FAKULTÄT FÜR !NFORMATIK

Analysemethoden für Weblogs im Rahmen der Entwicklung eines Blog-Suchmaschinen Prototyps

MAGISTERARBEIT

zur Erlangung des akademischen Grades

Magister der Sozial- und Wirtschaftswissenschaften

im Rahmen des Studiums

Informatikmanagement

eingereicht von

Bernd Jüptner
Matrikelnummer 0200642

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung:
Betreuer/Betreuerin: Ao.Univ.Prof. Dr. Mag. Dieter Merkl

Wien, 27.01.2009 _____ _____

(Unterschrift Verfasser/in) (Unterschrift Betreuer/in)

Eidesstaatliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen nicht benützt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Wien, 27. Jänner 2009

Bernd Jüptner

Danksagung

Zu Beginn möchte ich mich besonders bei meiner Familie bedanken, die mich in meiner Phase der wissenschaftlichen Auseinandersetzung mit einem spezifischen Thema voll und ganz unterstützt hat. Sie hat mir viele Stunden an zusätzlicher Arbeit abgenommen, damit ich meinen Fokus in höchstmöglichem Ausmaß auf die Umsetzung der Arbeit richten konnte. Die ständige Motivation durch meine Familie hat neben meiner eigenen entscheidend zum erfolgreichen Abschluss des Studiums beigetragen. In diesem Zusammenhang möchte ich auch ein großes Dankeschön an meine Freundin richten, die mir fortlaufend Mut zugesprochen hat und mir sehr viel Zuversicht in Hinblick auf die zeitgerechte Fertigstellung der Arbeit gegeben hat.

Ein weiteres Dankeschön möchte ich an dieser Stelle meinem verstorbenen Vater aussprechen. Auch er hat durch seine Anteilnahme und durch die Liebe zum Forschen, die er mir Zeit seines Lebens mit auf den Weg gegeben hat, mein Durchhaltevermögen sowie meinen Ehrgeiz bekräftigt. Die vorliegende Arbeit ist zu einem sehr großen Teil diesem mehr als nur liebenswürdigenden Menschen, der leider viel zu früh von uns gehen musste, gewidmet.

Bedanken möchte ich mich darüber hinaus auch bei meinem Betreuer Ao.Univ.Prof. Dr. Mag. Dieter Merkl, der mich einerseits auf das, in der Arbeit behandelte, Themenfeld aufmerksam gemacht hat und mir zweitens bei jeglichen Fragen hilfreich zur Seite gestanden ist.

Auf keinen Fall zu vergessen in dieser Danksagung ist mein bester Freund und ständiger Wegbegleiter Hermann Rauschmayr. Auch ihm möchte ich ein großes Dankeschön aussprechen, da er mich ebenfalls mit seiner fachlichen Expertise während des gesamten Prozesses begleitet hat.

Kurzfassung

Historisch gesehen reichen die Wurzeln von Weblogs bis in die Mitte der 90er Jahre zurück. Der Begriff „Weblog“ (Wortkreuzung aus den englischen Wörtern „World Wide Web“ sowie „Log“) wurde im Jahr 1997 durch Jørn Barger geprägt. Die ersten Weblogs tauchten in Form von Online-Tagebüchern im Internet auf und seit Ende der 90er Jahre zeichnete sich bereits ein rasantes Wachstum in der Anzahl an Weblogs (eine exakte Zahl zu definieren ist aufgrund von Messproblemen beziehungsweise Abgrenzungsproblemen relativ schwierig) ab. Eine Abschwächung dieses Wachstums ist in den kommenden Jahren nicht zu erwarten. Interessant zu beobachten in Zusammenhang mit Weblogs ist die altersmäßige und geschlechtliche Verteilung von Autoren innerhalb der Blogosphäre. Die selbstrekrutierende Umfrage „Wie ich blogge?!“ erbrachte im Oktober 2005 unter anderem das Ergebnis, dass unter den Weblog-Autoren die Altersgruppe zwischen 20 und 29 Jahren mit 41,8 Prozent die Mehrheit ausmacht, gefolgt von 24,3 Prozent bei der Altersgruppe zwischen 30 und 39 Jahren. Die Gruppe unter 20 Jahren macht 17,7 Prozent aus und nur 5,4 Prozent entfallen auf jene Blogger, die älter als 50 Jahre sind. Die verbleibenden 10,8 Prozent beziehen sich auf die Altersgruppe zwischen 40 und 49 Jahren. Darüber hinaus wurde im Rahmen dieser Studie festgestellt, dass der Männeranteil unter den Blog-Autoren bei 54,5 Prozent liegt, im Vergleich dazu der Frauenteil, der bei 45,6 Prozent liegt. Bei der Untersuchung von Weblogs gilt es weiters in Betracht zu ziehen, dass laut der 21. W3B-Studie von Fittkau & Maaß zwar drei Viertel der befragten Internet-Nutzer den Begriff „Weblog“ als vertraut ansehen, jedoch nur jeder Fünfte von ihnen das Blog Angebot zumindest gelegentlich nutzt. Aufgrund der vielen sozialen Netzwerke (beispielsweise zur Selbstdarstellung sowie zur Identitätskonstruktion der Weblog Autoren im Teenager Alter) die durch das Bloggen entstehen können und der Verschiebungen im gesellschaftlichen Öffentlichkeitsbereich beschäftigt sich die vorliegende Arbeit im ersten Teil besonders intensiv mit sogenannten Analysemethoden für Weblogs. Dem

Leser soll ein fundierter und weitreichender Überblick über die ausgewählten Analysemodelle gegeben werden. Darüber hinaus soll die Differenzierung und die Abgrenzung der verschiedenen, heutzutage existierenden Analyseansätze herausgearbeitet sowie transportiert werden. Im zweiten und gleichzeitig praktischen Teil der Arbeit wird ein Prototyp einer Weblog-Suchmaschine entwickelt. Diese Suchmaschine soll es dem User erleichtern gezielt nach einem bestimmten Thema zu suchen und eine qualifizierte Auswahl an Blogeinträgen zu dem gesuchten Thema anhand eines definierten Bewertungsschemas zu bekommen.

Aus diesen beiden Teilen der Arbeit wird ihre Forschungsfrage folgendermaßen formuliert: *In wie weit lässt sich ein Bewertungsschema als Folge eines Suchalgorithmus entwickeln, um Blogeinträge anhand eines Suchbegriffes in einer reliablen quantifizierten und vor allem qualifizierten Form im Rahmen bestehender Analysemethoden darzustellen?*

Abstract

Seen from a historical perspective the roots of weblogs reach down to the mid-nineties. The term „weblogs" (portmanteau of the english words „world wide web" and „logs") was coined by Jørn Barger in the year 1997. The first weblogs appeared in the form of online-diaries in the Internet and since the end of the nineties already a rapid growth in the quantity of weblogs (because of problems within the measurement or the delimitation it is difficult to define an exact figure) has been observable. In the next few years a decrease of this growth is not expected. In the context of weblogs the age-type and gender arrangement of authors within the blogosphere is really interesting to observe. The self-recruiting survey „How do I blog?!" adduced in October 2005 among other things that within weblog authors the age group between 20 and 20 years represents the majority with 41,8 percent, followed by 24,3 percent within the age group between 30 and 39 years. The group under 20 years minds 17,7 percent and only 5,4 percent fall to those blogger, who are older than 50 years. The remaining 10,8 percent refer to the age group between 40 and 49 years. In addition within the scope of this study it was observed that the part of the men among blog authors amounts to 54,5 percent, in comparison to the part of the women, which amounts to 45,6 percent. On further examination of weblogs it hast to be taken into account that according to the 21. W3B-study from Fittkau & Maaß indeed three quarters of the sampled internet users consider the term „weblogs" familiarly but only every fifth out of them uses the blog offer at least casually.

Due to the numerous social networks (in example for the self-expression or construction of the identity of teenager weblog authors) which could grow out of blogging and the displacements in the social public area the current thesis deals with analysis methods in the first part extraordinarily intensive. The reader should be provided with a profound and far-reaching overview about the selected analysis methods. In addition to this the differentiation and the

distinction of various, nowadays existing analysis approaches should be worked out as well as be transported.
A prototype of a weblog search engine will be developed in the second and simultaneously practical part of the thesis. This search engine should make it easier for the user to search pointedly for a specific topic and to get a qualified sample of blog entries related to the sought topic on the basis of a defined benchmark scheme.

Out of these two parts of the thesis the main research question is formulated as follows: *How far is it possible to develop a benchmark scheme as a result of a search-algorithm to constitute blog entries on the basis of a search term in form of a reliable quantified and especially qualified form within analysis methods?*

Inhaltsverzeichnis

Kapitel 1

1 Einleitung

In diesem ersten einleitenden Kapitel wird der interessierte Leser auf das in der vorliegenden wissenschaftlichen Arbeit zu untersuchende Forschungsthema vorbereitet. Das Kapitel 1.1 beinhaltet wesentliche Hintergrundinformationen und klärt den Leser somit über die allgemeinen Beweggründe für den wissenschaftlichen Diskurs auf. Im darauffolgenden Kapitel 1.2 wird auf die eigentliche Problemstellung im Rahmen der Arbeit eingegangen sowie die Hauptforschungsfrage definiert. In Kapitel 1.3 stehen Überlegungen zur Zielsetzung der Forschungsarbeit im Mittelpunkt. Darüber hinaus wird der inhaltliche Gliederungsaufbau der Arbeit im abschließenden Kapitel 1.4 detailliert beschrieben.

1.1 Motivation

Das aufstrebende Phänomen des Social-Web sowie der breitgefächerte Einsatz von Techniken wie Wikis, RSS Feeds, Social Bookmarks, etc. sind für einen signifikanten Umschwung in der Web-Anwendung verantwortlich [Bere07b].
In den vergangenen fünf bis zehn Jahren hat sich das onlinebasierte Medienformat namens „Weblogs“ (Wortkreuzung aus den englischen Wörtern „World Wide Web“ sowie „Log“) weltweit verbreitet und hat auf vielfältige Art und Weise Einzug im World Wide Web gefunden. Die Bandbreite der heutzutage existierenden Blogs reicht von sogenannten Unternehmensblogs, die darauf ausgerichtet sind firmeninterne Themen einer breiten Öffentlichkeit zugänglich zu machen, bis hin zu persönlichen Weblogs, welche eine Selbstdarstellungsplattform für Individualisten unter den Internetnutzern

darstellen. Die enorme Informationsflut, die sich dadurch entwickelt, bietet dem „Otto-Normal-User“ einerseits die Möglichkeit sich intensiv mit einer sehr großen Anzahl an unterschiedlichen Themengebieten auseinander zu setzen, macht es jedoch andererseits unglaublich schwierig, Zusammenhänge zwischen den einzelnen Blogs zu erkennen und somit die Suche nach einem bestimmten Thema effizienter zu gestalten. Der Schwerpunkt dieser Arbeit liegt einerseits darin, bestehende Analysemethoden für Weblogs kategorisch zu untergliedern, und dem Leser oder der Leserin einen fundierten Überblick über diese zu verschaffen.

1.2 Problemstellung

Da sich die gezielte Suche nach bestimmten Weblog-Artikeln oftmals sehr schwierig gestaltet, wird aufbauend auf diversere Parameter innerhalb der untersuchten Analysemethoden ein Prototyp einer derartigen Weblog-Suchmaschine realisiert. Anders als bei vielen populären Suchmaschinen sollen individuelle „Benutzerwünsche“ integriert werden. Der Prototyp soll es dem User ermöglichen eine qualifizierte Auswahl an Blogeinträgen zu dem gesuchten Thema zu bekommen. Die wissenschaftliche Forschungsfrage wird demzufolge folgendermaßen formuliert: *In wie weit lässt sich ein Bewertungsschema in Form eines Suchalgorithmus entwickeln, um Blogeinträge anhand eines Suchbegriffes in einer reliablen quantifizierten und vor allem qualifizierten Form im Rahmen bestehender Analysemethoden darzustellen?*

1.3 Zielsetzung der vorliegenden Arbeit

Grundsätzlich können in dieser Arbeit zwei verschiedene „Zielsetzungen" unterschieden werden: Der erste Teil beschäftigt sich neben der historischen Entwicklung von Blogs vor allem mit den vorhin erwähnten Analysemethoden für Weblogs. Verschiedene methodische Zugänge beziehungsweise Ansätze zu Blog Einträgen sollen genau untersucht werden. Diese Zielsetzung könnte man auch als „theoretisches Ziel" bezeichnen.

Die zweite Hälfte der Arbeit hat die Entwicklung des Suchmaschinen-Prototyps als erklärtes Ziel. Hier könnte man von der „praktischen Zielsetzung" der wissenschaftlichen Arbeit sprechen. Die Suchmaschinen-Webseite sollte im Wesentlichen ein Eingabefeld beinhalten, in welches der User beliebige Suchbegriffe eingeben kann. Nachdem der User auf einen der beiden „Suchen" Button geklickt hat wird eine kleine Auswahl an Blogeinträgen, die die notwendige Entsprechung zum Suchbegriff finden, unterhalb der Ergebnisse angezeigt. Die Suche über den einen Such-Button bezieht sich ausschließlich auf jene Blogeinträge, welche ihre emotionelle und informative Bewertung mittels einer Klassifizierung durch Testpersonen beziehen, während bei der Suche über den zweiten Such-Button zwar dieselben Blogeinträge verwendet werden, die Bewertung des emotionalen beziehungsweise informativen Gehaltes allerdings aus einer Klassifizierung durch einen Maschinen-Lernalgorithmus stammt. Die Suchergebnisse sind nach einem bestimmten qualitativen sowie quantitativen Bewertungsschema sortiert. Weiters soll im Rahmen des Vorverarbeitungsprozesses der Blog-Datensätze eine Kategorien-Klassifizierung durch Testpersonen im Vergleich zu einer Klassifizierung derselben Datensätze mittels eines Maschinenlern-Klassifizierungsverfahren analysiert werden.

1.4 Inhaltlicher Aufbau

Die vorliegende Arbeit gliedert sich in sechs Hauptkapitel, welche im Folgenden vorgestellt werden:

In Kapitel zwei wird primär auf die historische Entwicklung von Weblogs eingegangen. Zu Beginn wird abermals die Begriffsdefinition von Weblogs erklärt sowie die Ursprünge des Bloggens aufgegriffen. Darüber hinaus sollen die rechtlichen Grundlagen sowie die verschiedenen Typen von Weblogs untersucht werden.

Kapitel drei stellt eines der wesentlichsten Kapitel dieser wissenschaftlichen Arbeit dar. Zunächst wird ein Überblick über häufig verwendete Analysemethoden in Zusammenhang mit Weblogs gegeben. Diverse Methoden werden in verschiedene Kategorien unterteilt und detailliert beschrieben. Die vorgestellten Methoden sind besonders relevant für Kapitel vier, da diese in die Entwicklung des Blog Suchmaschinen Prototyps eingebunden werden.

Kapitel vier behandelt den praktischen beziehungsweise experimentellen Teil der Arbeit. Ein Blog Suchmaschinen Prototyp, welcher es dem User ermöglicht einen Suchbegriff einzugeben und als Ergebnis eine quantifizierte sowie qualifizierte Auswahl an Blogeinträgen zu bekommen, soll realisiert werden. In diesem Kapitel geht es vorerst darum, ein Konzept für die Suchmaschine zu erarbeiten. Fragen über mögliche Programmiersprachen, über den Suchalgorithmus, über das Design, etc. sollen geklärt werden. Screenshots über die finale Suchmaschinen-Seite sollen zur besseren Visualisierung dienen. Weiters setzt sich dieses Kapitel mit der tatsächlichen Umsetzung (dem Source Code) der Suchmaschine auseinander. Wesentliche Abschnitte der beteiligten Projektdateien werden in kommentierter Form dargestellt.

Den Abschluss der Arbeit bildet Kapitel fünf. In diesem Kapitel werden die Ergebnisse und Erkenntnisse, die im Rahmen der Untersuchung gewonnen werden konnten, sowie die Einsetzbarkeit des entwickelten Suchmaschinen-Prototyps strukturiert dargestellt. Im Anschluss daran wird ein Ausblick auf zukünftig denkbare Arbeiten im Bereich Weblog Forschung sowie im Bereich der Weiterentwicklung des Suchmaschinen-Prototyps gegeben.

Kapitel 2

2 Allgemeines zu Weblogs

Dieses Kapitel beschreibt zu Beginn die historische Entwicklung von Weblogs. Die wichtigsten Meilensteine in der Geschichte von Weblogs werden chronologisch dargestellt. Darüber hinaus geht dieses Kapitel auch auf die rechtliche Situation innerhalb der Blogosphäre ein. Im Anschluss daran wird ein Überblick über die verschiedenen Typen von Weblogs gegeben.

2.1 Historische Entwicklung

Bevor auf die eigentliche Geschichte von Weblogs eingegangen wird, sollen noch einmal die wichtigsten Begriffe erklärt werden: Die Termini „Blogs“ und „Weblogs“ werden heutzutage sehr oft synonym verwendet. „Weblog“ leitet sich, wie bereits in der Einleitung erwähnt, von der Wortkreuzung aus „World Wide Web“ und „Log“ (im ursprünglichen Sinn eine Art Online Tagebuch im WWW) ab. Beide Begriffe bezeichnen jedoch Webseiten, deren Inhalte (zumeist Texte, Bilder oder multimediale Inhalte) in umgekehrt chronologischer Reihenfolge dargestellt werden [Schm06].
In Zusammenhang mit Weblogs wird oftmals von der sogenannten Blogosphäre (engl. blogosphere) gesprochen. Diese beschreibt die Gesamtheit der Weblogs beziehungsweise der Blogger. Mit dem Begriff „Blogger“ werden im Wesentlichen Autoren assoziiert, welche regelmäßig im Internet ihre Beiträge verfassen und sich zu bestimmten Themen äußern.
Bezugnehmend auf die historische Entwicklung von Weblogs lässt sich festhalten, dass diese bis in die Anfänge des World Wide Web

zurückzuverfolgen ist. Barger definiert Weblogs als eine „Web page where a Web logger logs all the other Web pages she finds interesting“ [Bloo04].
Frühere Vorläufer von Weblogs existierten schon in der ersten Hälfte der 90er Jahre. Tim Berners-Lee (Begründer des World Wide Web) sowie zahlreiche Organisationen betreuten regelmäßig aktualisierte Webseiten, welche als Informationsfilter für den rasch wachsenden Content im Internet dienten. Diese Webseiten enthielten auch zahlreiche Verweise auf andere Online-Quellen. Solche kommentierten Linklisten gelten ebenfalls als eine der Wurzeln des heutigen Blog Begriffes. In den Anfängen wurden die meisten Weblogs mit der textbasierten Auszeichnungssprache HTML realisiert. Während das Internet jedoch seinen gesellschaftlichen Höhepunkt durchlebte, wurde die Gestaltung und Veröffentlichung von Weblogs zunehmend automatisiert. Ein sogenanntes CMS wurde entwickelt, um den Blog-Betreibern die Verfassung von neuen Beiträgen oder die Bearbeitung von verschiedenen Elementen eines Blogs zu erleichtern. Von diesem Standpunkt aus gesehen, kann ein CMS ebenfalls als „Ursprung“ für das heutige Blog-System betrachtet werden.

2.2 Technische Alternativen

Im Wesentlichen können zwei verschiedene Weblog-Techniken unterschieden werden:

I. Die „stand-alone" Variante:
Diese Alternative zieht in Betracht, dass der User das Softwarepaket für den Weblog auf seinem eigenen Server installiert. Einige sehr bekannte Anwendungen dieser Art sind „MoveableType", „Blosxom" und „Wordpress". Es existieren jedoch zahlreiche weitere Systeme, welche sich in ihrem Funktionsumfang und in ihrer Verbreitung sehr stark von einander abgrenzen (einen detaillierten Überblick über die verschiedenen Weblog-Softwarepakte liefert folgende Webseite: http://unblogbar.com/software/) [Schm06].

II. Die Weblog Hosting Variante:
Aufgrund der Tatsache, dass die Wartung von Weblog-Skripten (werden in Variante I vorgefertigt erstellt) ein gewisses technisches Know-How voraussetzt, haben sich verschiedene externe Dienstleister dazu entschlossen, dass bei ihnen ein Weblog eingerichtet werden kann auch ohne entsprechenden eigenen Server. Der User registriert sich in der Regel bei einem der Dienste und kann innerhalb von wenigen Minuten über sein eigenes Weblog verfügen. Die meisten gehosteten Weblogs können kostenlos genutzt werden, wobei von vielen Anbietern Werbungen eingeblendet werden. Erweiterte Funktionalitäten stehen meistens gegen ein monatliches Entgelt zur Verfügung. Namhafte Anbieter dieser Alternative sind beispielsweise *blogger* (http://www.blogger.de), *livejournal* (http://www.livejournal.com), *blogigo* (http://www.blogigo.de), *blogbar* (http://blogbar.de/), *twoday* (http://twoday.net/) und viele weitere. Diese Weblogseiten unterscheiden sich in folgenden Punkten: Auf der Startseite von *livejournal* sowie *blogigo* hat man bereits die Möglichkeit ein bestimmtes Themengebiet (bei *livejournal* erfolgt eine Untergliederung in die Kategorien „Leben", „Unterhaltung", „Musik"

„Kultur“, „Nachrichten und Politik“ sowie „Technologie“ und bei *blogigo* kann man aus den Hauptkategorien „Computer“, „Lifestyle“, „Leben&Romantik“, „Gesellschaft“, „Reisen&Freizeit“ und „Sonstiges“ auswählen, wobei diesen Hauptkategorien immer mehrere Unterkategorien zugeordnet sind wie zum Beispiel „Computer“ die Untergruppen „Internet“, „Hardware“, „Software“, etc. beinhaltet) aufzurufen und gezielt in diesem relevante Blogeinträge durchzulesen. Im Gegensatz dazu wirken *blogbar und blogger* relativ unstrukturiert und eine derartige Aufgliederung in Themengebiete ist nicht erkennbar. Bei *twoday* ist ebenfalls auf der Startseite eine Auswahl an Themengebieten nicht ersichtlich. Bezüglich der Gesamtanzahl an eingetragenen Blogs lässt sich festhalten, dass diese bei *twoday* im Dezember 2005 auf 18000 angewachsen ist und einen Zuwachs von 50 bis 100 Weblogs pro Woche aufweisen konnte. Auf der Startseite von *blogger* findet sich ganz unten ein kleines Feld mit einer Statistik über deren Bloganzahl (diese betrug 23422 am 21.11.2008). Was die Zahl an registrierten Benutzern anbelangt zählte *twoday* an die 41000 im Jahre 2005. Der Statistikseite von *livejournal* (http://www.livejournal.com/stats.bml, 21.11.2008) ist zu entnehmen, dass deren Anzahl an Benutzern 17193485 beträgt. Bei *blogbar* und *blogigo* ist es leider nicht möglich über deren Weblogseite Informationen über die Anzahl an eingetragenen Beiträgen sowie registrierten Benutzern zu beziehen. Sehr interessant allerdings zu beobachten ist, dass bei *livejournal* (ebenfalls deren Statistikseite zu entnehmen) beispielsweise in etwa doppelt so viele weibliche User (66,5 Prozent) registriert sind als männliche (33,5 Prozent). Dagegen sieht die soziodemographische Situation bei *twoday* so aus, dass Autoren Männer (56,2 Prozent) und Frauen (43,8 Prozent) in etwa gleich verteilt sind. Darüber hinaus liegt die am stärksten besetzte Altersgruppe bei *twoday* zwischen 19 und 35 Jahren (diese Daten stammen aus zwei Befragungen: zum einen aus einer Umfrage von Knallgrau aus dem Jahre 2004 [Schm06] und zum anderen aus der Umfrage „Wie ich blogge?!“ [Schm06]).

2.3 Rechtliche Rahmenbedingungen

In Zusammenhang mit Weblogs taucht sehr oft die spannende Frage nach der rechtlichen Verantwortlichkeit der Blogger beziehungsweise der Blogs auf. Im Folgenden werden die wichtigsten allgemeinen rechtlichen Rahmenbedingungen bezugnehmend auf die formalen und inhaltlichen Anforderungen an Blogs dargestellt, wobei sich diese Darstellung der Rechtssituation im Bereich von Blogs und Podcasts auf die deutsche Rechtsordnung bezieht. In Deutschland werden die medienrechtlichen Schranken für Blogs beziehungsweise Podcasts grundsätzlich durch das Telemediengesetz (TMG) sowie durch den Staatsvertrag über Rundfunk und Telemedien (RStV) gebildet [Wolf07]. Kurioserweise kommen die beiden Begriffe „Blog“ und „Podcast“ weder im TMG noch im RStV vor. Vielmehr wird in diesen Gesetzestexten generell von sogenannten Telemedien gesprochen. Diese Tatsache bringt bereits deutlich zum Ausdruck, dass die Wahrnehmung der rechtlichen Aspekte von Blogs einen starken subjektiven Charakter hat. Grundsätzlich werden im RStV drei Arten von Telemedien unterschieden, wobei die dritte Kategorie die spannendste Form darstellt:

- Telemedien, welche ausschließlich persönlichen oder familiären Zwecken dienen (Anbringung eines Impressums ist nicht erforderlich)
- Telemedien, welche über den privaten Bereich hinausgehen (Anbringung eines einfachen Impressums ist erforderlich)
- Telemedien, die journalistisch-redaktionell aufgebaut sind

Diese dritte Art der Telemedien findet, wie bereits erwähnt, die größte Aufmerksamkeit unter den genannten Kategorien und stellt deshalb gesteigerte inhaltliche sowie formale medienbezogene Anforderungen dar. Neben dem TMG und dem RStV finden weitere Gesetzestexte wie das Urheberrecht, das Wettbewerbsrecht, das Kennzeichenrecht sowie das Zivil- Straf- und öffentliche Recht ihre Anwendung auf Blogs [Wolf07]. In den

weiteren Abschnitten dieses Kapitels wird besonders darauf eingegangen, welche Bilder in Blogs gezeigt werden dürfen, in welcher Form das Impressum gehalten werden muss, welche Texte in Blogs veröffentlicht werden dürfen, welche urheberrechtlichen/ datenschutzrechtlichen Vorschriften zum Tragen kommen und welche Kriterien bei Werbung innerhalb eines Blogs beachtet werden müssen.
Auf die wesentliche Frage welche Inhalte in Blogs geschrieben und anschließend veröffentlicht werden dürfen lässt sich folgende Antwort finden: Die Grundregel für bestimmte Inhalte eines Blogs besagt, dass Tatsachenbehauptungen (besitzen eine objektive Beziehung zur Realität) der Wahrheit entsprechen müssen und sogenannte Meinungsäußerungen (eine subjektive Stellungnahme die sich nicht als wahr oder unwahr erweisen lässt) vertretbar sein müssen [Wolf07]. Ein weiterer journalistischer Grundsatz besagt, dass Nachrichten auf deren Inhalt, deren Herkunft sowie deren presserechtlicher Wahrheit (besonders detailliert im Pressekodex formuliert) vom Anbieter geprüft werden müssen, bevor diese veröffentlicht werden.

Bei der Betreibung von Blogs muss besonderes Augenmerk auf das sogenannte Impressum gelegt werden. Weblogs, welche rein persönliche beziehungsweise familiäre Zwecke überschreiten und als journalistisch-redaktionelle Blogs geführt werden, unterliegen einer vielfältigen Informationspflicht. Diese Anbieterkennzeichnung, welche im Idealfall leicht erkennbar (Link sollte klar als „Impressum“ gekennzeichnet sein) und unmittelbar verfügbar (Impressums-Link ist auf jede Unterseite des Blogs verlinkt) ist, sollte folgende Angaben enthalten:

- Name und Anschrift des Weblog Betreibers
- e-mail Adresse zur elektronischen Kontaktaufnahme
- gegebenenfalls eine Umsatzsteueridentifikationsnummer
- Handelsregister, Vereinsregister, etc.

- Ernennung einer verantwortlichen Person (Verantwortlicher im Sinne des Presserechts)

Bezugnehmend auf das Urheberrecht und das Datenschutzrecht innerhalb der Blogosphäre kann festgehalten werden, dass auf Blogs veröffentlichte Inhalte häufig dem Urheberrecht unterliegen und dass bestimmte datenschutzrechtliche Vorschriften aufgrund der Verarbeitung personenbezogener Daten (Informationen die sich auf eine identifizierbare natürliche Person beziehen) eingehalten werden müssen. Texte genießen genauso wie beispielsweise Fotografien urheberrechtlichen Schutz. Aufgrund dessen sind Blog-Einträge sofern sie die erforderliche eigene geistige Schöpfung aufweisen als urheberrechtlich schützenswerte Werke der Literatur anzusehen. Bei Verwendung von fremden geschützten Inhalten sind die notwendigen Nutzungsrechte einzuholen.
Datenschutzrechtlich darf der Diensteanbieter personenbezogene Daten nur dann erheben und verwenden, wenn sowohl das TMG als auch eine andere Rechtsvorschrift dies erlauben beziehungsweise der Nutzer selbst einwilligt [Wolf07].

Immer mehr Weblogs gehen über die reine textuelle Form hinaus und verwenden sowohl Bilder als auch Videos in ihrer Gestaltung. Bei sogenannten „Moblogs“ (zusammengesetzt aus den englischen Wörtern „mobile“ und „weblog“) handelt es sich um Weblogs, welche überwiegend aus, mit Handy-Kameras aufgenommenen, Bildern aufgebaut sind und Text lediglich als dokumentarische Ergänzung dient. Sowohl bei „Moblogs“ als auch bei herkömmlichen Weblogs muss auf das Persönlichkeitsrecht der Abgebildeten Bedacht genommen werden. Im Urheberrecht ist klar festgehalten, dass stets die Einwilligung der Personen, die auf einem Foto abgebildet sind, einzuholen ist bevor dieses veröffentlicht wird. Ausnahmen dieser Regelung stellen jedoch folgende Personen beziehungsweise Sachverhalte dar:

- Besonders bekannte Personen von öffentlichem Interesse (beispielsweise Politiker)
- Personen, die an öffentlichen Veranstaltungen teilnehmen (Demonstrationen, Sportveranstaltungen, Konzerte, etc.)
- Personen, die nur als Beiwerk im Bild dienen (zum Beispiel Personen in einer Landschaftsdarstellung)

Da Medien sehr stark meinungsbildend und polarisierend wirken können, kristallisieren sich heutzutage viele Weblogs heraus, welche als umfangreiche Werbeplattformen dienen. Das RStV regelt jedoch sehr genau welche Richtlinien dabei eingehalten werden müssen. Eine wesentliche Reglementierung ist die, dass Werbung innerhalb eines Weblogs als solche klar erkennbar und vom restlichen Inhalt der Angebote eindeutig getrennt sein muss. Eine deutliche Kennzeichnung ist beispielsweise bei Werbeformen wie Mietlinks unbedingt erforderlich, da der Link an sich keinen Aufschluss darüber gibt, ob es sich um eine werbliche oder redaktionelle Empfehlung handelt [Wolf07].

2.4 Typisierung von Weblogs

Eine allumfassende beziehungsweise allgemein gültige Kategorisierung von Weblogs zu treffen gestaltet sich als beinahe unmöglich. Stattdessen existieren zahlreiche unabhängige Versuche einer möglichen Typisierung. Stefan Buchner, Web 2.0 Content Specialist aus Zürich, teilt Weblogs folgendermaßen ein [Buch08]:

- Erzähl-Weblog: Geschichten mit literarischem Hintergrund
- Fach-Weblog: Diskussion und Austausch von Fachthemen
- Moblog: Beiträge von einem mobilen Gerät und meistens Fotos mit der Handy Kamera
- Photoblog: Abbildung des Alltags in Fotografien
- Corporate/Business-Weblog: Mitarbeiter schreiben im Namen einer Firma zu PR Zwecken

Jan Schmidt, Kommunikationswissenschaftler an der Universität Bamberg, hingegen unterscheidet bei politisch orientierten Blogs zwischen parteipolitischen und zivilgesellschaftlichen [Medi08].
Ein weiterer Ansatz einer strukturierten Klassifizierung von Weblogs stammt von Peter Wolff. Er definiert ganz allgemein die Bereiche Unternehmensblogs (Corporate Blogs), CEO-Blogs, persönliche Blogs, „normale Ich-Blogs“ und politische Blogs [Wolf07]. Zu Unternehmensblogs lässt sich sagen, dass sich ein Unternehmen die grundsätzliche Frage stellt, welche Absichten es mit dem Weblog verfolgen möchte bevor dieser tatsächlich gegründet wird. Unternehmensinterne Abteilungen wie Marketing, Vertrieb, Produktentwicklung, Public-Relations und viele andere sollten in den Weblog-Prozess eingebunden werden. Innerhalb der Kategorie der Unternehmensblogs kann eine Unterteilung in Themenblogs, Kampagnenblogs, Markenblogs, Wissensblogs und einigen weiteren getroffen werden: Themenblogs sollten, wie der Name bereits vermuten lässt,

idealerweise interne Themen wie Material-Kenntnisse, Aufbau der Unternehmenskultur, Mitarbeiterführung, etc. einer breiten Öffentlichkeit zugänglich machen [Wolf07]. Die Abgrenzung des Kampagnenblogs von anderen Blogarten fällt in der Regel nicht sehr einfach aus und deshalb wird er auch als Community Blog bezeichnet. Besonders im politischen Umfeld – im Zuge diverser Wahlkampagnen – etablierte sich dieser Blog. Ein Markenblog verfolgt das vorrangige Ziel die Kunden beziehungsweise Mitarbeiter auf die Präsenz und Lebendigkeit der Unternehmensmarke aufmerksam zu machen [Wolf07]. Sowohl die Mitarbeiter als auch die Kunden sollten aktiv in die Markenkommunikation involviert werden. Sogenannte Wissens- und Serviceblogs geben den Kunden die Möglichkeit Erfahrungswerte über die Produktbesonderheiten oder die richtige Handhabung eines Produktes auszutauschen. Auf die Integration der Serviceabteilung eines Unternehmens in dieses Weblog sollte auf keinen Fall verzichtet werden. Die nachfolgende Abbildung 2.1 verdeutlicht noch einmal sehr schön, welche Formen des Unternehmensblogs nach Peter Wolff unterschieden werden können:

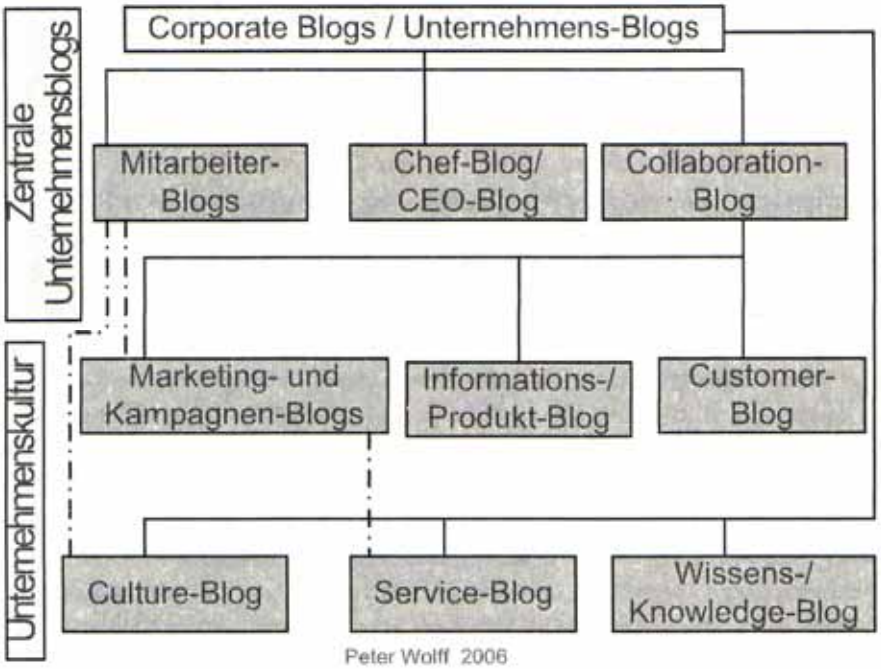

Abbildung 2.1 Unterteilung von Unternehmensblogs [Wolf07]

CEO-Blogs stellen einen Trend aus Amerika dar und definieren ein persönlich gehaltenes Weblog des Unternehmenschefs [Wolf07]. Die Erwartungshaltungen in diesem Blog liegen vorwiegend in den Bereichen Persönlichkeit, Authentizität sowie der Kompetenz. Zu den Vorreitern der CEO-Blogs zählen unter anderem diejenigen des Luftfahrtkonzerns BOEING, des Softwareunternehmens SUN sowie des Automobilkonzerns General Motors [Wolf07].

