

1 Wer bloggt was? Eine Analyse der deutschen Top 100-Blogs mit Hilfe von Cluster-Verfahren

Sebastian Schäfer, Alexander Richter und Michael Koch
Forschungsgruppe Kooperationsysteme an der Universität der
Bundeswehr München

1	Motivation	5
1.1	Analyse des Content von Social Software	5
1.2	Analyse der Blogosphäre	6
2	Cluster-Analyse für Blog-Inhalte	7
2.1	Hierarchisches Clustern	8
2.2	<i>k</i> -means Clusterverfahren	9
2.3	Clustern durch Dimensionsreduktion	10
2.4	Anwendung der Cluster-Analyse auf die deutsche Blogosphäre .	10
3	Ergebnisse und Interpretation	12
3.1	Datenqualität	12
3.2	Hierarchisches Clustern	12
3.3	Wortcluster	14
3.4	<i>k</i> -means Clusterverfahren	15
3.5	Clustern durch Dimensionsreduktion	18
4	Zusammenfassung und Ausblick	18

1 Motivation

Die Entwicklung vom Web 1.0 zum Web 2.0 bedeutet nicht nur, dass durch die Dezentralisierung und Beteiligung Vieler mehr Daten zu Interessen und Vernetzungen von Web-Nutzern bereitgestellt werden, sondern auch, dass diese nun leichter zugänglich sind – über APIs oder andere Standard-Schnittstellen wie RSS-Feeds.

Dadurch haben nun große Web-Plattformen kein Monopol mehr auf die Datenmengen, die zur Nutzung der „kollektiven Intelligenz“ (Surowiecki, 2005), zur Analyse von Trends etc. notwendig sind. Jeder Internet-Nutzer kann mit relativ geringem Aufwand auf die Information zugreifen, Analysen darüber fahren und sie zur Verbesserung eigener Dienste einsetzen (siehe z. B. Segaran, 2007).

In diesem Beitrag wollen wir dies am Beispiel der Analyse eines Ausschnitts der deutschen Blogosphäre demonstrieren. Dabei wählen wir das Medium Blog, da Blogs ein klassisches Beispiel für die Dezentralität des Web 2.0 sind und über RSS-Feeds einen guten Zugang zu den Daten bieten, die dezentral erfasst werden.

1.1 Analyse des Content von Social Software

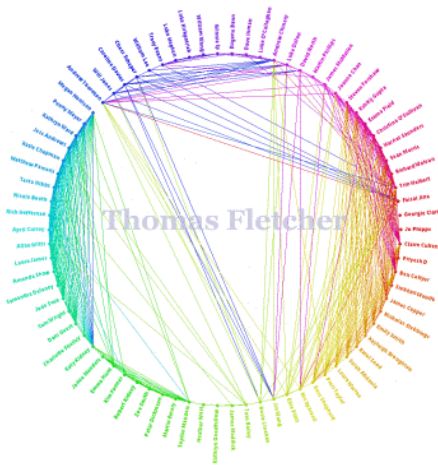
Social Software bietet aufgrund mehrerer Eigenschaften — z. B. der Personenzentriertheit der Dienste oder die Beteiligung Vieler an den Inhalten - eine sehr gute Datenbasis für die Gewinnung von Erkenntnissen im Sinne der „kollektiven Intelligenz“. Betrachtet man das ganze Spektrum von Social Software, bieten sich hierfür mehrere Möglichkeiten an, von denen einige kurz angesprochen werden sollen.

Nachdem im Zentrum von Social Software fast durchgehend Soziale Netzwerke stehen, spielen hier die Methoden der Sozialen Netzwerkanalyse (SNA) eine entscheidende Rolle (vgl. z. B. Wasserman und Faust, 1997; Scott, 1991). Dabei geht es grundsätzlich um die Analyse von (sozialen) Interaktionen zwischen Nutzern.

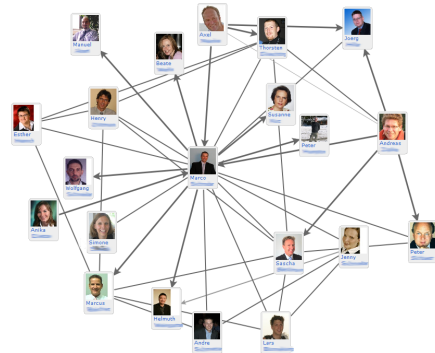
Ein Beispiel aus dem privaten Bereich sind die zahllosen Anwendungen auf der Plattform „Facebook“, die es ermöglichen die Verbindungen innerhalb des persönlichen Netzwerks anzeigen zu lassen (wie z. B. das „FriendWheel“ vgl. Abbildung 1.1, links). In Unternehmen kann die SNA eingesetzt werden um die Zusammenhänge und Potentiale bzgl. der Zusammenarbeit frühzeitig erkennen zu können. So ist es z. B. bei dem in der IBM zum Social-Networking eingesetzten Dienst „Fringe“ mit Hilfe der Erweiterung „Sonar“ sehr einfach möglich die Beziehungen zwischen Mitarbeitern zu analysieren und zu visualisieren (vgl. Abbildung 1.1, rechts).

Dasselbe gilt für Wikis: Müller nimmt in einem anderen Beitrag in diesem Band eine anschauliche Analyse virtueller Gemeinschaften in Corporate Wikis zur Förderung des selbstorganisierten Wissensmanagements vor. Ähnlich wie bei der Analyse von Social Networking Services ist das Ziel, gewünschte und ungewollte Entwicklungen frühzeitig zu erkennen und ggf. reagieren zu können. Blaschke nennt in seinem Beitrag mehrere weitere Veröffentlichungen zur Analyse und Visualisierung von Ko-Autoren-Netzwerken in Wikis.

Wer bloggt was?



(a) Facebook-Anwendung „FriendWheel“



(b) Ausschnitt aus „Fringe“ (IBM)

Abbildung 1.1: Visualisierung sozialer Netzwerke

Und auch – oder gerade – Weblogs/Blogs stellen aufgrund des kleinen Mikrouniversums, das jeder von ihnen für sich darstellt, ein ausgezeichnetes Ziel für die SNA dar. Hier setzt dieser Beitrag an.

