Text Mining an der Börse: Einfluss von Ad-hoc-Mitteilungen auf die Kursentwicklung

Myra Spiliopoulou
Universität Magdeburg
Institut für Technische und
Betriebliche Informationssysteme
39106 Magdeburg, Universitätsplatz 2
myra@iti.cs.uni-magdeburg.de

Anja Schulz
Humboldt-Universität zu Berlin
Institut für Bank-, Börsen- und
Versicherungswesen
10178 Berlin, Spandauer Straße 1
aschulz@wiwi.hu-berlin.de

Karsten Winkler
Handelshochschule Leipzig
Lehrstuhl für Wirtschafts-
informatik des E-Business
04109 Leipzig, Jahnallee 59
kwinkler@ebusiness.hhl.de

Zusammenfassung

Ein Grundpfeiler der modernen Kapitalmarkttheorie ist die Hypothese informationseffizienter Kapitalmärkte. Danach passen sich Kurse unverzüglich nach Bekanntgabe neuer, für die Bewertung dieser Aktien relevanter Informationen an. Börsennotierte deutsche Aktiengesellschaften sind nach § 15 Wertpapierhandelsgesetz (WpHG) verpflichtet, potentiell kursbeeinflussende Tatsachen zu veröffentlichen. Im Gegensatz zu strukturierten, tendenziell unkompliziert auswertbaren Daten sind Ad-hoc-Mitteilungen textuelle, d.h. unstrukturierte Dokumente. Wissensentdeckung in textuellen Datenbanken bzw. Text Mining ermöglicht eine automatisierte Analyse dieser betriebswirtschaftlich relevanten Textdaten. Für den Zeitraum von 01.01.1999 bis 31.12.2002 wurde im Rahmen einer empirischen Analyse untersucht, ob die von DAX100-Unternehmen veröffentlichten Ad-hoc-Mitteilungen tatsächlich bewertungsrelevante Informationen darstellen. Im Rahmen der Analyse wurde der SAS Enterprise Miner und ergänzend die DIAsDEM Workbench eingesetzt, um ein Vorhersagemodell für die Kursrelevanz von Ad-hoc-Mitteilungen zu generieren.

Keywords: Ad-hoc-Meldung, Kursrelevanz, Ereignisstudie, Text Mining, Textklassifikation, Börseninformationssystem, SAS Enterprise Miner.

1 Kursrelevanz von Ad-hoc-Mitteilungen

Der im Rahmen des zweiten Finanzmarktförderungsgesetzes neu geregelte § 15 Wertpapierhandelsgesetz (WpHG) legt die Ad-hoc-Publizität deutscher Aktiengesellschaften fest. Ad-hoc-Publizität verfolgt das Ziel, durch Reduzierung oder Beseitigung bestehender Informationsungleichgewichte zwischen Managern und Aktionären die Transparenz auf dem Kapitalmarkt zu erhöhen und eine korrekte

Kursbildung zu ermöglichen. Emittenten von im amtlichen oder geregelten Markt notierten Wertpapieren sind gemäß § 15 WpHG verpflichtet, die Geschäftsleitung der zuständigen Börsen, das Bundesamt für Finanzdienstleistungsaufsicht und die Kapitalmarktteilnehmer unverzüglich über potentiell kursrelevante Tatsachen bzw. Unternehmensnachrichten zu informieren. Die Deutsche Gesellschaft für Ad-hoc-Publizität mbH (DGAP) mit Sitz in Frankfurt am Main ist eine wichtige Institution, die Ad-hoc-Mitteilungen im Auftrag der Emittenten vorab an Börsen und Aufsichtsbehörde sowie nach einer festgelegten Frist an die angeschlossenen nationalen und internationalen Nachrichtenagenturen elektronisch übermittelt [2]. Abbildung 1 zeigt eine über das Informationsportal der Deutschen Gesellschaft für Ad-hoc-Publizität veröffentlichte Ad-hoc-Meldung.

Abbildung 1: Ad-hoc-Mitteilung der SAP AG vom 07.01.2000

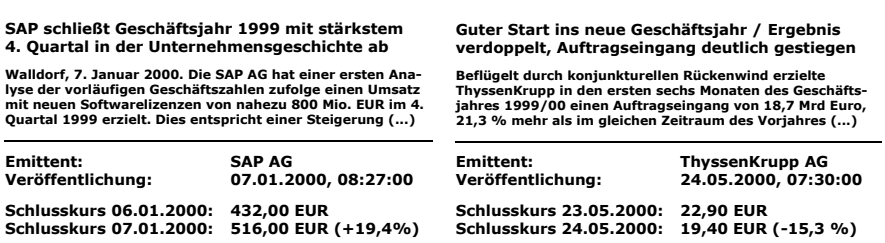
Ad-hoc-Meldungen sind öffentlich zugängliche Informationen, deren Inhalt nach dem Gesetz dazu geeignet sein sollte, die jeweiligen Aktienkurse wesentlich zu beeinflussen. Inwieweit Ad-hoc-Meldungen tatsächlich zu einer erheblichen Kursreaktion führen und somit kursrelevant sind, ist bisher noch ungeklärt. Auf einem informationseffizienten Kapitalmarkt sollte die Veröffentlichung von Ad-hoc-Meldungen unverzüglich durch Kauf- und Verkaufsentscheidungen der verschiedenen Handelsakteure zu einer entsprechenden Anpassung der Kurse führen. Somit kann von der tatsächlich realisierten Kursreaktion auf die Kursrelevanz des Meldungsinhaltes geschlossen werden. Dafür ist jedoch zumindest die Unterstellung der halbstarken Form der Informationseffizienz erforderlich, d.h. sämtliche historischen Marktdaten und alle öffentlich verfügbaren Informationen sind in den Aktienkursen berücksichtigt [5].
Seit der Gründung der DGAP am 1. Juli 1996 nutzen die Unternehmen dieses Informationsübermittlungsmedium sehr intensiv. So besitzen beispielsweise die etwa 30.000 in den Jahren 1999 bis 2002 über die DGAP veröffentlichten Ad-hoc-Meldungen ein Datenvolumen von etwa 94 MB. Jedoch weisen die im nächsten Abschnitt ausführlich vorgestellten empirische Studien darauf hin, dass