Im Gegensatz zu Unternehmensblogs bieten persönliche Blogs den Individualisten unter den Webusern die Gelegenheit, auf sich aufmerksam zu machen. Künstler, Politiker, Freiberufler „Ich-AGs" sowie „Otto-Normal-User" beispielsweise können durch vorzügliche individuelle Blogbeiträge besonderen Kultstatus innerhalb der Blogosphäre erlangen [Wolf07].
„Normale Ich-Blogs" bei welchen ein einzelner Blogger ein eigenes Blog betreibt, können insofern für ein Unternehmen interessant sein, wenn der Blogger Beiträge über Produkte, Mitarbeiter beziehungsweise das betreffende Unternehmen an sich verfasst. Aufgrund der enormen Flut an Weblogs gibt es durchaus Stimmen die argumentieren, dass es unmöglich sei die zahlreichen Blogs permanent zu kontrollieren beziehungsweise Einfluss auf diese zu nehmen.
Im Kontrast zu den „Otto-Normal-Bloggern" steht die Gruppierung der sogenannten politischen Blogger. Erste positive Erfahrungen zeichneten sich im Rahmen des letzten US-Präsidentenwahlkampfes ab. Im Lauf dieses Wahlkampfes stellte sich der Einsatz von Weblogs als sehr effiziente und kostengünstige Methode heraus, um einen Dialog mit den Bürgern und Bürgerinnen herzustellen. Es stellte sich weiters heraus, dass Blogs sehr gut dafür geeignet sind, um die eigene Partei oder auch Wahlkämpfer dementsprechend zu informieren und zu animieren als dies in klassischen Medien wie Zeitungen der Fall ist. Darüber hinaus bedienten sich die Wahlkampf-Kandidaten dieses neuen Kommunikationsinstrumentes um

Spenden einzusammeln [Wolf07]. Weiters, ähnlich wie bei den CEO-Blogs, entstanden auch in dieser Blog-Kategorie sehr unterschiedliche Lösungen durch deutsche Berufspolitiker. Dennoch fehlt in diesem Bereich, wie in einigen anderen, ein wesentliches Kriterium: die Möglichkeit ein Feedback abzugeben. Aufgrund der hier erwähnten Vorteile von Blogs liegt die Vermutung nahe, dass in absehbarer Zukunft immer mehr Politiker, auch vermehrt Kommunalpolitiker, die Möglichkeit eines Weblogs in Anspruch nehmen werden, um über den direkten Weg die eigenen Standpunkte einer breiten Öffentlichkeit zugänglich zu machen [Wolf07].

Bei den bisher vorgestellten Typisierungsvarianten wurde versucht Weblogs unter anderem anhand ihrer vielfältigen Themengebiete zu charakterisieren. [Zerf05] hingegen schlagen eine Typisierung der Blogger an sich vor. Sie definieren demzufolge vier Autorentypen, welche sich durch ihre vorherrschende Nutzungsweise unterscheiden:

- Beobachter/Kommentatoren: nehmen eher die Rolle des passiven Verfolgens einer Diskussion ein und verfassen sehr selten eigene Beiträge
- Autoren/Erzähler: betreiben meist ein privates Weblog um mit ihnen bekannten Personen in Kontakt zu treten
- Themenanwälte/Vernetzer: stellen spezifische Inhalte in ihrem Weblog zur Verfügung
- Botschafter/Moderatoren: diese Blogger-Gruppe ist am stärksten präsent beziehungsweise sichtbar, da sie eine Vielzahl von aktuellen Themen filtern sowie kommentieren [Schm06]

Solche Typologien entstehen größtenteils auf der Grundlage von beobachteten Praktiken, ohne jedoch auf kommunikationssoziologische Aspekte beziehungsweise Auswirkungen einzugehen. Aus diesem Grund wird im

folgenden Kapitel unter anderem ein kommunikationssoziologisches Analysemodell beschrieben, welches Bezug zu ausschlaggebenden Handlungsrahmenbedingungen nimmt. Die Tabelle 2.1 zeigt eine Auswahl an populären Weblogseiten und die URL, unter welcher diese im Internet zu finden sind.

Blog-Seite	URL
Livejournal	http://www.livejournal.com/
blog.de	http://www.blog.de/
Wordpress	http://de.wordpress.com/
twoday.net	http://twoday.net/
blogger.de	http://www.blogger.de/
Blogverzeichnis	http://www.blogverzeichnis.at/
blogbar.de	http://blogbar.de/

Tabelle 2.1 Auswahl namhafter Blog-Seiten

Kapitel 3

3 Analysemethoden für Weblogs

Kapitel drei beschäftigt sich thematisch mit dem Schwerpunkt Analysemethoden für Weblogs. Im Wesentlichen lassen sich derzeit zwei große Strömungen beziehungsweise Ansätze in Zusammenhang mit Blog-Analysemethoden erkennen [NiXi07]: Der erste Ansatz bezieht sich auf die Strukturverbindungen zwischen den Blogs. [Kuma03] beispielsweise entwickelten eine Methode um die Evolution der Blogger-Gemeinschaften zu erforschen. Der zweite Ansatz zieht den eigentlichen Inhalt existierender Weblogs als Grundlage für mögliche Analyseverfahren heran. Zahlreiche Wissenschaftler wie [Gruh04], [Glan04] oder auch [Dura06] beschreiben verschiedene Verfahren oder führen Experimentalreihen durch, um bestimmte blogübergreifende Tendenzen zu bestimmen, um Klassifizierungssysteme für die Stimmungslagen von Blog-Artikeln zu erforschen und um Aufschluss über weitere blogrelevante Anwendungsgebiete zu bekommen. Im Folgenden wird eine Auswahl an heutzutage existierenden Analysemethoden detailliert behandelt.

3.1 Kommunikationssoziologisches Analysemodell

Im Zuge der „New Media Studies“, ein interdisziplinäres Forschungsfeld welches die Wechselwirkung von technischen und sozialen Phänomenen observiert, stellen sogenannte neue Medien (email, World Wide Web, audio-visuelle Medien, etc.) ein soziotechnisches Geflecht aus Handlungen, Artefakten und Formen der sozialen Organisation dar [Schm06]. Trotz zahlreichen wissenschaftlichen Untersuchungen zum Thema „Weblogs“ fehlt

ein Analysemodell, welches handlungs- und netzwerkzentrierte Perspektiven in weblogbasierte Nutzungsepisoden integriert. Das folgende kommunikationssoziologische Analysemodell nach [Schm06] setzt sich mit der erfolgreichen Integration dieser Perspektiven auseinander und versucht auch Anknüpfungspunkte an andere Disziplinen zu ermöglichen. Die Grundlagen sowie gleichzeitig rahmende Strukturdimensionen dieses Analysemodells setzen sich aus drei großen Bereichen zusammen:

- Regeln: Besonders deutlich sichtbar beziehungsweise wahrnehmbar werden diese Regeln in ihrer Verdichtung als Rollen. Bei Weblogs ist eine Unterteilung in Autoren-, Kommentatoren- sowie Leserrollen naheliegend. Eine Studie nach [Rain05] zeigte unter anderem, dass sich die meisten amerikanischen Bürger unter dem Blog-Begriff nichts Konkretes vorstellen konnten. 38 Prozent hatten eine ganz gute Idee davon, die restlichen 62 Prozent konnten jedoch nichts mit dem Bloggen anfangen. Dabei ist es interessant zu beobachten, dass sich diejenigen, die sehr wohl mit Blogs vertraut sind, gebildet und starke Internet-Nutzer sind. Diejenigen allerdings, die nur begrenzt Ahnung von Blogs haben, stellen sich als Neulinge im Internet heraus und verfügen über ein niedrigeres Bildungsniveau. Neben dieser Aufgliederung in verschiedene Nutzerrollen kann eine weitere Einteilung in Adäquanzregeln und prozeduralen Regeln getroffen werden. Adäquanzregeln bestimmen vor allem die Medienwahl und können herangezogen werden, um Aufschluss über die Motive für das Führen eines Weblogs zu erlangen. Die nachstehende Abbildung 3.1 stellt die Rangfolge für die Motive dar.

N=4309	Prozent
Zum Spaß	70,8
Weil ich gerne schreibe	62,7
Um eigene Ideen und Erlebnisse für mich selbst festzuhalten	61,7
Um mich mit anderen über eigene Ideen und Erlebnisse auszutauschen	49,0
Um mir Gefühle von der Seele zu schreiben	44,5
Um mein Wissen in einem Themengebiet anderen zugänglich zu machen	33,4
Um mit Freunden und Bekannten in Kontakt zu bleiben	33,2
Um neue Bekanntschaften und Kontakte zu knüpfen	27,2
Aus beruflichen Gründen	12,7
Aus anderen Gründen	10,7

Abbildung 3.1 Motive für das Führen eines Weblogs [Schm06]

Prozedurale Regeln beziehen sich eher auf den Gebrauch des Mediums und demzufolge auf eine erfolgreiche Kommunikation. Eine Differenzierung dieser prozeduralen Regeln ist einerseits aufgrund ihrer Explizitheit (Allgemeine Geschäftsbedingungen sowie Nutzungsvereinbarungen von Weblogs, „blogging guidelines“, etc.) und andererseits aufgrund folgender drei analytischer Regeltypen möglich: Rezeptionsregeln beschäftigen sich vorwiegend damit, welche Inhalte auf Weblogs beziehungsweise anderen Quellen rezipiert werden. Sie schränken somit die Auswahl von Medieninhalten sehr stark ein. Die Entscheidung über die tatsächliche Rezeption wird jedoch nicht jedes Mal neu getroffen, sondern beruht auf Erfahrungen der Nutzer aus früheren Nutzungsepisoden [Schm06]. Publikationsregeln bieten den Weblogautoren gewisse Richtlinien an, welche Themen beispielsweise in einem Blog behandelt werden sollten oder in welcher Form multimediale Inhalte wie Bilder in das Blog integriert werden sollten. Das Identitätsmanagement spielt bei den Publikationsregeln eine große Rolle, da es in den meisten Fällen um die Selbstdarstellung des Autors gegenüber anderen Personen geht.

Der dritte Regeltyp wird als Vernetzungsregel bezeichnet. Aufgrund der Dynamik und Gestalt der Weblog Netzwerke sollte diese Regel gesondert betrachtet werden. Vernetzungsregeln beeinflussen im Allgemeinen die explizite Verlinkung innerhalb eines Weblogs. Sie gehen jedoch über den rein technischen Aspekt hinaus, da sie sehr wohl auch Einfluss auf die Gestalt von Beziehungen (diese können sowohl inhaltlicher als auch sozialer Natur sein) nehmen. Die Relationen innerhalb der Weblog Netzwerke werden durch verschiedene technische Kommunikationsmerkmale unterstützt: Im Zentrum dabei steht der Permalink, bei welchem auf der Startseite eines Blogs Beiträge gemeinsam erscheinen jedoch durch eigenständige URL's (Uniform Resource Locator) adressiert werden können [Schm06]. Auf diese Art und Weise kann eine gezielte Verlinkung erreicht werden. Beinahe gleichbedeutend für Gestaltung dichter Netzwerke ist die sogenannte Trackback Funktion. Diese Funktion (wurde ursprünglich für die Weblog-Software „Moveable Type" angeboten) löst das Problem der einseitigen Hyperlinks (ein Link von Seite A auf Seite B führt prinzipiell nicht wieder zur Seite A zurück). Andere technische Mittel wie etwa die Blogroll dienen zur Unterstützung des Aufbaus von Netzwerk Relationen, indem sie verschiedene Quellen miteinander vereinen. Gemeint sind hier Verweise auf andere Weblogs, welche meist in Form von Leseempfehlungen oder als Listen befreundeter Autoren vorliegen. Blogrolls befinden sich in der Regel auf der Startseite eines Weblogs und besitzen so gesehen einen stärkeren Stellenwert als Links in Einzelbeiträgen, da diese nach einer gewissen Zeit von der Startseite verschwinden. Abbildung 3.2 zeigt, welche Typen von prozeduralen Regeln nach [Schm06] unterschieden werden und welche Fragestellungen sie zum Inhalt haben.

Regeltyp	Inhalt	Kontext
Rezeptionsregeln	Welche Inhalte werden über welche Kanäle rezipiert?	Informations-management
Publikationsregeln	Welche Themen werden wie für das Weblog aufbereitet?	Identitäts-management
Vernetzungsregeln	Wann wird in welcher Form auf welche anderen Inhalte verwiesen?	Beziehungs-management

Abbildung 3.2 Typen von prozeduralen Regeln [Schm06]

- Netzwerke: Sie spielen insofern eine große Rolle, da Weblogs besonders den Aufbau sozialer Netze unterstützen. Sowohl in privaten als auch in wirtschaftlichen und politischen Organisationen werden Weblogs als eine wichtige Form sozialer computervermittelter Kommunikation über weite geographische Strecken eingesetzt. Besonders interessant zu beobachten ist, dass weblogbasierte soziale Netzte unter anderem die Formierung von Sozialkapital fördern. Unter Sozialkapital wird die Möglichkeit eines Akteurs verstanden innerhalb seiner Position in sozialen Geflechten bestimmte Ressourcen (emotionale Unterstützung, Informationsaustausch, etc.) zu mobilisieren [Schm06]. Dabei gilt es zu beachten, dass ein Akteur über umso mehr verbindendes Sozialkapital verfügen kann je stärker er in ein Netzwerk eingebunden ist, in welchem die Mitglieder untereinander bereits sehr eng in Verbindung stehen.

  Die Kanalisierung von Aufmerksamkeit, bei welcher Links innerhalb der Blogosphäre nicht gleichmäßig verteilt sind sondern einer bestimmten Gesetzmäßigkeit folgen, ist ebenfalls auf die Netzwerkbildung durch Weblogs zurückzuführen. Die Verteilung der Links gestaltet sich nach dem sogenannten „Power Law“ (auch als Potenzgesetz bezeichnet): Eine relativ kleine Anzahl an Weblogs aggregiert eine große Anzahl an eingehenden Links während die überwiegende Mehrheit von Weblogs eine geringe Zahl solcher Links beherbergt. Für vielfach vernetzte

Weblogs bedeutet dies ganz allgemein, dass sie einerseits eine höhere Aufmerksamkeit erzielen als schwach vernetzte Weblogs und dass sie mit hoher Wahrscheinlichkeit weitere Verlinkungen auf sich ziehen [Schm06].

- Software: Die dritte rahmende Strukturdimension innerhalb des kommunikationssoziologischen Analysemodells wird durch den Software-Code für Weblogs dargestellt. Als technische Basis besteht eine der Grundaufgaben der Software in der Eröffnung beziehungsweise Ausschließung von Handlungsmöglichkeiten. Die Software sollte auf keinen Fall als unwichtig im Weblog Prozess abgestempelt werden, da sie sehr wohl als Resultat von spezifischen Praktiken gesehen werden kann. Sie unterstützt einerseits die Selektion und Präsentation von Inhalten und andererseits die Vernetzung mit anderen Quellen [Schm06]. Dennoch sind der Weiterentwicklung im Bereich sozialer und technischer Innovationen keine wirklichen Grenzen gesetzt. Bei der Software-Entwicklung besteht jedoch das grundsätzliche Problem, dass bei den beteiligten Akteuren unterschiedliche Wissensstände vorhanden sind: den professionellen Programmierern mit ihrem Expertenwissen steht das Wissen der Alltagsnutzer gegenüber. Ein vorrangiges Ziel innerhalb der Entwicklungsarbeit besteht darin, diese Wissensdifferenzen abzugleichen. Am Beispiel der Weblog-Publishing Software „Wordpress“ lässt sich zeigen, dass in den Weblogs beziehungsweise in den Entwicklerforen soziale Netzwerke, welche von der Ad-hoc-Öffentlichkeit einer Frage und ihrer Antwort bis hin zu virtuellen Gemeinschaften engagierter Programmierer und Software-Tester reichen, entstehen [Schm06]. Mit Hilfe solcher Netzwerke kann aufgrund der auftretenden informationellen Unterstützung die angesprochene Wissensdifferenz zwischen den Experten und den Alltagsnutzern überbrückt werden. Ein weiteres Indiz dafür, dass der

Software-Code sehr wohl starken Einfluss auf Erwartungen, Regeln und Netzwerken innerhalb der Blogosphäre hat, ist der hohe kommunikative Aufwand der in der Entwicklung von Weblog-Produkten steckt und der bewusst von vielen Entwicklern regelrecht gesucht wird.

Zusammenfassend lässt sich sagen, dass sich in den letzten Jahren zahlreiche Einsatzfelder im Weblog-Bereich heraus kristallisiert haben, die allesamt nach einer differenzierten Analyse verlangen. Das kommunikationssoziologische Analysemodell wählt einen praxisorientierten Ansatz, da dieser einen Vergleich von unterschiedlichen Gebrauchsweisen und daraus resultierenden Folgen transparenter macht als andere Ansätze. Weiters werden individuelles Handeln und strukturelle Resultate gleichberechtigt in die eben erwähnten vergleichenden Elemente miteinbezogen.
Darüber hinaus zeichnet sich das kommunikationssoziologische Modell auch dadurch aus, dass es die drei, für eine computervermittelte Kommunikation ausschlaggebenden und rahmenden, Strukturdimensionen: die Software als technisches Merkmal, Regeln zum Gebrauch von Weblogs (Adäquanzregeln sowie prozedurale Regeln) und hypertextuelle beziehungsweise soziale Netzwerke aufgreift. Die nachstehende Abbildung 3.3 visualisiert sehr schön, wie die erwähnten Strukturdimensionen in Beziehung zueinander stehen. Die drei Strukturdimensionen „Regeln“, „Software“ und „Relationen (in der Auflistung als Netzwerke bezeichnet)“ stehen in wechselseitiger Beeinflussung zueinander und geben der individuellen Nutzungsepisode einen Rahmen vor.

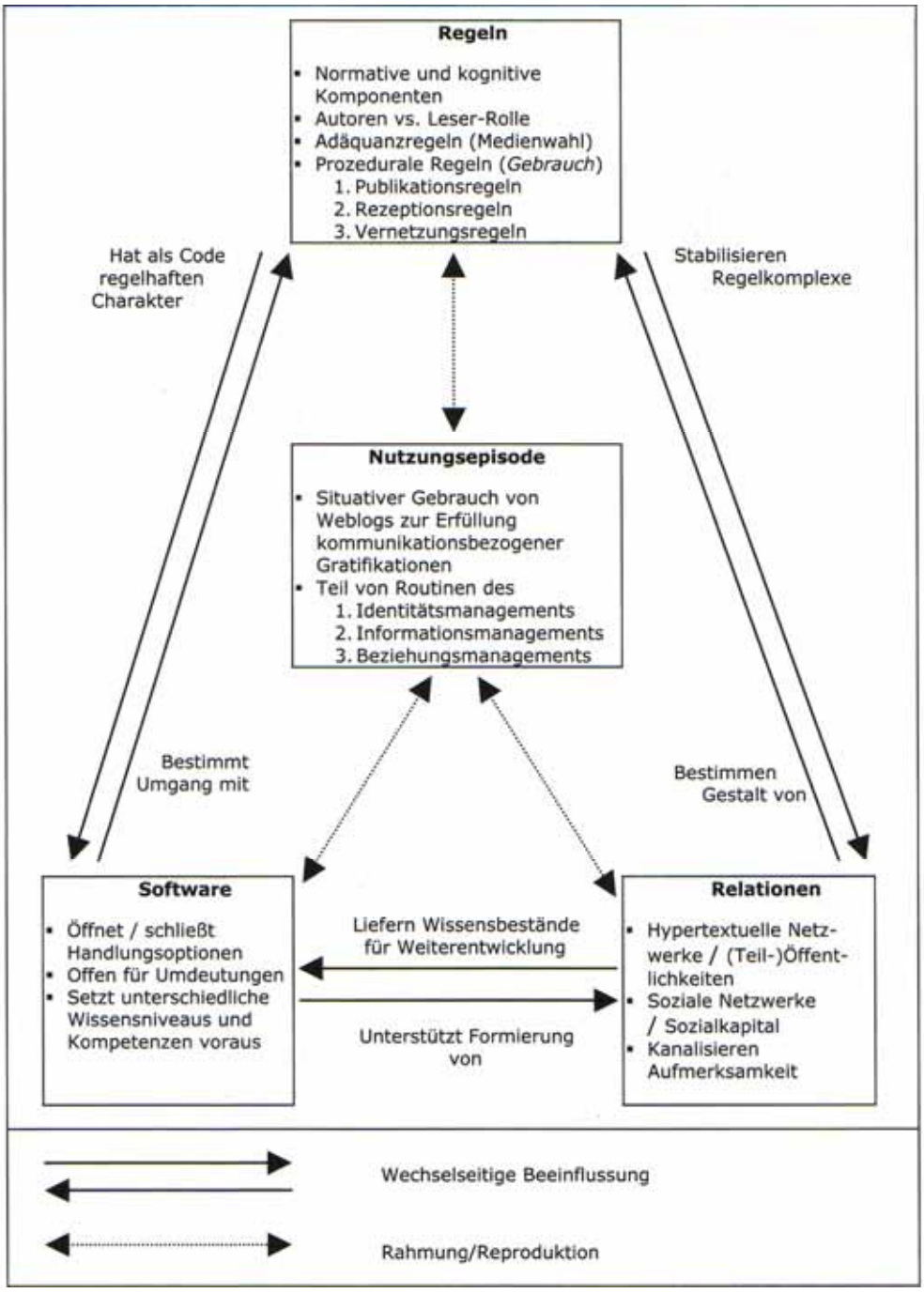

Abbildung 3.3 Das kommunikationssoziologische Analysemodell [Schm06]

3.2 Analyseverfahren zur Klassifizierung von Weblog-Artikel

Die folgenden Methoden beziehungsweise Ansätze beschreiben technische Alternativen um Blogs anhand ihrer eigentlichen Inhalte zu analysieren und zu charakterisieren. Auf diese Art und Weise soll es einerseits möglich sein, Weblog-Korpora gewissen Genres zuzuordnen sowie andererseits eine Klassifizierung, basierend auf Emotionen, zu finden.

3.2.1 Informative sowie Affektive Artikel innerhalb der Blogosphäre

Nach [NiXi07] können bei Inhalten von Blogs im Allgemeinen zwei Kategorien beziehungsweise Genres ausgemacht werden: Auf der einen Seite stehen diejenigen Inhalte, welche über Emotionen beziehungsweise Gedanken der Blogger handeln (diese werden als affektive Artikel bezeichnet). Diesen Inhalten können die sogenannten informativen Artikel gegenübergestellt werden, welche ausschließlich über diverse Technologien sowie verschiedene Arten informativer Nachrichten Aufschluss geben. [NiXi07] zielen mit ihrer Analysemethode und den dafür entwickelten Werkzeugen darauf ab, informative- von affektiven Blog-Artikel zu separieren. Das Genre-Erkennungsproblem wird von den Autoren als ein binäres Klassifikationsproblem betrachtet, welches mittels Text-Klassifizierungstechniken gelöst wird. Durch dieses Klassifikationsproblem entstehen jedoch einige wissenschaftliche Hürden:

- Die Definition von affektiven sowie informativen Artikel sollte mit den Intentionen der Blogger und Blog-Besucher übereinstimmen.
- Eine angemessene Auswahl an Trainings-Korpora sollte für beide Kategorien existieren.
- Algorithmen des maschinellen Lernens sollten sorgfältig ausgewählt und adaptiert werden, um dieses Problem mit einer hohen Effizienz zu lösen

Zur erfolgreichen Bewältigung dieser Hürden sieht der Ansatz zu ihrem Verfahren im Überblick so aus, dass zu aller erst eine Benutzer-Studie durchgeführt wird, um Beschreibungen über affektive sowie informative Artikel zu sammeln. Die eigentliche Definition der beiden inhaltlichen Kategorien erfolgt aufgrund einer Zusammenfassung der einheitlichsten Benutzerbeschreibungen. Weiters wird eine Anzahl an Artikel von traditionellen öffentlichen Webseiten gewählt, welche die Trainings-Auswahl der informativen Kategorie darstellt. Darüber hinaus werden einige Personen gebeten, Blog-Artikel anhand der beiden Genres zu kennzeichnen. Diese markierten Artikel werden in einen Trainings- sowie in einen Test-Teil aufgeteilt, von denen der Test-Teil zur Evaluation unterschiedlicher Algorithmen aus dem Bereich des maschinellen Lernens herangezogen wird.

Darüber hinaus wird die Performanz von drei Klassifizierungs-Algorithmen: der Naïve Bayes (NB), der Support Vector Machine (SVM) und der Algorithmus von Rocchio ausgewertet. Weiters wird die Leistung zweier Methoden zur Auswahl von Attributen (Chi-square und Information Gain) untersucht.
Im Anschluss an diesen Schritt werden drei große Studien durchgeführt, von denen die erste Studie eine emotionale sowie themenbezogene Klassifizierung von Blog-Artikel zum Inhalt hat. Die Ergebnisse aus dieser ersten Studie werden auf eine „vorhabengesteuerte“ Suchmaschine appliziert. In der dritten und letzten Phase wird eine Studie über die automatische Erkennung von hochqualitativen Blogs durchgeführt [NiXi07].
In den folgenden Abschnitten werden die, bisher im Überblick dargestellten, Ansätze innerhalb dieses Analyseverfahrens detailliert beschrieben.

Wie bereits erwähnt, steht am Beginn die Frage nach einer Definition von informativen und affektiven Blog-Artikel. Aufgrund dessen werden zahlreiche Benutzer darüber befragt, welche Arten von Blog-Inhalten sie bevorzugt lesen. Aus dieser Umfrage wiederum lassen sich einige wesentliche Charakteristika

über affektive sowie informative Artikel ableiten. Informative Artikel wurden von den befragten Personen folgendermaßen definiert:

- Technische Beschreibungen
- Objektive Kommentare über Ereignisse in der Welt
- „Vernünftiges" Wissen
- Nachrichten, welche ähnlich denjenigen auf traditionellen Webseiten sind

Als Beispiele für Artikel der affektiven Kategorie hingegen wurden einerseits „Tagebücher über persönliche Angelegenheiten" sowie andererseits „gefühls- und emotionseigene Beschreibungen" angeführt. Anhand dieser Definitionen über affektive- und informative Artikel wählten [NiXi07] einen menschlichen Evaluationsansatz um ihre Datensammlung zu bilden (diese ist unter „http://www.apexlab.org/apex_wiki/ia-blogdata" zugänglich).

Als nächster Schritt in der Studie nach [NiXi07] folgt die Effizienz-Bewertung bestehender Text-Klassifizierungs-Algorithmen sowie Attributauswahl-Algorithmen. Im Folgenden werden drei Klassifizierungs-Algorithmen, welche häufige Anwendung finden, detailliert beschrieben:

Naïve Bayes Algorithmus

Eine wichtige Domäne im Bereich des maschinellen Lernens ist die Dokument-Klassifikation. Dokumente werden durch Wörter, welche darin vorkommen, charakterisiert. Eine mögliche Art und Weise um maschinelles Lernen auf die Dokument-Klassifikation anzuwenden besteht darin, das Vorhandsein oder die Absenz von jedem Wort als Boolean-Attribut zu behandeln. Naïve Bayes (NB) bietet sich für diese Anwendung an, da er einen simplen jedoch sehr effektiven und schnellen Maschinenlern-Algorithmus darstellt [Witt05]. Der Naïve Bayes Klassifizierer hat den Satz von Bayes als Grundlage und seine Aufgabe besteht

darin ein Wahrscheinlichkeitsmodell aus Daten zu bauen und diese zu verwenden, um die Klassifikation neuer Beispiele vorauszusagen. Aus diesem Grund werden die zu klassifizierenden Daten in einen Trainingsdatensatz, an welchen der Algorithmus aufgebaut wird, und in einen unmarkierten Testdatensatz, auf welchen die Ergebnisse aus dem Trainingsdatensatz tatsächlich angewandt werden, unterteilt. Der Satz von Bayes für zwei Zufallsereignisse (A und B) lautet: $\Pr(A|B)=\frac{\Pr(B|A)x\Pr(A)}{\Pr(B)}$. $\Pr(A)$ beschreibt die unbedingte Wahrscheinlichkeit von Ereignis A und $\Pr(B)$ bezieht sich auf die unbedingte Wahrscheinlichkeit von Ereignis B. Diese beiden Wahrscheinlichkeiten werden als a priori Wahrscheinlichkeiten bezeichnet. $\Pr(A|B)$ ist die bedingte Wahrscheinlichkeit für ein Ereignis A unter der Bedingung, dass Ereignis B eingetreten ist. Diese beiden Wahrscheinlichkeiten werden a posteriori Wahrscheinlichkeiten genannt. Der Naïve Bayes kann für Lernaufgaben verwendet werden, bei welchen einerseits strukturierte Daten (zum Beispiel Tabellen) oder andererseits unstrukturierte Daten (Texte, Webdokumente, etc.) vorliegen.

Allerdings wird die Anzahl wie oft jedes Wort vorkommt nicht in Betracht gezogen obwohl diese Information sehr nützlich ist, wenn eine Kategorie für ein Dokument bestimmt werden soll. Stattdessen kann ein Dokument als ein „Beutel voller Wörter (engl. bag of words)“ betrachtet werden. Dieser „Beutel“ beinhaltet alle Wörter in einem Dokument, mit multiplem Auftreten eines Wortes, welches mehrfach vorkommen kann. Wort-Häufigkeiten können unter Anwendung einer modifizierten Form des Naïve Bayes (manchmal als multinomineller Naïve Bayes bezeichnet) angepasst werden. Die Formel für die Berechnung der Wahrscheinlichkeit eines Dokumentes E einer Klasse H lautet

$$\Pr\left[E\middle|H\right]\approx N!*\prod_{i=1}^{k}\frac{P_i^{n_i}}{n_i!}$$

unter folgenden Annahmen: $N=n_1,n_2,...,n_k$ ist die Anzahl wie oft ein Wort i in dem Dokument vorkommt und $P_1, P_2,..., P_k$ ist die Wahrscheinlichkeit um ein Wort i während einer Stichprobenerhebung von allen Dokumenten aus der Kategorie H zu erhalten. Weiters wird angenommen, dass die Wahrscheinlichkeit unabhängig von dem Kontext und der Position des Wortes innerhalb des Dokumentes ist.

Folgendes Beispiel demonstriert die praktische Anwendung des Naïve Bayes: Es existieren lediglich zweit Wörter (gelb und blau) und die Wahrscheinlichkeit einer speziellen Klasse ist $\Pr[gelb|H]=75\%$ und $\Pr[blau|H]=25\%$ in einer speziellen Klasse H. E ist das Dokument, bestehend aus (*blau*, *gelb* und *blau*) mit einer Länge $N=3$. Es gibt nun vier Möglichkeiten an Beuteln aus diesen drei Wörtern. Die Wahrscheinlichkeit für eine dieser Möglichkeiten anhand der obigen Formel lautet:

$$\Pr[\{gelbgelbgelb\}\mid H]\approx 3!*\frac{0{,}75^3}{3!}\,x\,\frac{0{,}25^0}{0!}=\frac{27}{64}=42{,}19\%$$

Analog dazu können auch die restlichen drei Wahrscheinlichkeiten (zum Beispiel von *gelb, gelb* und *blau*) berechnet werden.

In der Untersuchung von [NiXi07] wurde der Naïve Bayes Klassifikator auf die Text-Klassifizierung angewandt und kann folgendermaßen dargestellt werden:

$$P(c/d)=\frac{P(c)xP(d/c)}{P(d)}$$

c .....Text-Kategorie

d ..... Text-Dokument

$P(c)$ .....ursprüngliche Kategorie-Verteilung

Naïve Bayes kann dermaßen durch Suchen einer optimalen Kategorie, welche die nachfolgende Wahrscheinlichkeit $P(c|d)$ maximiert, konstruiert werden [NiXi07].

Rocchio Algorithmus

Dieser Algorithmus stellt einen am weitesten angewandten Lern-Algorithmus für die Text-Kategorisierung dar [Joac97]. Er wurde für die Informationswiedergewinnung entwickelt, kann jedoch auf die Text-Kategorisierung und Routing-Probleme adaptiert werden. Innerhalb dieses Algorithmus existieren folgende Parameter:

- TF (Term/Text Frequency): Gibt an wie oft ein Wort im Artikel vorkommt
- DF (Document Frequency): Anzahl der Dokumente in denen ein Wort vorkommt
- IDF: kennzeichnet die inverse DF

Trotzdem, dass der Rocchio Algorithmus intuitiv ist, weist er einige Probleme auf wie: heuristische Komponenten bieten sehr viele Design-Möglichkeiten an und es besteht wenig Anleitung wenn man diesen Algorithmus auf eine neue Domäne anwendet. Weiters ist es nicht klar, welche Heuristiken die besten Ergebnisse bei der Text-Kategorisierung liefern
Diese führen nach [Joac97] zu einer vergleichsweise niedrigen Klassifizierungs-Genauigkeit.
Die Funktionsweise des TFIDF Klassifikator sieht folgendermaßen aus: Der Algorithmus liefert eine Reihung an Dokumenten zurück jedoch ohne Bestimmung eines Schwellenwertes um eine Regel für die Klassenzugehörigkeit zu definieren. Aus diesem Grund muss der Algorithmus

dementsprechend adaptiert werden, um für die Text-Kategorisierung verwendet zu werden. Die hier vorgestellte Variante scheint nach [Joac97] die geradlinigste Adaption des Rocchio Algorithmus auf die Text-Kategorisierung und Domänen mit mehr als zwei Kategorien zu sein. Jedes Dokument d wird durch den Vektor $\vec{d} = (d^{(1)},...,d^{(|F|)})$ repräsentiert, sodass Dokumente mit ähnlichen Inhalten ähnliche Vektoren haben. Jedes Element d^i stellt ein eindeutiges Wort w_i dar. d^i für ein Dokument d wird aus einer Kombination der Statistiken $TF(w_i, d)$ und $DF(w_i)$ berechnet. Die *term frequency* $TF(w_i, d)$ gibt die Anzahl an, wie oft ein Wort w_i im Dokument d vorkommt. Die *document frequency* $DF(w_i)$ beschreibt die Anzahl an Dokumenten in welchen ein Wort w_i zumindest einmal vorkommt. Die *inverse document frequency* $IDF(w_i)$ kann wie folgt über die Dokumentfrequenz berechnet werden [Joac97]. Die inverse Dokumentfrequenz eines Wortes weist einen niedrigen Wert auf wenn das Wort in vielen Dokumenten vorkommt. Umgekehrt hat sie ihren höchsten Wert, wenn das Wort lediglich in einem Dokument enthalten ist. Die sogenannte Gewichtung $d^{(i)}$ eines Wortes w_i in einem Dokument d berechnet sich durch: $d^{(i)} = TF(w_i, d) * IDF(w_i)$

Diese Wortgewichtungs-Heuristik besagt, dass ein Wort w_i einen wichtigen Index-Term für ein Dokument d darstellt, vorausgesetzt es kommt regelmäßig darin vor (eine hohe *term frequency*). Das Lernen des Algorithmus wird durch das Kombinieren von Dokument-Vektoren erreicht, mit dem Resultat eines Prototypvektors $\vec{c}_j$ für jede Klasse C_j. Zu Beginn werden die normalisierten Vektoren von sowohl den positiven Beispielen einer Klasse als auch von den negativen Beispielen einer Klasse zusammengefasst. Der Prototypvektor wird als gewichtete Differenz aus jedem dieser Vektoren berechnet:

$$\vec{c}_j = \alpha \frac{1}{|C_j|} \sum_{\vec{d} \in C_j} \frac{\vec{d}}{\|\vec{d}\|} - \beta \frac{1}{|D - C_j|} \sum_{\vec{d} \in D - C_j} \frac{\vec{d}}{\|\vec{d}\|}$$

$\vec{d}$ .....Dokumente mit den Ausdrücken, welche nach dem TFIDF Schema gewichtet sind

D .....Trainingsset

$|c_j|$ .....Anzahl an Dokumenten in der Kategorie c_j

$|D - c_j|$ ....Anzahl an Dokumenten, welche nicht in der Kategorie c_j liegen

λ, β .....Parameter um den relativen Einfluss von positiven und negativen Trainingsbeispielen zu steuern (es wird empfohlen für $\lambda = 16$ und $\beta = 4$ zu verwenden)

Dieses Modell kann verwendet werden, um ein neues Dokument d' zu klassifizieren. Das Dokument wird ebenfalls unter Verwendung des obigen Schemas durch einen Vektor $\vec{d}'$ repräsentiert.

