A Deeper Look into Web-based Classification of Music Artists

Peter Knees, Markus Schedl, and Tim Pohle

Dept. of Computational Perception, Johannes Kepler University Linz, Austria
peter.knees@jku.at

Abstract. We extensively evaluate a Web-based approach to automatic classification of musical artists into genres and explore several directions for improvements. The investigated steps to advance the current approach include the finding of optimal query terms to initially retrieve a high fraction of relevant pages from the Web and furthermore the development of a filtering technique to identify and exclude irrelevant pages among the retrieved ones. Since these extensions do not yield any substantial improvement, we propose an alternative approach that performs equally well in the task of genre classification but reduces the effort to be made to a minimum. Based on the gained insights we can conclude that Web-based artist classification operates mainly on the basis of genre-relevant proper nouns and thus can be considered to be a special form of co-occurrence analysis.

1 Introduction

One of the central tasks of Music Information Retrieval is automatic organization of music. Although there are legitimate reservations (see, e.g. [1]), a very common approach is to classify music into genres and styles. Genres can serve as references in a discourse about music or they can be useful to discover similar artists. Furthermore, genres are also essential for evaluation of similarity measures, since they can serve as ground truth in cases where no other ground truth (e.g. opinions of humans) is available.

In general, automatic genre classification is based on audio signals (e.g. [2, 3]). Classification accuracies achieved with these approaches are far from being perfect, but in relation to judgments from different listeners they offer acceptable results [4]. In [5], we presented an approach based solely on features derived from Web data. Using artist-related Web pages suggested by Google, we extract $tf \times idf$ (*term frequency* $\times$ *inverse document frequency*) features which we use to classify the associated artists. Evaluation on a set of 224 artists from 14 genres gave classification accuracy values of up to 87%. Given the simplicity of the approach, these results suggest that possible improvements could be achieved by proper selection of parameters like query terms for acquisition of Web pages or usage of page filters to reduce noise (i.e. unrelated Web pages) in the data.

In this paper we evaluate the approach presented in [5] on a larger basis. Instead of one we use three genre taxonomies with different characteristics to get better insights into the behaviour of the approach. Furthermore, we aim at finding optimal query terms analytically and constructing a page filter from a set of Web pages manually labeled as either "noise" or "informative" to improve the approach (Sections 3.2 and 3.3). Since the obtained results are rather disappointing, i.e. the incorporated methods do not seem to improve accuracy of genre classification, we opt to uncover the reasons for these results by taking a deeper look at the extracted features. Based on the gained insights we can conclude that the categorization of musical artists based on Web-data is in fact another variant of co-occurrence analysis with genre-specific proper nouns. Finally, this finding is further supported by proposing a simplified approach which yields nearly equally good results (Section 4).

2 Related Work

As explained in Section 1, we will focus on the evaluation of our artist classification approach presented in [5]. However, before discussing this method in detail, we want to briefly describe related work.

2.1 Alternative Techniques

In [6], Whitman and Lawrence first used weighted term lists to describe artists and to calculate artist similarity. By querying a search engine with the name of the artist and the additional keywords "music" and "review" they aim at retrieving artist-related textual material from which they extract unigrams, bigrams, and noun phrases. Using the best setting, they achieve 88% accuracy at evaluation against artist similarity from the *All Music Guide*. In [7], Baumann and Hummel could improve this approach slightly by filtering Web pages. In [8], we estimated artist similarity based on the number of Web pages they co-occur on. Evaluation on the set from [5] yielded accuracy of about 85%. Geleijnse and Korst [9] opted for an unsupervised approach that predicts genres via text pattern matching. In very recent work, Levy and Sandler [10] demonstrated the usage of community-based tags (made available by last.fm[1]) also for genre classification.

