Generierung von neuartigem Wissen durch die Analyse von Weblogs

Vorstellung eines Forschungsvorhabens

Andreas Hilbert, Andreas Schieber

Technische Universität Dresden – Professur für Wirtschaftsinformatik
Business Intelligence Research

1 Einleitung

In den vergangenen Jahren wurde die Entwicklung des World Wide Webs von einem Thema beherrscht: der Interaktivität (KAISER (2009), S. 379). Darunter wird das Mitgestalten von Web-Inhalten durch den Nutzer verstanden. Im Zuge dieser Entwicklung hat sich vor allem die Anzahl der Weblogs stark erhöht (WORDPRESS (2009)). Mit der Weblog-Suchmaschine Technorati beispielsweise ließen sich bereits letztes Jahr Informationen aus über 112 Mio. Weblogs finden (TECHNORATI (2008)).

Der Begriff Weblog ist eine Wortneubildung aus den Worten Web und Log(buch) (FISCHER (2007), S. 7). In einem Weblog werden – vergleichbar mit einem Tagebuch – aktuelle Ereignisse und Entwicklungen aus der realen oder virtuellen Welt auf einer Webseite im Internet dokumentiert. Inzwischen hat sich für den Begriff Weblog auch die Kurzform Blog etabliert.

In den Weblogs wird überwiegend über Erlebnisse aus dem täglichen Leben berichtet, um sie mit anderen – etwa Freunden und Bekannten – zu teilen, und so in Verbindung zu bleiben (ZERFAß & BOGOSYAN (2007), S. 7). Inzwischen schreiben die Autoren jedoch nicht mehr nur über Privates, sondern auch über allgemein interessante Themen wie Politik oder Wirtschaft und tauschen untereinander Erfahrungen mit Produkten oder Unternehmen aus. In den Diskussionen werden auf diese Weise sowohl Meinungen als auch Anregungen geäußert, mit deren Hilfe sich die Produkte oder Dienstleistungen eines Unternehmens im Sinne des Kunden weiterentwickeln und verbessern lassen.

Daher führen immer mehr Unternehmen einen eigenen Weblog als weiteren Kommunikationskanal zu ihren Kunden ein, um von ihnen etwa Ratschläge zur Produktverbesserung zu erhalten (WRIGHT (2006), S. 25). Diese Entwicklung verfolgt auch die Studie Euro-

Blog, die Marketing-Experten zu ihrem Umgang mit Weblogs befragt. Seit Beginn der Untersuchung im Jahr 2006 steigt die Anzahl der Unternehmen kontinuierlich an, die Weblogs als Informationsquelle heranziehen und selbst zur Kommunikation nutzen (ZERFAß ET AL. (2007), S. 13). Viele dieser Unternehmen geben dabei an, dass neben dem eigenen Weblog auch Einträge in anderen Weblogs beobachtet werden müssen, um zeitnah z.B. auf Beschwerden oder Produktmängel reagieren zu können (ZERFAß ET AL. (2007), S. 22). Auch DAVIS & OBERHOLTZER (2008), S. 3 weisen darauf hin, dass Blogs bei der Sammlung von Kundenfeedback eine wichtige Informationsquelle sind. Die Analyse von Weblogs und deren Einträgen ist daher eine Ergänzung zur Marktforschung, die wertvolle Ergebnisse liefert.

Die Zahl der bestehenden Weblogs und Einträge ist jedoch inzwischen unüberschaubar geworden, so dass es nur in sehr geringem Umfang möglich ist, die Einträge relevanter Weblogs auszuwerten und wichtige Inhalte zu erkennen. In der Literatur finden sich daher zahlreiche Arbeiten, die sich mit der Analyse von Weblogs beschäftigen. Mit Methoden der Business Intelligence wie Data und Text Mining lassen sich wesentliche Einträge aus der unübersichtlichen Menge separieren. Als Datenstamm für diese Methoden müssen die Blog-Einträge zunächst im World Wide Web gefunden und extrahiert werden. Dazu ist ein System zu entwickeln, das Weblogs im Web identifiziert und deren Einträge speichert. Im Anschluss sollen diese Daten analysiert werden können, um Informationen, wie beispielsweise Kundenmeinungen, herauszufiltern.

2 Forschungsziele

Die Ziele der Forschungsarbeit lassen sich in einen theoretischen, erkenntniszielgeleiteten Teil sowie in einen konzeptionellen, gestaltungszielorientierten Teil gliedern; als Gestaltungsziel wird ein Konzept für ein Blog-Analysesystem erstellt, das anschließend prototypisch implementiert und evaluiert wird.

Der theoretische Bereich umfasst daher als Erkenntnisziel einerseits die Aufarbeitung des aktuellen Forschungsstands zu Methoden und Vorgehensweisen bei der Analyse von Weblogs und deren Einträgen. Andererseits beinhaltet dieser Teil auch eine Bestimmung möglicher Analyseziele, von denen einige in Abschnitt 4.4 vorgestellt werden.

Die Konzeption erläutert daraufhin den Aufbau eines Systems, das – im Sinne eines Management Support Systems – die Ergebnisse der Analysen geeignet darstellen bzw. den Anwendern bereitstellen kann. Im Rahmen der Analyse von Weblogs kommen beispielsweise Vergleiche von Diskussionen aus Blog-Einträgen und -Kommentaren in Betracht, z.B. zu welcher Zeit besonders über ein bestimmtes Thema diskutiert wurde. Analog kann für Unternehmen wichtig sein, die Verkaufszahlen ihrer Produkte mit den Diskussionen

in den Weblogs in Bezug zu bringen; in diesem Zusammenhang kann das Analysesystem als Messinstrument für Marketing-Aktivitäten dienen, wenn anhand der Diskussionsintensität erkannt werden kann, dass unterschiedliche Werbemethoden verschiedene Diskussionsintensitäten in der Blogosphäre bewirken.