Support Vector Machine Algorithmus

Der Support Vector Algorithmus (auch unter der nicht sehr gebräuchlichen Übersetzung „Stützvektormaschine“ zu finden) gehört wie die beiden vorhergehenden Algorithmen zur Gruppe der sogenannten Klassifikatoren. Er wurde im Jahre 1995 von Vladimir Naumovich Vapnik sowie seinen Kollegen (wie zum Beispiel Aleksei Chervonenkis) entwickelt und kann als ein leistungsstarker Lern-Algorithmus bezeichnet werden [Wiki08b]. Grundsätzlich bewirkt eine Support Vector Machine die Unterteilung von einer Menge an Objekten derart in Klassen, dass ein möglichst großer Bereich rund um die Klassengrenzen frei von Objekten entsteht. Sie wurde bereits vielfach erfolgreich auf das Feld der Text-Klassifikation angewandt und erwies sich im Zuge dessen als sehr wirkungsvoll. Die Schlüsselidee hinter Support Vector Machines sowie anderen Lernansätzen (wie die bereits beschriebenen

Klassifizierungs-Algorithmen) besteht darin, eine Trainingsmenge an markierten Instanzen zu verwenden, um die Klassifizierungsfunktion automatisch zu lernen. Abbildung 3.4 zeigt, dass die einzelnen Objekte beziehungsweise relevanten Wörter durch Vektoren im Vektorraum repräsentiert werden. Diese Repräsentation eines Dokumentes als ein Vektor an Wörtern wird typischerweise in der Informationswiedergewinnung vorgefunden. Für jedes Trainingsdokument wird die Text Frequency (TF) berechnet. Für die Text-Klassifikation werden oftmals binäre Attributwerte (das Wort kommt im Dokument vor und kommt nicht vor) verwendet.

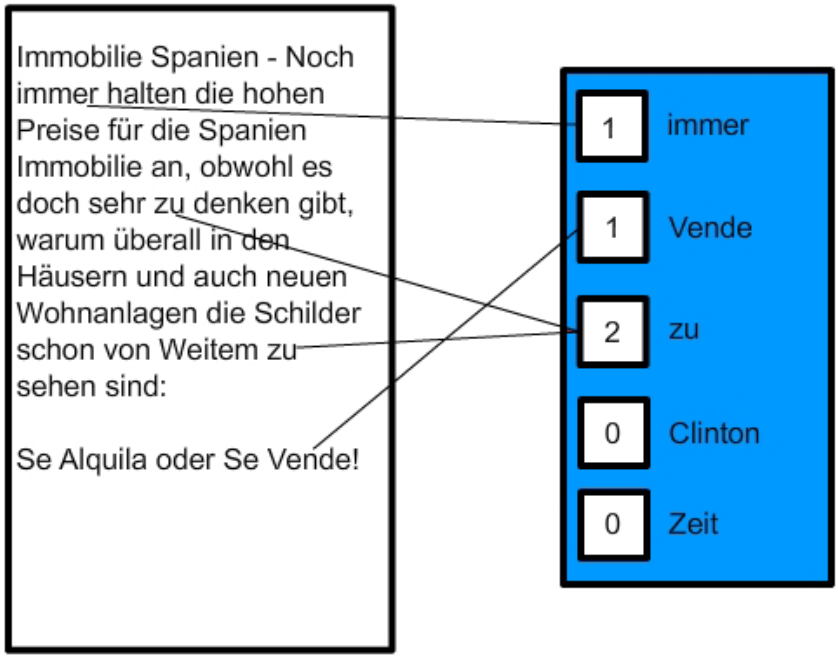

Abbildung 3.4 Text Frequency (TF)

Einfache lineare Modelle können wirkungsvoll für jene Klassifizierungen verwendet werden, bei welchen alle Attribute numerisch sind. Allerdings erweist sich als ein großer Nachteil, dass diese lediglich lineare Grenzen zwischen Klassen darstellen können. Support Vector Machines verwenden lineare Modelle um nicht lineare Klassen-Grenzen zu implementieren. Hierfür werden Eingabedaten unter Verwendung einer nicht linearen Abbildung transformiert. Der Instanzenraum wird sozusagen in einen neuen Raum transformiert [Witt05].

Die Aufgabe der Support Vector Machine besteht nun darin, eine mehrdimensionale Hyperebene in den Vektorraum zu integrieren, welche den Zweck einer Trennfläche, die die Trainingsobjekte in zwei Klassen unterteilt, erfüllt [Wiki08d]. Die maximale Rand-Hyperebene ist diejenige, welche die größte Teilung zwischen Klassen bewirkt. Jene Instanzen, welche am nächsten zur maximalen Rand-Hypereben liegen, werden als Support Vektoren bezeichnet [Witt05]. Es existiert zumindest ein Support Vektor für jede Klasse. Die Menge an Support Vektoren definiert eindeutig die maximale Rand-Hyperebene. Sind die Support Vektoren von zwei Klassen gegeben, kann die maximale Rand-Hyperebene sehr leicht konstruiert werden. Die verbleibenden Trainingsobjekte sind dabei irrelevant, da diese ohne Veränderung der Position und Orientierung der Hyperebene gelöscht werden können. Eine Hyperebene x, welche zwei Klassen von einander trennt, kann folgendermaßen geschrieben werden:

$$x = w_0 + w_1 a_1 + w_2 a_2$$

a_1, a_2 .....Attribut-Werte

w_i .....Gewichtungen, die gelernt werden müssen

Eine saubere Trennung durch die Hyperebene ist nur dann möglich, wenn die Objekte linear trennbar (kommen in realen Anwendungsfällen nicht immer vor) sind. Im Fall von nicht linear trennbaren Daten werden verschiedene sogenannte Kernel-Funktionen wie Polynom-Kernel, Sigmoid-Kernel, RBF (radial basis function)-Kernel, etc. verwendet, um nicht lineare Abbildungen zu implementieren. Abbildung 3.5 zeigt schematisch den Unterschied zwischen linear trennbaren und nicht linear trennbaren Daten.

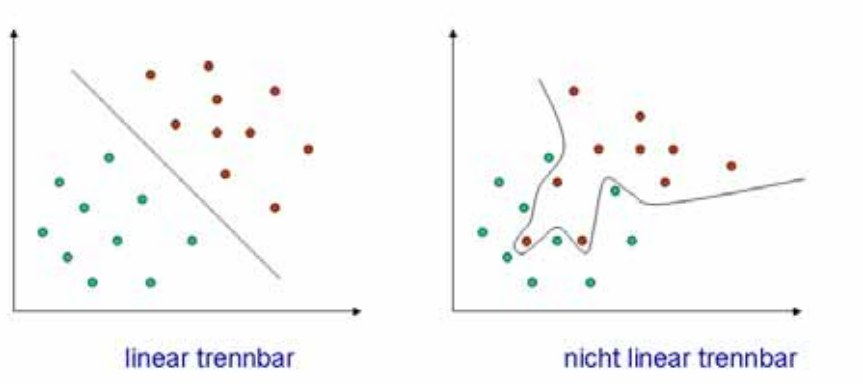

Abbildung 3.5 Linear-und nicht linear trennbare Objekte [Wiki08a]

Kernel-Funktionen führen den Vektorraum in einen Raum mit genügend hoher Dimensionszahl über und können somit selbst stark verschachtelte Vektormengen linear trennbar machen. Die trennende Hyperebene wird nun in diesem höher dimensionalen Raum bestimmt. Während der Rücktransformation in den niedriger dimensionalen Raum wird die lineare Hyperebene zu einer nicht linearen, die die Trainingsvektoren in zwei Klassen trennt. Allerdings ergeben sich bei diesem Prozess folgende Probleme: Die Hochtransformation ist einerseits sehr rechenintensiv und die Darstellung der Trennfläche im niedrig dimensionalen Raum ist andererseits komplex und damit praktisch unbrauchbar [Wiki08d]. Mit Hilfe von Kernel-Funktionen kann allerdings die Hyperebene im höher dimensionalen Raum beschrieben

werden. Somit ist es möglich, die Hin- und Rücktransformation umzusetzen, ohne sie tatsächlich rechnerisch ausführen zu müssen.
Abschließend lässt sich festhalten, dass selbst die schnellsten Trainings-Algorithmen für Support Vector Machines langsam in ihrer Anwendung auf nicht lineare Umgebungen sind, verglichen mit anderen Methoden wie beispielsweise „decision tree learners". Auf der anderen Seite produzieren Support Vector Machines oftmals sehr genaue Klassifizierer, da aufwändige und feine Grenzen gefunden werden können.

Die erste der drei bereits eingangs erwähnten Studien, welche von [NiXi07] im Rahmen ihrer Forschungsarbeit durchgeführt wurde, behandelt die Emotions- und Inhaltsklassifikation.
Aufgrund des enormen Wachstums an Blog Daten/Artikel (diese handeln zu einem erheblichen Teil von menschlichen Emotionen und Gefühlen) in den vergangenen Jahren sehen Wissenschaftler ihre Aufgabe darin, diese Artikel in vordefinierte Emotionskategorien einzuordnen. Der erste wesentliche Schritt innerhalb der Forschungsarbeit von [NiXi07] besteht darin, affektive Artikel aus dem gesamten Reservoir an Blogdaten zu extrahieren. Diese Methode verhilft im Wesentlichen dazu, einen Blog-Corpus mit rein emotionalen Inhalten zu konstruieren. Wie der nachstehenden Abbildung 3.6 zu entnehmen ist, werden im Rahmen von deren Analyseansatz einerseits informative Artikel in weitere verschiedene Themenkreise sowie andererseits affektive Artikel in weitere Typen an Emotionen klassifiziert.

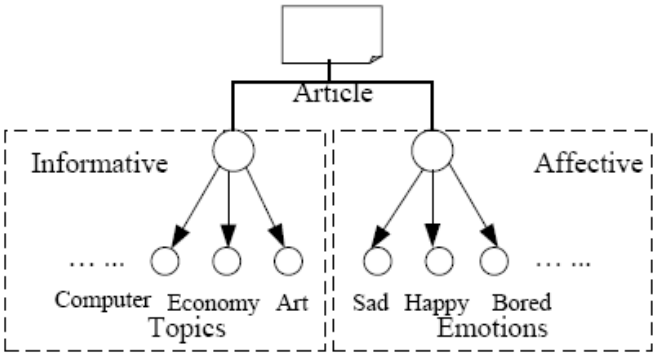

Abbildung 3.6 Ansatz zur Emotions- und Inhaltsklassifikation [NiXi07]

Um die Effektivität ihres Verfahrens zu untersuchen, führten [NiXi07] ein Experiment durch, bei welchem folgender Datensatz (dieser ist ebenfalls unter „http://www.apexlab.org/apex_wiki/ia-blogdata“ zu erreichen) verwendet wurde: 2494 Blog-Artikel (Trainingsdaten) wurden manuell anhand von zwei Emotionsrichtungen (positive und negative Artikel) bezeichnet. Weiters wurden willkürlich 75 Blogs von MSN Space mit insgesamt 1303 Artikel als Testdaten (ebenfalls manuell bezeichnet) ausgewählt. Artikel, welche eine Unklarheit ihrer Emotionsrichtung aufwiesen, wurden aus den Testdaten entfernt. Anschließend wurde die Performanz zweier Ansätze (I-Approach: eine erste Filterung der informativen Artikel unter Verwendung von informations-affektiven Klassifizierung und II-Approach: hier findet keine derartige Filterung der Artikel statt) zur Emotionsklassifikation miteinander verglichen, um die Effektivität des informations-affektiven Klassifizierungsansatzes zu zeigen. Tabelle 3.1 stellt die experimentellen Ergebnisse dar: Zuerst wurden nur die positiven und negativen Artikel als Testdaten für die Performanzmessung herangezogen. Im zweiten Schritt wurde der Klassifikator jedoch auf alle Testartikel (die informativen Artikel mit eingeschlossen) angewandt. Hierbei lässt sich erkennen, dass die Filterung – zuallererst von informativen Artikeln (I-Approach) – die Precision der Emotionsklassifikation signifikant erhöht und

nur einen kleinen Verlust in der Recall mit sich bringt. Die Precision ist in etwa ident mit derjenigen, bei welcher nur die positiven und negativen Artikel (II-Approach) getestet werden.

	Precision	Recall
Testing only on positive and negative articles	0.788	0.797
Testing on all articles		
I -Approach	0.762	0.785
II -Approach	0.596	0.797

Tabelle 3.1 Ergebnis aus den Ansätzen zur Emotionsklassifikation [NiXi07]

Die Ergebnisse aus diesem Experiment wurden auf die Entwicklung einer so genannten „Intentionsgesteuerten Weblog-Suchmaschine" angewandt.
Im Gegensatz zu „herkömmlichen" Suchmaschinen, bei welchen die die zurückgelieferten Blog-Einträge nach Datum oder Relevanz sortiert sind, konzentriert sich dieser Suchalgorithmus darauf, die verschiedenen Anforderungen der jeweiligen Benutzer gezielt miteinzubeziehen. Dies bedeutet beispielsweise, dass die eine Gruppe von Usern bei der Suchanfrage „IBM" sich Nachrichten über das Unternehmen an sich oder ihren Produkten erwartet während die andere Gruppe daran interessiert ist, einige Blogger-Kommentare oder Empfindungen über IBM oder ihren Dienstleistungen zu bekommen. Die beiden Genres (affektiv sowie informativ) an Blog Artikel werden als die beiden Rückgewinnungsabsichten betrachtet. Die informative Bedeutung eines Blog Artikels wird aus einem Zuversichtswert abgeleitet, welcher eine Bandbreite von -1 (entspricht einer starken affektiven Intention) bis 1 (entspricht einer starken informativen Intention) aufweist. Anschließend

folgt eine Neuordnung der Suchergebnisse entsprechend den vermischten Trefferzahlen der informativen Werte sowie der original gereihten Werte. Auf der Blog-Suchmaschinen Webseite nach [NiXi07] hat der User mittels einer slide bar die Möglichkeit, die Suchanfrage auf informative oder affektive Artikel zu fokussieren. Die Position dieser slide bar korrespondiert mit dem Wert des Parameters λ. Befindet sich die bar in der Mitte wird λ auf Null gesetzt. Befindet sie sich allerdings im linken oder rechten Teil, so wird λ auf den Bereich (0,1] und den Bereich [-1,0) gesetzt. Folgende Gleichung errechnet die vermischten Trefferzahlen:

$$S_{mixed} = \lambda * S_{if} + (1 - |\lambda|) * S_{origin}$$

S_{if} .....Zuversichtswert

S_{origin} .....Originalwert um die Suchergebnisse nach ihrer Relevanz oder Datum zu sortieren

λ .....Parameter um die Trefferzahlen der Suchergebnisse für eine Neuordnung neu zu berechnen

Die Demoversion dieser Suchmaschine kann unter der URL „http://infoset.apexlab.org“ erreicht werden. In dieser Version ist jedoch nur der Relevanzwert berücksichtigt.

Über die Ergebnisse der Untersuchungen nach [NiXi07] lässt sich folgendes sagen: Es wurde die Performanz von drei Klassifizierungs-Algorithmen analysiert. Unter Anwendung auf den verwendeten Datensatz (3309 informative sowie 3547 affektive Artikel) stellte sich heraus, dass der SVM-Algorithmus die anderen beiden Algorithmen in allen Messungen übertraf. Tabelle 3.2 visualisiert die unterschiedlichen Performances dieser drei Algorithmen. F1 ist der harmonische Durchschnitt der Precision und des Recalls. In diesem

Experiment gibt der Macro Durchschnitt eine gleiche Gewichtung zu jeder Kategorie und der Micro Durschnitt gibt eine gleiche Gewichtung zu jedem Dokument.

	Precision	Recall	MicroF1	MacroF1
NB	0.890	0.841	0.864	0.852
SVM	0.922	0.910	0.918	0.915
Rocchio	0.860	0.727	0.772	0.730

Tabelle 3.2 Performanzmessung der drei Klassifizierer [NiXi07]

Unter Verwendung der Support Vector Machine sowie des Information Gain konnten [NiXi07] weiters herausfinden, dass sich eine hohe Anzahl an Attributen nicht negativ auf die Performanz auswirkt, sondern die Effizienz signifikant verbessert wird. Darüberhinaus stellen [NiXi07] eine „vorhabengesteuerte" Suchmaschine dar, bei welcher die Suchergebnisse neu geordnet werden in Bezug auf ihren affektiven oder informativen Gehalt. Diese Suchmaschinen-Technik ist sowohl auf die Rückgewinnung von Artikel der gesamten Blogosphäre oder einer einzelnen Domäne anwendbar. Der Prozentsatz der informativen Artikel eines Blogs wurde verwendet, um die Qualität dieses Blogs, welche dabei helfen kann um nach hoch-qualitativen Blogs zu suchen, zu messen.

3.2.2 Die Benutzung von Weblog-Korpora zur Emotionsklassifikation

Mit steigender Popularität von Blogs setzte sich immer mehr der Wunsch der Blogger durch ihre Gefühle innerhalb eines Artikels (durch sogenannte Emoticons beispielsweise) auszudrücken. Blog-Einträge, welche diese Emoticons beinhalten, stellen einen sehr bequem nutzbaren Korpus in Bezug

auf die Verwendung als Klassifikation von Blog-Einträgen in emotionale Kategorien dar.

[Yang07a] greifen in ihren Untersuchungen von Weblog-Inhalten ebenfalls auf den SVM-Algorithmus sowie zusätzlich auf CRF (Conditional Random Field) Maschinenlern-Techniken zurück. Um die Effektivität statistischer Lernmethoden für die Emotionsklassifikation zu evaluieren, benötigen auch [Yang07a] sowohl einen Trainings-Datensatz als auch einen Test-Datensatz. Das Framework nach [Yang07a] kann durch Abbildung 3.7 folgendermaßen grafisch dargestellt werden: Es werden Blog-Einträge gesammelt, welche als Test- und Trainings-Datensätze dienen. Eine Lexikon-Erzeugungsmethode baut ein Emotionslexikon auf. Dieses bildet die fundamentalen Attribute für die Emotions-Klassifikation auf dem Satzlevel. Support Vector Machine Klassifizierer weisen den Sätzen jeweils eine Emotions-Kategorie zu. CRF zieht den Kontext (eine Sequenz von Sätzen) in Betracht, um eine geeignete Emotions-Kategorie zu bestimmen. Die trainierten Klassifizierer werden anschließend auf die Dokumente angewandt. Weiters werden Heuristiken vorgeschlagen, um die Emotion des Autors zu bestimmen.

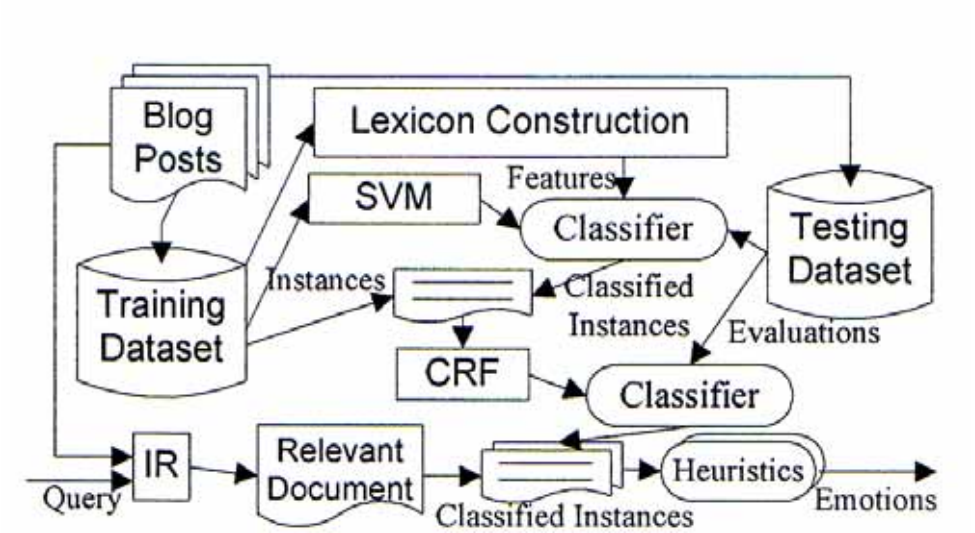

Abbildung 3.7 System der Emotionsanalyse nach [Yang07a]

Zu Beginn des Experimentes nach [Yang07a] werden Blog-Einträge gesammelt (der Datensatz besteht aus ca. 5 Millionen Blog-Einträgen), welche in weiterer Folge als Trainings-beziehungsweise Testdatensätze dienen. Die Emoticons (in etwa 14 Prozent von allen gesammelten Einträgen beinhalten markierte Emoticons) dienen als Indikatoren für die Emotionen, welche den Sätzen zugeordnet sind. In dem Experiment wurden zwei große Emotions-Kategorien – positive und negative – verwendet, wobei beide Kategorien in zwei weitere Subkategorien („fröhlich" und „freudig" sowie „verärgert" und „traurig") unterteilt sind. Die Emotions-Klassifikation geschieht folgendermaßen: Ein Satz S besteht aus n Termen $t_1, t_2, ..., t_n$. Eine Attributmenge FS für S ist definiert als $FS = \{t_k \mid t_k, 1 \le k \le n, t_k \in Lex\}$ (das Lexikon Lex wurde gebaut nach [Yang07b)]. Ein Emotions-Zuweisungs-Prozess wird in das Klassifikationsproblem transformiert: $S \xrightarrow{extraction} FS \xrightarrow{classification} \hat{e} \in \{e_1, ..., e_n\}$. Eine Attributmenge, welche aus einem Satz selektiert wird, wird auf eine Emotions-Kategorie abgebildet. Um die Emotions-Klassifizierer auf der Satzebene zu trainieren, wurde eine Version der SVM (LIBSVM) gewählt. Zusätzlich zur Auswahl einer Attributmenge aus einem Satz, wurde die Information aus nahegelegenen Sätzen miteinbezogen. Der CRF kontextbasierte Emotions-Klassifizierer gestaltet sich folgendermaßen: Um die Emotion e_t eines Satzes S_t zu bestimmten, berücksichtigt der neue Klassifizierer den benachbarten Kontext (Attribute und Emotionen von S_{t-1}). Das Klassifikationsproblem mit einem Kontext kann modifiziert werden zu: $S_1, ..., S_n \xrightarrow{extraction} FS_1, ..., FS_n \xrightarrow{classification} \hat{e}_1, ..., \hat{e}_n$.

Unter der Annahme von n Sätzen zur gleichen Zeit, werden die entsprechenden Emotionen zur selben Zeit zurückgeliefert. [Yang07a] definieren drei Kriterien um ein Dokument in einer emotionalen Kategorie zu klassifizieren:

- die Zuordnung der Emotion, welche am häufigsten in klassifizierten Sätzen aufscheint
- die Zuordnung der Emotion, welche in der längsten Serie an klassifizierten Sätzen, die die gleichen Emotionen aufweisen, vorkommt
- die Zuordnung der Emotion des zuletzt klassifizierten Satzes

Als Beispiel für die einzelnen Design-Kriterien führen [Yang07a] folgendes an: Ein Dokument könnte als „glücklich" bezeichnet werden da es aus mehreren „glücklichen" Sätzen besteht. Ein weiterer Grund dieses Dokument als „glücklich" einzustufen besteht darin, dass seine Betonung durchgehend auf „glücklichen" Sätzen liegt. Zu guter Letzt drückt der Autor durch ein „glückliches" Emoticon innerhalb seiner Abschlussstellungnahme seine „glückliche" Stimmung aus.

Zusätzlich zu der SVM und der CRF Methode, wurde ein Bayes-Klassifizierer verwendet, um einen Performanzvergleich mit der SVM durchführen zu können. Dieser Klassifizierer gebraucht eine bedingte Wahrscheinlichkeit eines jeden Wortes aus dem gebauten Emotions-Lexikon. Der Klassifizierer auf der Dokumentebene wird wie folgt definiert: Nach der Verarbeitung von 2586 Testdokumenten mittels des SVM Klassifizierers auf der Satzebene sowie dem CRF Klassifizierer auf der Kontextebene, enthält jedes Dokument eine Emotionsmarkierung an möglichen Sätzen. Die drei Kriterien nach [Yang07a] werden angewandt und für jedes Dokument wird eine Emotions-Kategorie vorgeschlagen.

[Yang07a] konnten folgende Ergebnisse aus ihrem Experiment gewinnen: Es wurde eine unterschiedliche Anzahl an Emotions-Schlüsselwörtern als Attribute verwendet, um die Emotions-Klassifizierer zu trainieren. Die Klassifizierer wurden auf den Testdatensatz angewandt. Es erfolgten vier Experimentreihen, von denen die ersten beiden (es wurden die beiden groben

Emotionskategorien „positiv“ und „negativ“ angewandt) zu nachfolgenden Resultaten führten: Unter der Verwendung von 50, 100 und 150 Emotionswörtern als Attribute stellte sich heraus, dass der SVM Klassifizierer beinahe in allen Fällen (Precision, Recall und F-Score) den Bayes-Klassifizierer übertraf. Einer Erhöhung der Anzahl an Schlüsselwörtern, welche als Attribute verwendet wurden, folgte ein gradweiser Rückgang der Precision. Die Recall- und F-Score Werte stiegen allerdings an. Es konnte weiters gezeigt werden, dass der CRF Klassifizierer bessere Ergebnisse liefert, wenn die Emotions-Kategorie des letzten Satzes verwendet wird. Die restlichen beiden Experimentreihen (hier wurden feinere Emotions-Kategorien mittels der vier, eingangs erwähnten, Emotions-Kategorien verwendet) lieferten jene Ergebnisse: Die maximale Attributgröße wurde auf 500 erhöht und es stellte sich heraus, dass dadurch sowohl die Performanz des SVM als auch des Bayes-Klassifizierers begünstigt wurde. Allerdings erzielt der Bayes-Klassifizierer unter der Integration von mehr als 150 Schlüsselwörtern als Attribute bessere Leistungen als der SVM-Klassifizierer. Darüber hinaus konnten [Yang07a] aufzeigen, dass CRF dann bessere Ergebnisse als SVM erzielt, wenn die Emotionen der vorhergehenden Sätze bekannt sind. Dem CRF-Klassifizierer ist es möglich, die Überleitung von Emotions-Kategorien von einem Satz zum nächsten zu lernen, während der SVM-Klassifizierer diese Lerneigenschaft nicht so gut beherrscht. Emotionelle Informationen von nahegelegenen Sätzen beeinflussen sich gegenseitig. Aus diesem Grund wäre ein kontextbewusster Klassifizierer sinnvoll.

3.2.3 Ein Genre-Analysemodell für Weblogs

Die Studie nach [Herr04] basiert auf der Voraussetzung, dass wiederkehrende elektronische Kommunikationspraktiken sinnvollerweise als Genres bezeichnet werden können. Darüber hinaus stützen sich [Herr04] in ihrer Untersuchung auch auf die Charakterisierung nach [Swal90], dass Genres als eine Klasse an kommunikativen Vorkommnissen, welche eine gemeinsame Menge an kommunikativen Zwecken, ähnlichen Strukturen, Inhalten und angestrebtem Leserkreis vorweisen, bezeichnet werden. Ausgehend von diesen Kriterien stellen Weblogs bis auf Widerruf insofern ein gutes Beispiel für einen Genrestatus dar, als dass sie einerseits benannt sind und auch andererseits dazu tendieren, alltägliche Strukturen und Substanzen aufzuweisen. [Herr04] versuchen im Rahmen ihrer Forschungsarbeit die Eigenschaften des „aufstrebenden" Blog-Genres ([Crow97] bezeichnen eine persönliche Homepage als ein Beispiel für ein „aufstrebendes" Web-Genre im Vergleich zu „reproduzierten" Genres an) zu charakterisieren und es in Hinsicht auf Offline-Genres [Eric02] zu platzieren. Eines der Hauptziele besteht darin, einen Beitrag darüber zu leisten, wie technologische Änderungen die Formation von neuen Genres, welche möglicherweise die Genre-Ökologie [Eric02] einer größeren Domäne wie beispielsweise dem Internet, einleiten können. Ausgehend von dem Klassifizierungsmodell nach [Kris02], welches Weblogs in vier verschiedene Basis-Gattungen einteilt, identifizieren [Herr04] Sub-Typen aus einer zufälligen Stichprobe an Webseiten, die sich selbst als Blogs bezeichnen. Abbildung 3.8 zeigt schematisch das Klassifizierungsschema nach [Kris02]. Es wird eine Klassifikation von Blogs in vier Basistypen (persönlich und thematisch sowie Individuum und Gemeinschaft) vorgeschlagen.

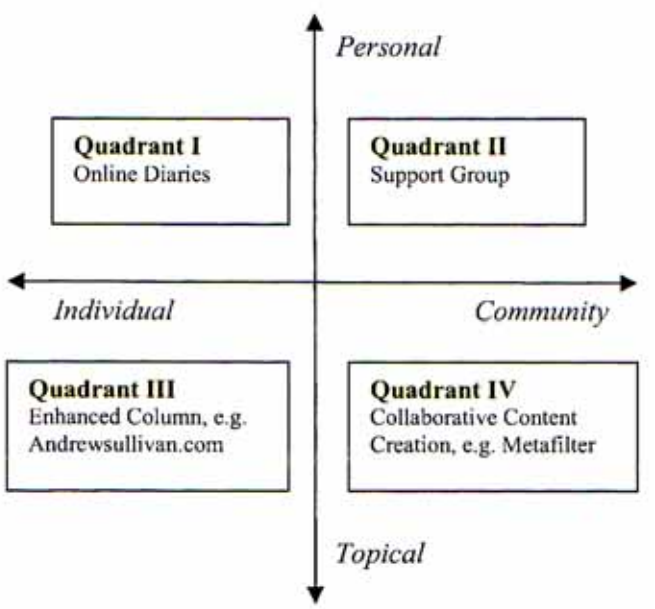

Abbildung 3.8 Typen von Weblogs nach [Kris02]

Ein weiteres Untersuchungsfeld innerhalb der Studie nach [Herr04] ergibt sich aus der Frage ob Blogs ein neu aufstrebendes oder ein reproduziertes Genre verkörpern. Um einen ausreichend großen Datensatz zu gewährleisten basiert die Studie auf einer zufälligen Stichprobe aus 203 Blogs, welche mit Hilfe der randomisierenden Funktion der Blog-Tracking Webseite „blog.gs“ angesammelt wurde. Um einen kohärenten Korpus zu schaffen, wurden einige dieser Blogs aus diversen Ausschließungsgründen weggelassen. Innerhalb der Analyse-Methodik setzten sich [Herr04] mit sogenannten Inhaltsanalysen (diese untersuchen den Inhalt von Blog-Einträgen und Kommentaren) [Baue02] auseinander, um strukturelle und funktionale Eigenschaften im Blog-Korpus zu identifizieren sowie zu quantifizieren. Die Kodierungs-Kategorien nach [Herr04] wurden durch mehrfache Instrumente bestimmt: Um das Genre innerhalb einer Benutzer-Community zu situieren, wurde nach demographischen Eigenschaften der Blog-Autoren (erste Bedeutung) in jenem Ausmaß kodiert, dass diese durch die Blogs an sich bestimmt werden können.

Weiters wurde in Betracht gezogen, wieviel an Information über den Blog-Autor in einem Blog enthalten ist. Dies wurde deshalb angestrebt, um einen Vergleich mit persönlichen Homepages, in welchen die Identität des Seiten-Eigentümer typischerweise im Mittelpunkt steht, ziehen zu können. Aufgrund dessen, dass der Zweck eines Blogs (zweite Bedeutung) ein Schlüsselkriterium darstellt, wurde nach dem gesamten Zweck eines Blogs kodiert. [Herr04] kodierten nach Filterblogs[1], persönlichen Journalblogs[2], K-logs[3], nach Blogs mit gemischten Zwecken[4] und nach restlichen Blogs.
Weiters wurden strukturelle Analysen von Blogs (dritte Bedeutung) durchgeführt. Die ausgewählten strukturellen Attribute wurden nach früheren inhaltsanalytischen Forschungen wie die Anzahl an Links, Bilder, Bestehen einer Suchfunktion und Werbungen, über Web-Genres bearbeitet.
Abschließend wurde nach drei Typen an temporären Informationen (vierte Bedeutung) kodiert, um Behauptungen über die Häufigkeit, mit welcher Blogs aktualisiert werden, zu evaluieren und um das durchschnittliche Alter von Blogs, welche zum Zeitpunkt der Studie im Internet verfügbar waren, zu bestimmen. Die drei Typen an temporären Informationen sind: Aktualisierungsintervall (zwischen dem jüngsten und dem vorhergehenden Eintrag), Alter des Blogs (bestimmt durch das Datum des ältesten Eintrags auf der Blog-Seite) und Neuheit der Aktualisierung (in Relation zum Zeitpunkt der Datensammlung).
Die Ergebnisse aus dieser Analyse, welche im Folgenden präsentiert werden, unterstützen einerseits vorangegangene Behauptungen über Blogs, rücken diese jedoch in anderen Zusammenhängen in ein völlig anderes Licht.

- Eigenschaften der Blog-Autoren: Um ein Genre innerhalb einer Benutzer-Community zu situieren, forschten [Herr04] nach

[1] Autoren verlinken auf und kommentieren den Inhalt anderer Webseiten

[2] Autoren berichten über ihr Leben sowie über ihre inneren Gefühle und Gedanken

[3] Die Abkürzung „K-Logs“ steht für Knowledge Management Weblogs. Häufig handelt es sich um Themenblogs.

[4] Gemischte Blogs vereinen die Funktionen von zwei oder mehreren dieser drei Blog-Typen

demographischen Eigenschaften der Blog-Autoren. Neben anderen wichtigen Erkenntnissen stellte sich heraus, dass Eigenschaften der Blog-Autoren - entgegengesetzt der anfänglichen Untersuchungen - sich nicht signifikant von den Demographien anderer öffentlicher Kommunikationsprotokolle im Internet (wie zum Beispiel Foren oder persönliche Webseiten) unterscheiden [Herr04]. Tabelle 3.3 gibt einen Überblick über die Eigenschaften von Blog-Autoren. Die Häufigkeit (engl. frequency) bezieht sich auf die insgesamt 203 Blogs, welche als Test-Datensatz verwendet wurden. Der Prozentwert (engl. percentage) wird aus der Gesamtanzahl an Individuen oder Blogs, bei welchen es möglich ist, nach einer Kategorie zu kodieren, berechnet. Ausgenommen sind jedoch „unbekannte" und andere problematische Instanzen (diese Berechnung des Prozentwertes trifft auch auf nachfolgende Tabellen zu).

Characteristic	Frequency	Percentage
One author	196	90.8
Male	110	54.2
Adult (20 years or older)	115	59.6
Student	73	57.5
Located in USA	104	69.8
Name on first page (other than pseudonym)	127	67.6
Other personal information on first page	108	54.0
Graphical representation on first page	34	17.5

Tabelle 3.3 Eigenschaften der Blog Autoren [Herr04]

Weiters konnten [Herr04] feststellen, dass sehr viele Blogger explizite persönliche Informationen auf der Startseite ihres Blogs integrieren. Ungefähr 92 Prozent der Blogger geben ihren Namen bekannt und in etwa 31 Prozent verwenden ihren vollen Namen. Ebenfalls interessant ist

die Beobachtung, dass das Geschlecht der Blog-Autoren je nach Alter sehr stark variiert. Unter den Bloggern (vorhandenes Wissen über das Geschlecht), welche als „Erwachsene" klassifiziert wurden, sind 63 Prozent männlich. Umgekehrt macht eine Mehrheit an „weiblichen Teenager" Blogger rund 58 Prozent aus.

- Zielsetzung des Blogs: Die Verteilung der verschiedenen Blog-Typen gestaltet sich im Zuge der Analyse nach [Herr04] so, dass in etwa 70 Prozent dem persönlichen Journal-Type entsprechen. Tabelle 3.4 stellt die Verteilung der verschiedenen Blog-Typen zusammenfassend dar. Vier von den insgesamt 203 getesteten Blogs konnten in Hinblick auf ihren Zweck nicht klassifiziert werden. Deshalb beträgt die Gesamtanzahl an Blogs, welche in dieser Kategorie analysiert wurden, 199.

Type	Frequency	Percentage
Personal journal	140	70.4
Filter	25	12.6
K-log	6	3.0
Mixed	19	9.5
Other	9	4.5
	199	100

Tabelle 3.4 Verteilung von Blog-Typen nach ihrem primären Zweck [Herr04]

Zusätzlich zu dieser Verteilung von Blog-Typen müssen die Geschlechts- und Altersunterschiede in Betracht gezogen werden. Persönliche Journalblogs werden von Bloggern allen Altersstufen sowie von beiden Geschlechtern verfasst. Konträr dazu konzentrieren sich ausschließlich männliche Erwachsene auf Filterblogs, K-Logs und gemischte Blogs.

