Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Seminararbeit

Nina Hälker

Text Mining für Newssites

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Nina Hälker

Text Mining für Newssites

Betreuender Prüfer: Prof. Dr. Kai von Luck

Eingereicht am: 14. September 2014

Inhaltsverzeichnis

1 Einleitung

1.1 Motivation

Data Mining, also das gezielte Durchsuchen der digitalen Kommunikation nach verwertbaren Inhalten, ist spätestens seit den Enthüllungen von Edward Snowden in aller Munde. Allerdings wird es nicht nur von Geheimdiensten, sondern auch von allen möglichen anderen netzaffinen Personen und Gruppen (z.B. im Gesundheitssektor, im Medienbereich, in der Wirtschaft, in NGOs) als Methode der Wissensgewinnung genutzt. Ein Teilbereich des Data Mining ist Text Mining. Text Mining bezieht sich ausschließlich auf das Auslesen von Textdaten.

Text Mining wird als Methode zur automatisierten Suche in und Aufbereitung von Texten derzeit im Bereich Journalismus diskutiert – unter anderem als Möglichkeit, innerhalb kürzester Zeit aus bestehenden Textarchiven neue Informationen zu extrahieren. Diese systematische, strukturierte, schnelle – und dadurch oftmals effektive – Möglichkeit, Textarchive nach noch unbekannten Informationen, Mustern etc. zu durchsuchen, birgt für den Bereich Journalismus Möglichkeiten, die noch keinesfalls voll ausgeschöpft sind.

In der vorliegenden Arbeit setze ich mich speziell mit Text Mining auseinander, da ich im Rahmen meines Studiums ein großes Interesse an der automatischen Auswertung von Texten gewonnen habe – mit dem Ziel, durch Algorithmen sinnvolle Dossiers erstellen zu können.

1.2 Ziel der Arbeit

Die Arbeit vermittelt einen Einblick in das Feld des Text Mining und zwei Anwendungsgebiete. Angerissen wird zudem die Frage, ob sich Aspekte der noch zu skizzierenden Ansätze sich auch für die Aufbereitung eines Artikelarchivs eignen könnten, für das ich im Rahmen der Masterarbeit nach geeigneten Text Mining Methoden suchen und sie testen werde. In der Masterarbeit soll untersucht werden, inwieweit Methoden des Text Mining geeignet sind (und welche Methoden), sinnhafte Dossiers zu erstellen und, welche Möglichkeiten sowie Schwierigkeiten sich dabei zeigen. Diese Arbeit ist damit als Einführung und Grundlage für die in den folgenden Monaten zu vertiefende Auseinandersetzung mit dem Thema im Rahmen der Masterarbeit zu betrachten.

2 Hauptteil

In dieser Arbeit werden drei Forschungspapiere vorgestellt. Aufgrund der uneinheitlichen Verwendung des Begriffs Text Mining wird zunächst eine Definition des Begriffs Text Mining vorgenommen, der sich Kroeze u. a. [2003] in ihrer Untersuchung »Differentiating Data- and Text-Mining Terminology« widmen. Im Anschluss daran werden zwei Ansätze zur automatisierten Extraktion von Texten vorgestellt: Die Studie von Agarwal u. Liu [2008] unter dem Titel »Blogosphere: Research Issues, Tools, and Applications« befasst sich mit dem Einsatz von Text Mining zur Analyse von Blogs bzw. Blognetzwerken, in der Studie von Ritter u. a. [2012], »Open Domain Extraktion from Twitter«, werden Möglichkeiten und Schwierigkeiten bei der automatisierten Analyse von Tweets vorgestellt.

