Alexander Mehler und Christian Wolff

Einleitung: Perspektiven und Positionen des Text Mining

1 Einleitung

Beiträge zum Thema *Text Mining* beginnen vielfach mit dem Hinweis auf die enorme Zunahme *online* verfügbarer Dokumente, ob nun im Internet oder in Intranets (Losiewicz et al. 2000; Merkl 2000; Feldman 2001; Mehler 2001; Joachims & Leopold 2002). Der hiermit einhergehenden „Informationsflut" wird das Ungenügen des *Information Retrieval* (IR) bzw. seiner gängigen Verfahren der Informationsaufbereitung und Informationserschließung gegenübergestellt. Es wird bemängelt, dass sich das IR weitgehend darin erschöpft, Teilmengen von Textkollektionen auf Suchanfragen hin aufzufinden und in der Regel bloß listenförmig anzuordnen.

Das auf diese Weise dargestellte Spannungsverhältnis von Informationsexplosion und Defiziten bestehender IR-Verfahren bildet den Hintergrund für die Entwicklung von Verfahren zur automatischen Verarbeitung textueller Einheiten, die sich stärker an den Anforderungen von Informationssuchenden orientieren. Anders ausgedrückt: Mit der Einführung der *Neuen Medien* wächst die Bedeutung digitalisierter Dokumente als Primärmedium für die Verarbeitung, Verbreitung und Verwaltung von Information in öffentlichen und betrieblichen Organisationen. Dabei steht wegen der Menge zu verarbeitender Einheiten die Alternative einer intellektuellen Dokumenterschließung nicht zur Verfügung. Andererseits wachsen die Anforderung an eine automatische Textanalyse, der das klassische IR nicht gerecht wird.

Der Mehrzahl der hiervon betroffenen textuellen Einheiten fehlt die explizite Strukturiertheit formaler Datenstrukturen. Vielmehr weisen sie je nach Text- bzw. Dokumenttyp ganz unterschiedliche Strukturierungsgrade auf. Dabei korreliert die Flexibilität der Organisationsziele negativ mit dem *Grad an explizierter Strukturiertheit* und positiv mit der *Anzahl* jener Texte und Texttypen (E-Mails, Memos, Expertisen, technische Dokumentationen etc.), die im Zuge ihrer Realisierung produziert bzw. rezipiert werden. Vor diesem Hintergrund entsteht ein Bedarf an *Texttechnologien*, die ihren Benutzern nicht nur „intelligente" Schnittstellen zur Textrezeption anbieten, sondern zugleich auf inhaltsorientierte Text-

analysen zielen, um auf diese Weise aufgabenrelevante Daten explorieren und kontextsensitiv aufbereiten zu helfen.

> *Das Text Mining ist mit dem Versprechen verbunden, eine solche Technologie darzustellen bzw. sich als solche zu entwickeln.*

Dieser einheitlichen Problembeschreibung stehen konkurrierende Textmining-Spezifikationen gegenüber, was bereits die Vielfalt der Namensgebungen verdeutlicht. So finden sich neben der Bezeichnung *Text Mining* (Joachims & Leopold 2002; Tan 1999) die Alternativen

- *Text Data Mining* (Hearst 1999b; Merkl 2000),
- *Textual Data Mining* (Losiewicz et al. 2000),
- *Text Knowledge Engineering* (Hahn & Schnattinger 1998),
- *Knowledge Discovery in Texts* (Kodratoff 1999) oder
- *Knowledge Discovery in Textual Databases* (Feldman & Dagan 1995).

Dabei lässt bereits die Namensgebung erkennen, dass es sich um Analogiebildungen zu dem (nur unwesentlich älteren) Forschungsgebiet des *Data Mining* (DM; als Bestandteil des *Knowledge Discovery in Databases* – KDD) handelt. Diese Namensvielfalt findet ihre Entsprechung in widerstreitenden Aufgabenzuweisungen. So setzt beispielsweise Sebastiani (2002) Informationsextraktion und Text Mining weitgehend gleich, wobei er eine Schnittmenge zwischen Text Mining und Textkategorisierung ausmacht (siehe auch Dörre et al. 1999). Demgegenüber betrachten Kosala & Blockeel (2000) Informationsextraktion und Textkategorisierung lediglich als Teilbereiche des ihrer Ansicht nach umfassenderen Text Mining, während Hearst (1999a) im Gegensatz hierzu Informationsextraktion und Textkategorisierung explizit aus dem Bereich des *explorativen* Text Mining ausschließt.

2 Sichten auf das Text Mining

Trotz der zuletzt erläuterten Begriffsvielfalt sind mehrere Hauptströmungen erkennbar, die teils aufgabenorientierte, teils methodische Kriterien in den Vordergrund ihres Text Mining-Begriffs rücken. Dabei handelt es sich IR-, DM-, methoden- und wissensorientierte Ansätze.

2.1 Die Information Retrieval-Perspektive

Bereits Jacobs (1992) konzipiert ein *textbasiertes intelligentes System*, das auf eine Verbesserung von Retrieval-Ergebnissen durch automatische Zusammenfassung von Texten, ihre Kategorisierung und hypertextuelle Vernetzung zielt und greift damit den in späteren Jahren im Bereich von Suchmaschinen erfolgreichen Ansätzen zur Analyse von Hypertextstrukturen vor (vgl. Salton et al. 1994; Allan 1997).

Mit dem Ansatz von Jacobs vergleichbar thematisiert Göser (1997) – in dieser Zeitschrift und als einer der Ersten im deutschsprachigen Bereich – das Text Mining aus der Perspektive des inhaltsbasierten, benutzerorientierten Information Retrieval.