Nach der Konzeption befasst sich der letzte Teil der Arbeit mit der prototypischen Umsetzung eines solchen Weblog-Analyse-Systems. Dabei soll – nach Möglichkeit mit Opensource-Tools – eine Implementierung geschaffen werden, welche die Untersuchung von Weblogs sowie die Darstellung der Ergebnisse ermöglicht.

Das Forschungsprojekt befindet sich noch am Anfang. Bisher wurde vor allem aktuelle Literatur zum Thema Blog Mining untersucht. Im Fokus standen dabei Vorgehensweisen und Methoden bei der Analyse von Weblogs. Die so gewonnenen Erkenntnisse sollen im Folgenden beschrieben werden. Dazu gehören zunächst die Definition des Begriffs Weblog sowie die Beschreibung der Eigenschaften eines Blogs. Im Anschluss wird das Ergebnis der Literaturanalyse in Bezug auf aktuelle Verfahren für das Blog Mining dargestellt und systematisch aufgearbeitet. Das Vorgehen bei der Analyse von Weblogs orientiert sich dabei an einem Standardprozess für Data Mining, der in jeder Phase an die spezifischen Anforderungen beim Blog Mining angepasst wird.

3 Eigenschaften von Weblogs

Zur Differenzierung zwischen gewöhnlichen Webseiten und Blogs muss der Begriff Weblog eindeutig definiert werden: Nach SCHMIDT (2006), S. 13 sind Weblogs „regelmäßig aktualisierte Webseiten, die bestimmte Inhalte (...) in umgekehrt chronologischer Reihenfolge darstellen. Die Beiträge sind einzeln über URLs adressierbar und bieten in der Regel die Möglichkeit, Kommentare zu hinterlassen. Dadurch sowie durch Verweise auf andere Weblogs (...) bilden sich Netzwerke von untereinander verbundenen Texten und Webseiten heraus; die Gesamtheit aller Weblogs wird auch als ‚Blogosphäre‘ bezeichnet.“ Darüber hinaus weist ein Blog weitere, teils optionale Eigenschaften auf (FISCHER (2007), S. 43ff.). So besteht jeder Eintrag aus einem Titel, dem Veröffentlichungsdatum und dem Text. Die Leser können zu jedem Eintrag Kommentare hinterlassen. Des Weiteren kann der Autor eines Eintrags angegeben werden, und jeder Eintrag kann einer oder mehreren Kategorien zugeordnet sein. Die meisten Blogs stellen einen RSS[15]-Feed zur Verfügung, der den Lesern automatisch mitteilt, wenn ein neuer Eintrag im Blog er-

[15] RSS: Really Simple Syndication (RSS ADVISORY BOARD (2009)). RSS-Feeds basieren auf dem Dateiformat XML und sind in ihrem Aufbau sehr einfach gehalten (FISCHER (2007), S. 46).

stellt wurde. In der sogenannten Blogroll sind Verknüpfungen zu anderen Blogs gesammelt, die der Blog-Autor regelmäßig besucht.

4 Entwicklung eines Vorgehensmodells für die Analyse von Weblogs

Eine Analyse von etablierten Standardprozessen[16] für Data Mining hat gezeigt, dass es sinnvoll ist, als Basis für das Vorgehen bei der Blog-Analyse den KDD-Prozess nach FAYYAD (1996) heranzuziehen. Im Gegensatz zum CRISP-DM, der die Aufgaben aus der Sicht des Managements betrachtet, liegt der Fokus des KDD-Prozesses auf der technischen Sicht. Dies zeigt sich darin, dass die Phasen Datenselektion und -aufbereitung explizit herausgestellt werden (siehe Abbildung 1). Gerade diese Phasen sind bei der Blog-Analyse als kritische Punkte anzusehen, da als Datenquelle lediglich Webseiten im World Wide Web, das bedeutet semi- oder unstrukturierte Daten, zur Verfügung stehen. Diese Daten kommen in der Analyse-Phase zunächst nicht für die klassischen Data-Mining-Methoden in Frage. Stattdessen werden vermehrt verwandte Techniken wie Text Mining, aber auch einfache statistische Analysen eingesetzt.

