Proceedings des KI-Praktikums der Veranstaltung
Management Support Systems III
an der Universität Osnabrück

Text Mining in wissenschaftlichen Publikationen

Wintersemester 2013/2014

Herausgeber
Prof. Dr.-Ing. Bodo Rieger
Axel Benjamins

Inhaltsverzeichnis

Geleitwort

Eine stetig steigende Anzahl von Informationen ist in unstrukturierten Daten, z. B. Textdokumenten, enthalten. Diese sind schwieriger automatisiert zu verarbeiten als strukturierte Daten. Zunehmend gewinnen jedoch gerade diese Informationen in den unstrukturierten Daten an Bedeutung zur Bildung von entscheidenden Wettbewerbsvorteilen von Unternehmen. Bei der erforderlichen automatisierten Verarbeitung unstrukturierter Daten konnte die aufstrebende Forschungsdomäne rund um das Text Mining bereits Erfolge erzielen.

Neben dem unternehmerischen Einsatz bietet sich eine Anwendung von Methoden des Text Minings auch im wissenschaftlichen Bereich an. Die Anzahl an wissenschaftlichen Publikationen nimmt kontinuierlich zu und eine manuelle Kategorisierung von Beiträgen oder eine inhaltliche Überprüfung werden immer aufwendiger. Das Methodenspektrum des Text Minings ist inzwischen breit genug gefächert, sodass unterschiedlichste Bereiche möglichst automatisiert analysiert oder aufbereitet werden können.

Die nachfolgenden Beiträge entstanden durch Studierende im Rahmen des KI-Praktikums der Veranstaltung „Management Support Systems III – Künstliche Intelligenz“ an der Universität Osnabrück zum Thema „Text Mining in wissenschaftlichen Publikationen“. Im ersten Beitrag wird durch die Analyse der verwendeten Literaturquellen eines Beitrages anhand des Rankings der Quellen eine qualitative Bewertung des Beitrages durchgeführt. Der zweite Beitrag unterstützt die Auswahl von Keywords für einen Beitrag durch eine automatische Generierung von potenziellen Stichwörtern. Ein inhaltlicher Überblick über mehrere Beiträge wird durch den dritten Beitrag in Form einer Tag Cloud erstellt. Im abschließenden vierten Beitrag werden Beiträge auf ausgewählte linguistische Qualitätsmerkmale automatisiert überprüft, sodass eine Bewertung eines Beitrags auf sprachlicher Ebene unterstützt werden kann.

Allen Beiträgen liegt eine prototypische Implementierung zugrunde, welche grob das Potenzial der jeweiligen Idee aufzeigt. Die prototypischen Umsetzungen und die Dokumentationen orientieren sich am international anerkannten CRISP-DM. Die Implementierungen wurden mit der Software RapidMiner 6 durchgeführt. Für die Bereitstellung der Lizenzen möchten wir uns bei der Firma RapidMiner GmbH bedanken.

Osnabrück, im September 2014

Prof. Dr.-Ing. Bodo Rieger

Axel Benjamins

Qualitative Bewertung wissenschaftlicher Publikationen anhand ihrer verwendeten Literaturquellen

Peter Biermanski, Kai Seifert

Abstract. *Die Projekt-Idee besteht im Abgleich der in den zu analysierenden Artikeln benutzten Literaturquellen mit einer Ranking-Liste von als qualitativ hochwertig eingestuften Journals und Konferenzen. Mit Hilfe bereits vorhandener Ranking-Listen (AIS, VHB, WKWI) wird eine synthetisierte Ranking-Liste erstellt, welche verschiedene Bezeichnungs-Varianten der enthaltenen Journals und Konferenzen beinhaltet, um die Treffergenauigkeit beim Abgleich zu erhöhen. Mit Hilfe von Text Mining und ggf. weiteren Methoden der KI wird der Inhalt der Literaturverzeichnisse aus den Dokumenten extrahiert und mit der Ranking-Liste verglichen. Zusätzlich soll bestimmt werden wo die Quellen jeweils im Text benutzt wurden, um anhand dessen ihre Relevanz zu bewerten. Schlussendlich soll eine qualitative Bewertung der Journals anhand der in ihren Artikeln genutzten Quellen erfolgen.*

1 Business Understanding

Derzeit existieren viele wissenschaftliche Rankinglisten, die die Qualität von Journals oder Konferenzen bestimmen. Dabei wird jedoch die Qualität eines einzelnen Artikels nicht explizit betrachtet bzw. mit eingeschlossen. Auch sind nicht alle Journals in Rankings abgebildet. Das Geschäftsziel dieser Analyse soll die Bestimmung der Qualität von wissenschaftlichen Artikeln (und deren Journals) sein. Der Qualitätsgrad wird hierbei ausschließlich anhand der verwendeten Quellen bestimmt. Es sollen für jeden wissenschaftlichen Artikel die Quellen extrahiert und mit einer synthetisierten Rankingliste aus mehreren Wirtschaftsinformatik-Rankinglisten von Journals und Konferenzen abgeglichen werden. Hohe Qualität bedeutet in diesem Fall, dass von den genutzten Quellen der Anteil an Journals / Konferenzen der Rankingliste möglichst groß ist. Erweiternd ist vorstellbar, dass überprüft wird, wo und ob im Text die Quellen benutzt werden. Dies bietet eine höhere Sicherheit hinsichtlich der tatsächlichen Verwendung der Quellen. Neben der Erkenntnis der Qualität lassen sich auch andere Anwendungsszenarien in Betracht ziehen. So ist es denkbar, dass statt der synthetisierten Rankingliste eine eigene Liste benutzt wird, um z. B. nur wissenschaftliche Artikel zu finden, die die ei-

genen favorisierten Journals benutzen (eine Art „Favoritenliste“). Das Qualitätskriterium in Form der Rankingliste lässt sich also beliebig anpassen oder austauschen.

MS	Name	Ziele	Termin
1	Business Understanding	Zieldefinition und Projektplan	08.01.14
2	Data Understanding	Bestimmung der Datenqualität der vorliegenden Daten	15.01.14
3	Data Preparation	Bereinigtes Datenset und synthetisierte Rankingliste von Journals / Konferenzen	22.01.14
4	Modeling	Vorgehensmodell, Analyse der Daten	02.02.14
5	Evaluation / Deployment	Interpretation der Ergebnisse, Limitationen, möglicher späterer Einsatz	05.02.14

Tab. 1: Meilensteinplan des Projekts

(Quelle: Eigene Darstellung)

Das Vorgehen dieser Analyse orientiert sich an dem Cross Industry Standard Process for Data Mining (CRISP-DM) Referenzmodell. Dieses unterteilt die Auswertung der Daten in folgende Phasen: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation und Deployment (vgl. Chapman et al. 2000, S. 10). Jeder Abschluss einer Phase bildet dabei eine Art Zwischenziel. Tab. 1 fasst die Zwischenziele in einem Meilensteinplan übersichtlich zusammen und zeigt den Umfang des Projekts.

2 Data Understanding

2.1 Art und Umfang der Ausgangsdaten

Bei den Ausgangsdaten handelt es sich um wissenschaftliche Artikel dreier englischsprachiger Journals aus den Jahren 2008 bis 2012. Konkret handelt es sich dabei um die Journals *International Studies of Management & Organization (ISMO)*, *Management International Review (MIR)* und *ACM Transactions on Database Systems (TODS)*. Die Artikel liegen einzeln im PDF-Format vor. Die Dateinamen entsprechen den Titeln der jeweiligen Artikel. Sie sind je Journal in Ordnern nach Jahr, Band und Ausgabe kategorisiert. Insgesamt liegen 450 Dateien vor:

ISMO: 104 Dateien

MIR: 202 Dateien

TODS: 144 Dateien

2.2 Qualität der Ausgangsdaten

Da es sich bei den Ausgangsdaten um publizierte, wissenschaftliche Artikel jeweils eines bestimmten Journals handelt, folgen diese einem einheitlichen, für das jeweilige Journal typischen, Aufbau. So sind Schriftgröße und -art, Zeilenabstände, Seitenränder, Kopf- und Fußzeilen sowie die Zitationsstile, zumindest auf den ersten Blick, durchgängig einheitlich.

Das Hauptaugenmerk bei der Analyse der Artikel ist auf deren Literaturverzeichnisse gerichtet, da in erster Linie die Quellenangaben in die Auswertung einbezogen werden müssen. Zu jeder Quellenangabe gehört die Herkunft, also das Buch, das Journal, die Konferenz oder die Internetseite, in welchem ein zitierter Beitrag erschienen und ggfs. nachlesbar ist. Genau diese Angabe wird im weiteren Verlauf für den Vergleich mit einer synthetisierten Ranking-Liste von hochklassigen, wissenschaftlichen Journals und Konferenzen genutzt.

Typischer Aufbau der Quellenangaben (am Beispiel eines Journalbeitrags):

- ISMO: Autorenname(n); Erscheinungsjahr; Titel (in Hochkommata); Journaltitel (kursiv); Band / Ausgabe; Seitenangabe
- MIR: Autorenname(n); Titel; Journaltitel (kursiv); Band / Ausgabe; Erscheinungsjahr (bei späteren Ausgaben in Klammern); Seitenangabe
- TODS: Autorenname(n); Erscheinungsjahr; Titel; Journaltitel (kursiv); Band / Ausgabe; Seitenangabe

Alle Elemente sind je nach Journal durch Punkte, Kommata oder Doppelpunkte voneinander getrennt und einzelne Elemente werden teilweise durch Hochkommata oder kursive Schrift hervorgehoben oder in Klammern gesetzt. Teilweise ändern sich Formatierungen jedoch auch innerhalb eines Journals im Laufe der Zeit. So wird beim MIR seit dem Jahr 2010 ein anderer Zitationsstil verwendet als zuvor. Damit sich diese Abweichungen im weiteren Verlauf der (semi-)automatisierten Verarbeitung nicht störend auswirken, sollen nicht nur die Bezeichnungen der Journals und Konferenzen, sondern die gesamten Informationen aus den Literaturverzeichnissen extrahiert werden. Auf den späteren Abgleich mit der Ranking-Liste hat dies keine Auswirkungen.

3 Data Preparation

3.1 Vorauswahl

Neben den wissenschaftlichen Artikeln der drei Journals enthalten die Ausgangsdaten teilweise spezielle in den Journals erschienene Artikel, welche für die bevorstehende Analyse irrelevant sind [Anzahl in eckigen Klammern]:

ISMO: Call for Papers [1x], Biographie (Biography) [1x], Vorworte (Prefaces) [9x], Einleitungen (Introductions / Guest Editors' Introductions) [7x], Autorenlisten (Author Indices) [2x]

MIR: Leitartikel (Editorials) [3x], Danksagungen (Acknowledgements) [3x], Buchempfehlungen (Biblio Services) [17x]

TODS: Vorstellungen von Konferenzbeiträgen (Introductions to Conferences) [3x]

Während des ersten Schrittes der automatisierten Verarbeitung mit dem RapidMiner sollen die Literaturverzeichnisse aus den wissenschaftlichen Artikeln herausgetrennt werden. Da die oben genannten speziellen Artikel jedoch (in den meisten Fällen) keine Literaturverzeichnisse enthalten, muss vorerst keine Vorauswahl getroffen werden. Jene Artikel werden im darauffolgenden Verarbeitungsschritt (durch ein Excel-Makro) automatisch entfernt. Nicht relevante Artikel, welche es dennoch durch den zweiten Verarbeitungsschritt schaffen, können auch noch nach diesem, jedoch mit geringerem Aufwand, manuell entfernt werden.

3.2 Datenformat

Das PDF-Format der Ausgangsdaten ist für eine strukturierte Weiterverarbeitung ungeeignet. Neben dem Text selbst werden auch Meta-Informationen über selbigen benötigt. Um Meta-Informationen über die Ausgangsdaten zu erhalten, werden diese mit Hilfe des Tools LA-PDFText (vgl. Ramakrishnan et al. 2012), welches auf die Extraktion von Textblöcken aus wissenschaftlichen Arbeiten ausgelegt ist, verarbeitet. Mit Hilfe des integrierten blockify-Befehls werden automatisiert XML-Dateien für jedes einzelne Dokument erstellt, welche Informationen über Schriftart und -größe sowie Position jedes einzelnen Wortes innerhalb des jeweiligen Dokuments liefern. Diese Informationen helfen im nächsten Schritt bei der Bereinigung der Daten, um das finale Datenset für die Analyse zu erhalten – reine Literaturverzeichnisse aller Artikel im Klartext.

3.3 Extraktion relevanter Daten

Zur Erreichung des Hauptziels, dem Abgleich der in den Quellen verwendeten Journals und Konferenzen mit einer Ranking-Liste, müssen zunächst die Literaturverzeichnisse aus jedem Artikel extrahiert werden. Diese können anschließend auf Journal- und Konferenzbezeichnungen der Ranking-Liste durchsucht werden. Dafür reicht es nicht aus, die Texte der Literaturverzeichnisse aus den PDF-Dateien „herauszuschneiden“, denn diese erstrecken sich in der Regel über mehrere Seiten und beinhalten somit Kopf- und / oder Fußzeilen sowie Seitenzahlen. Um den reinen Inhalt der Literaturverzeichnisse zu extrahieren, werden die mit LA-PDFText erzeugten XML-Dateien in den RapidMiner eingelesen und mit Hilfe eines, auf jedes Journal angepassten, Text Mining Prozesses automatisiert weiterverarbeitet. Mit Hilfe der Meta-Informationen zu jedem Wort können explizit nur die relevanten Informationen, also die Quellenangaben selbst pro Dokument, extrahiert werden.