nur einige der nach § 15 WpHG veröffentlichten Ad-hoc-Meldungen eine erhebliche Kursreaktion verursachen.
Ziel der vorliegenden Studie ist es, die von DAX100-Unternehmen veröffentlichten Ad-hoc-Meldungen im Zeitraum 1999 bis 2002 auf Grundlage ihres textuellen Inhalts, d.h. ohne zusätzliche Auswertung kapitalmarkttheoretischer Kenntnisse, entsprechend ihrer Kursrelevanz zu klassifizieren. Dabei wird Text Mining eingesetzt, um Beziehungen zwischen dem Inhalt einer Meldung und ihrer tatsächlich verursachten Kursreaktion zu entdecken bzw. ein Klassifikationsmodell für die Kursrelevanzprognose von Ad-hoc-Meldungen zu generieren. Auf Basis dieser Ergebnisse könnten kursrelevante Mitteilungen mit hoher Wahrscheinlichkeit in der täglichen „Informationsflut“ identifiziert werden.
Eine Reduzierung der Informationsmenge auf kursrelevante Meldungen trägt zur Vermeidung von Informationsüberlastung der Kapitalmarktteilnehmer durch kursirrelevante Unternehmensnachrichten bei. In einer Studie von Farhoomand und Drury aus dem Jahr 2002 gaben von 124 Managern etwa 37 Prozent (bzw. 27 Prozent) an, täglich (bzw. häufig) Symptome der Informationsüberlastung zu spüren, deren häufigste Auswirkungen Zeitverlust, Negativeffekte in Bezug auf die Qualität der eigenen Arbeitsleistung, verminderte Effizienz sowie Frustration und Stress sind [4]. Etwa die Hälfte der Befragten nutzt deshalb Techniken der Informationsfilterung zur Verminderung der Informationsüberlastung. Im Gegensatz zu profilbasierten oder kollaborativen Relevanzfiltern wird in diesem Beitrag ein kapitalmarktorientiertes Verfahren der automatisierten Informationsselektion angewendet. Dabei wird die Relevanz einer Nachricht durch die Aktienkursentwicklung des jeweiligen Unternehmens nach deren Veröffentlichung bestimmt und somit ein objektiver, marktbasierter Relevanzfilter eingesetzt.
Im nächsten Abschnitt werden bisherige Untersuchungen zu Kurseffekten von Ad-hoc-Mitteilungen dargestellt. Anschließend werden in Abschnitt 3 die durchgeführte Ereignisstudie zur Kursrelevanzbestimmung von Ad-hoc-Mitteilungen und die Ergebnisse des generierten Klassifikationsmodells vorgestellt. Der Beitrag schließt mit einer Zusammenfassung und einem Ausblick.

2 Bisherige empirische Untersuchungen

Für den deutschen Kapitalmarkt existieren bereits mehrere empirische Studien zur Informationswirkung von Ad-hoc-Meldungen. So untersucht Röder die Kursreaktionen von 912 über die DGAP veröffentlichten Ad-hoc-Meldungen im Zeitraum 01.07.1996 bis 30.06.1997, die zuvor in inhaltsbezogene Kategorien eingeteilt wurden [13]. Er stellt fest, dass die Kurse am stärksten durch Mitteilungen über Auftragseingänge beeinflusst werden. Zudem führen Informationen über Dividendenzahlungen vergleichsweise zu schwachen, dennoch statistisch signifikanten Kursbewegungen, während Ad-hoc-Mitteilungen zum Jahresergebnis eine überdurchschnittliche Kursbewegung verursachen. Ebenso scheint die Kursrelevanz der Ad-hoc-Meldung von der Marktkapitalisierung des anzeigenden Unternehmens abzuhängen. In der von Röder untersuchten Stichprobe reagieren Aktien mit niedriger Marktkapitalisierung stärker aber auch langsa-

mer, d.h. über den Veröffentlichungstag hinaus, auf positive oder negative Ad-hoc-Meldungen als Aktien des DAX- oder MDAX-Portefeuilles.
Andere empirische Studien legen Indizien vor, dass nicht jede Ad-hoc-Meldung eine kursbeeinflussende Wirkung besitzt. Nowak weist beisp. darauf hin, dass nicht mehr als ein Drittel aller Ad-hoc-Meldungen im Zeitraum von 01.01.1995 bis 31.12.1996 statistisch signifikant kursrelevant waren [12, S. 465]. Bei guten Marktbedingungen scheinen besonders Unternehmen des ehemaligen Neuen Marktes Ad-hoc-Mitteilungen als Werbemedium zu nutzen, da sie in diesen Zeiten tendenziell mehr Meldungen veröffentlichen als bei schlechter Marktlage. Demgegenüber zögern sie vermutlich die Veröffentlichung von negativen Unternehmensmeldungen hinaus [7, S. 13-14].
Die Autoren dieser Studien haben den Inhalt von Ad-hoc-Meldungen manuell entsprechend ihres Inhalts kategorisiert (z.B. Umsatz, Gewinn, Dividendenankündigung oder Kapitalerhöhung) und anschließend die Informationswirkung bzw. Kursrelevanz von Meldungen oder Meldungsgruppen untersucht. Im Gegensatz dazu erfolgt in diesem Beitrag keine ex-ante inhaltsbezogene Gruppierung von Ad-hoc-Meldungen. Es wird vielmehr Text Mining eingesetzt, um ein Klassifikationsmodell zur automatischen Kursrelevanzprognose zu generieren. Ad-hoc-Meldungen sollen ohne manuellen Eingriff mit großer Wahrscheinlichkeit einer der zwei Klassen *kursrelevant* bzw. *kursirrelevant* zugeordnet werden, um eine automatisierte Informationsselektion zu ermöglichen.
Abbildung 2 illustriert links eine positiv kursrelevante Ad-hoc-Meldung der SAP AG, die zu einer Kurssteigerung von 19,4 Prozent am Tag ihrer Wirksamkeit führte. Die Herausforderungen einer Kursrelevanzprognose wird in der auf den ersten Blick positiven Ad-hoc-Meldung der ThyssenKrupp AG auf der rechten Seite von Abbildung 2 angedeutet. Trotz scheinbar guter Unternehmensnachrichten in der gesamten Meldung ergab sich am Tag ihrer Wirksamkeit ein relevanter Kursverlust von 15,3 Prozent.



SAP schließt Geschäftsjahr 1999 mit stärkstem 4. Quartal in der Unternehmensgeschichte ab

Walldorf, 7. Januar 2000. Die SAP AG hat einer ersten Analyse der vorläufigen Geschäftszahlen zufolge einen Umsatz mit neuen Softwarelizenzen von nahezu 800 Mio. EUR im 4. Quartal 1999 erzielt. Dies entspricht einer Steigerung (...)

Emittent:	**SAP AG**
Veröffentlichung:	**07.01.2000, 08:27:00**
Schlusskurs 06.01.2000:	**432,00 EUR**
Schlusskurs 07.01.2000:	**516,00 EUR (+19,4%)**

Guter Start ins neue Geschäftsjahr / Ergebnis verdoppelt, Auftragseingang deutlich gestiegen

Beflügelt durch konjunkturellen Rückenwind erzielte ThyssenKrupp in den ersten sechs Monaten des Geschäftsjahres 1999/00 einen Auftragseingang von 18,7 Mrd Euro, 21,3 % mehr als im gleichen Zeitraum des Vorjahres (...)