- Temporäre Informationen: Blogs aus dem Testdatensatz nach [Herr04] wurden allesamt zwei Wochen vor der Datensammlung, entsprechend zu den Sammelkriterien, aktualisiert. Eine sehr repräsentative Messung ist die durchschnittliche Anzahl an Tagen, zwischen jüngsten und dem davor jüngsten Eintrag. Tabelle 3.5 ist, neben den anderen Ergebnissen aus diesen temporären Messungen zu entnehmen, dass der durchschnittliche Blog innerhalb der Studie nach [Herr04] 163 Tage alt ist und der älteste Blog bereits seit 990 Tagen existiert. Weitere 16 Prozent sind bereits älter als ein Jahr und insgesamt fünf Prozent weisen ein Alter von zwei Jahren auf.

Measure	Mean (days)	Mode (days)	Range (days)
Recency of update at time of data collection	2.2	0	0-11
Interval between two sequential entries	5.0	1	0-63
Age of blog	163.0	n/a[11]	0-990

Tabelle 3.5 Zeitliche Messungen nach [Herr04]

Diesen Zahlen liegt die Vermutung nahe, dass die Aufrechterhaltung eines Blogs eine nicht ganz triviale zeitliche Verbindlichkeit für viele Autoren darstellt. Persönliche Blogs zum Beispiel kommen über eine Dauer von sechs Monaten häufiger vor als andere Blog-Typen.

- Strukturelle Charakteristiken: Anhand dieser vierten Kodierungskategorie konnten [Herr04] ebenfalls einige wesentliche Aussagen im Rahmen ihrer Genre-Analyse treffen. Blogs weisen im Gegensatz zu anderen Genres weniger Bilder auf und nutzen generell

gesprochen das multimediale Webpotential vergleichsweise nicht sehr stark. Tabelle 3.6 zeigt die Häufigkeiten von strukturellen Attributen, von welchen angenommen wurde, dass sie charakteristisch für Blogs sind. Entgegengesetzt den anfänglichen Empfindungen nach [Herr04], dass sich ein Kalender in einer Sidebar auf der Startseite eines Blogs als ein typisches Blog-Attribut erweist, stellt es sich als weniger häufig (13 Prozent) heraus. In der gleichen Weise verhält es sich bei dem Attribut, welches der Leserschaft die Möglichkeit einräumt einen Eintrag zu kommentieren (43 Prozent).

Feature	Frequency	Percentage
Archives	139	73.5
Badges	138	69.0
Images	133	58.6
Comments allowed	85	43.0
Link to email blog author	63	31.3
Ads	48	25.1
Search function	35	18.5
Calendar	25	13.0
Guest book	9	4.5

Tabelle 3.6 Strukturelle Attribute nach [Herr04]

Bezüglich der Verlinkungen zu anderen Webseiten oder Blogs kristallisierte sich im Laufe der Studie heraus, dass einige Blogs zwar sehr viele Links beinhalten, das Ausmaß mit dem Blogs zu anderen Inhalten verknüpfen ist jedoch nicht so stark ausgeprägt wie es in vielen früheren allgemeinen Untersuchungen der Fall ist.

Abseits dieser Verlinkungsstrukturen wurden aufgrund der Charakterisierung von Blog-Einträgen jeweils die jüngsten Einträge aus der Datensammlung nach [Herr04] detailliert untersucht. Hierbei konnte festgestellt werden, dass der sogenannte „Header“ hauptsächlich Informationen über das Datum und den Titel des Eintrages enthält. Der

„Footer“ hingegen trägt typischerweise den Zeitpunkt des Eintrages, den Autorennamen und zu guter letzt Permalinks. Eine Verlinkung, um Kommentare zu schreiben beziehungsweise diese zu lesen, findet sich gewöhnlicherweise ebenfalls im „Footer“. In diesem Zusammenhang wurde von [Herr04] festgestellt, dass im Allgemeinen weitaus weniger Leser Kommentare zu Blog-Einträgen abgeben. Dies widerspricht somit den früheren Behauptungen über die Blog Interaktivität und Blog-Community. Tabelle 3.7 beschreibt die Arten von Informationen welche im „Header“ und im „Footer“ des jeweiligen Eintrages enthalten sind. Die letzte Zeile dieser Tabelle weist darauf hin, dass der durchschnittliche Eintrag in der Sammlung nach [Herr04] .3 Kommentare enthält und dass die Mehrheit an Einträgen keinen Kommentar empfangen hat.

Information contained		Frequency	Percentage
Header		331	99.5
	date	176	93.6
	title	84	44.7
	time	30	16.0
	author's name	21	11.2
Average number of header features per blog		*1.8*	
Footer		481	92.0
	time	148	78.7
	author's name	121	64.4
	internal links	109	58.0
	comments	61	32.4
	date	22	10.6
Average number of footer features per blog		*2.6*	
Number of comments per entry	mean .3	mode 0	range 0-6

Tabelle 3.7 "Header" und "Footer" der jeweiligen Einträge [Herr04]

Finale Observierungen durch [Herr04] spielen sich auf der Textebene von Blog-Einträgen ab. Der durchschnittliche Blog weist in etwa 210 Wörter auf und ist somit um einiges kürzer als eine Antwort email auf einer akademischen Diskussionsliste. Ein einzelner Satz in einem Blog-

Eintrag besitzt im Durchschnitt 13 Wörter und enthält demzufolge etwa drei Wörter weniger als private email Nachrichten auf einer Hochschule. Tabelle 3.8 fasst die finalen Messungen bezüglich der Struktur auf der Textebene von Blog Einträgen zusammen.

Measure	Total	Avg	Range
Words	42930	210.4	1-1262
Sentences or fragments	3260	16.0	1-117
Words per sentence		13.2	
Paragraphs	709	3.5	0 - 21
Words in quotations	3681	18.0	0 - 430
Quoted sentences/fragments	468	2.3	0 - 40
Quoted words per sentence		7.9	

Tabelle 3.8 Textmessungen im "Body" der Einträge [Herr04]

Zum Abschluss der Untersuchung griffen [Herr04] die Frage nach den Ursprüngen beziehungsweise der Herkunft von Blogs auf. Diverse Entwicklungen (Weblogs werden typischerweise mehrere Male pro Woche aktualisiert, Blogs können multimediale Inhalte integrieren, Weblog-Autoren behalten sich die äußerste Kontrolle über Bloginhalte ein, etc.) deuten darauf hin, dass Blogs nicht nur einer einzelnen Quelle entspringen sondern eine Kombination aus existierenden Genres darstellen. Anhand folgender Dimensionen: Aktualisierungshäufigkeit, Symmetrie des Kommunikationsaustausches und Multimodalität, können Weblogs als intermediäre Erscheinung zwischen Webseiten und asynchronem CMC (Computer Mediated Communication) angesehen werden. Abbildung 3.9 stellt schematisch dar, wie Weblogs die technologische Lücke zwischen Standard-Webseiten und CMC auffüllen.

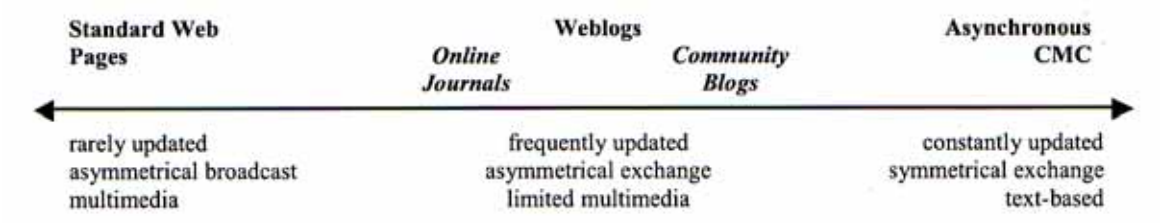

Abbildung 3.9 Weblogs als Kontinuum zwischen Webseiten und CMC [Herr04]

Diese intermediäre Charakteristik macht Blogs insofern sehr interessant für Benutzer, da sie in ihrer Rolle als Autoren einerseits Erfahrungen über soziale Interaktion sammeln können und andererseits die Kontrolle über diesen kommunikativen „Raum" inne haben. Als mögliche Konsequenz im Falle von Weblogs könnte das Potential einer Neuformung der Genre-Ökologie im Internet angesehen werden [Herr04].

3.3 Analysemodelle als Folge von Blog-Communities

Die starke Interaktion der Blogger untereinander sowie der temporale Verlauf von Blog-Artikel differenziert Weblogs eindeutig von anderen „Online-Plattformen“ [Qamr06]. Innerhalb der Blogosphäre kommt es oftmals zur Entstehung von dynamischen Communities, da Blogger sehr häufig gegenseitig auf ihre Einträge verlinken. Diese Communities können auch als soziale Netzwerke betrachtet werden, in welchen regelrecht emotionale Bindungen zwischen einzelnen Blogger-Gruppierungen heraus kristallisiert werden können. Zahlreiche Wissenschaftler machen sich das Vorhandensein der Blog- Communities zu Nutze um auf recht unterschiedliche Art und Weise Zusammenhänge zwischen Blog-Artikel beziehungsweise zwischen den einzelnen Bloggern zu erkennen.

3.3.1 Entdecken von Communities

Die Zielsetzung in der Studie nach [Zhou06] besteht darin, Weblog-Communities aus der Perspektive der sozialen Netzwerkanalyse zu betrachten.
In dem Modell von [Zhou06] erfolgt die Informationswiedergewinnung durch das Hervorheben der Intensität sozialer Verbindungen zwischen Weblogs. [Zhou06] definieren Blog-Communities als spezielle Typen von sozialen Netzwerken, innerhalb derer verschiedene Gruppen von Leuten unter bestimmten Beziehungsverhältnissen miteinander verbunden sind. Beispiele von solchen sozialen Netzwerken stellen Freundschaftsnetzwerke von Hochschulstudenten oder wissenschaftliche Mitverfasser-Netzwerke von Akademikern dar. Diese Gruppen werden auch als Akteure und die Beziehungen als Verbindungen bezeichnet. Im Falle von Weblog-Communities stellen Blogger die Akteure dar und die Querverweise (Hyperlinks) zwischen den Weblogs können als Verbindungen gesehen werden. Darüberhinaus

bildet jeder dieser Hyperlinks eine Kommunikationsinstanz zwischen zwei Bloggern. Die Anzahl an Instanzen wiederum deutet auf die Festigkeit der Verbindungen zwischen diesen Bloggern hin. Innerhalb einer ausgeprägten Community sollten Blogger feste Verbindungen mit anderen Mitgliedern aufweisen. Den Ausgangspunkt der Weblog-Community-Identifizierung bildet ein verbundener Blog-Raum mit relativ ähnlichen Themen. [Zhou06] entwickelten für ihre Untersuchungen einen speziellen Weblog-Crawler um die Blog-Räume zu konstruieren. Der Weblog-Crawler betrachtet eine Weblog-URL als „Saat" und fügt inkrementell Verbindungen zwischen den Weblogs aus der Sammlung hinzu. Dieser Basis-Algorithmus generiert eine sternförmige Darstellung mit der „Saat" im Mittelpunkt. Innerhalb dieses „künstlich" geschaffenen Blog-Raumes finden sich einige tausend einzigartiger Blogs, aus denen die Forschergruppe eine Community-Struktur ableiten möchte. Um die direkte Darstellung zu konstruieren, wurden zuerst vielfache Linien zwischen den beiden Eckpunkten entfernt und der Linienwert wurde als die Anzahl an Linien bestimmt. Dies kann als die Festigkeit der Verbindungen zwischen Weblogs interpretiert werden.

Um die verschiedenen Blog-Communities zu extrahieren, wurde der Insel-Partitionierungs-Algorithmus nach [Bata03] herangezogen. Eine Insel wird definiert als ein verbundenes kleines Sub-Netzwerk der Größe [*min*, *max*]. Dieser Algorithmus stellt eine hierarchische Clusterbildungsmethode dar. In der sortierten Reihenfolge verschmilzt dieser Algorithmus Knoten, um Sub-Netzwerken zu formieren, basierend auf gewöhnlichen Eckpunkten von Sub-Netzwerken. Alle Sub-Netzwerke bilden eine hierarchische Struktur mit möglicherweise einer gemeinsamen Wurzel. Im Rahmen der Fallstudie nach [Zhou06] wurden Weblogs von „Savas Parastadist" als „Saat" verwendet, um einen verbundenen Blog-Raum zu erhalten. Darin enthalten sind rund 3800 Weblogs mit über 33000 Linien. [Zhou06] erzielten folgende Ergebnisse innerhalb ihrer Studie: Zur Durchführung der Partitionierung und der Visualisierung wurde [Paje08] verwendet. 15 Communities mit jeweils eigenen

Thematiken konnten aus dem Blog-Raum identifiziert werden. Bei einer der untersuchten Communities handelt es sich um „general web services". Die Mehrheit an gefundenen Community-Mitglieder konnte durch gebräuchliche Suchmaschinen wie *GoogleBlogSearch*, *blogdex*, *daypop*, etc. nicht gefunden werden.

3.3.2 Ein Modell aufgrund des Social Hypertext

Das Modell nach [Chin06] untersucht miteinander verbundene Blogs als eine Form des social hypertext (das World Wide Web wird von Thomas Erickson als social hypertext bezeichnet, da persönliche Webseiten nicht nur nützliche Informationen transportieren sondern vorwiegend die Identität einer Person porträtieren [Eric96]) und skizziert Techniken für das Messen der Festigkeit einer Community innerhalb von Blogs. Weblogs sind in einem derartigen social hypertext einerseits durch explizite Links von einem Blog zu einem anderen eingebunden und andererseits durch Kommentare, welche von einem Blog-Autor an einem Eintrag eines anderen Blogs gemacht werden (diese können als Links interpretiert werden). [Chin06] verfolgen mit ihren Untersuchungen folgende Zielsetzungen: Als erstes wird ein Modell zur Entdeckung von Communities in Blogs vorgestellt, welches Blog-Verhalten in Betracht zieht und darüber hinaus Analysen und social network Analysen (oder nur Netzwerk-Analysen genannt) verbindet. Als zweites wird versucht, Verhaltensansätze, um Communities zu studieren, mit Netzwerk-und Verlinkungsansätzen zu verbinden. Dies soll es ermöglichen, Community-Strukturen zu kalibrieren, welche dazu verwendet werden können, um automatisch Communities in anderen Blogs zu finden und diese auch zu messen, jedoch ohne der Verwendung von Umfragen. Als drittes wird eine Messung zur Entdeckung von Communities vorgeschlagen. Diese Messung verwendet Netzwerk-Zentralität als eine Messung aus den social network Analysen.

Die nachfolgende Abbildung 3.10 illustriert sehr anschaulich, welche einzelnen Schritte von [Chin06] in ihrer Fallstudie durchgeführt wurden, um Blog-Communities in einem Set an zugehörigen Blogs zu evaluieren. Um Communities in Blogs ausfindig zu machen werden zum einen das Community-Gefühl (engl. sense of community, SOC) nach [Mill86][5] und zum anderen netzwerkanalytische Zentralitätsmessungen nach [Free78] verwendet. Der erste Schritt innerhalb der Methodik nach [Chin06] umfasst das Auffinden von zu untersuchenden Anwärter-Blogs nach möglichen Community-Anzeichen. In Schritt Nummer zwei wurde ein Crawler eingesetzt um die Blogs zu durchforsten und Verlinkungen zwischen den Blogs mittels der Einträge und Kommentare aufzuzeichnen. Der dritte Schritt dient zur Visualisierung des sozialen Netzwerkes von den Anwärter-Blogs unter Verwendung der Visualisierungssoftware von [Paje08]. In Schritt vier des Methodik Kreislaufes nach [Chin06] findet eine Befragung über das „Community-Gefühl" statt. Innerhalb dieser Befragung werden den Bloggern Fragen über Gefühle: Gehört ein User zur Community oder kann sich der User durch diese identifizieren?, Wird ein User durch die Community beeinflusst oder beeinflusst er selbst die Community?, Empfängt ein User Unterstützung durch andere Community-Mitglieder oder besitzt der User einen gewissen Staus innerhalb der Community? und Empfinden User eine gemeinsame Verbindung mit anderen Usern innerhalb der Community?, bezüglich des „Community-Gefühls" gestellt. Die insgesamt 12 gestellten Fragen konnten in Form einer Fünfpunkt-Likertskala (zur Auswahl standen insgesamt fünf Antwortmöglichkeiten, welche in ihrem Spektrum von „Ich stimme ganz entschieden zu" bis hin zu „Ich stimme ganz und gar nicht zu" reichten)

[5] Im Rahmen der Messungen und Kalibrierungen von Communities kristallisiert sich das verbreitete Motiv heraus, dass Mitglieder einer Community ein „Community-Gefühl" wahrnehmen. Nach [Mill86] existieren vier Attribute des „Community-Gefühls": Gefühle über die Gruppen-Zugehörigkeit (engl. membership), Gefühle über den Einfluss in einer Gruppe (engl. influence), Gefühle über die Stärkung der Bedürfnisse (engl. reinforcement of needs) und Gefühle über geteilte emotionale Verbindungen (engl. shared emotional connection).

beantwortet werden konnten. Tabelle 3.9 stellt die Fragen innerhalb der „Community-Gefühl“ Umfrage dar.

Question code	Question
Q1	I think this blog is a good one to read
Q2	Readers of this blog do not share the same values
Q3	Other readers and I want the same thing from this blog
Q4	I can recognize the names most readers who post comments in this blog
Q5	I feel at home in this blog
Q6	Very few other readers of this blog know me
Q7	I care about what other blog readers think of my actions
Q8	I have no influence over what this blog is like
Q9	If there is a problem in this blog, there are people here who can solve it
Q10	It is very important to me to be a reader of this blog
Q11	Readers of this blog generally don't get along with each other
Q12	I expect to stay a reader here for as long as I can

Tabelle 3.9 Fragen der "Community-Gefühl" Umfrage [Chin06]

Schritt fünf besteht aus der Identifizierung möglicher Communities aufgrund von strukturellen Analysen, mit dem Schwerpunkt auf Zentralitätsmessungen.
Schritt sechs beruht auf der Synthese der netzwerkanalytischen Zentralitätsmessungen und den „Community-Gefühl“ Ergebnissen. Daraus lassen sich Blogs bestimmen, welche Teil einer Community sein könnten. Abschließend ist die Menge an identifizierten Blog-Communities in Schritt Nummer sieben als Resonanz für die Anwärter-Blogs vorgesehen und wird dazu benutzt, um Empfehlungen für neue Verlinkungen zwischen den einzelnen Blogs auszusprechen, zwecks des Anwachsens von Communities innerhalb des social hypertext.

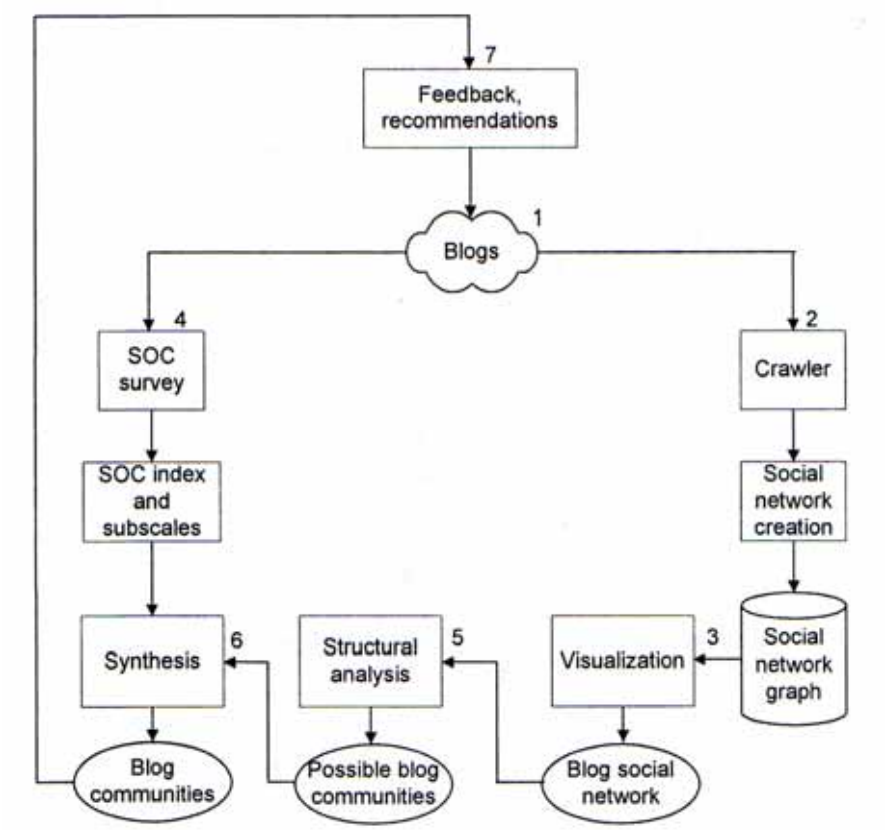

Abbildung 3.10 Methodik zur Identifizierung von Blog Communities [Chin06]

Diese Forschungsmethodik nach [Chin06] wurde auf eine Fallstudie eines unabhängigen Musik-Blogs angewandt. In den folgenden Abschnitten wird zu Beginn auf den ausgewählten Blog innerhalb der Studie eingegangen. Danach wird die Methodik beschrieben, mit welcher das soziale Netzwerk erstellt wurde. Weiters folgt eine Beschreibung der durchgeführten „Community-Gefühl" Umfrage. Als nächstes wird das „Community-Gefühl" von strukturellen Analysen – unter Verwendung des social hypertext Modells nach [Chin06] – identifiziert. Ausgehend von den SOC-Ergebnissen wird bestimmt, ob eine Beziehung zwischen dem „Community-Gefühl" Ergebnis für einen Blog und seiner Einbindung in jenes „Community-Gefühl", welches aus den strukturellen Analysen durch [Chin06] identifiziert wurde, besteht. Somit kann

die Umwandlung der Menge an möglichen Blog-Communities in bestimmte Blog-Communities verfeinert werden.

[Chin06] kreierten einen unabhängigen Musik-Blog (erster Schritt), um eine Community darin analysieren zu können. Um das Interesse und die Motivation der Community anzuregen, wurde ein media player, ein Photoalbum, ein Bewertungssystem und das „Lied der Woche" hinzugefügt. Das Bewertungssystem beispielsweise erlaubt es dem User Lieder nach einer Fünfpunkt-Skala, welche Besucher ermutigen ihre Beurteilungen innerhalb der Kommentare abzugeben und ein bestimmtes Lied als „Lied der Woche" auszuwählen, zu bewerten.

Um den Musik-Blog zu durchforsten wurde ein Blog-Crawler entwickelt (zweiter Schritt). Dieser durchkriecht den Blog auf zwei Unterteilungsgraden für „Eingangsverbindungen (Blogs welche zum Ziel-Blog verlinken)" sowie für „Ausgangsverbindungen (Blogs welche von dem Ziel-Blog aus verlinken)". Im Rahmen des „Crawl-Prozesses" wurde einerseits die Blog-URL des durchforsteten Blogs als auch andererseits die Blog-URL des Kommentators auf diesem Blog in dem Format (*durchforstete Blog-URL*, *URL des Kommentators*) aufgezeichnet. Anschließend wurde die Häufigkeit für jede Aufzeichnung berechnet und die Blog-URLs wurden anonymisiert. Das soziale Netzwerk (Schritt drei) wurde aus den Ergebnissen des Blog-Crawlers in UCINET[6] erstellt. Abbildung 3.11 illustriert die Visualisierung des Netzwerkes (beinhaltet 604 Blogs) bis auf zwei Separationsgrade aus dem Musik-Blog nach [Chin06].

[6] UCINET ist ein umfangreiches Programm für die Analyse von sozialen Netzwerken und anderen Näherungsdaten. Die Software beinhaltet Dutzende an Netzwerkanalyse-Routinen wie „Zentralitätsmessungen", „Positionsanalyse-Algorithmen", etc. [Ucin08]

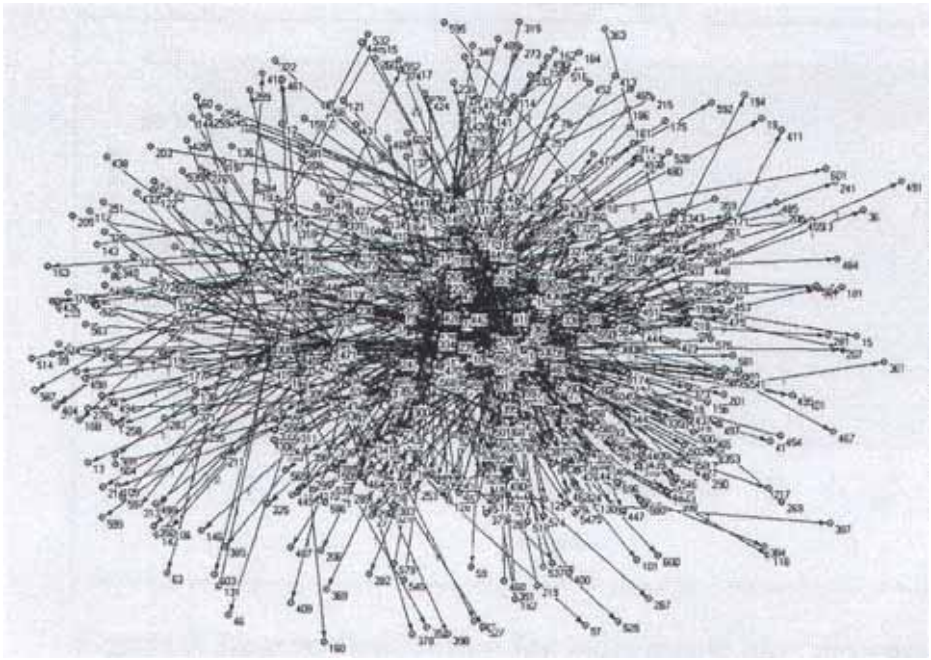

Abbildung 3.11 Visualisierung des Netzwerkes [Chin06]

Die Umfrage über das „Community-Gefühl“ (Schritt vier im Methodik-Kreislauf nach [Chin06]) wurde folgendermaßen durchgeführt: Die Umfrage wurde online auf der Musik-Blogseite zur Verfügung gestellt. 15 Personen füllten die Umfrage vollständig aus. Bei den Umfrageteilnehmer zeigte sich, dass die meisten zwischen 21 und 30 Jahren alt sind, die Mehrheit davon eher der Blog-Leserschaft als den Blog-Autoren angehört und keine regelmäßigen Kommentare abgegeben werden. Daraus lässt sich schließen, dass intensive Konversationen innerhalb des Netzwerkes, welches den Musik-Blog umgibt, nur sehr selten stattfinden werden. Dies wiederum bedeutet, dass es nur sehr wenige Communities geben wird. [Chin06] postulieren, dass die Communities klein und sehr stark fokussiert sein werden. Allerdings wird zugleich betont, dass die Stichprobengröße niedrig ist und keine konkreten Behauptungen über bestimmte Communities aufgestellt werden können.

Zur Visualisierung der Community-Struktur wurden strukturelle Analysen (Schritt fünf) durchgeführt. In diesem Zusammenhang wurden die Indikatoren aus dem social hypertext Modell nach [Chin06] angewandt: In diesem social hypertext Modell stellen sich [Chin06] die Frage, welche Eigenschaften und Indikatoren verwendet werden können, um ein „Community-Gefühl" innerhalb von Blogs zu entdecken. Ein möglicher Ansatz besteht darin, soziale Netzwerke zu visualisieren und dabei Strukturen zu bestimmen, welche auf eine Charakteristik aus dem „Community-Gefühl" deuten. Die daraus resultierenden Indikatoren werden als Visualisierungs-Indikatoren bezeichnet. Die Information, welche durch diese Visualisierungs-Indikatoren bereitgestellt wird, kann durch Indikatoren sozialer Netzwerkanalysen ergänzt werden. Tabelle 3.10 zeigt auf, wie „Community-Gefühl" Messungen mit den Visualisierungs-Indikatoren und den Netzwerkanalyse-Indikatoren verbunden werden können. Diese Tabelle kann bezüglich der Ausrichtung der Visualisierung und sozialen Netzwerkanalyse-Indikatoren mit den vier SOC-Eigenschaften als eine Erweiterung des Synthese-Schrittes (Nummer sechs des Methodik-Kreislaufs; siehe Abbildung 3.10) gesehen werden.

SOC characteristic	Visualization indicator	Social network analysis indicator
Membership	Reciprocal links with each directed link having frequency $\geq k$ Star network Hierarchical reduction Degree distribution	Degree centrality
Influence	Brokers or bridges	Betweenness centrality
Reinforcement of needs	Reciprocal links where each directed link has frequency $> m$ where m is the need threshold	Closeness centrality
Shared emotional connection	Triangles Completely connected graphs	k-cores

Tabelle 3.10 Social hypertext model nach [Chin06]

Die Anwendung des social hypertext Modells dient der Identifizierung von Strukturen, welche auf ein „Community-Gefühl“ in dem Netzwerk des Musik-Blog (siehe Abbildung 3.11) hinweisen. Für jede der vier Charakteristiken (siehe Tabelle 3.10) des „Community-Gefühls“ wurde der Visualisierungs-Indikator und der Indikator aus der sozialen Netzwerkanalyse, welche durch [Paje08] unterstützt wurden, angewandt, um Strukturen, welche wiederum auf ein „Community-Gefühl“ hinweisen, darzustellen. Bezüglich der „membership“ Eigenschaft stellte sich heraus, dass es schwierig ist einen Vorschlag darüber zu machen, ob sich eine Art von Community in diesem Netzwerk befindet. Aus diesem Grund wurde eine tief gehende Analyse unter Verwendung reziproker Verlinkungen[7] durchgeführt [Chin06]. Abbildung 3.12 stellt das Netzwerk für die Bestimmung von Zugehörigkeiten innerhalb des Musik-Blogs graphisch dar. Es wurden diejenigen Blogs visualisiert, welche einen direkten Link mit einer Mindesthäufigkeit von zwei haben. Mittels Aufsummierung der Häufigkeiten von jedem direkten Link in dem reziproken Link, wurde jeder direkte Link in einen reziproken umgewandelt. Jene Blogs mit null oder einem Nachbarn wurden entfernt. Daraus ergaben sich 54 Knoten mit dem Musik-Blog nach [Chin06] als Knoten 29 in der Mitte.

[7] Unter einem reziproken Link ist folgendes Beispiel zu verstehen: Ein Eintrag von Blogger A wird von Blogger B kommentiert (dies ist durch den Link A->B gekennzeichnet) und Blogger B schreibt einen Eintrag in Blog B, welcher von Blogger A kommentiert wird (dies wiederum wird durch den Link B->A angegeben.

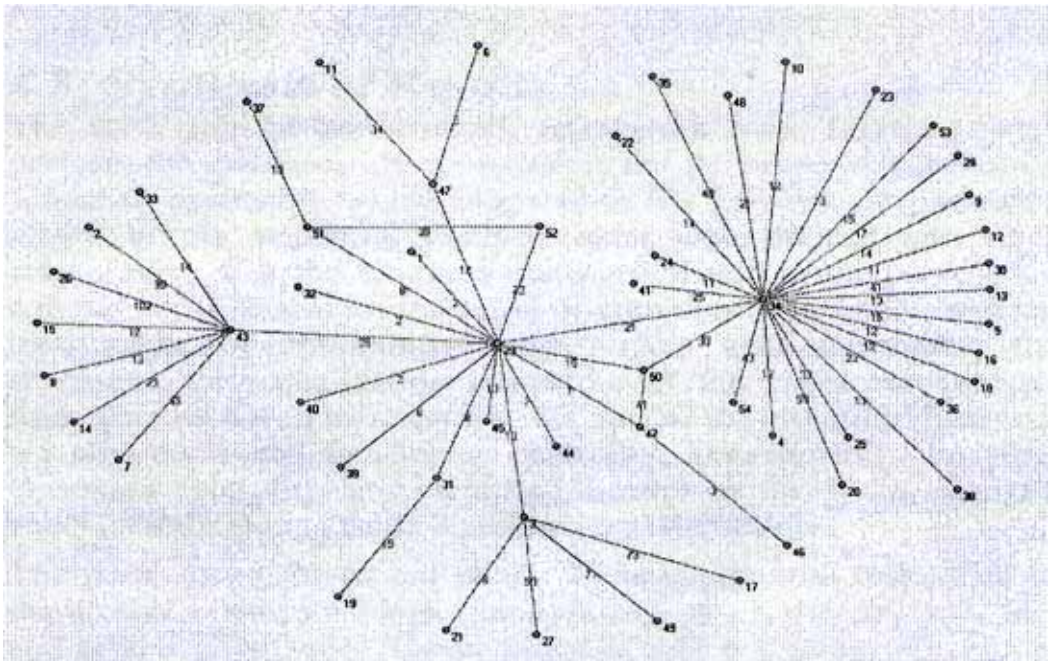

Abbildung 3.12 Netzwerk zur Identifizierung von Zugehörigkeiten in dem Musik-Blog [Chin06]

Es scheint nun offensichtlich, dass hier soziale Strukturen vorhanden sind, welche auf mögliche Communities in diesem Netzwerk hindeuten. Abbildung 3.13 zeigt mögliche Communities unter Verwendung der Sternnetzwerk-Struktur.

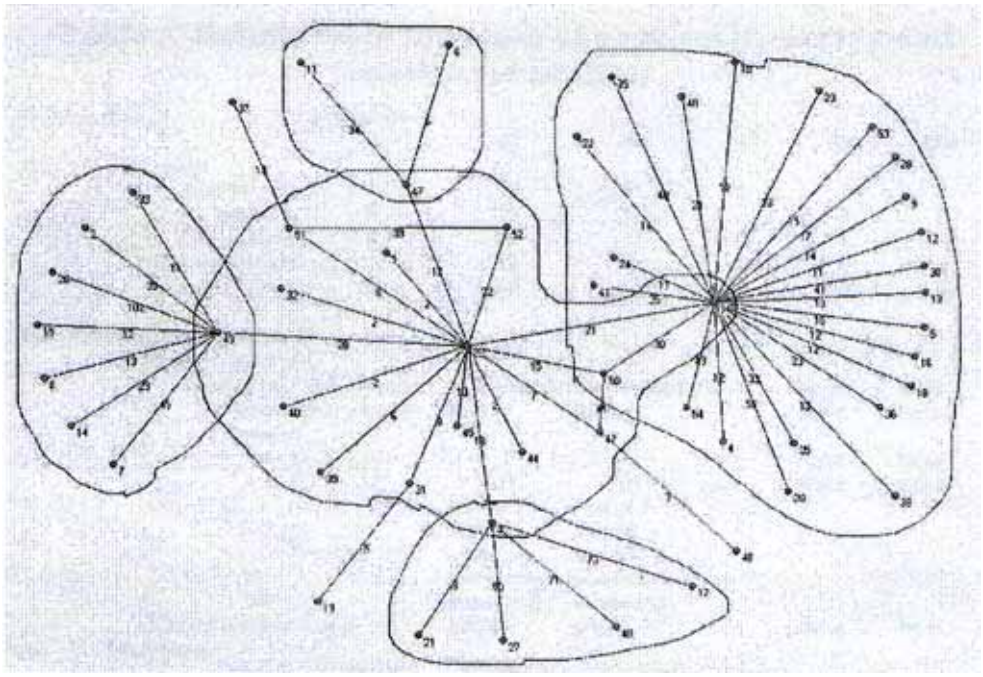

Abbildung 3.13 Mögliche Communities unter Verwendung der Sternnetzwerk-Struktur aus dem Musik-Blog Netzwerk [Chin06]

Blogs mit den Nummern 43, 34, 3 und 47 wurden als „Makler" identifiziert, da diese als Brücken zwischen Mitgliedern in ihrer Gruppe und dem Musik-Blog fungieren. Aus diesem Grund erscheinen die Sternnetzwerke dieser Blogs als mögliche Communities.

Anhand des Modells nach [Chin06] stellen Dreiecke und vollkommen verbundene Schaubilder die Strukturen für die Identifikation von geteilten emotionalen Verbindungen dar. In dem Musik-Blog wurden solche Dreiecke als vollkommen verbundene Schaubilder gefunden. Abbildung 3.14 stellt die Extraktion dieser Dreiecke dar.

