TECHNISCHE UNIVERSITÄT WIEN

VIENNA UNIVERSITY OF TECHNOLOGY

D I P L O M A R B E I T

Automatische Klassifikation von Musikkünstlern basierend auf Web-Daten

Ausgeführt am Institut für

Computational Perception

der Johannes-Kepler-Universität Linz

unter der Anleitung von

Univ-Prof. Dipl.-Ing. Dr. Gerhard Widmer

durch

Peter Knees

Wien, November 2004

Unterschrift

Kurzfassung

Musikorganisation stellt in einer Zeit der stark wachsenden Verbreitung digitaler Musik eine der zentralen Herausforderungen dar. Ein bewährtes Mittel ist die Einteilung von Musik in Genres. In dieser Arbeit wird ein Ansatz zur automatischen Klassifikation von Musikkünstlern unter Verwendung von Text Categorization Methoden vorgeschlagen. Konkret werden von Suchmaschinen empfohlene Webseiten analysiert, um daraus Beschreibungen von Künstlern, in Form von Worthäufigkeiten, zu extrahieren. Zur Klassifikation kommen hauptsächlich Support Vector Machines zum Einsatz.

Die präsentierten Experimente umfassen die Evaluierung des Klassifikationsprozesses anhand einer Taxonomie von 14 Genres mit insgesamt 224 Künstlern, die Erprobung von Filtermethoden zur Steigerung der Qualität der zugrunde liegenden Daten, sowie die Abschätzung des Einflusses der Fluktuationen des Internets auf die Klassifikation durch Auswertung einer Langzeitstudie über eine Zeitspanne von knapp einem Jahr. Anhand dieser Experimente wird untersucht, wie viele Künstler zur Definition des Konzepts eines Genres erforderlich sind, welche Suchmaschine und Suchanfrage am geeignetsten sind, welche Klassifikationsgenauigkeit erwartet werden kann und ob der Ansatz als Ähnlichkeitsmaß für Künstler tauglich ist.

Inhaltsverzeichnis

1 Einleitung **4**
1.1 Über die Einteilung von Musik 4
1.2 Genres 5
1.3 Inhalt dieser Arbeit 9

2 Grundlagen **12**
2.1 Audiosignal-basierte Ansätze 12
2.1.1 Ähnlichkeitsmaße 12
2.1.2 Genre-Klassifikation 14
2.2 Verarbeitung kultureller Informationen 15
2.2.1 Collaborative Filtering 15
2.2.2 Strukturierte Metadaten 16
2.2.3 Unstrukturierte Metadaten 16
2.2.4 Songtexte 19

3 Datengewinnung **21**
3.1 Suchmaschinen 21
3.2 Suchanfragen 22
3.3 Filtermethoden 23
3.3.1 Webseiten-Profile 24
3.3.2 Eingesetzte Filter 25

4 Merkmalsextraktion **30**
4.1 Vorverarbeitung 31
4.2 Termauswahl 32
4.3 Termgewichtung 35

5 Klassifikation **37**
5.1 Support Vector Machines 37
5.2 k-nearest neighbor Klassifikator 40

6 Evaluierung **41**
6.1 Künstlerklassifikation 42
6.1.1 Auswertungen der Experimente mit Google 44
6.1.2 Auswertungen der Experimente mit Yahoo! 49

6.2 Einsatz als Ähnlichkeitsmaß 52
6.3 Abschätzung der zeitbedingten Varianz 54

7 Zusammenfassung und Ausblick 59

Literaturverzeichnis 62

A Anhang 67
A.1 Zuordnung von Künstlern zu Genres 67
A.2 Typische χ^2-gewichtete Wortlisten 69
A.3 Listen der Stop-Words 84
A.4 Ergebnisse aus Knees et al., 2004 86

1 Einleitung

Das „Verstehen“ von Musik wird allgemein als kognitiver Prozess angesehen. Da die im Gehirn ablaufenden Prozesse möglicherweise niemals vollständig formalisierbar sind, ist Modellbildung der einzige Weg, um diesbezüglich Forschung betreiben zu können. Erst der Einsatz von Computern erlaubt die Realisierung komplexer Modelle. Ziel soll es sein, Computer so zu programmieren, dass sie in der Lage sind, Musik zu analysieren, zu vergleichen, Empfehlungen abzugeben – generell Aussagen über Musik zu tätigen, wie sie auch von Menschen getätigt werden können. Programme, die diese Aufgaben bewältigen können, haben ein breites Spektrum an Anwendungsmöglichkeiten, darunter automatisches Niederschreiben von Noten zu Melodien, Wiedererkennen von Musik, automatisches Beschreiben von Musik, sowie Einteilung, Organisation und Visualisierung großer Musikarchive. Auch Ansätze in Richtung einer möglichen „Objektivierung“ können, sofern überhaupt anstrebenswert, durch den Computer entwickelt werden. Vor allem die breite Vielfalt an musikalischen Stilrichtungen und ständig anwachsende Zahl von veröffentlichten Musikstücken stellen neue Herausforderungen an die Art und Weise wie Musik organisiert und zugänglich gemacht werden kann.

1.1 Über die Einteilung von Musik

Auch wenn die Wahrnehmung von Musik ein rein subjektiver Vorgang ist, gibt es seit jeher Versuche, Musik „allgemeingültig“ (im Sinne eines möglichst breiten gesellschaftlichen Konsens) zu beschreiben. Damit einher gehen auch Bestrebungen musikalische Merkmale zu finden, anhand derer eine Klassifikation durchgeführt werden kann. Mehrere Teildisziplinen der Musikwissenschaft beschäftigen sich mit „der Systematisierung von Musik nach intersubjektiven Kriterien“, wie es in *Wikipedia*, der freien Enzyklopädie [44] genannt wird. Demnach werden verschiedene Ordnungsdimensionen zur Einteilung herangezogen:

- *Art der Beteiligten*: z.B. Vokalmusik oder Instrumentalmusik
- *Herkunft*: z.B. europäische Abstammung oder Weltmusik
- *Verwendungszweck*: z.B. funktionale oder autonome Musik

- *Religiöse Motivation*: z.B. sakrale oder weltliche Musik
- *Tonsystem*: z.B. Neunton-, Zehnton- oder Zwölftonmusik
- *Tonalitätsvorstellung*: z.B. tonale oder atonale Musik
- *Menge der Beteiligten*: Solo, Duett, Trio, etc.
- *„Wertmäßige“ Einteilung*: E-Musik oder U-Musik

Vor allem die Unterscheidung zwischen E-Musik und U-Musik, zwischen „ernster“ Musik und „Unterhaltungsmusik“, stellt einen Computer vor eine unlösbare Aufgabe. Da dieses Merkmal unmöglich objektivierbar ist, ist es auch unter Musikwissenschaftern umstritten. Ungeachtet dessen, offenbaren auch die meisten anderen Kriterien Schwächen: sie sind für Laien oft nur schwer nachvollziehbar und meistens nicht von Interesse. So sehr diese Ordnungsdimensionen aus akademischer Sicht sinnvoll sein mögen, für den gebräuchlichen Umgang mit Musik sind sie kaum geeignet.

1.2 Genres

Aus diesem Grund herrscht im Alltag meist eine andere Form der Einteilung vor: die Einteilung in Genres. Der Begriff Genre ist ein Synonym für Gattung, für den Bereich Musik kann man diesbezüglich spezifizieren und konkretisieren (Definition aus [45]):

> *Musikgenres sind Kategorien, die Musik enthalten, die einen bestimmten Stil oder die bestimmte Elemente gemeinsam hat.*

In dieser Definition wird ein weiterer, für die Gruppierung ähnlicher Musik wichtiger Aspekt vernachlässigt – die Beeinflussung der Beurteilung durch kulturelle Faktoren. Auf diesen Punkt wird später noch genauer eingegangen.

Des weiteren wird oft auch noch zwischen Genres als umfassenderen Kategorien und Stilen als feiner definierten Strömungen innerhalb eines Genres unterschieden. Diese Unterscheidung wird in dieser Arbeit nicht getroffen, im Folgenden bezeichnen die Begriffe „Genre“ und „Stil“ gleichermaßen Kategorien, die obenstehender Definition genügen.

Viele der Genres orientieren sich durchaus an den oben genannten Ordnungsdimensionen (nach [45]). So sind Genres wie „Indianische Musik“ geographisch definiert, Genres wie „Barockmusik“ großteils durch einen Zeitabschnitt und andere, wie das vorwiegend in den Vereinigten Staaten verbreitete „Barbershop“, durch technische Anforderungen, wie Zusammensetzung der Gruppe und Art des Gesangs. Manche Genres sind hingegen sehr vage definiert und manchmal auch von „Experten“ (Kritikern) erdacht, wie „Post-Rock“.

Der Vorteil von Genres besteht darin, dass sie eine meist einfache und überschaubare Einteilung bieten, die als grobe Richtlinie zum Auffinden ähnlicher Musik dienen kann. Diese einfache Form der Strukturierung erleichtert etwa die Kommunikation zwischen Menschen über Musik, da mit verbreiteten Genrenamen meist bestimmte charakteristische Musikarten assoziiert werden: „Kennst du die Band XY? Die machen so etwas wie Countrymusik mit elektronischen Einflüssen.“ Alleine dieses einfache Beispiel demonstriert, dass es (vor allem in der Populärmusik) einige bestimmte Genres gibt, die allseits bekannt sind (über deren Charakteristika Konsens herrscht) und die folglich Fixpunkte im Diskurs über Musik darstellen. „Country“ und „Elektronische Musik“ aus dem vorangegangenen Beispiel sind solche, andere sind beispielsweise „Jazz“, „Blues“, „Reggae“ oder „Hip Hop“. Genres bieten somit die einfache aber durchaus mächtige Möglichkeit, Musik durch Zugehörigkeit zu bekannten Kategorien zu beschreiben. Oftmals werden Genres eingesetzt, um überhaupt einen Ansatzpunkt zur Organisation von Musik zu bekommen. Ein Anwendungsbeispiel hierfür wäre die Anordnung von CDs in Musikgeschäften. In [29] wird hierzu festgestellt, dass dabei typischerweise eine Taxonomie in vier Ebenen benutzt wird:

1. Eine Hauptebene mit *„globalen“ Musikkategorien* (Klassische Musik, Jazz, Rock, etc.)
2. Eine Ebene mit spezifischen *Unterkategorien* (z.B. Hard Rock in Rock)
3. Typischerweise die alphabetische *Sortierung* nach Künstlernamen
4. Die *Alben* (oder allgemein Tonträger), die dann die Musik enthalten

Das Prinzip eines „top-down“-Ansatzes zur Organisation von Musik, d.h. die Spezialisierung Ebene für Ebene in einer hierarchischen Taxonomie, ausgehend von groben Kategorien, findet sich häufig, wenn es darum geht, Benutzer oder Kunden bei der Erforschung großer Musikbestände zu unterstützen. Vor allem bei den seit den letzten Jahren enorm an Popularität gewinnenden Online-Music-Stores werden Genre-Hierarchien eingesetzt, um die verfügbaren Musikstücke zu strukturieren. Dabei wird natürlich viel Bedacht auf möglichst korrekte und „effiziente“ Taxonomien gelegt, da schließlich der erzielte Gewinn von der Zahl der verkauften Stücke abhängt. Neben dem Interesse des durchschnittlichen Heimanwenders, seine Sammlung einfach und aufgrund der anwachsenden Datenmengen am besten automatisch ordnen zu lassen, besteht somit auch ein reales wirtschaftliches Interesse an guter Organisation großer Musikdatenbanken.

Gerade in diesem Bereich zeigt sich aber schnell, dass Genres nicht wirklich in der Lage sind, die zahlreichen Facetten von Musik ansatzweise zu erfassen. Genres stellen nur eine vereinfachte Sicht auf das Problem dar. Das Argument, dass Genres flexibel an verschiedene Problemstellungen angepasst werden können und ihre Berechtigung schon alleine durch ihre weite Verbreitung beziehen, erweist sich gleichzeitig als ihr größtes Manko: Genre-Taxonomien sind inkonsistent. Pachet und Cazaly haben anhand von drei öffentlich verfügbaren Genre-Taxonomien diesbezüglich Untersuchungen angestellt [29]. Die wichtigsten Kritikpunkte werden im Folgenden zusammengefasst:

- Anzahl und hierarchische Gliederung variieren stark: 430, 531 oder 719 Genres werden in 16, 5 oder 18 Meta-Genres zusammengefasst.

- Zur Bezeichnung der Genres wird kein einheitliches Vokabular verwendet: nur 70 Worte kommen in allen drei Taxonomien vor. Genres, die wie „Rock“ oder „Pop“ breite musikalische Spektren abdecken, sind unterschiedlich definiert.

- Viele Stücke und Künstler lassen sich nicht eindeutig zuordnen.

- Als eine der schwerwiegendsten Schwachstellen wird die variable Semantik der hierarchischen Verbindungen zwischen Genres angesehen. Mehrere mögliche Bedeutungen wurden ausfindig gemacht:

  - *Genealogische Bedeutung*: Darstellung einer musikalischen Entwicklung, z.B. ist laut einer Taxonomie Disco aus Pop entstanden und daher ein Kind von Pop
  - *Geographische Bedeutung*:
z.B. International – Afrika – Algerien
  - *Zusammenfassende Bedeutung*: z.B. R'n'B/Soul – Soul
  - *Wiederholungen*: diese werden oft eingesetzt, wenn eine Bezeichnung sowohl für eine übergeordnete Kategorie, als auch für ein konkretes Genre benutzt wird.
  - *Historische Abschnitte*: Barock, Klassik und Impressionismus als Kinder von Historische Abschnitte, das wiederum Kind von Klassischer Musik ist.
  - *Genrespezifische Bedeutungen*: z.B. Einteilung nach Tanzstilen

Dies führt naturgemäß zu vielen Redundanzen und inkonsistenten Klassifikationen, weswegen die Strukturen in dieser Form nicht auf automatische Systeme umgelegt werden können. Allerdings wird richtigerweise noch darauf hingewiesen, dass die strukturellen Inkonsistenzen für Menschen kein Problem darstellen, da der Mensch in der Lage ist, mit verschiedenen Bedeutungsebenen im Aufbau umzugehen.

[45] führt noch weitere Gründe dafür an, gegenüber Genres skeptisch zu sein. Dazu gehört allgemein, dass jeder Versuch, Musik zu kategorisieren zu einem gewissen Grad artifiziell ist, da Künstler Musik in allen Stilen, die sie wollen, produzieren, ohne sich dabei damit zu beschäftigen, in welchem Genre sie arbeiten. Es wird auch die Meinung eines kritischen Musikers zitiert, der den Standpunkt vertritt, dass generell die Einteilung von Musik nutzlos ist, da Genres nur Werkzeuge sind, die zur Kommerzialisierung dienen und aus der komplexen persönlichen Vision eines Künstlers eine Handelsware machen. Damit wird die Sichtweise, dass Genres oft nur zu Marketingzwecken und nicht zur Unterscheidung musikalischer Kategorien eingesetzt werden, zum Ausdruck gebracht.

In [4] stellen Aucouturier und Pachet fest, dass „Musik-Genre ein übeldefinierter Begriff ist, der in keiner intrinsischen Eigenschaft der Musik begründet ist, sondern vielmehr von kulturellen extrinsischen Gewohnheiten abhängt“. Keiner der angeführten Kritikpunkte ist von der Hand zu weisen und es herrscht wohl Einigkeit darüber, dass Genres bestenfalls ein Zwischenschritt in puncto Musikorganisation sein können.

Eine aktuelle Studie zeigt, dass für die meisten Personen Ordnungsdimensionen wie Genre nur von untergeordnetem Interesse sind [13]. Da sich private Musiksammlungen im vergleichsweise kleinen Rahmen halten, ist hier der Bedarf nach „allgemeinen“ Kriterien zur Organisation nicht gegeben. Vielmehr tendieren Personen dazu, für verschiedene Anlässe oder Lokalitäten jeweils spezifisch angepasste Organisationsformen zu wählen. Als Anwendungsgebiet für die in dieser Arbeit vorgestellten Verfahren zeichnen sich daher eher große, weite musikalische Bereiche abdeckende Datenbestände ab, deren manuelle Annotierung zu zeit- und arbeitsintensiv wäre.

Ungeachtet dessen, in welcher Art und Weise Organisation und Visualisierung von Musikarchiven zukünftig stattfinden werden, ist bereits jetzt klar, dass der Begriff der Ähnlichkeit in diesen Systemen eine zentrale Rolle einnehmen wird. Dabei wird zweifellos dem Vergleich des Klanges zweier Stücke eine besonders hohe Bedeutung zukommen. Dies kann aber nicht der einzige Aspekt sein, der in die Bewertung einfließt. Eines der Hauptanwendungsgebiete für Ähnlichkeitsmaße heute (und in Zukunft sicher noch verstärkt) sind Empfehlungssysteme (*engl. Recommendation Systems*), die anhand eines Musikstückes in einer Datenbank jene anderen Stücke suchen, von denen man annimmt, dass sie dem Benutzer gefallen werden. Der pure Klang von Musikstücken ist hier sicher wichtig, es darf allerdings nicht außer Acht gelassen werden, dass musikalische Vorlieben nicht ausschließlich aufgrund der Musik entstehen. Ein sicher sehr wichtiger Faktor hier ist das *sozio-kulturelle Umfeld* des Menschen, in dessen Kontext er betrachtet werden muß. Vor allem in der globalisierten Welt, in der Produkte und Marken weltumspannend positioniert und präsentiert werden und in der Marketing ein dominanter Faktor ist (vgl. Kritik an Genres oben), spielt das „Image“, das ein Produkt hat (und als solches ist Musik zu verstehen) eine entscheidende Rolle, d.h., bei der Meinungsbildung über Musik, Musikrichtungen und Künstler hat das öffentliche Erscheinungsbild enormen Einfluss. Dies wirkt sich wiederum darauf aus, was als „ähnlich“ empfunden wird. Recommendation Systems, deren Aufgabe es ist, das musikalische Wertesystem des Benutzers möglichst gut zu erfassen und zu modellieren, müssen daher auch verstärkt *kulturelle Aspekte* in ihre Bewertungen miteinbeziehen. Die Ausprägungen des kulturellen Wissens, das es zu erfassen gibt, können vielfältig sein. Informationen über Künstler, der Inhalt von Liedtexten oder auch nicht naheliegende Formen wie die Ähnlichkeit von Fan-Gruppen rund um den Künstler, all das kann aussagekräftig sein und, sofern erfassbar, in das Modell einfließen. Diese verschiedensten Aspekte fasst Brian Whitman in [42] unter dem Begriff *community metadata* zusammen. Es ist nahe liegend, dass versucht wird, kulturelles Wissen in erster Linie aus dem Internet zu beziehen. Nicht nur, dass Informationen auf diesem Wege am einfachsten erfassbar sind, sondern auch die Tatsache, dass das Internet im Moment wohl jenes Medium ist, das gesellschaftliche Veränderungen und Trends am schnellsten widerspiegelt, macht die Analyse von Web-Daten so attraktiv.

So wichtig es ist, Ähnlichkeit konstatieren zu können, so problematisch ist es, festzustellen, ob diese „objektiv“ korrekt ist. Die einzige diesbezüglich Wahrheit, die *ground truth*, kann wiederum nur das Empfinden des Menschen bzw. der möglichst breite Konsens der Gesellschaft sein. Aber auch abgesehen davon, stellt sich die Frage, wie Ähnlichkeit automatisiert evaluiert werden kann. Eine einfache Möglichkeit ist, die Klassifikationsgenauigkeit in Genres auszuwerten. Dies bringt zwar einige der oben angesprochenen Probleme mit sich, stellt aber immer noch eine der praktikabelsten und tauglichsten Methoden dar.

1.3 Inhalt dieser Arbeit

Aufbauend auf einem bereits publizierten Paper ([20]), in dem erste Ergebnisse, die aus Überlegungen zu dieser Arbeit resultierten, präsentiert wurden, werden eine Methode zur automatischen Klassifikation von Musikkünstlern in Genres und ein Ansatz zur Entwicklung eines Ähnlichkeitsmaßes vorgeschlagen, die ausschließlich auf Textdaten, die von gewöhnlichen Webseiten extrahiert werden, basieren. Dabei werden die in [20] vorgestellten Methoden und Resultate ausführlich beschrieben, sowie Verbesserungen im Bereich der Datengewinnung, wie Einsatz von Filtern, eingebracht, mit deren Hilfe die Qualität des Systems noch gesteigert werden kann.

Die Grundidee des präsentierten Verfahrens besteht darin, Künstler in eine vordefinierte Taxonomie von n Genres einzuordnen. Jedes dieser n Genres wird dadurch repräsentiert, dass m typische Künstler dazu angegeben werden

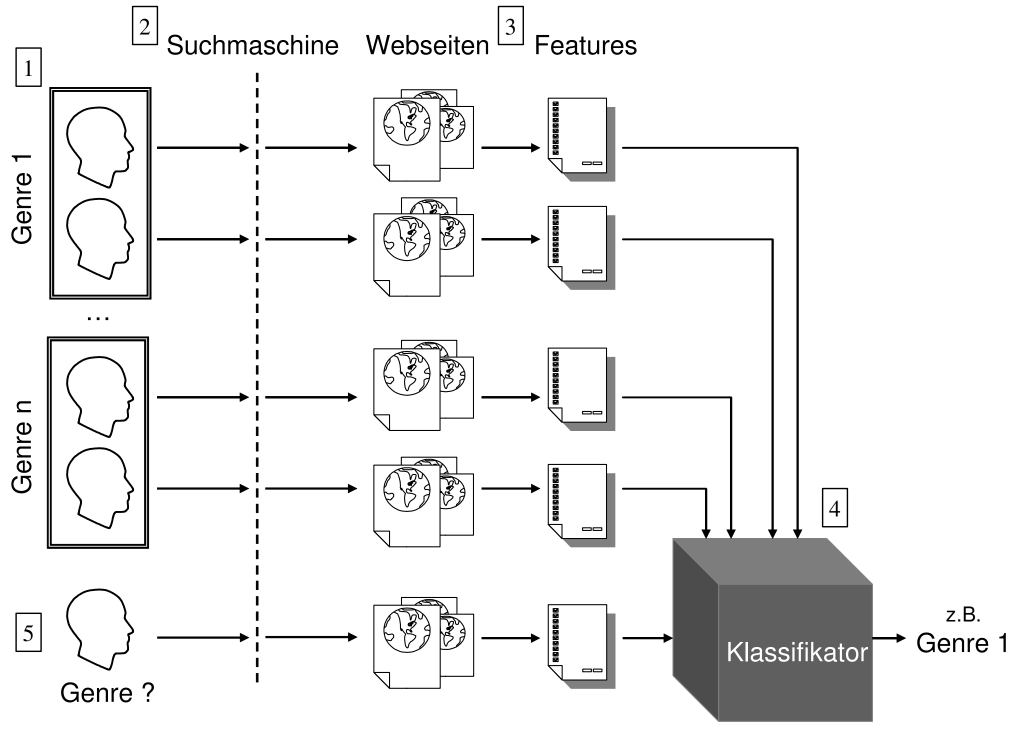

Abbildung 1.1: Überblick über das vorgestellte Klassifikationsverfahren. Voraussetzung ist eine Menge von n Genres, zu denen typische Künstler bekannt sind (1). Für jeden Künstler werden von Suchmaschinen empfohlene Webseiten heruntergeladen (2). Aus diesen Webseiten werden charakteristische Merkmale (Features) extrahiert (3), mit deren Hilfe ein Klassifikator trainiert wird (4). Neue Künstler (5) können nun durch den Klassifikator einer Klasse zugeordnet werden (hier beispielsweise Genre 1).

(zur Vereinfachung wird hier für alle Genres die selbe Anzahl an Künstlern vorausgesetzt). Die Klassifikation neuer, unbekannter Künstler erfolgt durch einen automatischen Klassifikator (in erster Linie Support Vector Machines), der durch die vorgegebenen Künstler die Konzepte der Genres lernt. Ein Künstler wird durch Merkmale beschrieben, die aus Webseiten extrahiert werden, die sich mit dem jeweiligen Künstler befassen. Die Funktionsweise des Ansatzes ist schematisch in Abbildung 1.1 dargestellt.

Diese Arbeit ist wie folgt organisiert. In Kapitel 2 werden bestehende Ansätze aus dem Bereich der signalbasierten Audioverarbeitung und dem Bereich der Metadaten-Analyse besprochen. Mit Kapitel 3, in dem darauf eingegangen wird, wie die diesem Verfahren zugrunde liegenden Daten gewonnen werden, beginnt die detaillierte Beschreibung der präsentierten Methode. Kapitel 4 befasst sich mit der Verarbeitung der Daten und der Extraktion von typischen Merkmalen. In Kapitel 5 werden kurz die Grundlagen jener Machine Learning-Verfahren, die dann in Kapitel 6 zum Einsatz kommen, erläutert. Kapitel 6 befasst sich, anhand ausführlicher Experimente, eingehend mit den Fragen „Wieviele Künstler muss man einem Computer vorgeben, um das Konzept eines Musikgenres ausreichend definieren zu können?“, „Welche Suchmaschine bringt die brauchbarsten Ausgangsdaten?“, „Welche Suchanfragen erzielen die besten Ergebnisse?“, „Welche Klassifikationsgenauigkeit kann man erwarten?“ und „Wie sehr eignet sich der Ansatz als Ähnlichkeitsmaß für Künstler?“. Weiters wird untersucht, wie sehr die Ergebnisse der präsentierten Verfahren von „Fluktuationen“ des Internet, also kurzfristigen Änderungen in den Daten, beeinflusst werden. In Kapitel 7 werden die wichtigsten Punkte dieser Arbeit nochmals zusammengefasst und Schlüsse gezogen, aus denen sich Perspektiven für zukünftige Verbesserungen ergeben.