3.4 Verarbeitung der XML-Dateien mit dem RapidMiner

Für jedes der drei Journals wurde ein spezieller Text Mining Prozess im RapidMiner erstellt, um jedes Literaturverzeichnis in einen zusammenhängenden Fließtext umzuformen. Das Grundprinzip dieser Prozesse ist immer dasselbe (*RapidMiner-Operatoren in Klammern*):

1. Alle XML-Dateien eines Journals einlesen und mit Metadaten verknüpfen (*Process Documents From Files*)
    a. Teil des Literaturverzeichnisses mit Hilfe von typischen Merkmalen, wie der Überschrift „References“ aus jedem Dokument heraustrennen (*Cut Document*)
    b. Alle Dokumente zu einem Dokument zusammenfügen, da der Tokenize-Operator nicht mit mehreren umgehen kann (*Combine Documents*)
    c. Dokumentinhalt in einzelne Einheiten zerlegen (*Tokenize*)
    d. Notwendige Einheiten herausfiltern (*Filter Tokens by Content*)
    e. Einheiten oder Teile davon ersetzen (*Replace Tokens*)
    f. Die Anzahl von Quellenangaben zählen (*Extract Token Number*)
2. Ausgabe der reinen textuellen Inhalte aller Literaturverzeichnisse inkl. Metainformationen (u. a. Dateinamen und Labels) sowie der Anzahl der Quellenangaben als ResultSet sowie als Excel-Datei (*Write Excel*)

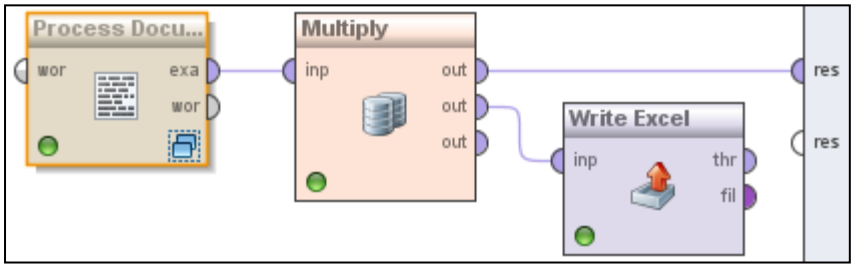

Abb. 1: RapidMiner-Hauptprozess

(Quelle: Eigene Darstellung)

Abb. 1 zeigt den RapidMiner-Hauptprozess, welcher für alle Journals gleich ist. Im Detail weichen die Prozesse jedoch voneinander ab, da die Formate und Layouts jedes Journals individuell unterschiedlich sind und die Kopf- / Fußzeilen auf unterschiedliche Art und in unterschiedlicher Reihenfolge herausgefiltert werden müssen. Bei den Artikeln des ISMO kann die Kopfzeile bspw. herausgefiltert werden, da sie immer in Schriftgröße 9 formatiert ist, die relevanten Inhalte des Literaturverzeichnisses jedoch in Schriftgröße 8. Beim MIR ist diese Vorgehensweise nicht möglich, da alle Inhalte dieselbe Schriftgröße haben. Deshalb werden hier die Inhalte der Kopfzeilen mit Hilfe ihrer Position auf der Seite ermittelt und entfernt. Dazu kommt ein regulärer Ausdruck zum Einsatz, der jedes Element anhand seiner y1- und y2-Koordinaten identifiziert:

*y1="[1-4][0-9]"\s.*y2="[1-4][0-9]"*

Die y-Koordinaten der Kopfzeilen-Elemente liegen bei diesem Journal typischerweise bei 32 bzw. 42. Der reguläre Ausdruck filtert nach Koordinaten, die einen Wert zwischen 10 und 49 haben. Damit werden auch evtl. Abweichungen von der Regel abgefangen. Da verschiedene, überflüssige Elemente (Tokens) der XML-Dateien entfernt werden müssen, wiederholen sich die Teilprozesse 1.c. und 1.d. je nach Bedarf und vereinzelt werden auch Teile innerhalb bestimmter Token ersetzt (siehe Teilprozess 1.e.).
Für die Ermittlung der Anzahl von Quellenangaben pro Literaturverzeichnis wird die Tatsache genutzt, dass zu jeder Quellenangabe typischerweise die Angabe des Erscheinungsjahres gehört. Nachdem die Schritte 1.a. – 1.e. abgearbeitet worden sind, liegen die Literaturverzeichnisse im Klartext vor, wovon alle Elemente einzelne Token repräsentieren. Mithilfe des Operators *Extract Token Number*, welcher wiederum nach regulären Ausdrücken suchen kann, werden nun alle Token gezählt, welche die Form einer Jahreszahl haben:

([1][9][0-9]{2})|([2][0][0-1][0-9])

Dieser Ausdruck schließt alle Elemente ein, welche zwischen 1900 und 2019 liegen und schließt damit wiederum weitestgehend alle vierstelligen Seitenangaben aus, welche ebenfalls ein typischer Bestandteil von Quellenangaben sind und somit das Ergebnis geringfügig verfälschen können. Als Konsequenz dieser Vorgehensweise kann somit nicht garantiert werden, dass nicht auch vereinzelt Seitenangaben im Bereich zwischen 1900 und 2019 als Jahreszahl interpretiert werden und somit die tatsächliche Anzahl an Quellen marginal erhöhen. Abb. 2 zeigt beispielhaft den RapidMiner-Unterprozess für das MIR, welcher die Teilprozesse 1.a. – 1.f. beinhaltet.

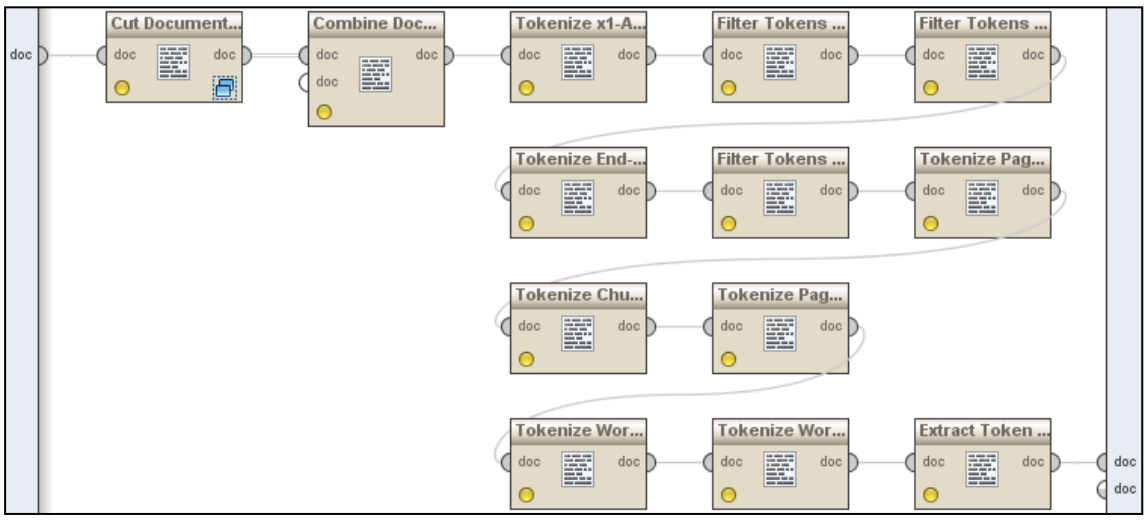

Abb. 2: RapidMiner-Unterprozess von *Process Documents From Files*
(Quelle: Eigene Darstellung)

Als Ergebnis der drei RapidMiner-Prozesse wird jeweils eine Excel-Datei erzeugt, welche pro Zeile jeweils ein Literaturverzeichnis, den Dateinamen des entsprechenden Artikels und die Gesamtzahl seiner Quellen enthält. Leider werden beim Schreiben der Excel-Dateien durch den RapidMiner die Inhaltsverzeichnisse teilweise abgeschnitten. Um dies zu korrigieren und im selben Zug noch überflüssige Leerzeichen zwischen den einzelnen Wortelementen zu entfernen, kann die komplette Spalte mit den Literaturverzeichnissen aus dem ExampleSet, welches im RapidMiner ausgegeben wird, manuell in die Excel-Tabelle kopiert (und somit die potentiell fehlerhaften Daten überschrieben) werden. Die Daten in den entstehenden drei Excel-Dateien bilden schließlich die Grundlage für die Auswertung der Quelleninhalte in Verbindung mit der Ranking-Liste.

3.5 Ranking-Liste

Die Herkunft der verwendeten Quellen der einzelnen wissenschaftlichen Artikel soll im nächsten Schritt mit einer Ranking-Liste verglichen werden, welche eine Auswahl hochklassiger Journals und Konferenzen beinhaltet. Mit Hilfe dieser Liste kann beurteilt werden, wie viele der benutzten Quellen eines Artikels aus diesen Journals bzw. Konferenzen stammen. In Verbindung mit der Gesamtzahl von Quellen innerhalb eines Artikels können Rückschlüsse auf die Gesamtqualität der benutzten Literatur pro Artikel bzw. pro Journal (ISMO, MIR, TODS) gezogen werden.

Für die Erstellung einer synthetisierten Ranking-Liste wurden die Top 25 Journals nach AIS-Ranking (vgl. AIS 2010), darüber hinaus die mit den Prädikaten A+, A und B bewerteten Journals und Konferenzen nach VHB-Teilranking Wirtschaftsinformatik und Informationsmanagement (vgl. VHB 2008) sowie die mit A bewerteten Journals und Konferenzen nach WKWI-Ranking (vgl. WKWI 2008) benutzt. Das Ergebnis bildet eine Liste der Top 50 Journals und Konferenzen aus dem Bereich der Wirtschaftsinformatik im Excel-Format. Um eine möglichst hohe Trefferquote beim Abgleich der Journals bzw. Konferenzen der Ranking-Liste mit den Literaturverzeichnissen zu gewährleisten, wurden ihre jeweiligen Bezeichnungen um weitere Varianten ergänzt, die bei der folgenden Auswertung in einem Phrasenkatalog berücksichtigt werden.

4 Modeling

In diesem Teil des Projekts sollen die zuvor bereinigten und aufbereiteten Daten analysiert werden. Hierzu sind mehrere Schritte notwendig, um von den Daten bis hin zur Beantwortung der Qualitätsfrage zu gelangen. Das folgende Vorgehensmodell veranschaulicht die nötigen Schritte der Analyse:

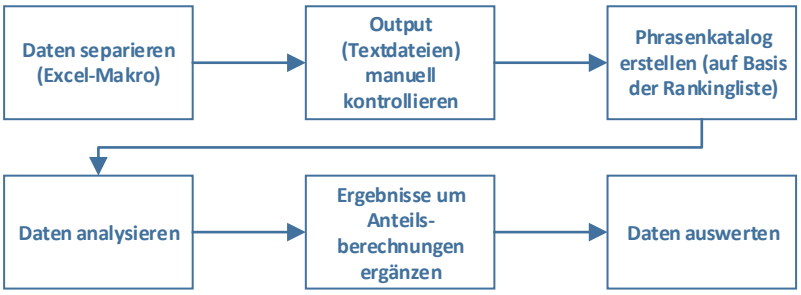

Abb. 3: Vorgehensmodell der Datenanalyse

(Quelle: Eigene Darstellung)

Die Analyse basiert im Allgemeinen auf einem Phrasenkatalog und einer Sammlung von Dokumenten (mit den Quellen als Inhalt). Da die Übereinstimmungen zwischen den Quellen der wissenschaftlichen Artikel und der Rankingliste betrachten werden, erscheint ein Phrasenkatalog sehr sinnvoll. Jeder Journal- / Konferenzname kann als Phrase gespeichert und schließlich im Quellentext gesucht werden. Mit Hilfe dieses Katalogs lassen sich zudem verschiedene syntaktische Varianten der Journals / Konferenzen abbilden. Durch die Beachtung mehrerer Schreibweisen wird das Ergebnis deutlich repräsentativer. Die Ergebnisse dieser Suche lassen sich visuell oder als Tabellen abbilden und evaluieren. Zur Bearbeitung dieser Analyse bietet sich die Software QDA Miner mit dem Wordstat-Plugin an (Provalis Research 2013). Nachfolgend ist das Analysemodell im Detail erklärt.

Wie im vorherigen Kapitel beschrieben, erhalten wir nach der Bereinigung der Daten pro Journal eine Excel-Datei mit allen Quellen, der Quellenanzahl und dem Namen zu jedem wissenschaftlichen Artikel. Darauf aufbauend lässt sich für jeden Artikel analysieren, welche Quellen aus Journals oder Konferenzen der Rankingliste stammen. Ziel ist es, jeden Quellentext pro Artikel einzeln in den QDA Miner einzulesen, um die Analyse später zu vereinfachen (näheres hierzu später im Text). Demnach soll für jede Zeile aus der Excel-Datei ein Textdokument erstellt werden, welches als Inhalt die gesamten Quellen und als Dateinamen den Titel des wissenschaftlichen Artikels hat. Hierfür dient ein Excel-Makro, welches durch jede Zeile der Excel-Datei iteriert und Textdokumente mit den entsprechenden Inhalten erzeugt. Abb. 4 zeigt den Code des Makros. Im abgebildeten Beispiel wird bis zur 145. Zeile der Zellinhalt in eine Textdatei geschrieben. Der Anfang ist in Zeile zwei, da die erste Zeile nur die Spaltenbeschreibung enthält und somit irrelevant ist.

```
Sub Create_Files()

'Quellentext
Dim record As String

'Soviele Dateien schreiben, wie es Zeilen gibt (Ende manuell anzupassen)
For zeile = 2 To 145
     Open "C:\Users\Acer\Desktop\KI\ACMTransDB\" & Cells(zeile, 2) & ".txt" For Output As #1
     'Schreibe Zelleninhalt in die Datei
     record = Cells(zeile, 1)
     Print #1, record
     'Datei schließen
     Close #1
Next zeile

End Sub
```

Abb. 4: Excel-Makro zur Überführung der Quellen in einzelne Textdokumente (Quelle: Eigene Darstellung)

Nach Ausführen des Makros lassen sich die Textdateien im angegebenen Ordner betrachten. Im Einzelfall müssen überflüssige Dateien manuell entfernt werden (z. B. „preface.txt“, welche keinen sinnhaften Inhalt aufweist). Nach dieser kurzen Sichtung der Textdokumente können diese in den QDA Miner hineingeladen werden. Hierbei bietet es sich an, für alle drei Journals jeweils ein Projekt mit den entsprechenden Textdokumenten der jeweiligen Artikel zu erstellen. So sind später auch Rückschlüsse bezüglich der Journals möglich. Bevor die Inhaltsanalyse vorgenommen wird, muss zuerst der Phrasenkatalog im Wordstat-Plugin erstellt werden. Für jedes Journal / jede Konferenz aus der Rankingliste wird eine Kategorie erstellt. In dieser können dann syntaktische Varianten der Journal- / Konferenznamen hinzugefügt werden (Abb. 5). Groß- und Kleinschreibung sowie Sonderzeichen werden nicht beachtet, wodurch die Trefferquote noch weiter erhöht wird. Neben diesem Katalog lässt sich zudem ein weiterer erstellen, welcher bspw. nur die Top 25 Journals / Konferenzen der Rankingliste beinhaltet. So kann auf einfache Art und Weise weiter differenziert werden.

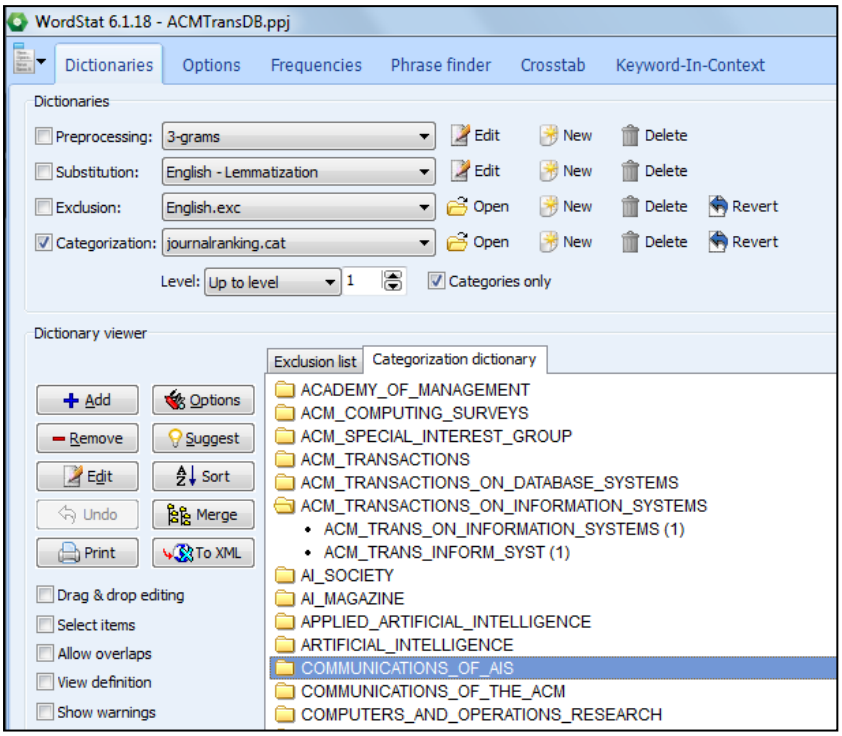

Abb. 5: Phrasenkatalog mit den Journals / Konferenzen der erstellten Rankingliste (Quelle: Eigene Darstellung)

Nachdem der Katalog erstellt worden ist, lässt sich die Inhaltsanalyse für jedes Projekt durchführen. Hierbei ist wichtig, dass die „FILE“-Variable (Dokumentname, z. B. „Datei.txt“) mit in die Analyse übergeben wird (auswählbar im Dialogfenster, wenn man die Inhaltsanalyse startet). Dies bietet die Möglichkeit, die Treffer pro Dokument anzuzeigen. Da die Namen der generierten Textdateien jeweils dem Titel des wissenschaftlichen Artikels entsprechen, bedarf es keiner zusätzlichen Metadaten oder ähnliches, um die Titel der Beiträge als Spaltenbeschriftung anzuzeigen.