Ansätzen dieser Art ist die Auffassung gemeinsam, dass das Text Mining der Verbesserung des Information Retrieval mittels Textzusammenfassungen und Informationsextraktion diene. Obgleich mehrere Ansätze das IR als Konstituente des Text Mining-Prozesses identifizieren, besteht weitgehend Einigkeit darüber, dass IR und Text Mining verschiedene Bereiche darstellen. Diese kritische Abkehr bringt unter anderem folgende Perspektive zum Ausdruck:

2.2 Die Data-Mining-Perspektive

Fayyad et al. (1996a, b) beschreiben *Knowledge Discovery in Databases* (KDD) als einen Ansatz zur Identifikation von *„valid, novel, potentially useful, and ultimately understandable patterns“*, der neben Datenaufbereitungs-, Evaluations- und Interpretationsschritten explorative Datenanalysen in Form des *data mining* umfasst.

Eine wiederkehrende Interpretation des Text Mining besteht nun darin, dieses als *Data Mining auf textuellen Daten* zu definieren (Rajman & Besançon 1998). Text Mining bedeutet demgemäß kein verbessertes Information Retrieval, sondern die Exploration von (interpretationsbedürftigen) Daten aus Texten. In Analogie hierzu beschreibt Kodratoff (1999) *Knowledge Discovery in Texts* (KDT) als Exploration von „nützlichem“ Wissen aus Texten. Ein vergleichbarer Ansatz stammt von Losiewicz et al. (2000), die in ihrem Modell IR-, IE-, KDD- und Visualisierungskomponenten vereinigen. All diesen Ansätzen ist gemeinsam, dass sie trotz der Analogie zum KDD die Unterscheidung von KDT (Gesamtprozess) und Text Mining (Teilprozess) ebenso vermissen lassen, wie eine Definition der für das KDD zentralen Begriffe des *Wissens*, der *Nützlichkeit* und der *Verständlichkeit*.

2.3 Die methodische Perspektive

In ihrem Leitartikel zum Themenheft *Text Mining* der Zeitschrift *KI* bezeichnen Joachims & Leopold (2002) das Text Mining als *„eine Menge von Methoden zur (halb-)automatischen Auswertung großer Mengen natürlichsprachlicher Texte“*, womit sie als Folge der reklamierten Multidisziplinarität seine Methodenpluralität betonen. Das Einsatzgebiet dieser Methoden sehen sie in der partiellen, fehlertoleranten und in der Regel statistischen Textanalyse, ob zu dem Zweck der Textkategorisierung, der Informationsextraktion und Textzusammenfassung oder der Visualisierung von Textrelationen. Im Zentrum dieser Konzeption steht die Feststellung der methodischen Unselbstständigkeit des Text Mining: Als ein Sammelbegriff subsumiert es vielfältige Textanalysemethoden, auf deren Weiterentwicklung und Integration fokussiert wird.

2.4 Die wissensorientierte Perspektive

Im Gegensatz hierzu zielt Hearst (1999a) auf die wissensorientierte Eingrenzung des Text Mining, und zwar unter expliziter Abgrenzung von Ansätzen der korpusanalytischen Computerlinguistik und des inhaltsbasierten Information Retrieval. Hearst betont die vielfach kritisierte (Wiegand 1999) Metapher des „Goldschürfens“. Sie definiert Text Mining als textbasierte Datenanalyse zur Exploration von *„heretofore unknown“*, *„never–before encountered information“* in Bezug auf jene „realweltlichen“ (nicht aber sprachlichen) Zusammenhänge, welche die Texte annahmegemäß thematisieren. Unter Absehung von ihrem Vorverarbeitungsstatus bilden Information Retrieval (IR), Informationsextraktion (IE), und Textkategorisierung (TK) folglich keine Kernbestandteile des Text Mining, da sie keine Information *explorieren*, sondern bloß Textmengen mittels Indexmengen erschließen (IR), vorgegebene Schemata mit ihren textuellen Instanzen abgleichen (IE) bzw. Texte auf vordefinierte Kategorien abbilden (TK).

Dabei ist allerdings zu verdeutlichen, dass IR, IE und TK jeweils im Kern funktional definiert sind und mit diesen Konzepten kein Hinweis auf eine konkrete Umsetzungsmethode gegeben ist: Ein Text Mining-Verfahren kann in diesem Sinn durchaus geeignet sein, für ein IR-System geeignete Beschreibungsterme zu ermitteln oder inhaltliche relevante Querbezüge zwischen verschiedenen Termen zu beschreiben.

Anstatt das Text Mining begrifflich weiter einzugrenzen, nennt Hearst Musterbeispiele, die als Prüfsteine für die „Mining“-Tauglichkeit von Textanalyse-Systemen dienen sollen. So verweist sie auf Zitationsanalysen, die zeigen, dass Patente weitgehend auf öffentlich finanzierter Forschung beruhen. Ein weiteres

Beispiel bildet die Analyse von Patientenakten, die kausale Zusammenhänge zwischen der Nichteinnahme von Spurenelementen und Syndromen belegen. Im Zentrum dieser Fallbeispiele steht die Überlegung, dass die jeweils explorierte Information in keinem der analysierten Texte isoliert thematisiert wird, sondern erst durch die Analyse mehrerer Texte zu gewinnen ist.