2 Grundlagen

Die Grundlagen dieser Arbeit stammen nicht primär aus dem Bereich der Musikwissenschaften. Vielmehr entspringen die direkt dieser Arbeit vorangegangenen Ansätze einem relativ jungen Forschungsbereich der Informatik: dem *Music Information Retrieval* (MIR). Ziel ist, umgangssprachlich formuliert, das „Herausholen“ von Information aus Musik. Konkret manifestiert sich dies unter anderem in Bestrebungen formale und computationale Methoden zur Analyse, Repräsentation, Modellierung, Klassifikation und Clustering von Musik zu finden. Dieses rapide wachsende Gebiet (siehe [10]) umfasst im Wesentlichen folgende Teilbereiche:

1. Verarbeitung von Musik, die in symbolischer Form vorliegt
2. Verarbeitung von Musik, die als Signal vorliegt
3. Verarbeitung von kulturellen Informationen (*community metadata*)
4. Benutzer-orientierte Entwicklung von Schnittstellen

Dazu muss festgestellt werden, dass die Grenzen zwischen diesen Bereichen nicht immer klar zu ziehen sind und viele der bisher vorgestellten Ansätze Methoden aus verschiedenen Bereichen miteinbeziehen und kombinieren. Die Grundlagen dieser Arbeit sind im Bereich „Verarbeitung von kulturellen Informationen“ anzusiedeln. Obwohl der direkte Zusammenhang von Punkt 2 und dieser Arbeit nicht gegeben ist, sollen zumindest ansatzweise die grundlegenden Ergebnisse und Fortschritte, aber auch die Grenzen auf diesem Gebiet erläutert werden, um zu zeigen, warum der Verarbeitung kultureller Information in Zukunft noch größere Bedeutung zukommen kann.

2.1 Audiosignal-basierte Ansätze

2.1.1 Ähnlichkeitsmaße

In den letzten Jahren wurde eine Vielzahl von Verfahren zur Berechnung von Ähnlichkeit zwischen Audiosignalen publiziert (z.B. [1, 2, 7, 8, 16, 25, 32, 34, 39]). Aucouturier und Pachet stellen in [5] fest, dass allen vorgeschlagenen Ansätzen im Wesentlichen dieselben Verarbeitungsschritte in unterschiedlichen Variationen mit verschiedenen Parametern zugrundeliegen:

1. Eine Unterteilung des Audiosignals in (typischerweise jeweils zur Hälfte) überlappende Segmente (*frames*) von 20-50ms
2. Berechnung von Merkmalen (sog. *Features*) für jedes Frame, typischerweise durch MFCCs[1].
3. Berechnung eines statistischen Modells der Verteilung der MFCCs
4. Vergleich der Modelle

Die präsentierten Verfahren unterscheiden sich beispielsweise in der Anzahl der gewählten MFCCs (im Bereich von 8 bis 20). Auch sonst weisen die Ansätze trotz ihrer prinzipiellen Ähnlichkeit Differenzen auf. In [32] werden 5 Ansätze verglichen, in denen einige der verschiedenen Methoden zum Einsatz kommen:

- **Logan und Salomon** [25]: Hier basiert die Feature-Extraktion auf 19 MFCCs und resultiert in 16 Werten, die mithilfe von k-means Clustering ermittelt werden. Zur Berechnung der Distanz zwischen zwei Modellen wird die Earth-Mover's-Distanz, ein Verfahren aus der Bildverarbeitung, eingesetzt.
- **Aucouturier und Pachet** [2]: Der Hauptunterschied zu Logan und Salomon ist, dass bedeutend weniger MFCCs, nämlich 8, Verwendung finden. Weiters wird statt eines k-means Clusterings ein Gauss'sches Mischmodell (GMM) zur Modellierung der Features eingesetzt. Der Vergleich der Modelle wird probabilistisch durchgeführt, indem aus der Verteilung eines Stückes gesampled wird und errechnet wird, wie wahrscheinlich die Samples von der Verteilung eines anderen Stücks generiert werden könnten.
- **Spectrum Histograms** [31]: Den Berechnungen liegt nicht direkt das Signal zugrunde, sondern eine Sone/Bark-Repräsentation des Signals, die sich an der Physiologie des menschlichen Ohrs orientiert und daher stärker die für den Menschen wichtigen Frequenzbereiche berücksichtigt. Die Beschreibung eines Musikstücks geschieht über Histogramme, die angeben, wie oft bestimmte Lautstärke-Level in den Frequenzbändern erreicht werden. Der Vergleich von Stücken erfolgt über Distanzberechnung 1000-dimensionaler Vektoren im euklidischen Raum, die durch Anwendung einer Principal-Component-Analysis (PCA) komprimiert werden können, wodurch das Verfahren effizienter als andere Ansätze ist.

[1]Mel-Frequency cepstrum Coefficients; die Koeffizienten einer nicht-linearen Frequenztransformation des Signals in die perzeptuelle Mel-Skala

- **Periodicity Histograms** (z.B. [39]): Die Grundidee ist es, nur regelmäßig wiederkehrende Taktschläge, unabhängig von der Frequenz zu beschreiben. Abgesehen von diesem Unterschied ist dieser Ansatz ähnlich zu jenem der Spectrum Histograms.

- **Fluctuation Patterns** [34]: Der Unterschied zu den Periodicity Histograms besteht darin, dass hier eine teure Berechnung (die des sog. Comb-Filters) durch eine Fast-Fourier-Transformation (FFT) ersetzt wird.

In [5] wird für alle Publikationen aus diesem Bereich festgehalten, dass ein Verfahren präsentiert wird, dessen Anwendbarkeit gezeigt wird und dessen Ergebnisse „vielversprechend" seien, wodurch impliziert wird, dass durch Feinabstimmung der Parameter nahezu „perfekte" Ergebnisse erzielt werden könnten. Um dies näher zu beleuchten, wird das in [2] vorgeschlagene Verfahren durch systematisches Variieren aller Parameter optimiert. Die Autoren kommen zu dem Schluss, dass eine Verbesserung des Verfahrens, evaluiert gegen eine ground truth von Metadaten des All Music Guide, über eine Präzision von 65% hinaus nicht möglich ist. Für die Qualität der Methoden spielen die meisten Parameter nur eine untergeordnete Rolle, auch die verschiedenen Verfahren weisen im Allgemeinen ähnliche Ergebnisse vor. Es wird somit als große Herausforderung angesehen, konzeptuell gänzlich andere Verfahren zu entwickeln, die in der Lage sind, die erreichte Schranke zu überwinden.

2.1.2 Genre-Klassifikation

Die für Ähnlichkeitsmaße gewonnenen Features können prinzipiell auch eingesetzt werden, um automatische Genre-Klassifikatoren zu erlernen. Nachdem sich die dafür vorgestellten Ansätze kaum von denen zur Berechnung von Ähnlichkeitsmaßen unterscheiden, werden hier nur ausgewählte Aspekte der Arbeiten aus dem Bereich der Audiosignal-basierten Genre-Klassifikation behandelt, nämlich die zum Einsatz kommende Genre-Taxonomie und die erzielte Klassifikationsgenauigkeit.

In einer der ersten Arbeiten über Musik-Klassifikation benutzen Tzanetakis, Essl und Cook [39] 6 Genres (Klassische Musik, Country, Disco, Hip Hop, Jazz und Rock), wobei Klassische Musik wiederum in Choral-, Orchester-, Klaviermusik und Streichquartett unterteilt wird. In [38] wird diese Taxonomie um Blues, Reggae, Pop und Metal erweitert. Des weiteren wird Jazz in 6 Subkategorien unterteilt (Bigband, Cool, Fusion, Piano, Quartet und Swing). In den Experimenten werden die Unterkategorien individuell evaluiert. Für die 10 allgemeinen Kategorien wird eine Klassifikationsgenauigkeit von 61% erzielt. In [9] wird eine hierarchisch strukturierte Taxonomie mit 13 verschiedenen Musikgenres vorgeschlagen.

Andere Arbeiten befassen sich im Allgemeinen mit kleineren Mengen von Genres. In [46, 37] werden 4 Kategorien (Pop, Country, Jazz und Klassische Musik) benutzt und Klassifikationsgenauigkeiten von 93% und 89% werden erreicht. In [27] kommen 7 Genres (Jazz, Folk, Elektronische Musik, R&B, Rock, Reggae und Vocal) zum Einsatz und die Gesamtgenauigkeit ist 74%.

2.2 Verarbeitung kultureller Informationen

2.2.1 Collaborative Filtering

Der „einfachste" Weg zur automatischen Empfehlung von Musik ist der des *collaborative filtering*. Beim collaborative filtering werden neuen Benutzern „Einheiten" (im konkreten Fall der Musik: Stücke, Alben, Künstler) auf Basis der Präferenzen anderer Benutzer mit ähnlichem Geschmack empfohlen (u.A. [18]). Ein collaborative filtering System ist beispielsweise in der Lage, Benutzer A die Band *The Rolling Stones* zu empfehlen, da er die selben Platten gekauft hat wie Benutzer B, der außerdem auch noch Platten von *The Rolling Stones* gekauft hat. Es werden also Profile von Benutzern angelegt (und adaptiert), um dann gleichartiges Kundenverhalten enttarnen zu können. Solche Systeme sind weit verbreitet, da sie einerseits oftmals brauchbare Resultate liefern und andererseits auf eine Vielzahl von Gebieten angewandt werden können (z.B. bei Amazon[2]), bedürfen aber einer gewissen Anzahl an Kunden, deren Kaufverhalten dann evaluiert werden kann. Um die Qualität solcher Systeme zu verbessern, können diese noch um einen Mechanismus erweitert werden, der Daten aus dem Web einbezieht, die durch spezielle Webcrawler gewonnen werden [17].

Um die Notwendigkeit der Benutzerstatistiken umgehen zu können, wird von Cohen und Fan [11] ein Verfahren vorgeschlagen, das Daten, die als Basis für collaborative filtering Techniken dienen können, mittels Webcrawlern aus Websites gewinnen kann und sogenannte „Pseudo-User" generiert. Im Gegensatz zu [17] sind die Webcrawler allerdings nicht auf besondere Strukturen in den Websites angewiesen. Als Ausgangspunkt dienen gewöhnliche kommerzielle Suchmaschinen. Um ähnliche Künstler entdecken zu können, werden Websites über Altavista[3] gesucht, die den Namen eines Künstlers enthalten. Danach werden jene Seiten, die zumindest doppelt in den Suchresultaten (d.h. bei zumindest zwei verschiedenen Künstlern) vorkommen, nach

[2] http://www.amazon.com
[3] http://www.altavista.com

gemeinsamen Vorkommnissen von Künstlernamen durchsucht und – nach einigen Verarbeitungsschritten – „Pseudo-User“ konstruiert. Es wird gezeigt, dass diese Form des „fiktiven“ collaborative filtering durchaus zu besseren Resultaten führen kann, als die klassische, auf realen Benutzerstatistiken beruhende Form.

2.2.2 Strukturierte Metadaten

In [30] wird von Pachet, Westerman und Laigre ein Ansatz zur Berechnung der Ähnlichkeiten von Künstlern und Songs präsentiert, der das gemeinsame Auftreten auf CD-Compilations, in Playlists von Radiostationen und auf Webpages analysiert. Dabei wird angenommen, dass zwei Lieder ähnlich sind, wenn sie sich auf dem selben CD-Sampler befinden (als Quelle für die Alben dient der Web-Dienst CDDB[4], der Inhalte von CDs in strukturierter Form zur Verfügung stellt), sie bei einer Radiostation direkt hintereinander gespielt wurden oder auf der selben Webpage erwähnt werden. Als eines der Hauptprobleme erweist sich hier das Auffinden von Künstler- und Songname, da es in keiner der drei Quellen eine eindeutige Regel zur Abtrennung und Anordnung dieser gibt. Basierend auf den extrahierten Ähnlichkeiten können allgemein populäre Künstler gefunden, sowie durch Clustering-Techniken grobe Genre-Strukturen entdeckt werden. Die Erkenntnis, dass Gruppen von, in Bezug auf Genres, ähnlichen Künstlern ohne weiteres Vorwissen gefunden werden können, wobei nur kulturelle Daten miteinbezogen werden, wird auch durch [3] unterstützt.

2.2.3 Unstrukturierte Metadaten

Die bisher angesprochenen Ansätze setzen, abgesehen von der Auswertung gemeinsamer Vorkommnisse auf Websites in [30], alle eine gewisse Struktur in den Daten voraus. Es existiert allerdings auch eine Reihe von Möglichkeiten, kulturelles Wissen aus allgemeinen Internetseiten und frei verfügbaren Plattenkritiken zu gewinnen. Viele der vorgestellten Arbeiten bedienen sich bewährter Methoden aus dem Bereich des „klassischen“ Information Retrieval.

In [42] schlagen Whitman und Lawrence Methoden für unüberwachtes Lernen von Textprofilen von Musikkünstlern vor, die für Klassifikation und Empfehlungen, sowie als Ergänzung zu collaborative filtering eingesetzt werden können. Die zugrunde liegende Idee ist es, Anfragen an Suchmaschinen zu stellen, um Webseiten, die etwas mit den Künstlern zu tun haben, zu er-

[4]http://www.gracenote.com

halten. Nach dem Herunterladen dieser Seiten werden Text- und Sprachcharakteristika extrahiert (auch hier spricht man wieder von *Features*; vgl. 2.1.1), die dann den Künstler beschreiben. Anhand dieser Beschreibungen werden Vergleiche angestellt und Ähnlichkeiten berechnet. Es werden Experimente auf einem Satz von 400 Künstlern durchgeführt, die über einen Zeitraum von drei Wochen die populärsten in der Musiktauschbörse OpenNap waren. Zur Evaluierung werden als ground truth (als „Wahrheit") Daten des All Music Guide[5] herangezogen. Dabei handelt es sich um eine umfassende Datenbank über Musik, die von „Experten" erstellt wird. Von Vorteil ist hierbei, dass es zu jedem in der Datenbank befindlichen Künstler eine Liste von ähnlichen und verwandten Künstlern gibt. Auch wenn diese Liste nicht immer nachvollziehbar sein mag und sich auch die Frage stellt, wer die Personen sind, die diese Listen editieren, bietet dieses „Expertenwissen" immer noch genug Referenz, um die Qualität eines Ähnlichkeitsmaßes evaluieren zu können. Um an Metadaten, die mit dem Künstler in Verbindung stehen, zu gelangen, werden Anfragen an die Suchmaschine Google[6] geschickt, die aus dem Namen des Künstlers sowie den Wörtern *music* und *review* bestehen. Dadurch erhofft man sich, Seiten zu bekommen, die sich auch tatsächlich mit der Musik auseinander setzen. Solche Seiten sind in erster Linie Plattenkritiken (engl. *reviews*). Aus dem textlichen Inhalt werden dann die Features durch Extraktion von *a.* Unigrammen (Einzelwörtern), *b.* Bigrammen (jeweils Sequenzen von zwei aufeinander folgenden Worten) und *c.* Part of Speech-Tagging (kurz PoS-Tagging) mit anschließendem Noun Phrase Chunking (kurz NP-Chunking) gewonnen. Als PoS-Tagging versteht man einen Vorgang, in dem zu jedem Wort in einem Satz seine grammatikalische Funktion innerhalb des Satzes ausfindig gemacht wird und unter NP-Chunking, dass Nominalphrasen, also Nomina und jene Satzteile, die die Nomina genauer beschreiben (meist Adjektive), extrahiert werden. Die Autoren entschließen sich außerdem dazu, Adjektiven eine höhere Bedeutung zukommen zu lassen, da sie diese als die „beschreibendsten" Terme (Wörter) ansehen. Für jede Form von extrahierten Features werden „Überlappungswerte" der Ausprägungen zwischen den Künstlern berechnet, die Aufschluss über die Ähnlichkeit geben. Eine Grundlage für diese Werte ist der $tf \times idf$ Wert (siehe 4.3). Außerdem schlagen Whitman und Lawrence eine weitere Art zur Berechnung von Künstlerähnlichkeit vor. Ausgehend von der OpenNet-Tauschbörse kann man Künstler dann als ähnlich betrachten, wenn sie sich oft gemeinsam in Sammlungen von Benutzern befinden. Dabei muss auch berücksichtigt werden, dass sehr populäre Künstler in vielen Sammlungen vorkommen (vgl. Abschnitt 2.2.2).

[5] http://www.allmusic.com
[6] http://www.google.com

Aufbauend auf dieser Arbeit haben Baumann und Hummel in [6] Verfahren zur „Verbesserung“ der zugrunde liegenden Daten angewandt. Der erste Schritt, den sie durchführen, ist das Entfernen aller empfangenen Seiten, die größer als 40kB sind, da davon ausgegangen wird, dass Seiten, die größer sind, nicht nur Rezensionen enthalten können. Andere Inhalte sollen nicht verwertet werden. Des weiteren werden aus Tabellen nur Zellinhalte verarbeitet, die zumindest einen ganzen Satz, der aus mindestens 60 Buchstaben besteht, enthält. Damit sollen Werbeinhalte weitestgehend vermieden werden. Außerdem werden Punkte vergeben, je nachdem, ob sich die Worte *music*, *review* oder, im besten Fall, der Künstlername in der URL, im Titel oder am Beginn der Website befinden. Websites, deren erreichte Punkte unterhalb eines gewissen Schwellenwerts liegen, werden ignoriert.

Seinen Ansatz aus [42] weiterverfolgend, hat Whitman, diesmal gemeinsam mit Smaragdis, gezeigt, dass die gewonnenen Ähnlichkeitswerte auch dazu verwendet werden können, Künstler in 5 Stilrichtungen (Heavy Metal, Contemporary Country, Hardcore Rap, Intelligent Dance Music und R&B) zu klassifizieren [43]. Dazu wird eine Abwandlung des k-NN Klassifikators (siehe Abschnitt 5.2) benutzt. Eine der Erkenntnisse dieser Arbeit ist, dass der Einsatz von community metadata für bestimmte Genres, wie Intelligent Dance Music, gut funktioniert, für andere, wie Hardcore Rap, allerdings nicht. Nachdem ein ebenfalls präsentiertes, auf akustischen Merkmalen basierendes Verfahren mehr oder weniger die komplementären Stärken und Schwächen aufweist, wird vorgeschlagen, die beiden Ansätze zu kombinieren, um optimale Resultate zu erzielen. In [20] wird gezeigt, dass der dieser Arbeit zugrunde liegende Ansatz bessere Resultate erzielt, als der Metadaten-basierte Teilansatz aus [42].

Eine der wichtigsten Arbeiten zum Thema Evaluierung von Ähnlichkeitsmaßen ist [8] von Berenzweig, Logan, Ellis und Whitman. Darin werden verschiedene Maße, sowohl aus dem Bereich der akustischen Verarbeitung, als auch aus dem Bereich der kulturellen Information verglichen. In einer vorangegangenen Arbeit ([15]), die nur kulturelle Faktoren behandelt, fassen Ellis, Whitman, Berenzweig und Lawrence einige der Probleme, die unmittelbar mit der „objektiven“ Evaluierung von Ähnlichkeit verknüpft sind, zusammen:

- *Individuelle Abweichungen*: Die Beurteilung verschiedener Personen von Ähnlichkeit von Künstlern ist inkonsistent. Außerdem sind Urteile sehr stimmungsabhängig. Dazu verändern sich Geschmäcker im Laufe der Zeit und es besteht die Gefahr, dass Musikrichtungen, die für Hörer nicht interessant sind, generell als „klingt alles gleich“ bewertet werden.

- *Mehrere Sichtweisen*: Je nach Bewertung der Wichtigkeit bestimmter Faktoren, ergeben sich andere Perspektiven auf die Ähnlichkeit. (vgl. Kapitel 1)
- *Asymmetrie*: Beispielsweise kann über viele (auch eher unbekannte) Künstler ausgesagt werden, dass sie wie *The Beatles* klingen, allerdings wird in den wenigsten Fällen jemand behaupten, dass *The Beatles* so klingen wie ein unbekannter Künstler. Der Grund dafür ist, dass *The Beatles* als Prototyp für eine bestimmte Art von Musik dienen.
- *Variabilität und Vielfalt*: Im Laufe ihrer Karriere ändern Künstler oft ihren Stil. Auch innerhalb eines einzigen Albums werden oft viele verschiedene Stilrichtungen vereint.

Allerdings sind die Autoren davon überzeugt, dass trotz aller Unterschiede eine „durchschnittliche“ Urteilsfindung, mit der die meisten Personen einverstanden wären, konstruiert werden kann. Um die verschiedenen Ähnlichkeitsmaße evaluieren zu können, wird als ground truth die Meinung von Menschen herangezogen. In einer Untersuchung, die per Internet als eine Art Spiel durchgeführt wird, werden Personen befragt, welcher Künstler aus einer vorgegebenen Liste ihrer Meinung nach am ähnlichsten zu einem vorgegebenen ist. Insgesamt wurden etwa 10000 Antworten gesammelt, damit ist es allerdings nicht möglich, alle Distanzen zwischen Künstlern erschöpfend zu erfassen. Um ein Maß für die Kompetenz der Befragten zu bekommen, werden hin und wieder zufällig generierte Bandnamen zur Auswahl gestellt. Wählt ein Teilnehmer eine dieser Bands, kann davon ausgegangen werden, dass sein Fachwissen als nicht allzu hoch einzustufen ist. Evaluiert werden die Ansätze, die bereits in vorangegangenen Arbeiten vorgestellt wurden: Erdös-Distanz durch die „ähnliche Künstler“-Relation des All Music Guide (Expertenmeinungen), gemeinsames Auftreten in Benutzersammlungen des OpenNap Netzwerks, Ähnlichkeit basierend auf Webtext-Analyse, sowie das gemeinsame Auftreten in Playlists, die im Internet verfügbar sind. Es zeigt sich, dass, evaluiert gegen die ground truth der Benutzermeinung, das Expertenwissen am besten abschneidet, gefolgt von der OpenNap-Ähnlichkeit, die, wie sich später zeigt, sehr hohe Übereinstimmung mit dem Auftreten in Playlists hat. Die Ähnlichkeiten, die auf Webtext basieren, erzielen hier keine besonders guten Resultate.

2.2.4 Songtexte

Ein weiterer Ansatz zur Beschreibung von Musik ist die Auswertung von Songtexten. Obwohl diese Vorteile gegenüber anderen Formen von Metadaten haben, wie z.B., dass es eine (eingeschränkt) eindeutige Zuordnung Musikstück – Text gibt und Texte einfach zu bekommen sind, gibt es bisher kaum Arbeiten, die sich ernsthaft mit dieser Form beschäftigt haben. Eine Ausnahme hierzu bildet [24], wo auf automatisch aus dem Internet[7] bezogene Songtexte (engl. *lyrics*) standard-Textverarbeitungsansätze angewandt werden, um ihren semantischen Gehalt zu charakterisieren. Für ca. 400 Künstler (zu Vergleichszwecken die selben wie in [42]) wurden ca. 15000 Songtexte heruntergeladen. Die Künstler wurden anhand von Informationen des All Music Guide in 5 Genres eingeteilt (Reggae, Country, Newage, Rap und Rock), um dann alle Lieder eines Genres als Gesamtheit zu betrachten. Es zeigt sich, dass Genres wie Rock und Country ein ähnliches Vokabular verwenden, Genres wie Newage oder Rap allerdings durch die benutzten Wörter unterschieden werden können. In Bezug auf die in [15] durchgeführte Benutzerbefragung liefern Songtexte bei weitem schlechtere Ergebnisse als Audiobasierte Verfahren. Nachdem Klassifikationsfehler aber in unterschiedlichen Bereichen beobachtet werden können, wird die Einbeziehung von Songtexten als gute Ergänzung zu bestehenden Ansätzen angesehen.

[7]http://www.azlyrics.com

3 Datengewinnung

Die Daten, die dem in dieser Arbeit präsentierten Ansatz zur automatischen Klassifikation von Musikkünstlern zugrunde liegen, sind ausschließlich textbasiert. Anhand von Informationen zum Künstler bzw. allgemein von Texten, die mit einem Künstler in Verbindung stehen, werden typische Merkmale extrahiert, die zur Beschreibung des Künstlers verwendet werden. Dazu wäre es möglich, auf spezielle Informationen wie Biographien oder andere von Experten erstellten Zusammenfassungen zurückzugreifen. Hierbei stellt sich aber wieder die Frage nach der Objektivität, da Musikanthologien sehr oft von der Sichtweise der Autoren geprägt sind. Um so etwas wie Objektivität erreichen zu können, muss versucht werden, die allgemeine Wahrnehmung eines Künstlers in der Gesellschaft zu erfassen. Dazu eignen sich mehrere, unter Umständen verschiedene Sichtweisen besser als starr vorgegebenes Expertenwissen. Deshalb werden im Zuge der Datengewinnung für das hier vorgestellte Verfahren Internetseiten benutzt, die etwas über den Künstler aussagen. Es wird also auf das *kollektive Wissen* des Internet zurückgegriffen.

3.1 Suchmaschinen

Für jeden Künstler wird das Internet mit zwei verschiedenen Suchmaschinen durchsucht. Es handelt sich dabei um die Suchmaschinen Google und Yahoo![1].

Um die Dienste von Google für eigene Zwecke nutzen zu können, bietet Google die sogenannte Google API an, deren Einsatz sich für Projekte aus dem Bereich Music Information Retrieval empfiehlt [48]. Bereits in [6] ist diese Schnittstelle zum Einsatz gekommen. Google ermöglicht es Programmentwicklern, nach einer Registrierung, aus mehreren Programmiersprachen auf die angebotenen Services zuzugreifen. Allerdings ist die Anzahl der Anfragen pro Benutzer auf 1000 per Tag limitiert, was vor allem bei größeren Datensätzen zu Problemen führen kann. Dies wird noch dadurch verstärkt, dass pro Anfrage maximal 10 Ergebnisse zurückgeliefert werden. Ein Vorteil besteht darin, dass auf die „Cached Sites“ von Google zugegriffen werden kann. Damit ist es möglich, auch wirklich jene Seiten und Informationen zu bekommen, aufgrund derer Google diese Seite als relevant bezüglich der

[1] http://www.yahoo.com

Suchanfrage eingestuft hat, auch dann, wenn der ursprüngliche Anbieter gerade nicht erreichbar ist. Bei der Durchführung dieser Arbeit wird jedoch auf diese Möglichkeit verzichtet, da auch diese Funktion das Kontingent an erlaubten Anfragen belastet und somit enormen Zeitaufwand nach sich ziehen würde. Der Einsatz der Google API empfiehlt sich allerdings nicht nur, weil die Einbindung problemlos erfolgt, sondern auch, weil die automatisierte Abfrage von Google abseits der zur Verfügung gestellten API untersagt ist.