1.2 Analyse der Blogosphäre

Wie auch bei anderen Dienstklassen stützen sich viele Anwendungen von SNA in der Blogosphäre auf direkte Verlinkungen zwischen Blogs (und damit meist Personen). Neben Blogrolls werden dabei vor allem Kommentierungen und Trackbacks zwischen Blogs ausgewertet und daraus verschiedene Arten von Verweisgraphen erstellt, wie wir sie im vorherigen Abschnitt schon gezeigt haben.

Hierzu liegen zahlreiche Studien vor. Ein interessanter Überblick zu Veröffentlichungen in denen z. B. strukturelle Merkmale weblogbasierter Netzwerke auf Grundlage von Stichproben untersucht werden (wie die Analyse von länder- oder sprachspezifischen Blogosphären) oder die sich mit thematisch zusammenhängenden Weblognetzwerken befassen findet sich bei Schmidt (2008).

Wir wollen in diesem Beitrag einen anderen Ansatz aufzeigen. Im Weiteren wird am Beispiel eines Ausschnitts der deutschen Blogosphäre gezeigt, wie sich mit Hilfe einfacher Cluster-Verfahren verschiedene Datenquellen im Internet an Hand ihrer Inhalte thematisch klassifizieren lassen. Ziel ist es zum einen, die einfache Anwendbarkeit und Leistungsfähigkeit der verwendeten Verfahren zu demonstrieren, und zum anderen Erkenntnisse über die Themenbereiche der deutschen Blogger-Landschaft zu gewinnen.

Dazu wurden die RSS-Feeds ausgewählter deutscher Top-Blogs im Hinblick auf aktuell diskutierte Themen und insbesondere der dabei enthaltenen Wörter analysiert. Zur Klassifizierung wurden einfache Standardverfahren verwendet, die im Folgenden kurz beschrieben werden. Weitere Ausführungen dazu finden sich beispielsweise bei Backhaus et al. (2000) oder Moosbrugger und Frank (1992). Segaran (2007) wendet ähnliche Verfahren auf die englische Blogosphäre an.

Im Folgenden werden zunächst die drei in der Studie angewandten Methoden der Cluster-Analyse und das konkrete Vorgehen (Abschnitt 2) vorgestellt. Anschließend werden die Ergebnisse der Anwendung der Methoden auf die einen ausgewählten Teil der „deutschen Blogcharts“ erläutert und interpretiert (Abschnitt 3). Eine Zusammenfassung sowie ein Ausblick auf weitere mögliche Analyse-Schritte schließen diesen Beitrag ab (Abschnitt 4).

2 Cluster-Analyse für Blog-Inhalte

Unter einer Cluster-Analyse versteht man im Allgemeinen strukturentdeckende, multivariate Analyseverfahren zur Ermittlung von Gruppen (Clustern) von Objekten, deren Eigenschaften oder ihre Ausprägungen bestimmte Ähnlichkeiten aufweisen (vgl. Wikipedia:Clusteranalyse). Es wird also nicht auf Strukturen von (Interaktions-)Graphen gearbeitet, sondern auf konkreten Inhalten, die zwar Blogs und somit Personen zugeordnet werden können, zwischen denen aber noch keine Beziehungen erkannt worden sind. Ergebnis der Cluster-Analyse sind dann Beziehungen zwischen den Inhalten (und damit den Blogs bzw. Personen) in der Form der Zugehörigkeit zu gemeinsamen Clustern bzw. der „Nähe“ zueinander.

Clusterverfahren können damit einen empirischen quantitativen Zugang zum Benutzergenerierten Inhalt des Web 2.0 eröffnen, wobei die Datenerhebung – z. B. im Gegensatz zu Umfragen (quantitativ) und Interviews (qualitativ) – eher indirekt in Form von unabhängigen Beobachtungen erfolgt.

Heute existiert eine Vielzahl von Cluster-Analyseverfahren, die beispielsweise zur automatischen Klassifikation oder zur Mustererkennung eingesetzt werden.

In der vorliegenden Studie wurden aufgrund der Einfachheit und der Anschaulichkeit der Ergebnisse die drei folgenden Cluster-Methoden angewendet:

- Hierarchisches Clustern
- k -means Clusterverfahren
- Clustern durch Dimensionsreduktion

Zur Anwendung der Cluster-Methoden liegen die zu clusternden Elemente in Form von mehrdimensionalen Vektoren vor, welche die Merkmalsausprägungen (hier die Vorkommenshäufigkeit eines bestimmten Wortes) beschreiben. Um die Vektoren dann vergleichen zu können, wird ein Ähnlichkeitsmaß bzw. eine Distanzfunktion benötigt.

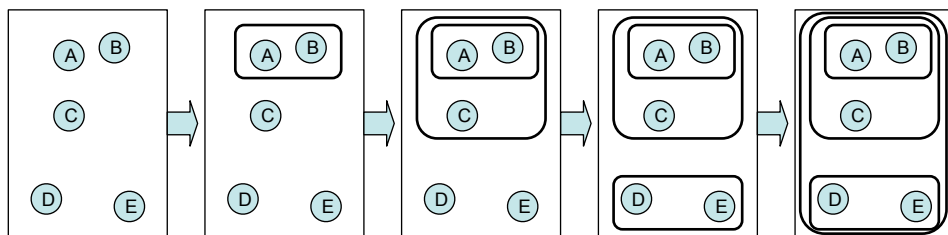


Abbildung 1.2: Hierarchische Clusterung von fünf Elementen in der Ebene

Neben der klassischen euklidischen Distanz bietet sich hierfür (u. a.) der Korrelationskoeffizient nach Pearson an, der auch für die Berechnungen in diesem Beitrag verwendet wird. Dieses Maß liefert insbesondere bei nicht normalisierten Daten die besseren Ergebnisse. Dadurch wird die möglicherweise unterschiedliche Länge von Blogbeiträgen ausgeglichen, da je nach Textlänge charakteristische Wörter in entsprechender Häufigkeit zu erwarten sind.