3 Zwei Grundpositionen

Die Verschiedenheit dieser vier Konzeptionen lässt erahnen, dass sich das Text Mining erst zu formieren beginnt, ohne auf einen bereits anerkannten Text Mining-Begriff zurückgreifen zu können. Dies betrifft in gleicher Weise das zugehörige Methoden- und Aufgabenspektrum. Dennoch lassen sich zwei Grundpositionen ausmachen, welche das Spektrum bestehender Text Mining-Ansätze aufspannen:

3.1 Methodenorientierte Ansätze

Das untere Ende des Spektrums bestehender Mining-Begriffe bilden *methodenorientierte Ansätze*. Sie untersuchen, welche Methoden welche Textanalyse-Aufgaben mit welchem Erfolg zu lösen erlauben, und zwar in Ergänzung, Erweiterung oder Ersetzung von herkömmlichen Methoden des Information Retrieval, der Informationsextraktion oder der Textzusammenfassung.

Im Zentrum steht die Konzeption von Methoden entlang der Prämisse, dass wegen des Fehlens bzw. der unzureichenden Skalierbarkeit von Verfahren zur automatischen Generierung propositionaler Textrepräsentationsmodelle statistische, textoberflächenstrukturelle Analysen unumgänglich sind. Dies betrifft insbesondere Situationen, in denen textuelle Massendaten zu analysieren sind, wie sie im Rahmen der Presse-, Wissenschafts- und betrieblichen Kommunikation anfallen.

Diese Massendaten sind mittlerweile vielfach webbasiert zugänglich und liegen in einer überschaubaren Zahl gängiger, mehr oder weniger strukturierter Formate vor (Office-Formate, das Portable Document Format (PDF), die HyperTextMarkup Language (HTML), zunehmend auch als XML-Dateien (extensible Markup Language)). Vor diesem Hintergrund erweist sich das Web Mining als eine Weiterentwicklung des Text Mining, was weiter unten erläutert wird.

Pragmatisch gesprochen werden massendatentaugliche Ansätze bevorzugt, die (zwar nur) partielle Analysen (dafür aber) zuverlässig und fehlertolerant produzieren, und zwar gegenüber solchen Ansätzen, die zwar (tiefen–)semantische

Analysen erlauben, aufgrund ihrer Arbeitsweise aber weder massendatentauglich noch ausnahmetolerant sind. Folgerichtig werden für die konzeptionierten Methoden nur im statistischen Sinne, nicht aber im diskurssemantischen Sinne explorative Qualitäten gefordert. Anstatt also zu beanspruchen, „verborgene realweltliche Zusammenhänge" anhand von automatischen Textanalysen zu rekonstruieren, werden Texte in einer Weise analysiert, die es Rezipienten der Analyseergebnisse ermöglichen soll, relevante Zusammenhänge effizienter zu *entdecken* oder auch nur zu *identifizieren*.

Diese Perspektive macht deutlich, dass Text Mining-Verfahren in vielen Fällen keine eigenständige Anwendung konstituieren bzw. eine vorgegebene Aufgabenstellung vollständig zu lösen in der Lage sind, sondern dass vielmehr erst die Kopplung z. B. mit intellektuellen Überarbeitungsverfahren ein wunschgemäßes Ergebnis der Textexploration liefert. Dies wird am Beispiel des *ontology engineering* deutlich, das auf die Exploration von (normativen) Wissensstrukturen aus großen Textmengen zielt. Obwohl derzeit kein Text Mining-Verfahren in der Lage sein dürfte, sozusagen „auf Knopfdruck" eine Ontologie zu generieren, können Ergebnisse des Text Mining intellektuell weiterverarbeitet und z. B. mit Hilfe geeigneter Ontologie-Editoren optimiert werden (vgl. dazu Böhm et al. 2002).

Die Last der Exploration nützlicher, unerwarteter Information liegt unter dieser Perspektive auf Seiten der Rezipienten, und für diese Sichtweise erscheint die Metapher des Schürfens durchaus angemessen, da ein gefundener Rohdiamant ohne Weiterverarbeitung (mit anderen Methoden) nur wenig Nutzen aufweist.

3.2 Wissensorientierte Ansätze

Hearsts Vision eines realweltliche Zusammenhänge anhand von Textanalysen selbstständig explorierenden Systems bildet das obere Ende des Text Mining-Spektrums. Die Explorationslast liegt nun umgekehrt auf Seiten des „künstlichen" Text Mining-Systems.

Es ist evident, dass dieser Ansatz an ein propositionales Textrepräsentationsmodell gebunden ist, das Explorationsresultate über Ähnlichkeitsvergleiche textueller Einheiten auf der Basis des strukturindifferenten *Bag-of-words*-Modells des IR hinaus erwartbar macht. Ein Paradebeispiel bilden Anstrengungen zum automatischen Aufbau von so genannten Ontologien und ihre Nutzbarmachung im Zusammenhang des *Semantic Web* (Fensel et al. 2003; Handschuh & Staab 2003). Dem hiermit einhergehenden höheren Automatisierungsanspruch steht allerdings der Mangel an bereits etablierten Systemen und Verfahren gegenüber.

Abgesehen von der Problematik des Begriffs der automatischen Informations- bzw. Wissensexploration (Wiegand 1999) stellt sich jedoch die Frage, ob hier nicht auch dann ein uneinlösbarer Anspruch vorliegt, wenn nicht von Text Mining, sondern korrekter von *explorativer Textdatenanalyse* – von einer Anwendung von Verfahren der explorativen Datenanalyse auf textuelle Daten also – gesprochen wird (Mehler 2004b, a).