2.2 Evaluated Approach

For our artist to genre classification approach, we perform similar steps as Whitman and Lawrence [6]. After querying the search engine, we retrieve the 50 top-ranked Web pages, remove all HTML markup tags, and remove stop words. For each artist a and each term t appearing in the retrieved pages, we count the number of occurrences of term t (term frequency tf_{ta}) and the number of pages the term occurred in (document frequency df_t). These are combined using the

[1] `http://www.last.fm`

term frequency × inverse document frequency ($tf \times idf$) function [11]. The term weight per artist is computed as

$$w_{ta} = \begin{cases} (1 + \log_2 tf_{ta}) \log_2 \frac{N}{df_t}, & \text{if } tf_{ta} > 0, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where N is the total number of pages retrieved. Since the goal is solely genre classification, the whole task can be treated as text categorization problem. Thus, the additional information from the genre assignments can be utilized to reduce the dimensionality of the feature space by performing term selection. To this end, the χ^2 test is applied. The χ^2 test measures the independence of a term t from a category c and quantifies therefore the ability of this term to discriminate between a category and all others (for a detailed discussion of the χ^2 test and its role in text categorization see e.g. [12]) and is calculated as

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

where A is the number of documents in c which contain t, B the number of documents not in c which contain t, C the number of documents in c without t, D the number of documents not in c without t, and N is the total number of retrieved documents. The 100 terms with highest χ^2 value per category are joined into a global list that determines the remaining feature dimensions. Finally, cosine normalization is performed on the feature vectors.

3 Enhancement Approaches

In this section we will describe our extensive evaluation of the approach from [5]. First, we will briefly discuss the characteristics of the different genre taxonomies we use for evaluation. In the next section, we aim at finding "optimal" constraining terms for the search engine queries. In Section 3.3 we try to improve classification accuracy by rejecting uninformative Web pages. Finally, we discuss the obtained results and investigate possible reasons.

3.1 Test Collections

The test taxonomy used in [5], which we will call *c224a*, consists of 224 very popular artists from 14 broad and well known genres. Clear advantages of this taxonomy are the easy comprehensiveness of assignments of artists to genres and the balanced number of artists in the genres (each genre contains 16 artists). On the other hand, results could be too optimistic due to the popularity and thus the nearly unexhaustable number of available Web pages for each contained artist. For this reason, we also included two other taxonomies from the literature, namely the *uspop2002* taxonomy used e.g. in [13] and the so called *in-house* collection used in [14] which we will call *c103a* in the following, since it contains

Taxonomy *c224a* (14 genres)

genre	no. of artists
Electronica	16 (7.4%)
Country	16 (7.4%)
Folk	16 (7.4%)
Punk	16 (7.4%)
Rap/HipHop	16 (7.4%)
Heavy Metal	16 (7.4%)
Rock'n'Roll	16 (7.4%)
Reggae	16 (7.4%)
Classical	16 (7.4%)
Blues	16 (7.4%)
Alternative Rock/Indie	16 (7.4%)
Jazz	16 (7.4%)
RnB/Soul	16 (7.4%)
Pop	16 (7.4%)
total	224 (100%)
baseline	7.4%

Taxonomy *uspop2002* (10 genres)

genre	no. of artists
Jazz	2 (0.5%)
Electronica	19 (5.8%)
Rap	28 (7.0%)
New Age	3 (0.8%)
Vocal	2 (0.5%)
Rock	293 (73.3%)
R&B	34 (8.5%)
Reggae	2 (0.5%)
Country	13 (3.3%)
Latin	4 (1.0%)
total	400 (100%)
baseline	73.3%

Taxonomy *c103a* (22 genres)

genre	no. of artists
Bossa Nova	4 (3.9%)
Heavy Metal/Thrash	5 (4.8%)
Downtempo	4 (3.9%)
Melodic Metal	5 (4.8%)
Blues	4 (3.9%)
Celtic	5 (4.8%)
Eurodance	6 (5.8%)
Folkrock	5 (4.8%)
Jazz	4 (3.9%)
Death Metal	4 (3.9%)
Trance	5 (4.8%)
Acid Jazz	4 (3.9%)
Reggae	3 (2.9%)
Punk	6 (5.8%)
A cappella	4 (3.9%)
German HipHop	6 (5.8%)
DnB	5 (4.8%)
Italian	5 (4.8%)
Jazz Guitar	5 (4.8%)
Hardcore Rap	6 (5.8%)
Trance2	4 (3.9%)
Electronica	4 (3.9%)
total	103 (100%)
baseline	5.8%

Table 1. Distribution of genres in the taxonomies used for evaluation.