2.1 Text Mining vs. Data Mining – Begriffsdefinitionen

Anlässlich der unterschiedlichen und oftmals ungenauen Verwendung von Begriffen wie Text Mining, Data Mining und Knowledge Discovery haben sich Kroeze u. a. [2003] mit der Geschichte und der Entwicklung dieser Begriffe auseinandergesetzt. Ausgehend von einem grundlegenden Artikel von Hearst – Untangling Text Data Mining – aus dem Jahr 1999, in dem dem Problem nachgegangen wird, Konzepte und Begriffsnutzungen zu klären, ist das Anliegen von Kroeze u.a., Hearsts Ausführungen sowohl kritisch zu hinterfragen als auch weiter zu präzisieren.

Während der Begriff Knowledge Discovery den gesamten Prozess der Gewinnung von Wissen durch das Auslesen von Daten bezeichnet – angefangen bei der Beschaffung und Bereinigung der Daten bis hin zum Auslesen und dem Sammeln der Ergebnisse – wird Data Mining als ein Teilbereich der Knowledge Discovery betrachtet: Data Mining bezeichnet das Durchsuchen riesiger Datenmengen nach Mustern – wobei zu Beginn der Suche noch unklar ist, was gefunden werden könnte. Text Mining wiederum ist als ein Teilbereich des Data Mining zu sehen, der sich ausschließlich mit dem Auslesen von textlichen Inhalten befasst. Kroeze u. a. [2003] definieren Text Mining wie folgt:

> The discovery of knowledge from databases sources containing free text is called text mining.

Das Interesse und der Zweck von Text Mining kann in etwa wie folgt umrissen werden:

> Tell me something I didn't know but would like to know.

Aufgabe des Text Mining ist es, Texte unterschiedlichster Formate auszulesen und Informationen aus diesen Daten zu extrahieren. Die Aquise bzw. Beschaffung von neuem, nützlichem, bedeutsamen und gültigen Wissen im Sinne von Knowledge Discovery kann also durch Methoden des Text Mining geschehen.

Auch Cios u. a. [2007] merken an, dass, wenngleich sowohl in der Theorie als auch in der Praxis Data Mining oft synonym mit Knowledge Discovery verwendet wird, der Begriff Data Mining lediglich einen Schritt im Prozess der Knowledge Discovery beschreibt. Ihnen zufolge ist Knowledge Discovery der Oberbegriff für Methoden, die dem Auffinden von bislang unbekannten Informationen dienen.

Hippner u. Rentzmann [2006] verwenden den Begriff Data Mining ebenfalls als Oberbegriff für spezifischere Ausprägungen wie z.B. Text Mining. Als wichtigen Unterschied zwischen Data- und Text Mining benennen sie die sich voneinander unterscheidende Datenbasis: Während beim Data Mining strukturierte Daten analysiert werden, arbeite man beim Text Mining, wie der Name schon vermuten lässt, mit Textdateien, die zumeist als semi- oder unstrukturierte Daten vorliegen. Hippner und Rentzmann präzisieren ihre Ausführungen wie folgt:

> [Textdateien verfügten überwiegend] über eine implizite Struktur, die aus der Grammatik resultiert, und – je nach Textdokument – über eine explizite Struktur, die sich z.B. aus Titel/Untertitel und Absätzen erschließen kann.

Kroeze u. a. [2003] weisen darauf hin, dass es im Bereich der Knowledge Discovery eine Erweiterung des Forschungsfeldes gegeben habe: Vom (nahezu) ausschließlichen Auslesen von strukturierten Datenbanken habe sich das Feld in jüngerer Zeit um das Auslesen von (semi- bis unstrukturierten) Textdatenbanken erweitert und sei damit zu einem umfassenderen Forschungsfeld geworden:

> Until recently computer scientists and information system specialists concentrated on the discovery of knowledge from structured, numerical databases an data warehouses. However, much, if not the majority, of available business data are captured in text files that are not overtly structured.