Dem Verzicht auf explorative Textanalysen à la Hearst steht eine Vielzahl erprobter und etablierter Methoden gegenüber – vgl. hierzu Hotho et al. (2005) (in diesem Band). Umgekehrt existieren kaum massendatentaugliche Anwendungen, die den Hearstschen Ansprüchen genügen. Offenbar besteht also ein – schon aus der KI-Forschung her bekannter – *trade-off* zwischen Massendatentauglichkeit, Fehlertoleranz und Robustheit auf der einen und analytischem, semantischem Auflösungsvermögen auf der anderen Seite. Der Aspekt der Massendatenanalyse verweist dabei ebenso wie das Schlagwort des *Semantic Web* auf einen Anwendungsbereich des Text Mining, der unter der eigenständigen Bezeichnung *Web Mining* firmiert.

4 Web Mining

Vor dem Hintergrund der unzähligen Menge verfügbarer Webseiten, ihrer Strukturen und Änderungsraten sowie der zahllosen Nutzer und ihrer heterogenen Informationsbedürfnisse problematisieren Kobayashi & Takeda (2000) die beschränkten Möglichkeiten des klassischen Information Retrieval im Web. Hiermit ist ein Aufgabendruck angesprochen, der oben für das Text Mining als richtungsweisend ausgemacht wurde. Dies erlaubt es, mit dem *Web Mining* einen Ausblick auf einen wichtigen Anwendungsbereich des Text Mining zu geben, wobei mit Kosala & Blockeel (2000) drei Teilbereiche zu unterscheiden sind:

4.1 Web Content Mining

Das *Web Content Mining* zielt auf ein verbessertes Browsing mit Hilfe von Verfahren des inhaltsorientierten Information Retrieval (Landauer & Dumais 1997), der Textkategorisierung und -klassifikation wie auch mit Hilfe von annotationsbasierten Abfragesprachen im Rahmen strukturierter Retrieval-Modelle. Ein Paradebeispiel bildet die Suchmaschine *Vivísimo* (Stein & zu Eissen 2004), die Clustering-Verfahren zur Strukturierung von Retrieval-Ergebnissen einsetzt. Anders als die Textkategorisierung und -klassifikation rekurrieren ihre hyper-

textuellen Entsprechungen jedoch auf eine erweiterte Merkmalsselektion, indem sie HTML-Tags (und insbesondere Metatags), DOM-Strukturen[1] und benachbarte Webpages inkorporieren.

4.2 Web Structure Mining

Das *Web Structure Mining* zielt auf die Typisierung von Webdokumenten unter anderem auf der Basis ihrer Linkstrukturen. Ein Paradebeispiel bildet die Ermittlung von Webpages als Kandidaten für *hubs* und *authorities* (Kleinberg (1999), vgl. auch Brin & Page (1998); Page et al. (1998); Lifantsev (1999)). In diesem Zusammenhang ist die Kategorisierung von *web hierarchies*, *directories*, *corporate sites* und *web sites* (Amitay et al. 2003) von Ansätzen zu unterscheiden, die auf die Segmentierung *einzelner* Webpages zielen (Mizuuchi & Tajima 1999). Diesen mikrostrukturellen Analysen stehen makrostrukturelle Betrachtungen der Topologie des Webs gegenüber. So untersucht beispielsweise Adamic (1999) Kürzeste-Wege- und Clusterungseigenschaften von Webpages unter dem Begriff des *Small Worlds*-Phänomens wie es für soziale Netzwerke kennzeichnend ist (Milgram 1967).

4.3 Web Usage Mining

Das *Web Usage Mining* bezieht sich schließlich auf die Analyse des Rezeptionsverhaltens von Web-Nutzern. Hierzu werden unter anderem Zipfsche Modelle herangezogen (Zipf 1949; Cooley et al. 1999). Im Kern sagen diese Modelle aus, dass quantitative Indikatoren der Rezeption webbasierter Dokumente dem semiotischen Präferenzgesetz der Ordnung nach der Wichtigkeit (Tuldava 1998) folgen. In diesem Sinne existiert beispielsweise eine sehr geringe Zahl von Webpages, die häufig angesteuert und lange rezipiert werden. Ihr steht eine große Zahl von Seiten gegenüber, die selten angesteuert und in der Regel nur sehr kurz rezipiert werden, wobei zwischen beiden Bereichen ein fließender Übergang beobachtbar ist, der insgesamt eine extrem schiefe Verteilung erkennen lässt.

Soweit das Web Usage Mining lediglich auf Nutzungsinformation bezüglich besuchter Webseiten (Zuordnungen von Nutzern und Adressen) zurückgreift, überschreitet es die Schwelle zur Textexploration i. e. S. allerdings noch nicht.

1 *Document Object Model*.

4.4 Fazit

Mit dem Web Mining steht dem Text Mining ein breites Bewährungsfeld offen, wobei Menge und Struktur der verfügbaren Webdokumente die Entwicklung stärker strukturorientierter Ansätze erwarten lässt. Dabei dürfte der Konflikt zwischen Massendatentauglichkeit auf der einen und semantischem Auflösungsvermögen auf der anderen Seite, der oben an der Unterscheidung von methoden- und wissensorientierten Verfahren festgemacht wurde, nur durch eine stärkere *computerlinguistische* und zugleich *textlinguistische* Fundierung zu lösen sein.

Der Grund für diese Einschätzung ist darin zu sehen, dass die Ablösung oder doch wenigstens Ergänzung des strukturindifferenten *Bag-of-words*-Modells sich an *Textstruktur*-Modellen orientieren sollte, deren Instanzen nachgewiesenermaßen effizient explorierbar sind. Das Resultat einer solchen Fundierung könnte ferner zeigen, welche äußerst engen Grenzen wissensorientierten Mining-Ansätzen gesetzt sind. Die Kritik der Metapher des *Goldschürfens* bzw. der textbasierten Wissensexploration nimmt diese Grenzziehung im Grunde genommen bereits vorweg (Wiegand 1999; Weber 1999).