103 artists. Due to a high number of in some cases very similar genres populated with only few, not always well known artists, *c103a* is a particularly difficult taxonomy. More detailed characteristics of the taxonomies can be found in Table 1.

3.2 Optimizing Query Constraints

Our assumption is that we could improve accuracy and robustness by incorporating only Web pages that contain valuable information. In the best case, we could formulate the search engine queries in a such a manner that the number of uninformative (noisy) result pages is negligible. More precisely, we aim at finding certain terms that occur frequently on "informative" pages and rarely on "uninformative" pages to constrain the result space by adding these terms to the query. Since we cannot define if a Web page is "informative" formally, we labeled a set of pages manually and induced the relevant terms from this information. To this end, we collected 35 artists covering a broad musical spectrum who do not occur in any of the taxonomies. We downloaded the first 100 Web pages that were returned for queries consisting of the artist names enhanced only by the term "music". From these Web pages we sampled 20 per artist randomly, leading to an overall set of 700 Web pages. This set was presented to 3 "experts" who had to judge whether each page contains valuable information about music and the artist in their personal opinion or just links, commercials, ringtones or other irrelevant content. To prevent the introduction of a subjective bias on the assigned labels, for further experiments, we only incorporated pages that were judged equally by all 3 experts. This reduced the size of the page set to 538, from which 340 pages were rated negatively and 198 positively ("informative").

Since the only (or at least the major) criterion for a search engine to include a page in the result set is the occurrence of the query terms, we extracted lists of occurring terms for each document (document frequency). To measure the discriminatory potential of each term, we performed the χ^2 test. Furthermore, we calculated the χ^2 value for all pairs of terms. Since we can not only constrain search results by requiring certain terms to occur, but also by prohibiting the occurrence of terms, for each pair of terms, we tested all 4 combinations ($+term1 + term2$, $+term1 - term2$, $-term1 + term2$, $-term1 - term2$). The resulting ranking of the most promising terms pointing to positively rated pages can be found in Table 2.

It can be seen that the occurrence of the terms "like" and "work" indicates useful pages as well as the absence of terms like "mp3", "download", "videos", or on-line store vocabulary like "cart", "prices", or "login". We assume the high importance for the term "like" to be caused by phrases similar to "People who like artist A, also like artist B" and also phrases like "Band C sounds like Band D", which occur frequently on informative pages.

To systematically evaluate the performance of the statistically derived query terms in conjunction with our genre classification method, we conduct 50-fold cross validation on all 3 taxonomies. For these experiments we use a total of 10 different query settings, namely "artist name +music", "artist name +music" augmented by "+review", by "+genre +style", by "+biography", and by the 6 top ranked additional terms from Table 2. For classification we used *Support Vector Machines (SVM)* and the *Nearest Neighbor classifier (NN)*. Results can be found in Table 3.

From our experiments we can see that the least restricted query setting "+music" yields the best results throughout all experiments. Furthermore, queries containing the term "like" perform well, even better than those suggested in previous work ("+music +review" and "+music +genre +style"). These results justify our approach to find better suited query terms analytically. However, even if some of our proposed query settings outperform those from the literature, best

terms/term combinations	χ^2 value
+like -mp3	0.278
+like	0.277
+like -videos	0.272
+work -prices	0.271
+work -mp3	0.269
+work -services	0.268
+like -download	0.267
+like -tickets	0.265
+like -cart	0.262
+like -login	0.261
+like +time	0.261
+work	0.260
+like -prices	0.260
+work -format	0.259
+work -health	0.258
+like +people	0.257

Table 2. Terms with highest χ^2 value to discriminate between informative and uninformative pages. Only terms or combinations of terms that result in positive rated Web pages are given.

results are achieved if no other constraints than "+music" are used, making a discussion about the helpfulness of our method obsolete.