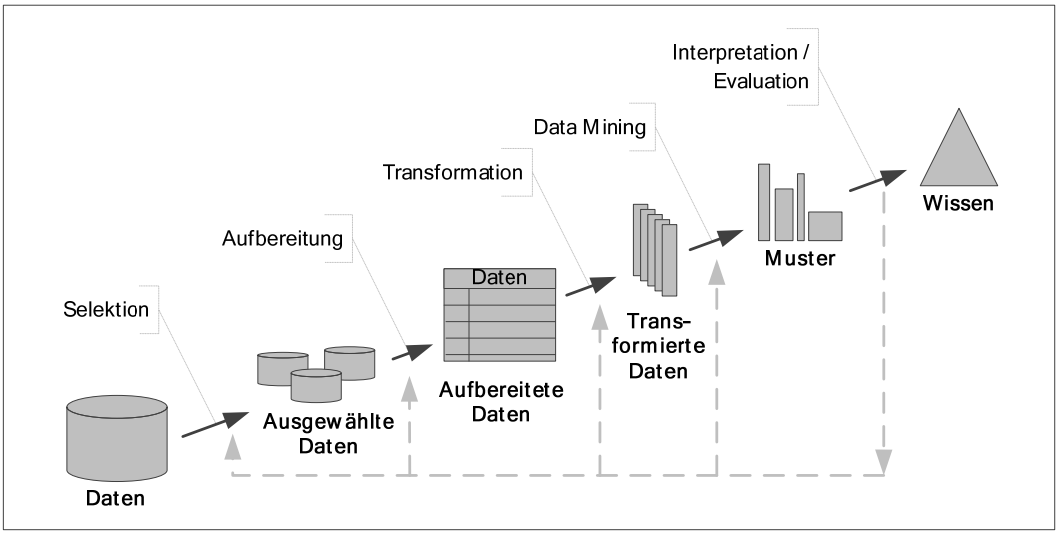

Abbildung 1: Der KDD-Prozess nach FAYYAD (1996), S. 6

Im Anschluss werden die einzelnen Phasen beschrieben und an die speziellen Anforderungen im Rahmen des Blog Mining angepasst.

[16] KDD nach FAYYAD (1996) und CRISP-DM nach CHAPMAN ET AL. (2000)

4.1 Phase 1: Selektion von Weblogs und Einträgen

In der Selektionsphase werden die Daten ausgewählt, die später analysiert werden sollen. Im Gegensatz zu traditionellen Data-Mining-Projekten liegen diese Daten nicht in strukturierter Form vor. Die relevanten Weblogs müssen erst identifiziert werden, bevor einzelne Einträge dieser Weblogs gespeichert werden können.

Identifikation von Weblogs im WWW

Die Herausforderung bei der Identifikation von Weblogs ist, sie von gewöhnlichen Webseiten zu separieren. Dazu müssen spezifische Kriterien herangezogen werden, die eine Differenzierung zwischen Weblogs und Webseiten aufgrund des Aufbaus bzw. des Inhalts erlauben.

Eine Möglichkeit bei der Identifikation von Weblogs ist die Verwendung eines Crawlers. Ein Crawler durchsucht eine Webseite nach Verknüpfungen zu anderen Webseiten und speichert die gefundenen Links. Nachdem er eine Seite vollständig untersucht hat, wird die nächste URL aus dem Speicher abgerufen und durchsucht (LIU (2007), S. 273f.). In der Regel erhält der Crawler eine Liste mit URLs[17], die er der Reihe nach abarbeitet (LIU (2007), S. 274). Damit der Crawler nicht auch URLs von gewöhnlichen Webseiten sammelt, muss er anhand von besonderen Kriterien bestimmen können, ob die Seite ein Weblog ist oder nicht. In der Literatur beschäftigen sich daher einige Autoren mit den speziellen Anforderungen, die ein Crawler für die Suche nach Weblogs erfüllen muss (HURST & MAYKOV (2009), S. 1615; SRIPHAEW ET AL. (2008); YU ET AL. (2008), S. 213).

Eine weitere Option zur Identifikation ist die Nutzung von sogenannten Ping-Diensten. Ein Ping-Dienst ist ein Online-Verzeichnis, das zuletzt erschienene Blog-Einträge auflistet. Sobald der Autor einen neuen Eintrag im Blog speichert, wird ein Signal, der Ping[18], an das Online-Verzeichnis gesendet. Dieses nimmt den neuen Eintrag anschließend auf seiner Webseite auf. Da eine Auflistung eines Ping-Dienstes stets neue Einträge enthält, besteht der Vorteil dieser Methode darin, dass die gefundenen Blogs aktiv sind. Kritisch anzumerken ist jedoch, dass Ping-Dienste zunehmend dazu missbraucht werden, Spam-Blogs oder andere Webseiten bekannt zu machen, die keine Weblogs sind (HURST & MAYKOV (2009), S. 1617). Solche Webseiten müssen vor der Analyse z.B. mit regel- bzw. inhaltsbasierten Verfahren herausgefiltert werden (SRIPHAEW ET AL. (2008), S.

[17] URL: Uniform Resource Locator; eine Webseite wird über eine URL aufgerufen, z.B. http://www.tu-dresden.de (http://tools.ietf.org/html/rfc3986, Abruf am: 30.09.2009)

[18] SAUER (2007), S. 35 bezeichnet einen Ping als „eine lautmalerische Bezeichnung für das Senden eines Echosignals im Web.“

1617). Diese Filterung ist auch bei der Nutzung eines Crawlers erforderlich, um auszuschließen, dass Spam-Blogs analysiert werden.

Darüber hinaus existieren viele Blogs, in denen keine weiteren Einträge mehr veröffentlicht werden (ECK & PLEIL (2006), S. 91). In manchen Situationen ist es aber wichtig, nur aktive Weblogs zu identifizieren. KRAMER & RODDEN (2007) berechnen dazu für jeden Blog eine individuelle Zeitspanne, innerhalb der ein Eintrag erscheinen muss, damit der Blog als aktiv eingestuft wird.

In der Literatur finden sich Belege, dass Weblogs mit beiden Methoden identifiziert werden können. Beispielsweise haben GLANCE ET AL. (2004), S. 3 bei der Entwicklung des Blog-Analyse-Systems BlogPulse[19] mehr als 100.000 Weblogs mit einem Crawler gefunden. In einer weiteren Arbeit beschreiben GLANCE ET AL. (2005), S. 422, dass sie für ihren Datenstamm die Listen mehrerer Ping-Dienste zusammengelegt und so „about 300.000 updated weblogs per day“ GLANCE ET AL. (2005), S. 422 gefunden haben. BANSAL & KOUDAS (2007), S. 1412 beschreiben ein Verfahren, das beide Methoden kombiniert: Ein Crawler erhielt dabei eine Liste eines Ping-Dienstes mit aktiven Blogs und hat darüber weitere Blogs gefunden.

