Automatisierte Erstellung von Pressedossiers durch Textmining

Kontextualierung im journalistischen Umfeld

Masterseminar Ausarbeitung

Marcel Schöneberg

marcel.schoeneberg@haw-hamburg.de

Hochschule für Angewandte Wissenschaften Hamburg (HAW)
Fakultät für Technik und Informatik
Department Informatik

Februar 2015

Zusammenfassung

Dieses Dokument dient als Resümee für abgeschlossene, aktuelle und vor allem zukünftige Arbeiten des Autors auf dem Gebiet des Textmining. Hierzu wird zunächst kurz auf die Ergebnisse des Moduls 'Projekt 1' (Grundprojekt) im Masterstudiungang Informatik eingegangen. Dieses befasste sich mit Trenderkennung in sozialen Netzwerken. Darüber hinaus werden, als Kern, die Zielsetzungen der Masterarbeit des Autors abgesteckt und die zugehörigen grundlegenden Herangehensweisen vorgestellt. Die übergeordnete Vision der Arbeit ist hierbei die automatisierte Erstellung von Pressedossiers. Im Kontext dessen werden auch die bisherigen Fortschritte des Moduls 'Projekt 2' (Hauptprojekt), welches Grundlagen für die Masterarbeit liefert, angerissen. Den Schluss bildet eine Betrachtung der Chancen und Risiken der gestellten Aufgabe.

1 Einleitung

Dieses Dokument greift die bisherige Arbeit des Autors auf und stellt daraufhin die grundlegende Planung für die Masterarbeit des Verfassers vor. Das bearbeite Forschungsgebiet ist das Textmining, das konkrete Ziel ist die automatische Erstellung von Pressedossiers. Dieses wird auf fachlicher Ebene durch eine Domänenexpertin unterstützt, daher gibt es domänenübergreifende Aspekte welche zu beachten sind. Diese werden ebenfalls in dieser Arbeit grundlegend vorgestellt.

Diese Ausarbeitung gliedert sich in fünf Abschnitte. Nach dieser Einleitung wird zunächst wird ein Rückblick auf die bisherige Arbeit des Verfassers zum Thema Trenderkennung in sozialen Netzwerken gegeben.

Abschnitt drei bietet einen Überblick über die geplante Masterarbeit und stellt dazu Grundlagen vor, auf welchen weitere Arbeiten basieren. In diesem Zuge wird die ursprüngliche Idee sowie der Weg zu einem konkreteren Ziel vorgestellt. Darüber hinaus werden Fragen erarbeitet welche als Leitfaden für die Arbeit dienen sollen. Aufbauend darauf wird das geplante konkrete Vorgehen umrissen, sowie Kriterien vorgestellt welche die Bewertung von produzierten Ergebnissen ermöglichen sollen.

Daraufhin wird der bisherige Projektfortschritt dargestellt. Hierzu werden ebenso die genutzten Methoden und Werkzeuge kurz umrissen.

Den Abschluss bildet eine Betrachtung der Chancen und Risiken des geplanten Projektes.

2 Rückblick Grundprojekt: Grundlagen, Clustering, Ergebnisse

Der folgende Abschnitt umreißt kurz die Ergebnisse des Grundprojektes [Sch14, siehe:]. Dieses befasste sich mit der Thematik der Trenderkennung/Clustering in sozialen Medien. Hierzu schuf sich der Autor zunächst eine Arbeitsumgebung inklusive Testdaten, zudem wurden erste Experimente durchgeführt um u.A. die Qualität der Daten abzuschätzen.

2.1 Grundlagen

Als Ansatzpunkt für die Erkennung von Trends dienten so genannten 'Weak Signals', welche in einem Set von ca 27.000 Tweets aufgespürt werden sollten. Hierzu wurde zunächst über einen selbst entwickelten Crawler die Suche der Twitter Website benutzt um das Datenset zu generieren. Die ausgewählte Methode zur Analyse der Daten war ein Clusteringverfahren. Die verfolgte Arbeitshypothese bestand darin, dass entstehende Trends bzw. Themen sich in (ggf. hierarchischen) Clustern im Ansatz auffinden lassen und dass diese in späteren Arbeitsabschnitten über die Zeit verfolgt werden können.

2.2 Ergebnisse

Dieses gewählte Vorgehen führte aus Sicht des Autors zu diversen Erfolgen: Zunächst entstand das genannte Datenset (27.000 Tweets, Zeitraum: 2,5 Jahre, Eingrenzung: Hashtag: #piraten). Darüber hinaus sammelte der Autor Erfahrungen in der Vorverarbeitung eines solchen Datensatzes. Konkrete Schritte hierzu umfassten eine Normalisierung der Daten, Stemming, sowie allgemein Verarbeitung von natürlicher Sprache (NLP). Betrachtet man allerdings die konkreten Ergebnisse dieses Projektes ist zu sagen, dass die Resultate in keinster Weise zufriedenstellt waren, da die Cluster zum Großteil aus unzusammenhängenden Begriffen bestanden und nicht zielführend waren. Verschiedene Gründe können für dieses Ergebnis angeführt werden: Grundsätzlich sind Tweets inhärent kurz (max. 140 Zeichen), darüber hinaus ist die verwendete Sprache oft nur begrenzt maschinell verarbeitbar, da Abkürzungen sowie Tippfehler etc. die Verarbeitung erschweren. Darüber hinaus ist anzumerken, dass Clusteringverfahren oft recht undurchsichtig sind und man daher nicht direkt verfolgen kann wie ein bestimmtes Ergebnis entstanden ist.

Grundsätzlich ist anzumerken, dass eine Trenderkennung auf ähnlichen Daten durchaus gelingen kann. Hierzu ist allerdings oft eine aufwändige Vorverarbeitung, sowie komplexe Lernverfahren von Nöten [Nik12, vgl.]. Der Autor entschied sich aus den obigen Gründen daher den Fokus seiner Arbeit zu verlagern, allerdings im Gebiet des Textmining zu verweilen.

Nähere Angaben zum Verlauf, sowie den Ergebnissen des Projekts lassen sich dem Projektbericht ([Sch14]) entnehmen.