Regarding the different taxonomies, we see that the *c224a* set is not a big challenge for the approach in general. We can expect accuracies above 90% on this taxonomy. The approach also yields high accuracies on the *uspop2002* set (between 85% and 90%). For the *c103a* set, NN consistently yields better results than SVM. This seems to be a symptom of the collection's structure consisting of many categories with only few examples, making the learning of concepts impractical for SVMs.

3.3 Page Filtering

Another approach to reduce the number of noisy pages is the application of a page filter. An advantage over the query-constraint-approach is the possibility to incorporate additional information, e.g. features based on the structure of HTML pages. A drawback is the necessity to download every page to judge its quality. This consumes time and internet bandwidth.

In [7], page filters are successfully incorporated. These filters use arbitrarily chosen features like page length or the number of words in tables to reject certain pages. For our purpose we make again use of the manual classification of Web pages into *useful* and *useless* pages. For each rated Web page, we calculate cosine normalized $tf \times idf$ vectors. To exploit information about the structure of the pages, we calculate the *tag frequency distribution* as proposed in [15]. Using

	c224a		c103a		uspop2002	
	SVM	NN	SVM	NN	SVM	NN
+music	95.69	93.90	65.00	73.00	89.75	87.25
+music +review	92.69	83.40	60.00	70.00	86.50	85.25
+music +genre +style	90.90	89.10	58.00	63.66	87.25	85.75
+music +biography	91.19	84.70	58.33	68.66	89.00	80.75
+music +like -mp3	92.70	87.80	57.66	72.00	88.50	86.00
+music +like -videos	92.70	86.30	60.00	73.00	88.75	87.50
+music +like	94.90	91.99	59.66	72.66	89.25	85.00
+music +work -prices	89.99	83.20	52.33	59.00	86.50	84.25
+music +work -mp3	89.09	81.00	58.33	62.00	86.50	82.75
+music +work -services	89.49	83.70	56.66	57.00	87.50	83.75

Table 3. Classification results on 3 genre taxonomies using 10 different query settings. The given values are classification accuracies obtained via 50-fold cross validation (values in percent).

Algorithm 1 Page filtering scheme learned by JRip. "Informative" pages belong to class 1, useless pages to class 0.

if just >= 0.055528 and two >= 0.051821 **then**
  ⇒ class 1
else if <p> >= 0.03515 and <i> >= 0.042748 **then**
  ⇒ class 1
else if <p> >= 0.04258 and life >= 0.050582 **then**
  ⇒ class 1
else if work >= 0.075633 **then**
  ⇒ class 1
else if album >= 0.083651 and review >= 0.111605 and privacy <= 0.071766 **then**
  ⇒ class 1
else
  ⇒ class 0
end if

the combined $tf \times idf$ and tag frequency distribution vectors as features, we train a learning algorithm to distinguish between noisy and informative pages.

For our page filter we decided to train a model based on the inductive rule learner RIPPER (we use the WEKA implementation JRip). We expect a classification accuracy of around 83% from the resulting model (estimated via 10-fold cross validation on the training set). Although more sophisticated learning methods like SVMs are capable of reaching nearly 90%, we decided to use a rule learner because the resulting classification scheme is much more intuitive. The resulting rule set for filtering pages can be found in Algorithm 1.

We see that the term "work" tends to occur on valuable pages (cf. Section 3.2). Also occurrences of "album" and "review" indicate useful information. Furthermore, we can see that structural information (tag frequencies) is incorporated into the filter. To evaluate the performance of artist classification with