Identifikation von Einträgen

Nach der Identifikation von Weblogs müssen die Einträge und Kommentare in den gefundenen Webseiten ebenfalls identifiziert und anschließend extrahiert werden. Hierbei besteht zunächst die Herausforderung darin, den Text des Eintrags auf der Webseite zu lokalisieren.

Da Weblogs im World Wide Web als HTML-Code gespeichert sind, lässt sich der Seitenquelltext des Blogs untersuchen, um die Einträge zu finden. Jeder Eintrag enthält per definitionem einen Titel, das Veröffentlichungsdatum und den Text; außerdem sind die Einträge eines Weblogs chronologisch umgekehrt sortiert. Daher ist der Quellcode der Webseite auf Elemente zu untersuchen, die in dieses Schema passen. In diesem Zusammenhang kann das Document Object Model[20] (DOM) der Webseite zu Hilfe genommen werden. Dabei werden die HTML-Elemente der Webseite als Baumstruktur dargestellt; am Ende der jeweiligen Äste finden sich Textstellen wie Einträge, Titel, Datumsangaben und Kommentare, die auf diese Weise identifiziert werden können (siehe Abbildung 2).

[19] http://www.blogpulse.com/

[20] Das Document Object Model ist eine Programmierschnittstelle, mit deren Hilfe auf HTML- oder XML-Dokumente zugegriffen werden kann; online unter http://www.w3.org/DOM/.

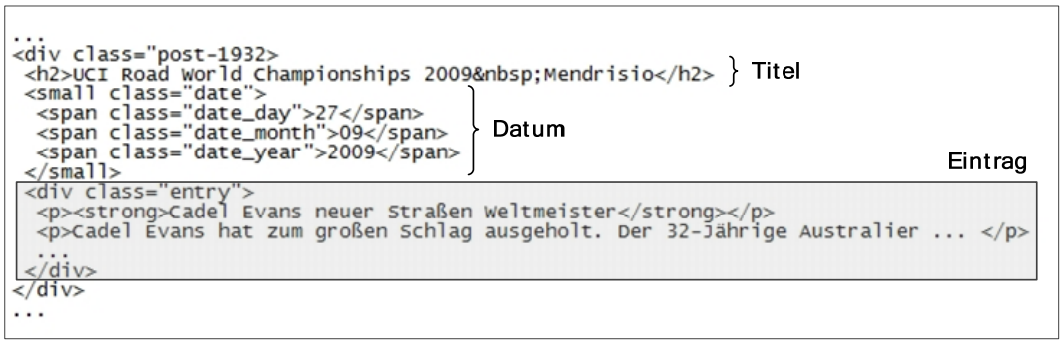

Abbildung 2: Quellcode eines Blog-Eintrags (in Anlehnung an Kirchgessner (2009))

Dieses Vorgehen kann zusätzlich unterstützt werden, wenn die Auszeichnungsregeln der Weblog-Programme berücksichtigt werden. Viele Blogs werden mit Hilfe von speziellen Programmen implementiert, die den HTML-Code für die Webseite des Blogs selbständig generieren. Jedes dieser Programme hinterlässt dabei gewisse Spuren im HTML-Code. Beispielsweise nutzt die Software WordPress das Quellcode-Element `<div class=“post“>…</div>`, um den Text des Blog-Eintrages darzustellen (ATTARDI & SIMI (2006), S. 3; siehe auch Abbildung 2). Durch diese Besonderheiten im HTML-Code können die Einträge und Kommentare auf der Webseite schneller identifiziert und gespeichert werden.

Viele Weblogs bieten ihren Lesern neben der Webseite des Blogs einen RSS-Feed an, der sie mit Informationen zu aktuellen Einträgen versorgt (siehe Abbildung 3). Diese Feeds können ebenfalls zur Identifikation von Einträgen und Kommentaren verwendet werden. Meistens beinhalten die Feeds alle wichtigen Informationen eines neuen Eintrags: Neben dem Titel und dem Veröffentlichungsdatum können im Feed auch der Name des Autors, die zugewiesenen Kategorien sowie der Inhalt des Eintrags selbst enthalten sein (RSS ADVISORY BOARD (2009)). Die Feeds liegen aufgrund des XML-Formats in maschinenlesbarer Form vor, so dass die übermittelten Informationen direkt weiterverarbeitet werden können (FISCHER (2007), S. 45). Voraussetzung dafür ist jedoch, dass die Feeds den gesamten Text des Blog-Eintrags beinhalten.

Die Literatur zeigt, dass die Analyse des HTML-Codes genutzt wird, um die Einträge aus dem Weblog zu extrahieren. CAO ET AL. (2008) beschreiben beispielsweise einen zweistufigen Prozess, der mit Hilfe des DOM Einträge und Kommentare voneinander trennt. Dabei wird zunächst der „main text“ (CAO ET AL. (2008), S. 2) isoliert. Der main text beinhaltet sowohl den Artikel des Blog-Autors als auch die Reaktionen darauf in Form der Kommentare. Wenn im DOM der Ast mit dem main text gefunden wurde, müssen im zweiten Schritt der Eintrag und die Kommentare voneinander getrennt werden. Dazu untersuchen die Autoren den identifizierten Ast weiter, indem sie die unterschiedliche

Verteilung von HTML-Code-Elementen im Eintrag und in den Kommentaren berechnen und vergleichen. Da ein Eintrag hauptsächlich Text beinhaltet, sind hier weniger Code-Elemente zu erwarten als bei Kommentaren, bei denen mindestens der Name, das Datum und der Kommentartext durch HTML-Code abgetrennt werden müssen.