3 Überblick: Masterarbeit

Die folgenden Abschnitte bilden den Kern dieser Ausarbeitung und umreißen die Masterthesis des Autors. Die Vision dieser Arbeit ist hierbei die automatisierte Erstellung von Pressedossiers auf Basis eines Artikelarchivs, sowie eines Leitartikels.

Zunächst wird die Motivation und Herkunft des Vorhabens geschildert, hierzu wird auch auf die vorhandene Datenbasis eingegangen. Im Weiteren wird daraufhin eine konkretere Zielsetzung erarbeitetet und darauf eingegangen welche konkreteren Vorhaben geplant sind. Darüber hinaus wird erläutert wie Ergebnisse bewertet werden können und welche weiteren Schritte denkbar sind.

3.1 Einführung in die Thematik

3.1.1 Vision

Die angestrebte Arbeit mit dem Ziel der automatisierten Dossiererstellung ist ein domänenübergreifendes Projekt welches sowohl die Informatik auf technischer Seite, sowie (im weitesten Sinne) die Journalistik auf fachlicher Seite umfasst. Aus diesem Grund wurde die ursprüngliche (mittlerweile abgewandelte) Vision von fachlicher Seite inspiriert. Diese Idee umfasste die Erstellung und Analyse von Presseerzeugnissen hinsichtlich verschiedener dargestellter Sichtweisen auf ein Thema. Ziel hierbei sollte es sein zu prüfen ob Textmining basierte Dossiers dem Wachsen einer gemeinsamen europäischen Erzählung dienen können [Hä14, vgl. auch:].

Die ursprüngliche fachliche Vision wurde im Dialog mit dem Autor zunächst herunter gebrochen, so dass das primäre Ziel nun eine, von Mitteln der Informatik gestützte, Erstellung von Pressedossiers ist. Diese Dossiers sollen der Kontextualisierung eines Ausgangsartikels (Leitartikel) dienen, so dass dieser für Domänenexperten aufbereitet wird (und in einem Kontext steht) um weitere Arbeiten zu erleichtern. Hierbei sind ausschließlich textbasierte Artikel und Dossiers im Fokus, so dass multimediale Inhalte außen vor bleiben.

Grundsätzlich stellt sich bei der vorliegenden Zielsetzung die Frage nach der Definition eines Dossiers. Die Beantwortung dieses Kernpunktes ist allerdings nicht trivial, so gibt es verschiedene (Experten-) Meinungen darüber was ein Dossier ausmacht ([Hä14, vgl.]). Die vom Autor verfolgte Arbeitsthese ist, dass ein Dossier zunächst eine Zusammenstellung von ähnlichen Artikeln im weitesten Sinne ist. Diese haben eine inhaltliche/semantische Nähe, welche mit Hilfe von Distanzfunktionen ermittelt werden kann. Dieser Ansatz soll als Grundlage für weitere Arbeiten dienen, er erfüllt allerdings nicht alle möglichen Anforderungen welche sich ergeben können. Aus diesem Grund ist die Zielsetzung eingeschränkt zu betrachten, dieses illustriert Abbildung 1.

Ausgehend von einem bestehenden Artikelarchiv und einem gegebenen Leitartikel (welcher als Referenz für die gesuchte Ähnlichkeit dient) sollen Vorschläge (basierend auf der in einer Blackbox berechneten Distanz) generiert werden. Diese werden einem Domänenexperten unterbreitet, dieser kann mit Hilfe der Vorschläge ein Dossier erstellen. Das geschilderte Vorgehen ermöglicht eine automatisierte Dossiererstellung ohne sich komplett auf die hinterliegende 'Intelligenz' (der Blackbox) oder den Domänenexperten verlassen zu müssen. Das Verfahren soll hierbei durch weitere Arbeiten fortlaufend verbessert werden, so dass der Domänenexperte im Optimalfall nur noch eine untergeordnete Rolle spielt.

3.1.2 Artikelarchiv

Für die Umsetzung der geschilderten Vision steht ein, für akademische Verhältnisse, großes Artikelarchiv zur Verfügung. Dieses basiert auf dem Eurozine Netzwerk (`www.eurozine.com`), einem Zusammenschluss von europäischen Kulturzeitschriften. Die vom Autor genutzten ca. 3700 Artikel wurden von professionellen Journalisten verfasst. Die Dokumente sind größtenteils (teils als Übersetzung) in englischen Sprache verfasst, darüber hinaus besteht weitergehend die Möglichkeit Metainformationen zu den Artikeln zu erhalten (z.B Verlinkungen auf die Inhaltsverzeichnisse der Ursprungszeitschrift, redaktionell erstellte Archive etc.). Die Artikel selber liegen Form von XML-Dateien vor und weisen eine Semistrukturiertheit auf, so können u.A. Informationen wie Autor, Kurzzusammenfassung, sowie Überschriften etc. direkt dem Dokument entnommen werden.