2.1 Hierarchisches Clustern

Beim hierarchischen Clusterverfahren wird schrittweise eine Hierarchie von Gruppen aufgebaut. Zu Beginn besteht jede Gruppe aus den Einzelementen, in diesem Fall den einzelnen Blogvektoren. Der Algorithmus fasst nun schrittweise immer die zwei Gruppen zusammen, die entsprechend des Ähnlichkeitsmaßes die kleinste Distanz zueinander aufweisen. Der Vektor der neu entstandenen Gruppe berechnet sich dann aus dem arithmetischen Mittel der beiden Gruppenvektoren. Das Verfahren wird so lange fortgesetzt, bis nur noch eine Gruppe existiert. Abbildung 1.2 verdeutlicht den Ablauf an Hand von fünf Elementen im zweidimensionalen Raum.

Als Ergebnis erhält man nicht nur eine hierarchische Clusterstruktur der Daten, sondern über die Distanzwerte auch Informationen wie weit die einzelnen Unter-Cluster eines Knotens bzw. die einzelnen Elemente auseinander liegen. Dies lässt sich am besten an Hand eines Dendrogramms visualisieren, bei dem die Länge der Kanten die Distanzen repräsentieren. Für das Beispiel aus Abbildung 2 ergibt sich ein Dendrogramm wie in Abbildung 1.3.

Als zusätzliche Möglichkeit lässt sich durch eine einfache Rotation der Matrix um 90 Grad ein Clustern der Wörter erreichen, d. h. es lassen sich Gruppen von thematisch ähnlichen Wörtern bilden. Allerdings ist in diesem Fall zu beachten, dass die Anzahl der zu clusternden Elemente deutlich größer ausfällt als die zur Verfügung stehenden Dimensionen und den dadurch aufgespannten Raum. Das Risiko eher fragwürdiger Cluster steigt damit deutlich an.

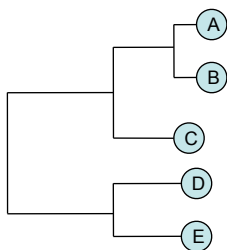


Abbildung 1.3: Darstellung des Beispiels aus Abbildung 1.2 als Dendrogramm

2.2 k -means Clusterverfahren

Hierarchische Clusterverfahren lassen sich durch Dendrogramme gut visualisieren, doch liegen die Interpretation und vor allem die Aufteilung in signifikant unterschiedliche Gruppen eher im Auge des Betrachters. Zudem erfordert die Berechnung eine Vielzahl Distanzberechnungen und Vergleichen, was insbesondere bei sehr großen Datensätzen zu langen Laufzeiten führt.

Eine Alternative bietet hierzu das k -means Clusterverfahren, welches die Daten in eine vorher festgelegte Anzahl von k Gruppen klassifiziert. Die Größe der einzelnen Gruppen richtet sich nach der Struktur der vorhandenen Daten. Kern des Verfahrens sind k Zentroide, d. h. Punkte im Raum, die das Zentrum eines Clusters repräsentieren. Die Zentroide werden zu Beginn zufällig in dem durch die Vektoren aufgespannten Raum verteilt. Der Algorithmus prüft nun für jedes Element (Blogvektor), welchem Zentroid es am nächsten liegt und weist es diesem Zentroid zu. Anschließend werden die Zentroide in den Mittelpunkt ihrer zugewiesenen Elemente verschoben. Diese Schleife wiederholt sich so lange, bis keine Veränderung mehr auftritt. Abbildung 1.4 verdeutlicht diesen Ablauf mit fünf Elementen und zwei Zentroiden im zweidimensionalen Raum.

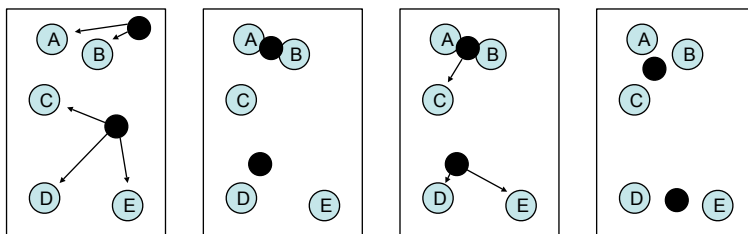


Abbildung 1.4: k -means Clustering von 5 Elementen mit 2 Zentroiden

Durch die zufällige Verteilung zu Beginn besitzt das Verfahren im Gegensatz zur hierarchischen Clusterung eine stochastische Komponente. Die Ergebnisse, d. h. die Inhalte der einzelnen Cluster, können sich daher von Lauf zu Lauf leicht unterscheiden. Als Gütekriterium dient hierbei die mittlere paarweise Distanz zwischen den Elementen innerhalb eines Clusters. Diese sollte deutlich unter der mittleren paarweisen Distanz über alle Elemente liegen. Auch kann es passieren, dass bei hinreichend großem k einzelne Cluster leer bleiben oder nur ein Element enthalten. Hier muss der Durchlauf gegebenenfalls neu gestartet werden.

2.3 Clustern durch Dimensionsreduktion

Als letztes Verfahren wurde eine so genannte Dimensionsreduktion auf die Daten angewendet. Gewöhnlich entzieht sich jeder Raum mit mehr als drei Dimensionen der menschlichen Vorstellungskraft. Es wurden daher eine Reihe von Verfahren entwickelt, die hochdimensionale Daten unter möglichst geringem Informationsverlust in drei, zwei oder gar einer Dimension darstellen und damit grundsätzliche Beziehungen zwischen verschiedenen Elementen veranschaulichen.

Die hier verwendete Technik ist unter dem Namen „multidimensionale Skalierung“ bekannt und reduziert die N -dimensionalen Blogvektoren auf eine Darstellung in der Ebene. Die Kernidee dabei ist für alle Elemente die paarweisen Abstände in der Ebene (gemessen durch die euklidische Distanz) mit den entsprechenden Werten der (multidimensionalen) Distanzfunktion überein zu bringen. Dazu werden zunächst für alle Elemente paarweise die Abstände über die Distanzfunktion (hier Pearson-Korrelation) als „Zieldistanz“ berechnet. Anschließend werden alle Elemente (Blogs) zufällig auf eine Ebene platziert.