Emittent:	**ThyssenKrupp AG**
Veröffentlichung:	**24.05.2000, 07:30:00**
Schlusskurs 23.05.2000:	**22,90 EUR**
Schlusskurs 24.05.2000:	**19,40 EUR (-15,3 %)**

Abbildung 2: Zwei kursrelevante Ad-hoc-Meldungen

Die Entwicklung des Aktienkurses ist als alleiniger Indikator für die Kursrelevanz einer Ad-hoc-Mitteilung nicht geeignet und wird deshalb in Abbildungen 2 nur zur Illustration des Konzepts verwendet. Für die Ermittlung der Kursrelevanz einer Ad-hoc-Mitteilung erfolgt in dieser Fallstudie eine Bereinigung der Kursentwicklung um Einflüsse des Kapitalmarktes. Das dabei angewendete Verfahren wird in Abschnitt 3.2 detailliert vorgestellt.

3 Kursrelevanzprognose von Ad-hoc-Mitteilungen

In diesem Abschnitt erfolgt zunächst die Beschreibung der verwendeten Datenbasis. Anschließend wird die Methode zur Schätzung der Kursrelevanz einer Ad-hoc-Mitteilung eingeführt und die Erzeugung eines Klassifikationsmodells zur automatisierten Kursrelevanzprognose von Ad-hoc-Meldungen vorgestellt.

3.1 Datenbasis der Fallstudie

Grundlage der Fallstudie sind alle Ad-hoc-Meldungen, die von der Deutschen Gesellschaft für Ad-hoc-Publizität im Zeitraum vom 1. Januar 1999 bis 31. Dezember 2002 im Auftrag der Emittenten veröffentlicht wurden. Diese enthalten neben der nach § 15 WpHG potentiell kursrelevanten Tatsache in Textform auch das Datum und die Uhrzeit der Veröffentlichung.

Tabelle 1: Datenbasis der Fallstudie (Teil 1)

Datenbasis I: Mitteilungen von Unternehmen mit Veröffentlichung durch die DGAP von 01.01.1999 bis 31.12.2002 in ca. 94 MB großer Textdatei	29.552
Mitteilungen in Englisch von Unternehmen mit Veröffentlichung durch die DGAP von 01.01.1999 bis 31.12.2002	- 8.332
Datenbasis II: Mitteilungen in Deutsch von Unternehmen mit Veröffentlichung durch die DGAP von 01.01.1999 bis 31.12.2002	= 21.220

Ad-hoc-Mitteilungen müssen nach § 15 WpHG in Deutsch, können jedoch auch zeitgleich in Englisch veröffentlicht werden. Einige Meldungen enthalten sowohl eine deutsche als auch eine englische Fassung des Inhalts. Wie in Tabelle 1 zusammengefasst, wurden deshalb zunächst Meldungen und Meldungsteile in englischer Sprache aus dem Archiv der insgesamt im betrachteten Zeitraum durch die DGAP veröffentlichten Mitteilungen entfernt. Dazu wurde eine angepasste Version des Programms TextCat [11] zur Sprachidentifikation eingesetzt.

Anschließend wurden, wie in Tabelle 2 dargestellt, ausschließlich Ad-hoc-Mitteilungen nach § 15 WpHG von 136 Unternehmen extrahiert, die zu irgendeinem Zeitpunkt zwischen 1999 und 2002 dem DAX100 angehörten. Dieser Aktienindex umfasst annähernd die einhundert bezüglich der Marktkapitalisierung größten und umsatzstärksten Unternehmen des Amtlichen Handels der Frankfurter Wertpapierbörse. Es erfolgt eine Beschränkung auf größere Unternehmen, da deren Aktien im Vergleich zu kleineren Unternehmen häufiger gehandelt werden. Somit steht am Tag der Wirksamkeit einer Ad-hoc-Meldung auch häufiger ein Transaktionskurs zur Verfügung, dessen Vorliegen eine wichtige Voraussetzung für die korrekte Bestimmung der Kursrelevanz von Ad-hoc-Meldungen ist. Zusätzlich erfolgte eine Bereinigung der Datenbasis um sämtliche Mitteilungen, die nicht aufgrund § 15 WpHG veröffentlicht wurden.

Tabelle 2: Datenbasis der Fallstudie (Teil 2)

Datenbasis II: Mitteilungen in Deutsch von Unternehmen mit Veröffentlichung durch die DGAP von 01.01.1999 bis 31.12.2002	21.220
Mitteilungen von Unternehmen ohne Listung im Index DAX100 zu irgendeinem Zeitpunkt zwischen 01.01.1999 und 31.12.2002	- 18.772
Datenbasis III: Mitteilungen (DAX100) in Deutsch mit Veröffentlichung durch die DGAP von 01.01.1999 bis 31.12.2002	= 2.448
Nicht auf Grundlage des § 15 WpHG veröffentlichten Mitteilungen (z.B. Unternehmensnachrichten oder Meldungen von sog. Director's Dealings)	- 134
Datenbasis IV: Ad-hoc-Meldungen (DAX100) in Deutsch mit Veröffentlichung durch die DGAP von 01.01.1999 bis 31.12.2002	= 2.314

Für die zur Ermittlung der Kursrelevanz erforderliche Renditeberechnung werden tägliche, um Dividenden und andere Kapitalmaßnahmen bereinigte Schlusskurse aus Datastream verwendet. Aus dieser kommerziellen Datenbank stammen ebenfalls die verwendeten Indexstände des Composite DAX (CDAX).