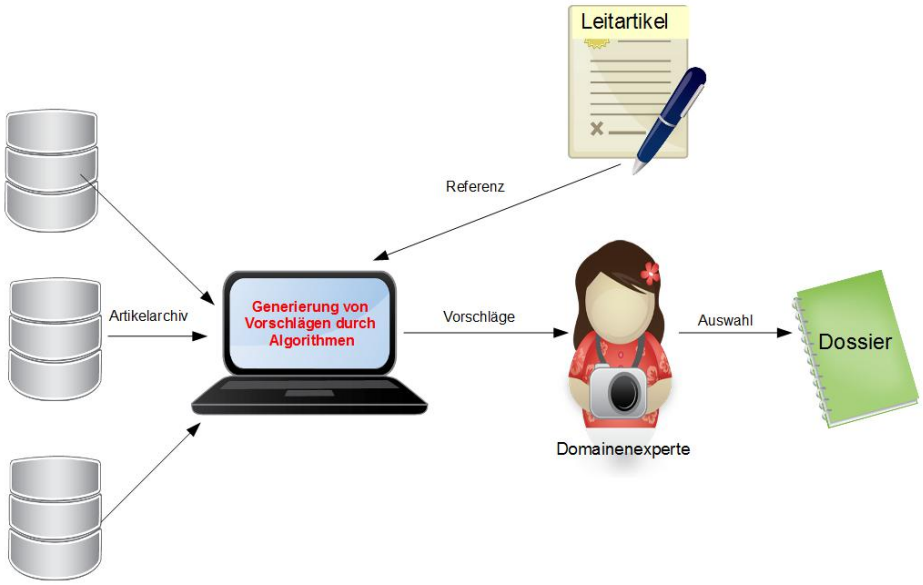


Fig. 1: Realisierbarer Workflow

Trotz der genannten Vorteile ist das Archiv nicht makellos, weshalb eine Vorverarbeitung von Nöten war. Zu erwähnen ist, dass nicht alle 3700 Artikel in englischer Sprache verfasst sind, darüber hinaus enthält das Archiv auch eine Reihe von Zusammenfassungen, Inhaltsangaben sowie Rezensionen bereits erschienener Artikel. Bei der Auswahl eines kleineren Testkorpus fielen darüber hinaus einige Gedichte auf, welche im Vergleich zum Rest eine extrem verkürzte Länge, sowie eine inhärent andere Art der Sprache verwenden. Aus technischer Sicht ist zu bedenken, dass das XML Markup nicht bei allen Artikel valide ist, dieses musste in einigen Fällen korrigiert werden.

3.2 Konkrete Fragen und Zielsetzung

Der folgende Abschnitt dient dazu eine Reihe an Fragen herauszuarbeiten, welche als Leitfaden der Masterarbeit genutzt werden sollen.

Den Kern bildet die Frage nach der Machbarkeit der automatisierten Erstellung von Pressedossiers. Die Antwort auf diese Frage hängt von einer Reihe von feingranulareren Fragen ab. Diese sind z.T. allerdings erst im Laufe der Arbeit zu beantworten. Daher ist es möglich, dass das gesetzte Ziel der Automatisierbarkeit schlussendlich nicht oder nur eingeschränkt zu erreichen ist.

Eine weitere Frage die sich direkt im Anschluss stellt befasst sich mit der Definition von Pressedossiers bzw. dem Nutzungsszenario auf fachlicher Seite. Was ein Dossier ist und wie dieses genutzt wird ist daher auch nur durch einen Domänenexperten zu beantworten. Diese Frage bildet daher die Basis für eine Schnittstelle, welche die fachliche mit der technischen Sicht verbindet. Diesen Verbindungspunkt bildet die Distanzfunktion, welche verschiedene Artikel in Bezug zueinander setzt und ihre Ähnlichkeit in Form eines Zahlenwertes repräsentiert. Für diese Funktion müssen fachliche Wünsche mit Hilfe von Mitteln der Informatik in Algorithmen überführt werden.

Dieser Punkt führt direkt zur nächsten Frage ob fachlich definierte Anforderungen an ein Dossier sich grundsätzlich in einen Algorithmus überführen lassen. Sind die fachlichen Wünsche im Allgemeinen eher auf eine Art 'Bauchgefühl' zurückzuführen bzw. schwammig formuliert, kann man dieses nicht umsetzen, hierzu bräuchte man konkretere Informationen.

Eine weitere eng verbundene Frage ist die nach der Machbarkeit der Umsetzung. Ein mögliches

Szenario könnte eine recht präzise Definition sein, welche allerdings aufgrund von mangelnder Datenbasis, beschränkter Ressourcen oder technischer Mittel etc. nicht innerhalb der geplanten Zeit umsetzbar ist.

Grundsätzlich ist darüber hinaus zu beantworten wie man ein generiertes Ergebnis bewertet um seine Qualität zu ermitteln bzw. eine Verbesserung oder Verschlechterung der Ergebnisse bei Änderungen zu erkennen.

Neben der Beantwortung der obigen Fragen wäre es wünschenswert einen zu entwickelten Algorithmus möglichst übersichtlich zu halten, so dass dieser durchaus (für Fachexperten) nachvollziehbar ist. Sofern dieses nicht gelingt entsteht eine Blackboxlösung welche umso schwerer auf ggf. auftretende weitere Wünsche anpassbar ist.

3.3 Vorgehen

Die folgenden Abschnitte beschreiben das vom Autor geplante Vorgehen und umreißen dabei auch den Prozess der Auswahl von konkreten Vorgehensweisen, sowie mögliche weitere Schritte.

3.3.1 Erste Ideen

Da diese Arbeit domänenübergreifend ist musste zunächst eine gemeinsame Basis erarbeitet werden, dieses geschah durch Diskussionen und das Sammeln und Abwägen von diversen Ideen. Diese umfassten beispielsweise das Finden von Facetten eines Artikels, sowie die Nutzung von Metainformationen (wie Autor, Datum der Veröffentlichung usw, [Nor12, Definition: S. 256] um einen bestmöglichen Überblick im Dossier zu gewähren. Auch wurde überlegt vorhandene Clusteringverfahren bzw. crowd-basierte Techniken anzuwenden um ein Dossier zu erstellen.

Die zunächst erarbeiten Ideen erwiesen sich beim Abwägen als nicht zielführend, da sie nicht die gewünschten Eigenschaften aufweisen. Zunächst sind einige Ideen (wie das Auffinden von Facetten) vergleichsweise komplex und sind daher nicht im ersten Schritt anzugehen. Ebenso ist die Nutzung von Metainformationen eine durchaus häufig genutzte Idee [FS06, siehe z.B. S. 65], welche allerdings nicht das Grundproblem löst. Darüber hinaus sind vorhandene Clustering- und/oder Lernverfahren oft sehr komplex und teils undurchsichtig für Endanwender [FS06, Seite: 64] und verhindern so auch die Erweiterbarkeit in Bezug auf journalistische Wünsche. Weiterhin besteht bei lernenden Systemen oft das sogenannte Coldstart-Problem [BOHG13, vgl. S. 5]. Dieses zeigt sich dadurch, dass nicht genug Lerndaten für sinnige Vorschläge vorhanden sind. Bei weiteren Recherchen wurde außerdem klar, dass viele vorhandene Ansätze auf Personalisierung bzw. auf crowd-basierte Lösungen setzen um Vorschläge für Nutzer zu generieren bzw. zu erlernen [KFD12, PL10, LCLS10, vgl.]. Dieses entspricht allerdings nicht den vom Autor gesetzten Zielen, da die angestrebten Dossiers möglichst automatisch zu erstellen sein sollen (widerspricht daher dem Crowdansatz) und zum Anderen zunächst nicht auf Journalisten personalisiert sein sollen. Dieses würde potenziell eine neutrale Einschätzung verhindern. Darüber hinaus gefährdet die Nutzung von externem Wissen (wie der Crowd) die bereits erwähnte Anpassbarkeit an journalistische Wünsche, da spezielle Anforderungen von journalistischer Seite nicht ohne weiteres auf z.B crowdbasierte Systeme übertragen werden können. Weiterhin leidet auch die Überschaubarkeit des Gesamtsystems und verwandelt es so zunehmend in ein (schlecht wartbares) Blackboxsystem. Die angesprochenen Verfahren können sich in späteren Projektphasen durchaus aus nützlich erweisen, sind allerdings im ersten Schritt nicht erwünscht.