In der Hauptschleife wird nun über mehrere Runden die „reale“ Distanz in der Ebene mit der angestrebten Zieldistanz verglichen und durch kontinuierliches Verschieben der Punkte versucht, den globalen Fehler, d. h. die Summe der Differenzen zwischen realer und gewünschter Zieldistanz, zu minimieren. Anschaulich lässt sich dies so vorstellen, dass in jeder Runde für jedes Element ein zweidimensionaler Kraftvektor berechnet wird, der sich aus seiner Fehlstellung zu seinen Nachbarn ergibt. Am Ende einer Runde werden alle Elemente ein kleines Stück in die Richtung ihres Kraftvektors verschoben und es werden erneut alle Fehlstellungen berechnet. Das Verfahren bricht ab, so bald keine Verbesserung mehr gefunden werden kann, d. h. der globale Fehler sich nicht weiter minimieren lässt. Der globale Fehler dient dabei auch als eine Art Gütemaß der gefundenen Lösung und wird durch die zufällige Platzierung zu Beginn bei jedem Durchlauf leicht variieren.

2.4 Anwendung der Cluster-Analyse auf die deutsche Blogosphäre

Die Datenbasis für die Analyse bildet eine Auswahl von deutschen Top-Blogs, die im Wesentlichen der Website Deutsche Blogcharts (www.deutscheblogcharts.de) entnommen wurde und auf der jede Woche die einhundert wichtigsten deutschen Blogs

ermitteln werden. Als Kriterium wird die Zahl der „Verlinkungen (Erwähnungen)“ eines Blogs innerhalb der Blogosphäre“ herangezogen, die sicherlich eine wichtige Kennzahl zur Bewertung von Relevanz und Bekanntheit eines Blogs darstellt. Eine alternative Bewertung bietet beispielsweise die Website www.metaroll.de, welche Nennungen in der Blogroll auf anderen Blogs auswertet. Zusätzlich wurden noch drei „hauseigene“ Blogs von Autoren der Forschungsgruppe Kooperationssysteme aufgenommen. Da aufgrund technischer Gründe nicht alle Blogs ausgelesen werden konnten, standen schlussendlich 76 Blogs zur Analyse zur Verfügung. (Eine Liste der URLs findet sich auf der Verlagsseite zum Buch unter www.vieweg.de).

In einem ersten Schritt wurden die RSS-Feeds der ausgewählten URLs eingelesen, welche je nach Blog die kompletten Beiträge oder aber nur einen Abstract umfassen. Nach Entfernung aller HTML-Tags wurden die Worthäufigkeiten der aktuellen Inhalte bestimmt. Um hierbei nur möglichst charakteristische Wörter zu betrachten, wird sich bewusst auf Substantive beschränkt, die im Gegensatz zum Englischen im Deutschen einfach an Hand von Großbuchstaben identifiziert werden können. Da dieses Kriterium allerdings auch zum Teil für andere Worte z. B. am Satzanfang zutrifft, wurde nach einem ersten Durchgang nach manueller Durchsicht eine Negativliste erstellt, mit der häufig verwendete „Satzanfangswörter“ herausgefiltert werden. Fortgeschrittenere Verfahren, die hier allerdings aus technischen Gründen (noch) nicht verwendet wurden, sind beispielsweise der Einsatz eines so genannten Stemming Verfahrens (vgl. Wikipedia:Stemming), um Worte, wie z. B. Blogs und Blog, auf ihre Grundform zurückzuführen und damit identisch einordnen zu können.

Als weiteres Kriterium wurden zwei Schranken festgelegt, mit denen die minimale als auch die maximale Worthäufigkeit über alle Blogs eingestellt werden kann. Beispielsweise macht es wenig Sinn, spezielle Worte, die nur in einem einzigen Blog vorkommen zu betrachten, andererseits bieten sehr häufig gebrauchte Worte kein Unterscheidungsmerkmal. In den Ergebnissen dieses Beitrags wurde nach einigem Ausprobieren die untere Grenze auf 5 %, die obere auf 60 % gesetzt, d. h. nur Substantive, die mindestens in 5 % und höchstens in 60% aller Blogs mindestens zweimal vorkommen, wurden berücksichtigt. Die Menge der charakteristischen Worte bietet abschließend auch eine Schranke für die tatsächlich geclusterten Blogs, indem nur Blogs klassifiziert werden, die mindestens 10 % der charakteristischen Worte in ihrem aktuellen Feed aufweisen.

Insgesamt konnten so 379 charakteristische Substantive identifiziert werden, so dass jedes Blog durch einen 379-dimensionalen Vektor repräsentiert wurde, der das Vorkommen eines jeden Wortes in diesem Blog enthält. Die dadurch aufgespannte 76×379 Matrix bildet die Datengrundlage für alle im Folgenden beschriebenen Verfahren.

Das ganze Verfahren wurde ohne spezielle Software innerhalb von zwei Wochen direkt in Python unter Rückgriff auf frei verfügbare Bibliotheken programmiert. Als Vorbild diente hierbei Segaran (2007). Damit zeigte sich schon sehr schön die Mächtigkeit und Verwendbarkeit heutiger Programmierumgebungen für solche Aufgaben.

Im ersten Schritt werden die Blogfeeds geparkt und die Worthäufigkeitsmatrix in ein Flatfile geschrieben. Alle Clusterverfahren greifen darauf als einheitliche Datenbasis

Wer bloggt was?

zu und speichern ihre Ergebnisse direkt als Bild (Dendrogramm, Bloglandkarte) oder Textfile (k -means Clustering).

3 Ergebnisse und Interpretation

In diesem Abschnitt werden die Ergebnisse des hierarchischen Clusters, der Abwendung des k -means Clusterverfahren und des Clusters durch Dimensionsreduktion der Reihe nach vorgestellt und interpretiert. Zuvor wird noch ein Blick auf die Qualität der erhobenen Daten geworfen.

