distribution channels for the automotive industry. This implies that forecasting accurate residual values for the vehicles is a major factor for determining monthly leasing rates: Either a systematic overestimation or underestimation of future residual values can incur large potential losses in resale value or, respectively, competitive disadvantages. In this paper, an operative DSS with the purpose of facilitating residual value related management decisions is introduced, with a focus on its forecasting capabilities. Practical implications are discussed, a multi-variate linear model and an artificial neural network approach are benchmarked and further, the effects of price trends and seasonal influences are investigated. The analysis is based on more than 150,000 data sets from a major German car manufacturer. We show that artificial neural network ensembles with only a few input variables are capable of achieving a significant improvement in forecasting accuracy.

E: It's Not a Bug, It's a Feature: How Visual Model Evaluation Can Help to Incorporate Human Domain Knowledge in Data Science

Zitat: Eilers, D., Köpp, C., Gleue, C., Breitner, M., 2017. It's not a Bug, it's a Feature: How Visual Model Evaluation can help to incorporate Human Domain Knowledge in Data Science. International Conference on Information Systems (ICIS) 2017 Proceedings.

Link: <http://aisel.aisnet.org/icis2017/DataScience/Presentations/15/>

Abstract: The question of how to incorporate human domain knowledge in practical data science projects is still a major challenge. While machine learning tasks are usually carried out by technically skilled data scientists, these analysts do not necessarily have the required domain knowledge concerning a particular business problem to explain certain phenomena. In real-world data science applications, this may result in models that do not adequately reflect relationships in the data. We address this issue by introducing a heat map based technique for model error visualization to facilitate discussions of the results between data scientists and domain experts. By discussing model errors with domain experts during the iterative analysis process, the generated insights can be used for engineering new features (explanatory variables) which better represent the problem and therefore improve the results. We demonstrate the visualization approach based on artificial data and in the context of a real-world industry example.

F: Decision Support for the Automotive Industry: Forecasting Residual Values Using Artificial Neural Networks (BISE)

Zitat: Gleue, C., Eilers, D., Mettenheim, H.-J. von, Breitner, M., (2018). Decision Support for the Automotive Industry: Forecasting Residual Values using Artificial Neural Networks. Business & Information Systems Engineering (BISE).

Link: <https://link.springer.com/article/10.1007/s12599-018-0527-3>

DOI: <https://doi.org/10.1007/s12599-018-0527-3>

Abstract: In the automotive industry, it is very common for new vehicles to be leased rather than sold. This implies forecasting an accurate residual value for the vehicles, which is a major factor for determining monthly leasing rates. Either a systematic overestimation or underestimation of future residual values can incur large potential losses in resale value or, respectively, competitive disadvantages. In this paper an operative Decision Support System is introduced for the purpose of facilitating residual value related management decisions with emphasis on its forecasting capabilities. We demonstrate the use of Artificial Neural Networks for this application in a case study based on more than 250,000 data sets of completed leasing contracts between 2011 and 2017 from a major German car manufacturer. The importance of determining price factors and the effect of different time horizons on forecasting accuracy are investigated and practical implications are discussed. In addition, we neither found a significant explanatory nor predictive power of external economic factors, which underlines the importance of collecting and taking advantage of vehicle-specific data or, in more general terms, the exclusive data of corporations, which is often only available internally.