Die durchgeführte Brainstorming zur Sammlung von Ideen brachte allerdings auch nützliche Gedanken und Paper in den Fokus. Zu diesen zählen z.B. das Overview Project `http://overview.ap.org`, dieses Opensource Tool ist ursprünglich dazu gedacht Journalisten in ihrer Arbeit mit großen Dokumentarchiven zu unterstützen. Die Fachexperten sollen damit die Möglichkeit haben Dokumente automatisch in (hierarchische) Themencluster einzusortieren und diese u.A zu visualisieren. Darüber hinaus wurden vom Autor diverse Paper über Vorschlagsysteme [LGS, BOHG13, BDD+12] sowie Distanzfunktionen [HRJM13, TCY09] gesichtet, welche interessante Anregungen enthalten und grundsätzlich förderlich sind in Bezug auf das Verständnis der Thematik sowie bekannter Hürden.

3.3.2 Konkrete Schritte

Die folgenden Abschnitte widmen sich der konkret verfolgten Strategie zur Erreichung des gestellten Ziels. Aufgrund des experimentellen Charakters des Projekts hat sich der Autor entschieden einen Ansatz zu nutzen welcher sich schrittweise dem Ziel nähert und mit jedem Schritt möglichst eine Verbesserung erreichen soll. Konkret soll hierbei die im Kern stehende Distanzfunktion durch neue Ansätze und (fachliche) Erkenntnisse verfeinert werden. Grundsätzlich beruht der Ansatz darauf, das der Korpus bzw. die einzelnen Artikel als 'Bag-Of-Words' behandelt wird, also einer Menge von Wörtern welche keine Semantik aufweisen [FS06, S. 68]. Darüber hinaus dient (zunächst) ein Leitartikel (*der* Artikel zum gewählten Thema) als Ausgangsbasis der Distanzfunktion.

Auf Basis des 'Bag-Of-Words' Ansatzes wird der Featurevektor (die abstrakte Repräsentation des Eingangsdokumentes, [FS06, S. 68]) definiert auf welchem die zu entwickelnde Distanzfunktion arbeiten soll. Der Vektor wird hierbei aus allen (nach einer Vorverarbeitung) vorhandenen Wörtern und deren Häufigkeit pro Artikel bestehen. Das geschilderte grundlegende Modell ist auch unter dem Namen 'Vector Space Model' (VSM) bekannt [FS06, LGS, vgl. S. 85]. Das Preprocessing [FPS96] der Daten wird zunächst nur aus einem Entfernen von 'stop words' (häufig vorkommende Wörter ohne konkrete Bedeutung für den Inhalt - z.B Artikel usw.), sowie einem 'Stemming' (Reduzieren von Wörtern auf ihren Stamm, [Nor12, siehe: S. 200 ff.]) bestehen. Dieses kann in späteren Iterationen den Bedürfnissen angepasst werden, so könnten weitere Vorverarbeitungsschritte von Nöten sein. Ebenso denkbar ist es, dass ein reiner 'Bag-Of-Words'-Ansatz nicht ausreicht sodass dieser erweitert werden muss.

Für die Berechnung einer Distanz zwischen Dokumenten soll im verfolgten Ansatz die Struktur der Artikel eine maßgebliche Rolle spielen. Diese umfasst wie bereits erwähnt u.A. eine Kurzzusammenfassung, den Titel eines Artikels, diverse Unterüberschriften für Paragraphen, sowie den eigentlichen Text. Die verfolgte Hypothese des Autors ist, dass diese Textanteile einen signifikanten Anteil der Informationen eines gesamten Artikels zusammenfassen. Demzufolge bieten sie eine gute Basis um ein Ähnlichkeitsmaß an ihnen fest zu machen. Aufgrund dieser Überlegung sollen diese Bestandteile im ersten Schritt mit jeweils einem spezifischen Gewichtungsparameter x_n belegt werden. Die jeweiligen Parameter drücken aus wie 'wichtig'/informationstragend die verschiedenen Abschnitte jeweils in Bezug zueinander sind. Die Gewichtungsparameter werden in die Berechnung des Featurevektors mit einbezogen.

Hierzu wird die **Worthäufigkeit** tf für **Wort** w, welches im Abschnitt mit dem **Parameter** x_n vorkommt, mit x_n multipliziert (aufgrund der These, dass das Wort für den Artikel ausschlaggebender ist als andere Wörter). Die **Gesamthäufigkeit** $tf_{ges}(w)$ eines Wortes w im Artikel ergibt sich daher als Summe über die gewichteten Vorkommen pro **Abschnitt** tf_{sec_n}:

$$tf_{ges}(w) = (x_1 * tf_{sec_1}(w)) + (x_2 * tf_{sec_2}(w)) + \cdots + (x_n * tf_{sec_n}(w)) \tag{1}$$

Die auf obige Weise erstellten Featurevektoren werden im nächsten Schritt mit einer einfachen Distanzfunktion (z.B. euklidische Distanz) verglichen (hierbei wird Artikel a_n jeweils mit dem vorgegeben Leitartikel verglichen). Der berechnete Wert stellt die Distanz der Artikel dar. Sowohl das genutzte Maß (Termfrequency (TF)) sowie auch die Distanzfunktion können beliebig erweitert werden. So kann man das Maß normiert werden und andere Distanzfunktionen genutzt werden (z.B. Cosine-Distanz).
Die auf diese Weise berechneten m 'besten' Artikel (kleinste Distanz zum Leitartikel) werden im ersten Schritt als Vorschläge an einen Fachexperten gegeben welcher diese in seine Überlegungen für ein Dossier mit einbeziehen kann.