Das Programm sucht in allen Dokumenten ausschließlich nach den Phrasen aus dem zuvor definierten Katalog. Die Treffer und deren Häufigkeiten lassen sich unter dem Reiter „Frequencies“ anzeigen. Interessanter für die Qualitätsbestimmung ist jedoch der Reiter „Crosstab“, in welchem die Phrasen aus dem Katalog gekreuzt mit den Dateien (durch die „FILE“-Variable) dargestellt sind und somit die Häufigkeiten pro Dokument angezeigt werden. Alle Statistiken lassen sich in Excel exportieren oder grafisch in Diagrammen visualisieren. Die in die Excel-Datei exportierten Daten lassen sich in einem nächsten Schritt weiter ergänzen. Bisher können immer noch keine Aussagen über die Qualität getroffen werden. Hierfür ist die Relation von Top-Quellen zu der Gesamtanzahl an Quellen pro Dokument zu berechnen. Der Übersichtlichkeit wegen wird die

Excel-Tabelle zunächst gedreht, sodass in jeder Zeile die Namen der wissenschaftlichen Artikel stehen und in jeder Spalte die der Journals / Konferenzen. Danach werden die Spalten „Summe Top Journals / Konferenzen“, „Quellen gesamt“ und „Anteil Top Journals / Konferenzen“ manuell hinzugefügt. Ersteres kumuliert die Häufigkeiten aller Treffer. „Quellen gesamt“ ist die Gesamtanzahl an Quellen des jeweiligen wissenschaftlichen Artikels und ist aus dem anfänglichen Datenset zu entnehmen. Durch die Bereinigung der Textdateien und dadurch, dass der QDA Miner nur solche Dokumente in der Tabelle anzeigt, welche Treffer enthalten, kann die Spalte mit den Quellenanzahlen nicht direkt aus dem Datenset übernommen werden. Der Einfachheit wegen ist es sinnvoll zunächst die Spalten „token number“ (Quellenanzahl) und „filename“ (Name) aus dem Datenset in die Excel-Datei mit den Häufigkeiten zu kopieren und anschließend alle überflüssigen Zeilen mit Dokumentnamen und der Quellenanzahl wieder zu löschen.

Der Anteil der Top Journals / Konferenzen lässt sich mit einer trivialen Excelformel berechnen, indem die Kumulation der Treffer durch die Gesamtanzahl der Quellen jeweils pro Dokument dividiert wird. Anschließend lässt sich noch der Mittelwert aller Anteile bestimmen, um so die Qualität des gesamten Journals messen zu können. Hierfür werden alle Anteile kumuliert durch die Gesamtanzahl der untersuchten Dokumente (Anzahl der Textdateien) dividiert. Durch diese Berechnungen ist es nun möglich, die Qualität pro Artikel sowie pro Journal zu messen.

5 Evaluation

Die Analyse hat mehrere Ergebnis-Dateien hervorgebracht. Es wurden drei Journals untersucht, für die jeweils eine Excel-Datei die Häufigkeiten und Anteile der Top-Quellen pro wissenschaftlichen Artikel angibt und je eine weitere, die die Häufigkeiten insgesamt pro Journal aufzeigt. Letztere sind auch als Balkendiagramme vorhanden. Da nur bei dem TODS Journal genügend Ergebnisse gefunden wurden, ist für dieses noch ein weiteres Balkendiagramm vorhanden, welches die Häufigkeit der gefundenen Topquellen pro Artikel anzeigt.

Bei dem TODS Journal sind insgesamt sechs Journals / Konferenzen aus der synthetisierten Rankingliste in den Artikeln gefunden worden. Hierbei sind „ACM Transactions on Database Systems“ und „IEEE Transactions“ die am häufigsten zitierten Top-Journals. Mit 23,53 Prozent sind in dem Artikel „An Information-Theoretic Analysis of Worst-Case Redundancy in Database Design_spatial“ anteilig gesehen die meisten Top-

Quellen verwendet worden. Unter Einbezug aller Artikel des Journals ergibt sich ein durchschnittlicher Anteil der Top-Quellen von 6,14 Prozent. Bei dem MIR Journal gab es lediglich zwei Journals / Konferenzen in allen Artikeln (Anteil aller Quellen: 0,05 Prozent). Ähnlich gering viel die Trefferquote bei dem ISMO Journal aus, wo nur eines gefunden wurde (Anteil aller Quellen: 0,03 Prozent). Im Umkehrschluss ist die Qualität dieser beiden Journals zumindest nach der hier verwendeten Rankingliste recht gering. Auch die untersuchten Artikel des TODS Journals verwenden nur zu knapp einem Viertel qualitativ hochwertige Quellen. Dieses Gesamtergebnis aller Journals ist jedoch zunächst kritisch zu betrachten. Die Qualität wird ausschließlich an der Quellenauswahl gemessen. Die synthetisierte Rankingliste umfasst nur Wirtschaftsinformatik-Journals / -Konferenzen von drei verschiedenen Rankinglisten (AIS, VHB, WKWI). Potentiell ist es durchaus möglich, dass die hier untersuchten Journals vorrangig aus Journals / Konferenzen anderer Fachgebiete zitiert haben (z. B. Informatik oder Management). Dies liegt nahe, da die Journals abgesehen von dem TODS Journal nicht direkt aus dem Gebiet der Wirtschaftsinformatik stammen. D. h. ausschließlich für dieses Journal ist das Ergebnis ausreichend repräsentativ.

Die Rankingliste ist demnach als eine Art Definition der Qualität zu betrachten und stellt keinen Anspruch auf Allgemeingültigkeit, da die vollständige Abdeckung aller Journals aus verschiedenen fachlichen Bereichen keineswegs gegeben ist. Es ist daher ggf. sinnvoll weitere Rankinglisten zu erstellen oder vorhandene zu benutzen und die Journals mit diesen abzugleichen. Je nachdem welches Ziel verfolgt wird, muss erwogen werden inwieweit durch die Rankingliste die Qualitätsfrage beantworten werden kann. In diesem Fall lässt sich also lediglich die Aussage treffen, dass zumindest teilweise Wirtschaftsinformatik-Top-Quellen in den Journals verwendet worden sind und dies zur Qualität des wissenschaftlichen Artikels beitragen kann.

Eine weitere Schwachstelle sind die Varianten der Journal- / Konferenznamen. Zwar werden mehrere Varianten berücksichtigt, jedoch ist es schwer wirklich alle abzudecken, sodass einige Quellen möglicherweise nicht gefunden werden. Auch in entgegengesetzter Richtung können Fehler entstehen. Das Journal „Information Systems“ kann aufgrund der allgemeinen Begriffe fälschlicherweise in anderen Angaben (z. B. Titel) gefunden werden. Diese Restriktionen führen zu einem gewissen Fehleranteil der Analyse.

Weiterhin existieren zwei Schwachstellen bei der Transformation der Daten. Es ist es nicht optimal, die Quellenanzahl aufgrund der Jahreszahlen zu zählen. Zwar werden nur

die Jahre zwischen 1900 und 2019 gezählt, jedoch kann es vorkommen, dass aus Büchern zitiert wird, welche Seitenzahlen in jenem Bereich haben. Dass Bücher mehr als 1900 Seiten besitzen und zudem aus genau diesem Intervall (1900 – 2019) zitiert wird ist höchstwahrscheinlich sehr selten, beeinflusst aber schlussendlich das Ergebnis. Ebenfalls entstehen durch die Transformation mittels LA-PDFText offensichtlich willkürliche Fehler. So wird das „i“ in jedem Dateinamen weggelassen und Textblöcke werden nicht immer korrekt in XML abgebildet (Reihenfolge einzelner Chunk-Tags).
Abschließend lässt sich noch die Frage der inhaltlichen Qualität der wissenschaftlichen Artikel diskutieren. Dadurch, dass nur das Quellenverzeichnis durchsucht wird, besteht kein Anspruch auf die wirkliche Nutzung der Quellen im Text. Es wird weder überprüft, wo und ob die Quellen im jeweiligen Text benutzt werden, noch ob die Zitate inhaltlich korrekt sind. Die Bestimmung der Qualität wissenschaftlicher Artikel aufgrund der Quellenangaben ist folglich zweifelhaft. In einem nächsten Schritt könnte noch einmal intensiv überprüft werden, inwiefern diese Aspekte umsetzbar sind. Trotz dieser Schwachpunkte lässt diese Analyse eine qualitative Tendenz zu, sofern das Fachgebiet der Rankingliste und das der wissenschaftlichen Artikel dasselbe ist.

6 Deployment

Das hier vorgestellte Modell zur Bewertung der Qualität von wissenschaftlichen Artikeln kann für wissenschaftliche Zwecke und Forschungen nützlich sein. Es lassen sich komplette Journals auf Basis der untersuchten Artikel bewerten. Neben der hier synthetisierten Rankingliste sind auch andere Listen integrierbar, sodass z. B. für jede Fachrichtung (Wirtschaftsinformatik, Informatik etc.) ein eigenes Ranking erstellt werden könnte. Durch die Auswertung der Journals kann dann schnell analysiert werden, ob die Artikel jenes Journals gute Quellen aus der jeweiligen Fachrichtung benutzen. Ebenso könnte eine Art Favoritenliste erstellt werden, in welcher von einem selbst präferierte Journals / Konferenzen enthalten sind. Um dies jedoch auch sicher zu gewährleisten müsste eine zusätzliche Überprüfung stattfinden, in welcher untersucht wird, wo welche Quellen auch wirklich im Text genutzt werden und inwieweit diese relevant für den Inhalt bzw. dessen Qualität sind.

7 Literatur

AIS. (2010). MIS Journal Rankings. Retrieved January 31, 2014, from http://start.aisnet.org/?JournalRankings

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide. *The CRISP-DM consortium*.

Provalis Research. QDA Miner: Wordstat. Retrieved January 14, 2014, from http://provalisresearch.com/products/content-analysis-software/

Ramakrishnan, C., Patnia, A., Hovy, E., & Burns, G. (2012). Layout-Aware Text Extraction from Full-text PDF of Scientific Articles. *Source Code for Biology and Medicine 7(1): 7*. doi:10.1186/1751-0473-7-7

VHB. (2008). Teilranking Wirtschaftsinformatik und Informationsmanagement. Retrieved January 31, 2014, from http://vhbonline.org/service/jourqual/jq2/teilranking-wirtschaftsinformatik-und-informationsmanagement/

WKWI. (2008). WI-Orientierungslisten. Retrieved January 31, 2014, from http://www.springerlink.com/index/10.1365/s11576-008-0040-2

Automatische Generierung und Verifizierung von Keywords für wissenschaftliche Publikationen

David Lübbing, Sebastian Osada

Abstract. *Die Suche nach wissenschaftlichen Artikeln in einer großen Literaturdatenbank kann sehr viel Zeit in Anspruch nehmen. Keywords, die möglichst genau aber gleichzeitig nicht zu abstrakt sind, helfen hier bei der Suche. Im Folgenden soll eine Möglichkeit gezeigt werden, relevante Keywords mithilfe von künstlichen Intelligenzen aus einem Dokument zu erzeugen. Zusätzlich soll festgestellt werden wie genau bzw. ob die Keywords der Dokumente überhaupt mit den Calls for Papers der relevanten Konferenz übereinstimmen.*

1 Business Understanding

Ziel der KI ist Computern Dinge beizubringen die Menschen derzeit besser können[1]. Dazu zählt auch die Stichwortgenerierung für Dokumente. In diesem Projekt wenden wir aktuelle Verfahren des Text Minings an, um Stichwörter (englisch: Keywords) aus den Texten dieser Dokumente zu extrahieren. Die Ziele der Analyse ergeben sich aus der Unterstützung des wissenschaftlichen Personals bei der Nutzung von wissenschaftlichen Texten hinsichtlich der Keywords, insbesondere von Dokumenten oder Dokumentensammlungen, die im Vorfeld nicht mit Stichwörtern versehen wurden. Auf der einen Seite sind Dokumente aufgrund der Keywords schneller in ein Themengebiet einordnen und auf der anderen Seite lassen sich relevante Texte in großen Dokumentensammlungen besser finden.

Durch den Abgleich der Keywords mit den relevanten Call for Papers ist es anschließend möglich zu untersuchen, inwiefern das Dokument oder besser seine Keywords zu diesem in Beziehung steht.

Ein Problem bei der Analyse von wissenschaftlichen Texten ist die Auswahl von Stichwörtern und somit Reduktion einer hohen Anzahl von eventuell relevanten Stichwörter. Nicht immer sind alle wissenschaftlichen Dokumente mit Stichwörtern versehen. Obwohl dies durch einen Menschen, der in dem Kontext der wissenschaftlichen Dokumen-

[1] Rich, E. (1983). Artificial Intelligence, S. 1.

te forscht eine zu lösende Aufgabe ist, benötigt dieser Prozess eine gewisse Zeit. So kann dies für ganze Dokumentensammlungen, die nicht mit Keywords verschlagwortet wurden, eine zeitintensive Aufgabe sein. Zudem werden Stichwörter in der Regel vom Autor erstellt. Eine dritte Person müsste diesen Text komplett erfassen um die Aggregation des Textes auf einzelne Stichwörter durchzuführen obwohl diese mit dem Thema nicht vertraut ist.

Die Ziele dieser Text Mining Analyse lauten wie folgt:

1) Analyse der Texte
    - Erkennen und Extrahieren der einzelnen Textabschnitte (Abstract, Introduction, Body, Conclusion)
2) Automatische Generierung der Keywords je Dokument
    - Generierung der Keywords je Dokumentenabschnitt anhand der Häufigkeit
    - Wertung anhand der Position im Text
3) Verifizierung von Keywords einer wissenschaftlichen Publikation anhand der relevanten Call for Papers
    - Reiner Abgleich der Stichwörter

Das Projekt wird innerhalb von 4 Wochen bearbeitet. Da der "Cross Industry Standard for Data Mining"[2] als Grundlage für die Projektphasen genutzt wird, besteht das Vorgehen innerhalb der Bearbeitungszeit grob aus 6 Phasen: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation sowie Deployment.

Für die Umsetzung des Text Mining werden die wissenschaftlichen Texte zunächst hinsichtlich Ihres Aufbaus analysiert, um diese sinnvoll im Text Mining-Prozess verwenden zu können. Dies soll mit dem Werkzeug LA-PDFText[3] geschehen, für das entsprechende Regeln definiert werden müssen. Anschließend sollen die extrahierten Abschnitte zur Nutzung im Text Mining-Prozess mithilfe des Werkzeugs RapidMiner 6[4] verwendet werden.

---

2 Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*.

3 Ramakrishnan, C., Patnia, A., Hovy, E. H., & Burns, G. A. (2012). Layout-aware text extraction from full-text PDF of scientific articles. *Source code for biology and medicine*, *7*(1).

4 (2007). RapidMiner - Predictive Analytics, Data Mining, Self-service, open ... Retrieved January 27, 2014, from http://rapidminer.com/.

2 Data Understanding

Um die Daten in den einzelnen Werkzeugen der Text Mining Analyse sinnvoll verwenden zu können ist das Verständnis der Struktur und Eigenschaften der vorliegenden Daten nötig. Die vorliegenden wissenschaftlichen Dokumente (Paper) sind Konferenzbeiträge der "Hawaii International Conference On System Sciences" Nummer 41 aus dem Jahre 2008 (HICSS 2008)[5].