Massendatengetriebene Ansätze (im Sinne eines *Text Data Mining*) und wissensorientierte Verfahren schließen keineswegs einander aus: Zum einen zeigen Entwicklungen innerhalb der Computerlinguistik der vergangenen Jahre, dass datenorientierte Verfahren ein unverzichtbares Werkzeug zur Rekonstruktion linguistischen Wissens bilden. Als Beispiele hierfür sind unter anderem das *data oriented parsing* (vgl. Bod et al. 2003), das POS-Tagging (vgl. Brants 2000) oder die latente semantische Analyse (Landauer & Dumais 1997; Schütze 1997) und semantische Räume (Rieger 1989) zu nennen. Auf der anderen Seite erlaubt die Rückkoppelung datenanalytischer Verfahren an explizite (linguistische) Wissensstrukturen die Verbesserung von Text Mining-Resultaten (vgl. Heyer et al. 2001). Hier liegt möglicherweise ein erhebliches Potential für die Optimierung der meist auf rein statistischen Methoden beruhenden Text Mining-Verfahren. Zu überlegen ist insbesondere, wie die Felder *Text Mining* und *Corpuslinguistik* angesichts ihrer sich überlappenden Gegenstandsbereiche noch fruchtbarer interagieren können (Heyer et al. 2005). Letztere befaßt sich bereits sehr viel stärker (und länger) mit Fragen der expliziten Strukturierung großer Textmengen, ihrer (linguistischen) Annotation und ihrer repräsentativen und standardisierten Zusammensetzung, Aspekte, die auch für Optimierung und Bewertung des Text Mining relevant sind. Im Licht des voranstehend zum Web Mining Gesagten ist dieses Potenzial dort besonders offensichtlich, wo das Web als Datengrundlage für die Corpuserstellung herangezogen wird (Kilgarriff & Grefenstette 2003).

Im Zusammenhang dieser Kombinationsmöglichkeiten wird sich das Text Mining auch dahingehend zu bewähren haben, inwieweit es über das „intelligente“ Information-Retrieval (Baeza-Yates & Ribeiro-Neto 1999) bzw. Formen der adaptiven Informationsextraktion (Wilks & Catizone 1999) hinausgeht, um mehr als ein Sammelbegriff für Methoden der explorativen Datenanalyse (Joachims & Leopold 2002) zu gelten, die auf textuelle Daten angewandt werden.

5 Überblick über das Themenheft

Das vorliegende Themenheft deckt das Spektrum methoden- und wissensorientierter Mining-Ansätze ab.

Andreas Hotho, Andreas Nürnberger und Gerhard Paass geben in ihrem Beitrag einen umfassenden Überblick über das Text Mining aus *methodischer Sicht*. Ausgehend von einer disziplinären Einordnung des Text Mining im Kontext verwandter Ansätze (wie Data Mining oder maschinelles Lernen) und Anwendungsbereiche (wie Information Retrieval, Informationsextraktion und Natural Language Processing) erläutern sie grundlegende Methoden der Vorverarbeitung und Repräsentation textueller Einheiten sowie ihrer automatischen Kategorisierung, Klassifikation und Informationsextraktion. Ein besonderes Augenmerk gilt dabei Methoden der Visualisierung von Analyseresultaten, womit der für das Mining kennzeichnende Aspekt der verständlichen Ergebnisaufbereitung angesprochen wird. Schließlich erläutern die Autoren die derzeit wichtigsten Anwendungsbereiche des Text Minings.

Ausgehend von dem Modell des *semantischen Raums* von Burghard Rieger (Rieger 1989) beschreibt Edda Leopold in ihrem Beitrag Verfahren zur Exploration von Ähnlichkeitsrelationen sprachlicher Einheiten. Dies betrifft die latente semantische Analyse ebenso wie ihre probabilistischen Erweiterungen. Als besonders vielversprechend erweisen sich dabei Versuche einer Verbindung von Kategorisierungs- und Klassifikationsverfahren mit Hilfe von *Support Vector Machines*, die Leopold zur Lösung des Dimensionenreduktionsproblems im Rahmen von semantischen Räumen einsetzt, ohne auf die Auswertung hochdimensionaler Merkmalsvektoren verzichten zu müssen.

Eine Synthese der methoden- bzw. wissensorientierten Perspektive schlagen Bloehdorn et al. mit dem Entwurf eines *Ontology-based Framework for Text Mining* vor. Sie gehen davon aus, dass sich Vor- und Nachteile der verschiedenen Perspektiven (massendatentauglich, ressourcensparsam, fehlerträchtig *versus* teuer, qualitativ und infolgedessen im Skopus beschränkt) nicht nur in Einklang bringen lassen, sondern sich sogar wechselseitig befruchten können. Ausgehend

von einer formalen Definition grundlegender ontologischer Konzepte stellen sie eine Systemarchitektur vor, in der vorhandenes ontologisches Wissen für ontologiebasierte Text Mining-Komponenten (Modul *TextToOnto*) fruchtbar gemacht werden können. Die Ontologie ist dabei selbst Erkenntnisziel (Anreicherung der Wissensstruktur, Lernen von Relationen) und Erkenntniswerkzeug, als die ontologischen Strukturen für Anwendungen wie Clusterung und Klassifikation zum Einsatz gebracht werden.