Für den Zugriff auf Yahoo! existiert eine solche API nicht, weshalb hier auf selbstgeschriebene Funktionen zurückgegriffen werden muss. Dazu muss zuerst die Struktur der bei Suchanfragen generierten URLs analysiert und für die eigenen Zwecke reproduzierbar gemacht werden. Über von der Programmiersprache Java zur Verfügung gestellte Methoden zum Herunterladen von Webseiten, wird dann eine normale Antwortseite von Yahoo! empfangen. Um daraus die Suchresultate extrahieren zu können, ist Wissen über den Aufbau dieser Seiten von Nöten. Besonders schwierig ist hierbei zum einen die Unterscheidung von Links zu Suchresultaten und Links zu Services von Yahoo! und Werbeeinschaltungen. Hinzu kommt noch eine Besonderheit von Yahoo! bei Suchergebnissen: diese müssen erst aus speziellen Links zu Seiten, die zur Erfassung des Benutzerverhaltens dienen, extrahiert werden.

3.2 Suchanfragen

Als Suchanfragen (*query string*) kommen drei verschiedene Varianten zum Einsatz:

- `"`*`Künstlername`*`"`
- `"`*`Künstlername`*`" music review`
- `"`*`Künstlername`*`" music genre style`

Dass der Künstlername zwischen doppelten Hochkommata steht, bedeutet, dass dieser Begriff als Phrase vorkommen muss, d.h. die eingeschlossenen Worte müssen unmittelbar in dieser Reihenfolge hintereinander auf der Website stehen. Damit soll verhindert werden, dass für Künstler wie z.B. *James Last* Websites gefunden werden, auf denen irgendwo im Text der Name *James* auftritt und an ganz anderer Stelle das Wort *last*, was vor allem bei englischen Texten sehr häufig vorkommen kann.

Mit der ersten Abfrage werden voraussichtlich keine allzu guten Ergebnisse erzielt werden können. Die Suche alleine nach dem Künstlernamen kann zu vielen Ergebnissen führen, die mit dem Künstler an sich gar nichts zu tun haben. Bei Bands wie *Bush* werden auf diese Art primär Seiten aus dem Bereich der Politik gefunden. In die Experimente wird diese Möglichkeit trotzdem mit einbezogen, da in dieser Arbeit auch der Unterschied zwischen Abfragen mit und Abfragen ohne unterstützende Suchterme evaluiert und quantifiziert werden soll. Der zweite Typ Abfrage versucht über die Hilfsterme *music* und *review* den Suchraum auf relevantere Dokumente zu beschränken. Diese Form der Suchanfrage wurde bereits in [42] benutzt. Die letzte Form versucht diesen Effekt über die Hilfsterme *music*, *genre* und *style* zu erzielen. Nachdem das Ziel des Verfahrens (unter anderem) die Klassifikation in Genres ist, wird versucht, auf diese Weise Künstlerinformationen, die speziell in Richtung Stilbeschreibung gehen, zu erhalten. Außerdem kann durch Benutzung verschiedener Suchterme deren Einfluss auf das Verfahren sichtbar gemacht werden.

Bei jedem Suchvorgang werden die 100 relevantesten Seiten angefragt, um für die Experimente auf einen ausreichend großen Datenbestand zurückgreifen zu können. Durchschnittlich sind davon bei Google knapp 96 Webseiten verfügbar und bei Yahoo! 84 (siehe Tabelle 3.1). Auf die restlichen muss aufgrund von nicht erreichbaren Servern, unauffindbaren Dateien etc. verzichtet werden. In den Prozeß der Merkmalsgewinnung (siehe Kapitel 4) werden jedoch maximal 50 Seiten miteinbezogen. Dies stellt einen Unterschied zu [20] dar, wo maximal 50 Suchresulate angefragt wurden, wovon dann nur diejenigen verarbeitet wurden, die auch verfügbar waren (im Durchschnitt 40). Hier wird nun der Fall simuliert, dass solange Suchergebnisse abgefragt werden, bis 50 verwertbare Seiten gefunden wurden (sofern dazu genug Seiten existieren). Soweit von den Suchmaschinen unterstützt, werden Suchergebnisse, die auf Dateien in einem anderen Format als HTML verweisen, ausgeschlossen. Die erhaltenen Ergebnisse unterliegen keinerlei Beschränkung in Bezug auf Sprache oder Herkunft.

3.3 Filtermethoden

Das Verfahren beruht darauf, möglichst viele relevante Websites zu erhalten und miteinzubeziehen. Um in späteren Schritten besonders präzise Charakteristika für Künstler extrahieren zu können (siehe Kapitel 4), sollte mit Bedacht entschieden werden, welche der verfügbaren Websites tatsächlich als Grundlage für weitere Verarbeitungsschritte dienen sollen. Dies stellt eine Erweiterung des Verfahrens aus [20] dar.

Prinzipiell gibt es eine Unzahl an Ansätzen zur Bewerkstelligung einer Qualitätskontrolle. Im Rahmen dieser Arbeit können die verschiedenen Möglichkeiten keinesfalls erschöpfend behandelt werden, vielmehr sollen die hier präsentierten Methoden dazu dienen, einen Hinweis zu bekommen, ob die vergleichsweise simplen Verfahren die Qualität des in [20] vorgestellten Ansatzes verbessern können.

	Google	Yahoo!
„*Künstlername*“	95.6	82.8
„*Künstlername*“ music review	95.9	84.0
„*Künstlername*“ music genre style	95.9	85.6
Durchschnitt	95.8	84.1

Tabelle 3.1: Anzahl der verfügbaren Seiten für die verschiedenen Suchmaschinen und Suchanfragen (Durchschnitt über je 224 Anfragen mit 100 Suchresultaten). Es zeigt sich ein deutlicher Unterschied zwischen Google und Yahoo!, der am ehesten dadurch zu erklären ist, dass Yahoo! für manche Anfragen gar nicht in der Lage ist, 100 Seiten zu empfehlen (Extremfall: nur 14 Seiten für Künstler *Alpha Blondie*) und von den vorgeschlagenen einige nicht zugänglich sind. (Zeitpunkt der Abfragen: Anfang November 2004)

3.3.1 Webseiten-Profile

Diese einfachen Verfahren bestehen im Wesentlichen darin, dass einige der verfügbaren Websites davon ausgeschlossen werden, zur Charakterisierung des Künstlers beizutragen. Konkret werden einfache Regeln erstellt, die sicherstellen sollen, dass nur „gewünschte“ Seiten miteinbezogen werden. Um festzustellen, welche Seiten den aufgestellten Kriterien genügen, wird zuallererst für jede Seite ein Profil erstellt. Die zur Profilbildung der Websites erfassten Eigenschaften, orientieren sich stark an den Vorschlägen aus [6]:

- Der Inhalt der Website (`<body>`) enthält den Namen des Künstlers (Boole'sches Attribut)
- Der Inhalt der Website enthält mehr als die Hälfte der unterstützenden Suchterme der Anfrage (Boole'sches Attribut)
- In den ersten 100 Wörtern des Inhalts der Website ist der Name des Künstlers enthalten (Boole'sches Attribut)
- In den ersten 100 Wörtern des Inhalts der Website sind mehr als die Hälfte der unterstützenden Suchterme der Anfrage enthalten (Boole'sches Attribut)
- Der Titel der Website (`<title>`) enthält den Namen des Künstlers (Boole'sches Attribut)
- Der Titel der Website enthält einen unterstützenden Suchterm der Anfrage (Boole'sches Attribut)

- Die URL der Website enthält den Namen des Künstlers (Boole'sches Attribut)
- Die URL der Website enthält einen unterstützenden Suchterm der Anfrage (Boole'sches Attribut)
- Anzahl der Worte auf der Website (Numerisches Attribut)
- Größe der Datei in Bytes (Numerisches Attribut)

Dabei muss noch erwähnt werden, dass sich „Inhalt“ hierbei immer auf den reinen Textinhalt der Webseite bezieht (vgl. Abschnitt 4.1), also nicht auf Informationen, die sich innerhalb der HTML-Tags befinden. Auch die Größe der Datei bezieht sich auf den reinen Text und gibt somit die Anzahl der Buchstaben an. Weiters soll noch erläutert werden, warum nach mehr als der Hälfte der Suchterme innerhalb des Textes gesucht wird. Für die Anfrage mit den Constraints *music* und *review* bedeutet dies, dass beide Terme vorkommen müssen, für jene mit *music*, *genre* und *style* hat dies zur Folge, dass nur zwei von drei Termen vorhanden sein müssen. Dies ergibt sich aus der Tatsache, dass je mehr Suchbegriffe bei einer Suchmaschine angegeben werden, die Wahrscheinlichkeit, dass die Resultate tatsächlich alle Begriffe enthalten, immer geringer wird. Konkret kann sogar festgestellt werden, dass (im Reintext) auf keiner der empfangenen Seiten, die mithilfe dieser drei Zusatzbedingungen gesucht wurden, alle drei Begriffe auftauchen. Durch die gewählte Vorgangsweise können somit gewisse Ansprüche an die Qualität der Seiten gestellt werden, ohne dabei durch unrealistische Forderungen die Bewertungen der Seiten niedrig anzusetzen. Für jene Seiten, die über die Suchanfrage ohne weitere Zusatzbedingungen gewonnen wurden, gelten alle Profilanforderungen, die sich auf die Suchterme beziehen, als erfüllt.

Anhand der Merkmale der Website-Profile können nun einfach Regeln erstellt werden, die angeben, ob eine Seite miteinbezogen werden soll oder nicht. Einige (unterschiedlich restriktive) Filtermöglichkeiten sollen im Rahmen dieser Arbeit eingesetzt werden und ihre Auswirkungen auf die Gesamtqualität des Systems evaluiert werden.

3.3.2 Eingesetzte Filter

Um eine einheitliche Nomenklatur einzuführen und in Folge kurz und eindeutig darauf Bezug nehmen zu können, werden die verschiedenen Filter mit F und einem Index bezeichnet. Als F_0 wird jener Ansatz bezeichnet, bei dem kein Filter zum Einsatz kommt, also einfach die Verwendung maximal der ersten 50 verfügbaren Suchergebnisse. Somit werden dabei auch keine Website-Profile ausgewertet.

F_1 bezeichnet einen relativ simplen und toleranten Filter, der nur Seiten zulässt, die

- den Namen des Künstlers enthalten,
- zwischen 20 und 10000 Wörter enthalten und
- deren Dateigröße zwischen 100 Byte und 200 kByte[2] beträgt.

Es mag zunächst nicht ersichtlich sein, warum durch das Kriterium „Seite muss den Namen des Künstlers enthalten“ Seiten ausgefiltert werden können, da ja bereits durch die Suchanfrage sichergestellt sein sollte, dass nur Seiten, die den Künstlernamen enthalten, in der Ergebnismenge enthalten sind. In der Praxis zeigt sich allerdings, dass sich darunter sehr wohl Seiten befinden, in deren Reintext-Repräsentation der Name des Künstlers nicht aufscheint. Dies ist einerseits bei Anfragen der Fall, die nur wenige Resultate zurückliefern, da die Suchterme (und damit auch der Künstlername) bei geringer Trefferanzahl auch ausgelassen werden können, um überhaupt Resultate präsentieren zu können. So ist es durchaus möglich, Seiten als Antwort zu bekommen, die zwar die zusätzlichen Hilfsterme enthalten, aber mit dem Künstler eigentlich nichts zu tun haben.

Aber auch bei Künstlern, die ausreichend mit relevanten Websites verknüpft sind, treten Seiten ohne dazugehörigen Namen auf. Der Hauptgrund dafür ist, dass viele Seiten die Information, welchen Inhalts sie sind, im Meta-Tag des HTML-Headers enthalten. Suchmaschinen werten diese Information aus, das Verfahren hier allerdings nicht. Das Ignorieren dieser Information lässt sich dadurch argumentieren, dass Meta-Tags meistens vom Programmierer der Website dazu mißbraucht werden, vermeintlich populäre Schlagworte zu integrieren, von denen erwartet wird, dass sie häufig Teil von Suchanfragen sind, sodass diese Websites öfters als relevante Suchresultate in Suchmaschinen angezeigt werden und dadurch auch öfter aufgerufen werden. Als Konsequenz finden sich oft Seiten, deren Meta-Tag aus ausgesprochen umfangreichen Wortlisten besteht, die mehr oder weniger jedes populäre Themenfeld umfassen. Zur Charakterisierung von Musikkünstlern sind aber nur Seiten von Interesse, deren Inhalt sich tatsächlich mit dem Künstler beschäftigt.

Ein anderer Grund für das Fehlen von Künstlernamen kann auch eine ausgefallene Schreibweise des Namens sein. Bespielsweise kann man Seiten der Band *Slayer* finden, in deren Titel der Name mit Abständen geschrieben

[2]Umrechnungsfaktor: 1 kByte $\equiv$ 1000 Byte

ist ($s\ l\ a\ y\ e\ r$). Offensichtlich sind die ausgefeilten Mechanismen der Suchmaschinen in der Lage, solche Muster zu erkennen, das einfache Verfahren, das dem Ansatz hier zugrunde liegt, behandelt solche Fälle nicht.

Die Beschränkungen in puncto Dateigröße und Wortanzahl sollen einen Schutz vor dem Einfluss völlig atypischer Webseiten bieten. Seiten, die kaum Wörter enthalten, werden in der Regel auch nicht viel Information beinhalten, Seiten, die deutlich länger als der Durchschnitt sind, behandeln unter Umständen sehr viele Themen bzw. einen sehr weit gefassten Themenkreis, was der Qualität und Präzision der daraus extrahierten Features nicht gerade zuträglich ist. Deutlich mehr Anforderungen an eine Website stellt Filter F_2. F_2 erlaubt nur Seiten, die

- den Künstlernamen enthalten,
- die zusätzlichen Suchterme enthalten,
- innerhalb der ersten 100 Wörter entweder den Künstlernamen oder die zusätzlichen Suchterme enthalten,
- entweder den Namen oder die Terme im Titel oder in der URL enthalten,
- zwischen 60 und 7500 Wörter enthalten und
- deren Dateigröße zwischen 550 Byte und 150 kByte beträgt.

Im Vergleich zu F_1 schränkt F_2 den Größenbereich weiter ein und fordert bereits die Existenz der zusätzlichen Suchterme. Weiters werden auch Anforderungen an die Position der Terme gestellt. F_3 bezeichnet den restriktivsten Filter unter den hier vorgestellten. F_3 akzeptiert nur Seiten, die

- den Künstlernamen im Titel enthalten,
- den Künstlernamen in den ersten 100 Worten enthalten,
- die zusätzlichen Suchterme in den ersten 100 Worten enthalten,
- einen Suchterm im Titel oder Namen oder Suchterm in der URL enthalten,
- zwischen 100 und 5000 Wörter enthalten und
- deren Dateigröße zwischen 1 kByte und 100 kByte beträgt.

Diese Filter stellen nur eine beliebige Auswahl dar und sollen dazu dienen, die prinzipielle Anwendbarkeit und Sinnhaftigkeit dieser Vorverarbeitung zu demonstrieren und aufzuzeigen, welche Formen der Restriktion sich als nützlich und welche als nutzlos, bzw. sogar qualitätsbeeinträchtigend erweisen können.

In den Tabellen 3.2, 3.3 und 3.4 finden sich Angaben über die verwertbaren Seiten nach Anwendung der Filter. Deutlich zeigt sich, dass für jene Anfrage, in der neben dem Namen des Künstlers keine weiteren Suchterme zur Verwendung gekommen sind (Tabelle 3.2), im Durchschnitt so gut wie keine Probleme mit der Anzahl der verwertbaren Seiten auftreten. Am Minimum kann man aber erkennen, dass die Einschränkungen der Filter teilweise starke Selektion mit sich bringen, sodass für manche Künstler kaum Seiten übrigbleiben, die den Filtern genügen. Dass der Durchschnittswert trotzdem auf einem vergleichsweise hohen Niveau bleibt, hängt damit zusammen, dass einige der Einschränkungen der restriktiveren Filter in diesem Fall keinerlei Bedeutung haben, da sich diese auf die Vorkommnisse der Suchterme auf der Seite beziehen und hier keine weiteren Suchterme zum Einsatz kommen (siehe Abschnitt 3.3.1).

Deutlich zeigt sich der Abfall der durchschnittlichen Anzahl der Seiten bei Filter F_3. Offensichtlich sind die Kriterien dieses Filters zu restriktiv, um in der Praxis eine ausreichende Anzahl von Webseiten zuzulassen. Mit durchschnittlichen Werten von ca. 5 Webseiten ist die Grundidee, nämlich eine möglichst breite und robuste Datenbasis, die die gesellschaftliche Wahrnehmung eines Künstlers repräsentiert, zu gewinnen, nicht verwirklichbar. Dazu kommt noch, dass für viele der Künstler überhaupt keine Seiten existieren, die die Kriterien des Filters F_3 erfüllen. Für manche Künstler existieren allerdings zumindest 50 solcher Seiten (ausgenommen für F_3 und die Terme *music* und *review*). Das resultierende Ungleichgewicht ist im Sinne möglichst gleicher Voraussetzungen für alle Künstler nicht wünschenswert. Abgesehen davon ist es auch nicht möglich, ohne zugrunde liegende Webseiten Charakteristika zu extrahieren.

	Google				Yahoo!			
	F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
Durchschnitt	50.00	49.93	49.93	44.15	50.00	49.67	49.60	34.36
min.	50	34	34	4	50	11	11	0
max.	50	50	50	50	50	50	50	50

Tabelle 3.2: Anzahl der verwertbaren Seiten der Suchanfrage `"`*`Künstlername`*`"` nach Anwendung der Filter. (Durchschnittsbildung über 224 Anfragen (siehe Anhang A.1), maximal 100 verfügbare Seiten, maximal 50 Seiten verarbeitet)

	Google				Yahoo!			
	F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
Durchschnitt	49.94	49.84	37.43	6.61	49.86	49.67	31.57	5.15
min.	36	32	0	0	18	13	0	0
max.	50	50	50	24	50	50	50	27

Tabelle 3.3: Anzahl der verwertbaren Seiten der Suchanfrage `"`*`Künstlername`*`" music review` nach Anwendung der Filter. (Durchschnittsbildung über 224 Anfragen (siehe Anhang A.1), maximal 100 verfügbare Seiten, maximal 50 Seiten verarbeitet)

	Google				Yahoo!			
	F_0	F_1	F_2	F_3	F_0	F_1	F_2	F_3
Durchschnitt	49.99	49.77	48.42	5.46	49.83	49.69	42.86	4.24
min.	49	22	11	0	13	10	3	0
max.	50	50	50	50	50	50	50	50

Tabelle 3.4: Anzahl der verwertbaren Seiten der Suchanfrage `"`*`Künstlername`*`" music genre style` nach Anwendung der Filter. (Durchschnittsbildung über 224 Anfragen (siehe Anhang A.1), maximal 100 verfügbare Seiten, maximal 50 Seiten verarbeitet)

4 Merkmalsextraktion

Bei der Merkmalsextraktion geht es nun darum, aus den erhaltenen Daten charakteristische Beschreibungen zu gewinnen, die, im Sinne des Verfahrens, effizient verarbeitbar sind. Die Repräsentationsform eines Problems hat enorme Auswirkung auf das eingesetzte Lernverfahren und die Fähigkeit zur Generalisierung, die erst das Wesen des Lerners ausmacht. Die Dokumente werden typischerweise in eine Attribut-Wert Repräsentation transformiert. Die eingesetzten Verfahren sind im Wesentlichen Standard-Verfahren im Information Retrieval, speziell im Bereich *Text Categorization*. Sebastiani beschreibt Text Categorization als den Vorgang des automatischen Erstellens (durch Methoden des Machine Learning) von automatischen Text-Klassifikatoren, d.h. von Programmen, die in der Lage sind, natürlichsprachige Texte einer Domäne D mit thematischen Kategorien aus einer vordefinierten Menge $C = \{c_1, ..., c_{|C|}\}$ zu kennzeichnen [36]. In diesem Fall sind die thematischen Kategorien Musikgenres.

Das hier vorgestellte Verfahren unterscheidet sich von anderen Text Categorization Aufgaben dadurch, dass nicht einzelne Texte thematischen Kategorien zugeordnet werden, sondern allgemein ein dynamischer Korpus an, nicht notwendigerweise im Sinne der Sprache wohlgeformten, Texten dazu dient, ein reales Konzept (konkret einen Musikkünstler) zu beschreiben. Die Textbeschreibungen der Musikkünstler werden dann vom Klassifikator einer Kategorie zugewiesen, daher kann man hier eigentlich nicht direkt von einem Text-Klassifikator sprechen, sondern vielmehr von einem Musikkünstler-Klassifikator, der sich Methoden von Text-Klassifikatoren bedient.

Nach Debole und Sebastiani ([14]) besteht die Konstruktion eines Text-Klassifikators im Wesentlichen aus zwei Phasen:

1. einer Phase der *Dokumentenindizierung*, d.h. des Erstellens von internen Repräsentationen für die Dokumente. Diese besteht typischerweise wiederum aus:

   (a) einer Phase der Termauswahl (*term selection*), d.h. einer Form der Dimensionalitätsreduktion, die darin besteht, eine Auswahl aus allen vorkommenden Termen zu wählen, so dass, wenn diese Auswahl als Repräsentation des Dokuments dient, davon erwartet wird, dass die besten Resultate, oder der beste Kompromiss zwischen Resultat und Effizienz, erzielt werden und

(b) einer Phase der Termgewichtung (*term weighting*), in welcher für jeden Term t_k, der in Phase 1a gewählt wurde, und für jedes Dokument d_j, ein Gewicht $0 \leq w_{kj} \leq 1$ berechnet wird, das, umgangssprachlich, aussagt, wie sehr Term t_k dazu beträgt, dass Dokument d_j anders ist als andere.

2. einer Phase der *Klassifikator-Initiierung*, d.h. der Herstellung des Klassifikators durch Lernen aus den internen Repräsentationen der Trainingsdokumente.

Auf Phase 2 wird in den Kapiteln 5 und 6 genauer eingegangen.

4.1 Vorverarbeitung

Aus den verwendeten 50 Seiten werden alle HTML Markup Tags entfernt, so dass nur der reine Textinhalt verarbeitet wird. Dazu werden alle Ausdrücke, die sich in spitzen Klammern befinden (zwischen `<` und `>`) gelöscht. Außerdem muss darauf geachtet werden, dass oftmals auch Text zwischen Tags nicht zum Inhalt gehört, wie bei `<style>` und `<script>`. Danach werden Sonderzeichen aus der HTML-Kodierung in die entsprechenden Zeichen umgewandelt (z.B. wird `&` durch `&` ersetzt) und einige weitere Ersetzungen zur Bereinigung und Vereinfachung des Materials durchgeführt, bevor der komplette Inhalt in Kleinbuchstaben konvertiert wird.

Weiters werden sehr häufige, allgemeine Worte, im Englischen spricht man von so genannten *Stopwords*, ignoriert. Stopwords sind die gebräuchlichsten Worte einer Sprache und da sie durch ihre Häufigkeit so gut wie keine Information tragen, stellen sie in der Verarbeitung nur Ballast dar, der bedenkenlos entfernt werden kann. Dieser Schritt kann bereits als Termauswahlschritt gesehen werden. Typische Beispiele für Stopwords sind im Englischen *a*, *and*, *or*, oder auch *the*, im Deutschen handelt es sich um Worte (Terme) wie *der*, *die*, *das*, *ein*, *eine*, *und* etc. Nachdem die Menge der Stopwords endlich ist, kann über bestehende Listen einfach ermittelt werden, ob ein Wort ein Stopword ist. Im konkreten Fall kommt eine Kombination von mehreren Listen zum Einsatz, darunter auch Terme aus den Sprachbeschreibungen von im Internet häufig eingesetzten Programmiersprachen, um auch Worte ausfiltern zu können, die möglicherweise aus übriggebliebenen Skriptfragmenten von fehlerhaft formulierten Webseiten stammen. Die vollständigen Listen samt Quellenangabe finden sich im Anhang A.3. Keine Berücksichtigung finden auch Terme, die nur aus einem oder aus mehr als 20 Buchstaben bestehen, und Zahlen.

4.2 Termauswahl

Aus einem Web-Crawl für einen Satz von ca. 200 Künstlern kann eine Liste von über 200000 verschiedenen Termen resultieren. Der Großteil davon sind Terme, die meist nur ein einziges Mal auftreten wie Tippfehler oder andere irrelevante Terme. Deshalb werden alle Terme, die nicht in zumindest 5 der 50 zum Künstler gehörenden Seiten auftreten, gelöscht. Das Ergebnis dieser Aktion ist, dass typischerweise zwischen 3000 und 10000 verschiedene Terme übrigbleiben, was die Weiterverarbeitung erheblich vereinfacht und beschleunigt. An dieser Stelle soll auch auf einen der Hauptunterschiede zu bestehenden Ansätzen wie [42, 6] hingewiesen werden, der darin besteht, dass hier keine Extraktion von n-grammen oder PoS-Tagging durchgeführt wird, um besonders „bedeutungstragende" Terme ausfindig zu machen. Stattdessen wird hier jeder auftretende Term miteinbezogen und gleichwertig behandelt, sofern er nicht durch Vorverarbeitungsschritte entfernt wurde. Außerdem wird hier kein Stemming, also die Reduktion von Worten auf ihren Wortstamm, vorgenommen, wie dies häufig in Information Retrieval Anwendungen der Fall ist.

Aus statistischer Sicht ist es problematisch, ein Klassifikations-Modell nur mit Hilfe von wenigen Trainingsbeispielen, die durch mehrere tausend Dimensionen beschrieben werden, zu lernen. Um diesem Umstand gerecht zu werden, muss versucht werden, die Anzahl an Termen (d.h. die Anzahl an Dimensionen) so weit wie möglich zu reduzieren. Deshalb kommen term selection Techniken zum Einsatz, um jene Untermenge von Termen auszuwählen, die als am nützlichsten zur kompakten Darstellung der Semantik der Dokumente erachtet wird. Dies wird normalerweise dadurch erreicht, dass jeder Term anhand einer *term evaluation function* (TEF) bewertet wird und dann jene Terme gewählt werden, für die diese TEF den höchsten Wert aufweist. Term selection-Verfahren dienen dazu, *overfitting*, also das Phänomen, dass Klassifikatoren dazu neigen, jene Daten mit denen sie trainiert wurden besser zu klassifizieren als andere, zu vermeiden.