Der Tagungsband ist wie folgt strukturiert:

- Proceedings Title Year
  - Track m
    - Minitrack n
      - Paper x

Die Beiträge der jeweiligen Tracks sind in einzelne Ordner einsortiert, die den Namen des Tracks tragen. Bei dem Abgleich der Tracktitel fällt auf, dass der Tracktitel nicht immer komplett übernommen wurde, sondern zum Teil Sonderzeichen entfernt wurden, wie zum Beispiel Bindestriche. Die laufende Nummer scheint die Struktur Tracks wiederzuspiegeln.

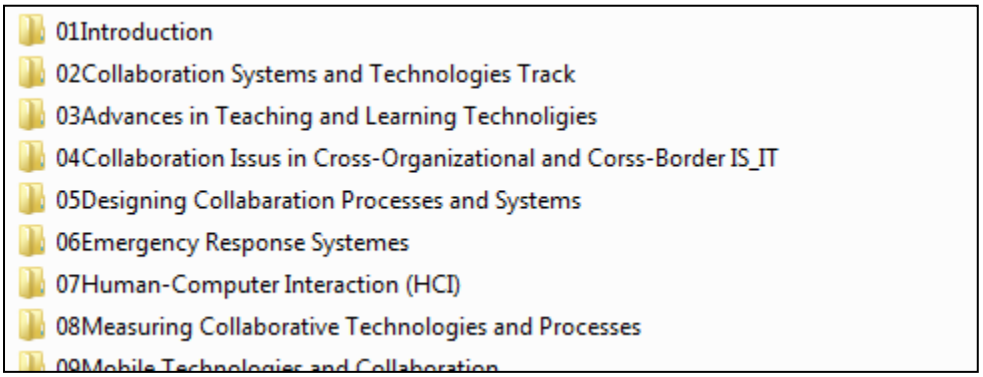

Abb. 6: Ordnerstruktur der Daten

(Quelle: Eigene Darstellung)

Die Beiträge liegen im PDF-Format vor und tragen den Titel der einzelnen Paper. Zu beachten ist, dass die Sonderzeichen in den Titeln im Dateinamen übernommen wurden. Dies kann bei der Verarbeitung zu Problemen führen

[5] (2008). HICSS-41 Highlights. Retrieved January 14, 2014, from http://www.hicss.hawaii.edu/hicss_41/41highlights.htm.

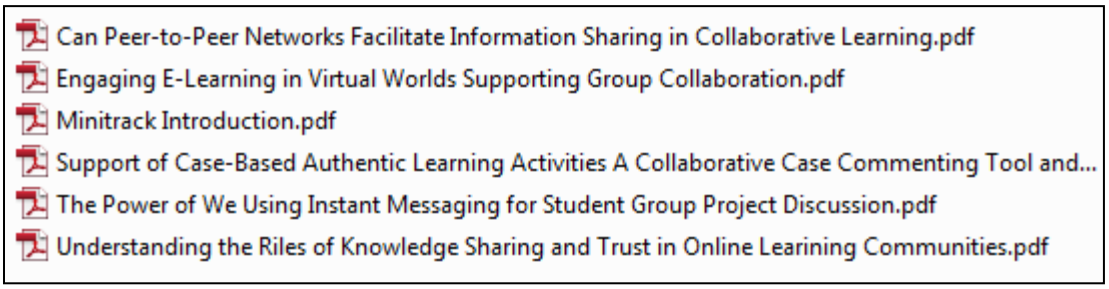

Abb. 7: Dateistruktur in einem Track

(Quelle: Eigene Darstellung)

Die einzelnen Konferenzbeiträge haben jeweils eine Gliederung, die allerdings nicht einheitlich ist. Die Seiten der Texte sind meistens zweispaltig aufgebaut und zum Teil durch Grafiken und Tabellen unterbrochen. Generell lässt sich folgender Aufbau erkennen.

- Abstract
- Introduction (nicht immer auch so benannt)
- Body
- Conclusion
- References
- Appendix (nicht immer vorhanden)

Bei der Untersuchung der Ordner fällt auf, dass nicht alle Ordner tatsächlich Dokumente enthalten. Zudem sind nicht alle vorhandene Dateien tatsächlich Paper, sondern zum Teil sind lediglich Einleitungen zu den Tracks beziehungsweise Minitracks in den Ordnern abgelegt.
Dies gilt es in der Phase der Data Preparation zu beachten.

3 Data Preparation

Um das Ziel des Abgleiches der ermittelten Keywords mit dem Call for Papers zu erreichen müssen zunächst die Call for Papers ermittelt werden. Durch eine kurze Internetrecherche können diese auf den zugehörigen Webseiten[6] der HICSS 41 gefunden werden. Allerdings sind die Daten nicht in einheitlicher Form vorhanden, sodass die vorhandenen Daten manuell sortiert werden müssen.

[6] (2008). HICSS-41 Call for Papers. Retrieved January 27, 2014, from http://www.hicss.hawaii.edu/hicss_41/cfp_41.htm

Da nicht alle Dateien tatsächlich Paper der HICSS sondern lediglich Einleitungen zu den Tracks sind, sollten diese von der Analyse ausgeschlossen werden.
Die Bereinigung der Daten erfolgt sowohl für die Paper als auch für die einzelnen Call for Papers.

Die einzelnen Call for Papers sind in mehreren Schritten zu bearbeitet:

- Extraktion der Daten in eine Textdatei
- Zuordnung der Call for Papers zu ihren jeweiligen Beiträgen
- Extrahierung von Keywords aus den Call for Papers
- Zusammenführen der Keywords der CfPs mit TrackID in Excel

Um die Paper und deren Textabschnitte sinnvoll verarbeiten zu können wird das Programm LA-PDFText[7,8] verwendet. Dieses Werkzeug ist ein Open Source Programm, welches eine zuvor definierte Regelbasis nutzt um den Text eines PDF-Dokuments und dessen einzelne Abschnitte, sogenannte Chunks, beziehungsweise Kapitel (Abstract, Body, Conclusion, etc.) zu erkennen und zu extrahieren.
Zunächst wurden mit LA-PDFText und dessen Werkzeug "debugChunkFeatures" die Textabschnitte der Dokumente grafisch in einer separaten Bilddatei je Seite ausgegeben. Zudem erstellt LA-PDFText eine Datei mit den Analyseergebnissen, worauf man anschließend, die Erstellung der Regeln aufbauen kann.
Dabei treten jedoch Fehler auf, da die Ordnerstruktur von dem Werkzeug nicht korrekt verarbeitet werden kann, sodass die Datenstruktur umgestellt wir. Die Dateien erhalten als führende Nummer die ID der jeweiligen Tracks. und werden alle in einen Ordner konsolidiert.
Nachdem die relevanten Regeln in der Datei "HICSS.drl" definiert sind, gibt das LA-PDFText-Werkzeug "blockifyClassify" einzelne XML-Dateien aus, welche zur Weiterverarbeitung in dem Data Mining Werkzeug RapidMiner genutzt werden können.
Doch ein weiteres Problem wird festgestellt. LA-PDF-Text in der vorliegenden Version gibt die Dokumente nicht vollständig als Chunk-Sections in XML aus. Eine Anpassung musste erfolgen, sodass alle definierten Chunk-Typen ausgegeben werden.

---

7 (2013). BMKEG/lapdftext · GitHub. Retrieved January 27, 2014, from https://github.com/BMKEG/lapdftext.

8 Ramakrishnan, C., Patnia, A., Hovy, E. H., & Burns, G. A. (2012). Layout-aware text extraction from full-text PDF of scientific articles. *Source code for biology and medicine*, *7*(1).

Ein zusätzliches Problem besteht darin, dass nicht klassifizierte Chunks den Body Elementen hinzugefügt werden. Dadurch werden zum Beispiel nicht korrekt erkannte Chunks (beispielsweise die Autoren) dem Body hinzugefügt. Durch die Text-Verarbeitung mittel Data Mining Verfahren lässt sich dieser Umstand jedoch zunächst ignorieren.

Und es besteht ein weiteres Problem, da LA-PDFText die Chunks nicht in einer konsistenten Weise ausgibt. So werden Body-Chunks als separate Sections mit einzelnen Absätze (<p>) ausgegeben; die anderen Chunk-Arten jedoch gemeinsam in einer Section ohne dedizierte Absätze. Dies muss in der Modellierungsphase berücksichtigt werden.

Insgesamt werden nicht alle 243 Dokumente fehlerfrei durch LA-PDFText analysiert. Zusätzlich zu den oben genannten und zum Teil gelösten Problemen werden nur 192 Dokumente von 243 insgesamt erfolgreich verarbeitet (79,012 %).

Nachdem nun eigentlich die Modeling Phase begonnen werden konnte, taucht bei der Verarbeitung ein weiteres Problem auf - die Kodierung der XML-Dateien, die XML-Kodierungs-Deklaration und darin enthaltener aus den Dokumenten extrahierten Zeichen stimmen nicht überein. Dies führt im RapidMiner zu Problemen, da die Inhalte nicht verarbeitet werden können. Die Analyse des Problems zeigt, dass durch LA-PDFText ANSI-Zeichen extrahiert, jedoch nicht umgewandelt werden. Da die XML-Deklaration jedoch mit UTF-8 angegeben ist, führt dies zu einer inkonsistenten Kodierungsbezeichnung, wodurch RapidMiner die Dateien nicht verarbeiten kann. Es werden zwei Lösungswege implementiert. Zunächst eine automatisierte Konvertierung mittels notepad++ und einem Python Script, welches zunächst die tatsächlichen Kodierung (ANSI) setzt und anschließend die Dateien in UTF-8 ohne BOM (Byte Order Mark) konvertiert. Die andere Variante ist mit zwei Kommandozeilen-Werkzeugen (Konvertierung: cscvt, BOM: RemoveBOM) realisiert und lässt sich somit im RapidMiner direkt einbinden.

Schließlich werden ca. 80% der Paper durch LA-PDFText analysiert und stehen als XML-Datei der Modellierung zur Verfügung.

Die Calls for Papers liegen vorbearbeitet als Excel-Datei vor und können für den Abgleich herangezogen werden

4 Modeling

Artificial Intelligence Methoden

- Text Mining

- aus großen Textbeständen explizites Wissen gewinnen
- Information Retrieval
  - Textstatistik
    - WDF
      - Within-document Frequency
    - IDF
      - Inverse Document Frequency
    - TF-IDF
      - Term Frequency–Inverse Document Frequency
      - benötigt einen Corpus, also mehrere "Dokumente" oder Textabschnitte über die der IDF gebildet werden kann
    - Word Co-occurence Statistical Information[9]
- Information Extraction
  - Text Extraction
    - Layout-Aware Text Extraction
  - Terminology Extraction
    - Automatic Keyword Extraction
    - mit Hilfe von Textvektoren auf Basis der Textstatistik
- Natural Language Processing
    - Tokenization
    - Stopwords

Zur Modellierung wird sowohl LA-PDFText (regelbasiert DROOLS) und für das Text Mining RapidMiner Studio 6 verwendet. Dieses Tool ist ein weit verbreitetes und in der Wissenschaft akzeptiertes Modellierungswerkzeug, welches mit vielen Datentypen vertraut und bereits Bausteine für die zu verwendenden Text Mining Verfahren bietet.
RapidMiner kann darüber hinaus um eigene Erweiterungen ergänzt werden, falls man die Analyse-Methoden später verfeinern oder ergänzen möchte.
Zusätzlich wurde für das Text Mining das Programm KNIME genutzt, welches ursprünglich an der Universität Konstanz entwickelt wurde. Mit diesem Programm kann die Keywordsuche vereinfacht werden, da bereits ein eigener Operator hierfür vorhanden ist. Ein weiterer Vorteil dieses Programmes ist, dass bei Änderungen nur die betroffenen Knoten neu ausgeführt werden, während die Ergebnisse der anderen Knoten zwischengespeichert werden. Allerdings ist die Bedienung nicht so intuitiv wie beim RapidMiner 6.
Zunächst wurde mit dem Read XML und Extract Information gearbeitet, welches je-

[9] Matsuo, Y., & Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, *13*(01), 157-169.

doch nicht zu den gewünschten Ergebnissen führte. So fügte Extract Information die extrahierten Informationen als Meta-Daten hinzu. Doch diese ließen sich nicht ohne weiteres weiterverwenden. Zudem stellten in diesem Fall die einzelnen Absätze bei den Body-Elementen ein Problem dar, da man diese nicht gemeinsam ohne weiteres extrahieren konnte.

Ein Problem, das bei der Verarbeitung mit dem RapidMiner Studio auftaucht ist, dass bei der Angabe von Verzeichnissen RapidMiner den Wert ".null" hinter den ausgewählten Verzeichnis-Namen hinzufügt und die Verarbeitung dadurch fehlschlägt.

Schlussendlich werden die Daten per "Loop Files" in den RapidMiner geladen. Dieser iteriert über alle Dateien eines Verzeichnisses die gewünschten Operationen. Nachdem die XML-Datei als Dokument eingelesen wurde, wird es zunächst in seine Bestandteile zerschnitten, sodass Abstract, Introduction, Body und Conclusion aus dem Dokument herausgefiltert werden. Dies geschieht über XPath-Abfragen, die die einzelnen Abschnitte an ihren Pfaden erkennen. Nach der Aufteilung werden die einzelnen Dokumente eines Artikels in einer Sammlung zusammengefasst. Für den Body gilt dies doppelt, da zunächst die einzelnen Body-Elemente zusammengeführt werden müssen, um diese dann mit den anderen Abschnitten zu verbinden.

Anschließend finden zwei Prozesse statt, in denen die Dokumente in ExampleSets umgewandelt werden. Hierbei werden einmal die TF-Methode und einmal die TF-IDF-Methode zur Attributsausgabe bzw. zum Ähnlichkeitsmaß ausgegeben. In beiden Versionen werden die Dokumente jedoch in Tokens (Wörter) aufgeteilt und nach einer Stoppwortanalyse auch auf ihren Wortstamm reduziert und in Kleinbuchstaben umgewandelt. Zusätzlich wurde die Möglichkeit eingebaut sogenannte n-grams herzustellen, also Wörter, die im Text zusammenstehen, und die Häufigkeit dieser mit auszugeben. Außerdem werden durch diesen Prozess nur die Top 15% der Wörter ausgegeben.

Es werden nun noch die ID des zugehörigen Tracks als Attribut ausgelesen, um ein späteres Zuordnen zu den Calls for Papers zu ermöglichen.

Da die Wörter momentan Attribute sind, sie jedoch besser als Liste auszuwerten sind, werden beide ExampleSets transponiert, sodass jeweils eine Liste mit den Stichwörtern und den jeweiligen TF- bzw. TF-IDF-Attributen entsteht. Leider sind einige Attribute, wie die Metadaten auch transponiert, was die Liste etwas unangenehm macht.

Bei dem ExampleSet der Term Frequency werden die Metadaten deshalb zunächst entfernt, um Rechenoperationen durchführen zu können. Dies geschieht mit der Aggregation der TF-Attribute der jeweiligen Abschnitte. Da sich die Term-Frequency aus der

Häufigkeit eines Terms und der Gesamthäufigkeit aller Terme eines Dokumentes berechnet, sind die Anteile der einzelnen Abschnitte normalisiert und eine hohe aggregierte Term-Frequency lässt darauf schließen, dass der Term in allen Abschnitten häufig vertreten ist. Deshalb wird die Liste auch nach diesem Attribut sortiert.