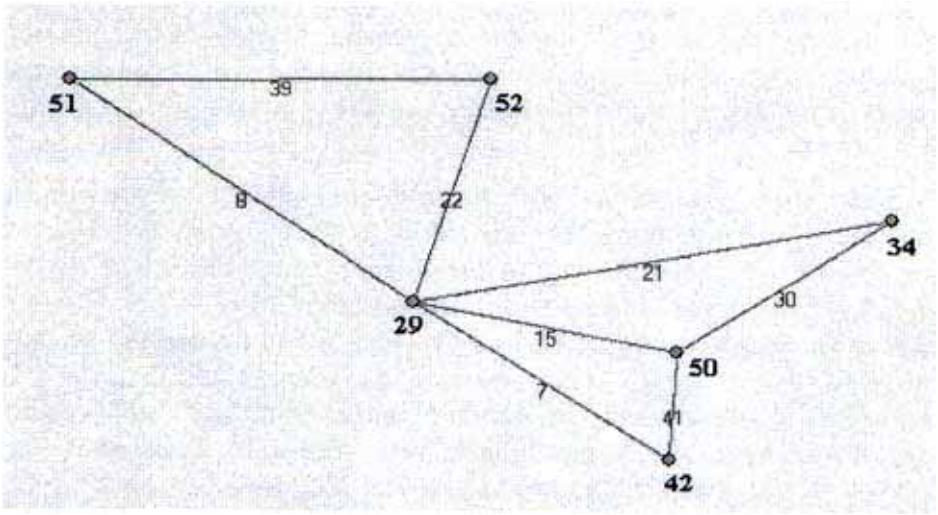

Abbildung 3.14 Gefundene Communities mittels der Identifizierung von geteilten emotionalen Verbindungen im Musik-Blog [Chin06]

Da die durchschnittliche Häufigkeit an Konversationen in dem Dreieck, welches den Musik-Blog (Blog 29) sowie die Blogs 34 und 50 beinhaltet, höher ist als in demjenigen Dreieck, welches die Blogs 29, 42 und 50 beinhaltet, kann es als kräftigere Community angesehen werden als das andere.

Abbildung 3.15 stellt das Netzwerk-Diagramm dar, welches durch die Kombination aus allen möglichen Communities, die unter Anwendung der Methodik nach [Chin06] gefunden wurden, erreicht wurde.

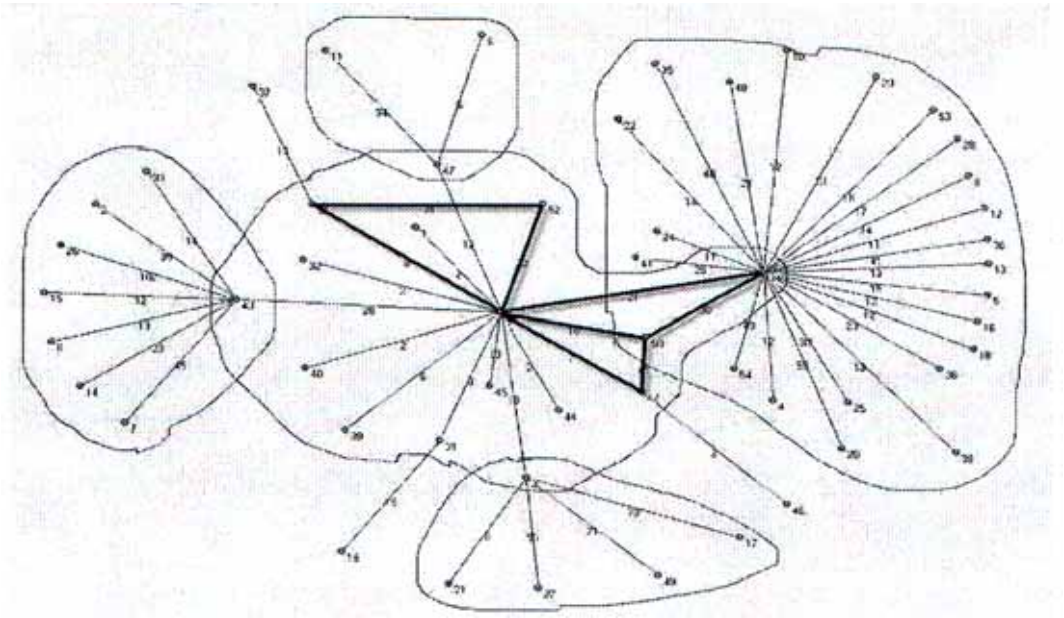

Abbildung 3.15 Identifizierung von Communities unter Verwendung der Visualisierungs-Indikatoren für das "Community-Gefühl" [Chin06]

Schritt Nummer sechs des Methodik-Kreislaufes nach [Chin06] bestand in der Analyse der „Community-Gefühl“ Werte. Diese wurden analysiert, um zu bestimmen, welche Blogs aus dem Netzwerk in Abbildung 3.15 – darüber hinaus in dem Strukturanalyse-Schritt (Nummer fünf) – tatsächlich einen Teil aus einer Community darstellen. Im Folgenden werden sowohl die Ergebnisse der „Community-Gefühl“ Umfrage als auch der Strukturanalyse beschrieben. Tabelle 3.11 stellt die Ergebnisse aus diesen beiden Schritten zusammenfassend dar. Es wurden sechs Blogs aus den insgesamt 15 Bloggern, welche an der Umfrage teilnahmen, herausgegriffen. Vier Blogs (45, 50, 34 und 343) stammen aus dem Netzwerk des Musik-Blogs. Die restlichen beiden Blogs (605 und 606) konnten nirgendwo gefunden werden. Dies hängt nach [Chin06] möglicherweise damit zusammen, dass diese Blogger lediglich Einträge und

Kommentare gelesen haben, jedoch keine Einträge aktiv im Musik-Blog geschrieben haben.

Analysis step	Blog number					
	45	50	34	343	605	606
Step #4: SOC index						
Membership	7	10	11	5	7	4
Influence	9	12	11	5	9	5
Need	9	13	12	8	9	10
Shared emotion	11	13	10	8	9	7
Total SOC index	36	48	44	26	34	26
Step #5: Structural analysis						
Visualization indicator						
Membership	Star network	Star network	Star network	None	None	None
Influence	Not a broker	Broker	Broker	None	None	None
Need						
min f	13	7	11	None	None	None
max f	13	41	59	None	None	None
Shared emotion	None	Part of 2 triangles which are completely connected graphs	Part of 1 triangle which is completely connected graph	None	None	None
SNA indicator						
Degree centrality (normalized)	0.0074627	0.0439469	0.0812604	0.000829		
Betweenness centrality (normalized)	0.0013794	0.016066	0.0760532	0		
Closeness centrality (normalized)	0.3528379	0.3802018	0.3880309	0.290602		
k-core	None	Part of two 2-cores	Part of one 2-core	None		

Tabelle 3.11 Ergebnisse aus der "Community-Gefühl" Umfrage und der Strukturanalyse [Chin06]

[Chin06] entdeckten im Rahmen ihrer Analyse, dass die Blogs 45, 50 und 34 in der Menge an Communities, welche aus dem Modell in Abbildung 3.13

identifiziert wurden, enthalten sind, während die restlichen Blogs nicht in Abbildung 3.13 gefunden werden konnten. In Anbetracht des SOC Gesamtindexes eines jeden Blogs aus Tabelle 3.11 könnte ein mögliches Kriterium für die Bestimmung ob ein Blog einem „Community-Gefühl" angehört, jenes sein, dass der jeweilige SOC Gesamtindex größer als 35 ist.
Die Umfrage nach [Chin06] erbrachte folgende Ergebnisse über das Blogger-Verhalten der in Tabelle 3.11 dargestellten Blogs: Blogger 45 schrieb mindestens einmal am Tag einen Blog-Eintrag und ist nur mit dem Musik-Blog verbunden. Seine Werte in den Charakteristiken aus dem „Community-Gefühl" sind im Vergleich zu den Bloggern 50 und 34 niedrig. Daraus folgt nach [Chin06] die Schlussfolgerung, dass Blog 45 Teil einer schwachen Community, welche durch den Musik-Blog geformt wird, darstellt.
Blogger 50 aktualisierte seinen Blog mehrere Male pro Woche. Abbildung 3.11 ist zu entnehmen, dass er mit den Blogs 34, 42 und 29 (Musik-Blog nach [Chin06]) verbunden ist. Sein SOC Gesamtindex ist sehr viel größer als der von Blogger 45. [Chin06] gehen davon aus, dass Blog 50 Teil einer starken Community, welche wiederum durch den Musik-Blog geformt wird, repräsentiert. Blogger 34 aktualisiert seine Blogs ebenfalls mehrere Male pro Woche. Abbildung 3.11 zeigt, dass er eine Verbindung zum Musik-Blog aufweist. Darüber hinaus wird durch das Sternnetzwerk identifiziert, dass er mit vielen anderen Blogs in seiner eigenen Community verbunden ist. [Chin06] schließen daraus, dass er Teil einer starken Community ist, welche durch das Sternnetzwerk identifiziert wurde und dem Dreieck, welchem er angehört.
Blogs 343 und 606 weisen den niedrigsten SOC Gesamtindex auf. Blogger 343 liest Blogs einmal am Tag, wobei er manchmal darauf verzichtet ein Kommentar an den gelesen Blogs abzugeben. Aus diesem Grund nehmen [Chin06] an, dass er nicht sehr viele Konversationen aufbauen wird. Dies erklärt auch den niedrigen SOC Gesamtindex. Analog verhält es sich bei den Blogs 605 und 606.

[Chin06] bedienten sich den Ergebnissen von der Visualisierung und der sozialen Netzwerkanalyse (siehe Tabelle 3.11), um zu untermauern ob die Blogs tatsächlich Teil einer Community darstellen. Die Ergebnisse aus diesen strukturellen Analysen lassen sich wie folgt zusammenfassen: Blog 45 ist Teil eines Sternnetzwerkes, welches um den Musik-Blog zentriert ist. Er weist den zweitniedrigsten Zentralitätsgrad[8] von allen Blogs aus der Umfrage auf. Dies bestätigt, dass er Zugehörigkeits-Gefühle hat. Er weist keine Einfluss-Gefühle auf, da er kein Makler ist und den zweitniedrigsten „Betweenness" Grad[9] hat. Er hat keine geteilte emotionale Verbindung, da er kein Teil eines Dreiecks oder eines vollkommenen Schaubilds ist. Aus diesem Grund folgt der Schluss [Chin06], dass Blog 45 keinem Teil einer Community angehört.

Blog 50 weist Zugehörigkeits-Gefühle auf, da er in dem Zentrum des Sternnetzwerkes gemeinsam mit Blogs 29, 34 und 42 ist. Weiters weist er den zweithöchsten Zentralitätsgrad auf. Er übt Einfluss auf andere Blogs, da er ein Makler mit dem zweithöchstem „Betweenness" Grad ist. Es kann validiert werden, dass Blog 50 Teil einer starken Community ist. Bei Blog 34 konnten [Chin06] aufzeigen, dass er Zugehörigkeits-Gefühle aufweist, da er im Zentrum seines Sternnetzwerkes positioniert ist und den höchsten Zentralitätsgrad von allen Blogs hat. Er hat Bedürfnis-Gefühle, da er viele Verbindungen mit gleichen Bloggern hat (mit der höchsten Frequenz von 59) und den höchsten Nähe Grad[10] aufweist. Blog 34 ist daher Teil einer starken Community [Chin06].

[8] Innerhalb der sozialen Netzwerkanalyse wird der Zentralitätsgrad als Indikator für eine Zugehörigkeit (membership) verwendet. Ein Knoten mit einem hohen Zentralitätsgrad hat viele direkte Verbindungen mit zahlreichen anderen Netzwerkknoten und kann andere beeinflussen.

[9] Ebenfalls im Rahmen der sozialen Netzwerkanalyse wird der „Betweenness" Grad als Indikator für den Einfluss (influence) verwendet. Ein Blog mit einem hohen „Betweenness" Grad übt einen hohen Einfluss auf andere Blogs in dieser Community aus.

[10] Der Nähe Grad scheint mit dem Konstrukt der Unabhängigkeit und der Stärkung der Bedürfnisse (reinforcement of needs) in Beziehung zu stehen. Ein Knoten C mit einem hohen Nähe Grad kann wirkungsvoll mit anderen Knoten in dem Netzwerk Kontakt aufnehmen, sodass Blog A, wenn er sich zu Blog C verbindet, andere Blogs über Blog C erreichen kann.

Blogs 343, 605 und 606 zeigten keine Zugehörigkeits-Strukturen, keine Bedürfnis-Strukturen, keine Einfluss-Strukturen und keine Strukturen an geteilten emotionalen Verbindungen. Blog 343 weist den niedrigsten Zentralitätsgrad, den niedrigsten „Betweenness" Grad und den niedrigsten Nähe Grad auf.

3.3.3 Zusammentragen von Blog-Inhalten auf Basis von Communities

[Qamr06] führen ihre Analysen von Blog-Artikel ebenfalls auf Blog-Communities sowie zusätzlich auch auf deren zeitliche Natur zurück. Innerhalb solcher Communities entwickelt sich eine gewisse Eigendynamik und es entstehen oftmals Diskussionen über bestimmte Themen. Noch präziser formuliert, korrespondiert eine Community nicht nur mit einem Thema, sondern vielmehr mit einem partikulären Standpunkt über dieses Thema. Aufgrund dessen, dass diese Diskussionsrunden häufig durch Antworten auf online- oder offline Geschehnisse ausgelöst werden, zeichnen sie sich durch eine limitierte zeitliche Dauer aus. Im Rahmen des Ansatzes von [Qamr06] wird ein Inhalts-Community-Zeit-Modell (engl. Content-Community-Time Model, CCT) vorgeschlagen, welches den Inhalt von Einträgen, deren individuelle Zeitstempel als auch die Community-Struktur der Blogs wirksam einsetzt, um dadurch die automatische Erkennung von Blog-Artikel zu ermöglichen. Als besonders wirkungsvoll wird dieses Modell in Zusammenhang mit dem Entdecken sogenannter „heißer Inhalte" gesehen [Qamr06]. Das Zusammentragen von Inhalten (engl. Story/Data Mining) wie auch von „hot stories" trägt dazu bei, zu erkennen beziehungsweise vorauszusehen welche Themen in verschiedenen Domänen und Communities von Interesse sind oder sein könnten. Darüber hinaus erstrecken sich Blog-Analysen immer mehr über das Feld der Marketing-Forschung, da Blog-Einträge oftmals freie Produktinformationen beinhalten.

Die Grundlage für diesen Ansatz nach [Qamr06] bildet folgende Problemstellung: Innerhalb der Blogosphäre entwickeln sich, wie bereits erwähnt, diverse Diskussions-Communities. Als besonders schwierig erweist es sich, zusammenhängende Diskussionen -in Anbetracht der fortlaufenden Zeit - aus diesen Communities zu extrahieren. Aus diesem Grund stellen [Qamr06] ein zeit- als auch Community sensitives Modell vor, um Blog-Einträge in bestimmte Geschichten zu gruppieren, während der Aushebelung bestimmter Blog-Charakteristiken wie Eintrags-Zeitstempel und Community-Strukturen. Dieses CCT-Modell wird von [Qamr06] als zweistufiges Verfahren betrachtet, welches im ersten Schritt Blog-Einträge zu Bündel entsprechend den Blogger-Communities gruppiert und anschließend Geschichten unter Verwendung einer Kombination aus Inhalt und Zeitstempel extrahiert. Zur Realisierung der Gruppierung von Blog-Einträgen in Geschichten wird ein wahrscheinlichkeits- sowie graphisches Modell verwendet. Die Geschichte wird dabei als versteckte Variable betrachtet während der Eintragsinhalt, die Zeitstempel und Links zwischen den Blogs als die beobachteten Variablen gesehen werden. Die Einträge werden zu Beginn nach der Community und dem Inhalt zusammengefasst. In jedem Bündel (engl. cluster) werden die Zeitstempel der Einträge zur Gruppierung dieser Einträge verwendet, damit aktuellere Einträge weniger wahrscheinlich mit älteren Einträgen gruppiert werden, sodass Einträge mit ähnlichen Inhalten jedoch mit weiter zeitlicher Aufteilung verschiedenen Geschichten zugeordnet werden. Die Zeit als solche kann sehr hilfreich für die Differenzierung verschiedener Inhalte sein. [Qamr06] führen in ihrem Artikel folgendes Beispiel an: Blog-Einträge, welche ähnliche Inhalte aufweisen, jedoch zeitlich weit auseinander verfasst wurden und folgedessen von verschiedenen Ereignissen ausgelöst wurden, sollten auch unterschiedlichen inhaltlichen Gruppierungen zugeordnet werden. Zwei inhaltlich ähnliche, jedoch zeitlich weit auseinander liegende, Blog-Einträge über den „iPod“ handeln möglicherweise von verschiedenen Generationen des „iPod“ und sollten nicht der gleichen Gruppe zugeordnet werden. Aus

diesem einfachen Beispiel lässt sich ableiten, wie wichtig die Berücksichtigung der Zeitstempel individueller Einträge in Bezug auf die Bündelung von Inhalten ist.

Im Folgenden wird das CCT Modell nach [Qamr06] beschrieben: Im Rahmen der Vorverarbeitung werden Blog-Einträge durchforstet und lokal gespeichert. Der Inhalt eines jeden Eintrags wird verarbeitet um Stopwörter[11] zu entfernen. Um eine Rückgewinnung der Einträge anhand von Abfrage-Schlüsselwörter zu ermöglichen, wird ein umgekehrter Index kreiert und der Inhalt von allen Einträgen wird in diesen Index eingefügt. Zusammen mit dem Inhalt, werden das Eintrags-Datum, die URL und die entsprechende Blog-URL im Index gespeichert. Um einen Community-Graphen für jeden Eintrag zu konstruieren, wurde eine Liste an Blogs, welche in dem entsprechenden Blog Querverweise innerhalb von *T* Tagen vor dem Eintrag an sich hatten, erzeugt. Diese Liste an verlinkten Blogs wurde ebenfalls im Index gespeichert, damit sie gemeinsam mit den Einträgen abgerufen werden kann. Der Community Graph wird verwendet um die Community-Struktur miteinzubeziehen, während CT (Community-Topic)-Cluster gefunden werden. Im eigentlichen ersten Schritt des CCT Modells wird die Inhalts- und Community Struktur untersucht und ein grobes Clustering durchgeführt. Die Menge an abgefragten Einträgen aus dem Testdatensatz wird in Community-Themen-Cluster unterteilt, sodass die Einträge innerhalb eines Clusters von einer Gruppe an Bloggern stammen, welche mit hoher Wahrscheinlichkeit ein ähnliches Interesse an den betreffenden Inhalten aufweisen beziehungsweise diese Inhalte auch online diskutieren [Qamr06]. Der zweite Schritt dient der Erkennung von Inhalten in jedem dieser CT-Cluster. Im zweiten Schritt geht es um die Entdeckung von Geschichten in jedem CT-Cluster. Hierbei wird der Inhalt abermals berücksichtigt, jedoch in einer derartigen Weise, um ein feineres Clustering zu

[11] Stopwörter werden als einfache geläufige Wörter bezeichnet, welche im täglichen Sprachgebrauch eingesetzt werden um Gedanken zu verbinden und um gebräuchliche grammatikalische Beziehungen zu erzeugen. Sehr viele Suchmaschinen beachten diese Wörter aufgrund von schnelleren Suchergebnissen nicht [Stop08].

erzeugen. Zusätzlich werden die Zeitstempel der jeweiligen Einträge benutzt um die Bündelung zu unterstützen. Wie bei den CT-Clustern ist es nicht notwendig, die Anzahl an Geschichten zu spezifizieren.

Im experimentellen Teil der Forschungsarbeit wurde das CCT Modell implementiert und dazu verwendet um Geschichten aus der Blogosphäre zu gewinnen. [Qamr06] legten eine Blog-Datenbank mit über einer Million Einträgen an. Bei denjenigen Blogs, bei welchen die Zeitstempel fehlten, wurden diese mittels Durchsehen der HTML Dateien oder der URL rekonstruiert. Es wurden insgesamt drei Fallstudien der Geschichten-Gewinnung unter Verwendung des CCT Modells durchgeführt. Die Abfrage Schlüsselwörter waren „Tsunami", „India China" und „Sony". Für jede Fallstudie wurden die obersten 200 Blog-Einträge verwendet, um diese in CT-Cluster und Geschichten zu gruppieren. Diese drei Fallstudien demonstrieren verschiedene Aspekte des Geschichten-Gewinnungsmodells nach [Qamr06]. Bei dem Schlüsselwort „Tsunami" wurden zwei CT-Cluster gefunden. Tabelle 3.12 zeigt diese beiden Cluster. Dargestellt werden die fünf häufigsten Wörter aus jedem Cluster (ausgenommen sind die Abfragewörter an sich).

CT	Most Frequent Words
1	katrina, donate, news, relief, hurricane
2	video, new, list, bittorrent, image

Tabelle 3.12 "Tsunami" CT-Cluster [Qam06]

CT-Cluster 1 beinhaltet Geschichten über das Tsunami Unglück während die Einträge aus dem CT-Cluster 2 Videos und Bilder bezüglich des Tsunamis bereitstellen. Während der Geschichts-Extraktionsphase wurden mehrere Geschichten aus dem CT-Cluster 1 extrahiert. Tabelle 3.13 zeigt die fünf häufigsten Wörter für diese Geschichten.

	Most Frequent Words	Time
1	katrina, hurricane, donate, relief, asian	Aug-Sep
2	news, donation, earhquake, relief, disaster	Dec-Feb
3	earthquake, island, report, photo, nicobar	Jul
4	help, relief, donate, countries, debt	Jan
5	sarvodaya, asia, million, raised, disaster	Aug-Sep

Tabelle 3.13 Geschichten in CT-Cluster 1 [Qamr06]

In CT-Cluster2 wird jeder Eintrag einer unterschiedlichen Geschichte zugewiesen. Eine Durchsicht dieser Inhalte zeigt, dass obwohl sie auf einer groben Ebene gleich sind, die Deutlichkeit der Einträge bedeutet, dass jeder Eintrag einer unterschiedlichen Geschichte entspricht.

Bei dem Schlüsselwort „India China“ wurden drei CT-Cluster gefunden. Tabelle 3.14 zeigt einige interessante Geschichten aus dem CT-Cluster 1 (repräsentiert durch die fünf häufigsten Wörter).

	Most Frequent Words	Time
1	time, story, rise, businessweek, tibet	Aug
2	industrial, elephant, chinese, growth, million	Jul-Sep

Tabelle 3.14 Geschichten in dem "India China" CT-Cluster 1 [Qamr06]

In diesem CT-Cluster 1 befindet sich eine Geschichte, welche Blogger, die einen Artikel über Indien und China (in dem Magazin „Business Week“ erschienen) diskutieren, beinhaltet. Eine Geschichte über den „Business Week“ Artikel findet sich ebenfalls im dem CT-Cluster 3. Tabelle 3.15 stellt

Geschichten aus dem CT-Cluster 3 dar (wiederum repräsentiert durch die fünf häufigsten Wörter).

	Most Frequent Words	Time
1	service, uranium, years, climate, companies, global	Jul-Aug
2	business, economy, outsource, service, infosys	Aug
3	businessweek, rise, chinese, economy, service	Aug

Tabelle 3.15 Geschichten in dem "India China" CT-Cluster 3 [Qamr06]

Dies demonstriert, dass verschiedene Communities über dieselben Geschichten diskutieren können. Ein User kann diese Geschichten über denselben Inhalt in verschiedenen Communities durchstöbern und verschiedene Perspektiven darüber erlangen [Qamr06].

Zusammenfassend lässt sich sagen, dass die gefundenen CT-Cluster als jene gesehen werden können, welche zu realen Online-Communities wie der Community der indischen Blogger gehören. Weiters wurde festgestellt, dass die gleiche Geschichte in verschiedenen Communities gefunden werden könnte, wie das Beispiel über den „Business Week“ Artikel zeigt. Ebenfalls konnten [Qamr06] observieren, dass der Einsatz von Zeitstempel wahrhaftig dazu beträgt, Geschichten besser zu finden, indem eine bessere Differenzierung, basierend darauf wann die Geschichten auftreten, möglich ist.

3.4 Bloggen als soziale Aktivität

Anders als bei den bisher vorgestellten Analysemodellen für Weblogs orientieren sich die nachfolgenden Ansätze an der gegenseitigen sozialen Interaktion der Blogger sowie an der Erforschung der Bedeutung von Blog-Leserschaften innerhalb der Blogosphäre.

3.4.1 Bloggen als Präsentationsplattform verborgener Inhalte

[Nard04] stützen ihre Thesen an den Ergebnissen einer durchgeführten ethnographischen Studie. Im Rahmen dieser Studie wurden vorwiegend drei Themenschwerpunkte untersucht:

- die Motivation für das Bloggen
- die Qualität der sozialen Interaktivität
- die Beziehung der Bloggerschaft zu ihrem Publikum/ Leserkreis

in dieser Blogstudie von [Nard04] wurden „lediglich" diejenigen Blogs analysiert, die von kleineren Gruppen beziehungsweise von individuellen Bloggern verfasst wurden. Nicht in Betracht gezogen wurden demzufolge stark frequentierte Blogs wie zum Beispiel Blogs über politische Kampagnen.
Eines der Hauptziele innerhalb der ethnographischen Analyse nach [Nard04] bestand darin, der Frage auf den Grund zu gehen, weshalb so eine breite Menschenmaße ihre persönlichen, in der Regel geheimen Tagebücher in einem derart öffentlichen Kommunikationsmedium, dem Internet, zur Schau stellt. Um eine Antwort oder mehrere Antworten auf diese Fragestellung zu bekommen, wird einerseits das Bloggen vom Gesichtspunkt des jeweiligen Bloggers (realisiert durch ethnographische Interviews) betrachtet und andererseits werden Blog-Einträge über einen kontinuierlichen Zeitraum

genauestens gelesen. Auf diese Art und Weise versuchen [Nard04] weiters zu analysieren, welche Faktoren für das Erzeugen und das Benutzen von Blog-Einträgen ausschlaggebend sind. Basierend auf den Ergebnissen aus dieser Studie werden Empfehlungen/ Vorschläge für eine optimierte Blog-Software gemacht. Im Folgenden wird die ethnographische Studie von [Nard04] detailliert beschrieben und es wird aufgezeigt, wie aus dem Bloggen eine soziale Aktivität entstehen kann.

Grundsätzlich kann die Studie in einen ethnographischen Tonband-Interviewteil sowie einen Text-Analyseteil untergliedert werden. Insgesamt wurden 23 gebildete Personen im Alter zwischen 19 und 60 Jahren zu Themen wie Beweggründe für das Eröffnen eines Blogs, Blog-Gewohnheiten, allgemeine Gedanken über das Bloggen, die Nutzung anderer Kommunikationsmedien (email, Telefon, Webseiten, etc.) und einigen anderen befragt. Eine der wichtigsten Erkenntnisse in Bezug auf Bloggen als soziale Aktivität ist das Charakteristikum, dass sowohl Blogs die Zuhörerschaft kreieren können, als auch der Leserkreis durchaus Blogs gestalten kann. In etwa 20 Prozent der Personen, die an der Studie teilnahmen, sagten aus, dass sie deshalb einen eigenen Blog ins Leben gerufen haben, da sie von anderen danach gefragt wurden. Ein noch viel höherer Prozentsatz startete seinen Blog in Folge einer Antwort auf eine direkte soziale Anfrage/ Aufforderung [Nard04].

Einige Blogger, welche angefangen haben ihre eigene Blogseite zu installieren, berichteten davon, dass sie plötzlich zu ihrer persönlichen „Stimme“ gefunden hätten, als sie realisierten, dass ihre Blogs tatsächlich von vielen anderen Mitmenschen gelesen werden. Ein anderer Teil der Befragten wiederum veränderte die Blog-Inhalte beziehungsweise wurde etwas vorsichtiger bei gewissen Äußerungen aufgrund diverser Rückmeldungen seitens der Leserschaft. Somit kann klar festgehalten werden, dass Blogger und der dazugehörige Leserkreis gemeinsam die soziale Natur des Bloggens

hervorrufen und dass Leser gleichermaßen Blogs formen wie die Autoren selbst.
Darüber hinaus berichtete ein Großteil der Teilnehmer aus der Studie nach [Nard04], dass Diskussionen über Blog-Themen häufig in anderen Medien (email, instant messaging, etc.) fortgesetzt wurden. Diese Ausdehnung der sozialen Interaktion auf andere Kommunikationsformen liefert gewissermaßen den Beweis dafür, dass Blogs als Teil eines größeren Bereichs sozialer Aktivität gesehen werden können.
Die weit verbreitete Betrachtung von Weblogs als so genannte Online-Tagebücher wird durch diese übergreifende soziale Interaktion ebenfalls in Frage gestellt. Um diese Behauptung zu untermauern folgten [Nadr04] der Aktivitätstheorie. Diese Theorie beschäftigt sich mit den Faktoren, die ein menschliches Handeln auslösen beziehungsweise motivieren. Folgende Faktoren waren für das Bloggen innerhalb der Studie verantwortlich, wobei manche Blogs auch von mehreren Faktoren motiviert wurden:

- anderen Personen eine aktuelle Information über bevorstehende Veranstaltungen zu geben
- das Ausdrücken gewisser Meinungen, um die eigenen Ideen mit anderen zu teilen und um andere zu beeinflussen
- das „Verlangen“ nach Rückmeldungen der Leserschaft
- das Benutzen des Blogs um den Schreib Prozess gewissermaßen durchzuarbeiten
- die Befreiung aus einer emotionalen Anspannung

Wie bereits eingangs erwähnt, versuchen [Nadr04] aufgrund der Ergebnisse und Erkenntnisse aus dieser ethnographischen Studie einige Verbesserungsvorschläge bezüglich des Designs sowie der Zweckmäßigkeit bestehender Blog-Systeme zu geben und einige kritische Anmerkungen in

Hinblick auf technologische Eigenschaften diverser Blogseiten zu treffen. Von zentralem Interesse unter den befragten Personen ist das Einbinden von Photos in persönliche Blogs. Bei vielen Blog-Systemen stellt das Einbinden von digitalen Photos eine regelrechte Herausforderung dar. „Blogger" beispielsweise verbot zum Zeitpunkt der Studie (April bis Juni 2003) das Posten von mehr als einem niedrigauflösenden Photo. Für das Einbinden von mehr als nur diesem einen Photo wird eine Gebühr eingefordert. Kein Teilnehmer aus Studie wäre jedoch bereit, diese Gebühr zu bezahlen. Stattdessen bedienten sich die meisten Personen anderer Alternativen wie der Eigenverwaltung ihrer Photos, Verfassen von Links zu Photo-Diensten und dem Entschluss keine Photos in ihrem Blog einzufügen.

Ein weiterer Kritikpunkt bestehender Blog-Systeme ist die Suchfunktion. Ein Durchstöbern entlang der zeitlichen Ebene gestaltet sich aufgrund der chronologischen Organisation relativ einfach. Die Suche nach bestimmten Informationen unter Einbeziehung anderer Indices erweist sich hingegen als weitaus schwieriger. Ein Grund dafür besteht darin, dass archivierte Blogs entweder chronologisch oder kategorisch abgespeichert und aufgerufen werden können, jedoch nicht in beide Richtungen. Abschließend konnten [Nadr04] einige weitere allgemeine Punkte bezüglich der Benutzerfreundlichkeit herauskristallisieren: Das Fenster zum Verfassen eines Blog-Eintrages ist in der Regel sehr klein. Folgedessen ist es nicht sonderlich gut geeignet um längere Einträge zu verfassen. Ein weiteres Problem stellt die Verwendung des Blogs als eigene Homepage dar. Dies ist insofern problematisch, da einige Interview-Teilnehmer versuchten ihren Blog als eigene Homepage zu verwenden, jedoch auf einige heikle Integrationswerkzeug-Aspekte wie zum Beispiel die Formatierung oder Frames in Zusammenhang mit der Verlinkung zu anderen Homepages auf ihrer Seite gestoßen sind.

3.4.2 Die Rolle der Leserschaft in ihrer Aktivität innerhalb des Blog-Prozesses

Im Vergleich zu vielen vorangegangenen Untersuchungen im Rahmen der Weblog-Forschung, welche allesamt den Fokus auf den Blog an sich oder die Blogger richten, beziehen [Baum08] gezielt die Weblog-Leser in ihre qualitative Studie mit ein, um ein besseres Verständnis des sozialen Gebrauchs des Bloggens. [Baum08] argumentieren dahingehend (unter ähnlichen Gesichtspunkten wie es in der so genannten „Leser-Antwort Theorie [Davi02][12]“ der Fall ist), dass ein vollkommenes Verständnis der Aktivitäten innerhalb des Bloggens in unmittelbarem Zusammenhang mit zahlreichen Analysen über mögliche Interpretationen sowie Antwortverhalten der Leserschaft auf individuelle Blog-Einträge steht. Die Leser dürfen auf keinen Fall als passive Rezipienten von Blog-Inhalten betrachtet werden, sondern müssen aktiv in ihrer Rolle in der gesamten Blogosphäre wahrgenommen werden. [Baum08] vermögen es zwar in ihrer qualitativen Studie nicht unbedingt Antworten auf Fragen wie: Wann, warum und wie entscheiden sich Leser Blogeinträge zu kommentieren? Welchen Gewohnheiten folgen Blog-Leser? zu finden, sondern tragen viel mehr dazu bei, einen Bezug zur Wichtigkeit der Leser in der Blog-Aktivität herzustellen.

Um verstärkte Kenntnis über das subjektive Erlebnis des Lesens von Blog-Einträgen zu bekommen, entschieden sich [Baum08] für qualitative sowie ethnographische Methoden. Für ihre Studie wurden insgesamt 15 Teilnehmer rekrutiert, wobei das Kriterium jenes war, dass mindestens fünf verschiedene Blogs zwei bis dreimal pro Woche gelesen werden mussten.

[12] Die Leser-Antwort Theorie (auch als Leser-Antwort Kritik bezeichnet) richtet seinen Fokus nicht nur auf die Literatur an sich sondern vielmehr auf die Antwort der Leserschaft darauf und auf die Interpretation des Textes. Diese Kritik argumentiert, dass die Literatur als Akt gesehen werden sollte, in welchem jeder Leser seine eigene, möglicherweise einzigartige, textbezogene Darbietung kreiert.

Tabelle 3.16 gibt einen Überblick über die individuellen Profile der Studien-Teilnehmer. Die in dieser Tabelle dargestellten Daten wurden über eine Online-Befragung gesammelt (bis auf eine Teilnehmerin nahmen alle Personen an der Befragung teil). Die Zielsetzung dieser Datensammlung besteht nicht darin statistische Interferenzen über Blog-Leser zu machen sondern zielt vielmehr darauf ab, ein Bild der verschiedenen Teilnehmer zu erzeugen. In der Spalte „tools" bedeuten die Zahlen folgendes: 1 steht für einen Webbrowser, 2 ist ein RSS Aggregat, 3 bedeutet email-Client, 4 bezeichnet eine Blog-Webseite und 5 bezieht sich auf Verlinkungen der Blog-Leser.

Pseudonym	Age	Gender	Occupation	Regular Blogs	Frequency	Example Blogs	Years Reading	Tools
Connie	22	F	--	--	Every Day	--	--	--
Fern	19	F	Student	1-2	Every Other Day	xanga.com, blogspot.com, livejournal.com	5-6 Years	4, 5, AIM Profiles
Selena	18	F	Student	6-10	2-3 Times a Week	greatestjournal.com, myspace.com, xanga.com, asianave.com	6-7 Years	1, 4, 5
Charles	24	M	Admin. Assistant	6-10	Several Times a Day	dailykos.com, boingboing.net, blogspot.com, slashdot.org, poplicks.com	6-7 Years	1, 4
Lillian	33	F	Graduate Student	20+	Every Day	blogspot.com, indigirl.com/blog, carrieoke.net, doggedknits.com	4.5 Years	2
Judith	20	F	Student	3-5	Every Other Day	myspace.com, xanga.com, facebook.com	3 Years	4
Jill	20	F	Student	6-10	Several Times a Day	livejournal.com, flickfilosopher.com/blog, ingliseast.typepad.com/ingliseast	5-6 Years	1
Cindy	19	F	Student	1-2	Several Times a Day	xanga.com, livejournal.com	5 Years	4
Patricia	20	F	Student	1-2	2-3 Times a Week	sibol.in, mochix.com	4 Years	1, 2, 5
Natalie	25	F	Legal Assistant	11-20	Every Other Day	perezhilton.com, blogspot.com, myspace.com, livejournal.com	10 Years	1, 4, 5
Tony	31	M	Graduate Student	3-5	Every Day	slashdot.org, fark.com, treehugger.com, somethingawful.com	6 Years	1, 3, iGoogle
Matthew	26	M	Graduate Student	11-20	Several Times a Day	blogspot.com, firejoemorgan.com, kugelmass.wordpress.com, sadlyno.com	6 Years	1, 2
Laura	27	F	Admin. Assistant	3-5	2-3 Times a Week	mypapercrane.com, blogspot.com, livejournal.com, bloesem.blogs.com	2 Years	1, 4
Cheryl	24	F	Graduate Student	3-5	2-3 Times a Week	fourfour.typepad.com, 2manadvantage.com, nydailynews.com/blogs/mets	2-3 Years	1
Krish	22	M	Student	3-5	Every Day	metblogs.com, kiruba.com, blogspot.com, aparnasblog.wordpress.com	8 Months	1

Tabelle 3.16 Profile der Studien-Teilnehmer [Baum08]

[Baum08] bedienten sich folgender drei Daten-Sammlungstechniken im Rahmen ihrer Untersuchung:

- Zwei semi-strukturierte Befragungen mit jedem Teilnehmer
- Eine Protokollierungs-Software um bestimmte Lesermuster zu verfolgen
- Eine Umfrage um Basisdaten wie beispielsweise Demographien zu gewinnen.