Ein Ansatz, der oft als TEF zum Einsatz kommt, ist die Verwendung des Chi-Quadrat-Tests (χ^2-Test). Der χ^2-Test ist ein statistisches Verfahren und wird in vielen Bereichen, vor allem in den Experimentalwissenschaften, eingesetzt, um zu messen, wie sehr sich konkrete Beobachtungen von dem unterscheiden, was ursprünglich aufgrund einer Hypothese vermutet wurde (bzw. wie unabhängig die Beobachtungen von der Hypothese sind). Im Falle der Text Categorization ist die Null-Hypothese H_0, dass das Auftreten eines bestimmten Terms von der Kategorie, in der er auftritt, unabhängig ist. Die Alternativhypothese H_1 lautet somit, dass das Auftreten eines Terms von der Kategorie abhängig ist. Die Kategorien sind im konkreten Fall, wie oben bereits erwähnt, Musikgenres, zu denen typische Künstler bekannt sind. Somit verlagert sich der Vorgang der Termauswahl von der Künstler- auf die Genreebene.

Der χ^2-Test gibt nun einen Wahrscheinlichkeitswert an, der aussagt, mit welcher Sicherheit die H_0 für einen Term und eine Kategorie verworfen werden kann. Nachdem das Interesse auf Termen liegt, die möglichst stark mit den Kategorien verbunden sind, werden jene Terme ausgewählt, die den höchsten Wahrscheinlichkeitswert aufweisen, da für diese mit der höchsten Wahrscheinlichkeit gesagt werden kann, dass die Unabhängigkeit zwischen Term und Kategorie nicht gegeben ist. Diese Terme sind jene, die das Genre am besten beschreiben und am besten geeignet sind, es von anderen Genres abzugrenzen. Typischerweise handelt es sich dabei um solche Terme wie die einzelnen Teile der Namen der Bands oder Künstler, die zur Definition des Genres herangezogen wurden, der Namen von ähnlichen Künstlern, Wörter, die in Song- oder Albentitel auftauchen, sowie auch allgemein mit dem Genre assoziierbare Begriffe. Im Anhang A.2 befinden sich exemplarische Termlisten, wie sie durch ein Experiment, das den χ^2-Test einsetzt, entstehen.

In [14] wird die Formel zur Berechnung probabilistisch formuliert. Wahrscheinlichkeiten werden auf einem Ereignisraum aus Dokumenten interpretiert und per Maximum-Likelihood geschätzt, also durch Auszählen der Auftreten von Termen im Trainingsset. Die Notation $P(\bar{t}, c)$ bezeichnet die Wahrscheinlichkeit, dass für ein beliebiges Dokument x, Term t nicht in x vorkommt und x zu Kategorie c gehört.

$$\chi^2_{tc} = \frac{\left[P(t,c)P(\bar{t},\bar{c}) - P(t,\bar{c})P(\bar{t},c)\right]^2}{P(t)P(\bar{t})P(c)P(\bar{c})} \tag{4.1}$$

Für automatische Berechnungen durch den Computer empfiehlt sich die Verwendung einer vereinfachten Form, die nicht auf Wahrscheinlichkeiten operiert, sondern auf Häufigkeitswerten. Deshalb kommt in der Praxis oft (z.B. [47]) eine vereinfachte Version zum Einsatz:

$$\chi^2_{tc} = \frac{N(AD - BC)^2}{(A+B)(A+C)(B+D)(C+D)} \tag{4.2}$$

wobei A die Anzahl der Dokumente aus Kategorie c, die Term t enthalten, B die Anzahl der Dokumente nicht aus Kategorie c, die Term t enthalten, C die Anzahl der Dokumente aus Kategorie c, die Term t nicht enthalten, D die Anzahl der Dokumente nicht aus Kategorie c, die Term t nicht enthalten und N die Gesamtanzahl an erhaltenen Dokumenten bezeichnet. Da N für alle Terme gleich ist, kann es ignoriert werden.

Laut Yang und Pedersen [47] ist einer der Vorteile des χ^2-Tests, dass χ^2 ein normalisierter Wert, also ein Wert im Intervall $[0,1]$, ist und somit Terme innerhalb einer Kategorie verglichen werden können. Allerdings ist diese Normalisierung nicht aufrecht zu erhalten, wenn ein Eintrag in der Häufigkeitstabelle unterbesetzt ist, also ein Term nicht so häufig auftritt. Deshalb ist die χ^2-Statistik dafür bekannt, für niedrigfrequente Terme nicht zuverlässig zu sein.

Mit Hilfe der χ^2_{tc}-Werte für jeden Term in jeder Kategorie gibt es verschiedene Strategien, um eine globale Liste von Termen auszuwählen, anhand derer alle Dokumente beschrieben werden. Ein einfacher Ansatz ist es, alle Terme, die die größte Summe oder den maximalen Wert über alle Kategorien aufweisen, auszuwählen, d.h. entweder Terme, die für alle Kategorien gut abschneiden oder jene, die besonders für eine Kategorie gut funktionieren, zu benutzen.

In dieser Arbeit kommt ein Ansatz zur Anwendung, in dem die n höchsten Terme für jede Kategorie selektiert und in einer globalen Liste vereinigt werden. Die besten Resultate werden erzielt, wenn dazu die besten 100 Terme pro Kategorie ausgewählt werden (siehe Kapitel 6). Wenn, wie in Kapitel 6 14 Genres zum Einsatz kommen, kann das in einer globalen Termliste von bis zu 14×100 Termen resultieren, sofern es keine Überlappungen in den Listen der verschiedenen Kategorien gibt.

Wie bereits in [20] bezeichnet die Notation C_n die Strategie, n Terme pro Kategorie auszuwählen (Cut-off-Index n). Im Fall von C_∞ werden keinerlei Terme anhand von χ^2_{tc}-Werten entfernt, was vor allem für die Benutzung der Merkmale in Ähnlichkeitsmaßen von besonderem Interesse ist, da in diesen Fällen die Genrezugehörigkeit eines Künstlers nicht bekannt ist. An dieser Stelle soll explizit darauf hingewiesen werden, dass das vorgestellte Verfahren zur Termselektion in Form des χ^2-Tests nur in Genre-Klassifikationsszenarien angewandt werden kann, wo für den Satz an Trainingsdaten die Zuordnung der Künstler zu Genres bekannt ist. Für andere Einsatzgebiete, wie Ähnlichkeitsberechnung ist diese Methode der Dimensionsreduktion nicht durchführbar, weshalb hier ein höherdimensionaler Raum benutzt werden muss, was einige erschwerende Umstände mit sich bringt (komplexere Berechnungen, höhere Anfälligkeit für overfitting etc.).

Ein anderer Ansatz zur Auswahl von Termen wäre eine lokale Strategie. Anstatt eine globale Termliste auszuwählen, wird für jede Kategorie eine individuelle Menge definiert und ein eigener Klassifikator, basierend auf der dadurch festgelegten Repräsentation, trainiert. Dieser Ansatz ist allerdings nicht so einfach auf Ähnlichkeitsmaße erweiterbar, für die ein globaler Vektorraum benötigt wird, weshalb er hier nicht weiter verfolgt wird.

4.3 Termgewichtung

Für die ausgewählten Terme wird im nächsten Schritt eine Gewichtung vorgenommen, die einen Indikator für die Wichtigkeit und den Beitrag eines Terms zur Beschreibung des Künstlers darstellen soll.

Es existiert eine Reihe von verschiedenen Gewichtungsmethoden. Allen gemein sind drei *Monotonie-Annahmen*, wie in [14] unter Bezugnahme auf [49] festgehalten wird:

1. Mehrfache Auftreten eines Terms in einem Dokument sind nicht weniger wichtig, als einzelne Auftreten (*TF-Annahme*), d.h. je öfter ein Term auftritt, um so mehr kann man davon ausgehen, das dieser Term eine wichtige Rolle spielt.

2. Seltene Terme sind nicht weniger wichtig als häufige (*IDF-Annahme*), d.h. Terme, die in sehr vielen Dokumenten auftreten, tragen geringe Information und sind daher als nicht so wichtig anzusehen.

3. Für die selbe Anzahl an Termübereinstimmungen sind lange Dokumente nicht wichtiger als kurze (*Normalisierungsannahme*).

Die ersten beiden Annahmen finden ihren Ausdruck in der *term frequency* $\times$ *inverse document frequency* ($tf \times idf$) Funktion. Um diese anwenden zu können, wird zuerst für jeden Künstler a und jeden Term t die Anzahl der Auftreten f_{ta} (*term frequency*) des Terms t in Dokumenten, die mit a assoziiert sind, und die Anzahl der Seiten df_t (document frequency), auf denen der Term aufgetreten ist, berechnet. Daraus wird der $tf \times idf$-Wert (hier in der *ltc*-Variante [35]) wie folgt berechnet:

$$tfidf_{ta} = tf_{ta} \log_2 \frac{N}{df_t} \tag{4.3}$$

wobei N die Gesamtanzahl der empfangenen Seiten bezeichnet und

$$tf_{ta} = \begin{cases} 1 + \log_2 f_{ta}, & \text{wenn } f_{ta} > 0, \\ 0 & \text{sonst.} \end{cases} \tag{4.4}$$

Die tf_{ta}-Komponente von Formel 4.3 unterstützt die TF-Annahme, während die $\log_2 \frac{N}{df_t}$-Komponente die IDF-Annahme unterstützt.

Nach Termauswahl und Termgewichtung wird jeder Künstler durch einen Vektor von Termgewichten T beschrieben. Danach werden die Gewichte normalisiert, so dass die Länge des Vektors 1 beträgt ($|T| = 1$). Dies erfüllt die dritte Monotonie-Annahme. Dadurch wird der Einfluss, den die Länge des

Dokuments andernfalls hätte, unterbunden, da lange Dokumente im Allgemeinen dazu neigen, die selben Worte immer und immer wieder zu wiederholen, was sich in höheren term frequencies niederschlagen würde. Bewerkstelligt wird dies durch *Kosinus Normalisierung*:

$$w_{ta} = \frac{tfidf_{ta}}{\sqrt{\sum_{u \in T} tfidf_{ua}^2}} \tag{4.5}$$

Die Normalisierung macht einen sinnvollen Vergleich zweier Vektoren erst möglich.

5 Klassifikation

Zur Klassifikation der Musikkünstler kommen in dieser Arbeit hauptsächlich Support Vector Machines (SVMs) zum Einsatz. Außerdem wird der k-nearest neighbor (k-NN) Klassifikator eingesetzt, um die Qualität der für Ähnlichkeitsmaße benutzten Merkmale zu evaluieren. Beide Klassifikatoren sind *überwachte* Machine Learning Verfahren, d.h. der Lerner wird üblicherweise in einer Trainingsphase mit Beispielen konfrontiert, von denen die Klassenzugehörigkeit bekannt ist. Dadurch wird automatisch ein Modell erstellt (Klassifikator), das in der Lage ist, zu neuen, unbekannten Beispielen eine Klassenzugehörigkeit vorauszusagen.

5.1 Support Vector Machines

Support Vector Machines (SVMs) sind ein vergleichsweise neues Lernverfahren, das von Vladimir Vapnik vorgestellt wurde [40, 41]. SVMs sind in der Computational Learning Theory verankert und eignen sich besonders zur effizienten Lösung hochdimensionaler Klassifikationsprobleme. Die Verwendung von SVMs auf dem Gebiet der Text Categorization wurde erstmals von Thorsten Joachims vorgeschlagen [19]. Joachims begründet in seiner Arbeit seine Vorgangsweise durch theoretische Überlegungen zu den Eigenschaften von Texten:

- *Hochdimensionaler Merkmalsraum:* Beim Lernen von Text-Klassifikatoren ergibt sich die Notwendigkeit, sehr viele Features verarbeiten zu müssen (typischerweise über 10000).

- *Kaum irrelevante Features:* Durch die Annahme, dass die meisten dieser Features bedeutungslos für die Semantik des Textes sind, kann man die Dimensionalität drastisch reduzieren. Dies wird klassischerweise durch Termauswahl bewerkstelligt (vgl. Abschnitt 4.2). Mit Hilfe eines Experiments zeigt Joachims aber, dass auch Klassifikationen nur mit den als am unwichtigsten erachteten Features deutlich besser verlaufen, als zufällige Klassifikationen. Es zeigt sich also, dass auch die „schlechtesten“ Features immer noch Information tragen und daher auch zu einem gewissen Grad als relevant angesehen werden müssen. Dies lässt vermuten, dass ein geeigneter Klassifikator in der Lage sein sollte, viele

Features zu kombinieren, man spricht davon, dass er in der Lage sein sollte, ein „dichtes“ Konzept zu lernen.

- *Feature Vektoren für Dokumente sind nur spärlich besetzt:* Für jedes Dokument gilt, dass der Vektor der (globalen) Features nur wenige Einträge besitzt, die nicht 0 sind (*Sparsity*).

- *Die meisten Text Categorization Probleme sind linear trennbar:* Viele der bekannten Textkorpora beinhalten Kategorien, die linear trennbar sind. Die auftretenden Probleme sind eher auf zweifelhafte Inhalte oder offensichtliche Fehlklassifikationen der Personen, die den Korpus annotiert haben, zurückzuführen.

Aufgrund dieser Argumente ist der Einsatz von SVMs sehr erfolgversprechend, da deren Stärken in diesen Bereichen liegen. Wie bereits erwähnt, können SVMs sehr effizient mit hochdimensionalen Merkmalsräumen umgehen. Damit verliert auch der Einsatz von Termselektionsverfahren an Bedeutung, auch weil SVMs selbst bereits robust gegenüber overfitting sind. Weiters wird auf Arbeiten verwiesen, die theoretisch und empirisch belegen, dass ähnlich gelagerte Verfahren auf Probleme, die dichte Konzepte und sparse Instanzen aufweisen, erfolgreich anwendbar sind. Zuletzt wird darauf hingewiesen, dass es eine Grundidee der SVMs ist, lineare Diskriminatoren zu finden, um die Klassen voneinander zu trennen, womit alle Aspekte von Text Categorization Aufgaben abgedeckt sind. Ein weiterer Vorteil ist, dass keine Anstrengungen unternommen werden müssen, um Parameter zu optimieren, da durch die SVMs automatisch eine standard-Auswahl getroffen wird, was sich als am effektivsten erwiesen hat.

Zum Verlust der Bedeutung von Termauswahlverfahren muss allerdings angemerkt werden, dass die Anwendung des χ^2-Tests für das hier präsentierte Verfahren trotzdem wichtig ist, da dadurch viele für die Beschreibung der Kategorien tatsächlich unbedeutsame Terme eliminiert werden. Die von Webseiten gewonnenen Daten weisen viele Terme auf, die mit dem „Inhalt“ der Seiten nichts zu tun haben. Der χ^2-Test in Kombination mit der Gewichtung durch $tf \times idf$ trägt dazu bei, die Daten von diesem Rauschen zu befreien und die Genauigkeit des Verfahrens zu verbessern.

Die wesentliche Funktionsweise der Support Vector Machines beruht darauf, durch eine Kernel-Transformation das Problem in einen Raum zu verlagern, in dem dieses linear trennbar ist. Durch Wahl des richtigen Kernels können mit SVMs auch Klassifikatoren für polynomiell, durch Radial-Basis-Functions (RBF) und 2-Ebenen-Perceptrons trennbare Probleme gefunden werden. Wie oben bereits erwähnt, sind die meisten Text Categorization-Probleme linear trennbar, weshalb für diese Aufgaben ein linearer Kernel zum Einsatz kommt. Das Erlernen eines SVM-Klassifikators besteht nun darin, unter all den möglichen Trennebenen ($(n-1)$-dimensionale Hyperebenen im n-dimensionalen Raum) zwischen den Klassen jene zu finden, die die Klassen mit dem weitest möglichen Abstand voneinander trennen (*Largest margin classifier*). Dabei wird jene Ebene gewählt, die Äquidistanz zu den beiden am weitesten voneinander entfernten parallelen Ebenen, die die Beispiele von einander trennen, aufweist. Diese „optimale“ Entscheidungsebene wird alleine von jenen Trainingsbeispielen, die sich „am Rand“ ihrer Kategorie, also am nächsten zu Beispielen aus anderen Kategorien, befinden, bestimmt. Jene Trainingsbeispiele werden *Support Vectors* genannt. Dies soll durch eine Illustration für 2 Klassen im 2-dimensionalen Raum verdeutlicht werden (Abbildung 5.1).

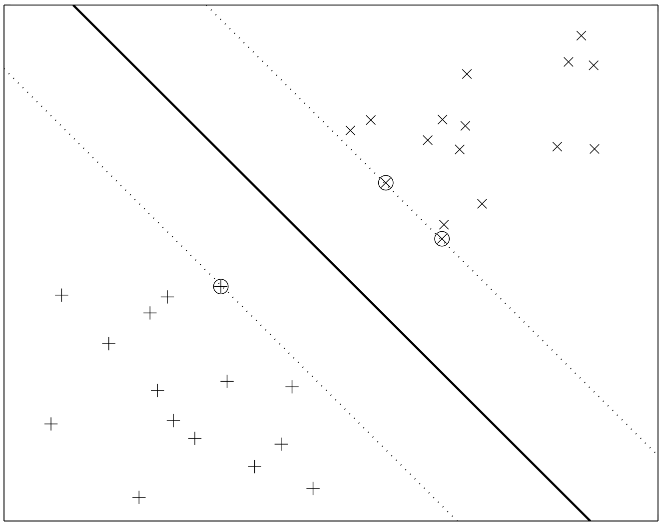

Abbildung 5.1: Als Entscheidungsgrenze zwischen den Klassen + und × wird von SVMs jene Hyperebene gewählt, die die Trainingsbeispiele der beiden Klassen mit maximalem Abstand trennt. Die Trainingsbeispiele, die sich am nächsten zur Hyperebene befinden, sind die *Support Vectors* (markiert durch Kreise).

Für die Durchführung der Experimente kommt der lineare Kernel der Bibliothek LIBSVM (Version 2.33)[1] in der Implementierung der Matlab OSU Toolbox (Version 3.00)[2] zum Einsatz.

[1]http://www.csie.ntu.edu.tw/~cjlin/libsvm
[2]http://www.ece.osu.edu/~maj/osu_svm

5.2 k-nearest neighbor Klassifikator

K-nearest neighbor (k-NN) ist ein sehr einfaches, aber sehr häufig eingesetztes Verfahren zur Klassifikation. In der erstmaligen Beschreibung dieses Verfahrens weisen Cover und Hart ([12]) nach, dass mit k-NN die Wahrscheinlichkeit eines Klassifikationsfehlers nach oben mit der 2-fachen Wahrscheinlichkeit eines Fehlers des Bayes-optimalen Klassifikators beschränkt ist. Das Grundprinzip hinter k-NN ist simpel, zu einer zu klassifizierenden Instanz wird das ähnlichste Trainingsbeispiel (wenn $k > 1$, die k ähnlichsten) gesucht und die Klasse dieses Beispiels vorausgesagt. k-NN ist ein Instance-based Learner, also ein Lernverfahren, dass seine Klassifikationen direkt auf Basis aller Trainingsbeispiele durchführt. Die Lernphase besteht daher nur aus dem Sammeln aller Trainingsbeispiele, der Prozess der Generalisierung der Beispiele wird solange verschoben, bis ein konkretes Beispiel klassifiziert werden soll. Diese Form von Lernalgorithmen wird auch als *lazy learner* bezeichnet.

Die Ähnlichkeitsberechnung erfolgt im einfachsten Fall, so auch hier, durch Berechnung der euklidischen Distanz zwischen zwei Bespielen:

$$dist(a,b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \tag{5.1}$$

wobei a und b Featurevektoren sind, n die Länge der Vektoren und die Notation a_i die Ausprägung der i-ten Dimension des Vektors a bezeichnet. Somit ergibt sich, dass sich jene Vektoren a und b am ähnlichsten sind, für die $dist(a,b)$ minimal ist. Für den Fall, dass $k > 1$ werden die Klassenzugehörigkeiten der k ähnlichsten Instanzen betrachtet und die häufigste unter ihnen vorausgesagt. In praktischen Anwendungen werden oft andere Distanzmaße eingesetzt, sowie Gewichtungen der Beispiele vorgenommen. Diesbezüglich finden sich genauere Informationen beispielsweise in [28] oder in [36].

Nachteil dieses Verfahrens ist, dass die Klassifikation ineffizient ist, da für jedes Beispiel die Ähnlichkeit zu allen Trainingsinstanzen berechnet werden muss und da besonders bei hochdimensionalen Problemen oft irrelevante Features die Klassifikation beeinträchtigen („Fluch der Dimensionalität“).

6 Evaluierung

Zur Evaluierung wurde auf die Genreeinteilung, die bereits in [20] Verwendung gefunden hat, zurückgegriffen. Es handelt sich dabei um eine Taxonomie, die 14 bekannte Genres umfasst. Zu jeder der Kategorien wurde manuell eine Liste von 16, für das jeweilige Genre typischen Künstlern (bzw. Bands) erstellt, die für die durchgeführten Experimente als ground truth dienen. Dabei ist es wichtig, Künstler zu wählen, die eindeutig einem bestimmten Genre zuordenbar sind. Um hier möglichst präzise und objektiv vorgehen zu können, wurden Künstler nur zugeordnet, wenn mehrere „Expertenmeinungen" bzgl. des Genres übereinstimmen. Dazu wurden die Informationen des All Music Guide, von Launch Yahoo![1] bzw. den Yahoo! Directories[2] und teilweise die überwiegende Mehrheit der Meinung auf Webseiten herangezogen. Die 14 Genres sind: Country, Folk, Jazz, Blues, R'n'B/Soul, Heavy Metal/Hard Rock, Alternative Rock/Indie, Punk, Rap/HipHop, Electronica, Reggae, Rock'n'Roll, Pop und Klassik. Dem Genre Jazz sind beispielsweise bekannte Künstler wie *Miles Davis*, *Dave Brubeck*, *Louis Armstrong* oder *Nat King Cole* zugeordnet, das Genre Punk wird z.B. durch Gruppen wie *Sex Pistols*, *Ramones*, *Bad Religion* oder *Sum 41* repräsentiert. Die vollständige Zuordnung Genre – Künstler findet sich im Anhang A.1.

Bei dieser Einteilung wurde darauf Bedacht gelegt, einerseits ein möglichst breites musikalisches Spektrum abzudecken (von „Klassik" bis zu beispielsweise „Electronica"), andererseits durchaus ähnliche Genres miteinzubeziehen (z.B. „Alternative Rock/Indie" und „Punk", die oftmals als ein einziges Genre angesehen werden), um zu sehen, wie sich das Verfahren in diesen Fällen verhält. Auch innerhalb der Genres wurde versucht, die Künstler so auszuwählen, dass dadurch möglichst viele Aspekte der jeweiligen Kategorie abgedeckt werden, ohne dabei den tatsächlichen Kernbereich zu verlassen. So befinden sich im Genre „R'n'B/Soul" etwa mit *James Brown* und *Marvin Gaye* Künstler, die klassisch als Künstler dieser Kategorie der 1970er-Jahre angesehen werden, während die Künstlerin *Alicia Keys* typischerweise eine Repräsentatin dieses Genres in seiner heutigen Bedeutung ist. Auch im Genre „Klassik" finden sich Repräsentaten aus verschiedensten Epochen, wobei hier noch die Besonderheit hinzukommt, dass nicht alle der genannten

[1] http://launch.yahoo.com
[2] http://dir.yahoo.com/Entertainment/Music/Genres

Persönlichkeiten Komponisten sind, sondern zwei davon Dirigenten. Dies soll unterstreichen, dass die Namen schlicht für Personen (oder Gruppen) stehen, die mit einer (und nur einer) bestimmten Musikrichtung assoziiert werden. Anhand dieser Assoziationen soll dann der „Charakter“ der Kategorie lernbar gemacht werden.

6.1 Künstlerklassifikation

Auf Grundlage der Genrezuordnungen als ground truth ist es nun möglich einen Klassifikator zu trainieren und seine Klassifikationen zu bewerten. Dazu können einige der 224 Künstler herangezogen werden, um die Genres zu definieren und andere um zu überprüfen, ob der Klassifikator tatsächlich jenes Genre voraus sagt, zu dem der Künstler zugeordnet ist. Dabei ist es entscheidend, dass Trainingsset und Testset disjunkt sind, d.h. dass keiner der Künstler, dessen Zugehörigkeit geprüft wird, dazu verwendet wurde, um das jeweilige Genre zu definieren, da Evaluationen auf dem Trainingsset im Allgemeinen zu optimistisch sind und daher einen falschen Eindruck der Qualität des Systems vermitteln. Das heißt, dass jene Künstler, mit denen evaluiert wird, aus dem Lernvorgang „herausgehalten“ werden müssen. Daraus resultiert die Bezeichnung Hold-out-Experiment. Um dabei zu vermeiden, dass die Resultate in erster Linie darauf beruhen, dass zufällig Künstler zum Trainieren gewählt wurden, die besonders positive oder negative Voraussetzungen mitbringen, sollte der Vorgang mehrmals durchgeführt werden. Deshalb werden alle Klassifikationsexperimente 50 mal durchgeführt und Mittelwert und Standardabweichung der Genauigkeiten als Kennzahlen zur Abschätzung des Verfahrens berechnet. Ein Experiment folgt folgendem Ablauf:

1. Zufällige Auswahl von 2 (bzw. 4 oder 8) Künstlern pro Genre als Trainingsset
2. Für jeden Künstler Berechnung des $tf \times idf$-Wertes, für jedes Genre Berechnung des χ^2-Wertes
3. u.U. Termauswahl (siehe unten)
4. Trainieren eines Klassifikators (SVM oder k-NN)
5. Messung der Genauigkeit des Klassifikators anhand der verbliebenen Künstler

Durch Versuchsanordnungen mit verschiedenen Trainingsset-Größen soll herausgefunden werden, wie viele Künstler notwendig sind, um ein Genre ausreichend zu definieren. Der Grund, warum Experimente ausgeführt werden, in denen ein Genre nur über 2 Künstler definiert ist, ist folgendes Anwendungsszenario: Die (private) mp3-Sammlung eines Benutzers ist in einer Verzeichnisstruktur abgespeichert, wobei davon ausgegangen wird, dass die Verzeichnisse mehr oder weniger Genres widerspiegeln. Aus den ID3-Tags der mp3s können die Künstlernamen extrahiert und damit ein „Verzeichnis"-Klassifikator trainiert werden. Für neue Stücke, die der Sammlung hinzugefügt werden sollen, kann nun automatisch entschieden werden, in welches Verzeichnis sie eingefügt werden sollen. Da in der Praxis die Anzahl der verfügbaren Künstler in einem Verzeichnis allgemein nicht besonders hoch sein wird, ist es von Interesse, wie sich das Verfahren verhält, wenn nur wenige Künstler als Prototypen benutzt werden können.