Matthias Dehmer schließlich thematisiert den Aufgabenbereich des *Web Structure Mining*. Ausgehend von einer kritischen Erörterung der Aussagekraft von Indizes von Hypertextgraphen leitet Dehmer zur Klassifikation solcher Graphen über. Die Grundlage hierfür bildet die Einsicht, dass Strukturvergleiche von Webdokumenten nicht länger an den summarischen Indizes ansetzen können, wie sie in der Frühphase der Hypertextmodellierung entwickelt wurden (Botafogo et al. 1992). Demgegenüber zielt Dehmer auf die Entwicklung von Maßen, welche die Ähnlichkeit von Hypertextgraphen automatisch bewerten können sollen.

6 Weiterführende Informationen

Text Mining ist eine noch junge wissenschaftliche, anwendungsorientierte Disziplin. Tabelle (1) gibt ein quantitatives Indiz und mag bei der Einordnung behilflich sein. Die Trefferhäufigkeiten für *Data Mining*, *Text Mining* und *Web Mining* in Google, Google Scholar und Inspec sprechen ein deutliches Bild.

	Google	Google Scholar	INSPEC
Data Mining	6.850.000	122.000	13.784
Text Mining	**301.000**	**4.180**	**409**
Web Mining	136.000	2.790	557

Tabelle 1: Trefferhäufigkeiten für Data Mining, Text Mining und Web Mining (Stand: Mai 2005).

6.1 Literatur zum Text Mining

Es kann aufgrund des voranstehend Gesagten kaum verwundern, dass bisher nur wenige Lehrbücher zum Text Mining vorliegen. Die nachfolgende Liste soll einen knappen Überblick zu den derzeit verfügbaren Werken geben:

- Als ein erstes Beispiel kann das weit verbreitete Data Mining-Lehrbuch von Witten & Frank (2000) gelten, das Text Mining zwar nur am Rande behandelt (Witten & Frank 2000, 331ff.), dafür aber eine Vielzahl analytischer Verfahren vorstellt, die auch für das Text Mining relevant sind.

- Aus computerlinguistischer Sicht empfehlenswert ist Manning & Schütze (2003). Die Autoren vermeiden zwar, das Konzept des *Mining* explizit einzuführen, aber ihr Anspruch „Statistical NLP as we define it comprises all quantitative approaches to automated language processing [...]“ (Manning & Schütze 2003, xxxi) und die damit verbundene ausführliche Behandlung auch der automatischen Verarbeitung von textuellen Massendaten macht dieses Lehrbuch zu einer nützlichen Einführung in Mining-relevante Verfahren. Aus der Sicht quantitativer Methoden innerhalb der Textlinguistik ist die Einführung von Altmann (1988) empfehlenswert, welche grundlegende Verteilungsmodelle zur Beschreibung quantitativer Merkmale textueller Einheiten erläutert, auch wenn dieses Buch sonst in keinem direkten Verhältnis zum Text Mining steht.

- Intensiv mit der systemischen Einordnung des Text Mining im Spannungsfeld von numerischer Datenanalyse, Information Retrieval und generischen Verfahren der Strukturidentifikation setzen sich Weiss et al. (2004) auseinander, wobei die Autoren zunächst von der grundsätzlichen Analogie des Text Mining zum Data Mining ausgehen („Text and documents can be transformed into measured values, such as the presence or absence of words, and the same methods that have proven successful for predictive data mining can be applied to text.“, (Weiss et al. 2004, v). Diese Einführung zeichnet sich weiterhin durch eine Sammlung praktischer Anwendungsstudien aus.

- Die Charakteristika des Web Mining als wichtigstem Anwendungsgebiet des Text Mining thematisiert (Chakrabarti 2002). Vertieft behandelt werden dort neben Fragen der Akquisition von Web-Dokumenten insbesondere Verfahren des maschinellen Lernens basierend auf hypertextuellen Datenbeständen. Die Darstellung der Verfahren wird durch die Beschreibung ausgewählter Anwendungen (*social network analysis, resource discovery*) ergänzt.

- Eine erste deutschsporachige Monographie zum Text Mining legen Heyer et al. (2005) vor, die vor dem Hintergrund zahlreicher anwendungsnaher Studien ein Gesamtbild des Text Mining-Prozesses skizzieren, das

neben statistischen Analyseverfahren für große Textcorpora auch linguistische Aspekte und traditionelle sprachliche Kategorien wie voranstehend angemahnt ins Blickfeld rücken.

Einige aus Workshops und Konferenzen hervorgegangene Sammelbände der letzten Jahre bieten eine gute Übersicht über aktive Forschungsfelder mit Bezug zum Text Mining; zu nennen sind hier Berry (2003), Franke et al. (2003) und Sirmakessis (2004). In ihnen steht weniger die systematische Erschließung des Gegenstandsbereichs Text Mining, sondern die Darstellung typischer Verfahren und Anwendungen im Mittelpunkt, von denen nachfolgend Beispiele genannt seien:

- Trenderkennung und Themenidentifikation durch Text Mining,
- Auffinden von Synonymen in Textcorpora,
- adaptives und kollaboratives Information Retrieval sowie
- Clustering und Merkmalsextraktion aus Texten.

6.2 Tagungen

So vielfältig wie die Anwendungsmöglichkeiten des Text Mining sind auch die Tagungen und Workshops, in denen sich einschlägige Beiträge finden:

- Konferenzen mit primär *computerlinguistischem* oder *sprachtechnologischem* Bezug – *International Conference on Computational Linguistics* (COLING), *Meeting of the (EuroAssociation for Computational Linguistics* (ACL, EACL), *International Conference on Linguistic Resources and Evaluation* (LREC), in Deutschland die *GLDV-Frühjahrstagung* (GLDV).
- Text Mining-Ansätze im Umfeld des *Data Mining* und des *maschinellen Lernens – International Conference on Machine Learning* (ICML), *European Conference on Machine Learning* (ECML), *International Conference on Knowledge Discovery and Data Mining* (KDD), *Principles and Practice of Knowledge Discovery in Databases* (PKDD), *International Conference on Data Mining, Text Mining and their Business Applications*.
- Da Text Mining-Verfahren mittlerweile auch in der *KI-Forschung* als wichtige Methode akzeptiert werden, finden sich in einschlägigen KI-Tagungen vermehrt Beiträge mit Text Mining-Bezug – *International Joint Conference on Artificial Intelligence* (IJCAI), *National Conference on Artificial Intelligence* (AAAI).