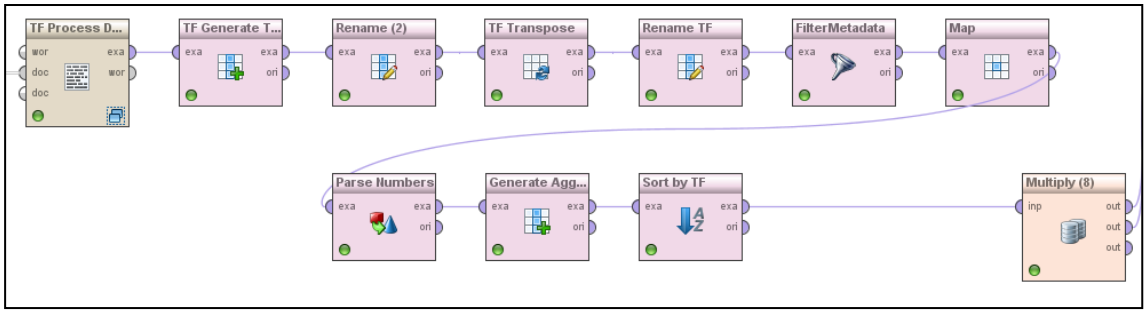

Abb. 8: TF-Prozess

(Quelle: Eigene Darstellung)

Da eine Aggregation bei TF-IDF keinen Sinn macht, werden nach diesem Prozess nur irrelevante Metadaten herausgefiltert. Zusätzlich wird aus diesem Prozess auch die WordList, in welcher die Gesamthäufigkeit eines Terms in einem Artikel extrahierbar ist, ausgegeben.

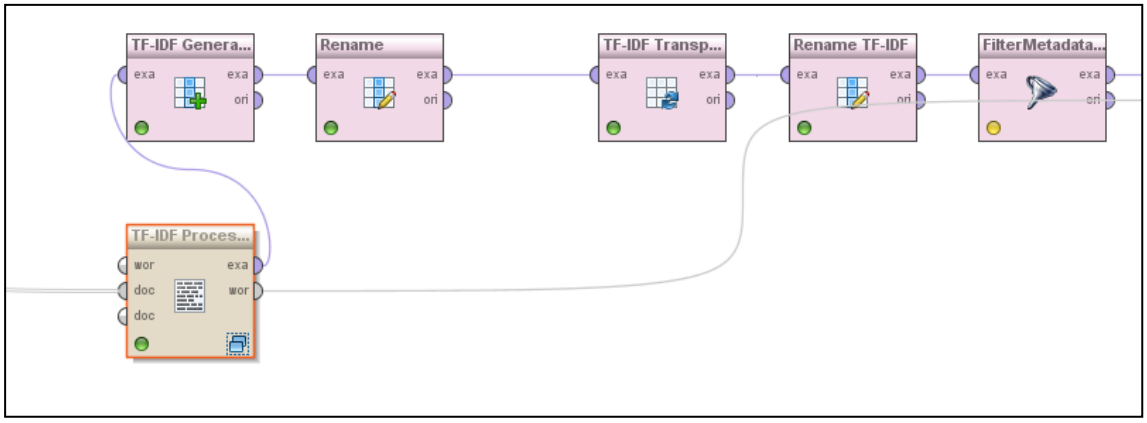

Abb. 9: TF-IDF-Prozess

(Quelle: Eigene Darstellung)

Anschließend werden die ExampleSets von TF und TF-IDF gejoined. Da im ExampleSet der Term Frequency keine Metadaten mehr vorhanden sind, wird ein outer join durchgeführt.

Die Wordlist, welche sortiert und auf die häufigsten 40 Terme reduziert wurde, wird nun mit diesem Gesamtdokument gejoined, um eine Gesamtübersicht über verschiedene Maße der wichtigsten Terme zu haben und somit eine Auswahl treffen zu können. Da die Maße des TF-TF-IDF-ExampleSets wichtiger sind, wird auch nach dessen Wortliste

gejoined, sodass somit nur die Gesamthäufigkeiten der WordList sowie die Anzahl der Dokumente, in denen der Term auftritt, angehängt werden.
Da dieser Gesamtprozess als CSV-Datei ausgegeben und über alle Artikel iteriert wird, entsteht eine Liste mit allen Artikeln und ihren wichtigsten Termen.

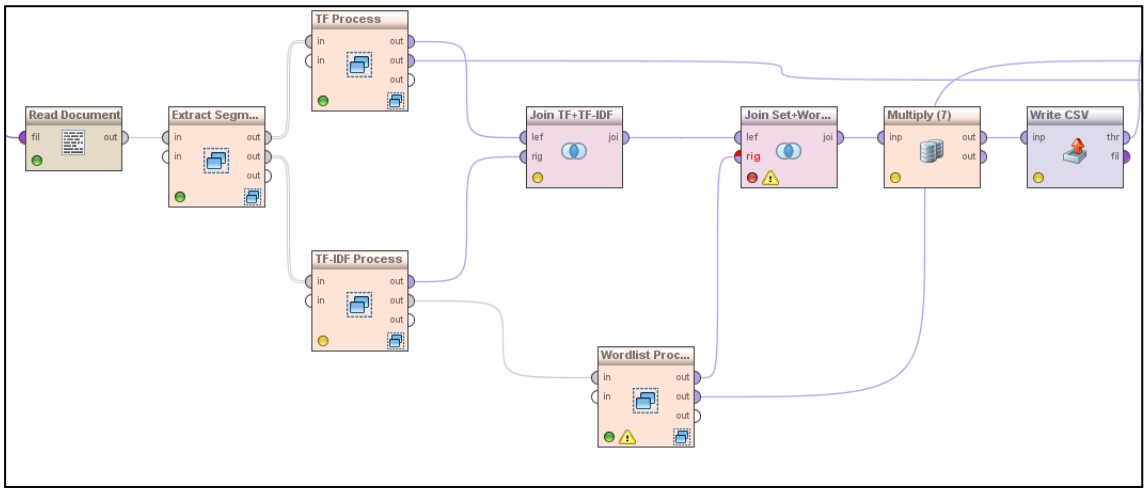

Abb. 10: Loop-Prozess
(Quelle: Eigene Darstellung)

Die CSV-Daten werden anschließend in einer Excel Datei über ein Makro verarbeitet.
Das Makro ist in Visual Basic for Applications geschrieben, beinhaltet bisher jedoch nur rudimentär die Funktionalität um einen Datensatz am Anfang der CSV-Datei einfach gegenüber dem Call for Paper-Text zu verifizieren.
Hier wäre noch weitere Entwicklung nötig um alle Dokumente zu verarbeiten.
Anschließend kann der Vergleichswert Ratio = “Anzahl der im CFP vorhanden Keywords” / “Anzahl der ausgewählten Keywords” gebildet werden und grafisch je Track-ID dargestellt werden.

5 Evaluation

Relevante Keywords sind auf Basis von statistischen Merkmalen identifiziert worden. Da eine Auswahl allein aufgrund statistischer jedoch fahrlässig wäre, sollte diese Liste den jeweiligen Autoren oder wissenschaftlichen Mitarbeitern vorgelegt werden um eine schnelle Auswahl treffen zu können.
Wenn man eine bestimmte Anzahl von Personen einen Text in Stichwörtern zusammenfassen lässt kommen sehr wahrscheinlich jeweils unterschiedliche Ergebnisse heraus. Wenn man zudem Personen diesen Vorgang durchführen ließe, die zum Teil über kon-

textuelles Wissen der untersuchten Texte verfügen und zum Teil nicht, würden wiederum anderen Ergebnisse erzeugt werden.

Computern fehlt die Möglichkeit Texte sinnvoll zu aggregieren ohne auf Algorithmen und statistische Methoden zurückzugreifen. Die Stärken des Computers liegen jedoch in der Verarbeitung einfacher und klar formalisierter Probleme und kommt in diesen Fällen schneller zu Ergebnissen. Ebenso können Vergleiche auf Wortebene schnell durchgeführt werden.

Also: → Analyse mit mehreren Verfahren und anschließendem Vergleich auf Gemeinsamkeiten.

- Kritische Beleuchtung von Schwachstellen und weitere Schritte:
    - eine semantische Überprüfung mit ggf. Synonymabgleich wäre sinnvoll
    - kontextuelles Wissen fehlt - aufgrund des Aufwands für die Erstellung eine Kontext-bezogenes Lexikon und des somit möglichen Abgleichs
    - Teilprogramme können nicht sinnvoll mit dem RapidMiner genutzt werden bzw. sind nicht für große Datenmengen ausgelegt
    - Call for Papers beinhaltet möglicherweise nicht die relevanten Keywords um eine aussagekräftige Analyse durchzuführen

6 Deployment

Der RapidMiner Prozess wurde um die Programme zur Anpassung der Kodierung ergänzt, sodass kein manueller Aufruf der Konsole nötig ist.

LA-PDFText sollte angepasst werden um

- die Daten Zeichensatzkonform auszugeben und
- die Daten in einer konsistenten XML-Struktur auszugeben, um somit die Verwendung mittels XPath zu vereinfachen.

7 Literatur

(2013). BMKEG/lapdftext · GitHub. Retrieved January 27, 2014, from https://github.com/BMKEG/lapdftext.

(2008). HICSS-41 Highlights. Retrieved January 14, 2014, from http://www.hicss.hawaii.edu/hicss_41/41highlights.htm.

(2008). HICSS-41 Call for Papers. Retrieved January 27, 2014, fromhttp://www.hicss.hawaii.edu/hicss_41/cfp_41.htm

Kroiß, A. (2010). Computerunterstützte Tagging-Verfahren für Dokumente im Web.

Matsuo, Y., & Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, *13*(01), 157-169.

Menaka, S. & Radha, N. (2013). Text Classification using Keyword Extraction Technique, *3*(12), 734–740.

Murfi, H. (2010). Machine Learning for Text Indexing.

Ramakrishnan, C., Patnia, A., Hovy, E. H., & Burns, G. A. (2012). Layout-aware text extraction from full-text PDF of scientific articles. *Source code for biology and medicine*, *7*(1).

(2007). RapidMiner - Predictive Analytics, Data Mining, Self-service, open ... Retrieved January 27, 2014, from http://rapidminer.com/.

Rich, E. (1983). Artificial Intelligence.

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2004). CO RI Automatic keyword extraction.

Uramoto, N., Matsuzawa, H., Nagano, T., Murakami, a., Takeuchi, H., & Takeda, K. (2004). A text-mining system for knowledge discovery from biomedical documents. *IBM Systems Journal*, *43*(3), 516–533. doi:10.1147/sj.433.0516

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*.

Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research*, *15*(1), 37–50. doi:10.1016/j.hitech.2003.09.003

Zhang, C., Wang, H., Liu, Y., Wu, D., Liao, Y., & Wang, B. (2008). Automatic Keyword Extraction from Documents Using Conditional Random Fields, 3.

Generierung einer Tag Cloud für wissenschaftliche Publikationen auf Basis von Keywords unter Beachtung der referenzierten Publikationen

Arne Karhof, Ali Farhat

Abstract. *Im Rahmen des KI-Praktikums soll in diesem Artikel die Möglichkeit untersucht werden, mithilfe des RapidMiner Studios 6 möglichst automatisiert eine Tag Cloud für wissenschaftliche Artikel zu erstellen. Dabei werden in einem ersten Schritt die "Keywords" sowie das Literaturverzeichnis eines Artikels automatisiert ermittelt. Die Keywords der im Literaturverzeichnis genannten Artikel werden ebenfalls ausgelesen. Während die Keywords des "Ursprungsartikels" die Mitte der Tag Cloud bilden, entscheidet die Häufigkeit des Vorkommens der verlinkten Keywords über deren Position und Gestaltung der Tag Cloud.*

1 Business Understanding

Das vordergründige Ziel dieses Beitrages soll es sein, das Ausmaß der Korrelation eines wissenschaftlichen Beitrages mit seinen referenzierten Artikel festzustellen. Dieses Ausmaß soll als Ergebnis einer Textanalyse visuell in Form einer Tag Cloud (vgl. Linderman, 2004) dargestellt werden. Als Faktor, der in diesem Beitrag als Kennzeichen zwischen „korreliert" beziehungsweise „nicht korreliert" entscheidet, wurden die sogenannten „Keywords" eines wissenschaftlichen Artikels gewählt. Diese Schlüsselwörter werden von den Autoren gewählt um die hauptsächlich behandelten Themengebiete ihrer Beiträge in knapper Form widergeben zu können. Nach eingehender Recherche ist bisher kein solches Verfahren bekannt. Es konnten einzig Artikel ausgemacht werden, die sich beispielsweise mit dem allgemeinen Netzwerk von referenzierter Literatur untereinander (vgl. Hargens, 2000), mit der Beziehung von Artikellänge und Literaturliste (vgl. Abt & Garfield, 2002) oder mit Korrelation zwischen Literaturliste und Zitationen im Text (vgl. Alimohammadi & Sajjadi, 2009) beschäftigen. Daher wird im Folgenden das Projektvorgehen ohne bestehende Anforderungen bestimmt und durchgeführt, auch eventuelle Problematiken bleiben somit unbekannt. Im Folgenden soll kurz erörtert werden, welche Schritte unternommen werden, um das Projektziel zu erreichen.

In einem ersten Schritt werden die Schlüsselwörter eines wissenschaftlichen Artikels, im Folgenden als Ursprungsartikel bezeichnet, ermittelt. Zudem wird das Literaturverzeichnis des Ursprungartikels in geeigneter Form ausgelesen, um die referenzierte Literatur identifizieren zu können. Dann wird mittels einer Datenbankabfrage bei einer großen Literaturdatenbank, in unserem Falle „Google Scholar“, möglichst automatisiert der entsprechende Link zum Publisher beziehungsweise Verlag ermittelt. Mit Hilfe dieses Links werden die Quelltextinformationen der referenzierten Veröffentlichung ausgelesen. Stellt der Verlag auf seiner Webseite Informationen bezüglich der Keywords eines Artikels zur Verfügung, können diese mit Unterstützung des RapidMiner Studios[10] ausgelesen und dem Ursprungsartikel zugeordnet werden. In einem finalen Schritt sollten somit sowohl die Keywords des Ursprungsartikels als auch die Keywords der referenzierten Artikel mit Hilfe eines Online–Tools als Tag Cloud dargestellt werden können.

2 Data Understanding

Als Grundlage für unsere Untersuchung stehen uns die Veröffentlichung des „Strategic Management Journals“ zur Verfügung. Dabei beschränken wir uns auf die Jahre 2008, 2009 sowie 2010. Insgesamt handelt es sich um einen Datenpool von 221 Dokumenten, die als PDF–Dateien vorliegen. Im Folgenden soll ein Eindruck der Struktur und des Aufbaus der PDF–Dateien vermittelt werden. Dazu werden in einem ersten Schritt die Dokumente manuell gesichtet und anschließend mit Hilfe des Kommandozeilenbasierten Tools LA-PDFText auch die Metadatenebene behandelt.