Es soll auch überprüft werden, wie sehr die Anzahl der benutzten Terme pro Genre die Qualität beeinflusst. Deshalb werden die Experimente, die auf Klassifikation durch SVMs abzielen, in drei Konfigurationen basierend auf den χ^2-Werten ausgeführt: mit einem Cut-off-Index von 100 (C_{100}), von 200 (C_{200}) und ohne Berücksichtigung der χ^2 Reihung, d.h. mit allen Features (C_∞). Für jene Klassifikatoren, die über die Eignung der Featurevektoren für Ähnlichkeitsmaße Aufschluß geben (k-NN), werden daher auch nur C_∞-Vektoren benutzt. Es wird mit $k = 3$ und $k = 7$ systematisch evaluiert[3].

In Anhang A.2 finden sich Listen von Worten, wie sie typischerweise durch das beschriebene Verfahren entstehen. Zu jedem Genre werden die 100 durch den χ^2-Test am höchsten eingestuften Terme eines Experiments angegeben. Dabei ist besonders auffällig, dass in keiner der Listen die Suchterme (*music review*) auftauchen. Daran ist erkennbar, dass χ^2 seinen Zweck erfüllt, da Worte, die überall vorkommen und daher keine Information tragen, entfernt werden. Weiters ist ersichtlich, dass sehr weit vorne in allen Listen die Namen der Künstler, die das Genre definieren, vorkommen. Dies ist nachvollziehbar, da gerade Eigennamen sehr spezifisch sind (und daher gute Diskriminatoren darstellen) und vor allem die Namen der Künstler besonders häufig auf den Webseiten auftreten. Auch sehr häufig sind bei Gruppen die Namen der Bandmitglieder (in Tabelle A.9 bei Rap/HipHop beispielsweise *flav*, *griff* und *terminator* als Mitglieder von *Public Enemy*), sowie allgemein ähnliche Künstler, da diese oftmals im Zusammenhang genannt werden (*busta rhymes*, *dr dre*, *melle*, *snoop dogg*, *outkast* etc.). Viele weitere, vielleicht nicht auf den ersten Blick nachvollziehbare Wörter stammen aus Titeln von Alben oder Liedern der jeweiligen Künstler (*eyez*, *supa dupa*, *wanksta*, *chronic* etc.). Die restlichen Terme sind meist die Namen der Genres (was als Bestätigung der

[3] Anmerkung: Experimente mit 2 Trainingsbeispielen und 7-NN stellen nicht wirklich eine sinnvolle Kofiguration dar, werden aber im Zuge einer vollständigen Evaluierung auch durchgeführt.

korrekten Zuordnung angesehen werden kann), sowie andere für das Genre typische Begriffe, teilweise in spezifischer Schreibweise (*gangsta*, *dj*, *mc*, *ya*, *da*, *pimp*, *hood*, *peeps*, *beats* etc.)

Alle beschriebenen Experimente werden für beide Suchmaschinen, alle drei Typen von Suchanfragen und die Filter F_0, F_1 und F_2 durchgeführt. Auf Experimente mit Features, die durch Filterung mit F_3 gewonnen werden können, wird verzichtet, da die Datenbasis hierfür nicht mehr als substantiell genug erachtet wird (vgl. Abschnitt 3.3.2). Jene Fälle, in denen aufgrund der Filter keine Seiten mehr zur Verfügung stehen, werden so gehandhabt, dass stattdessen die Featurevektoren der nächst niedrigeren Filterstufe eingesetzt werden.

In [20] wurden bereits Ergebnisse für Experimente mit den Suchanfragen *music review* und *music genre style* auf den Suchmaschinen Google und Yahoo! durchgeführt (siehe Anhang A.4). Dabei wurden, wie auch hier, 50 Holdout-Experimente mit SVM $C_1 00$, $C_2 00$ und C_∞, sowie 3-NN und 7-NN mit C_∞ für jeweils 2, 4 und 8 Trainingsbeispiele durchgeführt. In dieser Arbeit kommen Auswertungen für Anfragen ohne Zusatzterme, sowie die Anwendung der Filter hinzu. Insgesamt ergibt sich so ein Set von 270 verschiedenen Versuchsanordnungen, wobei jede dieser Anordnungen 50 mal durchgeführt wird, was eine Gesamtzahl von 13500 Klassifikator-Evaluationen (Trainings- und Testphase) ergibt.

6.1.1 Auswertungen der Experimente mit Google

Das markanteste Merkmal der Auswertungen für die Suchmaschine Google (Tabellen 6.1, 6.2 und 6.3) ist, dass alle Experimente bessere Resultate aufweisen, als die in [20] präsentierten Experimente (Tabelle A.15). Als Höchstwert für die Klassifikationsgenauigkeit kann ein Wert von 93% beobachtet werden (Google, F_0, SVM $C_1 00$, t8), was 6 Prozentpunkte über dem Maximalwert aus [20] liegt und in Anbetracht einer Baseline von etwa 7% ein ausgesprochen gutes Ergebnis ist.

Deutlich zeigt sich die plausible Tendenz, dass für alle Verfahren und alle Suchanfragen durch steigende Anzahl an Trainingsbeispielen die Klassifikationsgenauigkeit ansteigt. Jedoch ist es bemerkenswert, dass auch bei Vorgabe von 2 Künstlern pro Genre Klassifikationsgenauigkeiten von knapp 80% erreicht werden können. Weiters zeigt sich, dass in den meisten Fällen für die Ergebnisse der SVMs der Einsatz von durch Termauswahl beschränkte Featurevektoren zu besseren Ergebnissen führt. Allerdings sind die Unterschiede (speziell bei Experimenten mit 8 Trainingsbeispielen pro Genre) nicht allzu gravierend, was darauf zurückzuführen ist, dass SVMs besonders für hochdimensionale Klassifikationen geeignet sind (siehe Abschnitt 5.1).

	Google – Filter F_0								
	keine Zusatzterme			`music genre style`			`music review`		
	t2	t4	t8	t2	t4	t8	t2	t4	t8
SVM C_{100}	71±3.7	80±3.2	87±2.6	79±4.0	87±2.7	90±2.3	77±4.1	87±2.5	93±2.0
SVM C_{200}	67±4.8	80±3.0	87±2.7	79±4.4	86±2.4	91±2.3	76±3.8	87±2.6	93±2.2
SVM C_∞	69±4.1	79±3.2	84±2.8	78±4.1	86±2.7	90±2.1	77±4.3	87±2.8	92±2.3
3-NN C_∞	53±5.6	64±4.9	71±2.7	66±5.0	78±3.7	83±3.3	60±6.7	73±5.7	79±4.6
7-NN C_∞	35±7.1	64±4.9	74±3.5	43±8.1	76±3.3	83±2.8	41±8.9	72±5.6	81±4.8

Tabelle 6.1: Klassifikationsergebnisse auf einem Datenset von 14 Genres mit je 16 assoziierten Künstlern bei Benutzung von Google und F_0. Der erste Wert einer Zelle gibt den Mittelwert der Genauigkeit aus 50 Hold-out-Experimenten an. Der zweite Wert gibt die Standardabweichung an (Werte in Prozent). Die Größe des Trainingsets (Anzahl der Künstler pro Genre zur Genredefinition) ist mit t2, t4 und t8 bezeichnet.

	Google – Filter F_1								
	keine Zusatzterme			`music genre style`			`music review`		
	t2	t4	t8	t2	t4	t8	t2	t4	t8
SVM C_{100}	71±4.0	80±3.0	87±2.5	80±3.4	87±2.4	91±2.1	79±4.1	87±2.1	92±2.2
SVM C_{200}	67±3.9	81±3.6	88±2.4	81±3.8	87±2.3	91±2.0	77±3.7	88±2.5	92±2.2
SVM C_∞	70±4.5	80±3.3	85±2.2	79±3.9	86±2.6	91±2.2	78±3.9	88±2.0	92±2.2
3-NN C_∞	54±5.4	65±4.9	71±3.9	67±5.0	80±3.6	85±3.1	63±4.9	74±4.0	80±3.1
7-NN C_∞	36±8.9	66±4.8	75±3.7	43±8.5	77±3.5	85±2.9	44±9.7	73±4.6	82±3.0

Tabelle 6.2: Klassifikationsergebnisse auf einem Datenset von 14 Genres mit je 16 assoziierten Künstlern bei Benutzung von Google und F_1. Bezeichnungen wie in Tabelle 6.1.

	Google – Filter F_1								
	keine Zusatzterme			`music genre style`			`music review`		
	t2	t4	t8	t2	t4	t8	t2	t4	t8
SVM C_{100}	71±3.8	80±3.2	86±2.7	79±3.7	85±2.6	90±2.2	75±3.6	86±2.6	91±2.3
SVM C_{200}	68±3.9	80±3.2	87±2.4	79±4.0	86±2.8	91±2.5	69±5.5	84±2.9	91±2.2
SVM C_∞	69±4.5	78±3.6	84±2.9	78±3.7	85±2.7	91±2.8	73±4.7	85±2.7	91±2.0
3-NN C_∞	53±5.9	65±4.6	74±2.6	68±4.4	78±3.1	83±3.1	57±5.8	72±3.8	78±4.0
7-NN C_∞	35±7.6	65±4.8	75±4.1	43±8.8	76±3.3	84±2.8	42±7.3	71±4.5	80±3.3

Tabelle 6.3: Klassifikationsergebnisse auf einem Datenset von 14 Genres mit je 16 assoziierten Künstlern bei Benutzung von Google und F_2. Bezeichnungen wie in Tabelle 6.1.

Bezüglich der eingesetzten Suchterme kann festgestellt werden, dass Anfragen, die nur aus dem Namen des Künstlers bestehen, die schlechtesten Resultate erzielen, die Zusatzterme *music*, *genre* und *style* am zweitbesten abschneiden und mit den Zusatztermen *music* und *review* im Allgemeinen die besten Ergebnisse erzielt werden. Dies zeichnet sich allerdings nur für SVM-Klassifikationen in einer gewissen Deutlichkeit ab, für Klassifikationen, die mit k-NN durchgeführt wurden sind die Resultate durchwegs besser, wenn *music*, *genre* und *style* zum Einsatz kommen. Somit kann man keine allgemein gültige Aussage treffen, welche Zusatzterme besser geeignet sind, sondern nur festhalten, dass die Tendenz dahin geht, dass *music review* besser geeignet ist, wenn es um Genre-Klassifikation geht und *music genre style* wenn Anwendungsgebiete in Richtung Ähnlichkeit in Betracht gezogen werden. Auf die Ergebnisse der k-NN Experimente wird in Abschnitt 6.2 genauer eingegangen, da diese primär Aufschluss über die Qualität als Ähnlichkeitsmaß geben. Bei näherer Betrachtung fällt auf, dass der Einsatz der Filter kaum Verbesserungen in der Genauigkeit mit sich bringt. In einigen Fällen (z.B. *music review*, SVM, t8) verschlechtert der Einsatz der Filter sogar die Qualität der Verfahren (im konkreten Fall: F_0...93%, F_1...92%, F_2...91%). Dies lässt vermuten, dass einerseits der Einsatz von Filtern der vorgestellten Funktionsweise nur zu einer unnötigen Reduktion der zugrunde liegenden Daten führt, die durch sinnlose Restriktionen wertvolle Information vernichtet. Es zeigt sich allerdings auch, dass andere Konfigurationen durchaus (wenn auch in geringem Ausmaß) vom Einsatz zumindest des Filters F_1 profitieren. So steigt die Genauigkeit bei Experimenten mit *music genre style* und bei Experimenten mit k-NN. Nachfolgend wird auch anhand der Ergebnisse unter Verwendung der Suchmaschine Yahoo! deutlicher, dass der Filtereinsatz durchaus seine Berechtigung haben kann. Somit scheint es eher so zu sein, dass die Daten, die durch Einsatz von Google und der Suchterme *music review* gewonnen werden, bereits so gut sind, dass jede weitere Selektion der Daten die Qualität nur beeinträchtigt. Allerdings muss festgehalten werden, dass diesbezüglich sämtliche Formen der Interpretation spekulativer Natur sind, da statistisch gesehen zwischen diesen Ergebnissen kein signifikanter Unterschied besteht. In Tabelle 6.4 werden die Ergebnisse der Experimente über die Filter verglichen. Dabei werden mittels t-Test (Verwerfen der H_0 auf Signifikanzniveau 0.05) die Resultate von F_0 und F_1 (linke Seite der Tabelle) und von F_1 und F_2 (rechte Seite) auf signifikante Unterschiede getestet. Es zeigt sich ganz deutlich, dass der Einsatz von F_1 vereinzelt zu besseren Ergebnissen führt (s.o.), aber in den meisten Fällen keinerlei Auswirkung hat. Noch viel deutlicher ist allerdings ersichtlich, dass der Einsatz von F_2 in ausgesprochen vielen Fällen signifikante Verschlechterungen im Vergleich zu F_1 nach sich zieht. Der Filtereinsatz bei Google scheint nicht dafür verant-

	Google																	
	$F_0 \to F_1$									$F_1 \to F_2$								
	k.Z.t.			`mgs`			`mr`			*k.Z.t.*			`mgs`			`mr`		
	t2	t4	t8	t2	t4	t8	t2	t4	t8	t2	t4	t8	t2	t4	t8	t2	t4	t8
SVM C_{100}							+							−	−	−	−	
SVM C_{200}				+	+							−	−	−		−	−	−
SVM C_{∞}											−	−		−		−	−	−
3-NN C_{∞}					+	+	+					+		−	−	−	−	−
7-NN C_{∞}						+									−			−

Tabelle 6.4: Signifikanz der Resultatsunterschiede zwischen den Filtern F_0 und F_1 bzw. F_1 und F_2 bei Einsatz von Google. Die Abkürzungen *k.Z.t.*, `mgs` und `mr` stehen für *keine Zusatzterme*, `music genre style` bzw. `music review`. Der Eintrag + in einer Zelle bedeutet eine signifikante Verbesserung der jeweiligen Resultate durch Einsatz des stärkeren Filters, − eine signifikante Verschlechterung und kein Eintrag bedeutet, dass es keine signifikanten Unterschiede zwischen den Filtern gibt.

wortlich zu zeichnen, dass die Ergebnisse deutlich besser sind, als in [20]. Der Hauptgrund liegt daher vermutlich im Verfahren zur Seitengewinnung. Wurden in [20] nur die ersten 50 Suchergebnisse verwertet, wobei Seiten, die nicht verfügbar waren, schlichtweg ausgelassen wurden, was zu durchschnittlich 40 verwertbaren Seiten geführt hat, wird hier mit mehr Sorgfalt an die Datengewinnung herangegangen. In diesem Ansatz werden solange neue Resultate angefragt, bis 50 Webseiten verwertet werden können. Damit werden auch faire Bedingungen für das Erlernen der Künstler- und Genreprofile geschaffen, da nunmehr alle Künstler über dieselbe Anzahl an zugrunde liegenden Seiten definiert sind. Außerdem ist es durchaus im Bereich des Möglichen, dass Verbesserungen im Rankingmechanismus bei Google einen wichtigen Beitrag zur Steigerung der Gesamtqualität des Verfahrens geleistet haben. Zeitbedingte Änderungen werden in Abschnitt 6.3 intensiv untersucht.

Die Stärken und Schwächen des Verfahrens in seiner besten Konfiguration sind detaillierter mithilfe der Confusion Matrix in Abbildung 6.1 ersichtlich. Als äußerst positiv muss festgehalten werden, dass die Genres Jazz, Klassik und Rap/HipHop perfekt klassifiziert werden. Dies ist speziell für Rap/HipHop bemerkenswert, da gerade dieses Genre eines derjenigen war, für die in [43] der Metadaten-Ansatz keine zufrieden stellenden Ergebnisse lieferte. Für Jazz und Klassik fällt außerdem positiv auf, dass neben dem Recall auch die Precision 100% erreicht[4]. Das zeigt, dass die Konzepte dieser Genres

[4]Typische Kennwerte für die Qualiät von Klassifikatoren. Unter Recall versteht man den Anteil der korrekt klassifizierten Beispiele an allen klassifizierten Beispielen einer Katego-

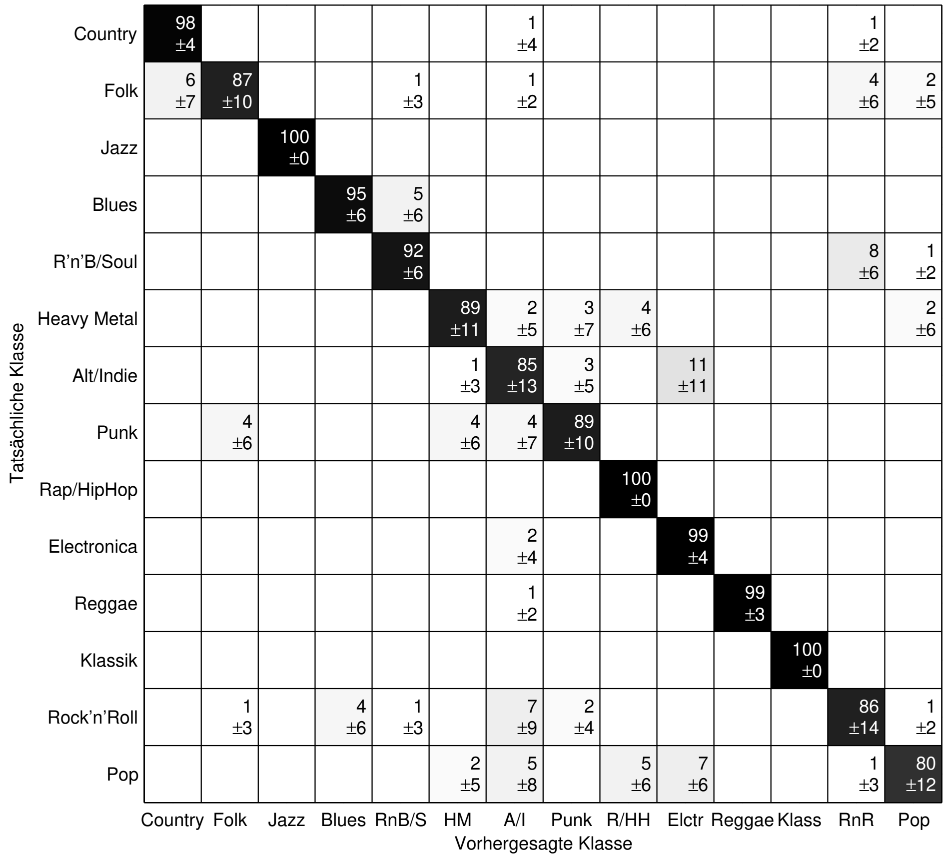

Abbildung 6.1: Confusion Matrix der Klassifikationsergebnisse von Google unter Benutzung der Terme *music review* mit SVM C_{100} und 8 Künstlern per Genre zum Training (Gesamtscore 93%). Werte in Prozent. Der obere Wert eines Feldes stellt den Mittelwert von 50 Hold-out Experimenten dar, der untere die Standardabweichung.

sehr gut von anderen unterscheidbar sind. Andere Genres mit einer sehr hohen Genauigkeit sind Country (98%), sowie Reggae und Electronica (je 99%). Die meisten Probleme treten bei den Genres Folk, Heavy Metal/Hard Rock, Alternative Rock/Indie, Punk, Rock'n'Roll und Pop auf. Für Folk zeigt sich, dass Verwechslungen hauptsächlich mit Country und Rock'n'Roll auftreten, wobei beide Fälle intuitiv erklärbar sind (bei Betrachtung der Künstlerlisten von Folk und Rock'n'Roll wird deutlich, dass es hier vor allem aufgrund der zeitlich bedingten Überschneidungen und Beeinflussungen der jeweiligen

rie, unter Precision den Anteil der korrekten Vorhersagen unter allen zu einer Kategorie als zugehörig vorausgesagten Beispielen. Eine genaue Beschreibung findet sich beispielsweise in [28]

Künstler zu Ähnlichkeiten kommt). Bei den drei Genres Heavy Metal/Hard Rock, Alternative Rock/Indie und Punk zeigt sich, dass hier Verwechslungen in allen Richtungen auftreten. Diese Genres sind einander zu ähnlich, um hundertprozentige Trennungen erwarten zu können. Das Genre Pop stellt sich mit nur 80% Genauigkeit als Problemfall dar. Dies mag in der Konzeption des Genres Pop liegen. Als Pop-Musik wird im Allgemeinen Musik angesehen, die in erster Linie populär ist (daher der Name). Das heißt eigentlich wird durch Verkaufscharts, Radiostationen etc. festgelegt, was Pop ist. Dass dies natürlich auch Musik umfasst, die eigentlich klar einem anderen Genre zuordenbar ist, versteht sich von selbst. Somit stellt sich die Frage, ob Pop als Genre überhaupt tauglich ist. Die Antwort ist dadurch gegeben, dass es als Genre zum Einsatz kommt und auch typische Künstler damit assoziiert werden. Allerdings wird durch die Verschwommenheit des Begriffs immer eine starke Durchmischung mit anderen Genres bestehen. Das Genre Pop stellt somit einen Unsicherheitsfaktor im Verfahren dar. So haben sich im Vergleich zu den Ergebnissen aus [20] alle Genres verbessert, nur das Genre Pop weist einen schlechteren Wert auf. Ähnlich geartet, wenn auch nicht mit so unvorhersehbaren Auswirkungen, ist der Fall des Genres Rock'n'Roll, das bei Betrachtung der Künstlerlisten auch als „Pop der 60er“ bezeichnet werden kann.

6.1.2 Auswertungen der Experimente mit Yahoo!

Für die Suchmaschine Yahoo! zeichnet sich ein anderes Bild ab. Zum ersten bleiben die Ergebnisse deutlich hinter jenen von Google zurück, zum zweiten ergeben sich völlig unerwartete Szenarien. Die maximal erreichte Klassifikationsgenauigkeit ist 89%. Interessanterweise treten diese besten Ergebnisse allesamt bei Experimenten auf, die auf Daten von Webseiten beruhen, die alleine mit dem Künstlernamen, ohne zusätzliche Suchterme angefragt wurden. Dies ist überraschend, da man eigentlich davon ausgehen kann, dass alleine die Suche nach beispielsweise *Kiss* oder *Hole* hauptsächlich Ergebnisse zur Folge hat, die mit Musik nicht zu tun haben. Allerdings zeigt der Vergleich mit den Ergebnissen von Google, dass auch Google ohne Suchterme ähnliche Resultate hervorgebracht hat. Diese erscheinen aber in Relation zu den Ergebnissen mit weiteren Suchtermen als nicht so bemerkenswert, wiewohl Ergebnisse um die 87% durchaus beachtlich sind. Beim Vergleich allein dieser Konstellation schneidet Yahoo! besser ab als Google. Somit stellt sich die Frage, warum die Ergebnisse durch Einsatz von Suchtermen im Verhältnis zu Google so stark zurückfallen. Eine Vermutung könnte dahin gehen, dass sich dieses Verhalten bereits bei der Anzahl der verarbeitbaren Webseiten angedeutet hat (siehe Abschnitt 3.3.2).

	Yahoo! – Filter F_0								
	keine Zusatzterme			`music genre style`			`music review`		
	t2	t4	t8	t2	t4	t8	t2	t4	t8
SVM C_{100}	71±3.9	81±3.0	88±3.1	73±3.6	80±2.7	84±2.4	66±5.3	78±3.5	86±2.4
SVM C_{200}	66±4.9	80±3.8	89±2.9	69±3.7	79±2.4	84±2.8	63±4.8	76±3.7	85±3.0
SVM C_∞	69±4.1	79±3.7	87±3.1	68±3.5	78±3.0	83±2.9	63±5.8	75±3.6	83±3.3
3-NN C_∞	51±6.2	64±5.0	72±4.1	54±5.1	67±3.7	72±2.9	45±5.7	58±4.5	66±3.6
7-NN C_∞	33±7.9	63±5.4	75±4.1	32±10.	66±4.4	75±2.8	38±8.1	59±4.6	67±3.8

Tabelle 6.5: Klassifikationsergebnisse auf einem Datenset von 14 Genres mit je 16 assoziierten Künstlern bei Benutzung von Yahoo! und F_0. Bezeichnungen wie in Tabelle 6.1.

	Yahoo! – Filter F_1								
	keine Zusatzterme			`music genre style`			`music review`		
	t2	t4	t8	t2	t4	t8	t2	t4	t8
SVM C_{100}	73±3.8	82±3.2	89±2.7	76±3.1	82±2.1	86±3.1	67±4.5	79±3.1	87±2.6
SVM C_{200}	69±5.1	81±3.2	89±2.6	74±3.3	81±3.0	86±3.2	65±5.2	79±3.0	87±3.4
SVM C_∞	71±4.3	81±3.4	88±2.7	74±3.5	81±2.6	86±3.2	65±5.3	76±4.0	83±4.6
3-NN C_∞	54±5.3	66±4.2	74±3.3	60±6.4	73±3.0	78±3.6	48±5.4	59±4.0	64±4.6
7-NN C_∞	35±7.2	65±3.9	76±3.3	37±8.6	72±3.4	79±3.1	40±6.4	60±5.0	68±4.3

Tabelle 6.6: Klassifikationsergebnisse auf einem Datenset von 14 Genres mit je 16 assoziierten Künstlern bei Benutzung von Yahoo! und F_1. Bezeichnungen wie in Tabelle 6.1.

