Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Seminararbeit

Jan Dennis Bartels

Text Mining

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Jan Dennis Bartels

Text Mining

Betreuender Prüfer: Kai von Luck

Eingereicht am: 28.02.2019

Inhaltsverzeichnis

1 Einleitung

Die meisten gespeicherten Informationen des Internets liegen in Textform vor, diese Informationen sind für konventionelle Data Mining Prozesse unbrauchbar da sie keine nutzbare Struktur besitzen.
Diese kaum genutzten Wissensrohstoffe werden von sogenannten Text Mining Prozessen aufbereitet um neue Informationen aus ihnen zu extrahieren. Heyer u. a. (2006)

1.1 Ziel der Arbeit

Diese Seminarbeit soll einen Einstieg in den Themenbereich des Text Mining bieten. Sie dient dazu den Themenbereich des kommenden Grundprojektes zu erforschen, das Grundprojekt soll sich der Anwendung von Text Mining mit zuhilfenahme eines lernenden Algorithmus widmen.

1.2 Begriffsdefinition

> Text mining [is a] process in which a user interacts with a document collection over time by using a suite of analysis tools.
> ..text mining seeks to extract useful information from data sources through the identification and exploration of interesting patterns.
>
> *Definition von Text Mining nach Feldman und Sanger (2006)*

Der Begriff Text Mining ist nicht eindeutig definiert, nach der Definition von Feldman und Sanger (2006) kann Text Mining als Prozess verstanden werden bei dem mithilfe von Analyse Tools interessante Muster erkannt werden und dieser zur Gewinnung nützlicher Information genutzt werden.

1.3 Abgrenzung zum Data Mining

Text Mining ist eine Art des Data Minings, der wesentliche Unterschied hierbei ist die Datenherkunft. Während beim Data Mining die Daten zumeist in struktorierter Form in Datenbanken oder anderen Datenquellen vorhanden sind, gibt es bei Texten keine derart starke Struktur. Textstrukturen sind nur durch die Regeln der Gramatik und Einteilungen in Blöcke, wie Paragraphen oder Kapitel, vorhanden. Für Computer sind diese Strukturen unbrauchbar und müssen vor der Verarbeitung vorbereitet werden. Cleve und Lämmel (2014)

1.3.1 Abgrenzung zum Web Mining

Web Mining ist die Anwendung Data Mining Methoden auf Daten aus dem Internet. Hierbei können verschiedene Themengebiete erforscht werden.

Web Content Mining Analyse von textuellen und multimedia Inhalten, sowie der Verlinkung zwischen Webseiten.

Web Usage Mining Analyse des Verhaltens von Nutzern des Internets, basiert oft auf Protokolldaten des Webservers. Es wird unterschieden zwischen ausschließlicher Analyse von Protokollen (Web Log Mining) und dem zusätzlichem Heranziehen anderer Datenquellen (Integrated Web Usage Mining)

Herausforderungen

Neben der fehlenden Datenstruktur gibt es beim Text Mining noch weitere abweichende Herausforderungen im Gegensatz zum Data Mining:
Zum einen gibt es viele Variationen wie man einen Satz aufbaut, verschiedene Satzformen verändern nicht unbedingt den Informationsgehalt.
Ein weiteres Problem stellen Synonyme und die Umgangssprache dar, während einem Mensch das Wort ”Sonnabend” als Abart des Wortes ”Samstag” geläufig ist kann ein Computer derartige Semantische Gleichheit nicht ohne weitere Hilfe vestehen. Cleve und Lämmel (2014)
Wortstämme sowie Vergangenheitsformen können von Algorithmen nicht ohne Vorarbeit erkannt werden. Als Beispiel können die Worte ”Connection”, ”connected” und ”connecting” allesamt auf das Verb ”connect” zurückgeführt werden.
EIne weitere Herausforderung mit der sich das Themenfeld des Text Mining auseinandersetzt ist die Rechtschreibung, falsch geschriebene Wörte könnten fehlerhaft interpretiert werden.

1.4 Anwendungsfälle von Text Mining

Ein paar Anwendungsfälle von Text Mining sind folgende:

Customer care Zur Unterstützung von Supportmitarbeitern könnte Text Mining eingesetzt werden um automatisch ähnliche Supportanfragen zu finden. Dieser Vorgang könnte bei entsprechender Datenmenge auch vollautomatisch ablaufen um häufig auftretende Supportanfragen ohne einen Supportmitarbeiter zu beantworten.