3.2 Ermittlung der Kursrelevanz von Ad-hoc-Mitteilungen

Die Kursrelevanz einer Ad-hoc-Meldung ist Zielvariable bzw. abhängige Variable, deren Ausprägung (d.h. *kursrelevant* bzw. *kursirrelevant*) bei neuen Mitteilungen durch das Klassifikationsmodell nur auf Basis des textuellen Inhalts und ggf. weiterer Metadaten zu prognostizieren ist. Für das Training eines Klassifikationsmodells wird ein sog. Trainingsdatensatz benötigt, bei dem die Ausprägung der Zielvariable für alle Datensätze ex-ante gegeben ist. Die Schätzung der Klassifikationsqualität erfolgt anschließend mittels des sog. Testdatensatzes, bei dem ebenfalls die Attributwerte der Zielvariablen bekannt sein müssen [3, S. 107-109]. In Ermangelung vorklassifizierter Trainings- und Testdaten muss deshalb in dieser Fallstudie zunächst die Kursrelevanz für Ad-hoc-Meldungen der Datenbasis IV (vgl. Tabelle 2) bestimmt werden. Mitteilungen mit geschätzter Kurswirkung können anschließend als disjunkte Datensätze anteilig für Training und Test eines Klassifikationsmodells verwendet werden.
In der vorliegenden Fallstudie wird zur Unterscheidung zwischen kursrelevanten und kursirrelevanten Meldungen die Ereignisstudien-Methodik genutzt. Die Ereignisstudie ist eine theoretisch fundierte Untersuchungsmethode der empirischen Kapitalmarktforschung, die sich besonders zur Bestimmung der kursbeeinflussenden Wirkung von Ad-hoc-Mitteilungen eignet [12, S. 450-451]. Dabei stellt die Veröffentlichung einer Ad-hoc-Mitteilung das Ereignis dar.
Im ersten Schritt muss der Ereignistag (Tag 0) bzw. der Tag der Wirksamkeit für jede Meldung festgelegt werden. Während Ad-hoc-Meldungen 24 Stunden am Tag abgegeben werden können, öffnet die Börse zu bestimmten Handelszeiten. Im Untersuchungszeitraum wurden diese Handelszeiten zweimal verlängert: am 18. September 1999 von 9:00-17:00 Uhr auf 9:00-17:30 Uhr und am 2. Juni 2000 auf 9:00-20:00 Uhr. Der Ereignistag entspricht in der Fallstudie dem Ver-

öffentlichungstag einer Ad-hoc-Meldung, wenn deren Veröffentlichung am gleichen Tag vor Ende der Handelszeit erfolgte. Bei Abgabe einer Ad-hoc-Meldung nach Beendigung der Handelszeit wird der nachfolgende Handelstag als Ereignistag definiert. Zur Isolierung ihrer Kurswirkung werden nur Mitteilungen berücksichtigt, deren Emittenten am gleichen Tag genau eine veröffentlichten.

Zur Bestimmung der durch eine Ad-hoc-Meldung verursachten Kursreaktion muss die gesamte Kursveränderung der Aktie am Ereignistag um die Kursreaktion bereinigt werden, die ohne Veröffentlichung der Ad-hoc-Meldung erwartet worden wäre. Zur Schätzung dieser erwarteten Rendite kann u.a. das Marktmodell verwendet werden, das einen bestimmten renditegenerierenden Prozess unterstellt [15]. Das Marktmodell nimmt an, dass die Rendite $R_{i,t}$ einer Aktie des Unternehmens, das die Ad-hoc-Mitteilung *i* veröffentlichte,[1] am Tag *t* aus einem unternehmensspezifischen Bestandteil a_i und einem von der allgemeinen Marktentwicklung abhängigen Bestandteil $b_i \cdot R_{CDAX,t}$ sowie aus der Störgröße $e_{i,t}$ besteht. Die allgemeine Entwicklung des Kapitalmarkts wird in dieser Fallstudie durch die Rendite des Composite DAX (CDAX) approximiert, der alle Unternehmen des Frankfurter Amtlichen Handels, des Geregelten Marktes und des Neuen Marktes umfasst.

$$\begin{aligned} R_{i,t} &= a_i + b_i \cdot R_{CDAX,t} + e_{i,t} \\ t &= -55, \ldots, -6 \\ i &= 1, \ldots, N \end{aligned} \tag{1}$$

Die Parameterwerte a_i und b_i der Formel (1) werden für jede der *N* Aktien über einem 50 Tage langen Schätzzeitraum von Tag –55 bis Tag –6 bestimmt. Der Schätzzeitraum muss zur Abbildung des gewöhnlichen Zusammenhanges zwischen der Rendite der Aktie *i* und des Marktportefeuilles so definiert werden, dass zum einen an diesen Tagen keine ungewöhnlichen Ereignisse die Aktien betreffen und dass zum anderen die zu untersuchende Ad-hoc-Meldung noch keine Kursreaktionen bewirkt hat. Aktien, für die weniger als 35 Renditen im Schätzzeitraum zur Verfügung stehen, werden aus der Untersuchung ausgeschlossen, um eine möglichst genaue Schätzung zu gewährleisten. Ebenfalls erfolgt aufgrund der Vermutung, dass keine Transaktionskurse vorliegen, ein Ausschluss von Aktien mit einer Rendite von 0% am Tag 0, am vorhergehenden oder nachfolgenden Tag. Diese Vorgehensweise ist notwendig, da bei Kursen aus Datastream nicht zwischen einem im Vergleich zum Vortag konstanten Kurs oder einem vom Vortag fortgeschriebenen Kurs unterschieden werden kann. Mit Hilfe der geschätzten Parameterwerte wird die erwartete Rendite am Ereignistag berechnet. Die Differenz aus gesamter Rendite $R_{i,0}$ und erwarteter Rendite am

[1] Im Folgenden wird zur besseren Veranschauung der Laufindex *i* der Aktie direkt zugeordnet, die von der jeweiligen Ad-hoc-Meldung betroffen ist.

Ereignistag $E(R_{i,0})$ stellt den Schätzwert für die kursbeeinflussende Wirkung der Mitteilung dar und wird als Überrendite oder abnormale Rendite $\hat{A}_{i,0}$ bezeichnet.

$$\hat{A}_{i,0} = R_{i,0} - E(R_{i,0}) \tag{2}$$

$$\hat{A}_{i,0} = R_{i,0} - (\hat{a}_i + \hat{b}_i \cdot R_{CDAX,0}) \quad i = 1,\ldots,N \tag{3}$$

Um eine Aussage darüber treffen zu können, ob die geschätzten Überrenditen statistisch signifikant von null abweichen oder nur zufällig um null schwanken, wird ein t-Test durchgeführt. Meldungen, die eine zum Niveau von 10% statistisch signifikante positive bzw. negative Überrendite bewirken, werden als positiv bzw. negativ kursrelevant bezeichnet. Kursirrelevant sind alle Ad-hoc-Meldungen mit Überrenditen, die nicht statistisch signifikant von null abweichen.