- Weitere relevante Konferenzen finden sich in den Bereichen *Information Retrieval* (*Conference on Research and Development in Information Retrieval* (SIGIR)), *Wissensmanagement* (*International Conference on Information and Knowledge Management* (CIKM), *International Conference on Knowledge Management* (I-Know)) sowie auf dem Gebiet *webbasierter Informationssysteme* (*International World Wide Web Conference* (WWW)) und der *automatischen Klassifikation* (*Annual Conference of the German Classification Society*).

Diese Breite an Konferenzen mit Text Mining-relevanten Inhalten zeigt, dass sich das Text Mining transdisziplinär etabliert hat, wobei Forscher aus den Bereichen Computerlinguistik, Informatik und verwandten Disziplinen zunehmend interdisziplinär kooperieren. Sie findet sich denn auch in dem vorliegenden Themenheft wieder, dessen Autoren aus den Bereichen Computerlinguistik und quantitative Linguistik sowie Informatik und Mathematik stammen.

Literatur

Adamic, L. A. (1999). The small world web. In S. Abiteboul & A.-M. Vercoustre (Eds.), *Proc. 3rd European Conf. Research and Advanced Technology for Digital Libraries, ECDL*, number 1696 in Lecture Notes in Computer Science (pp. 443–452). Berlin/Heidelberg/New York: Springer.

Allan, J. (1997). Building hypertext using information retrieval. *Information Processing and Management*, *33*(2), 145–159.

Altmann, G. (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Amitay, E., Carmel, D., Darlow, A., Lempel, R., & Soffer, A. (2003). The connectivity sonar: detecting site functionality by structural patterns. In *Proc. of the 14th ACM conference on Hypertext and Hypermedia*, (pp. 38–47).

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Reading, Massachusetts: Addison-Wesley.

Berry, M. W. (2003). *Survey of text mining*. New York: Springer.

Böhm, K., Heyer, G., Quasthoff, U., & Wolff, C. (2002). Topic map generation using text mining. *J.UCS - Journal of Universal Computer Science*, *8*(6), 623–633.

Bod, R., Scha, R., & Sima'an, K. (2003). *Data-Oriented Parsing*. Stanford: CSLI Publications.

Botafogo, R. A., Rivlin, E., & Shneiderman, B. (1992). Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Transactions on Information Systems*, *10*(2), 142–180.

Brants, T. (2000). TnT – a statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLP-2000)*, Seattle, WA.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, *30*, 107–117.

Chakrabarti, S. (2002). *Mining the Web: Discovering Knowledge from Hypertext Data*. San Francisco: Morgan Kaufmann.

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *Journal of Knowledge and Information Systems*, *1*(1), 5–32.

Dörre, J., Gerstl, P., & Seiffert, R. (1999). Text mining: Finding nuggets in mountains of textual data. In Chaudhuri, S. & Madigan, D. (Eds.), *The Fifth ACM SIKKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 398–401)., New York. ACM.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27–34.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996b). From data mining to knowledge discovery: An overview. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 1–34). Menlo Park, California: AIII Press/MIT Press.

Feldman, R. (2001). Mining unstructured data. In *Tutorial Notes for ACM SIGKDD 1999 International Conference on Knowledge Discovery and Data Mining*, (pp. 182–236). ACM.

Feldman, R. & Dagan, I. (1995). Knowledge discovery in textual databases (kdt). In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD'95)*, (pp. 112–117).

Fensel, D., Hendler, J., Lieberman, H., & Wahlster, W. (2003). *Spinning the Semantic Web. Bringing the World Wide Web to Its Full Potential*. Cambridge, Massachusetts: MIT Press.

Franke, J., Nakhaeizadeh, G., & Renz, I. (2003). *Text Mining, Theoretical Aspects and Applications*. Physica-Verlag.

Göser, S. (1997). Inhaltsbasiertes Information Retrieval: Die TextMining-Technologie. *LDV Forum*, *14*(1), 48–52.

Hahn, U. & Schnattinger, K. (1998). Towards text knowledge engineering. In *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98) and of the 10th Conference on Innovative Applications of Artificial Intelligence (IAAI-98)*, (pp. 524–531)., Menlo Park. AAAI Press.

Handschuh, S. & Staab, S. (2003). *Annotation for the Semantic Web*, volume 96 of *Frontiers in Artificial Intelligence and Applications*. IOS.

Hearst, M. A. (1999a). Untangling text data mining. In *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, 1999*.

Hearst, M. A. (1999b). User interfaces and visualization. In R. A. Baeza-Yates & B. Ribeiro-Neto (Eds.), *Modern Information Retrieval* chapter 10, (pp. 257–323). Addison Wesley.

Heyer, G., Läuter, M., Quasthoff, U., & Wolff, C. (2001). Wissensextraktion durch linguistisches Postprocessing bei der Corpusanalyse. In Lobin, H. (Ed.), *Sprach- und Texttechnologie in digitalen Medien. Proc. GLDV-Jahrestagung 2001*, (pp. 71–83).