Aufgrund dessen, dass der Fokus bestehender Literatur nur in geringem Ausmaß auf die Blog-Leserschaft gerichtet ist, gestaltet sich diese erste semi-strukturierte Umfrage größtenteils als explorativ und generativ. Mit dem ersten der beiden Interviews verfolgten [Baum08] unter Anderem den Zweck, zusätzliche interessante Themen und Inhalte unter den Blog-Lesergewohnheiten der befragten Personen herauszufinden. Im Folgenden werden die Erkenntnisse, welche [Baum08] im Rahmen ihrer Methode über die Lesergewohnheiten der Studienteilnehmer gewinnen konnten, dargestellt: 13 der befragten Personen gaben auf die Frage warum sie eigentlich Blogs lesen zur Antwort, dass es für sie eine Form des „Entspannens" darstellt, es als netter Zeitvertreib dient, es einfach eine bequeme Art des Nichtstuns ermöglicht, etc. Des Weiteren entwickelte sich das Lesen von Blogs als eine Art Gewohnheit unter den Studien-Teilnehmern im Laufe der Zeit. Einer der Studien-Teilnehmer (Krish), welcher erst seit acht Monaten Blogs liest, erklärte, dass das Abrufen von Blogs dem Abrufen von den eigenen e-mails entspricht. Diese Gewohnheit ähnelt derjenigen, welche auch von neun anderen Teilnehmern beschrieben wurde. Weiters stellte sich heraus, dass ein Gefühl der Informationsüberflutung, welche oftmals als Prämisse im Arbeitsprozess der Informationswiedergewinnung oder Suchtechnologien dient, nicht sehr geläufig unter den befragten Personen war. Lediglich zwei von den 15 Teilnehmern drückten ein Gefühl der Überflutung durch potenzielle Informationen, welche mittels Blogs erhältlich sind, aus. Die restlichen

Teilnehmer wiesen darauf hin, dass sie sich nicht dadurch gestört fühlten, wenn sie nicht die neuesten Einträge auf denjenigen Blogs, welche sie frequentierten, mit verfolgen konnten. Eine ebenfalls interessante Frage die innerhalb der Interviews nach [Baum08] aufgeworfen wird, ist nach der Synchronität. Computervermittelte Kommunikationen werden oftmals als synchrone (zum Beispiel Live-Video und Live-Audio Konferenzen) oder asynchrone (e-mail) Formen bezeichnet. In Bezug auf die Studienteilnehmer stellte sich heraus, dass diese Blogs auf keine zeitlich befindliche Art lesen. Ein Blog, welcher den jüngsten Eintrag in einem Blog repräsentiert obwohl er möglicherweise bereits ein paar Tage alt ist, wird mit höherer Wahrscheinlichkeit gelesen, als der viertletzte Eintrag, welcher vom vorhergehenden Tag stammt, in einem anderen Blog. Diese Art der Kommunikation ähnelt sogenannten „instant messaging“[13] Konversationen, bei welchen zeitliche Aussetzer zwischen den Konversationen nicht notwendigerweise einen Einfluss auf diese haben. [Baum08] schlagen den Begriff „non-chronous“ in diesem Zusammenhang vor, um Verfahren zu beschreiben, bei welchen individuelle Ereignisse in einem Kontext (in diesem Fall ein einzelner Blog) in der zeitlichen Reihenfolge in welcher diese auftauchen betrachtet werden, jedoch ohne Einbeziehung des spezifischen Zeitpunktes, in welchem diese Ereignisse eintreten.

Innerhalb des zweiten semi-strukturellen Interviews wurden die Teilnehmer darum gebeten, weitere von den spezifischen Themen, welche in der ersten Interviewrunde aufgetaucht sind, zu diskutieren. Alle Befragungen und Notizen aus den beiden Interviews wurden transkribiert und kodiert. Zu Beginn wurde eine offene Kodierung verwendet und anschließend folgte eine Überleitung zur axialen Kodierung [Baum08]. Die anfängliche Kodierung begann nach der

[13] Unter „instant messaging“ wird eine sofortige Nachrichtenübermittelung verstanden, bei welcher zwei oder mehrere Teilnehmer mittels Textnachrichten miteinander kommunizieren. Beispiele solcher „instant messaging“ Programme/ Protokolle sind Skype, ICQ, Windows Live Messenger, etc.

Beendigung der ersten Interviewrunde. Die Ergebnisse aus den Analysen dieser ersten Menge an Interviews unterstützten dabei, die zweite Interviewrunde anzuregen und zu dirigieren. Bezüglich der zu installierenden Protokollierungs-Software stellte sich heraus, dass die meisten Teilnehmer sich dagegen entschieden diese auf ihrem privaten PC zu installieren (lediglich fünf Personen installierten die Software erfolgreich) oder mit technischen Schwierigkeiten konfrontiert waren. Dennoch wurden die Analysen aus diesen Aufzeichnungen (engl. logs) verwendet, um Fragen für die zweite Interviewrunde zu generieren.

Im Folgenden werden die Ergebnisse, die [Baum08] im Rahmen ihrer Methodik gewinnen konnten und die sich von Resultaten aus früheren Arbeiten unterscheiden, dargestellt: Vorangegangene Arbeiten am Blogsektor haben verschiedene Elemente der Präsentation und Wahrnehmung übersehen. Möglicherweise hat sich dies aus deren Fokus auf die Blogger als sowohl Produzenten als auch Konsumenten von Blogs ergeben. Blog-Leser werden von Bloggern oftmals als eine nervtötende und anonyme Gruppe von Schleichern oder Anstiftern wahrgenommen, welche tölpelhafte soziale Situationen kreiert oder manchmal eine eindringende Anwesenheit präsentiert. Die Teilnehmer aus der Studie nach [Baum08] hatten allesamt gemeinsam, dass sie verschiedenartig in Abhängigkeit vom jeweiligen Blog kommentieren, anstiften oder schleichen würden. Bezüglich des Kommentierens gaben 11 der 15 Teilnehmer zur Antwort, dass sie semi-regulär auf Meinungsäußerungen oder Empfindungen treffen würden, welchen sie nicht zugestimmt haben. Lediglich vier Teilnehmer teilten Instanzen, bei welchen ihre Sichtweisen sich signifikant unterschieden und sie sich entschlossen, ihre Meinungsverschiedenheiten mittels Kommentaren auszudrücken. Jedoch würde nur einer von diesen vieren jene Kommentare, mit dem Ziel der Anstiftung einer tölpelhaften Situation oder der Invasion des Raumes von Bloggern, abgeben.

Während frühere Untersuchungen lediglich die Erwartungen, die Leser an die Blogger richten, beschrieben, stellte sich im Rahmen der Studie nach [Baum08] heraus, dass Leser das Gefühl haben, dass gewisse Erwartungen genauso auch an sie gerichtet werden. Gleichermaßen wie Blogger sich unter Druck fühlen Einträge zu aktualisieren, fühlten sich zehn Teilnehmer dazu verpflichtet, besonders die Blogs von Freunden oder jene Blogs, von denen sie den Eindruck haben einen Teil davon darzustellen, zu lesen oder zu kommentieren. Die Situation in Bezug auf die Erwartungen der Leser ist eine sehr Komplexe Sache. 13 Teilnehmer drückten ihre Erwartungen in Hinblick auf die Aktualisierungs-Häufigkeit, das visuelle Erscheinungsbild, die Angemessenheit und andere Aspekte von Blogs aus. Dies resultiert daraus, dass die Teilnehmer verschiedene Erwartungen an unterschiedlichen Blogs haben, genauso wie Leser verschiedene Blogs auf unterschiedliche Weise lesen. Wenn Leser zum Beispiel Kommentare auf großen Blogseiten machen, wird nur selten eine Antwort erwartet, während ein Kommentar auf einem Blog eines Freundes beinahe eine Erwiderung verlangt.

Obwohl sich die Studie nach [Baum08] auf Blog-Leser bezieht, sind es lediglich drei von den 15 Teilnehmern, die keinen eigenen Blog besitzen. Trotz der Tatsache, dass viele der Teilnehmer auch Blogger sind, können die hier beschriebenen Erkenntnisse auch auf die Blog-Leser angewandt werden, da es keinen Beweis in der Literatur gibt, dass Blogger existieren, welche keine Blogs lesen [Baum08]. Jedoch gestaltet sich eine Abweichung in der Tendenz, dass Nicht-Blogger lediglich populäre oder hoch-frequentierte Blogs lesen, wohingegen zehn von den 12 Blog-Leser, welche einen Blog führen, aus der Studie ihre Blogs dazu verwenden, um den Kontakt mit Freunden aufrecht zu halten.

3.5 Analysemodelle basierend auf Blog-Metadaten/ Blog-Tags

Allgemein formuliert dienen Tags (am ehesten mit den Begriffen Etikett und Markierung zu übersetzen) in der Informatik der zusätzlichen Kennzeichnung und Strukturierung zahlreicher Texte, Photos, Videos oder sonstiger Datenbestände.
Speziell im Internet werden Tags als relativ neue Art der erweiterten Kennzeichnung von Informationen - damit diese effizienter einsortiert und in späterer Folge leichter aufgefunden werden können - bezeichnet [Mars08]. In Zusammenhang mit Weblogs sind Tags folgendermaßen von Bedeutung: Auf den meisten Blog-Seiten finden sich in der Regel eine Vielzahl an Einträgen/ Artikel, welche diversen Kategorien zugeordnet sind. Das „Problem" besteht jedoch darin, dass die einzelnen Kategorien im Laufe der Zeit eine derart große Anzahl an Artikel beinhalten, sodass sich die gezielte Suche nach einem bestimmten Eintrag als durchwegs schwierig erweist. Hier bieten sich Tags als besonders probates Mittel an, da ein Artikel mit mehreren frei wählbaren Schlüsselwörtern gekennzeichnet werden kann. Somit kann ein Blog-Eintrag, welcher einer bestimmten Kategorie zugeordnet ist, mehrere zusätzliche Tags besitzen, die eine Auffindung leichter machen beziehungsweise der Kontext zu thematisch ähnlichen Inhalten hergestellt werden kann.

3.5.1 Die Analyse von Tags für ein Blog-Empfehlungssystem

Der Hauptuntersuchungsgegenstand in der Arbeit von [Haye07] beinhaltet das Verstehen der Dynamik von Beziehungen zwischen verschiedenen Themen und Benutzern innerhalb von Blogs, um eine plausible Erklärung für das Blogger-Verhalten zu konstruieren [Bere07b].
Aufgrund der Tatsache, dass sich die gesamte Blogdomäne aus vielen Millionen Dokumenten (diese werden regelmäßig aktualisiert) zusammensetzt, besteht seit geraumer Zeit der Bedarf diese Einträge anhand ihres Typs oder

ihrer Thematik zu organisieren beziehungsweise zu strukturieren. Clusterbildungen eignen sich in einem ersten Schritt sehr schön dazu, um solch große Datenbestände zu durchsuchen, um ähnliche Benutzerprofile in Empfehlungssystemen zu gruppieren, um Suchmaschinen-Ergebnisse zu organisieren, etc. [Haye07]. Das wesentliche Ziel eines qualitativen Clusterbildungsalgorithmus besteht darin, eine Menge an Datenpunkten dermaßen in eine Menge an Cluster zu partitionieren, dass Punkte aus dem selben Cluster nah beieinander liegen und Punkte aus verschiedenen Cluster weit voneinander entfernt sind. Folgende Gleichung verdeutlicht das Verhältnis zwischen der Intra- und Interclusterdistanz:

$$\mathrm{H}_r = \frac{I_r}{E_r} = \frac{\frac{1}{|S_r|}\sum_{d_i \in S_r} \cos(d_i, C_r)}{\cos(C_r, C)}$$

H_r .....Verhältnis zwischen der Intra- und Interclusterdistanz

r .....Cluster

S_r .....Menge an Instanzen von r

I_r .....Intraclusterähnlichkeit

C_r .....Cluster Schwerpunkt

E_r .....Interclusterähnlichkeit

$d_i \in S_r$ .....jeder Instanz

c .....Gesamter Datensatz

In direktem Zusammenhang mit der Strukturierung beziehungsweise der Empfehlung von neuen Blog-Einträgen stehen auch die bereits erwähnten Tags. [Haye07] definieren drei verschiedene Arten von Tags: A-Tags werden

als Hochfrequenz-Tags bezeichnet. Sie repräsentieren eine unabhängige Beschreibung von dem Clusterthema aus zwei oder mehr Bloggern. B-Tags kommen in mehreren Cluster zugleich vor. B-Tags können Stopwörtern gleichgesetzt werden, da sie ebenfalls unbrauchbar in Hinblick auf Indizierungs- und Rückgewinnungszwecke sind. C-Tags werden von keinem anderen Benutzer in einem Cluster wiederholt. Darüber hinaus werden sie nicht in der so genannten Tag Cloud-Ansicht dargestellt.

In Bezug auf die Anwendung von Clusterbildungsmethoden sowie den Gebrauch von Tags auf einen statischen Datensatz weisen [Haye07] darauf hin, dass dabei die dynamische Natur von Blogdomänen außer Acht gelassen wird. Der Grund weshalb die dynamische Struktur von Weblogs besonderer Aufmerksamkeit in diesem Zusammenhang bedarf kann folgendermaßen erklärt werden: In der Regel werden Blog-Einträge regelmäßig von den Autoren aktualisiert beziehungsweise werden neue Posts (im Englischen wird ein „Posting“ als eine Nachricht innerhalb von Newsgroups oder Foren bezeichnet) den jeweiligen Blogs hinzugefügt. Die Schlüsselfrage, die hier auftaucht, ist, ob die Beziehungen, welche durch eine Clusterlösung etabliert wurden, auch in einem nächsten Zeitfenster Gültigkeit besitzen. Eine weitere Schlüsselfrage ist, inwiefern die relevantesten und durchgängigsten Blogs, welche mit einem ganz bestimmten Thema assoziiert werden, identifiziert werden können [Haye07]. Mit verschiedenen Maßnahmen bezüglich Benutzer und Themenabweichungen innerhalb der Blogdaten wird über die Zeit hinweg versucht, Antworten auf diese Fragen zu finden.

Die Motivation der Forschungsarbeit nach [Haye07] liegt unter anderem darin, ein Mittel für die Kopplung von Ressourcen derart zu unterstützten, dass Benutzer relevantes themenbezogenes Material aus verschiedenen Quellen auffinden sowie verwenden können. Es geht hierbei um eine Art von Blog-Empfehlungssystem, in welchem registrierte User regelmäßig Beiträge oder

Tags von anderen Bloggern, die ähnliche Interessen aufweisen, empfohlen bekommen. Von den angestrebten Zielen sind sehr viele mit denen aus dem SIOC Daten Modell (http://www.sioc-project.org/) vergleichbar. Die SIOC Initiative zielt darauf ab, die Integration von Online-Community Informationen zu ermöglichen. Die Grundlage für das SIOC Modell bildet ein RDF-basiertes Schema, welches die Hauptkonzepte, die in Online-Communities gefunden werden, beschreibt [Haye07]. Derzeit können Online-Communities wie zum Beispiel Blogs als Inseln betrachtet werden, welche wertvolle Informationen beinhalten jedoch nicht sehr gut miteinander verbunden sind. Das SIOC Modell ermöglicht die Verkoppelung solcher Seiten und die Extraktion von reichhaltigerer Information aus verschiedenen Diskussionsdiensten. Abbildung 3.16 zeigt den Aufbau SIOC Daten-Modells im Überblick. Den Kern bildet die Ontologie. Darunter ist ein Wortschatz zu verstehen, welcher Konzepte beinhaltet. Diese Konzepte sind notwendig, um Information, welche auf Online-Community Seiten enthalten ist, auszudrücken [Sioc08]. Online Community-Seiten stellen der Außenwelt Information über ihre Struktur und ihre Inhalte bereit. Diese Information ist maschinenlesbar und durch die SIOC Ontologie strukturiert. Aufgrund dessen, dass diese Information innerhalb solcher Community Seiten präsent ist, besteht die einzige Notwendigkeit darin ein SIOC Export Plugin oder eine Erweiterung zu installieren. Diese Information kann von Werkzeugen, welche SIOC Daten interpretieren können, verwendet werden, um verwandte Informationen aus anderen Community-Seites vorzuschlagen.

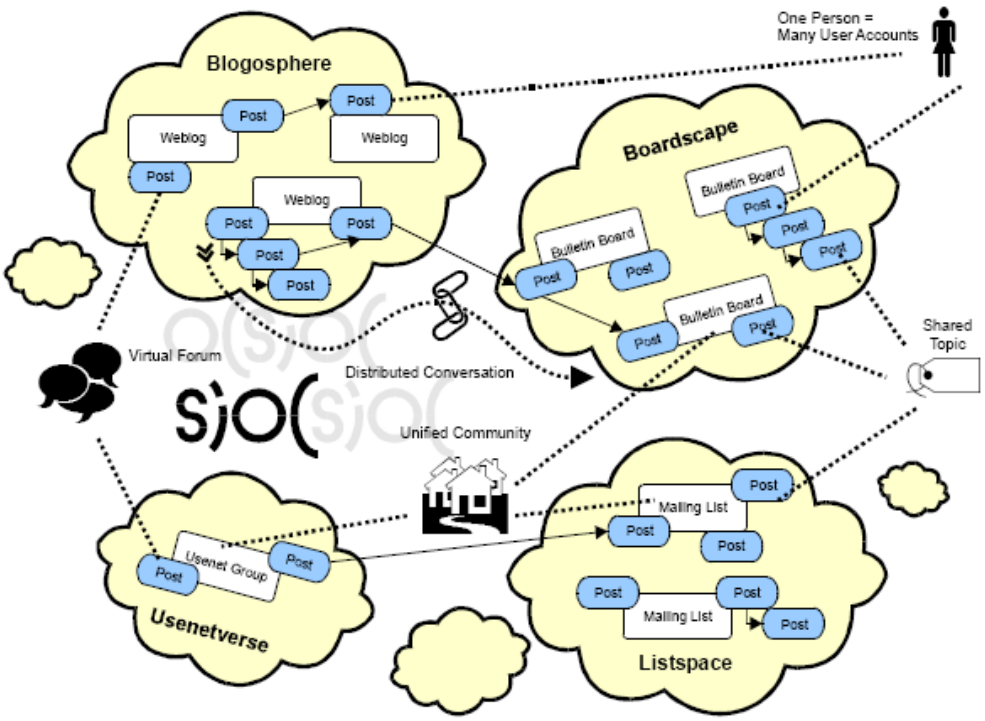

Abbildung 3.16 Erstellung von Verbindungen zwischen Diskussionswolken mittels SIOC [Sioc08]

[Haye07] versuchen demzufolge Dienste basierend auf Wissenserkennung aufzubauen, welche die maschinenlesbare Darstellung von Blog-Daten anreichern. Ein sogenannter Blog-Empfehlungsdienst akzeptiert SIOC Metadaten über einen Eintrag als Eingang und liefert eine Auswahl an Tags zurück, die die Blogger nach Möglichkeit verwenden. Abbildung 3.17 stellt eine Anwendung dieser Dienste in Form eines Tag-Empfehlungsdienstes dar: Ein großes Problem, mit welchem Blogger konfrontiert sind, ist die Auswahl an bedeutungsvollen Tags nachdem ein Eintrag geschrieben wurde. Innerhalb dieses Empfehlungsdienstes wird ein Eingangseintrag dem ähnlichsten Cluster zugeordnet und es wird eine Auswahl an A-Tags an den User zurückgeliefert. Ein zweiter Dienst würde ebenfalls eine Liste an A-Tags zum Eingangseintrag zurückliefern. Durch die Auswahl von einem oder verschiedenen A-Tags kann

der Blogger sicherstellen, dass sein Eintrag leichter zugeordnet oder durch andere Blogger oder sogar durch Leser von Community-Seiten aufgefunden werden kann. Dieser Sachverhalt stellt gleichzeitig eines der Ziele des SIOC Gerüsts dar. Unter Verwendung des SIOC Export-Zusatzmoduls kann der Blogger seine Eintrags-Metadaten durch Informationen, welche von dem Empfehlungssystem empfangen werden, anreichern.

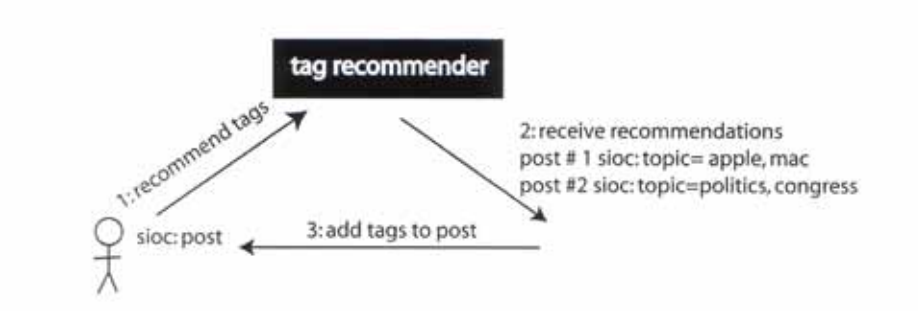

Abbildung 3.17 Einfaches SIOC Empfehlungssystem [Haye07]

Abschließend lässt sich festhalten, dass [Haye07] im Rahmen ihres Analysemodells (der Blog-Datensatz für die Untersuchungen wurde aus rund 13500 Blogs zusammengetragen) zu der Erkenntnis gekommen sind, dass die Blogosphäre durch eine Vielzahl an Bloggern gekennzeichnet ist, die häufig von einem Thema zum nächsten „wandern". Des Weiteren konnte gezeigt werden, wie diverse Tag-Informationen dazu verwendet werden können, um das Ergebnis einer Clustering-Lösung zu verfeinern.

3.5.2 Tags können mehr als nur Metadaten darstellen

Für viele Menschen stellen Tags laut [Bere07a] nach wie vor ein sehr großes Verständnisproblem dar. Häufig wird die Frage gestellt ob Tags den so genannten Metadaten gleichzusetzen sind beziehungsweise welche Bedeutung sie überhaupt mit sich tragen. [Bere07a] präsentieren in ihrer Forschungsarbeit eine Definition sowie einen empirischen Beweis für die Behauptung, dass Tags nicht nur reine Metadaten sondern vielmehr Inhalt darstellen. Die Analyse nach [Bere07a] beruht auf einer „Multi-Kommentator-Klassifizierung" eines Standard-Blogkorpus (dieser Korpus war ein Offert von Blogpulse für den „Weblogging Ecosystem Workshop 2006") unter Einbeziehung des WordNet Domänen-Bezeichnungssystems (engl. WordNet Domain Labels System, WND), der Entwicklung eines Systems aus Text-Klassifikationsmethoden (basierend auf WordNet und WND) und auf der qualitativen vergleichbaren Analyse dieser Klassifizierungen.

Tags spielen innerhalb von Weblogs eine spezielle Rolle und werden sowohl von den Autoren in Form von Zusatzinformationen zu den Blog Texten als auch von Personen, die diese gekennzeichneten Inhalte wieder- oder weiterverwenden (zum Beispiel auf Webseiten wie „http://flickr.com/" oder „http://delicious.com/"), erstellt. Der Hauptbeitrag der Analyse nach [Bere07a] besteht darin, Antworten auf folgende Fragen zu geben:

- Tags unterscheiden sich von traditionellen Klassifikations-Schlüsselwörtern. Sind Tags jedoch auch anders zu betrachten als Metadaten in ihrer Beziehung zu dem zugrundeliegenden Inhalt?

- Kann eine Untersuchung dieser Fragestellungen dazu beitragen, um das Verständnis dahingehend, worüber Inhalte wirklich handeln, zu vertiefen?

- Welche Voraussetzungen entstehen für Suchmaschinen, wenn mit gekennzeichneten Materialien gearbeitet wird? Auf welche Weise können Suchmaschinen verbessert werden, um individuelle Benutzer auf der einen Seite und Gruppen auf der anderen Seite zufriedenzustellen?

- Welche empirischen und computerbasierten Methoden eignen sich dazu, um die vorangegangen Fragen bestmöglich zu analysieren? [Bere07a]

Im Folgenden wird der Aufbau der Studie nach [Bere07a] detailliert beschrieben: Die anfängliche Zielsetzung der Forschungsarbeit bestand darin, die Qualität verschiedener Methoden für die Textinhalts-Klassifikation an Blogs zu beurteilen. Genau genommen, waren [Bere07a] an Methoden interessiert, welche einen „kalten Start" ermöglichen. „Kalter Start" bedeutet, dass keine markierten Daten für das Klassifizierungslernen notwendig sind. Methoden, welche auf der Semantik basieren, stellen eine naheliegende Wahl für diesen Sachverhalt dar. Solche Methoden können dazu verwendet werden, um einen unbekannten Blog-Korpus in eine kleine und handliche Anzahl an sinnhaften Kategorien zu organisieren, welche die Suche unterstützen können. [Bere07a] bedienten sich an dem Korpus[14], welcher im Jahre 2006 dem „Weblog Ecosystem Workshop" von Blogpulse angeboten wurde, um sich auf eine standardisierte Menge an Texten zu beziehen. Es wurde eine zufällige Stichprobe aus 100 Blog-Einträgen aus diesem Korpus ausgewählt. Diese Einträge wurden am vierten Juli 2006 geschrieben, waren auf Englisch und wurden durch ihre Autoren markiert. Die Gründe nach [Bere07a] für diese Auswahl sind folgende: Es wurden deshalb englische sowie von ihren Autoren markierte Artikel ausgewählt, um sicherzugehen, dass Standard-

[14] Die Webseite dieses Workshops ist unter folgender URL erreichbar: „http://www.blogpulse.com/www2006-workshop/#data"

Textanalysewerkzeuge im Stande sind diese Daten aufzubereiten und dass Tags analysiert ohne fehlende Daten (lediglich 24,1 Prozent der Einträge auf dem großen Korpus besitzen überhaupt Tags und lediglich 68 Prozent von diesen Tags sind verwendbar) analysiert werden können. Die relativ kleine Anzahl an ausgewählten Einträgen beruht auf der Sicherstellung einer hoch-qualitativen Zusammenarbeit durch die freiwilligen Kommentatoren.

Das Basismodell der Inhaltsklassifikation bildete die Zuweisung von einem oder mehreren semantischen Etiketten zu einem Blog-Eintrag. Hierfür wurden zwei Ressourcen verwendet: Zum Einen „WordNet[15]“ und zum Anderen Domänenkennzeichnungen von IRST (ermöglicht eine Zusammenführung zwischen WordNet „Synsets“ und 165 taxonomisch strukturierten Domänen wie zum Beispiel: Zahnmedizin bildet eine Art von Medizin, Medizin wiederum ist als eine angewandte Wissenschaft zu betrachten). Ein Begriff kann verschiedenen Domänen angehören, je nachdem welche Bedeutung durch ihn angezeigt wird. Als „Methoden“ wurden fünf Informationssystem-Absolventen herangezogen, welche sich für das Kennzeichnen des Blog-Korpus bereit erklärten. Die sechste Methode stellte ihre Gesamtbeurteilung dar. Die Gesamtbeurteilung wurde entworfen, um eine Bezugs-Klassifikation einzurichten, gegen welche die verbleibenden automatisierten Methoden bewertet werden konnten. Den Absolventen wurden verschiedene Klartext-Dateien, welche die Korpus-Einträge zum Inhalt hatten, vorgelegt. Unter schriftlicher Anweisung mussten die Absolventen eine willkürliche Anzahl an Einträgen mit Tags kennzeichnen. Jedem Kennzeichner war es erlaubt, eine beliebige Anzahl an Domänen pro Eintrag festzusetzen, wobei zwischen null und drei empfohlen waren. Daraus resultierten 500 Bezeichnungsmengen, von denen 23 leer waren, 304 eine Domäne beinhalteten und 160 zwei Domänen beinhalteten. Das Kennzeichnungs-Verhalten erwies sich als ziemlich

[15] Stellt ein computerbasiertes Englischlexikon dar. Hauptwörter, Verben, Adjektive und Adverbien sind in Mengen an kognitiven Synonymen („synsets“) gruppiert. Jedes „synset“ drückt ein eindeutiges Konzept aus. „synsets“ sind durch die Bedeutung der konzeptuellen Semantik und durch lexikalische Beziehungen verkettet [Word08].

verschiedenartig: 13 aus den 23 leeren Mengen sind auf den Kennzeichner Nummer drei zurückzuführen, während Kennzeichner Nummer eins und fünf alle Einträge markiert haben. Die durchschnittliche Anzahl an Domänen pro Eintrag variierte zwischen 2,08 (Kennzeichner Nummer drei) und 2,65 (Kennzeichner Nummer zwei). Die fünf Kennzeichnungsmengen wurden zusammengefügt, um an einer Übereinstimmungs-Klassifikation für jeden Eintrag anzukommen. Für jeden Blog-Eintrag wurde die durchschnittliche Ähnlichkeit (der fünf Kommentatoren) aus den individuellen Kennzeichnungsmengen der Kommentatoren für die Übereinstimmungs-Klassifikation verwendet. Unterschiedliche Messungen wurden zur Bewertung der Ähnlichkeit herangezogen, wobei [Bere07a] lediglich die Ergebnisse unter Verwendung des Jaccard-Koeffizienten in ihrer Arbeit darstellen. Dieser Koeffizient definiert die Ähnlichkeit zwischen zwei Kennzeichnungsmengen. Die Ähnlichkeit der Kommentatoren zu der Übereinstimmungs-Klassifikation wich zwischen 0,31 (Kommentator Nummer drei) und 0,47 (Kommentator Nummer zwei) ab.

Die exakte Bestimmung darüber, was genau die Kennzeichnung eines jeden Kommentators motivierte, ist schwierig. Folgendes Beispiel zeigt die unterschiedliche Art und Weise der Kennzeichnung eines Blog-Artikels. Einer der Einträge aus dem Datensatz sah inhaltlich folgendermaßen aus:

„Tags: Radio & TV, Islam.
GetReligion.
Todays New York Times includes this report about Sleeper Cell, a 10-part Showtime series about a faithful Muslim named Darwyn (yes, we get it) who infiltrates a terrorist group. The Times mentions the producers goal of high realism, but also must grant that, while some Muslim FBI agents exist , theres no way to know if any such agent has infiltrated a terrorist cell. Still, its easy to sympathize with series star Oded Fehr (pictured), an Israeli actor playing a terrorist, and Cyrus Voris, one of the producers, as they discuss the shows

idealism: "You learn there are peace-loving souls in every religion," said Mr. Fehr, who once served in the Israeli military. "We have to respect and strengthen the peace-believers, and hopefully find a way turn the terrorists." In that sense, the production, for all its violence - including the Sopranoesque rubout of a cell member by his fellow crew – is perhaps most ambitious for the idealism that courses through it. "I dont know if a guy like Darwyn is out there somewhere in the U.S.," said Mr. Voris, a creator of the show. "But I sure hope so. Talk about wish fulfillment." Besides which, Showtime has a bit to atone for while it promotes the hostilities of Penn & Teller toward all things religious (including Christopher Hitchens whipping post, Mother Teresa)"

(Permalink: http://www.getreligion.org/?p=894).

Bei diesem Eintrag, in dem es hauptsächlich um Themen wie Religion, Fernsehen und Politik ging, wurden sehr unterschiedliche Tags vergeben. Kommentator eins klassifizierte diesen Eintrag als „Religion" und „TV", während Kommentator drei diesen als „Politik" einordnete. Die Tagmethode spiegelt klar die Wahl von Kommentator eins („Religion" und „TV" sind die einzig verfügbaren Informationen für diese Methode). Weiters bietet lediglich der Anfang des ersten Satzes Hinweise für diese Interpretationen. Die unterschiedlichen Ergebnisse aus diesem Blog können auch dahingehend interpretiert werden, dass der Kommentator eins unmittelbar nach Beginn des Lesens versuchte, eine Meinung über den Inhalt zu formen, während Kommentator drei bis zum Ende las. Diese Annahme wird durch die Klassifikation der beiden Kommentatoren des folgenden Eintrages noch weiter ausgebaut:

Tags: Art.

Cool Hunting. I first wrote about Ludwika Ogorzelec s Space Crystallization Cycle after seeing her show here in NYC last February. Her prolific installation of site specific cellophane lattice has graced a broad range of settings since the series

began a couple years ago. The latest... farmland. Farming With Mary is a Queensland Australia project that brought environmental artists from all over the globe to the farming community. Ludwika installed three pieces, each comprised of about 5km of cellophane, on a farm in Tuchikoi in the Mary Valley Region. She also installed one piece in Noosa Woods. Pictures after the jump
(Permalink: http://feeds.feedburner.com/ch?m=57)

Kommentator eins klassifizierte diesen Eintrag als „Kunst" und Kommentator drei als „Photographie". Die Wahl der Kennzeichnung „Kunst" ist sowohl für den Kommentator als auch für die tagbasierenden Methoden offensichtlich. Allerdings kommt nur ein einziger Hinweis in diesem Artikel auf Photographie vor. Die Interpretation von Kommentator drei erfordert also Kenntnis über Authoring-Konventionen im Blograum. Durch diesen Sachverhalt wird von [Bere07a] betont, dass Texte verschiedenartige Bedeutungen für verschiedene Personen haben können. In Bezug auf die Frage, wie Tags zusätzliche Inhalte hinzufügen können, argumentieren [Bere07a] dermaßen, dass nicht immer Informationen, die im Text fehlen, angefügt werden, sondern dass gewisse Meinungsaspekte, welche für manche Leser weniger relevant erscheinen beziehungsweise nicht präsent genug im Blog-Körper erscheinen, hervorgehoben werden. Dies ist mit der eindeutig machenden Funktion von Metadaten verwandt, jedoch ist es nicht dasselbe. Blogs sind darüber hinaus schwerer zu kategorisieren als Nachrichten. Ein möglicher Grund könnte darin liegen (abgesehen von der „Text-Qualität" und den referentiellen Unterschieden), dass der Blog-Inhalt an sich fließender und leserabhängiger als Nachrichtentexte ist. (Antworten auf die zweite der eingangs gestellten Fragen).
Weiters konnten [Bere07a] im Rahmen ihrer Analyse zeigen, dass Autoren-Tags unterschiedliche semantische Informationen beinhalten als der Body eines Blog-Eintrags (Antwort auf die erste Frage). In dieser Bedeutung entsprechen sie der folgenden Definition: In einem Korpus an Einträgen

bestehend aus Body-Elementen (text, title, etc.) und Autoren-Tags, können Tags unter folgenden Bedingungen als Inhalt und nicht als Metadaten gesehen werden. Die erste Bedingung lautet, dass Tags eine niedrige Ähnlichkeit mit dem Body in der Art und Weisen haben müssen, dass Body-Eigenschaften nicht dazu verwendet werden können, um Tags zu prognostizieren und umgekehrt. Die zweite Bedingung bezieht sich darauf, dass die Kombination aus Body und Tags besser dazu geeignet ist, die menschliche Übereinstimmungs-Klassifikation von Inhalten vorherzusagen als entweder Body oder Tags alleine.