Zu den Unklarheiten der Begriffsdefinition gehört Kroeze u. a. [2003] zufolge auch, dass der Begriff Text Mining oftmals so gelesen würde, dass er entweder das Entdecken (*discovery*) bzw. Auffinden von Texten in den Mittelpunkt stelle oder das Erforschen von Texten (*exploration*) auf der Suche nach wertvollen, bislang verborgenen Informationen. Text Mining bezeichnet ihnen zufolge jedoch beides: eine Dokumentensammlung zu untersuchen und dabei Informationen zu finden, die nicht in einem einzelnen Dokument enthalten sind. Es handelt sich also um etwas, das erst durch das Bündeln bzw. die Synthese von Daten möglich wird.

Zweck und Möglichkeiten von Text Mining unterscheiden sich damit vom (allgemeineren) Bereich des Information Retrieval: Während es bei Letzterem bereits ein Wissen um die gesuchte Information gibt, also schon vor dem Suchprozess bekannt ist, nach was gesucht wird, gibt es dieses Wissen beim Text Mining nicht – im Vordergrund steht dort, aus den vorhandenen Daten neues Wissen und neue Erkenntnisse aufgrund der Erkennung von Mustern zu generieren. Während also die Aufgabe beim Information Retrieval darin besteht, die gesuchte Information (bzw. ihren Ort) zu finden, ist die Aufgabe beim Text Mining, neue Informationen zu finden.

Kroeze u. a. [2003] führen die Differenzierung der Begriffe insofern weiter aus, als dass sie einen Unterscheidung zwischen non-novel-, semi-novel- und novel Text Mining Investigation treffen, wobei sie diese Kriterien den Bereichen Information Retrieval, Knowledge Discovery und dem von ihnen vorgeschlagenen neu zu besetzenden Begriff Knowledge Creation zuweisen. Sie kommen zu folgendem Fazit:

> If non-novel text-mining investigation is information retrieval, and if semi-novel text-mining investigation is knowledge discovery, then novel text-mining investigation should be knowledge creation.

Kroeze u. a. [2003] gehen davon aus, dass novel Text Mining Investigation vollständig neues Wissen hervorbringt – im Sinne von Knowledge Creation – und dass das eine Fähigkeit ist, die üblicherweise von Menschen vollbracht wird, nicht von Maschinen. Da sich das Ergebnis der novel Text Mining Investigation stark von den Ergebnissen, die durch Information Retrieval und durch Knowledge Discovery erzielt werden, unterscheidet, schlagen eine neue Begrifflichkeit vor: Intelligent Text Mining. Weil der Prozess der Knowledge Creation sich zwingend die bestehenden Möglichkeiten künstlicher Intelligenz zunutze macht, die menschliche Intelligenz zu simulieren versucht, schlagen die Autoren den Begriff Intelligent Text Mining vor. Kurz gehen sie in diesem Zusammenhang ein auf die Funktion des Natural Language Processing (NLP) im Intelligent Text Mining: die inhärente Struktur von Texten muss aufgedeckt und die darunter liegenden linguistischen Strukturen müssen erforscht werden, um daraus

syntaktische und semantische Darstellungen des Textes zu ermöglichen. Der Einsatz von Natural Language Processing im Text Mining ermöglicht, Diskurse bzw. Diskussionsverläufe zu entdecken und die Struktur von Texten jenseits ihres offensichtlichen Gehalts zu untersuchen. Für die später noch anzureißende Frage, ob und welche Text Mining Methoden sich für automatisierte Dossiererstellung eignen könnten, ist dieser Aspekt von Interesse. Beispielsweise könnten Texte dahingehend untersucht werden, ob sie der Popkultur oder der Wissenschaft zugeordnet werden können. Texte könnten ggfs. auch auf Kriterien untersucht werden, die Aufschluss darüber geben könnten, ob ein Text aus Ost- oder aus Westeuropa kommt – möglicherweise ist dies durch die Verwendung bestimmter Begriffe, etc. herauszufinden.