Weitere geplante konkrete Schritte sind die Einführung eines Kategoriensystems welches in die Erstellung des Featurevektors mit einbezogen wird. Die zu ermittelnden Kategorien sollen dazu dienen die zunächst rudimentäre Distanzfunktion zu verbessern. Die hinterliegende Überlegung ist, dass eine ein beliebiger Presseartikel mindestens einer Kategorie zuzuordnen ist, diese kann ein signifikates Merkmal sein welches sich direkt auf die Ähnlichkeit zweier Artikel auswirkt. Nach dieser Überlegung sollte sich ein Artikel der Kategorie 'Informatik' stark von einem Text zum Thema 'Sozialwissenschaft' unterscheiden. Konkret soll zunächst eine kleine Menge von Kategorien ausgewählt werden, für die jeweils eine Liste mit themenspezifischen Schlagwörtern existiert. Das Vorkommen eines dieser Wörter soll entsprechend gewichtet werden und so die später berechnete Distanz beeinflussen. Eine weitere denkbare Möglichkeit ist es einem Artikel anhand der Schlagwortlisten und dem entsprechenden Wortvorkommen eine Metain-

formation (die Kategorie) zuzuweisen. Dieses hätte den Vorteil, dass die Kategorie nicht im Featurevektor 'untergeht' sondern als eigenständige Information erhalten bleibt.

Ein aufbauender Schritt auf den geschilderten Methoden ist das hinzuziehen von Informationen eines Fachexperten. Dieser soll Wissen darüber liefern, welche Aspekte bei der Erstellung von Dossiers auf fachlicher Seite benutzt werden, so dass diese nach Möglichkeit in die Distanzfunktion eingearbeitet werden können. Die Wissensfindung auf fachlicher Seite gehört allerdings nicht in den Aufgabenbereich des Autors sondern wird von einer Domänenexpertin durchgeführt [Hä15, Masterthesis in Bearbeitung]. Aus diesem Grund kann der Autor an dieser Stelle noch nicht auf etwaige Ergebnisse dieser Untersuchung eingehen.

Zu beachten ist, das die momentan geplante Distanzfunktion recht simpel ist. Das Thema Distanzfunktionen bildet allerdings einen eigenen Bereich und soll daher hier nur kurz angerissen werden. Informationen zu diesem Thema lassen sich z.B. [HRJM13, Sip14] entnehmen. Darüber hinaus sind diverse Aspekte in den Überlegungen des Autors präsent, dazu zählen z.B. verschiedene Relevanzmaße, um mit längeren Artikeln bzw. Mengen von häufigen Wörtern umzugehen, Beispiele für solche Relevanzmaße sind z.B. Termfrequency und TF-IDF. Darüber hinaus ist zu bedenken, dass es mehrere mögliche Kriterien zum Thema Distanz gibt, hierzu kann neben Worthäufigkeiten auch eine Themenbreite und Neuheit von Artikeln zählen [BOHG13, vgl. S. 3]. Viele dieser Aspekte gehen allerdings über den Rahmen dieser Arbeit hinaus und können eher von fachlicher Seite beantwortet werden, weswegen sie zunächst eine untergeordnete Rolle spielen.

3.3.3 Bewertungskriterien für Ergebnisse

Der folgende Abschnitt soll eine weitere Frage klären welche im Kontext der Arbeit wichtig ist, diese behandelt die Bewertung von gewonnenen Ergebnissen. Der geplante Workflow (vgl. Graphik 1) produziert Vorschläge für einen Fachexperten, welche dieser verwerten kann. Allerdings ist es für den Autor dieser Arbeit wichtig zu wissen ob etwaige Veränderungen an der 'Blackbox' des Algorithmus zu Verbesserungen führen. Zu diesem Zweck ist es geplant die redaktionell erstellten Focalpoint-Zusammenfassungen (`http://www.eurozine.com/comp/FocalPoints.html`) des vorliegenden Artikelarchivs zu nutzen. Diese enthalten jeweils ca. 30 Artikel zu einem Thema. Hierbei existieren verschiedene Themenbereiche, so dass mehrere Focalpoints zur Verfügung stehen. Diese Sammlungen stellen zwar kein konkretes Dossier dar, bieten allerdings eine Basis für die Validierung und Verifikation der Ergebnisse, da sie verschiedene Artikel mit einem gemeinsamen Schwerpunkt zusammenfassen. Diese Eigenschaft ist laut per Definition die Grundvorraussetzung eines Dossiers (vgl. z.B. Duden). Für den Aufbau eines Testkorpus wird ein Focalpoint (ca. 30 Artikel) mit der gleichen Anzahl von zufälligen Artikeln aus dem Archiv verschmolzen. Die von der Blackbox vorgeschlagenen Artikel (ausgehend von einem Leitartikel) sollen nun im besten Fall genau die Artikel des Focalpoints sein. Dieses ergibt sich aus der Eigenschaft der Focalpointartikel als Positivbeispiele, sowie den zufällig zusammengestellten Artikeln, welche als Negativbeispiel dienen.

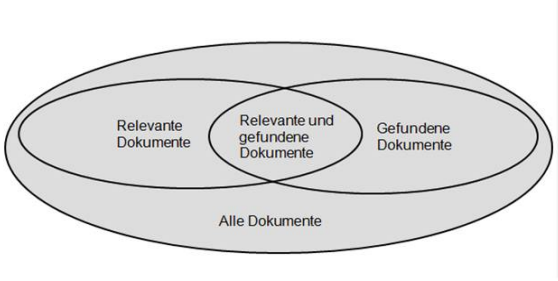


Fig. 2: Precision-Recall Venndiagramm [PR-]

Eine Möglichkeit die Ergebnisse des Vorschlagssystems in Zahlenwerte zu übertragen sind die Werte Recall (Trefferquote) und Precision (Genauigkeit) aus dem Umfeld des Information Retrieval [FS06, vgl. u.A. S. 79]. Hierbei benötigt man (im konkreten Fall) die Menge der 'gefundenen' Dokumente, sowie die Menge der tatsächlich relevanten Treffer (Artikel des Focalpoints). Der Recall drückt hierbei aus wie viele relevante

Ergebnisse aus der Gesamtmenge ausgewählt wurden, während die Precision ein Maß ist, welches aussagt wie viele der ausgewählten Ergebnisse relevant sind. Dieses wird noch einmal verdeutlicht in Abbildung 2. Des weiteren werden im folgenden Abschnitt erste Erfolge durch den vorhandenen Experimentalaufbau aufgezeigt.