2.1 Manuelle Sichtung der Daten

Gegeben durch die Vorgaben der „Strategic Management Society“ (vgl. Strategic Management Society) sind alle Artikel strukturell gleich aufgebaut. Auf der Titelseite der Artikel wird neben einer knappen Widergabe wichtiger Journalinformationen der Titel, hervorgehoben durch eine größere Schriftart, die Autoren, als Kapitälchen, sowie kurze Informationen zu den jeweiligen Autoren hinterlegt. Darauf folgt der Abstract, der rechtsbündig und kursiv dargestellt wird. Dem Abstract folgt der eigentliche Text des Artikels, dieser wird, bis auf einige Ausnahmen wie Formeln, Zitate oder Tabellen- und Abbildungsbezeichnungen durchgehend einheitlich dargestellt. Ab Beginn des eigentlichen Textes wird das Dokument vertikal halbiert, dies zieht sich bis zu einem eventuell

[10] Vgl. http://rapidminer.com/products/rapidminer-studio/, abgerufen am 27.01.2014

vorkommenden Anhang durch das vollständige Dokument. Weiterhin interessant ist für die Intention unseres Beitrages, dass die Keywords des Artikels immer auf der ersten Seite platziert werden. Die Keywords befinden sich dabei auf der linken Dokumentenhälfte und stehen häufig direkt oberhalb eines Textfeldes, der die Korrespondenz mit einem Mitautor näher beschreibt. Da hier keine klare Trennung zwischen Keywords und Korrespondenzangaben besteht, könnte sich dies als negativer Aufwandstreiber für unser weiteres Vorgehen herausstellen. Der weitere Aspekt, der für unsere spätere Bearbeitung noch relevant werden wird, ist das Literaturverzeichnis, das in dem uns vorliegendes Journal mit der Kapitelüberschrift „References" eingeleitet wird. Auch die Literaturliste wird vertikal getrennt dargestellt, einzelne Beiträge folgen Folgender Struktur: „*Autor1, Autor2. Jahr. Titel. Journal/Buch/etc."*. Die Schriftgröße ist unwesentlich kleiner als der Rest des Artikels, dies soll für unsere weitere Bearbeitung allerdings nicht von Relevanz sein. Zu nennen wären noch die Seitenzahlen. Diese ziehen sich durch das gesamte Dokument, eine Ausnahme bildet hier nur die Titelseite. Auf jeder Seite wechselt dabei die Ausrichtung der Seitenzahl, immer in der Kopfzeile des Dokuments, jeweils von links- auf rechtbündig bzw. umgekehrt. Weiterhin befindet sich in jeder Fußzeile ein Copyright-Vermerk. Die Fuß- sowie Kopfzeilen könnten insofern für die spätere Bearbeitung relevant sein, als dass sie vom rein strukturellen Fluss das Literaturverzeichnis unterbrechen.

2.2 Metadaten-Ebene

Durch eine Analyse der PDF Dateien mittels des Tool LA-PDFText ließen sich wichtige Erkenntnisse für die weitere Bearbeitung über die Qualität der Daten und der Struktur gewinnen. Dabei wurden für einige, zufällig ausgewählte Dokumente alle Funktionen von LA-PDFText genutzt, um möglichst viele Informationen gewinnen zu können. Da bei der manuellen Sichtung bereits festgestellt wurde, dass alle Dokumente denselben Aufbau besitzen, wurde auf eine komplette Analyse aller Dokumente mit jeweils allen Funktionen verzichtet. Unter Berücksichtigung des Umfangs soll im Folgenden nur auf die für unsere Bearbeitung relevanten Erkenntnisse eingegangen werden.

Durch Nutzen der Softwarefunktionen von LA-PDFText wurden verschiedene Dokumente in einem ersten Schritt ohne weitere Modifikationen durch Regeldateien oder Ähnliches als Bilddateien widergespiegelt, einmal als sogenannte „Block Images", ein anderes Mal als „*Section Images*". Durch Untersuchung der *Section Images* konnte festgestellt werden, dass die Annahme aus der manuellen Sichtung korrekt war, nämlich

dass keine klare Trennung zwischen den Keywords und dem Korrespondenzblock besteht, sodass an dieser Stelle später weitere Modifikationen vonnöten sein werden. Positiv zu nennen ist, dass das Literaturverzeichnis bereits ohne Veränderungen korrekt erkannt wird. Selbstverständlich bezeichnet LA-PDFText dieses noch nicht korrekterweise als Literaturverzeichnis, allerdings wird es bereits als zusammenhängende Entität verstanden. Die Möglichkeit, dass die Fuß- bzw. Kopfzeile eventuell den Textfluss stören könnte, bewahrheitet sich nicht. Die Seitenzahlen, der Copyright-Vermerkt sowie andere Angaben werden korrekterweise als nicht zum Textfluss zugehörigen Informationen erkannt. Die durch die *Section Images* gewonnen Informationen lassen sich auch nochmals als XML–Datei extrahieren, an dieser Stelle wird dann auch Text mit abgebildet. Dabei wird der gesamte Text, der beispielsweise der Klasse „Footer“ zugeordnet werden konnte, im entsprechenden XML–Kindeselement gespeichert. Dabei kam es allerdings ohne weitere Modifikation durch LA-PDFText zu Brüchen hinsichtlich der Literaturliste. So wurde diese, obwohl vorher durch die *Section Images* korrekt zugeordnet, teilweise gerade an Stellen mit Seitenumbrüchen verschiedenen XML–Kindeselementen unterstellt, sodass für eine weitere Bearbeitung eine Modifikation der ursprünglichen Konfiguration von LA-PDFText vonnöten werden würde.

2.3 Datenqualität

Insgesamt schätzen wir die Güte der Daten als sehr gut ein, vor allem da die Dokumente durch die strenge Regulierung seitens der Strategic Management Society klar und beinahe durchgehen gleich Strukturiert sind. Durch die manuelle Sichtung sowie der Bearbeitung der Daten mit Hilfe von LA-PDFText konnten nur wenige Defizite ausgemacht werden. Zu nennen seien an dieser Stelle einige PDF–Dokumente, die offensichtlich fehlerhaft dem Quellverzeichnis zugeordnet wurden und dadurch nicht der eigentlichen strukturellen Konvention entsprachen, u.a. mehrere Errata sowie ein fehlerhaft Veröffentlichter Beitrag. Eben dieser wurde durch Veröffentlichung ein ebenfalls in den Quelldateien vorkommendes Erratum konstatiert. Da diese Dokumente aber aufgrund des Ausmaßes ihrer Fehlerhaftigkeit bereits schnell als solche erkannt werden konnten und im Verhältnis betrachtet selten auftraten, wird die Datenqualität an dieser Stelle insgesamt immer noch als hoch betrachtet.

3 Data Preparation

Durch das bereits oben genannte seltene Auftreten von fehlerhaften Dokumenten konnten diese manuell aus der Datenquelle entfernt werden. Da ansonsten die Struktur der Artikel durchgehenden gleichbleibend ist und alle Artikel sowohl Keywords als auch ein Literaturverzeichnis enthalten, wurden alle nicht fehlerhaften Artikel als Datengrundlage ausgewählt. Im Zuge der Modellierungsphase wurde allerdings, in Anbetracht der Laufzeit verschiedener Prozesse, eine zufällig gewählte Teilmenge der Quelldaten gewählt, erst im abschließenden Schritt wird der Prozess an allen Datensätzen durchgeführt.

Wie in Kapitel zwei erörtert, besteht ein Problem hinsichtlich der Zuordnung der Literaturliste zu unterschiedlichen XML–Kindeselementen. Zur Lösung dieser Problematik bietet LA-PDFText die Zuordnung von Regeldateien zum Kommandozeilenbefehl an, die im weiteren Vorgehen genutzt wurde. Insgesamt stützt sich die Bearbeitung des Vorgehens an der Funktion *blockifyClassify.exe*, die openAccess–kompatible XML Dateien erstellt. Dadurch konnte erreicht werden, dass sowohl die Keywords, als auch die referenzierte Literatur vollständig in einem Kindeselement einer dem Dokument zugeordneten XML–Datei hinterlegt werden konnten. Leider gab es keine Lösung, auch nur diese Informationen zu hinterlegen, sodass z.T. weitere unnötige Textangaben mit in das Kindeselement einflossen, wie beispielsweise der oben genannte Korrespondenzblock. Diesem Problem wurde während der Modellierungsphase begegnet, in der mit Hilfe regulärer Ausdrücke und Iterationen über den Textblock die benötigten Informationen extrahiert wurden.

Durch die gewonnenen Vorteile der erstellten XML Dateien bilden diese nun für die weiteren Schritte unsere Datenquelle, sodass eine Bearbeitung der PDF Dateien nicht stattfinden muss. Somit besteht unser finales Datenset aus insgesamt 212 XML Dateien, bestehend aus zwei Kindeselementen, wobei letzteres die für uns relevanten Informationen bereithält.

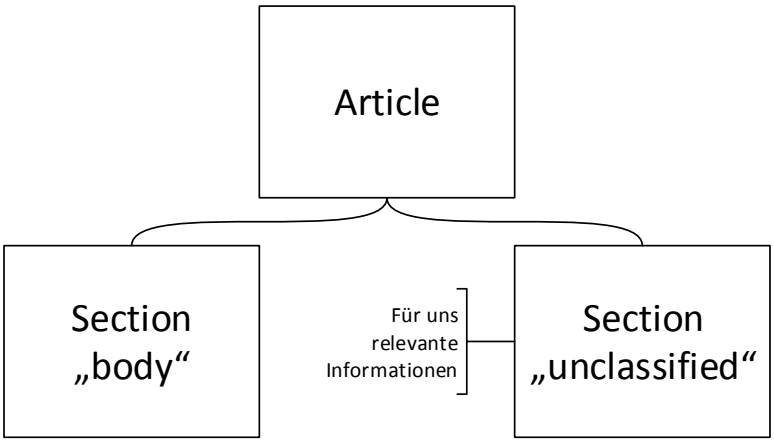

Abb. 11: Aufbau einer XML Datei der Datenquelle

(Quelle: Eigene Darstellung)

4 Modeling

Die gesamte Modellierungsphase wurde mittels des RapidMiner Studio durchgeführt. Dabei wurden sieben (Teil)Modelle erstellt, um das gewünschte Resultat zu erzielen. Drei Teilprozesse beschränken sich auf das Auslesen der Quellverzeichnisse, zwei Teilprozesse erstellen die benötigten Google Scholar URLs und wiederrum zwei Teilprozesse lesen die Quelldaten von Online-Informationsdienste für wissenschaftliche Publikation aus, um die Keywords der referenzierten Artikel zu extrahieren. Im Folgenden wird der Aufbau der drei unterschiedlichen Kategorien erläutert.

4.1 Auslesen der Quellverzeichnisse

Neben dem Auslesen der Quellverzeichnisse wird in diesem Teilschritt außerdem eine weitere Datenbereinigung durchgeführt, die wie in Abschnitt drei bereits erläutert während der Modellierungsphase vorgenommen werden musste.

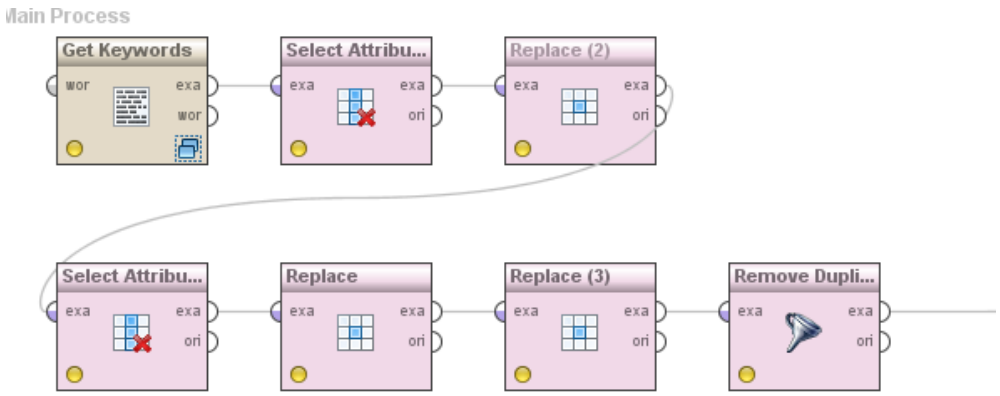

Abb. 12: Lesen der Quelldateien und Datenbereinigung
(Quelle: Eigene Darstellung)

Im ersten Teilschritt wird über den Quellordner iteriert, sodass jedes XML Dokument dem Example Set hinzugefügt werden kann. Bereits im „*Get Keywords*“ Operator werden dabei erste nötige Datenbereinigungen durchgeführt und die Keywords ausgelesen. Alle weiteren Prozessschritte führen entweder weitere Bereinigungen durch (beispielsweise das Entfernen von Umbrüchen, die durch das Token „*\n*“ in den Daten hinterlegt sind), oder beseitigen weitere unerwünschte Teildatensätze wie den Korrespondenzblock. Anschließend an den abgebildeten Prozess wird die Literaturliste extrahiert. Auch hier sind dann im dritten Teilprozess wiederrum Modifikationen notwendig, sodass in der später resultierenden Zieldatei jeder referenzierter Artikel in einer eigenen Spalte gespeichert wird. Dabei wird der ursprünglichen Zitationsform „*Autor1, Autor2. Jahr. Titel. Journal/Buch/etc.*“ gefolgt, wobei der letzte Abschnitt „*Journal/Buch/etc.*“ entfernt wurde, da bereits mit den ersten drei Angaben zufriedenstellende Suchergebnisse bei Google Scholar getroffen werden konnten. Das Resultat dieses Teilprozesses ist eine Excel Datei, die nun alle Keywords, den Titel des Ursprungsartikels und seine jeweiligen Referenzen Zeilenweise festhält.

4.2 Erstellen der Google Scholar Links

Für unseren weiteren Schritt ist es notwendig, aus den extrahierten referenzierten Artikeln gültige URLs zu erstellen, die folgende Form besitzen: „*http://scholar.google.de/scholar?hl=en&q=Querypart1+Querypart2*“.

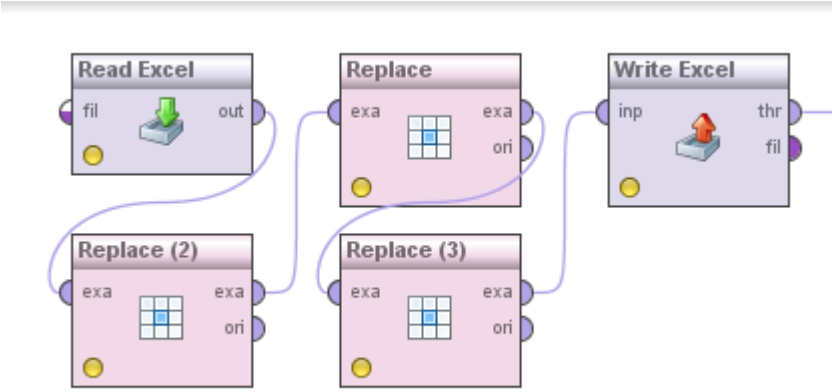

Abb. 13: Erstellen der Google Scholar URLs
(Quelle: Eigene Darstellung)

Auch hier finden wieder Textmodifikationen innerhalb der Attribute statt, sodass beispielsweise aus dem ursprünglichen Datensatz „*Baye MR Crocker KJ Ju J 1996 Divisionalization franchising and divestiture incentives in oligopoly“* die valide URL *„http://scholar.google.de/scholar?hl=en&q=Baye+MR+Crocker+KJ+Ju+J+1996+Di Divisionalizati+franchising+and+divestiture+incentives+in+oligopoly“* entsteht. Eine Überprüfung der URL in einem Webbrowser zeigt als ersten Suchtreffer auch bereits den gewünschten Artikel an. In einem zweiten Schritt sollten nun die erstellten URLs iterativ die Funktion „*Crawl Web*“, die RapidMiner mittels eines Plug-Ins bereitstellt, auslösen, sodass durch Untersuchung der Linkstruktur der Link auf den entsprechenden Onlinedienst für wissenschaftliche Arbeiten extrahiert werden kann. Leider konnte der Web Crawler des RapidMiners keinen zufriedenstellenden Daten liefern. Während die Crawlingfunktion, beispielhaft an anderen Webseiten implementiert, die gewünschten Quelldaten lieferte, lieferte der RapidMiner trotz erfolgreichen Durchlaufens keine Daten. Um trotzdem das gewünschte Resultat zu erzielen, wurde eine eigene Web Crawling Implementierung mittels Java durchgeführt sowie zwei proprietäre Programme verwendet. Leider konnte keine Vorgehensweise die gewünschten Google Scholar Ergebnislinks extrahieren. Obwohl an dieser Stelle keine zufriedenstellende Lösung gefunden wurde, wird dennoch die weitere Modellierung beschrieben, da wir an dieser Stelle das eigentlich automatisierte Vorgehen exemplarisch anhand eines PDF Dokuments durchgeführt haben.