2.2 Text Mining zur Untersuchung der Blogosphäre

Kennzeichnend für das Web 2.0. ist die niedrige Publiktionsbarriere: Ohne technisches Wissen können Menschen im WWW interagieren – Webseiten bauen, Bloggen und sich in Social Media Netzwerken jedweder Couleur beteiligen. Aus Content-Konsumenten wurden Content-Produzenten. Insbesondere Blogs bieten seit längerer Zeit durch die niedrigschwellige Art der Veröffentlichung vielfältige Möglichkeiten, Informationen, Wissen und Meinungen einer breiteren Öffentlichkeit zur Verfügung zu stellen.

Agarwal u. Liu [2008] untersuchten Ansätze zur Analyse der Blogosphäre. Blogosphäre bezeichnet in diesem Zusammenhang das Universum existierender Blogs – sowohl individuelle (single-author-) als auch community (multi-authored-) Blogs. Die massive Nutzung von Blogs führte zu einer Zunahme kollektiver Wissensproduktion. Agarwal u. Liu [2008] sprechen in diesem Zusammenhang von einer Art »open source intelligence« und von kollektiver Weisheit. Die Autoren gehen der Frage nach, welche Methoden eine Analyse der Blogosphäre ermöglichen, so dass das kollektiv produzierte Wissen systematisch geborgen werden kann. Sie benennen dabei sieben verschiedene Forschungsfacetten – Modeling, Clustering, Mining, Community Discovery and Factorization, Influence and Propagation, Trust and Reputation und Spam Blog Filtering. Die drei aus meiner Sicht bedeutsamsten Ansätze skizziere ich im folgenden Teil:

1. **Modeling:** Das Modellieren der Blogosphäre im Sinne einer Modellerstellung dient Agarwal u. Liu [2008] zufolge dazu, einen genaueren Einblick in die Beziehungen zwischen Blogs und in die Interaktion der Blogger untereinander zu bekommen. Darstellbar ist das Ergebnis einer solchen Analyse beispielsweise als Graph, durch den Nähe und Distanz der Blogs und Blogger untereinander visualisiert werden können. Die Schwierigkeit, die diese Art der Darstellung jedoch birgt, ist die der Nichtdarstellbarkeit der Pro-

zesshaftigkeit der Interaktion zwischen Bloggern und der kurzfristig stattfinden könnenden Verschiebungen, welche von Agarwal u. Liu [2008] folgendermaßen skizziert werden:

> ... the highly dynamic and and »short-lived« nature of the blog posts could not be simulated by the web models.

Mit einem Graphen können diese kurzlebigen Dynamiken nicht in adäquater Weise dargestellt werden. Diese sind jedoch ein grundlegendes Merkmal der Blogosphäre, da sich die Verhältnisse zueinander, die Positionierung einzelner Blogs in der Blogosphäre grundlegend verschieben können. So müsste unbedingt auch die Temporalität der jeweiligen Position eines Blogs im Netzwerk abbildbar sein.

2. Clustering: Eine übliche Methode seitens der Blogger, Blogs bzw. Beiträge zu verschlagworten, ist das Taggen von Blogposts. Tags ermöglichen so eine thematische Strukturierung innerhalb eines Blogs – ähnlich einem Schlagwortverzeichnis in einem Buch. Vergleichbar, nur automatisiert und blogübergreifend, funktioniert das Clustern von Blogs in sinnhafte Cluster bzw. zumeist thematisch sortierte Gruppen. Agarwal u. Liu [2008] weisen in puncto Clustern auf das Problem hin, dass

> human labeled tags are are good for classifying the blog posts into broad categories while they were less effective in indicating the particular content of a blog post.