3.1 Datenqualität

Als erster Qualitätscheck wurde die Worthäufigkeitsmatrix untersucht. Hierbei zeigt sich, dass jedes Blog im Mittel 92, d. h. ca. 25%, der charakteristischen Substantive mindestens einmal enthält. Diese Besetzungsdichte ist jedoch nicht gleichmäßig verteilt, sondern schwankt mit einer Standardabweichung von 49,2 und damit einem Variationskoeffizienten von ca. 0,53. Dies ist vermutlich auf die stark unterschiedliche Länge und der Themenbreite einzelner Feed zurückzuführen und sollte insbesondere bei einer detaillierten Analyse zwischen einzelnen Blogs berücksichtigt werden. Für eine allgemeine (Grob-)clustering wie in diesem Beitrag sollten sich dadurch keine großen Verzerrungen ergeben.

Als nächstes wurden die Blogvektoren paarweise an Hand der Distanzfunktion verglichen. Wünschenswert ist vor allem eine nicht zu kleine Standardabweichung bezüglich der Distanzwerte eines Vektors zu den anderen. Diese Schwankung dient als Zeichen dafür, dass die Inhalte unterschiedliche Ähnlichkeiten zu den übrigen aufweisen und somit geeignet geclustert werden können. Im vorliegenden Fall ergibt sich eine mittlere Distanz zwischen den Vektoren von 0,929 (d. h. eine Korrelation von 0,071 nach Pearson) und eine mittlere Standardabweichung von 0,087. Letztere kann als „ausreichend“ bezeichnet werden, d. h. die Verfahren sind grundsätzlich anwendbar, allerdings ist auch mit einigen Schwankungen bei zufallsbehafteten Verfahren zu rechnen.

3.2 Hierarchisches Clustern

Abbildung 1.5 zeigt einen Ausschnitt aus dem Dendrogramm, welches aus dem Ergebnis des hierarchischen Verfahrens erstellt wurde. Dabei handelt es sich offensichtlich um Blogs, die sich in ihren Beiträgen fast ausschließlich oder überwiegend mit den Themen Web 2.0 und Internet auseinandersetzen. Der thematische Zusammenhang dieses Cluster wird bereits anhand der Namen der angegebenen Blogs deutlich, die meistens direkten Bezug zum Thema „Web 2.0“ aufweisen. Der „S-O-S SEO Blog“ (SEO=Search Engine Optimization) und der „Google Watchblog“ (also ein Blog, der sich mit eben Google auseinandersetzt) stehen sich thematisch sehr nahe, während

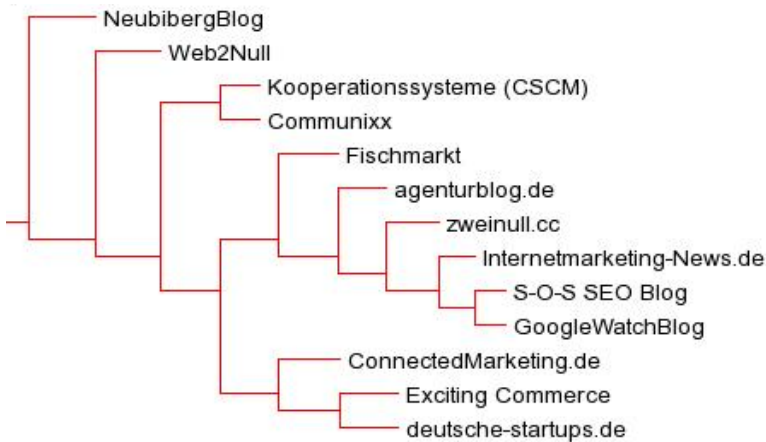


Abbildung 1.5: Web 2.0 und Internet Cluster

der „NeubibergBlog“ und „deutsche-startups“ den größten thematischen Abstand innerhalb des Clusters aufweisen.

Einen Hinweis auf die gute Aussagekraft des Verfahrens ist zudem die gemeinsame Gruppierung der „hauseigenen“ Blogs „Kooperationssysteme (CSCM)“ und „Communix“, deren Beiträge zu einem großen Teil vom selben Autor verfasst wurden.

Abbildung 1.6 zeigt demgegenüber zwei kleinere Cluster. Auf der linken Seite ist ein Cluster zu sehen, dessen zugehörige Blogs sich fast ausschließlich mit Software und Webdesign befassen. Nicht dazugehörig scheint der „law blog“. Eine mögliche Erklärung wäre, dass eine Vielzahl der Berichte im „law blog“ sich mit rechtlichen Fragen der Website-Gestaltung befassen. Auf der rechten Seite befinden sich mehrere Blogs, die sich überwiegend mit Themen rund um Apple Produkte, wie beispielsweise dem iPhone und iPod befassen.



Abbildung 1.6: Software-/Webdesign- und Apple-Blogcluster

Wer bloggt was?