		c224a				*c103a*				*uspop2002*			
		unfiltered		*filtered*		*unfiltered*		*filtered*		*unfiltered*		*filtered*	
	pg.	SVM	NN	SVM	NN	SVM	NN	SVM	NN	SVM	NN	SVM	NN
+music	10	92.07	91.12	92.49	86.60	66.81	72.63	64.99	72.00	86.74	84.75	87.50	83.75
	25	92.01	86.62	95.61	91.14	68.72	71.63	61.00	66.90	87.25	85.74	88.24	83.75
	50	95.17	91.95	93.41	91.12	64.81	72.81	60.09	65.81	89.00	86.25	87.50	84.75
+music	10	92.92	85.23	92.03	83.91	62.09	66.90	57.18	63.18	86.50	84.49	88.00	84.75
+review	25	93.79	87.47	91.58	85.71	60.18	66.90	54.18	72.81	88.75	86.50	86.99	85.25
	50	92.92	83.85	92.01	81.24	56.27	64.18	52.45	68.09	85.50	87.00	85.00	84.50
+music	10	86.20	80.79	83.03	79.94	52.18	54.45	55.18	55.45	84.25	79.49	81.99	80.49
+genre	25	90.25	83.10	90.23	84.44	56.18	62.90	54.18	60.09	86.25	82.49	87.00	83.00
+style	50	92.47	88.02	89.90	85.29	56.18	63.81	51.36	65.72	86.75	85.50	86.75	86.00
+music	10	93.33	87.01	92.01	88.37	68.81	74.72	61.27	66.99	87.50	85.74	88.00	84.49
+like	25	93.81	92.47	92.92	90.67	67.90	72.72	58.18	70.81	88.24	85.00	87.50	85.00
	50	93.37	90.27	92.94	86.60	60.00	72.63	52.27	66.72	88.00	85.00	86.75	83.75

Table 4. Classification results on 3 genre taxonomies using 4 different query settings. The given values are classification accuracies obtained via 10-fold cross validation (values in percent).

rejection of pages regarded useless, we again perform cross validation on four query settings. Another assumption we want to test is if the number of necessary pages to extract robust artist features decreases when using a page filter. Thus, we perform each experiment 6 times, using 10, 25, or 50 unfiltered pages, as well as 10, 25, or 50 filtered pages (if available in the first 100 pages) for feature calculation. The results of our evaluation can be found in Table 4.

The most obvious finding is the inconsistency of results caused by page filtering. In opposition to our expectations, the filtering of Web pages leads to a loss of accuracy in many cases. The most plausible explanation for this is that the classification approach relies on factors that were not taken into consideration by our experts at rating stage. The "informativeness" of the pages was judged from a human point of view, leading thus not necessarily to better results or features for the automatic classification.

Based on the results reported by Baumann and Hummel [7], it is unexpected not to yield improvements by incorporating page filtering. By excluding very long pages and eliminating sections that do not seem to contain meaningful text, as well as removing pages that do not contain (enough of) the original query terms in URL and/or title, their genre classification approach yields improved results. A possible explanation for this difference may be that their filtering approach is very targeted at identifying Web pages that contain music reviews which are then further processed by means of *Natural Language Processing* techniques, i.e. *Part-of-Speech Tagging*, to extract meaningful nouns or adjectives or combinations thereof. Since feature extraction and classification scheme are rather different from our approach, the findings from [7] may not be directly applicable to our approach. To further investigate these differences, the effects of both filtering approaches would have to be evaluated on both classification approaches.

3.4 Discussion

On the basis of our experiments, we can see that genre classification of musical artists based on Web data yields respectable results. We can also see that even different parameter settings lead to similar results. Putting additional knowledge into the process of feature extraction may even worsen classification accuracy. From this we conclude that successful classification does not rely on the assumed factors (i.e. incorporating many Web pages that contain "useful" information about the artist).

To gain deeper insights we have to take a look on the features used for classification (Tables 5 and 6). These examples illustrate the effects of the χ^2 test. Since only the most discriminative terms are included, the major part of the features is tied to proper nouns like artist names, their works or locations (see Table 7). For genre "Jazz", we can observe a high number of terms originating from album or track names, whereas for "Country", we find many culturally related terms such as names of dedicated TV or radio stations or Country Music awards (mostly acronyms). For both genres, we find only few typical musical terms, such as instruments, and no adjectives describing style or mood of the genre's music.[2]