Agarwal u. Liu [2008] weisen in ihrer Studie auf eine Untersuchung hin, bei der alternativ zum oben beschriebenen Taggen die Methode tf-idf (term frequency-inverse document frequency) eingesetzt worden ist, bei der die drei meistbenutzten Wörter eines jeden Blogposts extrahiert und analysiert werden. tf-idf wird auch von Overview genutzt, einer open source Software, die zum automatischen Durchsuchen großer Datenmengen und deren Aufbereitung entwickelt wurde. Auf der Website werden Anwendungsmöglichkeiten wie folgt beschrieben:

> Read and analyze thousands of documents super quickly. Full text search, topic modeling, coding and tagging, visualizations and more. All in an easy-to use, visual workflow.

Das Ergebnis des Einsatzes von tf-idf in der von Agarwal u.a. zitierten Studie zeigte, dass das automatische Clustern der Blogposts eine sehr viel bessere Aussagekraft hinsichtlich der konkreten Inhalte der einzelnen Artikel aufwies als die human labeled Blogposts.

Das sinnhafte Clustern von Blogs wurde in einer weiteren Studie, auf die die Autoren verweisen, durch eine unterschiedliche Gewichtung der verschiedenen Teile eines jeden Beitrags – Titel, Inhalt und Kommentare – versucht: Doch auch Ansätze wie dieser kommen aufgrund der Schlagwortbasierung (bzw. -fixierung) schnell an ihre Grenzen. Als eine empfehlenswerte Methode zum Clustern von Blogs nennen die Autoren WisClus, eine Methode, die sich das bereits erwähnte kollektive Wissen bzw. die kollektive Weisheit der Bloggercommunity in Kombination mit Schlagworten zunutze macht:

> They … construct the *category relation graph* to merge different categories and cluster the blogs that belong to these categories. … The similarityy between two categories is computed using the number of blogs that simultaneously uses these categories as their blog labels.

So kann die Ähnlichkeit bzw. Nähe unterschiedlicher Kategorien in Form eines Graphen komplexer visualisiert werden als es z.B. beim vorgestellten Modeling möglich ist. Die Autoren betonen, dass die Ergebnisse bei Ersterem deutlich besser ausfallen, da das Clustern, das collective wisdom berücksichtigt, präziser ist als das keywordbasierte Clustern.

3. Influence in Blogs and Propagation: Produktentscheidungen werden in den meisten Fällen infolge von Empfehlungen durch Freunde, Familie oder Experten getroffen – diese gängige Praxis, die bekannt ist als »word of mouth«, ist für den Bereich Produktmarketing sehr wichtig: Sogenannte Influencer sollen ausfindig gemacht werden, um so den Kontakt zum potenziellen Käufer zu bekommen. Auch viele Blogger gelten als Experten für den Themenbereich, über den sie bloggen. Verschiedene Studien haben untersucht, mit welchen Methoden einflussreiche Blogger ausfindig gemacht (und gezielt angesprochen) werden können und worin genau der Einfluss von Bloggern besteht: Sind es die direkten Empfehlungen in Form von expliziten Blogposts, sind es Links im Sinne von direkten Produktempfehlungen oder handelt es sich um andere Aspekte (wie Leserkommentare o.a.), die den Einfluss ausmachen? Agarwal u. Liu [2008] fassen die Aufgabe, die dabei gestellt wurde, folgendermaßen zusammen:

> Some try to find influential blog sites in the entire blogosphere and study how they influence the external world and within the blogosphere.

Agarwal u. Liu [2008] bilanzieren, dass es schwierig ist, einflussreiche Blogs bzw. Blogger – also die Influencer bzw. market-mover – ausfindig zu machen. Gründe dafür sind, dass es sich bei der Rezeption von Blogs um eine Art long tail Nutzung handelt. Ebenso betonen sie, dass der Einfluss von Bloggern – so die Ergebnisse der Studien – nicht

zwangsläufig mit ihrer Aktivität zusammenhängt, also einflussreiche Blogger nicht unbedingt viel und täglich bloggen. In die Analyse über den Einfluss von Blogs muss zwingend eine komplexe Gewichtung von Themen, Kommentaren, Häufigkeit des Bloggens, Verlinkung der Blogs u.a. Aspekte eingehen. Dazu komme, dass aufgrund des schnellen Wachstums der Blogosphäre es zunehmend schwieriger sei, die Blogosphäre als Ganzes eingehend zu analysieren.