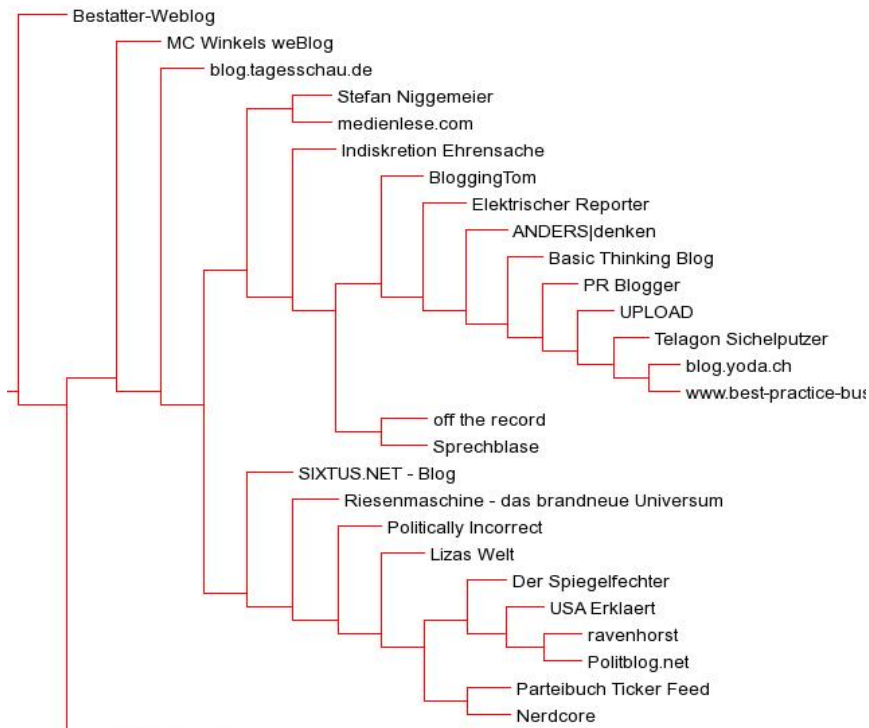


Abbildung 1.7: Der Cluster der Blogs mit allgemeinen Themen

Einen weiteren wesentlich umfangreicheren Cluster stellt eine Gruppe von Blogs dar, die über allgemeine Themen berichten (vgl. Abbildung 1.7). In diesem sind auch Blogs vertreten, die zwar – nach Erfahrung der Autoren - einen hohen Anteil an Beiträgen zum Web 2.0 haben („PR Blogger“, „Basic Thinking Blog“) aber offensichtlich in ihren Beiträgen auch über weitläufigere/allgemeine Themen berichten. Interessant ist auch der Teil-Cluster „Politik“ im unteren Bereich, dessen thematischer Zusammenhang wiederum an der Namensgebung der Blogs sichtbar wird.

Auf eine Darstellung des gesamten Dendrogramms über alle Blogs wurde hier aus Platzgründen verzichtet. Alle Grafiken des Beitrages stehen auf der Verlagsseite des Buchs unter www.vieweg.de zum Download bereit.

3.3 Wortcluster

Wie oben beschrieben kann durch eine einfache Drehung der Matrix um 90 Grad mit dem hierarchischen Clusterverfahren auch eine Klassifizierung nach Worten erreicht werden. Abbildung 1.8 zeigt einige Ausschnitte von zusammenhängend geclusterten

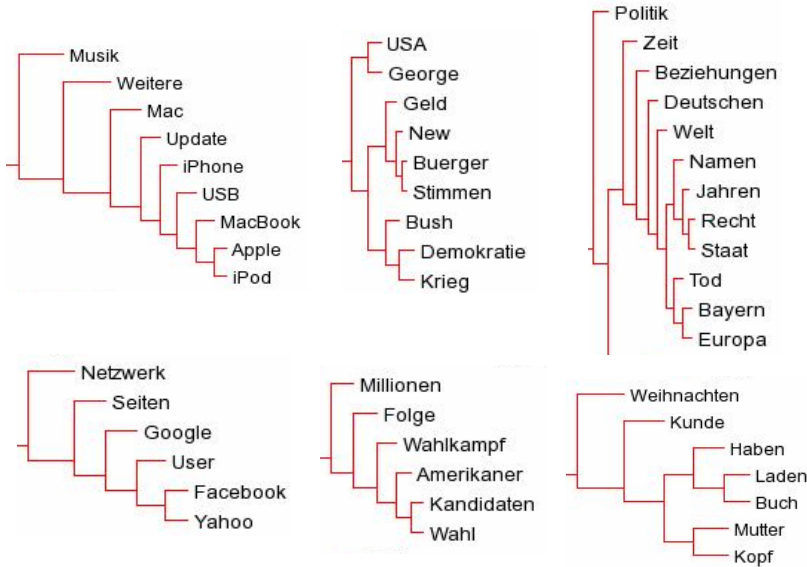


Abbildung 1.8: Einige Beispiele der generierten Wortcluster

Worten, die sich unseres Erachtens nach erstaunlicherweise gut mit dem allgemeinen Begriffsverständnis decken. Die Wortcluster geben schnell einen Überblick über aktuelle Themen und Zusammenhänge und lassen bei entsprechendem Hintergrundwissen bereits den kommentierten Sachverhalt errahnen. Neben den durchaus als „sinnvoll“ einzustufenden Clustern gibt es allerdings – bedingt durch den starken Dimensionsunterschied (s. o.) – auch einige eher sinnfreie Cluster, wie beispielsweise in Abbildung 1.8 ganz rechts unten.

3.4 k -means Clusterverfahren

Das k -means Clusterverfahren wurde für verschiedene Größen zwischen 3 und 10 Cluster durchgeführt. Im Folgenden werden die Ergebnisse für 6 und für 9 Cluster vorgestellt (vgl. Tabellen 1.1 und 1.2), da diese Größe in den Durchläufen im Allgemeinen eine gute Qualität aufweisen. Als Qualitätsmerkmal wurde, wie in den Tabellen angegeben, sowohl die mittlere Distanz der Elemente in einem Cluster als auch die durchschnittliche Clustergröße verwendet. Wie erwartet, liegen die Werte überall deutlich unter dem oben angegebenen Gesamtdurchschnittswert von 0,929 – mit Ausnahme des dritten Clusters in Tabelle 1.1, hier sind – wohl eher zufällig bedingt – thematisch sehr unterschiedliche Blogs gruppiert worden.

Insgesamt zeigen sich erwartungsgemäß ähnliche Ergebnisse wie auch der hierarchischen Clusteranalyse. Cluster 1 gruppiert sowohl die „Apple-affinen“ Inhalte als auch