Social media In Social Media Kanälen ist es aufgrund der enormen Datenmengen oft nicht möglich für das Unternehmen relevante Kommentare bzw. Reaktionen von Nutzern zu erkennen. Hier könnte die Technologie helfen um Stimmungsbilder zu erzeugen oder unrelevante Daten zu filtern.

Sicherheit / Anti Terror Text Mining lässt sich auch für eine Gefahreneinschätzung nutzen, hierbei könnten beispielsweise Staatliche Einrichtung Social Media Kanäle oder erbeutete Kommunikationsprotokolle analysieren um terroristische Aktivitäten aufzudecken.

Spam Erkennung Text Mining könnte genutzt werden um eine automatische Filterung von unseriösen Emails vorzunehmen. Hierzu könnte ein Algorithmus antrainiert werden welcher den Schreibstil oder Informationsgehalt von Spam erkennt und so einschätzt ob eine gegebene Mail seriös ist.

2 Hauptteil

2.1 Ablauf von Text Mining

Abbildung 2.1: Text Mining Prozess nach Hippner und Rentzmann (2006)

Der Ablauf von Text Mining nach Hippner und Rentzmann (2006) gliedert sich wie folgt:

Aufgabendefinition Festlegung der Problemstellung und der Ziele des Text Mining

Dokumentselection Ausgehend von den jeweiligen Zielen müssen potentiell relevante Dokumente identifiziert werden.

Dokumentaufbereitung Die Aufbereitung der Dokumente stelle eine der größten Herausforderungen des Text Mining dar, dieser Vorgang wird in 2.2 näher erläutert.

(Text) Mining Methoden Nach der Aufbereitung können abwandlungen der regulären Data Mining Methoden auf die Terme angewand werden. Dies wird in 2.3 näher vorgestellt.

Interpretation / Evaluation Bewertung handlungsrelevanter Ergebnisse des Minings

Anwendung Anwendung der Ergebnisse, beispielsweise reagieren auf erkannte Kunden oder Marktentwicklungen.

2.2 Linguistische Datenaufbereitung

Dieser Ansatz des Text Mining konstruiert eine fehlende Datenstruktur, damit diese in späteren Schritten von Text Data Mining Methoden genutzt werden kann.

Ziel dieser Aufbereitungsschritte ist es ebenfalls die Komplexität des Textes zu verringern, hierfür werden sich Techniken des Themenfeldes Natural language Processing genutzt.

2.2.1 Morphologische Analyse

Während der Morphologischen Analyse werden Worte auf ihren Wortstamm zurückgeführt sowie Flexionsformen aufgelöst.
Populäre Ansätze zur Wortstammreduzierung sind zum einem das Stemming, zum anderen die Lemmatization.

Stemming

Der Vorgang des Stemming versucht durch die gezielte Entfernung von Wortteilen, wie Präfixen und Suffixen, ein gegebenes Wort auf seinen Wortstamm zu reduzieren.
Ein bekannter Algorithmus ist der von Martin F. Porter, der Algorithmus entfernt häufig genutzte Suffixe beziehungsweise formt diese um. Porter (1997)
Ein Problem das diese Herangehensweise mit sich bringen kann ist die Erzeugung von falschen Worten, es ist möglich dass Wörter entstehen die es so nicht gibt oder das Wörter mit falscher Bedeutung interpretiert werden. Wird beispielsweise das Wort ”designate” von einem Stemming algorithmus verarbeitet so könnte der Algorithmus das Wort auf ”design” runterbrechen. Die Bedeutung ist aber eine ganz andere: Während ”designate” bedeutet etwas zu bestimmen ist die Bedeutung des Verbs ”design” etwas zu entwerfen oder planen.

Lemmatization

Ein anderer Ansatz ist der Lemmatization-Ansatz, hierbei wird eine Art Wörterbuch genutzt um verschiedene Wortformen und Wörter gleicher Bedeutung auf eine Grundform zurückzuführen. Vorteile dieses Ansatz ist das es praktisch keine falschen Rückführungen gibt, allerdings ist dieser Ansatz nur so gut wie die zugrundeliegende Datenbasis: Alle Wortformen müssen definiert werden.

2.2.2 Syntaktische Analyse

In der Syntaktischen Analyse geht es eine Einteilung der Wörter in ihre Satzbausteine. Hierzu wird ein sogenannter Part-of-Speech-Tagger genutzt, dieser basiert in der Regel auf einem Lexica mit Definitionen welche Wortarten ein Wort annehmen kann, sowie Daten über kontextabhängige Wortarten.
Auf diesen annotierten Satzbausteinen setzt ein Parser an welcher die Stellung der Wörter im Satz ermittelt.