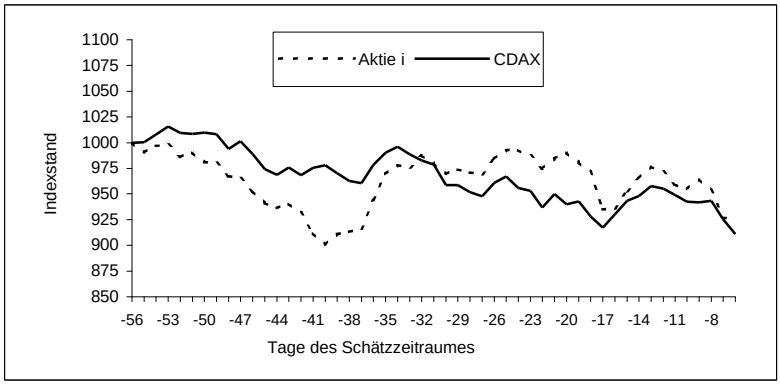


Abbildung 3: Tägliche Indexentwicklung einer Aktie *i* und des CDAX im Schätzzeitraum (Ausgangsbasis: Tag –56 = 1000)

Die obigen Ausführungen werden in den Abbildungen 3 und 4 an einem Beispiel graphisch verdeutlicht. Hierfür wird angenommen, dass die Rendite einer Aktie *i* am Ereignistag 3% beträgt, während das Marktportefeuille in Form des CDAX eine Rendite von 1% besitzt. Im Schätzzeitraum von Tag –55 bis Tag –6 weist die Aktie *i* gegenüber dem CDAX die in Abbildung 3 dargestellte tägliche Kursentwicklung auf. Abbildung 4 gibt für diese Zeitperiode die zugehörigen Renditekombinationen zwischen Aktie *i* und CDAX als Punktwolke an. Um die Rendite der Aktie *i* am Tag 0, die ohne Veröffentlichung der Ad-hoc-Meldung erwartet worden wäre, prognostizieren zu können, wird aus den verfügbaren Renditekombinationen das Marktmodell geschätzt. Da die Rendite des CDAX am Tag 0 annahmegemäß 1% ist, berechnet sich nach dem geschätzten Marktmodell ($\hat{a}_i = -0{,}002$, $\hat{b}_i = 0.799$) eine erwartete Rendite für die Aktie *i* von 0,8%. Die tatsächliche Rendite der Aktie *i* ist allerdings 3%, dargestellt in Abbildung 4

als Kreuz. Die sich daraus ergebene Überrendite von 2,2% ist auf einem Niveau von 10% statistisch signifikant von null verschieden, da die realisierte Renditekombination der Aktie *i* und des CDAX am Ereignistag außerhalb des Konfidenzintervalls liegt, das das geschätzte Marktmodell umgibt.

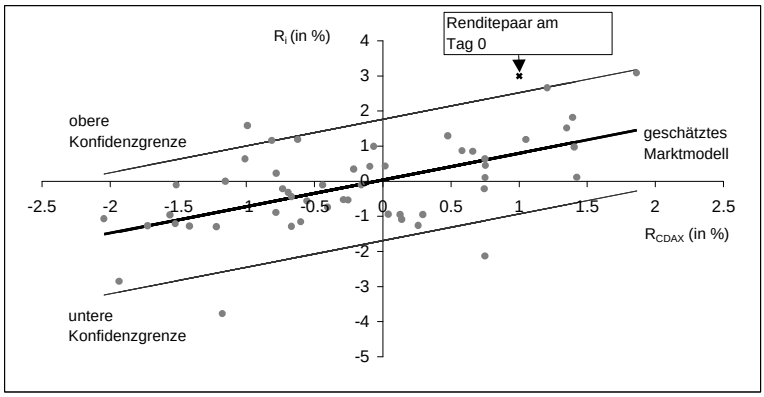


Abbildung 4: Schätzung des Marktmodells und des Konfidenzintervalls auf Basis der Renditekombinationen der Aktie *i* und des CDAX im Schätzzeitraum

Tabelle 3: Datenbasis der Fallstudie (Teil 3)

Datenbasis IV: Ad-hoc-Meldungen (DAX100) in Deutsch mit Veröffentlichung durch die DGAP von 01.01.1999 bis 31.12.2002	2.314
Ad-hoc-Meldungen mit Veröffentlichungsdatum, an dem deren Emittenten mehr als eine Ad-hoc-Meldung veröffentlichten	- 320
Ad-hoc-Meldungen, für deren Kursrelevanzschätzung weniger als 35 Renditen in Schätzperiode oder Nullrenditen im Ereignisfenster vorliegen	- 534
Datenbasis V: Bereinigte Ad-hoc-Meldungen (DAX100) in Deutsch mit Veröffentlichung durch die DGAP von 01.01.1999 bis 31.12.2002	= 1.460

Tabelle 3 gibt einen Überblick über die Schritte der Datenbereinigung, die während der Ermittlung der Kursrelevanz für Ad-hoc-Mitteilungen der Datenbasis IV vorgenommen wurden. In der Datenbasis V gibt es unter den 1.460 (100%) für die Kursrelevanzprognose verwendbaren Ad-hoc-Meldungen 235 (16,1%) zum Niveau von 10% statistisch signifikant positiv bzw. 161 (11,0%) negativ kursrelevante sowie 1.064 (72,9%) kursirrelevante Ad-hoc-Mitteilungen.

3.3 Aufbereitung der Ad-hoc-Meldungen

Die Vorverarbeitung der bereinigten Ad-hoc-Mitteilungen (Datenbasis V) umfasst den Einsatz der DIAsDEM Workbench [6; 17], um zusätzliche Metadaten in Form häufiger semantischer Konzepte zu extrahieren. Das dabei genutzte DIAsDEM Vorgehensmodell zur semantischen Auszeichnung anwendungsspe-

zifischer Textarchive beinhaltet einen interaktiven und iterativen Prozess der Wissensentdeckung. Dessen Ziel ist die inhaltsbezogene Annotation von strukturellen Textelementen (z.B. Sätzen) mit XML-Textmarken und die Ableitung einer XML-Dokumenttypdefinition. Die DTD beschreibt dabei die semantische, d.h. die inhaltliche Struktur des Archivs. Abbildung 5 illustriert auszugsweise links die in der Abbildung 1 gezeigte Ad-hoc-Mitteilung als inhaltsbezogen ausgezeichnetes XML-Dokument sowie rechts die darin referenzierte XML DTD.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE AdHocMitteilung SYSTEM 'AdHocMitteilung.dtd'>

<AdHocMitteilung> SAP schließt Geschäftsjahr 1999 mit
stärkstem 4. Quartal in der Unternehmensgeschichte ab.
Walldorf, 7. Januar 2000. <QuartalUmsatz Company="1010;
SAP AG" AmountOfMoney= "800000000 EUR"> Die SAP AG hat
einer ersten Analyse der vorläufigen Geschäftszahlen zu-
folge einen Umsatz mit neuen Softwarelizenzen von nahezu
800 Mio. EUR im 4. Quartal 1999 erzielt. </QuartalUmsatz>
<ErhoehungKennzahl> Dies entspricht einer Steigerung ge-
genüber dem Vorjahresquartal von ungefähr 40% und ge-
genüber dem Vorquartal von rund 150%. </Erhoehung
Kennzahl> <QuartalUmsatz> Der Gesamtumsatz im 4. Quar-
tal 1999 wuchs um rund 25% gegenüber dem Vorjahres-
quartal. </QuartalUmsatz> (...) Genaue Angaben zum vor-
läufigen Ergebnis des Geschäftsjahrs 1999 wird die SAP AG
am 24. Januar 2000 veröffentlichen. </AdHocMitteilung>
```