Darüber hinaus kann auch auf das Wissen eines Domänenexperten zugegriffen werden um Ergebnisse zu prüfen bzw. um zu verstehen warum ein Resultat anders ausfällt als erhofft.

3.3.4 Weitere mögliche Schritte

Auf der Basis der ursprünglichen Vision, sowie dem beschriebenen Brainstorming von Ideen gibt es noch eine Reihe von weiteren möglichen Schritten welche in Zukunft verfolgt werden können. Zu diesen zählen u.A. die Nutzung von Metainformationen, welche des Archiv zur Verfügung stellt (beispielsweise Autoren, Verlinkungen von Inhaltsverzeichnissen etc.). Darüber hinaus können im späteren Projektverlauf auch linguistische Verbesserungen (z.B. die Nutzung von Ontologien [FS06, siehe: S. 197]) oder eine generell erweiterte Vorverarbeitung der Daten von Interesse sein. Im Weiteren gilt es auch die ursprüngliche Vision der Erkennung von Facetten eines europäischen Themas nicht aus den Augen zu verlieren. Diese und mögliche weitere Ansätze liegen derzeit allerdings nicht im Fokus des Autors und sollen nur als Ausblick und weitere Motivation für interessierte Personen dienen.

4 Stand des Experimentalaufbaus

Im Folgenden soll kurz der aktuelle Fortschritt des Projekts geschildert werden. Dieses umfasst diverse Vorarbeiten am Dokumentenkorpus, die technische Grundlage zur Ermittlung von Distanzen mit Hilfe von Rapidminer, sowie erste Experimente.

Korpusaufbau

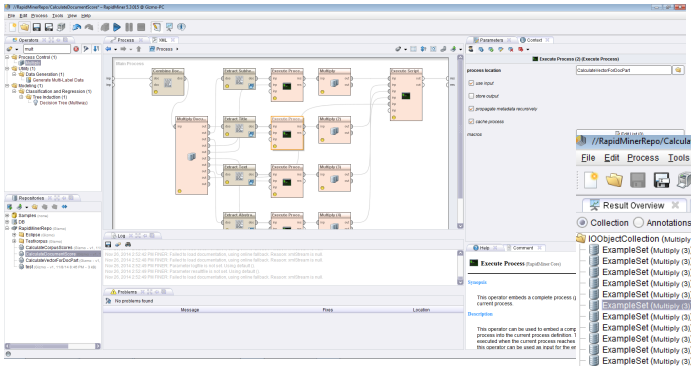
Nachdem zusammen mit der Domänenexpertin Nina Hälker ein gemeinsames Verständnis der Zielsetzung entwickelt worden ist wurde ein Korpus entworfen, welcher als Grundlage für alle weiteren Arbeiten dient. Diese Dokumentenzusammenstellung besteht aus Artikeln des benutzten Pressearchivs und dient zur Überprüfung der gewonnenen Ergebnisse. Hierzu umfasst das Testset Artikel des Focalpoints 'Demokratie', sowie zufällig ausgewählte Archivdokumente. Diese sollen als Vergleichsbasis dienen und nicht dem Focalpoint angehören. Hierbei sollen die Focalpointartikel möglichst genau die vom System vorgeschlagenen Artikel sein. Das Mischungsverhältnis von Artikeln des Focalpoints und den zufälligen Negativbeispielen soll 50:50 betragen, so dass das Vorschlagssystem eine Erfolgsrate von 50% (zufällige Einteilung von falsch und richtig) übertreffen sollte.

Weitere Testkorpora befinden sich in Planung, diese sollen u.A. dazu dienen die zu findenden Ergebnisse (z.B. Gewichtungsfaktoren) gegen einen anderen Korpus zu prüfen, um zu testen wie stabil diese Werte sind und zu untersuchen wie stark sich der Lösungsansatz verallgemeinern lässt.

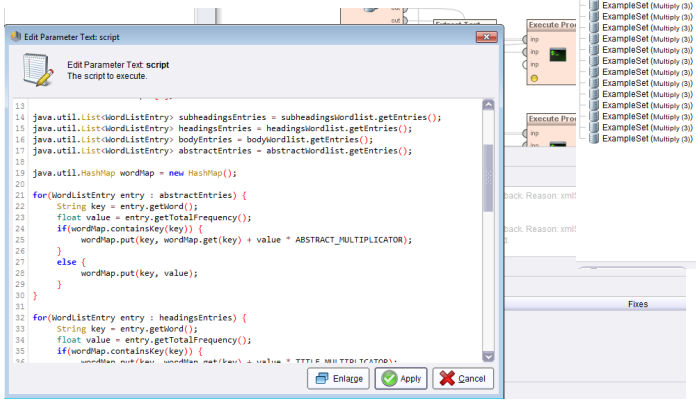
Rapidminer

Darüber hinaus musste zunächst eine Analyseumgebung geschaffen werden, dieses geschah mit Hilfe des Tools 'RapidMiner' (www.rapidminer.com). Abbildung 3 zeigt einige Ausschnitte des momentan existierenden RapidMiner-Projekts.