Heyer, G., Quasthoff, U., & Wittig, T. (2005). *Wissensrohstoff Text. Text Mining: Konzepte, Algorithmen, Ergebnisse*. Bochum: W3L.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. *LDV-Forum*, *20*(1), 19–63.

Jacobs, P. S. (1992). Introduction: Text power and intelligent systems. In P. S. Jacobs (Ed.), *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval* (pp. 1–8). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Joachims, T. & Leopold, E. (2002). Themenheft: Text-Mining. Vorwort der Herausgeber. *Künstliche Intelligenz*, *2*, 4.

Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, *29*(3), 333–347.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, *46*(5), 604–632.

Kobayashi, M. & Takeda, K. (2000). Information retrieval on the web. *ACM Computing Surveys*, *32*(2), 144–173.

Kodratoff, Y. (1999). Knowledge discovery in texts: A definition and applications. In Rás, Z. W. & Skowron, A. (Eds.), *Proceedings of the 11th International Symposium on Foundations of Intelligent Systems (ISMIS '99)*, (pp. 16–29)., Berlin/Heidelberg/New York. Springer.

Kosala, R. & Blockeel, H. (2000). Web mining research: A survey. *SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining*, *2*(1), 1–15.

Landauer, T. K. & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Lifantsev, M. (1999). Rank computation methods for web documents. Technical Report TR-76, ECSL, Department of Computer Science, SUNY at Stony Brook, Stony Brook/NY.

Losiewicz, P., Oard, D. W., & Kosthoff, R. N. (2000). Textual data mining to support science and technology management. *Journal of Intelligent Information Systems*, *15*, 99–119.

Manning, C. D. & Schütze, H. (2003). *Foundations of Statistical Natural Language Processing* (6. Aufl. ed.). Cambridge, Massachusetts: MIT Press.

Mehler, A. (2001). Aspects of text mining. From computational semiotics to systemic functional hypertexts. *Australian Journal of Information Systems*, *8*(2), 129–141.

Mehler, A. (2004a). Automatische Synthese Internet-basierter Links für digitale Bibliotheken. *Osnabrücker Beiträge zur Sprachtheorie*, *68*, 31–53.

Mehler, A. (2004b). Textmining. In H. Lobin & L. Lemnitzer (Eds.), *Texttechnologie. Perspektiven und Anwendungen* (pp. 329–352). Tübingen: Stauffenburg.

Merkl, D. (2000). Text data mining. In R. Dale, H. Moisl, & H. Somers (Eds.), *Handbook of Natural Language Processing* (pp. 889–903). New York: Dekker.

Milgram, S. (1967). The small world problem. *Psychology Today*, *61*, 60 – 67.

Mizuuchi, Y. & Tajima, K. (1999). Finding context paths for web pages. In *Proceedings of the 10th ACM Conference on Hypertext and Hypermedia*, (pp. 13–22).

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, Stanford Digital Library Technologies Project, Stanford/CA.

Rajman, M. & Besançon, R. (1998). Text mining — knowledge extraction from unstructured textual data. In Rizzi, A., Vichi, M., & Bock, H.-H. (Eds.), *Advances in Data Science and Classification: Proc. of 6th Conference of International Federation of Classification Societies (IFCS-98)*, (pp. 473–480)., Berlin/Heidelberg/New York. Springer.

Rieger, B. (1989). *Unscharfe Semantik: die empirische Analyse, quantitative Beschreibung, formale Repräsentation und prozedurale Modellierung vager Wortbedeutungen in Texten*. Frankfurt a.M.: Peter Lang.

Salton, G., Allan, J., & Buckley, C. (1994). Automatic structuring and retrieval of large text files. *Communications of the ACM*, *37*(2), 97–108.

Schütze, H. (1997). *Ambiguity Resolution in Language Learning: Computational and Cognitive Models*, volume 71 of *CSLI Lecture Notes*. Stanford: CSLI Publications.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47.

Sirmakessis, S. (2004). *Text Mining and its Applications*. Number 138 in Studies in Fuzziness and Soft Computing. Berlin, DE: Springer-Verlag.

Stein, B. & zu Eissen, S. M. (2004). Automatische Kategorisierung für Web-basierte Suche - Einführung, Techniken und Projekte. *KI - Künstliche Intelligenz*, *18*(4), 11–17.

Tan, A.-H. (1999). Text mining: The state of the art and the challenges. In *Proc. of the Pacific Asia Conference on Knowledge Discovery and Data Mining PAKDD'99*, (pp. 65–70).

Tuldava, J. (1998). *Probleme und Methoden der quantitativ-systemischen Lexikologie*. Trier: WVT.

Weber, N. (1999). *Die Semantik von Bedeutungsexplikationen*, volume 3 of *Sprache, Sprechen und Computer/Computer Studies in Language and Speech*. Frankfurt am Main: Lang.

Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. J. (2004). *Text Mining. Predictive Methods for Analyzing Unstructured Information*. New York: Springer.

Wiegand, H. E. (1999). Wissen, Wissensrepräsentation und Printwörterbücher. In Heid, U., Evert, Lehmann, E., & Rohrer, C. (Eds.), *Proceedings of the 9th Euralex International Congress, August 8.-12. 2000, Stuttgart*, (pp. 15–38)., Stuttgart. Institut für maschinelle Sprachverarbeitung.

Wilks, Y. & Catizone, R. (1999). Can we make information extraction more adaptive. In Pazienza, M. T. (Ed.), *Information Extraction. Towards Scalable, Adaptable Systems*, (pp. 1–16)., Berlin/Heidelberg/New York. Springer.

Witten, I. H. & Frank, E. (2000). *Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco: Morgan Kaufmann.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology*. Cambridge/MA: Addison-Wesley.