Obviously, the presence of artist names and/or well known album and track titles are already strong indicators for the associated genres. It is clear that these informations lead to high separability (cf. the results of nearest neighbor in the preceding sections). Thus, genre classification degrades in fact to an extended co-occurrence analysis, i.e. to predict membership of a category, one only has to examine the occurrence of prototypical examples for that category. We see that the applied Information Retrieval approaches are not capable of capturing the essence of a genre by extracting meaningful words or concepts (i.e., capturing its intensional definition), but, based on an initial set of artist examples, harvest additional examples (including synonyms, popular tracks, and culturally related institutions). In the end, we get a classifier that makes decisions based on extensional definitions of genres, i.e. based on enumerations of prototypical examples (cf. [3] for a discussion on this). To further support the finding that the presence of prototypical artist names is basically the backbone of this sort of genre classification, we modified the feature extraction step and conducted additional experiments as reported in the next section.

---

[2] Although one can often find "descriptive" adjectives in such term lists, they are generally extracted from album or track titles rather than from descriptions of the music, e.g., "*Paranoid*" by Black Sabbath or "*Vulgar* Display of Power" by Pantera (cf. [5]). Nevertheless, these descriptions are in many cases still very valuable as composers already tend to describe their impressions or motivations or want to emphasize the desired message when giving a name to a piece.

duotones	teo	saxophonist	mobley	balakrishna
nefertiti	gil	mulligan	trumpeter	dameron
adderley	bartz	pangaea	blakey	flagelhorn
relaxin	saxophonists	concierto	cookin	ife
agharta	tingen	bess	alton	leu
songbird	modal	bitches	wynton	melonae
silhouette	lechafaud	breathless	porgy	mikkelborg
gorelick	macero	amandla	najee	palle
konitz	jarrett	frelimo	bop	saeta
tutu	cannonball	stitt	kilimanjaro	sitarist
eckstine	orea	harmon	improvisations	sivad
filles	grover	adderly	airto	shorter
nonet	kenny	airegin	sidemen	blackhawk
steamin	mabry	cleota	cosey	soprano
sketches	mtume	lascenseur	prestige	dewey
decoy	ascenseur	milesdavis	sanborn	miles
brew	boplicity	szwed	bebop	charlap
dejohnette	siesta	tadd	davis	sharrock
zawinul	aranjuez	yesternow	albright	birdland
lorber	freeloader	liebman	badal	ipanema

Table 5. 100 highest χ^2 ranked terms for genre "Jazz" from *uspop2002*.

mcgraw	traditionalist	curb	ropin	newnan
cma	dunn	gallimore	buffett	helplessly
nashville	gentry	faith	mcentire	viacom
garth	chely	tulsa	hill	dolly
brooks	tnn	alan	raye	messina
chesney	acm	maines	trisha	collin
leann	chicks	tritt	parton	robison
rimes	honky	martina	loveless	seidel
shania	tonk	aboutcountry	outlaw	tennessee
toby	montgomery	funstuff	clint	daryle
cyrus	keith	gac	getcha	cmas
twain	honkytonk	chattahoochee	oklahoma	patsy
country	andrews	shockn	cowboylyrics	tippin
fireflies	tim	haggard	tonks	flatts
cmt	mcbride	lila	hgtv	tn
strait	shave	martie	hgtvpro	jacked
deana	opry	merle	yearwood	nascar
dixie	wade	mccann	spaces	paisley
achy	reba	entertainer	stroud	diffie
breaky	kenny	cowboy	hayes	tillis

Table 6. 100 highest χ^2 ranked terms for genre "Country" from *uspop2002*.

Category	*Jazz*	*Country*
artist name, nickname, etc.	51	58
album/track title	32	11
instrument, role	8	1
location or institution (e.g., club, award, radio station)	2	21
record label	2	0
genre, subgenre, style	5	8
adjective	0	0
unrelated	0	1

Table 7. Categories of terms and frequencies of categories among the 100 highest χ^2 ranked terms for the genres "Jazz" and "Country" (cf. Tables 5 and 6).