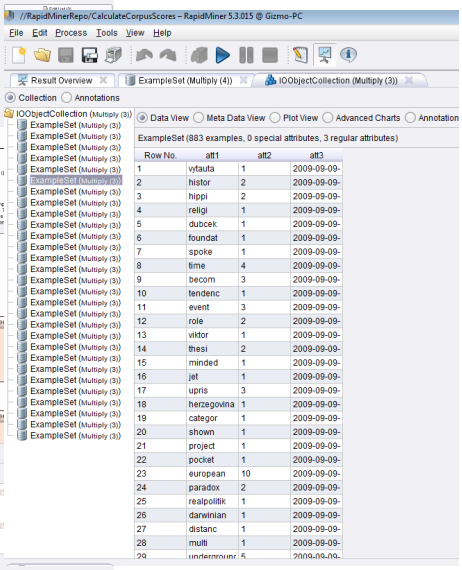
Fig. 3: Einblicke in RapidMiner



(a) Process



(b) Script



(c) Resultwindow

Abgebildet sind unter anderem ein Rapidminer-Prozess (eine Ansammlung von nacheinander ablaufenden Verarbeitungsschritten) - in diesem Fall ein Teil der Berechnungskette des 'Bag-Of-Words'. Darüber hinaus zeigt die Abbildung auch Ausschnitte eines Groovy-Skripts innerhalb der Rapidminer-Umgebung. Dieses und weitere selbst entwickelte Codefragmente erweitern die vorhandenen Fähigkeiten der Analyseumgebung im Sinne des Autors. Ebenfalls zu sehen ist eine Rapidminerergebnistabelle (konkret ein 'Bag-Of-Words' für ein Dokument).

Erste Erfolge

Die bisherigen Ergebnisse ermöglichten es eine Vorverarbeitung der Daten durchzuführen, diese umfasst u.A. eine Normalisierung der Daten, sowie ein Stemming. Darüber hinaus sind erste Versuche mit verschiedenen Distanzfunktionen samt einer vorherigen Gewichtung lauffähig.

Die Ergebnisse dieser Versuche werden mithilfe der Kriterien Precision und Recall (vgl. Abschnitt 3.3.3) evaluiert. Diese Werte werden zur visuellen Auswertung in Precision/Recall Diagramme übertragen (siehe Abbildung 4). Hierbei wird der Recall schrittweise gesteigert (d.h. mehr Dokumente gehen in die Berechnung der Ergebnisse ein), diese Methodik gehört zu den Standards der Bewertung von (sortierten) Information-Retrieval/Suchmaschinen Ergebnissen [MR02, vgl.]. Darüber hinaus wird auch die durchschnittliche Precision (bei steigendem Recall) eines Verfahrens zur Bewertung herangezogen.

Näheres zu den Ergebnissen des Projekts wird dem entsprechenden Projektbericht zu entnehmen sein.

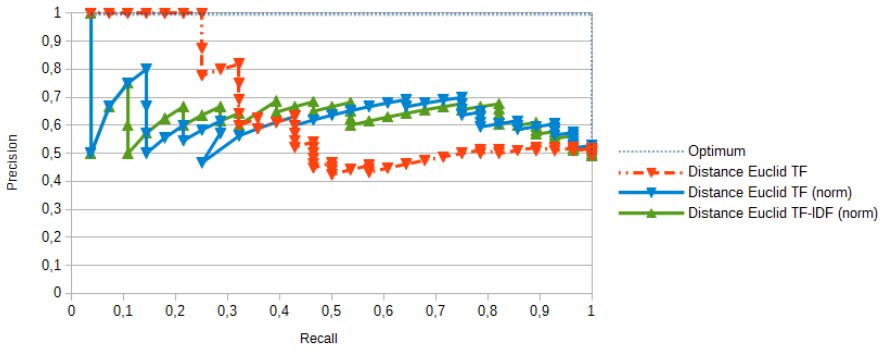


Fig. 4: Precision-Recall Verlauf mit ansteigendem Recall (x-Achse) (selbst erstellt)

Ein weiteres wichtiges Ergebnis der bisherigen Arbeit ist ein Grundverständnis der Analyseumgebung samt ihrer Möglichkeiten, Beschränkungen und Hürden. Dieses wird die zukünftige Arbeit am Projekt erleichtern.

5 Chancen und Risiken des Vorhabens

Der folgende Abschnitt widmet sich den Chancen, sowie den Risiken die dieses Projekt beinhaltet. Diese Liste stellt keine vollständige Aufzählung dar, sondern listet die dem Autor bekannten Kernpunkte auf.

5.1 Chancen

Betrachtet man das Projektziel und die vorhandene Literatur so stellt sich heraus, dass das Interesse an journalistischer Kontextualisierung nicht neu ist. Dieses zeigt sich u.A. durch diverse Vorschlagssysteme für journalistische Erzeugnisse [YLZ14, CR12, BDD+12, z.B]. Ein weiterer bedeutender Punkt kann durch den Ausspruch 'Editors don't scale' auf den Punkt gebracht werden. Journalisten skalieren nicht mit der Menge der Informationen welche sie verarbeiten müssen. Daher ist es nötig der Informationsflut entgegenzutreten - z.B. mit Mitteln der Informatik. Dass dieses Interesse real ist zeigt auch der Studiengang 'Next Media' (`http://nextmedia-haw.de`) der HAW Hamburg. Dieser, durch das Bundesministerium für Bildung und Forschung geförderte Studiengang, soll Journalisten die Möglichkeit bieten sich mit Methoden der Informatik vertraut zu machen um zukünftige Entwicklungen besser nutzen zu können.

Die Forschung des Autors soll den aktuellen Forschungserfolgen einen weiteren Baustein hinzufügen und hat hierzu die Chance auf eine gute Datenbasis zurückzugreifen.

5.2 Risiken

Durch den experimentellen Charakter des Projektes ergeben sich auch zahlreiche Risiken. Grundsätzlich ist, wie erläutert, die Abgrenzung des Begriffs Dossier nicht trennscharf [Hä15]. Darüber hinaus besteht auch die Gefahr, dass auch Experten keine einheitliche Meinung zur Definition eines Dossiers haben, ebenso denkbar ist auch, das diese ihr Wissen nicht konkret weitergeben können. Dieses könnte sich darin äußern, dass eine Dossiererstellung aus ihrer Sicht eher intuitiv geschieht, ohne dabei Regeln zu folgen welche der Autor nutzen kann. Das Problem der Befragung von Experten wurde u.A. in [McD83] beschrieben. Möglich wäre auch, dass die Fachexperten durchaus Informationen liefern können, diese allerdings zu unspezifisch (oder im Gegenteil: Zu komplex) sind als das diese sich mit Hilfe der Informatik realisieren ließen.