GLANCE ET AL. (2005) nutzen ebenfalls die Quellcode-Informationen zur Identifikation der Einträge, haben ihr System aber nach anfänglichen Problemen erweitert: Wenn vorhanden, sollen zuerst RSS-Feeds verarbeitet werden, bevor der Quellcode analysiert wird.

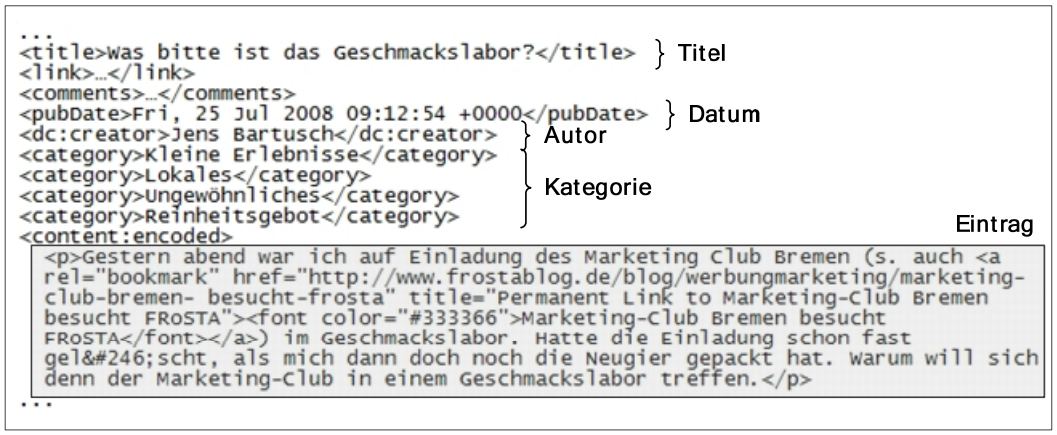

Abbildung 3: Quellcode eines RSS-Feeds mit HTML-Code-Elementen (in Anlehnung an FROSTA (2008))

Experimente der Autoren zeigen, dass die Extraktion der Einträge mit Hilfe der RSS-Feeds erwartungsgemäß schneller und besser möglich ist als mit der Quellcode-Analyse (CAO ET AL. (2008), S. 8; GLANCE ET AL. (2005), S. 423). Durch die überwiegend einheitliche Struktur der RSS-Feeds können die benötigten Informationen zudem mit überschaubarem Aufwand extrahiert werden. Dabei muss der Feed jedoch den gesamten Inhalt des Eintrags übermitteln. RSS-Feeds werden zwar von der Mehrheit der Blogs angeboten, jedoch nicht von allen. Daher ist ein zweistufiges System empfehlenswert, das zunächst versucht, den RSS-Feed zu analysieren und im Zweifel eine Extraktion direkt von der Webseite des Blogs vornehmen kann.

4.2 Phase 2: Aufbereitung der Daten

Nach der Identifikation der Weblogs und der Extraktion ihrer Einträge werden die gesammelten Informationen in einer Datenbank gespeichert; dadurch ist gewährleistet, dass die Informationen aus Blogs dauerhaft und performant untersucht werden können. Dafür muss vor der Speicherung eine strukturierte und konsistente Datenbasis geschaffen werden.

Dabei ist im ersten Schritt der HTML-Code aufzubereiten. Hierbei ist besonders darauf zu achten, dass Code-Fragmente – etwa für Umlaute oder Sonderzeichen – aus dem Quelltext entfernt werden, um spätere Analyseergebnisse nicht zu verfälschen. Im Ergebnis liegen der Titel des Eintrags, das Datum und der Text sowie optionale Angaben wie der Name des Autors oder die Bezeichnung der Kategorie separat als reiner Text vor.

Im zweiten Schritt werden die Daten vereinheitlicht. Das Ziel ist dabei, Formate und Texte in eine einheitliche Form zu bringen. Dazu zählen z.B. die Datumsangaben, die Schreibweise der Autorennamen und die Bezeichnung der Kategorie.

An dieser Stelle soll darauf hingewiesen werden, dass im Zuge der Aufbereitung auch linguistische Methoden zum Einsatz kommen, welche die Texte analysieren. Für weiterführende Darstellungen sei z.B. auf die Werke von LIU (2007) sowie FELDMAN & SANGER (2007) verwiesen.

4.3 Phase 3: Transformation der gespeicherten Daten

Nach der Aufbereitung steht eine strukturierte Datenbasis für die Analysen zur Verfügung. Je nach Analysemethode kann eine weitere Transformation der Daten erforderlich sein. So können aus den bestehenden Attributen neue Variablen gebildet werden (z.B. ein Attribut, das die Schlüsselwörter eines Eintrags enthält) bzw. Fehlwerte oder Ausreißer eliminiert werden. Diese verfahrensabhängige Transformation wird in der dritten Phase des Prozesses durchgeführt und unterscheidet sich im Bezug auf die Aufgaben nicht von der Phase des KDD-Prozesses.

4.4 Phase 4: Analyse der Weblogs

Für die Analyse von Weblogs stehen unterschiedliche Methoden zur Verfügung. Diese Methoden lassen sich in zeitraumbezogene Längsschnitt- und zeitpunktbezogene Querschnittanalysen unterscheiden (KUSS 2007, S. 41ff.).