4 Simplified Genre Classification

Since artist names play a very important role in the presented classification method, we could directly intend to capture mainly the names of related artists. Calculation of similarity would then basically consist of finding overlapping artist names. To this end, we modify the query scheme to cause the results to contain names of similar artists (comparable to the approach presented in [9]). We accomplish this by adding either "related artists" or "similar artists" (as whole phrase) to the artist name. As a consequence, we obtain Google result pages pointing frequently to on-line music portals and artist profile pages. Instead of downloading each of the suggested sites and merging them together in order to calculate $tf \times idf$ representations, now we simply extract our $tf \times idf$ features directly from the Google result page. This page contains enough information on similar artists, since Google always displays snippets of the corresponding Web pages that surround the occurrence of the query terms. In our case, we are presented some information that surrounds the artist name as well as text surrounding the terms "similar artists", i.e. related artists. Thus, we circumvent the necessity of downloading each proposed page by extracting the relevant terms directly from the Google summarizations. A further advantage of exploiting this "artist digest" is that each artist is assigned exactly one Web page. Opposed to the initial approach where term frequency and document frequency are slightly redefined to suite the requirements of the $tf \times idf$ assumption, this allows for the original interpretation of the $tf \times idf$ weighting.

We examine this new approach on both proposed query settings and with result pages containing a different number of result snippets (10, 50, or 100). To further test the applicability of this approach also for lesser known artists (which do not occur on many music portals) we also evaluate a genre taxonomy containing around 2000 artists extracted from the *All Music Guide* (see Table 8). Table 9 contains the evaluation results as well as the best results achieved with the initial approach ("+music") for comparison.

The simplified approach yields results equally good as those achieved with the approach from [5], but with considerably less effort. Thus, we have shown that genre classification heavily relies on occurrences of proper names, enabling us to reduce the effort to be made. Our approach is applicable also for lesser known artists, since there obviously exist enough pages with information on artist similarity in the Web even for them. For the specific task of genre classification, the presence of prototypical artists for a genre on the result page seems to be sufficient for predictions.

5 Conclusions and Outlook

In this paper, we extensively evaluated our approach to musical artist classification into genres based on Web data. We investigated its behavior on different evaluation sets and tried to enhance its performance by incorporating techniques to improve the quality of the underlying features. From the obtained results, we

Taxonomy *amg1994a* (9 genres)	
genre	no. of artists
Jazz	811 (40.7%)
Electronica	95 (4.7%)
Rap	41 (2.1%)
Blues	188 (9.4%)
Heavy Metal	270 (13.5%)
Folk	81 (4.1%)
RnB	202 (10.1%)
Reggae	60 (3.0%)
Country	246 (12.3%)
total	1994 (100%)
baseline	40.7%

Table 8. Distribution of genres in the additional taxonomy used for evaluation of the simplified genre classification approach.

		c224a		*c103a*		*uspop2002*		*amg1994a*	
	results	SVM	NN	SVM	NN	SVM	NN	SVM	NN
+music	50	95.69	93.90	65.00	73.00	89.75	87.25	95.82	94.11
+"similar artists"	10	89.69	79.69	56.66	61.66	85.25	77.00	92.35	78.92
	50	95.10	93.49	71.33	72.33	88.25	80.75	89.37	96.57
	100	95.69	93.49	68.33	72.33	87.75	78.25	95.87	89.43
+"related artists"	10	92.59	86.60	54.00	58.66	87.00	80.25	94.41	86.57
	50	94.69	91.59	65.66	71.66	96.32	86.66	89.00	81.75
	100	94.99	90.10	68.66	73.66	89.75	81.50	95.92	87.57

Table 9. Classification results of the proposed simple data retrieval approach on 4 genre taxonomies. The given values are classification accuracies obtained via 50-fold cross validation (values in percent).

gained the insight that the already good results can not be easily improved with the proposed methods. Furthermore, we presented a new approach to accomplish the same task with considerably less effort. Thus, based on the result pages from Google, we can extract features that are equally well suited for genre classification of artists.