Grundsätzlich ist auch zu bedenken, dass diese Arbeit auf aufeinander aufbauenden und von einander abhängigen Teilen besteht, dieses bringt die Gefahr mit sich, dass beim Fehlschlag eines der Teile das gesamte Projekt in Mitleidenschaft gezogen wird.

Trotzdem können derartige Ergebnisse zumindest zu Teilerfolgen verhelfen und somit die Gesamtfragestellung um neue Erkenntnisse erweitern und zukünftige Arbeiten erleichtern, indem potenzielle Probleme aufgezeigt werden.

Literatur

[BDD+12] Bancu, Cristian ; Dagadita, Monica ; Dascalu, Mihai ; Dobre, Ciprian ; Trausan-Matu, Stefan ; Florea, Adina M.: ARSYS – Article Recommender System. In: **Proceedings of the 2012 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing**. Washington, DC, USA : IEEE Computer Society, 2012 (SYNASC '12). – ISBN 978–0–7695–4934–7, 349–355

[BOHG13] Bobadilla, J. ; Ortega, F. ; Hernando, A. ; GutiéRrez, A.: Recommender Systems Survey. In: **Know.-Based Syst.** 46 (2013), Juli, 109–132. http://dx.doi.org/10.1016/j.knosys.2013.03.012. – DOI 10.1016/j.knosys.2013.03.012. – ISSN 0950–7051

[CR12] Chhabra, Sidharth ; Resnick, Paul: CubeThat: News Article Recommender. In: **Proceedings of the Sixth ACM Conference on Recommender Systems**. New York, NY, USA : ACM, 2012 (RecSys '12). – ISBN 978–1–4503–1270–7, 295–296

[FPS96] Fayyad, Usama M. ; Piatetsky-Shapiro, Gregory ; Smyth, Padhraic: From Data Mining to Knowledge Discovery: An Overview. In: **Advances in Knowledge Discovery and Data Mining**. 1996, S. 1–34

[FS06] Feldman, Ronen ; Sanger, James: **The Text Mining Handbook**. Cambridge University Press, 2006 http://dx.doi.org/10.1017/CBO9780511546914. – ISBN 9780511546914. – Cambridge Books Online

[Hä14] Hälker, Nina: **Dienen textminingbasierte Dossiers dem Wachsen einer gemeinsamen europäischen Erzählung?** 2014

[Hä15] Hälker, Nina: **Masterarbeit**. 2015. – Arbeitspapier

[HRJM13] Harispe, Sébastien ; Ranwez, Sylvie ; Janaqi, Stefan ; Montmain, Jacky: Semantic Measures for the Comparison of Units of Language, Concepts or Entities from Text and Knowledge Base Analysis. In: **CoRR** abs/1310.1285 (2013). http://arxiv.org/abs/1310.1285

[KFD12] Kirshenbaum, Evan ; Forman, George ; Dugan, Michael: A Live Comparison of Methods for Personalized Article Recommendation at Forbes.Com. In: **Proceedings of the 2012 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II**. Berlin, Heidelberg : Springer-Verlag, 2012 (ECML PKDD'12). – ISBN 978–3–642–33485–6, 51–66

[LCLS10] Li, Lihong ; Chu, Wei ; Langford, John ; Schapire, Robert E.: A Contextual-bandit Approach to Personalized News Article Recommendation. In: **Proceedings of the 19th International Conference on World Wide Web**. New York, NY, USA : ACM, 2010 (WWW '10). – ISBN 978–1–60558–799–8, 661–670

[LGS] Lops, Pasquale ; Gemmis, Marco de ; Semeraro, Giovanni: In: **Recommender Systems Handbook**

[McD83] McDermott, John P.: Extracting Knowledge From Expert Systems. In: Bundy, Alan (Hrsg.): **IJCAI**, William Kaufmann, 1983, 100-107

[MR02] Manning, Christopher ; Raghavan, Prabhakar: Text Retrieval and Mining (CS27A) - Lecture 8 / Stanford. Version: 9 2002. https://web.stanford.edu/class/cs276a/handouts/lecture8.pdf. 2002. – Vorlesung

[Nik12] Nikolov, Stanislav: Trend or No Trend: A Novel Nonparametric Method for Classifying Time Series. (2012), 11. http://dspace.mit.edu/bitstream/handle/1721.1/85399/870304955.pdf

[Nor12] North, Matthew: **Data mining for the Masses**. 2012. – ISBN 978–0615684376

[PL10] Peng, Chi-Chieh ; Liu, Duen-Ren: Combining Reputation and Content-based Filtering for Blog Article Recommendation in Social Bookmarking Websites. In: **Proceedings of the 12th International Conference on Electronic Commerce: Roadmap for the Future of Electronic Business**. New York, NY, USA : ACM, 2010 (ICEC '10). – ISBN 978–1–4503–1427–5, 8–14

[PR-] **Precision-Recall diagram**. http://wikis.gm.fh-koeln.de/wiki_ir/uploads/InformationRetrieval/Recall/ir_bild.jpg

[Sch14] Schöneberg, Marcel: Erkennung von Trends in sozialen Netzwerken. Version: 10 2014. http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2014-proj/schoeneberg.pdf. 2014. – Projektbericht

[Sip14] Sippel, Sigurd: Recommendations for cocktail recipes. (2014), 11. http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2014-sem/sippel/folien.pdf

[TCY09] Tekli, Joe ; Chbeir, Richard ; Yetongnon, Kokou: Survey: An Overview on XML Similarity: Background, Current Trends and Future Directions. In: **Comput. Sci. Rev.** 3 (2009), August, Nr. 3, 151–173. http://dx.doi.org/10.1016/j.cosrev.2009.03.001. – DOI 10.1016/j.cosrev.2009.03.001. – ISSN 1574–0137

[YLZ14] Yang, Ming ; Li, Ying-ming ; Zhang, Zhongfei(Mark): Scientific articles recommendation with topic regression and relational matrix factorization. In: **Journal of Zhejiang University SCIENCE C** 15 (2014), Nr. 11, 984-998. http://dx.doi.org/10.1631/jzus.C1300374. – DOI 10.1631/jzus.C1300374. – ISSN 1869–1951