	Yahoo! – Filter F_1								
	keine Zusatzterme			`music genre style`			`music review`		
	t2	t4	t8	t2	t4	t8	t2	t4	t8
SVM C_{100}	72±4.4	82±2.8	89±2.2	75±3.8	82±2.2	87±2.4	68±3.7	79±2.9	87±2.9
SVM C_{200}	69±5.0	81±2.6	89±2.7	75±4.3	83±2.3	88±2.5	62±5.2	76±3.4	87±3.1
SVM C_∞	71±4.8	81±3.1	87±2.8	74±4.2	82±2.3	86±2.7	66±4.6	77±2.8	87±3.2
3-NN C_∞	55±5.8	66±4.0	73±3.6	63±4.4	75±3.3	79±2.9	51±6.1	65±3.8	75±3.6
7-NN C_∞	39±7.9	66±4.7	75±3.5	39±9.3	73±3.6	83±2.6	36±9.2	66±4.0	78±3.6

Tabelle 6.7: Klassifikationsergebnisse auf einem Datenset von 14 Genres mit je 16 assoziierten Künstlern bei Benutzung von Yahoo! und F_2. Bezeichnungen wie in Tabelle 6.1.

	Yahoo!																	
	$F_0 \rightarrow F_1$									$F_1 \rightarrow F_2$								
	k.Z.t.			`mgs`			`mr`			*k.Z.t.*			`mgs`			`mr`		
	t2	t4	t8	t2	t4	t8	t2	t4	t8	t2	t4	t8	t2	t4	t8	t2	t4	t8
SVM C_{100}	+			+	+	+			+						+			
SVM C_{200}	+	+		+	+	+		+	+					+	+	−	−	
SVM C_{∞}	+	+		+	+	+												+
3-NN C_{∞}			+	+	+	+	+		−				+	+		+	+	+
7-NN C_{∞}		+	+	+	+	+				+					+	−	+	+

Tabelle 6.8: Signifikanz der Resultatsunterschiede zwischen den Filtern F_0 und F_1 bzw. F_1 und F_2 bei Einsatz von Yahoo!. Bezeichnungen wie in Tabelle 6.4

Auch wenn die Durchschnittswerte nur gering unter jenen von Google liegen, sieht man, dass die minimale Anzahl an verwertbaren Webseiten deutlich unter jener von Google liegt. Dies legt nun die Vermutung nahe, dass vor allem das Ungleichgewicht bei der Anzahl der Webseiten für verschiedene Künstler problematisch ist. Wie bereits im vorangegangenen Abschnitt (6.1.1) diskutiert, hat sich das Verfahren dort vermutlich hauptsächlich durch die Schaffung gleicher Ausgangsbedingungen für alle Beispiele verbessert. Offensichtlich stellt es für Google ein geringeres Problem dar, als für Yahoo!, ausreichend Webseiten zu finden, auf denen die Suchterme enthalten sind. Im Fall, wo nur der Name gesucht wird, ist Yahoo! ebenbürtig, wenn nicht sogar überlegen, sobald die Anfragen komplexer werden, steigt bei Google die Gesamtqualität des Verfahrens und bei Yahoo! sinkt sie.

Ein weiterer markanter Unterschied zu Google ist, dass der Einsatz der Filter hier in vergleichsweise sehr vielen Fällen signifikante Verbesserungen mit sich bringt (Tabelle 6.8). Erstaunlicherweise führt nicht nur der Einsatz des Filters F_1 zu besseren Ergebnissen, sondern auch noch der des Filters F_2. Dies lässt wieder Rückschlüsse auf die Qualität der zugrunde liegenden Webseiten zu. Offenbar sind diese von Haus aus nicht besonders hochwertig bzw. im Sinne dieser Aufgabe geeignet, so dass die Nachselektion entscheidend störende Seiten entfernt.

Ansonsten zeigt sich auch bei Yahoo! das erwartete Bild, dass die Klassifikationsgenauigkeit mit steigender Anzahl an Trainingsbeispielen steigt. Im Vergleich mit den Ergebnissen aus [20] (siehe Tabelle A.16) zeigt sich keine eindeutige Tendenz in Richtung Verbesserung oder Verschlechterung.

6.2 Einsatz als Ähnlichkeitsmaß

Um festzustellen, wie sehr sich die gewonnenen Features der Künstler dazu eignen, um in einem Ähnlichkeitsmaß Verwendung zu finden, empfiehlt sich die Betrachtung der Ergebnisse jener Experimente, die mit k-NN ohne χ^2-basierten Cut-off durchgeführt wurden. In der Praxis haben Ähnlichkeitsmaße ein viel breiteres Einsatzspektrum, als Genre-Klassifikationen (vgl. Kapitel 1). Ein Metadaten-basiertes Ähnlichkeitsmaß könnte beispielsweise im Islands of Music Projekt Einsatz finden, wo unterschiedliche Sichtweisen auf Musik kombiniert werden, um interaktives Browsen in Musikkollektionen zu ermöglichen [33].

Die Zuordnung von Künstlern zu Genres kann als ground truth für die Evaluierung von Ähnlichkeit dienen. Die Evaluierung zielt darauf ab, festzustellen, ob zu einem Künstler ein ähnlicher (der ähnlichste) gefunden wird. Das Kriterium der Ähnlichkeit kann damit darauf reduziert werden, ob der vermeintlich ähnliche Künstler aus demselben Genre stammt, da die zugrunde liegende Annahme ist, dass sich Künstler innerhalb eines Genres alle ähnlich sind.

Klassifikationsergebnisse von bis zu 85% (Tabelle 6.2, *music genre style*, t8) sind ein sehr ermutigendes Ergebnis und zeigen die prinzipielle Anwendbarkeit. Auch hier erscheint Google wieder als bessere Wahl, allerdings erweisen sich die Zusatzterme *music genre style* in diesem Fall offensichtlich als besser geeignet, als *music review*. Das Maximalergebnis von 85% kommt zustande, indem Filter F_1 eingesetzt wird, es handelt sich dabei um einen der seltenen Fälle, in denen die Anwendung eines Filters bei Google eine signifikante Verbesserung mit sich bringt. Aber auch bei Benutzung anderer Filter und der Terme *music review* kann man Klassifikationsergebnisse von rund 80% erwarten. Ein Nachteil ist, dass dieses Evaluationsverfahren unter Umständen zu pessimistisch ist, da zwar festgestellt werden kann, ob ein Künstler ähnlich zu anderen des selben Genres ist, aber nicht, ob nicht vielleicht eine größere Ähnlichkeit zu einem Künstler eines anderen Genres besteht. Damit kann es passieren, dass durch das Ähnlichkeitsmaß enttarnte Ähnlichkeiten als Fehlklassifikation eingestuft werden. Um diese Fälle berücksichtigen zu können, müssten Ähnlichkeiten zwischen den Genres explizit angegeben werden.

Um die Anwendbarkeit als Ähnlichkeitsmaß weiter zu testen, wurde eine Self-Organizing Map mit allen 224 Künstlern trainiert (Abbildung 6.2). Eine Self-Organizing Map (SOM) ist ein spezielles Neuronales Netz (Kohonen-Netz, benannt nach seinem Erfinder), das oft zur Visualisierung eingesetzt wird [22]. Die SOM ist sowohl ein Verfahren zur Multi-dimensionalen Ska-

KLASSIK (16)	JAZZ (7)	RNBSOUL (10)	rnbsoul (4) pop (3) folk (1) raphiphop (1) electro (1)	RAPHIPHOP(14)	REGGAE (15)
JAZZ (6) rocknroll (1)	jazz (1)	altindie (1) pop (1)	POP (6) reggae (1)	raphiphop (1) pop (1)	electro (1) pop (1)
ROCKNROLL (5) jazz (1) rnbsoul (1)	ROCKNROLL (5) blues (1) rnbsoul (1)	country (1)	pop (3)	electro (2) pop (1)	ELECTRO (12) altindie (1)
BLUES (15)	COUNTRY (12) folk (1)	FOLK (10) country (3) jazz (1)	ALTINDIE (5) ROCKNROLL (5) folk (4) punk (1)	PUNK (5) altindie (3) heavymetal (2)	HEAVYMETAL (14) PUNK (10) ALTINDIE (6)

Abbildung 6.2: SOM trainiert mit allen 224 Künstlern. Die Anzahl der Künstler des jeweiligen Genres, die durch eine Unit repräsentiert werden ist in Klammern angegeben. Genrenamen in Großschreibung sollen Units hervorheben, die besonders viele Künstler eines Genres repräsentieren.

lierung, als auch zum Clustering und bildet hoch-dimensionale Vektoren auf eine 2-dimensionale Karte so ab, dass ähnliche Vektoren nahe beisammen platziert werden (auf die gleiche Unit oder benachbarte). Da die SOM nur ein Ähnlichkeitsmaß benötigt, aber keine Zuordnung von Künstlern zu Genres, kann die SOM dazu verwendet werden, die inhärente Struktur in den Daten zu finden. Dies ist besonders von Vorteil, wenn es darum geht, automatisch Musikkollektionen zu organisieren und zu visualisieren [31, 33]. Wie im Falle der k-NN kommt als Ähnlichkeitsmaß auch hier die euklidische Distanz (Formel 5.2) zum Einsatz. Als Grundlage dienen jene Daten, die bei der k-NN Klassifikation am besten abgeschnitten haben (Google, *music genre style*, F_1, C_∞). Für die Durchführung der Experimente wird die Matlab SOM Toolbox[5] verwendet.

Die SOM spiegelt einige der Ergebnisse der Confusion Matrix wider. Klassik (Unit links oben) ist klar von allen Genres getrennt, alle Klassik-Künstler werden auf diese eine Unit abgebildet. Auch Reggae (rechts oben) und Blues

[5] http://www.cis.hut.fi/projects/somtoolbox

(links unten) lassen sich sehr gut von den anderen Genres unterscheiden (jeweils 15 von 16 Künstlern auf einer Unit). Weiters ist zu sehen, dass auch das Genre Rap/HipHop eine hohe Konsistenz aufweist (14 Künstler auf einer Unit; neben Reggae).

Rechts unten auf der Karte zeigt sich, dass Heavy Metal/Hard Rock, Punk und Alternative Rock/Indie sehr stark überlappen. In den drei Units am rechten unteren Rand befinden sich 46 von 48 Künstlern aus diesen Genres. Außerdem tritt noch eine Überschneidung von Alternative Rock/Indie mit dem Genre Rock'n'Roll auf. Auch dieses Verhalten hat sich bereits in der Confusion Matrix offenbart. Ein besonders interessanter Aspekt der SOM ist die globale Anordnung, die zumindest einen Eindruck von der Ähnlichkeit der Genres untereinander gibt, auch wenn dies nicht quantifizierbar ist. Hier zeigt sich, dass das ähnlichste Genre zu Klassik Jazz ist, was in Anbetracht der anderen Genres sicherlich zutrifft. Des weiteren zeigt sich sehr eindrucksvoll, dass der Bereich mit Heavy Metal, Punk und Alternative die maximale Distanz zu Klassik auf der Karte aufweist. Auch die Nachbarschaft von R'n'B/Soul mit Rap/HipHop ist nachvollziehbar, da diese Genres einerseits geschichtlich miteinander verknüpft sind und sich vor allem in den letzten 10-15 Jahren gegenseitig beeinflusst haben. Außerdem erscheint die Nachbarschaft von Blues sowohl mit Rock'n'Roll, als auch mit Country plausibel, da diese Genres (gemeinsam mit dem benachbarten Folk) speziell in den 1960er und 1970er Jahren starken Einfluß aufeinander gehabt haben.

Als letzter Punkt soll die Verbreitung des Genres Pop hervorgehoben werden. Auf die Probleme, die bei der Definition von Pop entstehen, wurde bereits in Abschnitt 6.1.1 eingegangen. In der SOM zeigt sich sehr gut, dass Pop sehr zentral liegt und auf viele umliegende Units verteilt ist. Dies kann als Beleg dafür angesehen werden, dass für Pop keine klar umrissene musikalische Definition existiert, bzw. dafür, dass die für dieses Genre ausgewählten Künstler, auch wenn sie allesamt Pop-Künstler sind, keine homogene Gruppe darstellen. Besonders interessant ist noch, auf welche Units sich Pop verteilt. Grob gesehen bewegt sich Pop im Bereich zwischen R'n'B/Soul, Rap/HipHop und Electronica, was angesichts der Verkaufscharts der letzten 15 Jahre (speziell in den USA) durchaus als korrekt angesehen werden kann.

6.3 Abschätzung der zeitbedingten Varianz

Bekannterweise sind Inhalte im Internet kurzlebig. Auch die Reihungen von Suchergebnissen durch Suchmaschinen ändern sich häufig. Die Gefahr, zu einem bestimmten Zeitpunkt eine tendenziöse Sicht des Internet zu bekommen, besteht. Arbeiten, die sich mit dieser Problematik auseinander setzen,

Youssou N'Dour (79)			Stacie Orrico (79)		Robbie Williams (79)		Daft Punk (79)
							Strokes (79)
			Alicia Keys (78)	Alicia Keys (1)			
Mozart (79)							Michael Jackson (79)
					Eminem (9)		
Pulp (77)	Pulp (2)	Sublime (3)	Sublime (76)		Eminem (70)		Marshall Mathers (79)

Abbildung 6.3: SOM trainiert auf den Daten der Suchanfrage *music review*. Pro Künstler 79 Datenpunkte aus einer Zeitspanne von 11 Monaten. Die Zahl unter dem Namen eines Künstlers, gibt an, wie viele Ergebnisvektoren von verschiedenen Tagen auf die selbe Unit abgebildet werden.

sind beispielsweise [21] und [23]. Diese Möglichkeit stellt somit auch ein reales Risiko für den präsentierten Ansatz dar, da alle erzielten Resultate unter Umständen nur für den Moment, in dem die Daten für die Experimente gewonnen wurden, gültig sind. Das heißt, um die Anwendbarkeit des Verfahrens tatsächlich zu zeigen, ist es nötig, die Zeitabhängigkeit abzuschätzen. Um den Einfluss der Fluktuationen der Web-Inhalte auf die Datenrepräsentation der Künstler, konkret auf die $tf \times idf$-Vektoren, erfassen zu können, wurden über einen Zeitraum von 11 Monaten, beginnend mit dem 18. Dezember 2003, jeden Tag Anfragen an Google geschickt (ausgenommen während einer zweiwöchigen Unterbrechung rund um den Jahreswechsel 2003/04).

Die Anfragen umfassen 12 Künstler aus verschiedenen Genres, die beliebig gewählt wurden. Für jeden Künstler wurden Anfragen mit den Suchtermen *music review* und *music genre style* geschickt. Von den wichtigsten 50 Seiten wurden alle verfügbaren einbezogen, um die $tf \times idf$-Vektoren (ohne χ^2 Termauswahl) zu berechnen. Aus Gründen der Berechnungseffizienz wird von den Daten jeder vierte Tag ausgewertet, was insgesamt 79 Momentaufnahmen pro Künstler und Suchanfrage ergibt. Für den Fall, dass an einem Tag keine Ergebnisse verfügbar waren (Suchmaschine nicht erreichbar) werden die Daten des nächsten Tages herangezogen.

Pulp (79)		Sublime (57)		Alicia Keys (55)	Alicia Keys (24)		Mozart (79)
		Sublime (22)					
Strokes (79)			Eminem (9)			Stacie Orrico (1)	Stacie Orrico (78)
Michael Jackson (2)			Eminem (7)	Eminem (42)			
Michael Jackson (76)	Michael Jackson (1)			Eminem (21)			
Daft Punk (79)		Robbie Williams (79)		Marshall Mathers (79)			Youssou N'Dour (79)

Abbildung 6.4: SOM trainiert auf den Daten der Suchanfrage *music genre style*. Bezeichnungen wie in Abbildung 6.3.

Zur Beobachtung der Varianz wurde für beide Formen der Anfragen eine SOM mit allen jeweiligen Vektoren trainiert. Die Ergebnisse sind in den Abbildungen 6.3 und 6.4 dargestellt.

Beispielsweise zeigt sich für beide Suchanfragetypen, dass alle 79 $tf \times idf$-Vektoren für die Künstler *Robbie Williams*, *Youssou N'Dour*, *Daft Punk*, *Strokes*, *Marshall Mathers* und *Mozart* auf nur jeweils eine einzige Unit abgebildet werden. Dies ist bereits ein starkes Indiz dafür, dass die Daten über 11 Monate hinweg relativ gleich bleiben. Die Vektoren für *Eminem* und *Marshall Mathers* (*Eminem*s bürgerlicher Name) sind benachbart, auch wenn in Abbildung 6.3 eine Unit dazwischen freigeblieben ist. Erwähnenswert ist auch die Tatsache, dass es zwischen Künstlern keine Überlappungen gibt, d.h. jede Unit repräsentiert maximal einen Künstler. Auffällig sind jedoch die Unterschiede zwischen den Suchanfragen. So weist *music genre style* offenbar stärkere Schwankungen auf, als *music review*. Besonders deutlich wird dies bei den Künstlern *Michael Jackson* und *Eminem*. Bei letzterem erstrecken sich die Vektoren über vier Units, soviele, wie in keinem anderen Fall. Es ist anzunehmen, dass die Tatsache, dass über *Eminem* und seine Plattenveröffentlichungen im Verlauf der 11 Monate sehr intensiv berichtet wurde, auch Auswirkungen auf die Suchresultate hat. Offenbar treten diese Änderungen bei *music genre style* stärker zu Tage, was darauf zurückzuführen ist, dass durch diese Suchterme keine konkrete Form von In-

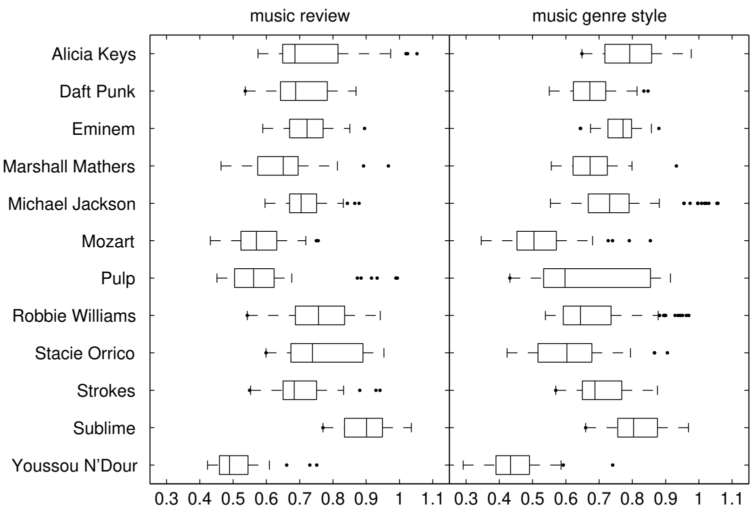

Abbildung 6.5: Box-Whisker-Plots zeigen die Varianz der Daten über die Zeit. Die x-Achse beschreibt die relative Distanz zwischen dem Mittelwert eines Künstlers über die Zeit und jedem einzelnen Tag, normalisiert über die durchschnittliche Distanz zwischen den Vektoren von *Eminem* und *Marshall Mathers*. Die Boxen haben Linien beim unteren Quartil, beim Median und beim oberen Quartil. Die Whisker sind verlängerte Linien aus den Enden der Box, die das Ausmaß der restlichen Daten anzeigen sollen. Werte außerhalb der Whisker sind Ausreißer.

halt auf Webseiten gesucht wird, sondern allgemein Seiten, die sich mit dem Künstler und der Art seiner Musik beschäftigen. Somit ist diese Form wohl anfälliger für Variationen in den Daten.

Zusätzlich wird die Abweichung über die Zeit noch wie folgt berechnet. Aus den 79 Vektoren $\{\mathbf{v}_{ad}\}$, die zu Künstler a gehören, wobei d den Tag, an dem die Seiten angefragt wurden bezeichnet, wird der Durchschnittsvektor $\overline{\mathbf{v}}_a$ gebildet. Für jeden Künstler wird die tägliche Distanz zu diesem Mittelwert berechnet ($d_{ad} = ||\mathbf{v}_a - \mathbf{v}_{ad}||$). Die Ergebnisse für *music review* und *music genre style* sind in Abbildung 6.5 dargestellt. Die Distanzen werden so normalisiert, dass der durchschnittliche Abstand zwischen *Eminem* und *Marshall Mathers* 1 beträgt.

Die Resultate zeigen, dass die Abweichungen vom Mittel allgemein deutlich kleiner als 1 für alle Künstler sind. Es gibt allerdings auch Ausnahmen. Beispielsweise sind einige der Resultate für *Michael Jackson* bei *music genre style* ziemlich verschieden von den anderen. Wie bereits zuvor erwähnt, ist dies möglicherweise auf Änderungen in der Berichterstattung über den Künstler zurückzuführen. Im Falle von *Michael Jackson* könnte dies mit einem gegen ihn angestrengten Gerichtsverfahren in Zusammenhang stehen.

In beiden Fällen erweist sich der afrikanische Künstler *Youssou N'Dour*, als derjenige mit den konstantesten Resulaten. *Youssou N'Dour* ist einer breiten Öffentlichkeit im Jahre 1994 durch seinen Hit *Seven Seconds* bekannt geworden. Die Vermutung, die konstanten Resultate lassen sich alleine darauf zurückführen, dass er im Jahr 2004 keine öffentlichkeitswirksamen Aktivitäten gesetzt hat (abgesehen von der Auslosung der Afrika-Qualifikationsgruppen für die Fußball-Weltmeisterschaft 2006), kann dadurch widerlegt werden, dass selbiges auch für die Alternative-Ska-Punk-Gruppe *Sublime* gelten müsste, die sich nach dem Tod ihres Sängers 1996 aufgelöst hat. Für *Sublime* sind die Variationen allerdings bedeutend größer.

Abschließend kann festgehalten werden, dass es in den benutzten Seiten signifikante Variationen gibt. In der SOM ist ersichtlich, dass diese Variationen aber nicht derart markant sind, dass Überschneidungen oder gar Verwechslungen zwischen den Künstlern auftreten. Änderungen in den Daten über die Zeit sind nicht per se ein Übel, das es zu verhindern gilt, sondern einer der Hauptgründe, warum die Exploration von Web-basierten Daten so interessant ist. So wie sich die öffentliche Meinung verändert und weiterentwickelt, so soll auch die Repräsentation an die aktuellen Gegebenheiten angepasst werden können. Dies scheint aufgrund der Eindrücke der Langzeitstudie möglich, ohne dass dadurch völlig unvorhersehbares Verhalten entsteht. Daher kann davon ausgegangen werden, dass der Klassifikationsprozeß nicht übermäßig negativ von zeitbedingten Schwankungen beeinflusst wird. Jedoch sind weitere Untersuchungen nötig, um die Auswirkungen auf größere Gruppen von Künstlern abschätzen zu können.

7 Zusammenfassung und Ausblick

In dieser Arbeit wurde ein Ansatz zur automatischen Klassifikation von Musikkünstlern anhand von Web-basierten Daten vorgestellt. Anhand mehrerer Experimente wurden folgende Erkenntnisse gewonnen. Mithilfe einer 14 Genres umfassenden Taxonomie und 8 zu jedem Genre zugeordneten Musikkünstlern ist es möglich, Klassifikationsgenauigkeiten von 93% zu erzielen, wobei nur Informationen von Seiten aus dem Internet Verwendung finden. Anhand der Ergebnisse lässt sich argumentieren, dass die Suchmaschine Google für diese Aufgabe besser geeignet ist, als die Suchmaschine Yahoo!. Außerdem hat sich gezeigt, dass mit Suchanfragen der Form `"`*`Künstlername`*`" music review` die besten Klassifikationsergebnisse erzielt werden. Zur weiteren Verbesserung des Verfahrens aus [20] wurde der Einsatz von einfachen regelbasierten Filtern zur Steigerung der Qualität der zugrundeliegenden Daten vorgeschlagen. Es hat sich gezeigt, dass Daten, die mit Hilfe der Suchmaschine Google gewonnen wurden, einer solchen Filterung nicht bedürfen, diese sogar schädlich für das Verfahren sein kann. Bei Verwendung von Daten, die über die Suchmaschine Yahoo! bezogen werden, erweist sich der Einsatz hingegen deutlich als qualitätssteigernd. Als besonders bemerkenswert ist die Tatsache anzusehen, dass auch mit nur jeweils 2 Künstlern zur Definition eines Genres Klassifikationsgenauigkeiten von rund 80% erreicht werden können, was vor allem deshalb als wichtige Erkenntnis gesehen werden kann, da in praktischen Anwendungen nicht immer von einer großen Zahl an bekannten Künstlern in jeder Kategorie ausgegangen werden kann. Weiters wurde gezeigt, dass die gewonnenen Features auch für die direkte Verwendung in einem Ähnlichkeitsmaß geeignet sind, was dem Verfahren ein breiteres Spektrum an Anwendungsmöglichkeiten eröffnet. Zu guter letzt wurde anhand einer Langzeitstudie über einen Zeitraum von beinahe einem Jahr gezeigt, dass die täglichen Fluktuationen des Internets die Anwendbarkeit des Verfahrens nicht entscheidend behindern.