Tabelle 1.1: Ergebnis einer k -means Clusterung mit $k = 6$

Cluster	Blog
Cluster 1 (10 items, Dist. 0,841)	Der Spiegelfechter, Irgendwas ist ja immer, Lupe, der Satire-Blog, Mac Essentials Newsfeed, Nerdcore, Parteibuch Ticker Feed, Politically Incorrect, fsklog, praegnanz.de, stereopoly NewsFeed
Cluster 2 (5 items, Dist. 0,712)	Communixx, Kooperationssysteme (CSCM), PR Blogger, Sprechblase, Web2Null
Cluster 3 (5 items, Dist. 0,940)	ConnectedMarketing.de, Lustige Videos, NachDenkSeiten – Die kritische Website, Politblog.net, law blog
Cluster 4 (24 items, Dist. 0,892)	1x umruehren bitte, Bestatter-Weblog, Das Kopfschuettel-Blog, Der Shopblogger, Exciting Commerce, F!XMBR, F5, Gluehweinjunkies, Indiskretion Ehrensache, Lizas Welt, MC Winkels weBlog, Medienrauschen, das Medienweblog, Riesenmaschine – das brandneue Universum, SIXTUS.NET – Blog, Spreeblick, Stefan Niggemeier, USA Erklaert, Webkrauts, annalist, blog.tagesschau.de, deutsche-startups.de, fudder – neuigkeiten aus freiburg, medienlese.com, ravenhorst
Cluster 5 (24 items, Dist. 0,879)	ANDERS denken, Basic Thinking Blog, Blogs! Buch Blog, Blogschrott.net – Web 2.0 – Yannick Eckl, Buchhaendleralltag und Kundenwahnsinn, Der Schockwellenreiter (RSS-Feed), Dr. Web Weblog, Elektrischer Reporter, Fischmarkt, GoogleWatchBlog, Internetmarketing-News.de, Karriere-Bibel, NeubibergBlog, S-O-S SEO Blog, Telagon Sichelputzer, UPLOAD, Wortfeld, agenturblog.de, imgriff.com, netzpolitik.org, off the record, pixelgraphix, ricdes dot com, zweinull.cc
Cluster 6 (8 items, Dist. 0,722)	BloggingTom, Blogwiese, Frank Helmschrott, Peruns Weblog, Software Guide, blog.yoda.ch, bueltge.de [by:ltge.de], www.best-practice-business.de/blog

Tabelle 1.2: Ergebnis einer k -means Clusterung mit $k = 9$

Cluster	Blog
Cluster 1 (21 items, Dist. 0,893)	1x umruehren bitte, Bestatter-Weblog, Blogschrott.net – Web 2.0 – Yannick Eckl, Buchhaendleralltag und Kundenwahnsinn, Das Kopfschuettel-Blog, Der Shopblogger, Exciting Commerce, F!XMBR, F5, Gluehweinjunkies, Indiskretion Ehrensache, Irgendwas ist ja immer, Lizas Welt, MC Winkels weblog, SIXTUS.NET – Blog, Spreeblick, Stefan Niggemeier, blog.tagesschau.de, imgriff.com, medienlese.com, netzpolitik.org
Cluster 2 (3 items, Dist. 0,766)	Politblog.net, USA Erklaert, law blog
Cluster 3 (4 items, Dist. 0,523)	Frank Helmschrott, Peruns Weblog, Software Guide, bueltge.de [by:ltge.de]
Cluster 4 (4 items, Dist. 0,843)	Der Schockwellenreiter (RSS-Feed), Lustige Videos, Parteibuch Ticker Feed, Webkrauts
Cluster 5 (15 items, Dist. 0,815)	ANDERS denken, BloggingTom, Communixx, ConnectedMarketing.de, Elektrischer Reporter, Karriere-Bibel, Kooperationsysteme (CSCM), PR Blogger, Sprechblase, UPLOAD, Web2Null, annalist, blog.yoda.ch, fudder – neuigkeiten aus freiburg, www.best-practice-business.de/blog
Cluster 6 (5 items, Dist. 0,656)	Fischmarkt, Mac Essentials Newsfeed, fscklog, praegnanz.de, stereopoly NewsFeed
Cluster 7 (14 items, Dist. 0,833)	Basic Thinking Blog, Blogs! Buch Blog, Blogwiese, Dr. Web Weblog, GoogleWatchBlog, Internetmarketing-News.de, Neubi-bergBlog, S-O-S SEO Blog, Telagon Sichelputzer, Wortfeld, agenturblog.de, off the record, ricdes dot com, zweinull.cc
Cluster 8 (5 items, Dist. 0,859)	Der Spiegelfechter, Lupe, der Satire-Blog, NachDenkSeiten – Die kritische Website, Nerdcore, deutsche-startups.de
Cluster 9 (5 items, Dist 0,898)	Medienrauschen, das Medienweblog, Politically Incorrect, Riesenmaschine – das brandneue Universum, pixelgraphix, ravenhorst

einige kritisch/satirische Inhalte, wie sie auch im Gesamt-Dendrogramm beieinander stehen. Cluster 2 beinhaltet die typischen Web 2.0-Blogs. Cluster 3 weist wie oben bereits erwähnt eher weniger thematischen Zusammenhang auf. Cluster 4 ließe sich z. B. mit Politik und Allgemeines überschreiben. Ähnlich Cluster 5, wobei hier wohl eher der Web 2.0 Fokus verstärkt vorhanden ist. Cluster 6 zeigt wie oben Blogs, bei denen es sich thematisch vorwiegend um Software und Bloggen an sich dreht.

Bei neun Clustern wird das Feld erwartungsgemäß weiter aufgeteilt und es ergeben sich vor allem auch „Minicluster“ mit drei bis fünf Elementen, die dann – zumindest stellenweise – eine äußerst niedrige Distanz zueinander aufweisen. Auffallend ist hier beispielsweise Cluster 3, der sich so in gleicher Weise auch in Abbildung 1.6 (links) wieder findet.

3.5 Clustern durch Dimensionsreduktion

Als Ergebnis dieses Verfahrens ergibt sich eine Art Landkarte des untersuchten Ausschnitts der deutschen Blogosphäre. Thematisch ähnliche Blogs stehen enger beieinander, unterschiedliche Themen eher weiter entfernt. Ferner finden sich im Zentrum eher allgemeine Blogs, die Ähnlichkeiten mit vielen anderen aufweisen und andererseits stehen Blogs zu Spezialthemen eher an den Außenbereichen.

Die folgenden Abbildungen zeigen einen Ausschnitt aus der Gesamtkarte der analysierten Blogs. Auf eine vollständige Darstellung wurde aus Platzgründen verzichtet. Abbildung 1.9 zeigt ein eher an politischen Themen orientierten Bereich mit den typischen Vertretern Politblog.net, USA Erklärt, Spiegelfechter etc.

Weiter rechts davon (vgl. Abbildung 1.10), d. h. im Zentrum der Karte finden sich eher die allgemeineren Blogs, mit einem breiten Fokus, die über aktuelles Tagesgeschehen berichten, wie beispielsweise imgriff, Blogwiese, Indiskretion Ehrensache etc.