2.2.3 Semantische Analyse

Letzlich gibt eine Semantische Analyse rückschlüsse auf Kontextbasiertem Wissen, nehme man das Wort ”Bank”, dies kann je nach Kontext eine Sitzgelegenheit oder ein Geldinstitut sein.

2.3 Text Mining Methoden

Klassifikation

Klassifikation bescheibt den Vorgang wenn Texte oder Textbausteine anhang gegebener Regeln in vorgegebene Kategorien eingeteilt werden. Hierbei könnte eine Einteilung nach Fachgebieten oder fachlichem Niveau von Dokumenten erfolgen. Feldman und Sanger (2006)

Clustering

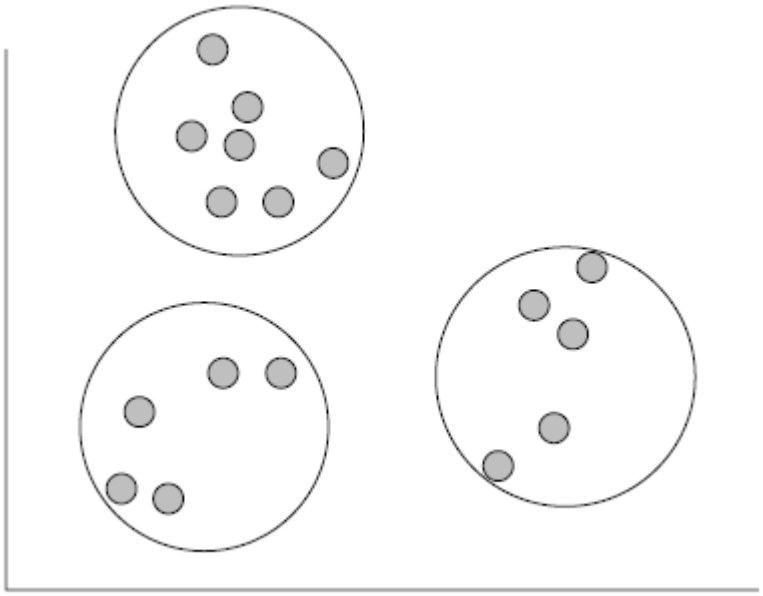

Abbildung 2.2: Visualisierung von Clustern

Unter Clustering versteht man das Verfahren vobei Textsammlungen in Untermengen unterteilt werden. Diese Teilmengen enthalten Texte welche sich auf Grund von gegebenen Kriterien ähneln. Im Umkehrschluss bedeutet dies, das Texte aus verschiedenen Clustern sich stark voneinander unterscheiden lassen.

Unterscheiden lässt sich dieses Verfahren zur Textklassifikation darin, dass die Klassifikationsschemen beim Clustering nicht vorgegeben sind.
Als ein Anwendungsbeispiel für Clustering könnte eine Extraktion von Textmerkmalen stattfinden. Diese Textmerkmale könnten zum Beispiel der fachliche Inhalt von Texten oder der Schreibstil sein.
Feldman und Sanger (2006)

Musteranalysen

Für die Analyse von Texten ohne Struktur sowie zuvor annotierten Texten kann ein Musterbasierendes Verfahren vorteilhaft sein. Grundlage hierfür sind häufig Reguläre Ausdrücke. Heyer u. a. (2006)
Nachfolgend werden zwei Ansätze vorgestellt:

Syntaktische Muster Bei dieser Analyse wird mit einem Regulärem Ausdruck ein Satz, oder Satzteil, gegeben über welchen bestimmte Informationen gesucht werden. Beispielsweise kann nach Oberbegriffen und deren Unterbegriffen gesucht werden, hierzu könnte ein Satzbau verwendet werden wie derartige Konstrukte oft geschrieben werden.
Ein Beispiel:
[...] Säugetiere wie Menschen und Kühe [...]
Ein Regulärer Ausdruck könnte wie folgt definiert sein:
`[NOMEN] wie [NOMEN] und [NOMEN]`
Wobei das erste Nomen der Oberbegriff und die folgenden Unterbegriffe sind.
Dieser Ausdruck könnte dann auf den Text angewandt werden und so einige Definitionen für Ober und Unterbegriffe extrahiert werden.