Nichts desto trotz bringt der Einsatz von Webdaten auch Beschränkungen und Nachteile mit sich. Der wohl schwerwiegendste Nachteil besteht darin, dass dieser Ansatz extrem von den zugrunde liegenden Suchmaschinen und der Annahme, dass die vorgeschlagenen Webseiten stark in Zusammenhang mit den Künstlern stehen, abhängt. Mit der Qualität der Suchmaschine steht und fällt daher die Präzision des Verfahrens. Auch wenn ansatzweise die Möglichkeit besteht, eindeutig falsche Vorschläge zu ignorieren, bietet dies noch keine ausreichende Gewissheit, dass das Verfahren trotz zweifelhafter Datengrundlage funktionieren kann. Dass die nachträgliche Bewertung von Webseiten problematisch ist, wurde in der Arbeit gezeigt. Die hier erprobten Ansätze zur Qualitätsbewertung stellen mit Sicherheit nicht die beste Form der Entscheidung dar, jedoch muss auch angeführt werden, dass die Möglichkeiten, nicht relevante Webseiten zu ignorieren, im Allgemeinen sehr limitiert sind, sofern nicht domänenspezifisches Wissen eingesetzt werden kann. Ein Beispiel soll dies illustrieren. Um Webseiten für die Heavy Metal-Band *Slayer* zu finden, wurde an Google die Suchanfrage `¨slayer¨ +music +genre +style` geschickt. In den Wortstatistiken zu den zugehörigen Seiten gab es auffällig hohe Frequenzen der Terme *vampire* und *buffy*. Filter, wie die vorgestellten, aber auch „intelligentere“ Filter, die über die Möglichkeit verfügen, zu überprüfen, ob es sich um Seiten handelt, die überhaupt mit Musik zu tun haben, sind nicht in der Lage, zu verhindern, dass Seiten einbezogen werden, die sich mit dem Soundtrack der TV-Serie „Buffy The Vampire Slayer“ beschäftigen. Das domänenspezifische Wissen hätte in diesem Fall darin bestehen können, zur Suchanfrage die Bedingung −`buffy` hinzuzufügen, um solche Seiten von vornherein auszuschließen. Ein ähnliches Verhalten zeigt sich beispielsweise bei Suchanfragen für die Band *Son Goku*, die vor allem Seiten zurückliefern, die sich mit dem Soundtrack der Anime-Serie „Dragonball“ beschäftigen, da die Band nach einem der Charaktere der Serie benannt ist. Dabei handelt es sich um schwerwiegendere Probleme, als typische Verwechslungen bei Bands wie *Kiss* oder *Bush* mit Seiten, die gar nichts mit Musik zu tun haben, da oben genannte Beispiele sehr wohl Seiten liefern, die sich um Musik drehen. Das Hauptproblem liegt, verkürzt gesagt, darin, dass Künstlernamen keine eindeutigen Bezeichner sind.

Dies ist umso mehr eine Gefahrenquelle, da, wie aus den Wortlisten im Anhang A.2 hervorgeht, die Namen von Künstlern und Bands eine besonders wichtige Rolle spielen. Daher ist es auch möglich, dass Künstler mit häufigen Namen Opfer von Fehlklassifikationen werden. Ist beispielsweise das Genre Pop durch *Michael Jackson* und *Janet Jackson* definiert, so werden Seiten, die den Term *jackson* enthalten, eher als Pop klassifiziert. Dies beinhaltet auch Seiten des Country-Künstlers *Alan Jackson*. Ein ähnliches Problem ist Rap-Künstler *Nelly*, dessen Name ein Teilstring der Ethno-Pop-Künstlerin *Nelly Furtado* ist. Auch Bandnamen wie jener der Electronic-Künstler *Daft Punk* können zu Fehlinterpretationen verleiten, da der zweite Teil des Namens häufiges Auftreten des Terms *punk* verursacht. Das bloße Auftreten eines Terms ist natürlich nicht entscheidend für die gesamte Klassifikation, jedoch sollten Möglichkeiten gefunden werden, um derartige Tendenzen weitestgehend vermeiden zu können.

Ein Ansatz, um vor allem das Problem der Namensüberschneidungen zu behandeln, wäre die Benutzung von Noun-Phrases (wie in [42] vorgeschlagen). Auch wäre es möglich, Künstlernamen als besondere Identifikatoren und nicht als normale Terme anzusehen. Dies wirft allerdings neue Probleme auf (z.B., dass die Liste aller Künstlernamen bekannt sein müsste) und könnte auch wichtige Information entfernen. So ist es durchaus wünschenswert, wenn andere, ähnliche Künstler auf Seiten vorkommen, da Beschreibungen oft unter Bezugnahme auf andere erfolgen und diese Namen auch als Feature-Dimension Teil des Klassifikationsprozesses werden.

Andere Verbesserungsvorschläge sind der minimale Einsatz von Stemming, also der Reduzierung von Worten auf ihren Wortstamm. Eine einfache Möglichkeit dazu, die die Daten geringfügig genauer machen kann, ohne aber deshalb gravierend in die Semantik der Texte einzugreifen, wäre etwa die Reduzierung von Plural-Wörtern auf die Singularform, wenn der Singular auch in der Liste enthalten ist.

Weiters bieten sich bessere Formen von Filtern an. Eine Möglichkeit, starke Ausreißer identifizieren zu können, wäre die Berechnung des Mittelwerts über beispielsweise 50 Seiten mit darauf folgender Neuauswertung, in der Seiten, die über ein gewisses Maß vom Mittelwert abweichen, ignoriert werden. Ein anderer Lösungsansatz, der möglicherweise in der Lage ist, oben beschriebene Probleme zu umgehen, wäre das Hinzufügen von Songtiteln der Künstler zu den Suchanfragen. Erste Ergebnisse zeigen, dass damit zwar die Treffgenauigkeit für relevante Seiten enorm ansteigt, die Anzahl der verfügbaren Seiten aber mitunter drastisch sinkt. Außerdem beschränkt sich der Inhalt meist nur auf einzelne Songs, Videos oder Songtexte. Eine erfolgsversprechende Methode wäre die Kombination mehrerer Suchanfragen mit verschiedenen Songtiteln zur Konstruktion eines Filters, der dann auf die altbekannten Suchresultate angewendet werden kann. Damit könnte es möglich sein, die Präzision der Anfragen mit Songtitel und die Vielfalt der Anfragen ohne zu kombinieren.

Des weiteren soll in Zukunft auch die Klassifikation in hierarchisch strukturierte Genre Taxonomien wie in [9] oder [26] untersucht werden. Unter Zuhilfenahme beispielsweise des Wissens, dass sich in der Hierarchie benachbarte Genres ähnlich sind, könnte die Klassifikationsgenauigkeit vor allem bei sehr schwer unterscheidbaren Kategorien verbessert werden. Andere Perspektiven für zukünftige Verfahren beinhalten die Einbeziehung zusätzlicher Information von Google (die erstgereihte Seite sollte relevanter als die fünfzigste sein), das Experimentieren mit anderen Formen von Suchanfragen und die Kombination mit Audio-basierten Ansätzen.

Literaturverzeichnis

[1] Jean-Julien Aucouturier and François Pachet. Finding songs that sound the same. In *Proceedings of the IEEE Benelux Workshop on Model Based Processing and Coding of Audio*, pages 91–98, Lueven, Belgium, 2002. University of Lueven.

[2] Jean-Julien Aucouturier and François Pachet. Music similarity measures: What's the use? In *Proceedings of the Third International Conference on Music Information Retrieval (ISMIR'02)*, pages 157–163, Paris, France, October 2002. IRCAM.

[3] Jean-Julien Aucouturier and François Pachet. Musical genre: A survey. *Journal of New Music Research*, 32(1), 2003.

[4] Jean-Julien Aucouturier and François Pachet. Representing musical genre: A state of the art. *Journal of New Music Research*, 32(1):83–93, 2003.

[5] Jean-Julien Aucouturier and François Pachet. Improving timbre similarity: How high is the sky? *Journal of Negative Research Results in Speech and Audio Sciences*, 1(1), 2004.

[6] Stephan Baumann and Oliver Hummel. Using cultural metadata for artist recommendation. In *Proceedings of the 3rd WedelMusic Conference*, September 2003.

[7] Adam Berenzweig, Dan P.W. Ellis, and Steve Lawrence. Anchor space for classification and similarity measurement of music. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'03)*, Baltimore, MD, 2003. IEEE.

[8] Adam Berenzweig, Beth Logan, Daniel P.W. Ellis, and Brian Whitman. A large-scale evaluation of acoustic and subjective music similarity measures. In *Proceedings of the Third International Conference on Music Information Retrieval (ISMIR'03)*, Washington DC, 2003.

[9] Juan José Burred and Alexander Lerch. A hierarchical approach to automatic musical genre classification. In *Proceedings of the 6th International Conference on Digital Audio Effects (DAFX)*, London, UK, 2003.

[10] Donald Byrd and Michael Fingerhut. The History of ISMIR – A Short Happy Tale. *D-Lib Magazine*, 8(11), November 2002.

[11] William W. Cohen and Wei Fan. Web-collaborative filtering: Recommending music by crawling the web. *WWW9 / Computer Networks*, 33(1-6):685–698, 2000.

[12] Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, IT-13(1), January 1967.

[13] Sally Jo Cunningham, Matt Jones, and Steve Jones. Organizing digital music for use: An examination of personal music collections. In *Proceedings of 5th International Conference on Music Information Retrieval (ISMIR'04)*, pages 447–454, Barcelona, Spain, October 2004.

[14] Franca Debole and Fabrizio Sebastiani. Supervised term weighting for automated text categorization. In *Proceedings of SAC-03, 18th ACM Symposium on Applied Computing*, pages 784–788, Melbourne, US, 2003. ACM Press, New York, US.

[15] Daniel Ellis, Brian Whitman, Adam Berenzweig, and Steve Lawrence. The quest for ground truth in musical artist similarity. In *Proceedings of 3rd International Conference on Music Information Retrieval (ISMIR'02)*, Paris, France, 2002.

[16] Jonathan T. Foote. Content-based retrieval of music and audio. In C. Kuo, editor, *Proceedings of SPIE Multimedia Storage and Archiving Systems II*, volume 3229, pages 138–147, Bellingham, WA, 1997. SPIE.

[17] Nathaniel Good, J. Ben Schafer, Joseph A. Konstan, Al Borchers, Badrul Sarwar, Jon Herlocker, and John Riedl. Combining collaborative filtering with personal agents for better recommendations. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, pages 439–446, 1999.

[18] William Hill, Lawrence Stead, Mark Rosenstein, and George Furnas. Recommending and evaluating choices in a virtual community of use. In *Proceedings of ACM CHI'95*, pages 194–201, 1995.

[19] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*, pages 137–142, 1998.

[20] Peter Knees, Elias Pampalk, and Gerhard Widmer. Artist classification with web-based data. In *Proceedings of 5th International Conference on Music Information Retrieval (ISMIR'04)*, pages 517–524, Barcelona, Spain, October 2004.

[21] Wallace Koehler. A longitudinal study of web pages continued: A consideration of document persistence. *Information Research*, 9(2), 2004.

[22] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, 3rd edition, 2001.

[23] Steve Lawrence and C. Lee Giles. Accessibility of information on the web. *Nature*, 400(6740):107–109, 1999.

[24] Beth Logan, Andrew Kositsky, and Pedro Moreno. Semantic analysis of song lyrics. In *Proceedings of IEEE International Conference on Multimedia and Expo 2004*, Taipei, Taiwan, June 2004.

[25] Beth Logan and Ariel Salomon. A music similarity function based on signal analysis. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'01)*, Tokyo, Japan, 2001.

[26] Cory McKay and Ichiro Fujinaga. Automatic genre classification using large high-level musical feature sets. In *Proceedings of 5th International Conference on Music Information Retrieval (ISMIR'04)*, pages 525–530, Barcelona, Spain, October 2004.

[27] Martin F. McKinney and Jeroen Breebaart. Features for audio and music classification. In *Proceedings of the Fourth International Conference on Music Information Retrieval (ISMIR'03)*, pages 151–158, Baltimore, MD, 2003.

[28] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[29] François Pachet and Daniel Cazaly. A taxonomy of musical genres. In *Proceedings of RIAO 2000 Content-Based Multimedia Information Access*, Paris, France, 2000.

[30] François Pachet, Gert Westerman, and Damien Laigre. Musical data mining for electronic music distribution. In *Proceedings of the 1st WedelMusic Conference*, 2001.

[31] Elias Pampalk, Simon Dixon, and Gerhard Widmer. Exploring music collections by browsing different views. In *Proceedings of the Fourth*

International Conference on Music Information Retrieval (ISMIR'03), pages 159–166, Baltimore, MD, 2003. John Hopkins University.

[32] Elias Pampalk, Simon Dixon, and Gerhard Widmer. On the evaluation of perceptual similarity measures for music. In *Proceedings of the Sixth International Conference on Digital Audio Effects (DAFx-03)*, London, UK, September 8-11 2003.

[33] Elias Pampalk, Simon Dixon, and Gerhard Widmer. Exploring music collections by browsing different views. *Computer Music Journal*, 28(3), 2004. In press.

[34] Elias Pampalk, Andreas Rauber, and Dieter Merkl. Content-based organization and visualization of music archives. In *Proceedings of the ACM Multimedia*, pages 570–579, Juan les Pins, France, December 1-6 2002. ACM.

[35] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[36] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[37] Xi Shao, Changsheng Xu, and Mohan S. Kankanhalli. Unsupervised classification of music genre using hidden markov model. In *IEEE International Conference of Multimedia Explore (ICME04)*, Taibei, Taiwan, China, 2004.

[38] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[39] George Tzanetakis, Georg Essl, and Perry Cook. Automatic musical genre classification of audio signals. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR'01)*, 2001.

[40] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[41] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, Chichester, GB, 1998.

[42] Brian Whitman and Steve Lawrence. Inferring descriptions and similarity for music from community metadata. In *Proceedings of the 2002 International Computer Music Conference*, pages 591–598, Göteborg, Sweden, September 2002.

[43] Brian Whitman and Paris Smaragdis. Combining musical and cultural features for intelligent style detection. In *Proceedings of the 3rd International Conference on Music Information Retrieval*, pages 47–52, Paris, France, October 2002.

[44] Wikipedia. Die freie Enzyklopädie. http://de.wikipedia.org.
Stand: 1. November 2004.

[45] Wikipedia. The free encyclopedia. http://www.wikipedia.org.
Stand: 1. November 2004.

[46] Changsheng Xu, Namunu C. Maddage, Xi Shao, and Qi Tian. Musical genre classification using support vector machines. In *Proceedings of the International Conference of Acoustics, Speech & Signal Processing (ICASSP03)*, Hong Kong, China, 2003.

[47] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In D. H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420, Nashville, US, 1997. Morgan Kaufman Publishers, San Francisco, US.

[48] Mark Zadel and Ichiro Fujinaga. Web services for music information retrieval. In *Proceedings of 5th International Conference on Music Information Retrieval (ISMIR'04)*, pages 478–483, Barcelona, Spain, October 2004.

[49] Justin Zobel and Alistair Moffat. Exploring the similarity space. *ACM SIGIR Forum*, 32(1):18–34, 1998.

A Anhang

A.1 Zuordnung von Künstlern zu Genres

Country – Johnny Cash, Willie Nelson, Dolly Parton, Hank Williams, Faith Hill, Dixie Chicks, Garth Brooks, Kenny Rogers, Tim McGraw, Hank Snow, Brooks and Dunn, Lee Hazlewood, Kenny Chesney, Jim Reeves, Roger Miller, Kris Kristofferson

Folk – Bob Dylan, Joni Mitchell, Leonard Cohen, Joan Baez, Townes van Zandt, Pete Seeger, Suzanne Vega, Tracy Chapman, Tim Buckley, Steeleye Span, Woodie Guthrie, Donovan, Cat Stevens, John Denver, Don McLean, Crosby Stills & Nash

Jazz – Miles Davis, Dave Brubeck, Billie Holiday, Duke Ellington, Django Reinhardt, Glenn Miller, Ella Fitzgerald, Louis Armstrong, Nat King Cole, Herbie Hancock, Nina Simone, John Coltrane, Charlie Parker, Count Basie, Thelonious Monk, Cannonball Adderley

Blues – John Lee Hooker, Muddy Waters, Taj Mahal, John Mayall, Big Bill Broonzy, BB King, Howlin' Wolf, Willie Dixon, Blind Lemon Jefferson, Blind Willie McTell, Mississippi John Hurt, T-Bone Walker, Etta James, Lightnin' Hopkins, Otis Rush, Albert King

R'n'B/Soul – James Brown, Marvin Gaye, Otis Redding, Solomon Burke, Sam Cooke, Aretha Franklin, Al Green, The Temptations, The Drifters, Fats Domino, The Supremes, Isaac Hayes, Alicia Keys, Erykah Badu, India Arie, Jill Scott

Heavy Metal/Hard Rock – Iron Maiden, Megadeth, Slayer, Sepultura, Black Sabbath, Anthrax, Alice Cooper, Deep Purple, Def Leppard, AC/DC, Judas Priest, Kiss, Metallica, Pantera, Queensryche, Skid Row

Alternative Rock/Indie – Nirvana, Beck, Smashing Pumpkins, Radiohead, Belle and Sebastian, Alice in Chains, Echo and the Bunnymen, Sonic Youth, Weezer, Pearl Jam, Foo Fighters, Hole, Bush, The Smiths, Depeche Mode, Jane's Addiction

Punk – Patti Smith, Sex Pistols, Sid Vicious, Ramones, Bad Religion, The Clash, Nofx, Dead Kennedys, Buzzcocks, Green Day, Blink 182, Sum 41, The Misfits, Rancid, Screeching Weasel, Pennywise

Rap/HipHop – Eminem, Dr Dre, Public Enemy, Missy Elliot, Cypress Hill, 50 Cent, Run DMC, Grandmaster Flash, 2Pac, Snoop Dogg, Jay-Z, Busta Rhymes, LL Cool J, DMX, Ice Cube, Mystikal

Electronica – Aphex Twin, Daft Punk, Kraftwerk, Chemical Brothers, Fatboy Slim, Basement Jaxx, Carl Cox, Moloko, Paul Oakenfold, Prodigy, Armand van Helden, Moby, Massive Attack, Mouse on Mars, Jimi Tenor, Underworld

Reggae – Bob Marley, Jimmy Cliff, Peter Tosh, Ziggy Marley, Sean Paul, Alpha Blondie, Shaggy, Maxi Priest, Shabba Ranks, UB40, Inner Circle, Desmond Dekker, Capleton, Bounty Killer, Eddy Grant, Black Uhuru

Rock'n'Roll – The Rolling Stones, The Animals, The Faces, The Kinks, The Who, Elvis Presley, Chuck Berry, Little Richard, Jerry Lee Lewis, Buddy Holly, Bo Diddley, Bill Haley, Chubby Checker, The Yardbirds, Carl Perkins, Gene Vincent

Pop – Madonna, Britney Spears, N'Sync, Justin Timberlake, ABBA, Michael Jackson, Janet Jackson, Prince, Spice Girls, Christina Aguilera, Robbie Williams, Nelly Furtado, Avril Lavigne, Jennifer Lopez, O-Town, Shakira

Klassik – Wolfgang Amadeus Mozart, Ludwig van Beethoven, Johann Sebastian Bach, Joseph Haydn, Johannes Brahms, Frederic Chopin, Antonin Dvorak, Gustav Mahler, Franz Schubert, Antonio Vivaldi, Richard Wagner, Herbert von Karajan, Yehudi Menuhin, Georg Friedrich Händel, Tchaikovsky, Giuseppe Verdi

A.2 Typische χ^2-gewichtete Wortlisten

Im Folgenden finden sich typische Wortlisten, wie sie unter Verwendung von Google mit den Zusatztermen *music review* durch Gewichtung mit dem χ^2-Test entstehen. Die Listen, die sich in den Tabellen A.1 bis A.14 finden, sind alle unter Verwendung von 8 Künstlern eines Genres entstanden (die Namen sind der jeweiligen Tabellenbeschreibung zu entnehmen). Dabei ist es wichtig darauf hinzuweisen, dass alle diese Tabellen miteinander in Verbindung stehen, da die jeweiligen Genrerepräsentationen in den angeführten Konstellationen gegeneinander gelernt wurden. Richtigerweise müsste die Beschreibung einer Tabelle auch immer auf alle anderen Genres und deren Künstler verweisen, da diese mitverantwortlich an der Zusammensetzung der wichtigsten 100 Terme sind. Das heißt, dass die nachfolgenden Tabellen als zusammengehörig gesehen werden müssen.

Die Werte der Tabellen sind so skaliert, dass der höchste Wert 100 ist. Worte, die mit * markiert sind, waren Teil der jeweiligen Suchanfragen.

100	*hank	23	gaines	13	chug
75	tonk	23	shania	13	unanswered
75	honky	23	cheatin	13	rodgers
73	*hazlewood	23	outlaw	13	boots
72	nashville	22	ole	13	vince
72	country	21	ropin	13	buryin
66	*parton	21	dunn	13	stonewall
65	bluegrass	21	fences	13	unsaveable
57	*garth	20	cowboys	13	escott
57	*brooks	19	*miller	13	scotia
53	*nelson	19	jennings	12	walkin
53	*dolly	19	gill	12	banjo
46	opry	19	kristofferson	12	tennessee
41	*snow	18	lug	12	jimmie
37	grass	18	lovesick	12	outlaws
35	lefty	18	oklahoma	12	kris
34	merle	17	dollywood	12	clint
34	lonesome	17	scarecrow	12	sundown
33	sparrow	17	patsy	12	partons
33	haggard	17	dang	11	wynette
32	*willie	17	*williams	11	origami
31	cowboy	16	fiddle	11	flatulence
30	jolene	16	ledoux	11	hazlewoods
29	twain	16	tonkin	11	bocephus
28	yearwood	16	acuff	11	husbands
28	trisha	15	rhumba	11	buffalo
27	reba	15	*hill	11	*roger
27	waylon	15	emmylou	11	breathe
27	movin	14	strait	11	krauss
25	frizzell	14	wacka	11	halos
25	mcentire	14	coladas	11	cash
24	mcgraw	14	yodel	11	tn
24	faron	13	cline		
24	tubb	13	toby		

Tabelle A.1: Die 100 Terme mit den höchsten χ^2-Werten für das Genre *Country* definiert durch 8 Künstler (Dolly Parton, Faith Hill, Garth Brooks, Hank Snow, Hank Williams, Lee Hazlewood, Roger Miller, Willie Nelson).

100	*baez	22	jakon	16	tambourine
92	*steeleye	21	songwriters	15	desolation
81	*joni	21	parcel	15	dreamland
78	folk	21	rogues	15	marlene
78	*suzanne	21	schoolyard	15	roads
78	*vega	20	objects	15	teaser
77	*joan	20	*tim	15	zimmerman
67	*mitchell	19	muslim	15	headshots
60	maddy	19	freewheelin	15	pegrum
47	*denver	19	solitude	15	jonimitchell
45	yusuf	19	sefronia	15	deutschendorf
43	moonshadow	19	majiks	15	lorca
43	fairport	19	georgiou	15	colvin
42	*dylan	19	majikat	15	starsailor
39	songwriter	19	rhymer	15	rocky
38	folkie	19	annie	15	homesick
37	a&m	19	*bob	14	woodstock
37	islam	18	poems	14	desire
36	tillerman	18	underwood	14	convention
32	hejira	18	*cat	14	hark
30	d'arbanville	18	troubadour	14	greenwich
28	carthy	18	maggie	14	greetings
28	firecat	17	prayers	13	kemp
28	*stevens	17	froom	13	amelia
28	guthrie	17	blackleg	13	calypso
26	luka	17	genockey	13	folksinger
26	lawns	17	gaudete	13	caramel
25	diner	17	harries	13	moisty
24	singer	17	blacksmith	13	tabor
23	morning	17	changin	13	barleycorn
23	hutchings	16	woody	13	katmandu
23	hissing	16	*buckley	13	demetre
22	frae	16	blowin		
22	majik	16	foreigner		

Tabelle A.2: Die 100 Terme mit den höchsten χ^2-Werten für das Genre *Folk* definiert durch 8 Künstler (Bob Dylan, Cat Stevens, Joan Baez, John Denver, Joni Mitchell, Steeleye Span, Suzanne Vega, Tim Buckley).

100	*ellington	22	monk	13	straighten
82	jazz	22	turk	13	*miller
68	*adderley	22	bop	13	perdido
65	*brubeck	22	oscar	13	rondo
64	*duke	22	verve	12	solos
56	*nat	21	chattanooga	12	hubbard
52	*cannonball	21	serenade	12	waller
48	*armstrong	20	davis	12	stan
46	*parker	20	hines	12	tatum
45	trumpet	20	cornet	12	musicians
44	*louis	19	mingus	12	gershwin
39	saxophonist	19	trumpeter	12	*herbie
38	sax	18	goodman	12	coltrane
37	alto	18	piano	12	ella
37	*cole	17	swinging	12	beige
35	bebop	16	blakey	12	saxophonists
35	swing	16	headhunters	12	corea
34	*charlie	16	mercer	11	satin
34	saxophone	15	wynton	11	pianists
33	pianist	15	clarinet	11	hoagy
32	trombone	15	morello	11	rock
29	basie	15	orchestra	11	signatures
28	bandleader	15	tuxedo	11	1930s
27	gillespie	14	1940s	11	chameleon
27	benny	14	choo	11	junction
26	miles	14	fitzgerald	11	mood
26	quintet	14	henderson	11	40s
25	desmond	14	satchmo	11	improvisation
25	thelonious	14	mulligan	11	marsalis
24	dizzy	14	trombonist	10	gonsalves
23	strayhorn	14	*hancock	10	milestones
23	zawinul	14	savoy	10	evans
23	*glenn	13	rockit		
22	hodges	13	quartet		

Tabelle A.3: Die 100 Terme mit den höchsten χ^2-Werten für das Genre *Jazz* definiert durch 8 Künstler (Cannonball Adderley, Charlie Parker, Dave Brubeck, Duke Ellington, Glenn Miller, Herbie Hancock, Louis Armstrong, Nat King Cole).

100	*hooker	19	johnson	12	raitt
86	*lightnin	19	jefferson	11	bloomfield
78	blues	18	robert	11	*john
71	*muddy	18	harmonica	11	dawkins
63	*hopkins	17	lonnie	11	donto
62	*waters	17	guitarists	11	bobbin
58	*mctell	17	coochie	11	curley
54	mississippi	17	chicago	11	taj
50	bluesman	17	pinetop	10	memphis
48	*albert	17	catfish	10	stormy
42	chess	16	vaughan	10	buddy
41	delta	16	hoochie	10	matriarch
40	bluesmen	16	walker	10	dixon
39	*otis	16	slide	10	mcdowell
37	*etta	16	healer	10	*rush
36	broonzy	16	spann	10	electric
32	mayall	16	mckinley	10	stax
31	*blind	16	crawlin	10	patton
30	*willie	16	brownie	10	lipscomb
28	cray	15	*lee	10	wallflower
26	chillen	15	crosscut	10	scorsese
25	mojo	15	*james	10	vestapol
24	howlin	15	piedmont	10	lomax
23	musselwhite	14	mannish	10	delia
23	statesboro	14	lemon	9	eric
21	thorogood	14	dimples	9	ike
21	bone	14	stevie	9	plantation
20	elmore	14	guitar	9	mama
20	cotton	13	margolin	9	clarksdale
20	*king	13	boogie	9	guitarist
19	lucille	13	mcghee	9	williamson
19	clapton	13	gambler	9	stovall
19	morganfield	12	freddie		
19	*bb	12	sametto		

Tabelle A.4: Die 100 Terme mit den höchsten χ^2-Werten für das Genre *Blues* definiert durch 8 Künstler (Albert King, BB King, Blind Willie McTell, Etta James, John Lee Hooker, Lightnin' Hopkins, Muddy Waters, Otis Rush).