2.3 Text Mining zur Analyse von Twitter

Auch aus Twitter können Informationen extrahiert und zu neuem Wissen zusammengefügt werden. Tweets als Datenbasis für den Einsatz von Miningverfahren zu nutzen, hat Vor- und Nachteile. Zunächst: Aufgrund der von Twitter zur freien Verfürgung gestellt API ist es problemlos möglich, Daten jenseits individueller Timelines zu nutzen. Ein Grund, Tweets als Datenbasis zu nutzen, ist Ritter u. a. [2012], die sich mit dem Thema »Open Domain Event Extraction from Twitter« auseinandergesetzt haben, zufolge:

Die Kürze: Tweets zwingen dazu – aufgrund der auf maximal 140 Zeichen begrenzenten Länge – selbst komplexere Themen in wenigen Worten auf den Punkt zu bringen.

Zwei weitere Vorteile, die Ritter u. a. [2012] neben der Kürze von Tweets nennen, sind:

Die einfache Struktur: Die Struktur von Tweets ist häufig simpel und dadurch u.U. gut zu analysieren. Tweets sind (meistens) in klarer, einfacher Sprache geschrieben, der Satzaufbau ist oftmals klassisch aufgeteilt in Subjekt - Prädikat - Objekt.

Die hohe Anzahl der in kurzer Zeit zu einem Thema versendeten Tweets: Das Volumen der in einem kurzen Zeitintervall zu einem Thema versendeten Tweets ist um ein Vielfaches höher als z.B. die Anzahl der zum gleichen Thema in kurzer Zeit erscheinenden Artikel auf Newssites. Insbesondere bei Großveranstaltungen twittern oft zeitgleich sehr viele Menschen über eine identische Situation – ein Umstand, der sich gut dazu eignet, Keywords zu extrahieren, da davon auszugehen ist, dass es eine nennenswerte Häufung von bestimmten Begriffen geben wird.

Alle drei Aspekte tragen dazu bei, dass Tweets eine gute Datengrundlage liefern, wenn es darum geht, eine semantische Analyse vorzunehmen. Insbesondere zum Aspekt der Menge lässt sich für die Analyse festhalten, dass die durch die zu vermutende so entstehende inhaltliche Redundanz in Form vieler nahezu identischer Tweets die Auswertungsmöglichkeiten unterstützt.

Festgehalten werde muss allerdings, dass die genannten Vorteile auch Nachteile bergen (oftmals die Kehrseite der eben genannten Vorteile), die für die Analyse von Tweets zusätzliche Herausforderungen bedeuten:

Zerstückelter Inhalt: Die Informationen in Tweets sind oftmals zerstückelt und ungenau – auch als »noisy« bezeichnet.

Unwichtige Informationen: Neben der Tatsache, dass insbesondere bei grossen Events, Breaking News, etc. viele nahezu identische Tweets versendet werden, gibt es immer auch eine nicht unerhebliche Anzahl an Tweets, deren Inhalt unwichtig (im Sinne von »nicht zum analysierten Thema passend«) ist.

Verknappte Sprache: Die erzwungene Kürze von Tweets bringt oftmals eine verknappte Sprache mit sich, die den Inhalt schwerer auslesbar machen. Dies können Abkürzungen sein, aber auch das Weglassen von Vokalen.

Festgehalten werden kann also, dass zu Beginn der Analyse Tweets mit zerstückeltem und themenfremdem Inhalt sowie verkürzter Sprache im Zuge der Analyse zunächst erkannt und dann herausgefiltert werden müssen.