```
<?xml version="1.0" encoding="ISO-8859-1"?>

<!ELEMENT AdHocMitteilung ( #PCDATA |
AbschluszMasznahmePositiv | Abschreibung |
AdHocMitteilung | Akquisition | AkquisitionUmsatz |
AktieErhoehung | Ansprechpartner | AufsichtsratBeschlusz
| AufsichtsratDividende | AufsichtsratFunktionstraeger
| AufsichtsratJahresabschlusz | AufsichtsratZustimmung
| Auftragseingang | AuftragseingangPositiv |
AussageNegativ | ... | Gewinn | GewinnJeAktie |
GewinnPlanung | GewinnPositiv | GewinnRueckgang |
GewinnUmsatz | Grundkapital | ... | Zwischenbericht )* >

<!ELEMENT AbschluszMasznahmePositiv (#PCDATA)>
<!ELEMENT Abschreibung (#PCDATA)> ...
<!ELEMENT Zwischenbericht (#PCDATA)>

... <!ATTLIST QuartalUmsatz Company CDATA #IMPLIED>
<!ATTLIST QuartalUmsatz AmoutOfMoney CDATA #IMPLIED> ...
```

Abbildung 5: Semantisch annotierte Ad-hoc-Meldung mit XML DTD

Im Ergebnis dieser Fallstudie ist auch eine erste Tendenzaussage zu treffen, ob eine Verwendung der entdeckten semantischen Konzepte (XML-Textmarken) die Klassifikationsgüte erhöhen kann. Dazu werden auf Grundlage von Datenbasis V (vgl. Tabelle 3) zwei Klassifikationsmodelle mit jeweils unterschiedlichen unabhängigen Variablen trainiert. In beiden Szenarien A und B werden neben dem textuellen Mitteilungsinhalt die nominal skalierte Branchenkennung und die Firma des Emittenten sowie die ordinal skalierte Stunde, der Wochentag und die Kalenderwoche der Veröffentlichung zur Prognose der Kursrelevanz verwendet. Im Szenario B werden außerdem die auf Satzebene entdeckten 129 semantischen Konzepte als zusätzliche, strukturierte Metadaten genutzt. Diese binärskalierten Indikatorvariablen zeigen dabei die Existenz bzw. Nichtexistenz des entsprechenden Konzepts in der jeweiligen Ad-hoc-Mitteilung an.

3.4 Klassifikation von Ad-hoc-Mitteilungen

Für die Klassifikation von Ad-hoc-Mitteilungen hinsichtlich ihrer Kursrelevanz wurde der SAS Enterprise Miner [14] aus der Vielzahl gegenwärtig am Markt verfügbarer allgemeiner Data-Mining-Software [18] und dedizierter Text-Mining-Applikationen [9] ausgewählt. Der SAS Enterprise Miner integriert nahtlos spezielle Textanalysefunktionen (z.B. Textaufbereitung, Dimensionsreduktion und Segmentierung von Dokumenten) und anerkannte, nicht textoptimierte Data-Mining-Algorithmen (z.B. Klassifikationsverfahren). Abbildung 6 zeigt das nach vielen Iterationen des Wissensentdeckungsprozesses letztendlich für die Klassifikation der Ad-hoc-Mitteilungen in beiden Szenarien verwendete Text-Mining-Diagramm. Die Knoten *Eingabedaten*, *Datenpartitionierung*, *Text Mining* für Textzerlegung und Dimensionsreduktion sowie *Variablenselektion* und *Oversampling* des Graphen sind dabei der Aufbereitungsphase des Prozes-

ses der Wissensentdeckung in textuellen Datenbanken zuzuordnen. Der Knoten *Logistische Regression* dient der Musterentdeckung bzw. der Generierung von Klassifikationswissen zur Unterscheidung von relevanten und irrelevanten Meldungen. Eine Bewertung der Qualität des gelernten Klassifikators und somit eine Nachbereitung der Ergebnisse ermöglicht der abschließende Knoten *Evaluation*. Die gerichteten Kanten des Graphen symbolisieren den Fluss von Trainings-, Test- und Metadaten zwischen den einzelnen Verarbeitungsschritten.

Abbildung 6: Text-Mining-Diagramm zur Klassifikation von Ad-hoc-Mitteilungen

Die Eingabedaten für beide Szenarien enthalten neben den in Abschnitt 4.3 genannten strukturierten Variablen den unstrukturierten, textuellen Inhalt der 1460 bereinigten Ad-hoc-Meldungen in Datenbasis V. Aufgrund der geringen Anzahl von Meldungen mit geschätzter Ausprägung der Zielvariable wurden in beiden Szenarien jeweils 90 Prozent der Ad-hoc-Mitteilungen für Trainings- und 10 Prozent für Testzwecke verwendet. Dabei wurden zufällige, aber geschichtete Trainings- und Testpartitionen erzeugt, in denen der Anteil relevanter Meldungen jeweils dem in der Grundgesamtheit entspricht.
Der Knoten *Text Mining* kapselt einerseits elementare Textaufbereitungsfunktionalität. Diese umfasst die Zerlegung textueller Inhalte in individuelle, sämtlich kleingeschriebene Wörter, die Entfernung von Interpunktionszeichen und die Eliminierung sinnleerer Wörter anhand einer angepassten Stoppwortliste mit 385 Termen wie z.B. „ein“, „die“ und „etwa“. Außerdem wurde eine spezielle Synonymliste mit 182 Termen eingesetzt, um synonyme Ausdrücke wie z.B. „Gewinnanstieg“ und „Gewinnerhöhung“ auf ein gemeinsames semantisches Konzept abzubilden. Auf die Identifizierung benannter Entitäten (z.B. Personen oder Datumsangaben) wurde bewusst verzicht, um ein möglichst allgemeingültiges Klassifikationsmodell zu generieren. Trainings- und Testarchiv wurden danach in eine binäre Wort-je-Dokument-Matrix überführt. Durch Multiplikation der binären Worthäufigkeiten mit der Testgröße des χ^2-Unabhängigkeitstests auf gemeinsame Verteilung des jeweiligen Worts und der Zielvariable wurden Terme zusätzlich entsprechend ihrer Korrelation mit der Zielvariablen gewichtet.
Andererseits kapselt der Knoten *Text Mining* auch die erforderliche Dimensionsreduktion. Die eingesetzte Singulärwertzerlegung ist ein Verfahren der latentsemantischen Analyse, das die Anzahl der Dimensionen je Dokument auf etwa ein- bis zweihundert numerische Attribute reduzieren kann [16, S. 337-341; 1]. Diese Methode der linearen Algebra verringert die Dimensionalität der gewichteten Wort-je-Dokument-Matrix durch Transformation in eine semantisch komprimierte Ergebnismatrix, die lediglich statistisch bedeutsame Inhalte enthält. Dadurch wird die Relevanz bedeutungsloser Wörter gemindert und die Identifikation von relevanten Termen und Termkombinationen ermöglicht.