Ein Vergleich von verschiedenen Text-Analysemethoden und Kommentatoren erlaubt es, Tag-Text-Beziehungen besser zu verstehen. Diese Beziehungen könnten Hinweise darüber geben, wie verschiedene Personen unterschiedliche Texte lesen. Diese Beobachtung ermöglicht ein besseres Verständnis über die Popularität von Tags Dritter wie beispielsweise *del.icio.us* und der Popularität von Tag-Auswahl basierend auf Communities. Eine Ausbeutung solcher Observierungen könnte Suchmaschinen und Community-Finder eleganter gestalten. Das strukturelle Verständnis von Leseweisen hinter Tag-Präferenzen könnte simplen Anpassungs- oder gemeinschaftlichen Filterungsoptionen überlegen sein. Suchmaschinen sollten Sorge sowohl für individuelle User als auch User-Gruppen, über welche sie Informationen besitzen, und für anonyme User, tragen. Allerdings traten User-Subgruppen auf, für welche Tags oder Body-Elemente als besonders brauchbare Inhalts-Indikatoren für Inhalt darstellen [Bere07a]. Diese Subgruppen ignorieren sogar einschlägige andere Elemente in ihrer Beurteilung worüber ein Blog handelt. Der Designprozess von Suchmaschinen oder Kennzeichnungs-Empfehlungssystemen sollte auf diese Tatsache besondere Rücksicht nehmen (Antwort auf die dritte Frage).

Auf die vierte der eingangs gestellten Fragen, konnten [Bere07a] folgende Antworten finden: Die Extraktion von Merkmalen aus verschiedenen „Beutel

voller Wörter“ ermöglicht Hinweise über Inhalt von Blog-Einträgen. Im Besonderen war die Analyse von Autoren-Tags und Body-Hauptwörtern nützlich. Allerdings wird die erforderliche Bearbeitung um gute Resultate (wie zum Beispiel POS-Kennzeichnung[16]) zu erzielen oftmals durch die niedrige syntaktische Qualität oder die sich dynamisch ändernde Netzsprache von Blogs behindert. Darüber hinaus muss der referentielle Inhalt eines Blog-Eintrages, welcher beispielsweise durch Hyperlinks erzeugt wird, in Betracht gezogen werden.

Die Studie nach [Bere07a] dient als gute Ausgangsbasis, wobei größere Stichproben von Kommentatoren und Inhalten notwendig wären. Um die gefundenen Annahmen über Lese-Strategien und Inhaltsbewertung mehr zu untermauern sind als zukünftige Forschungsarbeiten psycho-linguistische Methoden geplant. Weitere Erkenntnisse könnten durch experimentelle Studien über die Auswirkung von Merkmalen wie Sprache, Autor/ Leser Demographien und Tag-Systemattributen gewonnen werden.

[16] POS-Kennzeichnung beschreibt den Prozess der Zuordnung von einem Teil der Sprache wie einem Hauptwort, einem Verb, einer Präposition, einem Adverb oder anderen lexikalischen Klassen-Kennzeichnern zu jedem Wort in einem Satz [Aukb08].

Kapitel 4

4 Konzeption des Blog Suchmaschinen Prototyps

Wie bereits in der Kurzfassung sowie in Kapitel 1.2 erwähnt, bestand der praktische beziehungsweise experimentelle Forschungsteil der vorliegenden Arbeit darin, einen Prototyp einer Weblog-Suchmaschine zu entwickeln, welcher es dem User erleichtern soll gezielt nach einem bestimmten Thema zu suchen und eine qualifizierte Auswahl an Blogeinträgen zu dem gesuchten Thema anhand eines definierten Bewertungsschemas zu bekommen.
In diesem Kapitel wird zunächst die Herangehensweise an diese Problemstellung detailliert beschrieben. Wichtige theoretische Überlegungen bezüglich eines „effizienten" Suchalgorithmus sowie einer darauf aufbauenden Ergebnisstruktur sollen im Folgenden skizziert werden.

4.1 Aufbereitung der Blog-Testdatensätze

Noch bevor mit dem Entwicklungsprozess des Suchmaschinen-Prototyps begonnen werden kann, wurden folgende Überlegungen bezüglich der Vorverarbeitung der Daten angestrebt: Einerseits ist eine ausreichend große Anzahl an Blog-Einträgen notwendig, um den Suchalgorithmus in einer realistischen Umgebung anwenden zu können, Zu diesem Zweck wurden mehrere Blogseiten-Betreiber wie *blogger.de*, *twoday.net*, *blogverzeichnis.at*, etc. sowie zahlreiche Autoren, deren wissenschaftliche Artikel in dieser Arbeit analysiert wurden, um Testdatensätze gebeten. Dankenswerterweise erklärten sich die Betreiber von *blogbar.de* dazu bereit einen SQL-Dump[17] von rund 1000 älteren Blog-Einträgen (aus dem Jahre 2005) zu Verfügung zu stellen.

[17] Als SQL-Dump wird ein Programm bezeichnet, mit welchem eine Sicherheitskopie eines SQL Servers in Form einer Textdatei erstellt werden kann.

Da es sich bei der zu entwickelnden Suchmaschine um einen Prototyp handelt, wurde es nicht in Erwägung gezogen, die Suchfunktion dermaßen zu implementieren, dass in der Ergebnisstruktur verschiedene populäre Blogseiten miteinbezogen werden. Dies würde sich auch deshalb als sehr komplexer Vorgang erweisen, da gezielt auf die HTML-Struktur der jeweiligen Blogseite zugegriffen werden müsste, um die eigentlichen Blog-Einträge herauszufiltern. Aufgrund dessen, dass diese HTML-Struktur bei jeder Blogseite anders aussieht und die Einträge folgedessen immer irgendwo anders zu finden sind, wird bereits die darin liegende Schwierigkeit ersichtlich. Um diesen Prozess auf den Prototyp zu adaptieren wurde nach einer Möglichkeit des direkten Zugriffs auf die jeweiligen Blog-Einträge (ebenfalls notwendig um in weiterer Folge effektiv mit diesen arbeiten zu können) gesucht. Zu dessen Realisierung wurde im ersten Schritt eine MySQL Datenbank (eine der populärsten Open-Source-Datenbanken) angelegt. Bei MySQL handelt es sich genau genommen um ein relationales Datenbankverwaltungssystem, welches die Grundlage für viele dynamische Webauftritte bildet. MySQL wird sehr häufig in Verbindung mit PHP oder dem Apache Webserver verwendet. Darüber hinaus bietet die Open Source Anwendung „phpMyAdmin“ (Version 2.11.4) die Möglichkeit, solche MySQL-Datenbanken oder komplette MySQL-Server zu verwalten. Die Administration erfolgt bequem über HTTP (bezeichnet allgemein ein Protokoll zur Übertragung von Dateien über Netzwerke; findet besondere Anwendung bei der Kommunikation im Web) mit einem Webbrowser. Tabelle 4.1 zeigt die Tabellenstruktur innerhalb der Datenbank für den zu entwickelnden Blog-Suchmaschinen Prototyp.

name	nr	inhalt	datum	views	titel	scientific	emotional	scientific_e	emotional_e
Bernd	1010	Der Streifzug…	2005-08-31 09:26:07	450	Screensport: der Tag …	6	5	6	5
Bernd	1286	Damit stehen …	2005-01-13 11:46:14	510	Morientes geht …	10	2	9	2
Bernd	1510	Vor knapp einem…	2005-01-11 10:47:24	192	ALBA feuert…	4	3	4	3
Bernd	1300	Die Schlagzeilen…	2005-09-16 08:09:27	320	Zeilensport mit …	5	5	4	6
Bernd	2005	Wenn jemand...	2005-09-19 07:49:56	150	NASNs Pressemitteilung …	7	1	8	1

Tabelle 4.1 Tabellenstruktur der Blog-Datenbank

Insgesamt wurden zehn Spalten in dieser Tabelle angelegt, von denen die meisten eigentlich selbsterklärend sein sollten. Die Spalte „name“ (Typ: VARCHAR) gibt Auskunft über den Autor eines Blog Eintrags. Die nächste Spalte namens „titel“ (Typ: TEXT) trägt die Überschrift eines Artikels. Die dritte Spalte namens „inhalt“ (Typ: LONGTEXT) enthält die zur Verfügung gestellten Blog-Einträge. Als nächstes wurde eine Spalte „datum“ (Typ: DATETIME) angelegt, welche den individuellen Zeitstempel beinhaltet. Die Spalte „nr“ (Typ: BIGINT) dient als sogenannter Primärschlüssel, damit die Werte in der Tabelle eindeutig bestimmt werden können. Besondere Aufmerksamkeit verdienen vor allem die Spalten „views“ (Typ: INT; enthält fiktive Werte), „scientific“ (Typ: INT) und „emotional“ (Typ: INT) sowie die Spalten „scientific_e“ (Typ: INT) und „emotional_e“ (Typ: INT). Sie stellen wesentliche Parameter im Suchalgorithmus dar. In Kapitel 3.5 wurde bereits darauf

eingegangen, welche tragende Rolle Tags in Zusammenhang mit Blog-Einträgen spielen können. Diese Tatsache wurde auch in der Entwicklung des Suchmaschinen-Algorithmus berücksichtigt. Die Werte in der „scientific" und „scientific_e" Spalte geben Auskunft darüber wie wissenschaftlich/ informativ ein Eintrag geschrieben ist, während hohe Werte in der Spalte „emotional" und „emotional_e" auf eher affektive Inhalte hindeuten. Der Benutzer hat somit die Möglichkeit seine Suche beziehungsweise die zurückgelieferten Ergebnisse gezielt in Richtung wissenschaftlicher oder emotionaler Blog-Artikel zu lenken.
Die Blog-Testdatensätze wurden direkt in die bereits angelegte Datenbank integriert. Mittels der „Importieren" Funktion von „phpMyAdmin" ist es relativ problemlos möglich eine derartige Textdatei zu importieren. Da die weiter oben beschriebene Datenbank bereits vor dem Import der Datensätze aus *blogbar.de* erstellt wurde, bestand die einzige Schwierigkeit darin, die Reihenfolge der Spalten innerhalb der Testdatenbank von *blogbar.de* derart zu verändern, dass sie mit derjenigen Reihenfolge innerhalb dieser primär angelegten Datenbank übereinstimmen. Diese Umstrukturierung war auch deshalb notwendig, da zu Projektbeginn bereits mit der angelegten Datenbank gearbeitet wurde und die Reihenfolge der Tabellenspalten in einigen Projektdateien ausschlaggebend ist.

Der zweite Schritt besteht darin, diese Blog-Einträge für die spätere Verwendung innerhalb der Suchmaschine aufzubereiten. Wie oben erwähnt, wurde die Bedeutung von Tags im Rahmen des zu entwickelnden Such-Algorithmus durch zwei Tag-Kategorien in der Datenbank berücksichtigt. Um die Tag-Spalten „scientific" und „emotional" sowie „scientific_e" und „emotional_e" mit Werten zu befüllen, wurden zwei verschiedene Methoden gewählt: Zum einen wurden 20 Personen gebeten, jeweils 10 bis 15 Einträge zu lesen und anschließend in den beiden Feldern (scientific und emotional) ihre subjektive Bewertung auf einer Skala von eins bis 10 über den

wissenschaftlich/ informativen und affektiven Inhalt des jeweiligen Blog-Artikels einzutragen. Nach Drücken des „Weiter“ Button kommen die Testpersonen zum nächsten zu bewertenden Eintrag und die von ihnen eingegebenen Werte werden in der Datenbank gespeichert. Diese Blog-Labelling Seite ist unter folgender URL erreichbar: *http://www.youth-art.net/bernd/bloglabel.php.* Aus dieser ersten Bewertungsmethode resultierten 220 Blog-Einträge. Dieselben 220 Einträge (jedoch ohne Markierung) wurden mit Hilfe der Data Mining Software WEKA und dem darin implementierten Naïve Bayes Klassifizierungs-Algorithmus dieser Bewertung unterzogen, um auch die beiden Spalten „scientific_e“ und „emotional_e“ mit Werten zu befüllen. Im Rahmen des Klassifizierungsprozesses mittels WEKA wurde zu Beginn eine Datenbank-Verbindung zur angelegten Blog-Datenbank hergestellt. Als Inputdaten dienten die ersten 20 Blog-Einträge (die Spalten „inhalt“, „scientific“ und „emotional“), welche durch die Testpersonen klassifiziert wurden. Diese wurden in WEKA importiert. Es standen nun drei Attribute zur Verfügung, mit welchen der Klassifizierungsprozess realisiert werden kann. Abbildung 4.1 zeigt die Preprocess Ansicht in WEKA nachdem die ersten 20 Instanzen importiert wurden. Die Balkengrafik rechts unten stellt die Verteilung der Werte des jeweils ausgewählten Attributes dar.

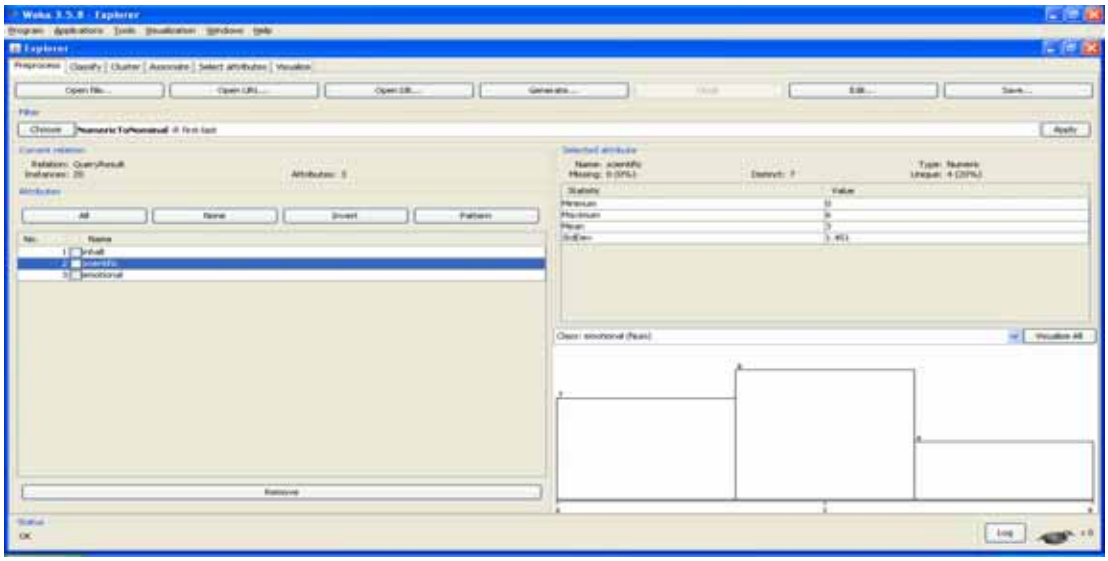

Abbildung 4.1 Preprocess Ansich in WEKA nach dem Import der Daten

Im nächsten Schritt muss einer der zahlreichen, in WEKA implementierten, Filter ausgewählt werden. Der „NumericToNominal“ Filter (dieser ist unter den unsupervised Attribut-Filter zu finden) wurde deshalb angewandt, da der Naïve Bayes Klassifizierer nicht mit numerischen Klassen umgehen kann. Der Reiter „Classify“ führt zur Auswahl eines Klassifizierungs-Algorithmus. Naïve Bayes wurde auf die 20 Instanzen, welche als Trainingsdatensätze dienen, angewandt. Die in einem Blog Eintrag vorkommenden Wörter werden auf die Kategorien „scientific“ und „emotional“ abgebildet. Es wurde je ein Klassifizierungsdurchgang mit dem „emotional“ Attribut als auch mit dem „scientific“ Attribut in ihrer Funktion als Klassenattribut durchgeführt. In dem Fall von „emotional“ als Klassenattribut erreichte der Klassifizierer 95 Prozent an korrekt klassifizierten Instanzen und lediglich 5 Prozent an inkorrekt klassifizierten Instanzen, und im Fall von „scientific“ waren es 90 Prozent an korrekt klassifizierten Instanzen und 10 Prozent an inkorrekt klassifizierten Instanzen. Das, auf diesen Trainingsdatensätzen, angewandte Modell wurde gespeichert, um es im darauffolgenden Schritt für die Klassifizierung der verbleibenden 200 Instanzen zu verwenden. Unter Anwendung auf die 200 Testdatensätze und „emotional“ als Klassenattribut erzielte Naïve Bayes 75 Prozent an korrekt klassifizierten Instanzen und 25 Prozent an inkorrekt klassifizierten Instanzen. Bei „scientific“ als Klassenattribut erreicht er 82 Prozent an korrekt klassifizierten Instanzen und 18 Prozent an inkorrekt klassifizierten Instanzen. Die durch den Klassifizierer vorausberechneten Werte wurden in die beiden Datenbank-Spalten „scientific_e“ und „emotional_e“ für alle 200 Blog-Einträge eingefügt.

Es stehen nun „zwei“ Blog-Testdatensätze zur Verfügung, wobei es sich bei dem einen um eine menschliche Zuordnung beziehungsweise Gewichtung der jeweiligen Einträge zu den beiden Tag-Spalten und bei dem anderen um eine Zuordnung durch einen maschinellen Klassifizierungs-Algorithmus handelt (die jeweiligen Blog-Artikel sind allerdings vollkommen identisch). In Kapitel fünf wird gezeigt, welche Blog-Einträge aus diesen „beiden“ Datensätzen als

Ergebnisse auf diverse eingegebene Suchbegriffe durch den Suchmaschinen-Prototyp zurückgeliefert werden.

4.2 Vorgehensweise im Rahmen des Suchmaschinen-Prototyps

Als nächster Schritt im Zuge der gedanklichen Forschungsarbeit am Sektor der Blog-Suchmaschinen Entwicklung wurde neben der Recherche nach existierenden Suchmaschinen vor allem die Frage nach den „wichtigsten" Suchparametern/ Suchkriterien in Zusammenhang mit Blogeinträgen aufgegriffen. In Kapitel drei wurden bereits einige wesentliche Analysemethoden, welche heutzutage eine breite Anwendung finden, vorgestellt. Ähnlich den Überlegungen nach relevanten Rahmenbedingungen für eine zweckmäßige Blog-Suchmaschine, basieren die vorgestellten Analyseverfahren größtenteils ebenfalls auf diesen Parametern. Betrachtet man einzelne Blog-Einträge beziehungsweise eine komplette blogosphärische Gemeinschaft lassen sich aufgrund diverser wissenschaftlicher und experimenteller Erkenntnisse folgende „erfolgsbestimmende" Faktoren bestimmen:

- Blog-Post: Einer der wichtigsten Parameter in diesem Zusammenhang beschreibt den eigentlichen Inhalt der jeweiligen Blogeinträge. Der, in dieser Arbeit entwickelte, Suchalgorithmus basiert unter anderem darauf, dass ein prozentualer Wert ermittelt wird, welcher angibt wie oft der abgefragte String (also der Suchbegriff) im Vergleich zur Gesamtanzahl der Zeichen des jeweiligen Blog-Eintrages vorkommt. Ein Eintrag, welcher beispielsweise nur 150 Zeichen lang ist, den abgefragten Begriff jedoch viermal enthält wird in der Regel informativer sein, als ein doppelt so langer Artikel indem der Suchbegriff nur halb so oft vorkommt. Dieser Schritt im Suchalgorithmus

ermöglicht bereits eine gefilterte Auswahl an informativen beziehungsweise relevanten (wobei diese Wertung als subjektiv empfunden werden kann) Blog-Einträgen.

- Views: Die Bezeichnung Views gibt diejenige Anzahl an, wie oft ein einzelner Blog-Eintrag aufgerufen (komplett durchgelesen oder nur überflogen) wurde. Da Blog-Einträge, die sehr viele Views aufweisen, mit großer Wahrscheinlichkeit einen hohen Informationsgehalt besitzen beziehungsweise von breitem Interesse für einen speziellen Leserkreis sind, beeinflusst genau dieser View-Wert das Ranking der Such-Ergebnisstruktur und stellt den nächsten entscheidenden Schritt innerhalb des zu entwickelnden Suchalgorithmus dar.

- Kommentare: Blog-Einträge bieten dem Leser die Möglichkeit einen Kommentar darüber abzugeben. Bei Kommentaren verhält es sich ähnlich wie bei Views. Auch in diesem Fall kann in einem weiten Anwendungsbereich davon ausgegangen werden, dass Einträge mit einer hohen Anzahl an Kommentaren interessanter, informativer oder zumindest kontroversieller sind als andere Artikel zum selben Thema. Eine quantitative sowie qualitative Bewertung einer Such-Ergebnisstruktur anhand von Kommentaren wäre demzufolge ebenfalls denkbar. Bei Kommentaren kommt jedoch noch der Aspekt einer Form von Community-Bildung hinzu. Aufgrund der verschiedenen Reaktionen auf einen bestimmten Eintrag können Rückschlüsse auf die Zuordnung des Lesers/ der Leserin zu einer Community getroffen werden.

- Soziale Netzwerke/ Communities: Blogosphärische Gemeinschaften bieten ihren Bloggern oftmals die Möglichkeit, soziale Netzwerke

jeglicher Art untereinander zu schließen. Die Kommentar-Funktion beispielsweise lässt, wie vorhin bereits erwähnt, Rückschlüsse auf Verbindungsstrukturen zu. Die Berücksichtigung dieser sozialen und inhaltlichen Verbindungen würde möglicherweise die Auswahl und die Qualität an Blog-Suchmaschinenergebnissen erhöhen. Auf der anderen Seite würde sich der Suchalgorithmus weitaus komplexer gestalten, was sich wiederum (bei großen Datensätzen) negativ auf die Berechnungszeit auswirkt.

- Erstellungsdatum: Als weiterer relevanter Parameter in Zusammenhang mit Weblogs ist der Zeitpunkt des Verfassens eines Eintrages zu sehen. Besonders bei technisch wissenschaftlichen Themengebieten besitzen ältere Beiträge in der Regel einen anderen Bezug zu diesem Thema als aktuelle Einträge. Die Integration der jeweiligen Zeitstempel in den Suchalgorithmus erweist sich als schwieriges Unterfangen, da die Umrechnung der verschiedenen Datumsangaben in einen prozentualen Wert nicht wirklich objektivierbar ist.

- Tags/ Metadaten: In Kapitel 3.5 wurde bereits auf die Relevanz von Tags eingegangen. Tags können ebenfalls starken Einfluss auf Blog-Einträge ausüben und unter Umständen für völlig unterschiedliche Sichtweisen auf ein und denselben Blog-Eintrag sorgen. Ein weiterer obligatorischer Task in der Entwicklung des Suchalgorithmus spiegelt die Berücksichtigung dieser Tags. Wie die Berücksichtigung dieser Parameter jedoch im Detail aussieht, wird weiter unten erklärt werden.

Im nachfolgenden Kapitel werden einerseits Entwicklungsstrategien sowie andererseits das Konzept für den Suchmaschinen-Prototyp anhand der vorgestellten Parameter besprochen. Entscheidungen bezüglich einer passenden Programmiersprache, einer benutzerfreundlichen grafischen Darstellung sowie einer „ausgewogenen" Berechnung durch den Suchalgorithmus werden getroffen. Darüber hinaus wird detailliert auf die Gesamtstruktur des zu entwickelnden Suchsystems eingegangen, um den Leser und Leserinnen einen fundierten Einblick in den systematischen Ablauf sowie in den Umfang des Projektes zu geben. Zum Abschluss werden exemplarisch einige Testergebnisse aus dem Suchmaschinen-Prototypen beschrieben.

4.3 Entwicklungsstrategie des Suchmaschinen-Algorithmus

Aufgrund der Vielzahl und Diversität an Parameter die es zu beachten gibt, gestaltet sich die Entwicklung einer effizienten und benutzerorientierten Suchmaschine als sehr komplexer Vorgang. Die Fragen, die unter anderem in diesem Zusammenhang auftauchen, lauten: Welche Parameter gilt es in den Suchalgorithmus zu integrieren (wurde bereits kurz erwähnt) beziehungsweise wie könnte eine sinnvolle und effiziente Gewichtung dieser Parameter innerhalb des Suchalgorithmus aussehen? Weiters stellt sich natürlich auch die Frage, wie dieser Prototyp technisch gesehen umgesetzt werden kann: Welche Programmiersprache kann beziehungsweise sollte verwendet werden? Wie erfolgt die Suchabfrage? Wie gestaltet sich die grafische Umsetzung der Suchergebnisse, damit diese möglichst benutzerfreundlich dargestellt werden (zum Beispiel die Anzeige im Webbrowser)?
Antworten auf all diese Fragen werden im Folgenden systematisch erarbeitet. Als kleine Inspiration während der Entwicklungsphase wurden einige

Suchmaschinen getestet. Tabelle 4.2 zeigt eine Auswahl an populären Blog-Suchmaschinen mit ihren zugehörigen URLs.

Blog-Suchmaschine	URL
Technorati	http://technorati.com/
Yahoo Search Blog	http://www.ysearchblog.com/
Google Blog-Such Beta	http://blogsearch.google.com/
Blog Search Engine	http://www.blogsearchengine.com/
Blogato	http://www.blogato.net/
Blog-Sucher	http://www.blog-sucher.de/
BlogPulse	http://www.blogpulse.com/

Tabelle 4.2 Auswahl einiger bekannter Blog-Suchmaschinen

Diese bekannten Suchmaschinen spiegeln sehr deutlich den momentanen Stand auf dem Gebiet der Blog-Suchmaschinen wider. Die Suche gestaltet sich teilweise recht unterschiedlich: Unter Verwendung des Suchbegriffes „Krieg" lieferte *blogato.de* sehr brauchbare Ergebnisse. Es werden Auszüge von Blog-Einträgen verschiedener Blogseiten angezeigt. Die Suche dauerte 0,454 Sekunden und ergab insgesamt 5692 Suchergebnisse. Die Suche auf *technorati.com* nach demselben Begriff dauerte zum Einen länger (1,7 Sekunden) lieferte jedoch 6464 Suchergebnisse. Allerdings waren die Ergebnisse alles andere als zufriedenstellend, da die ersten vier Artikel meiner Meinung nach nur sehr bedingt von dem Begriff „Krieg" handeln. Die Blog-Suchmaschine von Google hingegen erzielte zu dem abgefragten Begriff (ebenfalls „Krieg") weitaus suchrelevantere Ergebnisse (es wurden 424465

Ergebnisse in 0,12 Sekunden zurückgeliefert) als beispielsweise *blogsearchengine.com*, bei welcher zwar 35038 Ergebnisse gefunden wurden, die ersten drei bis vier Artikel wiederum nur sehr begrenzt das Thema „Krieg" beinhalten. Über *blog-sucher.de* gibt es zu sagen, dass die Anzahl an Suchergebnissen zwar gering ist (es wurden vier Einträge aus unterschiedlichen Blog-Seiten zurückgeliefert), die Qualität der gefundenen Blog-Einträge allerdings durchaus meinen Erwartungen entspricht. Alle vier Artikel (nur der erste Artikel-Link verweist auf eine leere Blog-Seite) behandeln den Begriff „Krieg" in einer erwarteten Form.

Die erste der eingangs gestellten Fragen die es zu beantworten gilt ist diejenige nach den technischen Gesichtspunkten der Blog-Suchmaschine. Derzeit existieren diverse Skriptsprachen wie Perl, Python, VBScript, Ruby, etc., welche allesamt mehr oder weniger für die Umsetzung dieses Projektes geeignet wären. Dennoch fiel die Wahl von Anfang auf eine weitere Skriptsprache namens PHP. Diese Sprache kann einerseits problemlos in HTML-Code eingebunden werden und andererseits, wie auch einige andere Sprachen, sehr einfach an eine Datenbank angeknüpft werden. PHP stellt sozusagen ein leistungsfähiges und doch auch benutzerfreundliches Werkzeug zum Erstellen dynamischer Webinhalte dar. Darüber hinaus läuft PHP sowohl unter Unix, als auch unter Windows und Mac OS X [Trac05] und wurde unter anderem auch aus Sympathiegründen gewählt. Abbildung 4.2 dient zur weiteren Beantwortung beziehungsweise einfachen Visualisierung der Frage nach der technischen Realisierung sowie der Struktur des Suchmaschinen-Prototyps. Die einzelnen .php Projektdateien sowie die angelegte MySQL-Datenbank wurden auf einem Testserver installiert. Dieser Testserver besteht einerseits aus einem WWW-Server, welcher die Verwaltung der .php Dateien übernimmt, sowie andererseits aus einem MySQL-Server, welcher die Datenbank verwaltet.

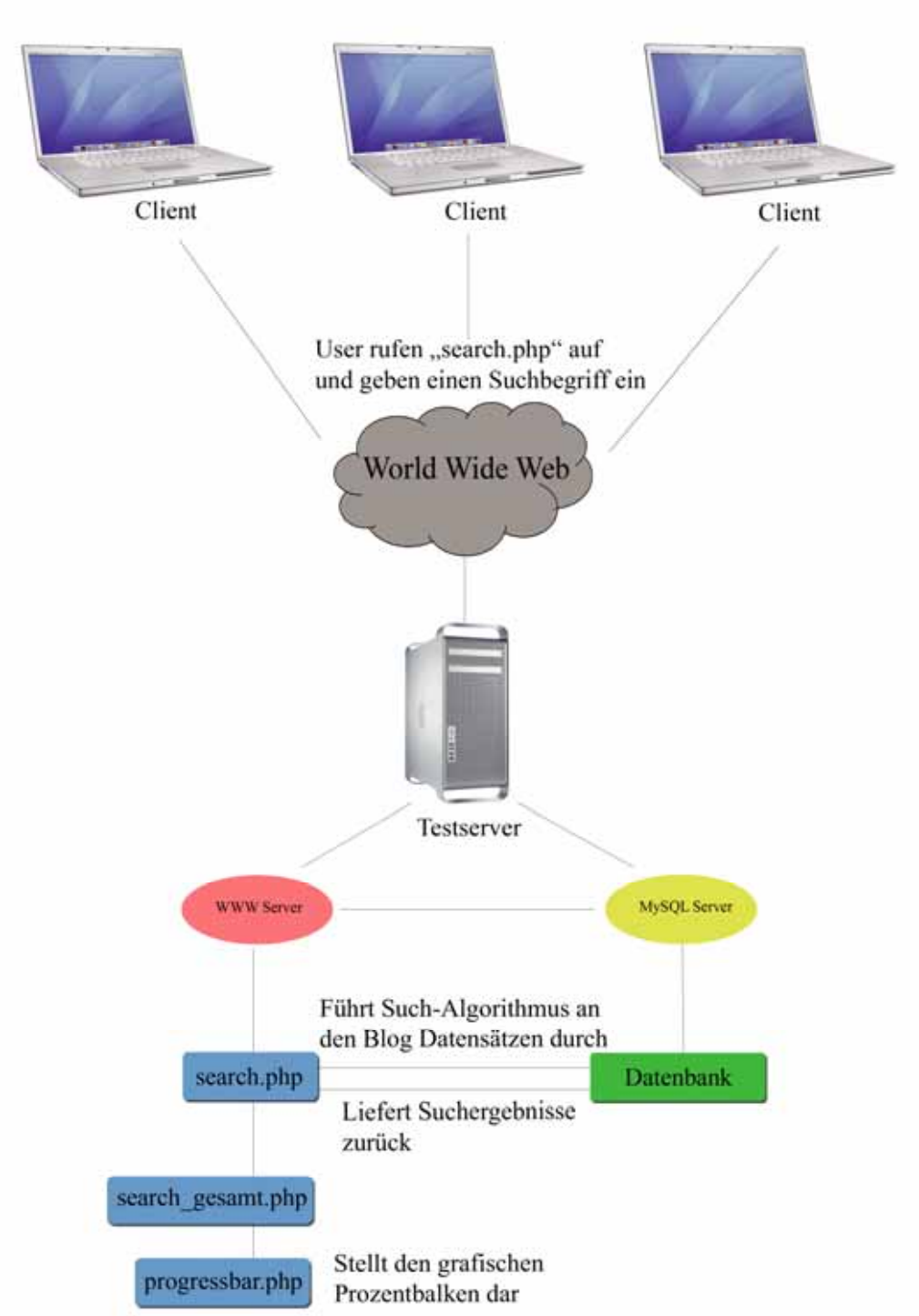

Abbildung 4.2 Gesamtstruktur des Suchmaschinen Prototyps

Dieser Grafik ist deutlich zu entnehmen, dass an der Umsetzung des Suchmaschinen Prototyps-Gesamtsystem mehrere einzelne Komponenten beteiligt sind.

Eingangs wurde bereits die Frage nach der eigentlichen Suchabfrage innerhalb des Suchmaschinen-Prototyps gestellt. Im Folgenden wird sukzessive erklärt wie der User von der Eingabe eines Suchbegriffes zur endgültigen Darstellung der Suchergebnisse kommt. Darüber hinaus wird detailliert auf den Suchalgorithmus eingegangen. Als erste wichtige Aufgabe und gleichzeitig Grundvoraussetzung für den, im Rahmen dieser Arbeit entwickelten, Suchalgorithmus steht die Befüllung der Datenbankspalten „scientific" und „emotional" sowie „scientific_e" und „emotional_e" im Vordergrund. Nachdem diese Spalten den beiden Kategorien mit entsprechenden Werten zugeordnet wurden (in Kapitel 4.1 beschrieben) kommt die wohl wichtigste Datei im gesamten Suchprozess ins Spiel. Der User ruft die Datei „search.php" auf dem Testserver auf (*http://www.youth-art.net/bernd/search.php*) und gelangt zum Sucheingabe-Feld. Auf dieser Webseite hat er/sie die Möglichkeit einen Suchbegriff einzugeben. Des Weiteren ist es möglich eine Wertung der Wichtigkeit in Hinblick auf eher wissenschaftliche beziehungsweise emotionelle Artikel zu vergeben. Diese Wertung wird maßgeblich in den Suchalgorithmus miteinbezogen. Werden diese beiden Felder jedoch nicht ausgefüllt, wird auch keinerlei Rücksicht auf die inhaltliche Tendenz der Blog-Einträge innerhalb des Abfrageprozesses genommen. Wird nun der „Suche1" oder der „Suche2" Button gedrückt folgt die Ausführung des eigentlichen Suchalgorithmus:

```
$proz = ($count_search*1000/$countwords) +
(($zeile[4]*100/$hoechsteviews)/1.5) +
$_GET['scientific']*$zeile[6]/6 + $_GET['emotional']*$zeile[7]/6
```

Der Algorithmus setzt sich aus insgesamt vier Berechnungen zusammen, welche additiv zu einem Gesamtergebnis führen. Sämtliche Multiplikations- beziehungsweise Divisionswerte wurden durch beispielhaftes Ausprobieren unterschiedlichster Suchbegriffe definiert. In Kapitel 4.2 wurde bereits erwähnt, dass das Verhältnis des abgefragten Suchbegriffes in einem Blog-Eintrag zu dessen Wort-Gesamtanzahl als entscheidendes Qualitätsmerkmal für die Suchergebnisse anzusehen ist. Der erste Klammerausdruck berechnet genau dieses Verhältnis. Ein weiterer Beurteilungsmaßstab innerhalb des Suchalgorithmus bezieht sich auf die Anzahl der Views. Hohe Werte in der „views" Datenbankspalte deuten in der Regel auf interessante Artikel hin. Im zweiten Klammerausdruck der Additionskette wird auf diesen View-Wert folgendermaßen Rücksicht genommen: Die Blog-Einträge in der Datenbank werden, genau wie im ersten Klammerausdruck, dahingehend gefiltert, ob sie den eingegebenen Suchbegriff enthalten. Kommt nun dieser Suchbegriff beispielsweise in 20 Blog-Einträgen vor, bezieht sich die Variable „$hoechsteviews" auf den höchsten View-Wert innerhalb dieser 20 Einträge („$hoechsteviews" bildet sozusagen die 100 Prozent Marke). Es wird bei jedem dieser Einträge („$zeile[4]" bezieht sich auf den jeweiligen „views" Wert) das individuelle Verhältnis des höchsten View-Wertes zu dem jeweiligen View-Wert berechnet.

Nach den ersten beiden Additionen werden bereits in etwa 70 Prozent der gesamten Suchrelevanz abgedeckt. Der mittlerweile mehrfach erwähnten Bedeutung von sogenannten Tags innerhalb eines benutzerorientierten Suchalgorithmus wird in den verbleibenden 30 Prozent des Gesamtergebnisses Rechnung getragen. Die beiden Tag-Spalten „scientific" und „emotional" werden durch die Variablen „$zeile[6]" und „$zeile[7]" (stellen jeweils 100 Prozent dar) in das mehrdimensionale Array eingetragen. Beide Berechnungen dürfen jeweils die 15 Prozent-Grenze nicht überschreiten, um im Rahmen der verbleibenden 30 Prozent auf das Gesamtergebnis zu

bleiben. Aus diesem Grund wird bei beiden Rechnungen die Division durch den Wert sechs durchgeführt.
Um die Ergebnisstruktur so übersichtlich als möglich zu gestalten, werden die fünf suchrelevantesten Suchbegriffe bei jedem Suchdurchlauf angezeigt. Wie bereits erwähnt, wird der Wert aus diesem Algorithmus in der Variable „$proz" gespeichert, welche der Datei „progressbar.php" übergeben wird. Dies ist deshalb notwendig, um die jeweilige prozentuelle Relevanz der gefundenen Blog-Artikel auch grafisch darstellen zu können.