Das Verfahren, das Ritter u. a. [2012] beschreiben, besteht unter anderem darin, Tweets POS zu taggen - das »part of speech«-Taggen ist als ein Abstraktionsschritt zu verstehen, um die Tweets maschinenlesbar zu machen. Es werden Begriffsentitäten und Event-Phrases extrahiert, zeitliche Ausdrücke beseitigt und die Events in Typen kategorisiert. Die Aussagekraft eines Tweets berechnet sich durch die (berechnete) Stärke der Verbindung zwischen der jeweiligen Begriffsentität, dem Datum und der Anzahl der gesendeten Tweets zu einem Begriff bzw. einem Thema.

Ritter u. a. [2012] beschreiben in ihren Ausführungen TwiCal, das erste open domain event-extraction and categorization system für Twitter. TwiCal arbeitet mit einer »4-Tupel representation« von Events und zerlegt Tweets in die folgenden Teile: entity, event phrase, date, type. Der Grund für diese Aufteilung liegt in der bereits erwähnten oftmals simplen Struktur von Tweets (Subjekt, Prädikat, Objekt) und ermöglicht so eine Analyse der einzelnen Teile eines Tweets.

3 Fazit und Ausblick

Die vorgestellten Text Mining Methoden vermitteln einen Eindruck davon, mit welchen Schwierigkeiten und Herausforderungen Text Mining umgehen muss. Im Hinblick auf die Masterarbeit ergibt sich daraus die Aufgabe, vor dem Einsatz einer bestimmten Methode zwingend zu klären, was der Zweck der Datenanalyse ist – also, was genau mittels Text Mining herausgefunden werden soll. Die vorgestellten Methoden betrachtend erscheint mir eine genauere Kenntnis der erwähnten Methode tf-idf für das Vorhaben einer Aufbereitung eines Artikelarchiven hin zu Dossiers sinnvoll. Auch Aspekte der Analyse von Tweets können ggfs. für die Masterarbeit hilfreich sein: Perspektivisch muss ich untersuchen, ob als Textkorpus für das Text Mining und das Ziel der Dossiererstellung bereits die Artikel-Header ausreichend sind oder ob andere Teile der Artikel zur Analyse herangezogen werden müssen (z.B. Tags, Abstracts oder ganze Artikel), um die Sinnhaftigkeit der zu erstellenden Dossiers zu garantieren. Beide vorgestellten Methoden zeigen, dass eine gute (differenzierte) Struktur des Textkorpus das Mining vereinfacht und zu (qualitativ) besseren Ergebnissen führt.

Literaturverzeichnis

[Agarwal u. Liu 2008] Agarwal, Nitin ; Liu, Huan: Blogosphere: Research Issues, Tools, and Applications. In: *SIGKDD Explor. Newsl.* 10 (2008), Mai, Nr. 1, 18–31. http://dx.doi.org/10.1145/1412734.1412737. – DOI 10.1145/1412734.1412737. – ISSN 1931–0145

[Cios u. a. 2007] Cios, K.J. ; Pedrycz, W. ; Swiniarski, R.W. ; Kurgan, L.: *Data Mining. A Knowledge Discovery Approach.* 2007

[Hippner u. Rentzmann 2006] Hippner, Hajo ; Rentzmann, René: Text Mining. (2006). https://www.gi.de/service/informatiklexikon/detailansicht/article/text-mining.html

[Kroeze u. a. 2003] Kroeze, Jan H. ; Matthee, Machdel C. ; Bothma, Theo J. D.: Differentiating Data- and Text-mining Terminology. In: *Proceedings of the 2003 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on Enablement Through Technology.* Republic of South Africa : South African Institute for Computer Scientists and Information Technologists, 2003 (SAICSIT '03). – ISBN 1–58113–774–5, 93–101

[Ritter u. a. 2012] Ritter, Alan ; Mausam ; Etzioni, Oren ; Clark, Sam: Open Domain Event Extraction from Twitter. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York, NY, USA : ACM, 2012 (KDD '12). – ISBN 978–1–4503–1462–6, 1104–1112