Im Rahmen der Längsschnittanalyse kommen Verfahren in Betracht, die Trends in der Blogosphäre aufdecken. Dabei werden die Blog-Einträge mit Methoden des Text Mining analysiert und häufig vorkommende Wörter daraus extrahiert. So sind Themengebiete erkennbar, welche die Aufmerksamkeit der Blog-Autoren auf sich ziehen. Die Analyse-Systeme BlogScope[21] und BlogPulse stellen dem Nutzer z.B. anhand von Diagrammen vieldiskutierte Themen im Verlauf der Zeit dar (BANSAL & KOUDAS (2007), S. 1411; GLANCE ET AL. (2004), S. 4f.).

[21] http://www.blogscope.net/

Im Gegensatz dazu beschreiben Querschnittanalysen die Ereignisse in der Blogosphäre zu einem bestimmten Zeitpunkt. Dazu gehören die Verfahren des Opinion Mining und der sozialen Netzwerkanalyse. Mit Hilfe des Opinion Mining können die Meinungen der Blogger z.B. zu Produkten aufgedeckt werden. Dabei das Produkt und seine Eigenschaften in den Einträgen gesucht und mit der dazu geäußerten Meinung des Autors verbunden. Diese Bewertungen können als Marktforschungsergebnis zur Weiterentwicklung eines Produkts verwendet werden. In diesem Rahmen haben GLANCE ET AL. (2005), S. 419 Blog-Einträge analysiert und konnten so die Meinungen zu PDAs, Pocket PCs und Smartphones verschiedener Hersteller auswerten.

Die soziale Netzwerkanalyse betrachtet die Blogosphäre im Ganzen. Dabei ist das Ziel, eine Segmentierung der Blogosphäre zu erhalten, indem Verknüpfungen zwischen den Blogs aufgedeckt und z.B. als Netz dargestellt werden. Durch die Analyse dieser Blog-Communities lassen sich einflussreiche Blog-Autoren innerhalb der Gemeinschaft identifizieren. Solche Analysen und deren Ergebnisse sind in der Literatur bereits beschrieben worden (AGARWAL ET AL. (2008); ZHOU & DAVIS (2006)). GLANCE ET AL. (2005), S. 420ff. zeigen mit Hilfe einer Fallstudie eine Kombination von Opinion Mining und sozialer Netzwerkanalyse: Ein Netz von Einträgen stellt eine Diskussion über einen PDA der Firma Dell dar. Dabei sind drei Segmente von Einträgen zu sehen, die jeweils unterschiedliche Eigenschaften des Geräts diskutieren. Ein „Drill-down“ in eines der Segmente offenbart dabei die schlechte Audioqualität des PDAs (GLANCE ET AL. (2005), S. 421).

4.5 Phase 5: Interpretation / Evaluation der Resultate

In der letzten Phase des angepassten KDD-Prozesses werden die Analyseresultate interpretiert und hinsichtlich ihrer Eignung evaluiert. Für den Fall, dass die Ergebnisse unzureichend sind, müssen die vorgelagerten Schritte wiederholt werden. Nach positiver Evaluation gehen die Ergebnisse in Wissen über und können so z.B. als Entscheidungsgrundlage für Marketing-Aktivitäten dienen.

4.6 Zusammenfassung in einem Vorgehensmodell für das Blog Mining

Der für die Blog-Analyse angepasste Prozess teilt sich wie der KDD-Prozess in mehrere, aufeinander folgende Phasen auf (siehe Abbildung 4).