Nach erfolgter Dimensionsreduktion wird der textuelle Inhalt einer Ad-hoc-Meldung durch maximal 200 numerische Singulärwerte beschrieben. Zusätzlich zeigen 100 neue Attribute das Auftreten der einhundert im gesamten Archiv am höchsten gewichteten Terme (z.B. „Wachstum“ und „veräußern“) in den jeweiligen Meldungen an. Der sich anschließende Knoten *Variablenselektion* kapselt die automatische Entfernung von Variablen vor der Modellbildung, die einen parametrisierbaren Schwellenwert der Korrelation mit der Zielvariable nicht überschreiten. Im Knoten *Oversampling* wird der Anteil kursrelevanter Meldungen auf 50 Prozent übergewichtet, da ein Klassifikationsmodell zur Vorhersage dieser unterrepräsentierten Klasse zu trainieren ist. Für die Klassifikation der Mitteilungen wurde ein Verfahren der logistischen Regression ausgewählt. Diese Methode prognostiziert die Wahrscheinlichkeit für das Auftreten des Attributwerts *kursrelevant* der Zielvariable bei gegebenen Attributwerten der Eingabevariablen. Ist diese Wahrscheinlichkeit größer gleich ein Schwellenwert (hier: ein Drittel), so wird die entsprechende Ad-hoc-Meldung als kursrelevant klassifiziert. Die Parameterwahl für Variablenselektion und logistische Regression kann hier aus Platzgründen nicht diskutiert werden.
Um trotz des erforderlichen großen Anteils an Trainingsdaten eine möglichst genaue Schätzung des Klassifikationsfehlers durchführen zu können, wurde eine 10-fache Überkreuz-Validierung durchgeführt [3, S. 109]. In zehn Durchläufen wurden jeweils 90 Prozent der verfügbaren 1.460 Datensätze aus Datenbasis V (vgl. Tabelle 3) für das Training und die verbleibenden 10 Prozent für die Bewertung des jeweiligen Klassifikators verwendet. Tabelle 4 fasst die Ergebnisse der Evaluation der Klassifikationsgenauigkeit für die in Abschnitt 4.3 erläuterten zwei Szenarien zusammen. Der durchschnittliche Klassifikationsfehler ist dabei das arithmetische Mittel der Klassifikationsfehler der zehn im Rahmen der Überkreuz-Validierung generierten Klassifikationsmodelle.

Tabelle 4: Schätzung des Klassifikationsfehlers der zwei Klassifikatoren auf Testdaten

10-fache Überkreuz-Validierung auf Datenbasis V	Szenario A	Szenario B
Durchschnittlicher Klassifikationsfehler und 95%-Konfidenzintervall [3, S. 111]	0,39 [0,365; 0,415]	0,39 [0,365; 0,415]
Durchschnittlicher Klassifikationsfehler der Klasse *kursrelevant* und 95%-Konfidenzintervall	0,57 [0,545; 0,595]	0,59 [0,565; 0,615]
Durchschnittlicher Klassifikationsfehler der Klasse *kursirrelevant* und 95%-Konfidenzintervall	0,32 [0,296; 0,344]	0,31 [0,286; 0,334]

3.5 Interpretation der Ergebnisse

Die in Tabelle 4 dargestellten Ergebnisse der Fallstudie verdeutlichen die Herausforderungen einer Kursrelevanzprognose von Ad-hoc-Meldungen. Der durchschnittliche Klassifikationsfehler von 39 Prozent in beiden Szenarien der Datenbasis V ist ggf. noch hinnehmbar. Zielstellung dieser Fallstudie ist jedoch

die automatisierte Selektion kursrelevanter Ad-hoc-Mitteilungen zur Vermeidung von Informationsüberflutung. Der zur Beurteilung des Zielerreichungsgrades zu betrachtende durchschnittliche Klassifikationsfehler der Klasse *kursrelevant* liegt mit etwa 57 Prozent in Szenario A bzw. 59 Prozent in Szenario B jedoch noch weit entfernt vom Sollzustand eines automatischen Relevanzfilters für Unternehmensnachrichten: Von sämtlichen kursrelevanten Mitteilungen im Testarchiv wurden im Durchschnitt nur 43 Prozent (Szenario A) bzw. 41 Prozent (Szenario B) tatsächlich als kursrelevant klassifiziert.
Der durchschnittliche Klassifikationsfehler bei der Identifizierung kursrelevanter Ad-hoc-Meldungen ist in Szenario A etwa 2 Prozent geringer als in Szenario B. Bei etwa gleichem durchschnittlichen Klassifikationsfehler scheint also die Einbeziehung der auf Satzebene entdeckten 129 semantischen Konzepte als zusätzliche, strukturierte Metadaten zu keiner Verbesserung der Klassifikationsgüte zu führen. Im Gegenteil, eine Einbeziehung dieser Indikatorvariablen (z.B. „AuftragseingangPositiv“ oder „GewinnRückgang“) als unabhängige Variablen führte im Szenario B dieser Fallstudie offenbar zu einer Verschlechterung des durchschnittlichen Klassifikationsfehlers der Klasse *kursrelevant* um 2 Prozent. Diese Tatsache könnte intuitiv auf die Einführung zusätzlicher Freiheitsgrade durch die 129 Indikatorvariablen zurückzuführen sein, da die zur Dimensionsreduktion angewandte Singulärwertzerlegung [1] ebenfalls Semantik repräsentierende Linearkombinationen von Termen identifiziert und Dokumente anschließend auf diesen orthogonalen Vektorraum abbildet.
Ein Grund für die besondere Schwierigkeit einer Kursrelevanzprognose könnte der oft mit dem Schlagwort „Informationsmüll“ artikulierte Vorwurf sein, viele Ad-hoc-Meldungen enthielten neben tatsächlich potentiell kursbeeinflussenden Tatsachen im Sinne von § 15 WpHG eher allgemeine Nachrichten oder sogar Werbung [10]. Ebenso scheint es gängige Praxis der Finanzmarktkommunikation vieler Unternehmen zu sein, negative Nachrichten einerseits beschönigend darzustellen oder diese andererseits in „umhüllende“ positive Meldungen einzubetten. Beide Präsentationstechniken erfordern jedoch ein Lesen des kundigen Adressaten „zwischen den Zeilen“, um die Bedeutung sprachlicher Konstrukte im kommunikativen Zusammenhang zu betrachten. Diese pragmatische Perspektive der Sprachanalyse wurde, im Gegensatz zur semantischen Untersuchungsperspektive durch Einsatz einer Synonymliste, in dieser Fallstudie jedoch nicht berücksichtigt.