100	*redding	17	cupid	10	bahamadia
96	*aretha	16	chain	10	bartholomew
78	*franklin	16	twistin	9	scherrie
74	*cooke	16	tenderness	9	vandellas
63	*supremes	16	garment	9	berns
61	*erykah	15	soulful	9	karma
60	*badu	15	floy	9	lauryn
58	*otis	15	smokey	9	abkco
58	motown	14	secular	9	mayfield
49	*burke	14	stax	9	gonna
46	*solomon	14	dozier	9	ain't
43	soul	13	chained	9	wilson
42	baduizm	13	d'angelo	9	curtis
42	r&b	13	oldies	9	respect
36	*domino	13	possum	9	birdsong
33	stirrers	13	mary	9	stepchild
33	*fats	13	atlantic	9	appletree
27	temptations	13	fa	9	gul
26	fallin	12	carla	8	poyser
26	heartburn	12	primettes	8	*keys
24	dock	12	singers	8	arms
24	*alicia	12	marvelettes	8	searchin
22	marvin	12	mavis	8	josephine
21	gladys	12	terrell	8	otherside
21	diana	11	sittin	8	lovelight
21	gaye	11	tramp	8	cozier
20	*sam	11	rimshot	8	love
20	gospel	11	cleva	8	atco
19	pickett	11	dionne	8	cookes
19	harlem	11	60s	8	volt
19	ross	11	covay	8	hangin
18	samsonite	11	wexler	8	pitiful
17	copa	10	cropper		
17	blueberry	10	gordy		

Tabelle A.5: Die 100 Terme mit den höchsten χ^2-Werten für das Genre *R'n'B/Soul* definiert durch 8 Künstler (Alicia Keys, Aretha Franklin, Erykah Badu, Fats Domino, Otis Redding, Sam Cooke, Solomon Burke, Supremes).

100	metal	21	turbo	14	wrathchild
59	halford	21	screaming	14	defenders
55	*queensryche	20	jugulator	14	chaos
53	thrash	19	hellion	14	killing
47	*maiden	18	q2k	14	youthanasia
47	heavy	18	geoff	14	rocka
41	*slayer	18	beast	14	powerslave
40	mustaine	18	owens	14	ezrin
39	mindcrime	18	hells	14	furnier
34	painkiller	18	bands	14	coop
33	brutal	18	ballbreaker	14	satanic
29	angus	18	dio	14	gers
28	*megadeth	17	manalishi	14	roorback
28	*judas	17	sanctuary	13	riffing
27	ripper	17	trooper	13	nightmare
27	tate	17	hallowed	13	speed
27	dickinson	17	degarmo	13	awaits
26	hell	17	frontier	13	araya
26	tipton	17	vengeance	13	ozzfest
25	*priest	17	voivod	13	bon
25	*ac	16	ellefson	13	freewheel
25	metallica	16	*dc	13	gods
25	rust	16	hanneman	13	rolla
25	rockenfield	16	bestial	13	janick
25	cavalera	16	operation	12	doom
24	*iron	15	riffs	12	tribe
24	morbid	15	downing	12	band
23	*sepultura	15	sabbath	12	meltdown
23	evil	15	nicko	12	brazilian
23	abyss	15	kisser	12	stained
22	deeds	15	*cooper	12	diabolus
21	roadrunner	14	thunderstruck	11	mcbrain
21	extinction	14	wilton		
21	dragontown	14	lucidity		

Tabelle A.6: Die 100 Terme mit den höchsten χ^2-Werten für das Genre *Heavy Metal/Hard Rock* definiert durch 8 Künstler (AC/DC, Alice Cooper, Iron Maiden, Judas Priest, Megadeth, Queensryche, Sepultura, Slayer).

100	grunge	35	bleach	19	ian
86	*bunnymen	35	*pearl	19	sliver
86	corgan	34	ament	19	wretzky
67	cuomo	34	grohl	19	vig
67	collie	32	gossard	18	courtney
66	cobain	31	disarm	18	seattle
65	mcculloch	31	pumpkin	18	unplugged
65	geffen	30	*beck	18	lips
61	pinkerton	30	*hole	17	*bush
59	*pumpkins	29	smells	17	loser
59	*nirvana	29	seas	17	krist
58	*smashing	28	vitalogy	17	siva
57	adore	28	soundgarden	17	iscariot
51	vedder	27	hash	17	jonas
50	siamese	27	ava	16	evergreen
47	iha	27	godrich	16	wishkah
47	odelay	26	*weezer	16	apologies
47	mutations	25	rivers	15	corduroy
46	kurt	24	ocasek	15	matt
46	gish	24	dgc	15	rossdale
45	maladroit	24	yield	15	aeroplane
44	crocodiles	23	mellon	15	bands
44	nevermind	23	alternative	15	mayonaise
40	sergeant	23	binaural	15	porcelina
39	lithium	23	rhinoceros	15	supermellow
38	utero	23	ballyhoo	15	bedbugs
38	*echo	23	cutter	15	freitas
37	chamberlin	22	pennyroyal	15	electrafixion
37	machina	22	mccready	15	insignificance
37	vultures	22	midnite	14	band
36	infinite	20	incesticide	14	*jam
35	d'arcy	20	novoselic	14	bushleaguer
35	cherub	20	angst		
35	sadness	19	guitars		

Tabelle A.7: Die 100 Terme mit den höchsten χ^2-Werten für das Genre *Alternative Rock/Indie* definiert durch 8 Künstler (Beck, Bush, Echo and the Bunnymen, Hole, Nirvana, Pearl Jam, Smashing Pumpkins, Weezer).

100	punk	18	ska	11	rockaway
64	*sid	17	pinhead	11	byo
61	*pistols	17	thunders	10	apathy
60	*weasel	17	vacant	10	glue
55	*rancid	16	ho	10	fuck
52	*vicious	16	spungen	10	clash
48	matlock	16	queers	10	wreck
46	*ramones	16	*sum	10	gabba
45	anarchy	15	jubilee	10	sniff
42	lydon	15	hellcat	10	redondo
42	punks	15	bands	10	trampin
41	*screeching	15	53rd	10	kkk
41	mclaren	15	valuum	10	nubs
38	ramone	15	sedated	10	hefe
36	bollocks	14	strummer	10	unconsciousness
31	rotten	14	daugherty	9	heroin
30	*nofx	14	whibley	9	dolls
27	*sex	14	sheena	9	lars
27	*patti	14	band	9	deryck
26	blitzkrieg	13	joey	9	ivy
26	swindle	13	brat	9	johnny
25	emo	13	belsen	9	fat
24	epitaph	13	petting	9	buzzcocks
24	errorism	13	grundy	9	winterland
20	julien	13	wolves	9	methadone
20	drublic	12	fury	8	marky
20	lookout	12	iggy	8	bark
19	indestructible	12	rock	8	overdose
18	gung	12	boogada	8	westwood
18	dee	12	bean	8	pistol
18	glen	12	frederiksen	8	nigger
18	infected	12	kaye	8	priestess
18	heebs	11	skate		
18	malcolm	11	hardcore		

Tabelle A.8: Die 100 Terme mit den höchsten χ^2-Werten für das Genre *Punk* definiert durch 8 Künstler (Nofx, Patti Smith, Ramones, Rancid, Screeching Weasel, Sex Pistols, Sid Vicious, Sum 41).

100	rap	23	kurtis	15	dr
89	*2pac	23	shakur	15	darryl
78	*grandmaster	22	busta	15	mixtapes
73	rappers	22	terrordome	14	furious
71	*mystikal	22	epmd	14	whodini
71	hop	22	muggs	14	patiently
69	rapper	21	rhyme	14	nas
65	gangsta	20	raps	14	mc
63	hip	20	yayo	13	terminator
57	*dmc	20	griff	13	obie
48	dre	20	*flash	13	mixx
46	thug	20	shady	13	eminem
44	rhymes	20	outlawz	13	trice
42	tupac	20	snoop	13	tarantula
40	pac	19	*run	13	makaveli
37	*cent	19	jay	13	2pacalypse
37	*cypress	19	sugarhill	12	unit
34	*elliot	19	ya	12	outkast
33	beats	19	dupa	12	ludacris
33	melle	19	da	12	wu
31	adidas	18	explicit	12	ja
30	*enemy	17	wanksta	12	pump
29	*missy	17	sen	12	flow
29	tryin	17	chronic	11	skool
29	flav	17	dj	11	killuminati
29	rapping	17	amaru	11	redman
27	eyez	17	dayz	11	dmx
27	dogg	17	pimp	11	notorious
26	rakim	16	mcdaniels	11	hood
26	supa	16	tha	11	peeps
25	temples	16	amerikaz	11	*hill
24	biggie	16	lil	11	banks
23	revolverlution	15	pe		
23	timbaland	15	cube		

Tabelle A.9: Die 100 Terme mit den höchsten χ^2-Werten für das Genre *Rap/HipHop* definiert durch 8 Künstler (2pac, 50 Cent, Cypress Hill, Grandmaster Flash, Missy Elliot, Mystikal, Public Enemy, Run DMC).

100	techno	23	autoditacker	17	decks
67	electronic	22	dodo	17	oakenfold
65	*underworld	22	synths	16	respoke
64	*armand	22	darren	16	actionist
63	*helden	21	vulvaland	16	cowgirl
59	*moloko	21	bodyrock	16	tahiti
52	beats	21	rockafeller	15	trainspotting
49	*fatboy	21	dancefloor	15	jumbo
45	*cox	21	sasha	15	porcelain
41	electronica	20	*carl	15	howlett
39	toma	20	tiesto	15	roisin
38	dj	20	skank	14	housemartins
37	*prodigy	20	acid	14	remix
36	idiology	20	phuture	14	palookaville
35	slippy	20	house	14	upstairs
34	*slim	19	daft	14	jockey
34	trance	19	breakbeat	13	remixes
34	dance	19	rave	13	*mars
34	andi	19	funk	13	sonig
32	*moby	19	*mouse	13	astralwerks
31	beaucoup	19	jilted	13	autechre
30	outgunned	19	idm	13	dyk
30	v2	18	jaxx	13	timo
29	outnumbered	18	moaner	13	turntables
29	djs	18	firestarter	13	fish
29	connector	18	emerson	12	bootsy
29	electro	18	puritans	12	toughest
26	werner	18	infants	12	flint
25	niun	18	hyde	12	push
24	ambient	17	loops	12	wipe
24	niggung	17	niobe	12	tong
23	ibiza	17	iaora	12	jbo
23	orbital	17	mix		
23	nkishi	17	liam		

Tabelle A.10: Die 100 Terme mit den höchsten χ^2-Werten für das Genre *Electronica* definiert durch 8 Künstler (Armand van Helden, Carl Cox, Fatboy Slim, Moby, Moloko, Mouse on Mars, Prodigy, Underworld).

100	reggae	15	*bob	9	duckie
83	*marley	15	dread	9	sinsemilla
68	dancehall	15	ganja	9	inna
57	jamaica	15	riddims	9	elephant
50	*shabba	15	maytals	9	trojan
48	wailers	15	selassie	9	gong
47	*uhuru	14	boombastic	9	ting
46	*capleton	14	dem	9	*grant
45	*ziggy	14	greensleeves	9	tubby
45	*ub40	14	rastaman	9	ghetto
45	*shaggy	14	vibes	8	tra
39	jamaican	13	bunny	8	sponji
37	jah	12	yellowman	8	caan
31	rasta	12	zion	8	dennis
31	buju	12	pon	8	cedella
30	sizzla	12	augustus	8	labour
30	banton	11	aswad	8	damian
29	kingston	11	yuh	8	tuff
27	beenie	11	bounty	8	rastafarian
24	ragga	11	cliff	8	minott
24	ras	11	makers	8	nuh
23	tosh	11	loverman	8	gal
23	dub	11	beres	8	mikey
23	*ranks	11	soca	8	wicked
21	riddim	10	maxi	8	kaya
19	vibration	10	luciano	8	jo'anna
18	*eddy	10	lexxus	7	africa
18	isaacs	10	hotshot	7	demus
18	wailer	10	sly	7	tok
18	babylon	10	prophet	7	abyssinians
17	dragonfly	10	gregory	7	eek
17	vp	10	spear	7	puma
15	toots	9	exodus		
15	cocoa	9	dunbar		

Tabelle A.11: Die 100 Terme mit den höchsten χ^2-Werten für das Genre *Reggae* definiert durch 8 Künstler (Black Uhuru, Bob Marley, Capleton, Eddy Grant, Shabba Ranks, Shaggy, UB40, Ziggy Marley).

100	*ludwig	49	sonatas	29	conducting
99	concerto	49	haydn	29	overture
91	*wagner	48	amadeus	28	quartets
90	*brahms	47	*antonio	28	largo
85	mozart	46	baroque	28	soloists
84	symphonies	44	mendelssohn	27	di
83	*johannes	44	conductors	27	rossini
79	op	43	tchaikovsky	27	*händel
79	symphony	43	*beethoven	27	debussy
76	orchestra	42	sonata	26	classical
74	composers	42	works	26	das
72	bach	40	karajan	25	liszt
70	conductor	40	schumann	25	libretto
69	handel	40	allegro	24	eroica
68	*georg	38	andante	24	von
67	vienna	36	cello	24	molto
66	*mahler	35	arias	24	grammophon
66	*gustav	35	soprano	24	italian
65	philharmonic	35	adagio	23	giovanni
64	composer	35	frideric	23	tristan
64	*giuseppe	35	composed	23	mezzo
62	opera	33	choral	23	hungarian
61	johann	33	puccini	23	minor
61	*verdi	33	conducted	23	ravel
60	*vivaldi	32	romantic	23	berlioz
60	*friedrich	31	chamber	23	giacomo
59	operas	31	nos	23	choir
59	schubert	31	deutsche	23	philharmonia
56	concertos	31	movements	23	anton
55	violin	31	chopin	22	berlin
55	strauss	30	naxos	22	symphonic
52	*antonin	30	scherzo	22	barenboim
52	wolfgang	29	orchestral		
51	*dvorak	29	german		

Tabelle A.12: Die 100 Terme mit den höchsten χ^2-Werten für das Genre *Klassik* definiert durch 8 Künstler (Antonin Dvorak, Antonio Vivaldi, Georg Friedrich Hndel, Giuseppe Verdi, Gustav Mahler, Johannes Brahms, Ludwig van Beethoven, Richard Wagner).

100	*yardbirds	26	busey	14	sods
96	*kinks	25	lola	14	lukather
74	*berry	24	baba	14	*elvis
65	*holly	24	*buddy	14	idan
60	mccarty	24	kink	13	pete
59	diddley	23	keith	13	rodford
58	roll	23	burdon	13	kronikles
48	townshend	23	stones	13	richwine
47	dreja	22	nazz	13	happenings
45	invasion	21	british	13	fooled
44	crickets	21	sue	12	hillbillies
43	*animals	20	blueswailing	12	baxter
42	daltrey	20	o'riley	12	rash
42	maybellene	20	beaty	12	psychedelia
39	lubbock	20	lennon	12	debris
38	relf	20	samwell	12	sgt
38	*chuck	19	pinball	12	scotty
38	bo	18	preservation	11	saboteur
38	shapes	18	gouldman	11	gypie
36	entwistle	18	vai	11	chirping
36	davies	18	elv1s	11	aron
33	sideways	17	muswell	11	roustabout
33	*faces	17	smokestack	11	gunne
33	goode	17	clovis	11	kontroversy
32	*presley	17	rock	11	60s
31	berrys	17	bopper	10	rockabilly
30	that'll	17	birdland	10	pie
30	mclagan	16	jailhouse	10	schoolgirl
28	everly	16	rave	10	celluloid
27	waterloo	15	graceland	10	fade
27	beatles	15	avory	10	sessions
27	quadrophenia	15	reminiscing	10	kenney
26	peggy	14	roger		
26	valens	14	guitarist		

Tabelle A.13: Die 100 Terme mit den höchsten χ^2-Werten für das Genre *Rock'n'Roll* definiert durch 8 Künstler (Buddy Holly, Chuck Berry, Elvis Presley, The Animals, The Faces, The Kinks, The Who, Yardbirds).

100	*aguilera	27	pop	18	marmalade
81	*shakira	27	posh	18	wants
78	*lopez	26	sexed	18	jerkins
70	*sync	26	halliwell	18	dance
69	*jennifer	26	chasez	17	ego
65	*christina	26	*girls	17	lens
61	*spice	26	bunton	17	thriller
59	genie	26	justin	17	xtina
56	escapology	26	colombian	17	sexy
51	spiceworld	25	ames	17	aqui
50	damita	25	spanish	16	lebanese
49	ladrones	24	emma	16	estn
49	*jackson	24	jc	16	anaconda
46	*janet	23	ciega	16	sexhibition
46	laundry	23	asi	16	objection
45	dirrty	22	wannabe	16	feelin
45	mariah	22	tearin	16	mj
44	donde	22	melanie	16	ginger
42	estan	22	estefan	16	timberlake
40	descalzos	21	octavo	16	jo
38	geri	21	dancer	16	ballads
37	pies	21	suerte	15	sexual
37	sporty	21	sordomuda	15	bojangles
36	britney	21	slolove	14	knebworth
35	*robbie	21	actress	14	mtv
33	celebrity	20	rope	14	moist
33	ojos	20	jacksons	14	sexuality
31	reflejo	20	landed	14	quedes
31	attached	19	chisholm	14	dnde
29	invincible	19	boyzone	14	columbian
28	estoy	19	carey	14	mebarak
28	spears	19	vas	14	creo
27	madonna	19	noche		
27	selena	19	latin		

Tabelle A.14: Die 100 Terme mit den höchsten χ^2-Werten für das Genre *Pop* definiert durch 8 Künstler (Christina Aguilera, Janet Jackson, Jennifer Lopez, Michael Jackson, N'Sync, Robbie Williams, Shakira, Spice Girls).

A.3 Listen der Stop-Words

- http://www.ranks.nl/tools/stopwords.html (Google Stopwords)

  I a about an are as at be by com de en for from how in is it la of on or that the this to was what when where who will with und the www

- http://unesdoc.unesco.org/ulis/stopwords.html

  A ABOUT AGAIN AI AINSI AL ALL ALSO AM AN AND ARE AS AT AU AUTRE AUTRES AUX AVEC AYANT B BE BEFORE BOTH BOX BY C CAN CAR CE CERTAIN CERTAINE CERTAINES CERTAINS CES CET CETTE CHACUN CHACUNE CHAQUE CO CON CORP COULD D DANS DE DEL DELA DEPUIS DER DES DIVERS DONC DONT DU E EACH EL ELLE ELLES EME EN ES ET ETC EVERY F FOR FROM FUR G GIVEN GMBH H HAS HAVE HE HORS HOW I IL IN INC INTO IS IT ITS J K L LA LAS LAST LE LES LEUR LEURS LOS LTD M MA MAIS MANY ME MES MIS MON MORE N NE NI NO NON NOT O OF OFF ON ONT OR OTHER OU OUT OUTSIDE OVER P PAR PARA PAS PLUS POR POUR PRE PTY Q QU QUE QUEL QUELQUE QUELQUES QUELLE QUELLES QUELS QUI R S SA SARL SE SES SHE SHOULD SINCE SO SONT SOUS SUB SUCH SUR T TA TE TES THAN THAT THE THEIR THERE THESE THEY THOSE THROUGH THROUGHOUT TO TOO TOUS TOUT TOUTES U UN UND UNE USE USES V W WELL WERE WHAT WHEN WHERE WHICH WHO WITH X Y Z ZU ZUM ZUR

- http://thomas.loc.gov/home/stopwords.html

  a about above across adj after afterwards again against albeit all almost alone along already also although always among amongst an and another any anyhow anyone anything anywhere are around as at be became because become becomes becoming been before beforehand behind being below beside besides between beyond both but by can cannot co could down during each eg either else elsewhere enough etc even ever every everyone everything everywhere except few first for former formerly from further had has have he hence her here hereafter hereby herein hereupon hers herself him himself his how however ie if in inc indeed into is it its itself last latter latterly least less ltd many may me meanwhile might more moreover most mostly much must my myself namely neither never nevertheless next no nobody none noone nor not nothing now nowhere of off often on once one only onto or other others otherwise our ours ourselves out over own per perhaps rather same seem seemed seeming seems several she should since so some somehow someone something sometime sometimes somewhere still such than that the their them themselves then thence there thereafter thereby therefor therein thereupon these they this those though through throughout thru thus to together too toward towards under until up upon us very via was we well were what whatever whatsoever when whence whenever whensoever where whereafter whereas whereat whereby wherefrom wherein whereinto whereof whereon whereto whereunto whereupon wherever wherewith whether which whichever whichsoever while whilst whither who whoever whole whom whomever whomsoever whose whosoever why will with within without would xsubj xcal xauthor xother xnote yet you your yours yourself yourselves

- Java Reserved Words

  abstract boolean break byte case catch char class const continue default delete do double else extends false final finally float for function goto if implements import in instanceof int interface long native new null package private protected public return short static super switch synchronized this throw throws transient true try typeof var void while with

- VBScript Reserved Words

  And As Boolean ByRef Byte ByVal Call Case Class Const Currency Dim Do Double Each Else ElseIf Empty End EndIf Enum Eqv Event Exit False For Function Get GoTo If Imp Implements In Integer Is Let Like Long Loop LSet Me Mod New Next Not Nothing Null On Option Optional Or ParamArray Preserve Private Public RaiseEvent ReDim Rem Resume RSet Select Set Shared Single Static Stop Sub Then To True Type TypeOf Until Variant Wend While With Xor

- JScript Reserved Words

  break case catch class const continue debugger default delete do else enum export extends false finally for function if import in instanceof new null return super switch this throw true try typeof var void while with

- HTML Reserved Words

  `a abbr acronym address applet area b base basefont bdo big blockquote body br button caption center cite code col colgroup dd del dfn dir div dl dt em fieldset font form frame frameset h1 h2 h3 h4 h5 h6 head hr html i iframe img input ins isindex kbd label legend li link map menu meta noframes noscript object ol optgroup option p param pre q s samp script select small span strike strong style sub sup table tbody td textarea tfoot th thead title tr tt u ul var nbsp`

A.4 Ergebnisse aus Knees et al., 2004

Zu Vergleichszwecken sind hier die wichtigsten Resultate, die in [20] publiziert wurden, abgedruckt.

Real Class	Country	Folk	Jazz	Blues	R&B/S	HM	A/I	Punk	R/HH	Elctr	Reggae	Class	R&R	Pop
Country	91 ±7	2 ±6					2 ±4			1 ±3			2 ±4	
Folk	11 ±8	55 ±22		1 ±3	5 ±8	2 ±5	9 ±7	1 ±3		1 ±3			15 ±16	1 ±3
Jazz		1 ±2	93 ±4		4 ±4								1 ±2	1 ±3
Blues				88 ±14	4 ±6								8 ±14	
R&B/Soul				1 ±3	75 ±12				4 ±7				14 ±9	6 ±8
Heavy Metal		2 ±4			1 ±2	75 ±16	9 ±11	6 ±11					2 ±4	6 ±8
Alt/Indie		1 ±3			1 ±3	7 ±10	57 ±23	13 ±14		12 ±9			1 ±4	8 ±10
Punk		1 ±2			1 ±2	8 ±11	11 ±12	69 ±14						10 ±7
Rap/HipHop					1 ±3				91 ±9	4 ±4				3 ±6
Electro							5 ±7	1 ±2		94 ±8				
Reggae					3 ±6		1 ±4		4 ±5	2 ±4	86 ±9		1 ±2	2 ±4
Classic												100 ±0		
Rock n' Roll	4 ±4	3 ±6		2 ±5	2 ±6	5 ±4	8 ±10	2 ±5		1 ±3			74 ±15	1 ±3
Pop		1 ±2			2 ±5	1 ±2	5 ±6	1 ±3	1 ±3	2 ±4			1 ±3	87 ±8

Classification Results

Abbildung A.1: Confusion Matrix der Klassifikationen unter Benutzung einer SVM mit Google, den Suchtermen *music review*, C_{100} und 4 Künstlern pro Kategorie zum Training. Beschriftung wie in Abbildung 6.1

	Google					
	`music genre style`			`music review`		
	t2	t4	t8	t2	t4	t8
SVM C_{100}	70±3.7	80±2.9	86±2.3	71±4.3	81±3.1	87±3.0
SVM C_{200}	67±3.8	78±3.0	85±2.7	68±4.3	79±3.3	86±2.6
SVM C_{∞}	67±3.8	77±3.1	84±3.0	69±4.7	79±3.5	84±2.7
3-NN C_{∞}	54±6.9	66±4.6	73±3.8	56±4.6	68±4.3	74±3.3
7-NN C_{∞}	39±7.7	67±3.7	75±3.0	43±8.2	68±4.5	77±3.7

Tabelle A.15: Resultate für die Suchmaschine Google aus [20]. Für Suchanfragen, die nur aus dem Namen des Künstlers ohne Zusatzterme bestehen, wurden keine Resultate erhoben. Ansonsten Bezeichnungen wie in Tabelle 6.1.

	Yahoo!					
	`music genre style`			`music review`		
	t2	t4	t8	t2	t4	t8
SVM C_{100}	61±4.3	72±3.1	79±2.9	65±4.9	78±2.9	87±2.6
SVM C_{200}	56±4.4	69±3.3	78±3.0	62±4.5	75±3.2	85±3.2
SVM C_{∞}	56±4.8	67±3.7	74±3.1	65±4.9	76±2.1	85±3.1
3-NN C_{∞}	39±6.1	51±5.6	58±3.9	51±6.9	62±4.7	71±3.7
7-NN C_{∞}	31±9.0	51±5.5	62±3.7	40±8.5	63±5.2	73±3.7

Tabelle A.16: Resultate für die Suchmaschine Yahoo! aus [20]. Für Suchanfragen, die nur aus dem Namen des Künstlers ohne Zusatzterme bestehen, wurden keine Resultate erhoben. Ansonsten Bezeichnungen wie in Tabelle 6.1.