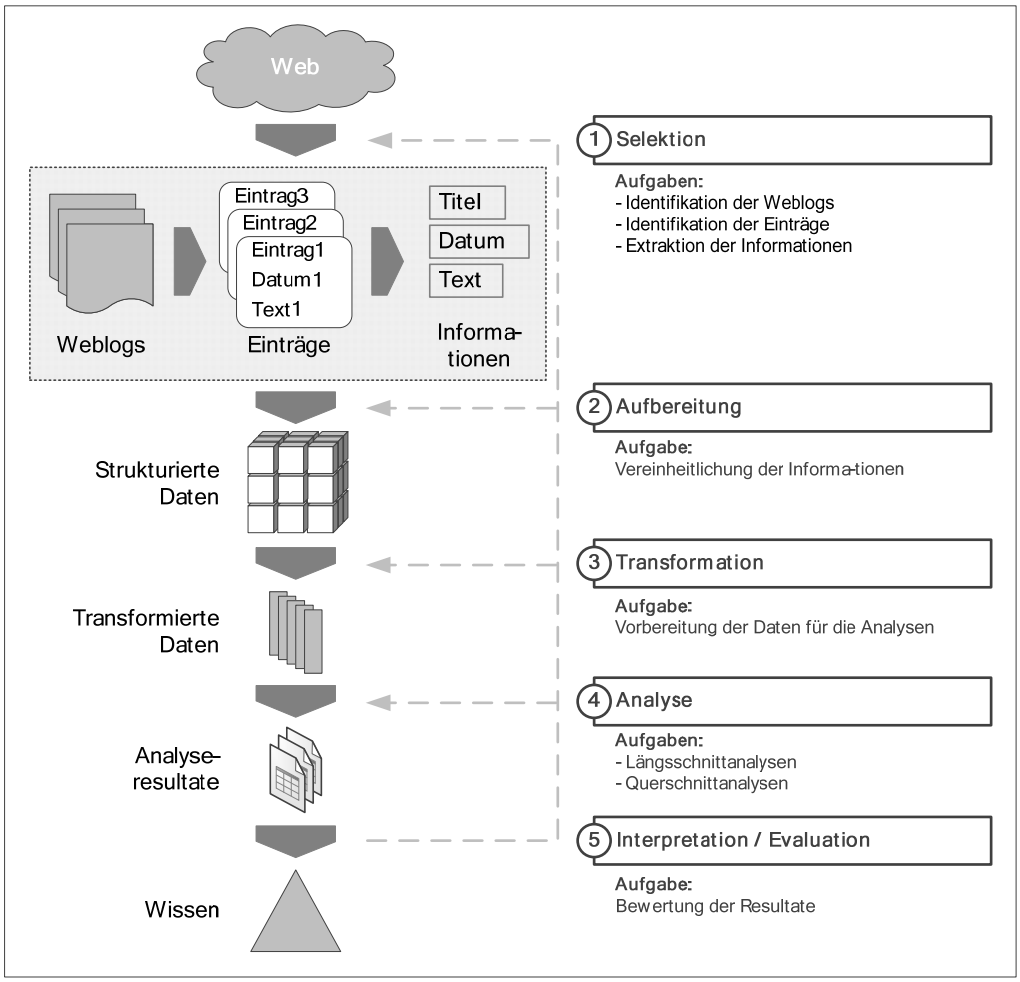

Abbildung 4: Das Vorgehensmodell für das Blog Mining
(in Anlehnung an FAYYAD (1996), S. 6)

Das Vorgehensmodell zeigt die Abfolge der fünf Phasen zur Durchführung eines Blog-Mining-Projektes. Sie beginnen mit der Datenerhebung und enden mit der Interpretation der Resultate. Im ersten Schritt werden im World Wide Web die Weblogs und Einträge identifiziert sowie extrahiert, die für die Untersuchung relevant sind. Diese Daten sind anschließend aufzubereiten, so dass sie in einer einheitlichen Form vorliegen. In der nächsten Phase findet – teils abhängig von den eingesetzten Analysemethoden – eine Vorverarbeitung der Blog-Einträge statt. Im Anschluss können die Informationen mit Hilfe unterschiedlicher Methoden ausgewertet werden. Sind nach der Evaluation und Interpretation der erzielten Resultate weitere Analysen nötig, müssen je nach Situation die Daten neu ausgewählt, aufbereitet oder transformiert werden. Bei erfolgreicher Evaluation gehen die Ergebnisse in Wissen über.

5 Fazit und Ausblick

Diese Arbeit beschreibt ein aktuelles Forschungsvorhaben, bei dem neues Wissen durch die Analyse der Blogosphäre gewonnen werden soll. Der Fokus liegt auf den Methoden der Business Intelligence, welche die dabei anfallenden Aufgaben unterstützen. In diesem Zusammenhang wurde das Vorgehen bei der Sammlung und Analyse von Daten im Rahmen des Blog Mining dargestellt. Dabei wurde eine Adaption des KDD-Prozesses für traditionelle Data-Mining-Projekte auf die besonderen Anforderungen bei der Analyse von Weblogs vorgenommen. Vor allem in den Phasen Selektion, Aufbereitung und Analyse (bzw. Modelling) sind Anpassungen erforderlich, um die spezifischen Aufgaben beim Blog Mining zu integrieren. Im weiteren Verlauf wurden Methoden der Datenerhebung, -speicherung und -analyse vorgestellt und deren erfolgreiche Anwendung anhand von Literaturbelegen nachgewiesen.

Es besteht jedoch weiterer Forschungsbedarf bei den jeweils dargestellten Schritten. Vor allem bei der Datenselektion müssen potente Alternativen zur Identifizierung von Weblogs geschaffen werden. Auch bei der Speicherung der Einträge ist weitere Forschungsarbeit nötig. Um die Daten in geeigneter Form für die Auswertungen bereithalten zu können, müssen dafür analytische Datenmodelle entwickelt werden.

Generell dürfen für das Forschungsvorhaben neue – blog-ähnliche – Technologien in der schnelllebigen Welt des World Wide Web nicht übersehen sehen werden. Dies bezieht sich vor allem auf sogenannte „micro-blogging“-Dienste wie Twitter (HONEYCUTT & HERRING (2009), S. 1). Twitter erlaubt dem Nutzer, kurze Nachrichten bis 140 Zeichen zu senden, die auf der Twitter-Webseite[22] aufgelistet werden. Damit sind diese Nachrichten – wie bei einem Weblog auch – öffentlich einsehbar. Aufgrund des hohen Nachrichtenaufkommens und aufgrund der Verwandtschaft mit Weblogs hat die Blog-Suchmaschine Technorati inzwischen eine Webseite (Twittorati[23]) entwickelt, mit deren Hilfe Twitter-Nachrichten gesucht werden können (TECHNORATI (2009)).

6 Literatur

Agarwal, N.; Liu, H.; Tang, L.; Yu, P. (2008): Identifying the influential bloggers in a community, in: ACM (Hrsg.): Proceedings of the International Conference on Web Search and Web Data Mining, ACM, New York, New York, S. 207-218.

Attardi, G.; Simi, M. (2006): Blog Mining through Opinionated Words, URL: http://trec.nist.gov/pubs/trec15/papers/upisa.blog.final.pdf, Abruf am: 10.09.2009.