4 Zusammenfassung und Ausblick

In diesem Beitrag wurde die prinzipielle Möglichkeit wie auch die Komplexität einer automatisierten, kapitalmarktbasierten Kursrelevanzprognose von textuellen Ad-hoc-Mitteilungen vorgestellt. Es wurde Text Mining eingesetzt, um kapitalmarktrelevante von kapitalmarktirrelevanten Ad-hoc-Mitteilungen automatisiert zu unterscheiden. Eine operative Umsetzung eines derartigen objektiven Relevanzfilters, d.h. des trainierten Klassifikationsmodells, könnte beisp. zur

Minderung der wahrgenommenen Informationsüberflutung bei den Nutzern von Börseninformationsdiensten beitragen. Die Qualität der automatischen Informationsselektion muss jedoch vor einer Einbindung in operative Systeme durch künftige Forschungsanstrengungen verbessert werden.
Ein wichtiger Ansatzpunkt zur Verbesserung der Klassifikationsqualität ergibt aus der in Abschnitt 4.5 dargestellten sprachlichen Komplexität von Ad-hoc-Mitteilungen. Als Ersatz einfacher Synonymlisten könnte beisp. ein anwendungsspezifischer Thesaurus erstellt und eingesetzt werden, der neben Synonymen auch weitere Beziehungen (z.B. „Umsatzrendite“ als Unterbegriff von „Rendite“) zwischen Wörtern und Konzepten abbildet. Gegenwärtig wird eine Stoppwortliste verwendet, um sinnleere Wörter nicht in die Wort-je-Dokument-Matrix aufzunehmen. Bei Einsatz eines Thesaurus könnte im Gegensatz dazu eine sog. Startwortliste erstellt werden, die nur gültige Deskriptoren bzw. Schlagworte für Ad-hoc-Meldungen enthält. Darüber hinaus sollte eine Auflösung der semantischen Bedeutung von Homonymen („Rücktritt“ als CEO oder „Rücktritt“ vom Fusionsvertrag) ebenso zur Senkung des durchschnittlichen Klassifikationsfehlers beitragen wie der Einbezug der pragmatischen Perspektive: Eine „Erhöhung“ des Gewinns ist z.B. im Gegensatz zur „Erhöhung“ des Verlustes durchaus positiv. Zusätzlich sollten andere, häufig zur Textklassifikation verwendete Algorithmen wie z.B. der Bayes-Klassifikator [3, S. 114-116] oder Supportvektormaschinen [8] eingesetzt werden.

Danksagungen

Die Autoren danken besonders Herrn Prof. Stehle, Ph.D., Institut für Bank-, Börsen- und Versicherungswesen der Humboldt-Universität zu Berlin für wertvolle Hinweise und die Bereitstellung des Zugangs zu Datastream. Der Deutschen Gesellschaft für Ad-hoc-Publizität mbH wird für die unkomplizierte Überlassung der Ad-hoc-Meldungen für Forschungszwecke herzlich gedankt.

Literaturverzeichnis

1. Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K. and Harshman, R. (1990): Indexing by Latent Semantic Indexing. Journal of the American Society for Information Science, 41(6), p. 321-407.
2. Deutsche Gesellschaft für Ad-hoc-Publizität mbH (Hrsg., 2002): Finanzmarkt-Kommunikation aus einer Hand. 5. überarbeitete Auflage, Frankfurt am Main. http://www.dgap.de/downloads/dgap_service_dt.pdf, Abruf am 2003-03-31.
3. Ester, M. und Sander, J. (2000): Knowledge Discovery in Databases: Techniken und Anwendungen. Berlin, Heidelberg, Springer-Verlag.
4. Farhoomand, A. F. and Drury, D. H. (2002): Managerial Information Overload. In: Communications of the ACM, 45 (10), pp. 127-131.

5. Eugene F. Fama (1970): Efficient Capital Markets: A Review of Theory and Empirical Work. In: Journal of Finance, 25 (2), pp. 383-417.
6. Graubitz, H.; Spiliopoulou, M. and Winkler, K. (2001): The DIAsDEM Framework for Converting Domain-Specific Texts into XML Documents with Data Mining Techniques. In: Proceedings of the First IEEE International Conference on Data Mining, San Jose, CA, USA, November/December 2001, pp. 171-178.
7. Güttler, A. (2001): Wird die Ad-hoc-Publizität korrekt umgesetzt? Eine empirische Analyse unter Einbezug von Unternehmen des Neuen Markts, Arbeitspapier, Fachbereich Wirtschaftswissenschaften, Johann Wolfgang Goethe-Universität: Frankfurt am Main.
8. Joachims, T. (2002): Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms. Boston, Dordrecht, London, Kluwer Academic Publishers.
9. Kamphusmann, T. (2002): Text-Mining: Eine praktische Marktübersicht. Düsseldorf, Symposion Publishing.
10. Kaserer, C. und Nowak, E. (2001): Die Anwendung von Ereignisstudien bei Ad-hoc-Mitteilungen: Zugleich Stellungnahme zu dem Beitrag „Die Informationswirkung von Ad hoc-Meldungen" von Klaus Röder. In: : ZfB - Zeitschrift für Betriebswirtschaft 71 (11), S. 1353-1356.
11. Noord, G. van (2002): TextCat. http://odur.let.rug.nl/~vannoord/ TextCat/index.html, Abruf am 2003-02-15.
12. Nowak, E. (2001): Eignung von Sachverhalten in Ad-hoc-Mitteilungen zur erheblichen Kursbeeinflussung. In: ZBB – Zeitschrift für Bankrecht und Bankwirtschaft, 13 (6), S. 449-465.
13. Röder, K. (2000): Die Informationswirkung von Ad hoc-Meldungen. In: ZfB - Zeitschrift für Betriebswirtschaft 70 (5), S. 567-593.
14. SAS Institute Inc. (ed., 2002): SAS Text Miner: Distilling Textual Data for Competitive Business Advantage, A SAS White Paper. Cary, NC, USA. http://www.sas.com/apps/whitepapers/whitepaper.jsp, Abruf am 2003-03-31.
15. Sharpe, W. F. (1963): A Simplified Model for Portfolio Analysis. In: Management Science, 9 (2), pp. 277-293.
16. Sullivan, D. (2001): Data Document Warehousing and Text Mining. New York, John Wiley & Sons.
17. Winkler, K. and Spiliopoulou, M. (2001): Semi-Automated XML Tagging of Public Text Archives: A Case Study. In: Proceedings of EuroWeb 2001 “The Web in Public Administration”. Pisa, Italy, December 2001, pp. 271-285.
18. Wilde, K. D.; Hippner, H. und Merzenich, M. (Hrsg., 2002): Data Mining: Mehr Gewinn aus Ihren Kundendaten. Düsseldorf, Verlagsgruppe Handelsblatt.