4.3 Auslesen der Keywords von Onlinediensten

Da der Teilschritt zwei nicht zufriedenstellend umgesetzt werden konnte, werden nun einige Schritte manuell vorgenommen, um das eigentlich erstrebte Vorgehen zu simulieren.

Durch die erfolgreiche Extrahierung der Autoren, des Jahres und des Titels können wir diese Informationen nun manuell der Excel Datei entnehmen und selbstständig bei Google Scholar suchen. Wir gehen nun dem ersten Link nach und kopieren die URL. Diese wird mittels des „Get Page“ Operators des RapidMiners als HTML Dokument in einem lokalen Ordner hinterlegt und später, wie in Abb. 14 ersichtlich, je nach Onlineanbieter unterschiedlich bearbeitet. Aufgrund der Vielzahl an unterschiedlichen Online Angeboten zu wissenschaftlichen Publikationen (SpringerLink, ScienceDirect, ACM Digital Library etc.) und deren jeweils unterschiedlicher HTML Struktur wurde an folgender Stelle die Implementierung auf zwei Dienste beschränkt: Der *Wiley Online Library*[11] sowie der *Pubs Online*[12].

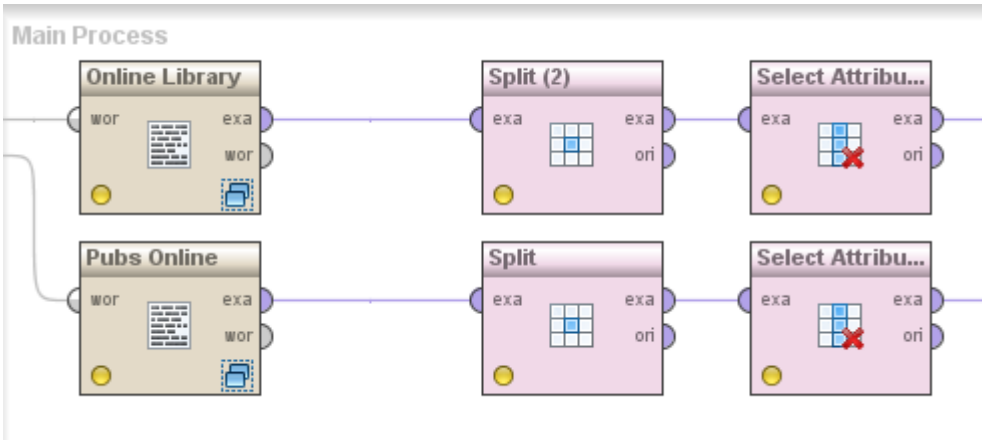

Abb. 14: Extraktion der Keywords von verschiedenen Online Angeboten
(Quelle: Eigene Darstellung)

Als Ergebnis wird eine Liste mit Keywords geliefert, die jeweils zu einem Ursprungsartikel gehören. Diese Liste wird nun, zusammen mit den Keywords des Ursprungsartikels, manuell bei einem Onlinetool zur Erstellung von Tag Clouds[13] eingefügt.

[11] Vgl. http://onlinelibrary.wiley.com/, aufgerufen am 04.02.2014
[12] Vgl. http://pubsonline.informs.org/, aufgerufen am 04.02.2014
[13] Vgl. http://www.wordle.net/, aufgerufen am 04.02.2014

5 Evaluation

In der folgenden Abbildung ist die resultierende Tag Cloud zu sehen, beispielhaft anhand der PDF Datei *„Old technology meets new technology complementarities, similarities, and alliance formation.pdf“*.

Abb. 15: Erstellte Tag Cloud

(Quelle: Eigene Darstellung)

Wie anhand der Schriftgröße zu erkennen ist, sind die hauptsächlichen Schlagworte des Ursprungsartikel sowie der referenzierten Publikation „alliances“, „capabilities“, etc. Da nicht alle Schritte automatisiert umgesetzt werden konnten und nicht alle Onlinedienste für wissenschaftliche Artikel implementiert wurden, stellt Abb. 15 letztendlich das Ergebnis aus den ursprünglichen Keywords und insgesamt 13 referenzierten Artikeln dar. Dadurch kam es eher selten zu einer Mehrfachnennung, sodass bereits eine Häufigkeit von vier Nennungen zu einer relativ großen Schriftart führten („pharmaceutical“). Um ein Ergebnis zu erzielen, dass eventuell repräsentativer ist, müssten noch weitaus mehr Onlinedatenbanken wie SpringerLink oder ScienceDirect implementiert werden, außerdem müsste das Schnittstellenproblem bzgl. Google Scholar gelöst werden.

6 Deployment

Eine vollautomatisierte Implementierung des gesamten Projektvorgehens ist unserer Meinung nach nicht möglich. Eine erste Hürde würden bereits Datenquellen darstellen, die nicht vom Strategic Management Journal bereitgestellt werden. Da gerade das Data Understanding und die Data Preparation viel Zeit in Anspruch nehmen, müsste eventuell eine einfachere Lösung für das extrahieren benötigter Informationen aus wissenschaftlichen Artikeln gefunden werden. Das weitere Vorgehen lässt sich größtenteils vollständig automatisiert implementieren, nur das Schnittstellenproblem zu Google Scholar konnte nicht gelöst werden. Auch das Einfügen der Keywords in einen Online Tag Cloud Generator erfolgt noch manuell, doch durch Einbindung von Datenbanken in der die benötigten Informationen hinterlegt werden sowie eventueller PHP Umsetzungen sollte auch dieser Schritt automatisiert ablaufen können. Erste Ansätze dazu wurden im Rahmen unseres Projektvorgehens bereits realisiert, allerdings stellte dies im Vergleich zu dem Online angebotenen Tool keine befriedigende Ersatzlösung dar, vor allem da das manuelle Hinzufügen der Keywords kaum Aufwand darstellte, die eine PHP Implementierung gerechtfertigt hätte.

7 Literaturverzeichnis

Abt, H. A., & Garfield, E. (2002). Is the Relationship between Numbers of References and Paper Lengths the Same for All Sciences? *Journal of the American Society for Information Science and Technology* (S. 53(13):1106-1112). Journal of the American Society for Information Science and Technology.

Alimohammadi, D., & Sajjadi, M. (25. Juni 2009). Correlation between references and citations. *Webology*.

Hargens, L. L. (2000). Using the Literature: Reference Networks, Reference Contexts, and the Social Structure of Scholarship. *American Sociological Review Vol. 65* (S. 846-865). Washington: American Sociological Association.

Linderman, M. (2. Dezember 2004). *The Spread of Weighted Lists*. Abgerufen am 14. 01 2014 von Signal vs Noise: http://37signals.com/svn/archives/000937.php

Strategic Management Society. (kein Datum). *Author Guidelines*. Abgerufen am 27. 01 2014 von Wiley Online Library: http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1097-0266/homepage/ForAuthors.html

Automatische Überprüfung ausgewählter linguistischer Qualitätsmerkmale in wissenschaftlichen Arbeiten

Jonas Jacobj, Fabian Otte

Abstract. *Das Ziel dieser Arbeit ist es, anhand eines in dieser Ausarbeitung modellierten Tools die linguistische Qualität von wissenschaftlichen Publikationen zu überprüfen. Hierzu wird ein Kriterienkatalog erstellt, der Qualitätsmerkmale wissenschaftlicher Arbeiten beinhaltet. Diese werden durch das zu erstellende Tool überprüft und ausgewertet. Zuletzt soll das Ergebnis der Analyse in Textform ausgegeben werden und die linguistische Qualität der untersuchten Publikation bewertet werden können.*

1 Business Understanding

Heutzutage existieren diverse Bewertungssysteme von Journals, welche die Qualität von wissenschaftlichen Publikationen einschätzen. Es mangelt allerdings an einer Möglichkeit, die linguistische Qualität von selbst erstellten oder im Internet verfügbare wissenschaftliche Arbeiten zu beurteilen und dadurch deren Gesamtqualität und wissenschaftlichen Nutzen einschätzen zu können. Daher soll in dieser Ausarbeitung ein Prototyp eines Analysetools entwickelt werden, welches die linguistische Qualität von wissenschaftlichen Publikationen untersucht. Dafür wird ein Kriterienkatalog aus Qualitätsmerkmalen definiert, die überprüft werden sollen. Die Ergebnisse werden anschließend ausgegeben und als Bewertungsgrundlage herangezogen.

Die Untersuchung der wissenschaftlichen Qualität und damit die Überprüfung der Qualitätsmerkmale erfolgt auf sprachlicher Ebene. Der Aufbau der zu untersuchenden Arbeit wird dementsprechend nicht beachtet. Ebenfalls unbeachtet bleiben Merkmale wie Schriftgröße, Zeilenabstand und Seitenrand, zumal es für diese Eigenschaften keinen allgemeingültigen Standard für wissenschaftliche Ausarbeitungen gibt, der als Bewertungsgrundlage herangezogen werden könnte. Anders verhält es sich bei den linguistischen Mitteln in wissenschaftlichen Arbeiten. So finden sich hier einige allgemeingültige Vorgaben, die zu beachten sind und die Qualität der Arbeit widerspiegeln.

Bei einer umfassenden Recherche im Internet stößt man auf viele Hinweise, Vorschläge, Leitfäden und Regeln, die bei der Erstellung von wissenschaftlichen Arbeiten umge-

setzt werden sollten. So finden sich Anleitungen zum wissenschaftlichen Schreiben beispielsweise in Büchern (Maja Jokic, 2006), Fachzeitschriften (Fachjournalist, 2010) sowie auf Internetseiten verschiedener Universitäten und Fachhochschulen (Fachhochschule Potsdam, 2004), (Universität Oldenburg, 2010), (Universität Augsburg, 2014), (Universität Freiburg), (Universität Jena, 2007), (Universität Dresden, 2006). Da die Quellen einige Grundregeln gemeinsam haben, wurde eine Liste angefertigt, in der jede gefundene Regel aufgenommen wurde. Nach abgeschlossener Quellenanalyse und einer vollständigen Liste von gefundenen Qualitätsmerkmalen wurden diese auf ihre Realisierbarkeit hin überprüft. Da die Analyse einiger Qualitätsmerkmale, wie beispielsweise die Überprüfung auf korrekte Ergänzung der hochsprachlichen Idiome, sich als zu komplex gestaltete, wurde eine Auswahl getroffen, die kontrolliert wird.

Folgende fünf Qualitätsmerkmale sollen letztlich die linguistische Qualität der wissenschaftlichen Arbeit identifizieren und der Bewertung des Textes dienen: Dazu gehört die Erstellung und Überprüfung eines Wortkatalogs, der Begriffe beinhaltet, die in einer wissenschaftlichen Arbeit zu vermeiden sind. Berücksichtigt sind hierbei Füllwörter, Umgangssprache, Unschärfe, umständliche Sprache, Verallgemeinerungen und Angstwörter. Des Weiteren wird untersucht, ob es aufeinander folgende Adjektive im Text gibt, die es zu vermeiden gilt. Ebenfalls wird die Satzlänge überprüft, indem zum einen die Wortanzahl zwischen zwei Satzendungen kontrolliert und zum anderen die Anzahl der verwendeten Kommas innerhalb des Satzes bestimmt wird, um verschachtelte Sätze zu identifizieren und das Textverständnis zu sichern. Darüber hinaus findet eine Überprüfung der Satzzeichen statt. Dabei wird der Text auf doppelte Leerzeichen sowie Ausrufezeichen durchsucht und es wird geprüft, ob geöffnete Klammern auch wieder geschlossen werden und dies in der richtigen Reihenfolge geschieht. Weiteres zu prüfendes Qualitätsmerkmal sind die Satzanfänge, welche auf Variabilität untersucht werden. Dazu wird der Satzanfang identifiziert und mit den drei nachfolgenden Satzanfängen verglichen, um Wiederholungen herauszustellen und den Textfluss zu ermitteln.

Um diese Textanalyse durchzuführen werden die Programme Eclipse, Layout-Aware PDF Text Extraction (LA-PDF-Text) sowie der Stanford Parser verwendet und im Rahmen agiler Softwareentwicklung auf das World Wide Web zurückgegriffen. Das LA-PDF-Text überführt die zu überprüfende wissenschaftliche Arbeit in ein passendes Format. Anschließend wird der Stanford Parser eingesetzt, welcher die grammatikalische Art der Wörter im Text identifiziert. Dies ist eine Voraussetzung für die Analyse

bestimmter Qualitätsmerkmale. Nach der Überprüfung soll schließlich eine Ausgabe erfolgen, die Auskunft darüber gibt, welche Qualitätsmerkmale inwieweit missachtet wurden. Auf dieser Grundlage erfolgt letztlich noch eine Bewertung der linguistischen Qualität der wissenschaftlichen Arbeit in drei Stufen.

2 Data Understanding

Als Datenbasis für das Analysetool werden die Publikationen der HMD – Praxis der Wirtschaftsinformatik aus den Jahren 2008 bis 2012 herangezogen. Die Sammlung umfasst 322 Dateien im PDF-Format, die im Durchschnitt je zehn Seiten lang sind. Die Dokumente beinhalten nicht nur Text, sondern auch eine Inhaltsübersicht, Tabellen, Abbildungen sowie ein Literaturverzeichnis, wodurch sich durchschnittlich acht Seiten Fließtext pro Datei ergeben. Zusammengefasst entspricht das 2576 Seiten Lauftext, die als Datenbasis zur Verfügung stehen.

Die HMD ist eine deutschsprachige Fachzeitschrift, die aktuelle Themen der Wirtschaftsinformatik behandelt. Herausgegeben von Fachleuten aus Industrie und Hochschule richtet sie sich an Führungskräfte der IT und Software-Ingenieure sowie Studierende und dient der laufenden Fortbildung im Bereich Wirtschaftsinformatik (Fachzeitungen). Anhand der Einstufung der HMD in die Kategorie „B“ von der „WI-Journalliste 2008“ (Wikipedia, 2013) kann auch die wissenschaftliche Qualität der Artikel als gegeben angesehen werden und somit als Grundlage für die spätere Parametereinstellung der Qualitätsmerkmale genutzt werden. Je nach Datenbasis ergeben sich entsprechend unterschiedliche Parametereinstellungen, um ein sinnvolles Ergebnis zu erzielen. Wird beispielsweise eine tadellose linguistische Qualität erwartet, sollte die Datenbasis aus Zeitschriften der Kategorie „A“ bestehen.

Bei der Sichtung der Daten wurde deutlich, dass die definierten Qualitätsmerkmale mit Ausnahme der inkorrekten Klammersetzung in vielen Fällen missachtet wurden. Dies ist für die linguistische Qualität nicht von Vorteil, hilft jedoch bei der Ausarbeitung des Analysetools, da ohne die Missachtung der Regeln die spätere Parametereinstellung sowie die Bewertung der wissenschaftlichen Qualität ohne einen durchschnittlichen Fehlerwert nicht sinnvoll möglich wäre.