[22] www.twitter.com

[23] www.twittorati.com

Bansal, N.; Koudas, N. (2007): BlogScope - A System for Online Analysis of High Volume Text Streams, in: Koch, C.; Gehrke, J.; Garofalakis, M.; Srivastava, D.; Aberer, K.; Deshpande, A.; Florescu, D.; Chan, C.; Ganti, V.; Kanne, C.-C.; Klas, W.; Neuhold, E. (Hrsg.): Proceedings of the 33rd International Conference on Very Large Data Bases, ACM, New York, New York, S. 1410-1413.

Cao, D.; Liao, X.; Xu, H.; Bai, S. (2008): Blog Post and Comment Extraction Using Information Quantity of Web Format, Chinese Academy of Sciences, Institute of Computing Technology, Beijing, China.

Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R. (2000): CRISP-DM 1.0 - Step-by-step data mining guide, URL: http://www.crisp-dm.org/CRISPWP-0800.pdf, Abruf am: 15.09.2009.

Davis, H.; Oberholtzer, M. (2008): What are they saying about us?, URL: http://greenfield-ciaosurveys.com/assets/pdfs/Davis_0108-Blogmining.pdf, Abruf am: 06.08.2009.

Eck, K.; Pleil, T. (2006): Public Relations beginnen im vormedialen Raum, in: Picot, A.; Fischer, T. (Hrsg.): Weblogs professionell, 1. Auflage, dpunkt-Verlag, Heidelberg, S. 77-94.

Fayyad, U. (1996): Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, California.

Feldman, R.; Sanger, J. (2007): The text mining handbook, Cambridge Univ. Press, Cambridge.

Fischer, E. (2007): Weblog & Co, 1. Auflage, VDM Müller, Saarbrücken.

Frosta (2008): RSS-Feed zum Blog, URL: http://www.frostablog.de/blog/feed, Abruf am: 25.07.2008.

Glance, N.; Hurst, M.; Nigam, K.; Siegler, M.; Stockton, R.; Tomokiyo, T. (2005): Deriving Marketing Intelligence from Online Discussion, Intelliseek Applied Research Center, Pittsburgh, Pennsylvania.

Glance, N.; Hurst, M.; Tomokiyo, T. (2004): BlogPulse - Automated Trend Discovery for Weblogs, Intelliseek Applied Research Center, Pittsburgh, Pennsylvania.

Honeycutt, C.; Herring, S. (2009): Beyond Micro-Blogging: Conversation and Collaboration via Twitter, Abruf am: 20.07.2009.

Hurst, M.; Maykov, A. (2009): Social Streams Blog Crawler, in: IEEE (Hrsg.): Proceedings of the 25th International Conference on Data Engineering, IEEE Press, S. 1615-1618.

Kaiser, C. (2009): Analyse von Meinungen in sozialen Netzwerken des Web 2.0, in: Hansen, H.; Karagiannis, D.; Fill, H.-G. (Hrsg.): Business Services: Konzepte, Technologien, Anwendungen, 1. Auflage, Österreichische Computer Gesellschaft, Wien, S. 379-388.

Kirchgessner (2009): UCI Road World Championships 2009 Mendrisio, URL: http://kirchgessner.wordpress.com/2009/09/27/uci-road-world-championship/, Abruf am: 04.10.2009.

Kramer, A.; Rodden, K. (2007): Applying a User-CenteredMetric to Identify Active Blogs, in: ACM (Hrsg.): Conference on Human Factors in Computing Systems, ACM, New York, New York, S. 2525-2530.

Liu, B. (2007): Web Data Mining, 1. Auflage, Springer-Verlag, Berlin, Heidelberg.

RSS Advisory Board (2009): RSS 2.0 Specification (version 2.0.11), URL: http://www.rssboard.org/rss-specification, Abruf am: 29.09.2009.

Sauer, M. (2007): Weblogs, Podcasting & Online-Journalismus, 1. Auflage, O'Reilly, Köln.

Schmidt, J. (2006): Weblogs, 1. Auflage, UVK Verlagsgesellschaft, Konstanz.

Sriphaew, K.; Takamura, H.; Okumura, M. (2008): Cool Blog Identification Using Topic-Based Models, in: IEEE/WIC/ACM (Hrsg.): International Conference on Web Intelligence and Intelligent Agent Technology, IEEE Press, S. 402-406.

Technorati (2008): Technorati: About Us, URL: http://www.technorati.com/about/, Abruf am: 21.08.2008.

Technorati (2009): We've Launched Twittorati - Discover Where Blogs and Tweets Converge, URL: http://technorati.com/weblog/2009/07/512.html, Abruf am: 20.07.2009.

WordPress (2009): Stats « WordPress.com, URL: http://en.wordpress.com/stats/, Abruf am: 29.09.2009.

Wright, J. (2006): Blog Marketing als neuer Weg zum Kunden, 1. Auflage, Redline Wirtschaft, Heidelberg.

Yu, F.; Zheng, D.; Zhao, T.; Cheng, X. (2008): Structure and Content Based Blog Pages Identification, in: Proceedings of the 5th International Conference on Fuzzy Systems and Knowledge Discovery, Vol. 2, S. 213-217.

Zerfaß, A.; Bogosyan, J. (2007): Blogstudie 2007, Universität Leipzig, Institut für Kommunikations- und Medienwissenschaft, Leipzig.

Zerfaß, A.; Sanhu, S.; Young, P. (2007): EuroBlog 2007: European Perspectives on Social Software in Communication Management, URL: www.euroblog2007.org/euroblog2007-results.pdf, Abruf am: 23.08.2007.

Zhou, Y.; Davis, J. (2006): Community Discovery and Analysis in Blogspace, International World Wide Web Conference Committee, Edinburgh, Scotland.