Daran unmittelbar angrenzend beginnt noch weiter rechts der Bezirk der eher technik- bzw. internetorientierten Blogs (vgl. Abbildung 1.11).

4 Zusammenfassung und Ausblick

Die angeführten Ergebnisse verdeutlichen, wie sich mit Hilfe einfacher Cluster-Verfahren verschiedene Datenquellen im Internet an Hand ihrer Inhalte thematisch einteilen lassen. Konkret wurden typische Vertreter des Web 2.0 bzw. Benutzergenerierten Inhalte in Form eines Ausschnitts der deutschen Blogosphäre untersucht und an Hand der verwendeten charakteristischen Substantive klassifiziert. Als Fazit lässt sich festhalten, dass die Verfahren ihre Tauglichkeit für die konkrete Anwendung erfolgreich unter Beweis stellen konnten und die Ergebnisse im Allgemeinen durchaus den Erwartungen entsprechen. Durch die Klassifizierung konnten beispielsweise typische Vertreter zu Themen rund um das Web 2.0 als auch einige eher am aktuellen Tagesgeschehen bzw. Politik orientierte Blogs identifiziert werden. Darüber hinaus gibt es eine Reihe von Spezialthemen, die von einer kleineren Gruppe von Blogs vornehmlich behandelt werden.

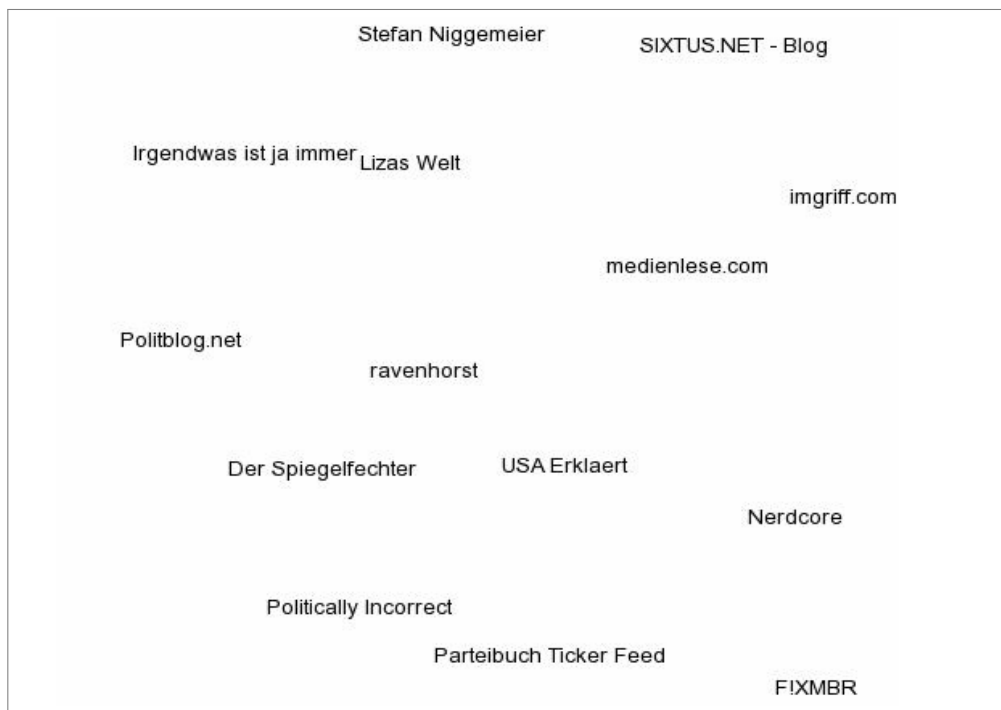


Abbildung 1.9: An politischen Themen orientierter Bereich der Blog-Landkarte



Abbildung 1.10: Allgemeine, themenvielfältige Blogs im Zentrum der Bloglandkarte

Wer bloggt was?



Abbildung 1.11: Der eher technik- und internetorientierte Bezirk der Bloglandkarte

Auch wenn Clustering naturgemäß eher eine sehr grobe und oberflächliche Einteilung darstellt, zeigen die eingesetzten Verfahren, wie schnell ein Überblick gewonnen und thematische Zusammenhänge erkannt werden können. Zudem handelt es sich um einen Vertreter des „unüberwachten Lernens“ (unsupervised learning), d. h. es sind keinerlei Vorab-Informationen für die Einteilung notwendig. Mit entsprechender technologischer Automatisierung lässt sich quasi auf „Knopfdruck“ innerhalb weniger Minuten ein aktueller Abzug der interessierenden Blogs generieren und die entsprechenden Clusterverfahren darauf anwenden. Neben einer Momentbetrachtung wie in diesem Beitrag, werden so beispielsweise auch Zeitreihenbetrachtungen möglich und es offenbaren sich leichter aktuelle Trends oder Veränderungen. Der Ansatz bietet damit auch eine Möglichkeit, um unternehmensinterne Bloglandschaften zu beobachten, zu analysieren und aktuelle Trends zu erkennen.

Literaturverzeichnis

- Backhaus, K., B. Erichson und W. Plinke (2000). *Multivariate Analysemethoden*. Berlin: Springer.
- Moosbrugger, H. und D. Frank (1992). *Clusteranalytische Methoden*. Bern: Huber.
- Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Review* 45, 167–256.
- Schmidt, J. (2008). Zu Form und Bestimmungsfaktoren weblogbasierter Netzwerke: Das Beispiel twoday.net. In C. Stegbauer und M. Jäckel (Hrsg.), *Formen der Kooperation in computerbasierten Netzwerken: Beispiele aus dem Bereich „Social Software“*, 71–93. Wiesbaden: VS-Verlag.
- Scott, J. (1991). *Social Network Analysis: A Handbook*. London: Sage.
- Segaran, T. (2007). *Programming Collective Intelligence*. San Francisco: O’Reilly.
- Surowiecki, J. (2005). *The Wisdom of Crowds – Why the Many are Smarter than the Few*. London: Abacus.
- Wasserman, S. und K. Faust (1997). *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.