Die Ergebnisse aus den Suchdurchläufen der jeweiligen Testpersonen unter Anwendung des entwickelten Such-Algorithmus werden im Folgenden beispielhaft dargestellt. Die angegebenen Relevanzwerte (Relevanz bedeutet in diesem Zusammenhang, dass diejenigen zurückgelieferten Blogeinträge mit den höchsten Relevanzwerten die eingegebenen scientific-und emotional Werte inhaltlich am besten widerspiegeln) sowie scientific/informative Werte (abgekürzt mit „s") und emotional Werte (abgekürzt mit „e") beziehen sich jeweils auf den ersten der zurückgelieferten Blog-Einträge.

Testperson1:
Eingegebener Suchbegriff: „Basketball"
Eingegebener Scientific Wert: 7
Eingegebener Emotional Wert: 1

Ergebnisse über „Suche1" Button: 73,90 Prozent Relevanz (s.: 6, e.: 6)
Ergebnisse über „Suche2" Button: 73,90 Prozent Relevanz (s.: 6, e.: 6)

Im Falle von „Basketball" waren die zurückgelieferten Ergebnisse den Erwartungen entsprechend, da diese durchwegs informative Artikel beinhalteten. Allerdings sind hier, wie auch bei einigen anderen Suchbegriffen, die zurückgelieferten Suchergebnisse aufgrund ihrer teilweise

unterschiedlichen Bewertungen der Tag-Spalten und den daraus resultierenden Relevanzwerten verschiedenartig angeordnet. Dies lässt sich sehr schön am Beispiel des Eintrages mit dem Titel „Nowitzki für 7-10 Tagen out“ zeigen: Bei den Suchergebnissen, welche durch Drücken des „Suche1“ Button zurückgeliefert wurden, steht dieser Eintrag (s.: 5 und e.: 1) an vorletzter Stelle während er (s.: 1 und e.: 1) bei den Suchergebnissen, die durch Drücken des „Suche2“ Button gefunden wurden (es handelt sich hierbei um exakt die gleichen Einträge), an letzter Stelle steht. Von der Testperson wurde dieser Eintrag als informativer und weniger emotionell bezeichnet als der Eintrag mit dem Titel „Screensport: Hasse mal'ne Mark?“. In Anbetracht dessen, dass dieser Artikel im Fall des „Suche1“ Button an letzter Stelle gereiht ist und im Fall des „Suche2“ Button an vorletzter Stelle gereiht ist, empfindet diese Testperson die Verteilung der Relevanzwerte über den „Suche1“ Button besser und kann sich in diesem einen Fall stärker mit der Zuordnung zu den beiden Kategorien durch eine menschliche Klassifizierung identifizieren als mit einer Klassifizierung durch den Maschinenlern-Algorithmus.

Testperson2:
Eingegebener Suchbegriff: „Fußball“
Eingegebener Scientific Wert: 10
Eingegebener Emotional Wert: 4

Ergebnisse über „Suche1“ Button: 83,25 Prozent Relevanz (s.: 10, e.: 3)
Ergebnisse über „Suche2“ Button: 80,88 Prozent Relevanz (s.: 7, e.: 2)

Der erste Eintrag, welcher über den „Suche1“ Button zurückgeliefert wird, ist wesentlich länger und etwas informativer als der erste Eintrag, welcher über den „Suche2“ Button zurückgeliefert wird. Dieser entspricht mehr den Erwartungen der gewählten Tag-Werte, welche darauf hindeuten, dass die

zurückgelieferten Ergebnisse stark in Richtung informativer Artikel tendieren sollen. Diese Testperson kann sich im Falle des ersten zurückgelieferten Blog-Eintrages ebenfalls stärker mit der Zuordnung zu den beiden Kategorien durch eine menschliche Klassifizierung identifizieren als mit einer Klassifizierung durch den Maschinenlern-Algorithmus. Sie empfindet den Wert 3, welcher für die scientific_e Kategorie berechnet wurde, als zu niedrig, da dieser Artikel einen sehr hohen Informationsgehalt aufweist.

Testperson3:
Eingegebener Suchbegriff: „Auto“
Eingegebener Scientific Wert: 2
Eingegebener Emotional Wert: 9

Ergebnisse über „Suche1“ Button: 75,41 Prozent Relevanz (s.: 8, e.: 3)
Ergebnisse über „Suche2“ Button: 75,41 Prozent Relevanz (s.: 8, e.: 3)

Im Großen und Ganzen wurden die Einträge, welche über den „Suche1“ Button gefunden wurden, als emotioneller betrachtet und entsprechen deshalb mehr den Erwartungen. Diese Tatsache spiegelt sich auch in den Relevanzwerten wider, da der letzte Eintrag der Ergebnisse aus „Suche1“ immer noch eine einen Relevanzwert von 62,2 Prozent aufweist im Vergleich zum letzten Eintrag der Ergebnisse aus „Suche2“, welcher einen Wert von 56,38 Prozent hat.

Testperson4:
Eingegebener Suchbegriff: „Manager“
Eingegebener Scientific Wert: 10
Eingegebener Emotional Wert: 1

Ergebnisse über „Suche1“ Button: 79,27 Prozent Relevanz (s.: 7, e.: 2)
Ergebnisse über „Suche2“ Button: 79,27 Prozent Relevanz (s.: 7, e.: 2)

Bei beiden Suchdurchläufen wurden exakt die gleichen Suchergebnisse zurückgeliefert. Dieser Testperson zufolge wurden gute, sachliche Ergebnisse, das heißt nüchterne Kommentare und keine Beschimpfungen, etc. erzielt. Bei diesen fünf zurückgelieferten Blog-Artikel wurden durch beide Klassifizierungsmethoden dieselben Tag-Werte vergeben beziehungsweise berechnet.

Testperson5:
Eingegebener Suchbegriff: „München“
Eingegebener Scientific Wert: 1
Eingegebener Emotional Wert: 10

Ergebnisse über „Suche1“ Button: 77,46 Prozent Relevanz (s.: 4, e.: 6)
Ergebnisse über „Suche2“ Button: 77,30 Prozent Relevanz (s.: 3, e.: 6)

Es wurden sowohl über den „Suche1“ als auch über den „Suche2“ Button dieselben Blog-Einträge gefunden, allerdings zeigt sich gleich beim ersten zurückgelieferten Artikel (dieser ist in beiden Fällen exakt der gleiche), dass dieser unterschiedliche Tag-Werte aufweist und sich dadurch auch ein anderer Relevanzwert aufgrund der Berechnung durch den Such-Algorithmus ergibt. Laut dieser Testperson wurden überwiegend informative Artikel gefunden, allerdings mit klarer emotionaler Färbung. Aufgrund der, durch die Testperson, eingegebenen Scientific- und Emotional Werte sind die zurückgelieferten Ergebnisse in beiden Suchdurchläufen nur bedingt zufriedenstellend.

Testperson6:
Eingegebener Suchbegriff: „Europameisterschaft“
Eingegebener Scientific Wert: 3
Eingegebener Emotional Wert: 9

Ergebnisse über „Suche1“ Button: 81,42 Prozent Relevanz (s.: 5, e.: 8)
Ergebnisse über „Suche2“ Button: 81,42 Prozent Relevanz (s.: 5, e.: 8)

Bei diesem Suchbegriff wurde über beide Such-Buttons der gleiche Blog-Artikel zurückgeliefert. Hier stimmen auch die Relevanzwerte überein. Dies ist ein Indiz dafür, dass bei beiden Klassifizierungsmethoden dieselben Tag-Werte vergeben beziehungsweise berechnet wurden. Der Kommentar der Testperson zum Inhalt dieses Artikels lautet: „Der gefundene Artikel ist ganz okay“.

Testperson7:
Eingegebener Suchbegriff: „liebe“
Eingegebener Scientific Wert: 2
Eingegebener Emotional Wert: 7

Ergebnisse über „Suche1“ Button: 76,80 Prozent Relevanz (s.: 2, e.: 8)
Ergebnisse über „Suche2“ Button: 71,09 Prozent Relevanz (s.: 7, e.: 2)

Es wurden zwar fünf gleiche Blog-Artikel zurückgeliefert, allerdings unterscheiden sich die einzelnen Einträge der beiden Datensätze in ihrem Relevanzwert von einander. So hat beispielsweise der erste zurückgelieferte Eintrag (s.: 2 und e.: 8) in dem einen Datensatz, welcher über den „Suche1“ Button gefunden wurde, einen Relevanzwert von 76,8 Prozent während der gleiche Eintrag (s.: 2 und e.: 2) aus dem Datensatz, welcher über den „Suche2“

Button zurückgeliefert wurde, einen Relevanzwert von 69,8 Prozent. Dieser Eintrag hat zwar nicht wirklich das Thema „Liebe" zum Inhalt, allerdings wurde er als weitaus emotionaler als informativer von dieser Testperson bezeichnet. Die berechneten Tag-Werte durch den Maschinenlern-Algorithmus korrespondieren somit nicht mit den Ansichten der Testperson, weshalb der hohe emotional Wert (resultierend aus der menschlichen Klassifizierung) in diesem Fall eindeutig die inhaltliche Tendenz dieses Blog-Artikels besser widerspiegelt.

Über die hier dargestellten Ergebnisse aus dieser kleinen Studie muss allerdings gesagt werden, dass diese nur bedingte Aussagekraft in Bezug auf die Bewertung der Qualität der zurückgelieferten Blog-Einträge durch die Testpersonen besitzen. Zum einen unterscheiden sich die Werte in den vier Tag-Spalten aufgrund des sehr akkuraten Naïve Bayes Klassifizierers teilweise nur sehr geringfügig von einander. Zum anderen ist es auch schwierig thematisch vollkommen unterschiedliche Suchbegriffe auszuprobieren, da die meisten der von *blogbar.de* zur Verfügung gestellten Blog-Einträge (zumindest die 220 als Testdatensätze verwendeten Einträge) von Sportthemen handeln. Die Suche nach Begriffen wie beispielsweise „Computer", wie von einer Testperson durchgeführt, liefert weder in dem Datensatz, in welchen die menschliche Klassifizierung stattgefunden hat noch in demselben Datensatz, jedoch mittels Maschinenlern-Klassifizierung, ein befriedigendes und erwartungsgemäßes Ergebnis. Nichts desto trotz konnten einige Unterschiede in Zusammenhang mit den beiden Klassifizierungsmethoden und den zurückgelieferten Blog-Einträgen festgestellt werden.

Kapitel 5

5 Abschließende Zusammenfassung

Dieses Schlusskapitel stellt die wichtigsten Ergebnisse und Erkenntnisse aus der vorliegenden Diplomarbeit dar. Darüber hinaus wird gegen Ende ein Ausblick auf zukünftige Weiterentwicklungen im Bereich des Suchmaschinen-Prototyps gegeben.

Aufgrund des enormen Wachstums beziehungsweise der steigenden Popularität von Weblogs in den vergangenen fünf bis acht Jahren, haben viele wissenschaftliche Arbeiten oder Projekte diverse Analyse- und Klassifizierungsmethoden für Weblogs zum Inhalt. In der vorliegenden Arbeit wurden bestehende Analysemethoden in verschiedene Kategorien unterteilt, um den Leser- und Leserinnen einen besseren Überblick über die möglichen Ansätze zu geben. Aus den jeweiligen, in dieser Arbeit verwendeten, Kategorien wie „Kommunikationssoziologie von Weblogs“, „Weblog-Communities“, „Weblogs im Kontext Sozialer Aktivitäten“, „Weblog-Inhaltsanalyen“ und „Weblogs in Zusammenhang mit Metadaten/ Tags“ wurden einige wichtige Forschungsarbeiten herausgegriffen und deren individuelle, teilweise sehr komplexen Analyseverfahren detailliert beschrieben. Teile aus diesen Methoden und Ansätzen wurden in späterer Folge als Anhaltspunkt, Designvorlage und Inspiration für die Entwicklung des Blog-Suchmaschinen Prototyps herangezogen.

Das vorrangige Ziel bei der Erstellung des Suchmaschinen-Prototyps lag darin, dem Benutzer die gezielte Suche nach bestimmten Begriffen oder Themen zu ermöglichen und eine qualifizierte Auswahl an Blogeinträgen zu dem

gesuchten Thema anhand eines definierten Bewertungsschemas zu bekommen. Die tatsächliche Realisierung beziehungsweise Programmierung erfolgte mittels der Skriptsprache PHP. Insgesamt drei .php Dateien sind an der Realisierung des gesamten Projektes beteiligt (siehe Abbildung 4.2). Die Vorzüge des konstruierten Suchmaschinen-Prototyps liegen nicht nur auf der Anwenderseite sondern auch auf der Entwicklerseite. Einer der großen Vorteile des gesamten Suchmaschinen-Prototyps liegt darin, dass er vollkommen modular und erweiterbar aufgebaut ist. Sollte es in zukünftigen Arbeiten zu einer Weiterentwicklung kommen, könnte das momentane System beinahe vollständig übernommen werden. Eine Adaption wäre beispielsweise nur in dem Teil der „search.php" Datei notwendig, in welchem der Suchalgorithmus definiert ist. Die übrigen Dateien würden jedoch in ihrer Struktur weitestgehend unberührt von dieser Änderung bleiben.

Wie in Kapitel 4.1 beschrieben, wurden 220 Blog-Einträge einerseits durch Testpersonen den beiden Tag-Kategorien (scientific und emotional) zugeordnet und Wertigkeiten zwischen eins und 10 vergeben und auf der anderen Seiten wurden dieselben nicht markierten 220 Artikel durch den Naïve Bayes Maschinenlern-Klassifizierungsalgorithmus diesen beiden Kategorien (scientific_e und emotional_e) zugeordnet. Anhand dieser „beiden Datensätze", welche sich lediglich (zumindest teilweise) durch die Wertigkeiten in diesen beiden Spalten unterscheiden, wurden dieselben Testpersonen gebeten einige Suchbegriffe einzugeben und ihre Suchabfrage zum Einen mittels des „Suche1" Buttons durchzuführen und zum Anderen mittels des „Suche2" Buttons. Die Suche über den „Suche1" Button bezieht sich auf jene 220 Blogeinträge, bei welchen die Werte in den beiden Tag Spalten durch die Testpersonen vergeben wurden und die Suche über den „Suche2" Button benutzt ebenfalls dieselben 220 Blogeinträge als Datensatz, allerdings unter Berücksichtigung der, durch den Naïve Bayes Maschinenlern-Klassifizierungsalgorithmus berechneten, Werte innerhalb der Spalten scientific_e und emotional_e.

Abschließend kann zum Thema Weiterentwicklungen gesagt werden, dass eine denkbare Ausbaustufe des Suchmaschinen-Prototyps einerseits die Integration von Communitybildungen (Permalinks beispielsweise spielen hier eine wichtige Rolle) innerhalb der blogosphärischen Gemeinschaft wäre sowie andererseits die Berücksichtigung der Aktualität der jeweiligen Blog-Einträge. In Kapitel 4.2 wurde jedoch bereits darauf hingewiesen, dass die Einbindung der individuellen Zeitstempel nach objektiven Gesichtspunkten sehr schwierig ist. Gerade bei technischen Artikel (betrachtet man die schnelllebige und rasante Entwicklung auf vielen technischen Gebieten und somit die unterschiedlichen Inhalte zur gleichen Thematik) wäre die Unterscheidung in aktuelle sowie veraltete Beiträge sinnvoll. Die Grenzziehung zwischen diesen beiden zeitlichen Kategorien liegt jedoch im Auge des Betrachters und kann nicht wirklich verallgemeinert werden.

Diverse wissenschaftliche Artikel stützen sich auf der Behauptung, dass innerhalb einer spezifischen fachlichen Community eine größere Anzahl an inhaltlich hochwertigen Blog-Einträgen existiert. Dies könnte folgendes bedeuten: Würde nun der Suchalgorithmus dahingehend erweitert werden, dass vorab eine Auswahl an passenden Communities zu dem eingegebenen Suchbegriff ausgeführt wird und anschließend die Abfrage nur noch innerhalb dieser relevanten Communities stattfindet, könnte die Qualität der zurückgelieferten Blog-Artikel noch mehr gesteigert werden.

Anhang

Auf der CD-ROM, die der vorliegenden Arbeit beigelegt ist, finden sich folgende Dokumente:

- Im Verzeichnis „Magisterarbeit“ liegt die Arbeit im PDF Format vor
- Im Verzeichnis „Poster“ befindet sich das anzufertigende Poster über die Magisterarbeit
- Das Verzeichnis „Projektdateien_Blog Suchmaschine“ beinhaltet die .php Dateien des Prototyps

Abkürzungsverzeichnis

CCT ..... Content-Community-Time Model
CEO ..... Chief Executive Officer
CMC ..... Computer Mediated Communication
CMS ..... Content Management System
CRF ..... Conditional Random Field
DF ..... Document Frequency
HTML ..... Hypertext Markup Language
HTTP ..... Hypertext Transfer Protocol
IDF ..... Inverse Document Frequency
IRST ..... Istituto per la Ricerca Scientifica
..... Tecnologica
RDF ..... Resource Description Framework
PHP ..... PHP: Hypertext Preprocessor
POS ..... Part-of-Speech
RSS ..... Really Simple Syndication (seit RSS 2.0)
SIOC ..... Semantically-Interlinked Online
..... Communities
SVM ..... Support Vector Machine
TF ..... Term/Text Frequency
WEKA ..... Waikato Environment for Knowledge
..... Analysis
WND ..... WordNet Domain Labels System
WWW ..... World Wide Web
XML ..... Extensible Markup Language

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[Aukb08] Au-Kbc: „POS Tagging“, http://www.aukbc.org/research_areas/nlp/projects/postagger.html, 5.12.2008

[Bart08] Rainer Bartel: „Blogs für alle: Das Weblog-Kompendium“, Smart Books Publishing AG, München, Deutschland, 2008

[Bata03] Vladimir Batagelj: „Analysis of large networks – Islands Presented at Dagstuhl seminar 03361”, Algorithmic Aspects of Large and Complex Networks Dagstuhl, 2003

[Baue02] M. W. Bauer: „Classical content analysis: A Review”, In M. W. Bauer & G. Gaskell (Eds.), Qualitative Researching with Text, Image, and Sound: A Practical Handbook, London: Sage Publications, 2000, pp. 131-151

[Baum08] Eric Baumer, Mark Sueyoshi, Bill Tomlinson: „Exploring the Role of the Reader in the Activity of Blogging“, Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, Florenz, Italien, 2008, pp. 1111-1120

[Bere07a] Bettina Berendt, Christoph Hanser: „Tags are not Metadata, but “Just More Content-to Some People”“, In Proceedings of the International Conference on Weblogs and Social Media, Boulder, Colorado, 2007

[Bere07b] Bettina Berendt, Andreas Hotho, Dunja Mladenič, Giovanni Semeraro: „From Web to Social Web: Discovering and Deploying User and Content
Profiles", Workshop on Web Mining, WebMine 2006, Springer-Verlag, Berlin, Deutschland, 2007

[Bere08] Bettina Berendt, Roberto Navigli: „Finding your way through blogspace: Using semantics for cross-domain blog analysis", Proceedings of the AAAI 2006 Symposium on Computational Approaches to Analysing Weblogs, Stanford, 2006, pp. 1-8

[Blek06] Alexander Blekas, John Garofalakis, Vasilios Stefanis: „Use of RSS feeds for content adaptation in mobile web browsing", ACM International Conference Proceeding Series; Vol. 134 archive Proceedings of the 2006 international cross-disciplinary workshop on Web accessibility (W4A): Building the mobile web: rediscovering accessibility?, 2006, pp. 79-85

[Bloo04] Rebecca Blood: „How blogging software reshapes the online community", Communications of the ACM, Volume 47, Issue 12, pp. 53-55

[Buch08] Stefan Buchner: „Weblogs",
http://www.stefanbucher.net/weblogfaq, 29.7.2008

[Chin06] Alvin Chin, Mark Chignell: „A Social Hypertext Model for Finding Community in Blogs", Conference on Hypertext and Hypermedia - Proceedings of the seventeenth conference on Hypertext and hypermedia, Odense, Dänemark, 2006, pp. 11-22

[Code08] Kerlins.net: „Coding Strategies“, http://kerlins.net/bobbi/research/nudist/coding/strategies.htm, 2.12.2008

[Crow97] Kevin Crowston, Marie Williams: „Reproduced and emergent genres of communication on the World-Wide Web", The Information Society, 1997, pp. 30-39

[Davi02] Todd F. Davis, Kenneth Womack: „Formalist Criticism and Reader-Response Theory", Palgrave, New York, 2002

[Dura06] K.T. Durant and M.D. Smith: „Mining Sentiment Classification from Political Web Logs", In Proceedings of Workshop on Web Mining and Web Usage Analysis of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (WebKDD-2006), 2006

[Eric96] Thomas Erickson: „The World-Wide-Web as social hypertext", Communications of the ACM – Volume 39, New York, USA, 1996, pp. 15-17

[Eric02] Thomas Erickson: „Making Sense of Computer-Mediated Communication (CMC): Conversations as Genres, CMC Systems as Genre Ecologies", In Proceedings of the 33rd Hawaii International Conference on System Sciences, 2000

[Free78] Linton C. Freeman: „Centrality in Social Networks: Conceptual Clarification, Social Networks 1: 1978/79, pp.215-239

[Glan04] Natalie S. Glance, Matthew Hurst, Takashi Tomokiyo: „BlogPulse: Automated Trend Discovery for Weblogs", WWW2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, New York, 2004 (http://www.blogpulse.com/papers/www2004glance.pdf)

[Gruh04] Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins: „Information Diffusion Through Blogspace", In Proceedings of the 13th International Conference on World Wide Web, 2004, pp 491-501

[Haye07] Conor Hayes, Paolo Avesani, Uldis Bojars: „An Analysis of Bloggers, Topics and Tags for a Blog Recommender System", From Web to Social Web: Discovering and Deploying User and Content Profiles, Springer-Verlag, Heidelberg, Berlin, 2007, pp. 1-20

[Herr04] Susan C. Herring, Lois Ann Scheidt, Sabrina Bonus, Elijah Wright: „Bridging the Gap: A Genre Analysis of Weblogs", Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences, IEEE Computer Society, Washington, USA, 2004, pp. 1-11

[Joac97] Thorsten Joachims: „A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization", In Proceedings of 14th International Conference on Machine Learning (ICML-97), 1997, pp.143-151

[Kris02] S. Krishnamurthy: „The Multidimensionality of Blog Conversations: The Virtual Enactment of September 11", In Maastricht, The Netherlands: Internet Research 3.0, 2002

[Kuma03] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins: „On the Bursty Evolution of Blogspace", In Proceedings of the 12th International Conference on World Wide Web, New York, USA, 2003, pp. 568-576

[Mars08] Marsianer.de: „Was sind Tags (Verschlagwortung)?", http://www.marsianer.de/flickr/schule/tag-verschlagwortung/, 4.10.2008

[Medi08] Medienpraxis>ch: „Diverse Kategorisierungsvorschläge für Weblogs", http://medienpraxis.ch/2005/06/26/welche-typen-von weblogs-gibt-es/, 29.7.2008

[Mill86] David W. McMillan, David M. Chavis: „Sense of Community: A Definition and Theory", Journal of Community Psychology, Vol. 14, No. 1 (1986), pp.6-23

[Nard04] Bonnie A. Nardi, Diane J. Schiano, Michelle Gumbrecht: „Blogging as Social Activity, or, Would You Let 900 Million People Read Your Diary?", Computer Supported Cooperative Work - Proceedings of the 2004 ACM conference on Computer supported cooperative work, Chicago, USA, 2004, pp. 222-231

[NiXi07] Xiaochuan Ni, Gui-Rong Xue, Xiao Ling, Yong Yu, Qiang Yang: „Exploring in the Weblog Space by Detecting Informative and Affective Articles", International World Wide Web Conference -

Proceedings of the 16th international conference on World Wide Web, New York, USA, 2007, pp. 281-290

[Paje08] Pajek: „Program for Large Network Analysis", http://vlado.fmf.uni-lj.si/pub/networks/pajek/, 28.11.2008

[Qamr06] Arun Qamra, Belle Tseng, Edward Y.Chang: „Mining Blog Stories Using Community-Based and Temporal Clustering", Conference on Information and Knowledge Management - Proceedings of the 15th ACM international conference on Information and knowledge management, Arlington, Virginia, USA, 2006, pp. 58-67

[Rain05] Lee Rainie: „The state of blogging", PEW Internet & American Life Project Data Memo, 2005 (http://www.pewinternet.org/pdfs/pip_blogging_data.pdf)

[Saee08] Naumann Saeed, Yun Yang: „Incorporating blogs, social bookmarks, and podcasts into unit teaching", Conferences in Research and Practice in Information Technology Series, Proceedings of the tenth conference on Australasian computing education - Volume 78, 2008, pp. 113-118

[Schm06] Jan Schmidt: „Weblogs: Eine kommunikationssoziologische Studie", UVK Verlagsgesellschaft mbH, Konstanz, Deutschland, 2006

[Sioc08] SIOC: „Datenmodell" http://sioc-project.org/files/1_a_sioc_executive_summary.pdf, 3.12.2008

[Stop08] RSS Pieces: „Stop Word List“, http://www.rsspieces.com/stop-word-list, 11.10.2008

[Swal90] John M. Swales: „Genre Analysis: English in Academic and Research Settings“, Cambridge University Press, 1990

[Trac05] Adam Trachtenberg: „Umsteigen auf PHP 5“, O'Reilly Verlag GmbH & Co. KG, Köln, Deutschland, 2005

[Ucin08] Ucinet: „Social Network Analysis Software“, http://www.analytictech.com/ucinet/ucinet.htm, 28.11.2008

[Weka08] Weka: „Data Mining Software in Java“, http://www.cs.waikato.ac.nz/ml/weka/, 28.11.2008

[Wiki08a] Wikipedia: „Linear-und nicht linear trennende Diskriminanzfunktionen“, http://upload.wikimedia.org/wikipedia/de/a/a0/Diskriminanzfunktion.png, 28.8.2008

[Wiki08b] Wikipedia: „Support Vector Machine“, http://de.wikipedia.org/wiki/Support_Vector_Machine, 22.11.2008

[Wiki08c] Wikipedia: „Tag Cloud“, http://de.wikipedia.org/wiki/TagCloud, 3.12.2008

[Wiki08d] Wikipedia: “Support Vector Machines”

http://de.wikipedia.org/wiki/Support_Vector_Machine,
6.12.2008

[Witt05] Ian H. Witten, Frank Eibe: „Data Mining: Practical Machine Learning Tools and Techniques“, Second Edition, Morgan Kaufmann Verlag, USA, 2005

[Wolf07] Peter Wolff: „Die Macht der Blogs“, 2. Überarbeitete und erweiterte Auflage 2007, Datakontext-Fachverlag GmbH, Frechen, Deutschland, 2007

[Word08] WordNet: „Eine lexikalische Englischdatenbank“,
http://wordnet.princeton.edu/,
5.12.2008

[Yang07a] Changhua Yang, Kevin Hsin- Yih Lin, Hsin-Hsi Chen: „Emotion Classification Using Web Blog Corpora“, Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Washington, DC, USA, 2007, pp. 275-278

[Yang07b] Changhua Yang, Kevin Hsin- Yih Lin, Hsin-Hsi Chen: „Building Emotion Lexicon from Weblog Corpora“, Proceedings of the 45th Annual Meeting of ACL, 2007, pp. 133-136

[Zerf05] Ansgar Zerfaß, Dietrich Boelter: „Die neuen Meinungsmacher. Weblogs als Herausforderung für Kampagne, PR und Medien“, Nausner & Nausner Verlag, Graz, Österreich, 2005

[Zhou06] Ying Zhou, Joseph Davis: „Community Discovery and Analysis in Blogspace", International World Wide Web Conference - Proceedings of the 15th international conference on World Wide Web, New York, USA, 2006, pp. 1017-1018

Glossar

Axiale Kodierung

Die axiale Kodierung beruht darauf, Kategorien und Sub-Kategorien, welche durch die offene Kodierung gefunden wurden, miteinander zu verknüpfen und als Hauptkategorien in Beziehung zu setzen [Code08]. Das Verhältnis einer Kategorie zu anderen Kategorien und Sub-Kategorien wird untersucht und es werden sukzessive Zusammenhangmodelle entwickelt.

Blogroll

Eine Vielzahl an heutzutage existierenden Weblogseiten beinhaltet die Blogroll. Die Blogroll enthält im Grunde genommen nichts anderes als Verweise auf andere Blogs, meist in Form von Leseempfehlungen oder Listen regelmäßig gelesener Autoren. Zahlreiche Online-Services erleichtern das Verwalten sowie das Einbinden der Blogroll in eine Blogseite. Die Blogroll ist in der Regel permanent auf der Startseite eines Weblogs sichtbar [Schm06].

CMS

Abkürzung für Content Management System: die Inhalte einer Webseite können auch ohne Programmierkenntnisse und getrennt vom Design in der Datenbank über den Webbrowser geändert beziehungsweise gewartet werden

Crawler

Unter einem Crawler, auch als Spider oder Roboter bezeichnet, versteht man ganz allgemein ein Programm, welches das World Wide Web auf eine bestimmte methodische und automatisierte Weise durchsucht. Ein Webcrawler stellt eine spezielle Form eines Crawlers dar: Es handelt sich um einen Softwareagenten, welcher als Ausgangsbasis eine Liste an URL's (diese werden auch „Saatgut" genannt) vorfindet und diese nacheinander besucht. Im

Rahmen dieses Prozesses werden die neu gefundenen Hyperlinks einer jeden Seite zur URL Gesamtliste hinzugefügt. Theoretisch könnten auf diese Art und Weise alle im World Wide Web erreichbaren Seiten gefunden werden. Faktoren wie das große Volumen des Internets, die Manipulation von Inhalten, etc. bereiten dem Crawlen jedoch große Schwierigkeiten und somit kann nicht der gesamte Inhalt des Internets erfasst werden.

HTML

Abkürzung für Hypertext Markup Language: dient zur Strukturierung von Texten, Bildern oder Hyperlinks in Dokumenten und wurde vom World Wide Web Consortium (W3C) bis zur Version 4.01 weiterentwickelt

Machine Learning (Maschinelles Lernen)

Stellt ein Teilgebiet der Informatik beziehungsweise dem Feld der „Künstlichen Intelligenz" dar und setzt sich mit Computerprogrammen auseinander, welche aus Daten lernen können. Es entsteht ein künstliches System, das aus Beispielen lernt und somit in der Lage ist Gesetzmäßigkeiten zu erkennen. Beispielhafte Algorithmen im Bereich des maschinellen Lernens sind das „Überwachte Lernen (supervised learning)", das „Unbeaufsichtigte Lernen (unsupervised learning)", das „Verstärkungs-Lernen (reinforcement learning)", und noch einige andere. Im Wesentlichen wird zwischen induktiven und deduktiven Methoden des Maschinellen Lernens unterschieden.

Offene Kodierung

Der Begriff offene Kodierung (auch als generative Kodierung bezeichnet) beschreibt den Prozess der Entwicklung von Konzeptkategorien und Themen, welche aus einem Datensatz entstehen [Code08]. Innerhalb des Datensatzes werden Themen lokalisiert und anfängliche Codes oder Etiketten werden zugewiesen, um in einem ersten Versuch die Masse an Daten in Kategorien zusammenzufassen. Der Prozess wird deshalb als offen bezeichnet, da die

Erforschung der Daten ohne frühere Annahmen darüber, was entdeckt werden könnte, erfolgt.

Permalink

Der Permalink wird auch als Permanentlink bezeichnet und spiegelt das „Problem" wider, dass Beiträge gemeinsam auf der Startseite eines Blogs erscheinen, jedoch durch eigenständige URLs erreicht und angesprochen werden [Schm06]. Permalinks werden häufig eingesetzt, um Verweise auf andere Einträge in Weblogs zu verlinken.

Precision (Genauigkeit)

Die Precision bezieht sich auf die Genauigkeit eines Suchergebnisses. In Bezug auf die Beurteilung eines Klassifikators spricht man auch von einem „Positiven Vorhersagewert". Die Precision wird weiters auch als Wahrscheinlichkeit, mit welcher ein gefundenes Dokument bedeutsam ist, definiert. Die Precision kann Werte zwischen null Prozent und 100 Prozent beziehungsweise zwischen 1 und 0 annehmen.

Recall (Abruf)

Der Recall (Richtigpositiv-Rate oder Sensitivität) bezeichnet die Vollständigkeit eines Suchergebnisses. Die Basis dafür bildet der Anteil an gefundenen relevanten Dokumenten innerhalb einer Suche. Der Recall kann, analog zur Precision, auch als Wahrscheinlichkeit, mit der ein relevantes Dokument gefunden wird, ausgedrückt werden. In Analogie zur Precision kann auch der Recall Werte zwischen null Prozent und 100 Prozent annehmen.

RSS Feeds

RSS entspringt der Familie des XML Dateiformats, welches dem Zweck der Zusammenfassung von Webseiten-Inhalten dient. Das RSS Format ist darüber hinaus auch plattform-unabhänging. Ein RSS Dokument (wird auch als „Feed"

bezeichnet) beinhaltet sämtliche zusammengefasste Texte als auch Metadaten wie beispielsweise Publizierungsdatum oder Urheberschaft. Mittels eines sogenannten RSS-Readers können RSS-Dateien gelesen werden und somit sämtliche Änderungen einer Webseite verfolgt werden.

Social Bookmarks

Werden im deutschen Sprachgebrauch als so genannte „Soziale Lesezeichen" beziehungsweise „Internet-Lesezeichen" bezeichnet. Im Wesentlichen geht es darum, Lesezeichen auf einer öffentlichen Webseite zu speichern und diese zu markieren (Verknüpfungen mit einem oder mehreren Schlüsselwörtern, welche diese beschreiben, abzuspeichern) [Saee08]. Social Bookmarks erlauben im Großen und Ganzen einen einfachen und schnellen Zutritt zu Online-Ressourcen. Zu einigen der bekanntesten Anbieter im englischsprachigen Raum zählen unter anderem „del.icio.us", „Digg", etc.

Tag Cloud

Eine Schlagwort-Wolke (engl. tag cloud) stellt eine Methode zur gezielten Indizierung von Webseiten wie Blogs oder Wikis dar. Weiters dient diese Methode zur Informationsvisualisierung, bei der eine Liste aus Tags alphabetisch sortiert flächig angezeigt wird [Wiki08c]. Die Darstellung erfolgt auf eine Art Wolke. Häufig vorkommende Wörter werden beispielsweise größer dargestellt.

Trackback

Die Trackback-Funktion löst grundsätzlich das Problem, dass ein Hyperlink prinzipiell nur einseitig ausgerichtet ist und ein Link von der einen Seite auf die andere Seite nicht wieder zur Ausgangsseite zurückführt [Schm06]. Weiters ist es Bloggern möglich festzustellen, ob jemand in einem anderen Weblog auf ihren eigenen Eintrag Bezug nimmt. Unter Verwendung von Trackback ist es

möglich, dass sich verschiedene Seiten (im Speziellen Autoren von Weblogs) über zusammengehörige Ressourcen austauschen.

WEKA

WEKA bietet eine Sammlung an Maschinenlern-Algorithmen, um Data Mining Aufgaben zu bewerkstelligen [Weka08]. Diese, in der Programmiersprache Java geschriebene, Software beinhaltet Werkzeuge für die Vorverarbeitung von Daten, Klassifizierung von Daten, Clustering von Daten, Visualisierung von Klassifizierungsergebnissen, etc. WEKA stellt eine freie Software dar, welche unter der General Public License (GNU) verfügbar ist.

Wiki

Der Begriff stammt ursprünglich aus dem Hawaiischen und kann als „schnell" übersetzt werden. Mittels eines Wikis, auch als WikiWeb bezeichnet, kennzeichnet eine Ansammlung an Webseiten, welche es jedermann erlaubt Inhalte zu modifizieren beziehungsweise bei der Inhaltsgestaltung mitzuwirken. Diese Modifikationen werden unter Verwendung von so genannten Markup Languages durchgeführt.

Wordpress

Unter Wordpress ist eine Weblog-Publishing Software zu verstehen, welche dem User vordefinierte jedoch individuell anpassungsfähige Designvorlagen zur Verfügung stellt.