Morphemmuster Bei der Morphologischen Struktur geht es um die Struktur von Wörtern, deren Aufbau und der Regelmäßigkeit der Wörter. Durch eine Anwendung einer Morphemmusterbasierenden Analyse lassen sich charakteristische Fachausdrücke und semantische Zusämmenhänge zwischen Wortformen erkennen.
Auf dieser Basis ist ein Clustering möglich.

Bootstrapping

Unter dem Namen Bootstrapping ist ein machinelles Lernverfahren benannt welches auf Basis einer Startmenge an Informationen sowie einiger Regeln iterativ neue Informationen findet. Dieser Absatz basiert auf dem Werk von Heyer u. a. (2006).
Im Grunde ist dies eine Art Klassifikation, da der Vorgang immer wieder auf den gegebenen Regeln und gefundenen Informationen neue Informationen sucht.
Probleme dieses Verfahrend entstehen wenn ”verunreinichtes Wissen” in folgenden Iterationen zu Fehlfortplfanzungen führt, dadurch wird die Menge an fehlerhaften Informationen zunehmend größer.
Beispielsweise könnte man versuchen mit Bootstrapping Namen aus Texten zu extrahieren. Namen setzen sich zusammen aus Vorname und Nachname, dadurch werden zwei Regeln erzeugt:

1. Eine Wortform vor einem Nachnamen ist ein Vorname
2. Eine Wortform nach einem Vornamen ist ein Nachname

Wenn nun der Text nicht nur Namen in voller Schreibweise (Vorname und Nachname) verwenden, würde die Methode auch alle anderen Worte die vor Nachnamen und nach Vornamen auftreten als Namen aufnehmen.
In darauf folgenden Iterationen werden auf Basis der falschen Wörter noch weitere falsche Wörter gefunden.
Um diesen Problemen entgegen zu wirken gibt es Abarten des Vorgehens:

Pendel-Algorithmus In dieser Abwandlung erfolgt ein zusätzlicher Verifikationsschritt bevor erkannte Muster als neues Wissen aufgenommen werden. Hierzu werden während der Suche alle möglichen Musten in eine Kandidatenmenge aufgenommen. Anschließend wird gepfüft ob ein gefundener Kandidat in anderen Mustern auftritt, anhand dieser Prüfung geschieht eine Klassifizierung mit einem Schwellwert. So werden unrelevante Muster nicht als Wissen weiter verwendet.

Mutual Bootstrapping Bei dieser Erweiterung des Bootstrapping wird die zuvor als statisch angesehene Regelmenge auch erweitert. Dies geschieht indem die Regelmenge auf Basis der Informationsmenge erweitert wird.

3 Fazit

Der überwiegende Teil an Informationen liegt in Textform vor, sei es in Forschungsartikeln, Beiträgen in Social Media Kanälen oder sonstigen Medien. Text ist essentiell für unsere Tagtägliche Kommunikation. Es wird zunehmend wichtiger Erkenntnisse aus diesen Quellen extrahieren zu können.
Durch die computergestützte Analyse der Flut an Informationen des Internets können aus dem vermeindlichem Informationsmüss wertvolle Wissensrohstoffe gewonnen werden. Heyer u. a. (2006)
Text repräsentiert Wissen, daher ist Text Mining als eine effiziente Methode zur Akquisition von Wissen anzusehen.

Literaturverzeichnis

[Cleve und Lämmel 2014] Cleve, J. ; Lämmel, U.: *Data Mining*. De Gruyter, 2014 (De Gruyter Studium). – ISBN 9783486720341

[Feldman und Sanger 2006] Feldman, Ronen ; Sanger, James: *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York, NY, USA : Cambridge University Press, 2006. – ISBN 0521836573, 9780521836579

[Heyer u. a. 2006] Heyer, Gerhard ; Quasthoff, Uwe ; Wittig, Thomas: *Text Mining: Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse*. 1., Aufl. W3l, 2006. – ISBN 3937137300

[Hippner und Rentzmann 2006] Hippner, Hajo ; Rentzmann, René: Text Mining. In: *Informatik-Spektrum* 29 (2006), Aug, Nr. 4, S. 287–290. – URL https://doi.org/10.1007/s00287-006-0091-y. – ISSN 1432-122X

[Porter 1997] Porter, M. F.: Readings in Information Retrieval. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1997, Kap. An Algorithm for Suffix Stripping, S. 313–316. – URL http://dl.acm.org/citation.cfm?id=275537.275705. – ISBN 1-55860-454-5

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, 28.02.2019 Jan Dennis Bartels