3 Data Preparation

Aufgrund der gegebenen Publikationen von HMD ist die Datenauswahl als gegeben anzusehen. Um mit diesen Daten arbeiten zu können, müssen sie allerdings noch in das

richtige Format gebracht werden. Hierzu wird zuerst das LA-PDF-Text eingesetzt. Dieses Programm bietet die Möglich eine PDF-Datei als einzigen Textblock in einer TXT-Datei auszugeben. Durch eine Java-Abfrage werden die Gliederung und das Literaturverzeichnis im Textblock identifiziert. Da diese für die Überprüfung der Qualitätsmerkmale nicht relevant sind, sondern der Text in den Kapiteln untersucht werden soll, werden diese nicht in die Zieldaten übernommen und die Quelldaten an dieser Stelle bereinigt. Ziel dieses ersten Schrittes der Datenverarbeitung ist es, einen Textblock zu erstellen, der möglichst nur aus Fließtext besteht. In der von LA-PDF-Text erstellten Ausgabedatei lassen sich nun die Qualitätsmerkmale Wortkatalog, Variabilität der Satzanfänge, Satzlänge sowie Satzverschachtelung und Zeichensetzung überprüfen. Um die Verwendung von Adjektiven ebenfalls analysieren zu können, muss die grammatikalische Art der Wörter bestimmt werden. Daher wird die Ausgabedatei des LA-PDF-Text an den Stanford Parser weitergegeben, der diese Aufgabe erledigt. Anstatt des Textes gibt der Stanford Parser eine Datei aus, welche ausschließlich die grammatikalische Art der Wörter und Satzzeichen benennt und das Ergebnis aneinanderreiht und ebenfalls als Textblock in Form einer TXT-Datei ausgibt. Diese kann nun dahingehend untersucht werden, ob im Text aufeinander folgende Adjektive vorkommen, um das entsprechende Qualitätsmerkmal zu analysieren.

Die Überprüfung der definierten linguistischen Qualitätsmerkale findet demzufolge nicht innerhalb eines Zieldatensatzes statt, sondern vielmehr in zwei Ausgabedateien. Zum einen der vom LA-PDF-Text, in der ein Textblock aus Fließtext enthalten ist und zum anderen der des Stanford Parsers, in der die grammatikalische Art der Wörter und Zeichen des Fließtextes ausgegeben werden.

4 Modeling

Im Falle der vorliegenden Daten und in Anbetracht deren Anwendungszweck, gibt es mehrere Möglichkeiten, um mittels Modeling der Daten zu das gewünschte Output zu erreichen. In jedem Fall müssen die folgenden Arbeitsschritte durchgeführt werden:

1. Umformung des vorliegenden Textes in ein Format, das maschinenverarbeitbar ist.
2. Überprüfung des umgeformten Textes auf definierte Qualitätsmerkmale
3. Bildung einer Wertung, die, einfach verständlich, Informationen über die Qualität der Arbeit zur Verfügung stellt.

Besonders wichtig für die Wahl einer Modeling-Methode sind hierbei der erste und zweite Arbeitsschritt. Der erste Schritt wurde dabei bereits in der Phase Data Preparation bearbeitet und dort zwei Zieldatensätze erstellt. Anhand des ersten Zieldatensatzes können vier der fünf Qualitätsmerkmale überprüft werden. Die Überprüfung von Adjektiv-Verwendung in Texten ließ sich hiermit allerdings nicht umsetzen. Hierfür müssen Sätze mit Tokens angereichert werden, die grammatikalische Informationen über ein Wort und seine Funktion innerhalb eines Satzes zur Verfügung stellen. Eine solche Tokenization lässt sich mittels eines Sprach Parsers oder linguistischer Lexika umsetzen. Das ursprünglich präferierte Werkzeug zur Modellierung, (RapidMiner, 2014), ist nicht in der Lage eine Tokenization durchzuführen. Um ein gut geeignetes Tool zu finden, wurde eine Sammlung linguistischer Tools der (AG Digital Humanaties des Institutes für Informatik an der Friedrich-Alexander Universität Erlangen-Nürnberg, 2014) untersucht. Ausgewählt wurde der Stanford Parser der (The Stanford Natural Language Processing Group, 2014). Dieser Parser identifiziert zuverlässig Satzteile in Texten, ist in deutscher Sprache verfügbar und kann problemlos in Java-Programme integriert werden. Somit ist der Stanford Parser für den Anwendungszweck gut geeignet. Um ganze Texte in den Stanford Parser zu übertragen und das gewünschte Output als TXT-Datei exportieren zu können, wurde in der Programmiersprache Java eine angepasste Methode zum parsen erstellt. Da viele Methoden, die in späteren Arbeitsschritten zur Abfrage von Qualitätsmerkmalen benötigt werden, bereits zum Zweck des Parsens implementiert werden mussten, wurde sich dafür entschieden, die komplette Implementierung des Arbeitsprozesses in Java vorzunehmen.

Nach der Entscheidung für eine Modeling-Methode, wurde diese angewandt und ein Prototyp erstellt, dessen Arbeitsweise im Folgenden erläutert werden soll.

Während der Entwicklung eines ersten Prototyps wurde die automatische Ausführung von LA-PDF-Text noch nicht implementiert, somit muss hier händisch gearbeitet werden. Die von LA-PDF-Text erstellten Fließtexte werden zunächst vom Prototyp eingelesen und aufbereitet, indem nicht benötigte Text-Teile entfernt werden. Bei diesen handelt es sich um das Inhaltsverzeichnis, sowie das Literaturverzeichnis. Ist dieser Schritt abgeschlossen, wird der Text mit dem Stanford Parser analysiert und die Ergebnisse der grammatikalischen Bezeichnung gesondert abgespeichert. Somit stehen für die Anwendung der Qualitätsmerkmale alle Rohdaten zur Verfügung. Zur Überprüfung der Merkmale wurden Regeln festgelegt, an die sich ein wissenschaftlicher Text halten sollte. Die Regeln sind als Textoperationen hinterlegt und werden nacheinander geprüft. Stellt

der Prototyp hierbei einen Regelverstoß fest, wird dies abgespeichert. Sind alle Regeln geprüft worden, wird ausgegeben in welchem Bereich Regelverstöße vorliegen und wie viele es gibt. Darüber hinaus wird dem Anwender eine Bewertung zur Verfügung gestellt, die darstellt, ob ein Text qualitativ wertvoll ist. Diese Bewertung wird aus der Anzahl an aufkommenden Fehlern gebildet.

5 Evaluation

5.1 Erfolge und Schwachstellen des Prototyps

Die wichtigste Feststellung, die im Hinblick auf den erstellten Prototyp zu machen ist, ist, dass dieser zuverlässig arbeitet und keinerlei Probleme beim Umgang mit wissenschaftlichen Arbeiten aufweist. Auch die Erkennung von Regelverstößen, sowie die Bildung von Wertungen sind schlüssig.

Allerdings weist der Prototyp auch noch Schwachstellen auf, für deren Behebung weitere Arbeit investiert werden könnte. Als Schwachstelle identifiziert wurde vor allem, dass die Regelerkennung nicht zu 100% genau arbeitet. Dies liegt darin begründet, dass das Output von LA-PDF-Text in wenigen Fällen nicht korrekt ist und die implementierten Regeln nicht alle auftretenden Wort- und Zeichenkombinationen zur Gänze abdecken. Hinzu kommen weitere, kleinere Fehler, die keine inhaltliche Bedeutung haben und in der Ausführung eines ersten Prototyps üblich sind.

Bei der Betrachtung von Fehlerausgaben und Wertungen ist zu beachten, aus welchem Kontext diese generiert worden sind. Die Strenge in der Anwendung der Regeln ist auf die Artikel der Datenbasis, also der Zeitschrift HMD mit einem Ranking von „B“, abgestimmt. Die Bewertung ob ein Artikel gut oder schlecht ist, muss also immer in Relation dazu betrachtet werden. Um eine kritischere oder weniger strenge Bewertung zu erhalten, müsste man die Datenbasis anpassen und die Grenzen für die Fehlererkennung und die Wertung neu festlegen.

5.2 Mögliche Funktionserweiterungen

Abgesehen von den in Abschnitt 5.1 angesprochenen Einschränkungen, gibt es auch Punkte, an denen bei zukünftiger Entwicklung angesetzt werden kann um die Bedienung und den Funktionsumfang angenehmer zu gestalten. Der wichtigste Aspekt ist hierbei die Erweiterung des Prototyps mit mehr Qualitätsmerkmalen. Bei den bisher untersuchten handelt es sich nur um eine Auswahl, folglich sind die daraus gewonnenen Wertungen nur Indiz für die gute oder schlechte Qualität eines Textes. Durch die Im-

plementation weiterer Merkmale ließe sich Aussage der Wertungen allerdings verbessern.

Hinzu kommt eine Durchschnittsbildung bisher analysierter Dateien. Dadurch kann angegeben werden, wie hoch die Qualität eines aktuell untersuchten Textes im Vergleich zu anderen untersuchten Texten ist. Dies würde bedeuten, dass eine alternative Wertungsform hinzugefügt wird, die stärker auf Erfahrungswerten beruht, als die bisherige Wertung. Hierbei muss allerdings beachtet werden, dass eine Durchschnittsbildung immer nur in der Domäne einer bestimmten Datenbasis durchgeführt wird.

Alternativ wäre die Bereitstellung mehrerer Strenge-Grade und Konfigurationen möglich. Durch die Untersuchung weiterer Zeitschriften mit unterschiedlichen Rankings könnte dem Anwender die Wahl gegeben werden auf welchem Niveau er seine Arbeit testen lassen möchte.

Um dem Anwender darüber hinaus noch die Möglichkeit zu bieten seine Arbeit aktiv zu verbessern, sollte es möglich sein eine Ausgabe zu machen, die beinhaltet an welcher Stelle im Text welche Art von Fehler aufgetreten ist.

6 Deployment

Zur Durchführung eines letztendlichen Deployments ist es wichtig, dass zunächst in 5.1 beschriebene Fehler behoben werden und außerdem die in 5.2 aufgezeigten Verbesserungsmöglichkeiten realisiert werden. Anschließend ist der Nutzungszweck des Tools wichtig, da die Nutzung des Stanford Parsers nur zu wissenschaftlichen Zwecken kostenlos ist. Darüber hinaus sollte dem Tool eine GUI hinzugefügt werden, die Anwendern eine einfache Nutzung ermöglicht und die Ausgabe möglichst verständlich gestaltet.

7 Literaturverzeichnis

AG Digital Humanaties des Institutes für Informatik an der Friedrich-Alexander Universität Erlangen-Nürnberg. (2014). *NLP Software*. Abgerufen am 04. 02 2014 von http://wwwdh.cs.fau.de/IMMD8/Services/lt/

Fachhochschule Potsdam. (19. 08 2004). *http://bibliothek.fh-potsdam.de/fileadmin/fhp_bib/dokumente/Schulungen/wissenschaftliches_Arbeiten/Leitfaden_wiss_arbeiten_Graetsch.pdf*. Abgerufen am 23. 01 2014 von http://bibliothek.fh-

potsdam.de/fileadmin/fhp_bib/dokumente/Schulungen/wissenschaftliches_Arbeiten/Leitfaden_wiss_arbeiten_Graetsch.pdf

Fachjournalist. (02 2010). *http://www.ifp.uni-mainz.de/files/wissenschaftlichesschreiben.pdf.* Abgerufen am 23. 01 2014 von http://www.ifp.uni-mainz.de/files/wissenschaftlichesschreiben.pdf

Fachzeitungen. (kein Datum). *http://www.fachzeitungen.de/seite/p/titel/titelid/1016446741*. Abgerufen am 27. 01 2014 von http://www.fachzeitungen.de/seite/p/titel/titelid/1016446741

Maja Jokic, R. B. (2006). *Qualität und Quantität wissenschaftlicher Veröffentlichungen.* Jülich: Forschungszentrum Jülich GmbH.

RapidMiner. (2014). *RapidMiner Studio*. Abgerufen am 04. 02 2014 von http://rapidminer.com/products-2/rapidminer-studio/

The Stanford Natural Language Processing Group. (2014). *The Stanford Parser: A statistical parser*. Abgerufen am 04. 02 2014 von http://nlp.stanford.edu/software/lex-parser.shtml

Universität Augsburg. (22. 01 2014). *http://www.wiwi.uni-augsburg.de/vwl/michaelis/pdf/leitfaden.pdf.* Abgerufen am 23. 01 2014 von http://www.wiwi.uni-augsburg.de/vwl/michaelis/pdf/leitfaden.pdf

Universität Dresden. (2006). *http://tu-dresden.de/die_tu_dresden/fakultaeten/fakultaet_forst_geo_und_hydrowissenschaften/fachrichtung_geowissenschaften/ig/lehrstuehle/meuropa/lehre/downloads/akl/WissArb/Wiss%20Arb%202.pdf.* Abgerufen am 23. 01 2014 von http://tu-dresden.de/die_tu_dresden/fakultaeten/fakultaet_forst_geo_und_hydrowissenschaften/fachrichtung_geowissenschaften/ig/lehrstuehle/meuropa/lehre/downloads/akl/WissArb/Wiss%20Arb%202.pdf

Universität Freiburg. (kein Datum). *http://www.zlb.uni-freiburg.de/info_gympo/studium/dateien/reader-wiss-arbeiten.* Abgerufen am 23. 01 2014 von http://www.zlb.uni-freiburg.de/info_gympo/studium/dateien/reader-wiss-arbeiten

Universität Jena. (2007). *http://www.schreibenlernen.uni-jena.de/opsismedia/dokumente/technik_wiss_arbeitens_neu_iwk.pdf.* Abgerufen am 23. 01 2014 von http://www.schreibenlernen.uni-jena.de/opsismedia/dokumente/technik_wiss_arbeitens_neu_iwk.pdf

Universität Oldenburg. (04 2010). *http://www.uni-oldenburg.de/fileadmin/user_upload/germanistik/download/Leitfaden_wiss_Schreiben_WJ_final.pdf*. Abgerufen am 23. 01 2014 von http://www.uni-oldenburg.de/fileadmin/user_upload/germanistik/download/Leitfaden_wiss_Schreiben_WJ_final.pdf

Wikipedia. (04. 03 2013). *http://de.wikipedia.org/wiki/HMD_Praxis_der_Wirtschaftsinformatik*. Abgerufen am 27. 01 2014

Zusammenfassung und Ausblick

Die Ergebnisse der einzelnen Beiträge zeigen, dass die Anwendung von Text Mining auf wissenschaftliche Publikationen enormes Potenzial bietet. Die unterschiedlichen Prototypen stellen eindrucksvoll unter Beweis, dass eine durchdachte Konzeption und gewissenhafte Umsetzung ein gutes Ergebnis liefern kann.

Unabhängig von den inhaltlichen Ergebnissen möchten wir an dieser Stelle besonders die Veranstaltungsform hervorheben. Im Rahmen der Master Veranstaltung „Management Support Systeme III – Künstliche Intelligenz“ haben die Studierenden ein Themengebiet erhalten („Text Mining in wissenschaftlichen Publikationen“) und durften sich selbst über vertiefende Themen Gedanken machen. So entstanden in Zusammenarbeit mit den Studierenden abwechslungsreiche Themen mit wissenschaftlichen Fragestellungen. Diese wurden im Anschluss über einen Zeitraum von sechs Wochen von den Studierenden eigenständig bearbeitet. Die Ergebnisse waren interessant und spannend. Vor allem überraschte die hohe Qualität der prototypischen Implementierungen. Trotz des vergleichsweise kurzen Bearbeitungszeitraumes wurden umfangreiche und vor allen Dingen funktionstüchtige Prototypen vorgestellt.

Das Feedback der Studierenden war durchweg positiv. Hierbei wurde vor allem die praktische Ausrichtung der Veranstaltung gelobt, sodass neben der theoretischen Konzeption auch mit einem Prototypen experimentiert und die Ergebnisse dadurch verfeinert werden konnten.

Auf Grund der guten bis sehr guten Ergebnisse haben wir uns entschlossen, die Dokumentation der Implementierungen in Form dieses Sammelbandes zu veröffentlichen. Eine Fortführung der Veranstaltungsform ist fest geplant und eine weitere Veröffentlichung der Ergebnisse als weiterer Sammelband ist angedacht.

Osnabrück, im September 2014

Prof. Dr.-Ing. Bodo Rieger
Axel Benjamins