From these findings, we can conclude that our genre classification approach heavily relies on the presence of typical artists names. The classification system used in our approach is built upon genre-typical examples, yielding genre models based on extensional definitions. As a consequence, these models can not be exploited to get deeper insights into "what makes a genre", apart from "who" makes it and which works seem to be relevant. While these models are sufficient for (in most cases) proper categorization, they are not descriptive in a sense that a person without any idea of the genre can imagine what the concept of that genre is like (at least not more as by reading the track listing of a CD). To build models that capture the intensional properties of genres (assuming they can be captured at all), one would have to focus on retrieving and extracting the concepts that "make up a genre". This would probably include detection of instruments, typical locations, historic information (e.g., decade, era), descriptions of musical properties, etc. One step in this direction would be to narrow down extraction of information from Web pages to noun phrases and adjectives as proposed by Whitman et al. (e.g., [6]). However, it is also necessary to predetermine the relevant categories of concepts that should be taken into consideration. With this prerequisite, one will unavoidably run into the problem of giving explicit definitions for genres herself and likely end up in a discussion on the shortcomings and insufficiencies of genres in general (cf. [1]). Hence, from our point, it is doubtable whether it is feasible at all to perform automatic genre classification for arbitrary taxonomies without using extensional definitions.

For future work, it seems more promising to pursue related tasks like artist similarity, co-occurrence analysis, and prototypical artist detection [16] which also allow for a broader spectrum of application scenarios. The valuable information present on the Web could be used to offer the user qualitative information about artists, e.g. biographies automatically created by means of text summarization. Furthermore, approaches for Web-entity detection could be issued to discover new artists. This would be a straight forward extension to existing co-occurrence techniques, capable of revealing new information.

6 Acknowledgments

This research is supported by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung (FWF) under project numbers L112-N04 and L511-N15.

References

1. Pachet, F., Cazaly, D.: A taxonomy of musical genres. In: Proceedings of RIAO 2000 Content-Based Multimedia Information Access, Paris, France (2000)
2. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing **10**(5) (2002) 293–302
3. Aucouturier, J.J., Pachet, F.: Representing Musical Genre: A State of the Art. Journal of New Music Research **32**(1) (2003)
4. Lippens, S., Tzanetakis, G., Martens, J.P., Leman, M., Baets, B., Meyer, H.: A Comparison of Human and Automatic Musical Genre Classification. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2004, Montreal, Quebec, Canada (2004)
5. Knees, P., Pampalk, E., Widmer, G.: Artist Classification with Web-based Data. In: Proceedings of 5th International Conference on Music Information Retrieval (ISMIR'04), Barcelona, Spain (October 2004) 517–524
6. Whitman, B., Lawrence, S.: Inferring descriptions and similarity for music from community metadata. In: Proceedings of the 2002 International Computer Music Conference, Goteborg, Sweden (September 2002) 591–598
7. Baumann, S., Hummel, O.: Using cultural metadata for artist recommendation. In: Proceedings of the 3rd WedelMusic Conference. (September 2003)

8. Schedl, M., Knees, P., Widmer, G.: A Web-Based Approach to Assessing Artist Similarity using Co-Occurrences. In: Proceedings of the 4th International Workshop on Content-Based Multimedia Indexing (CBMI'05), Riga, Latvia (June 2005)
9. Geleijnse, G., Korst, J.: Web-based artist categorization. In: Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR'06), Victoria, Canada (October 2006) 266 – 271
10. Levy, M., Sandler, M.: A semantic space for music derived from social tags. In: Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR'07), Vienna, Austria (2007)
11. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management **24**(5) (1988) 513–523
12. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In Fisher, D.H., ed.: Proceedings of ICML-97, 14th International Conference on Machine Learning, Nashville, US, Morgan Kaufman Publishers, San Francisco, US (1997) 412–420
13. Berenzweig, A., Logan, B., Ellis, D.P., Whitman, B.: A large-scale evaluation of acoustic and subjective music similarity measures. Computer Music Journal **28**(2) (2004) 63–76
14. Pampalk, E., Flexer, A., Widmer, G.: Improvements of audio-based music similarity and genre classification. In: Proceedings of the 6th International Conference on Music Information Retrieval, London, UK (2005) 628–633
15. Cruz, I.F., Borisov, S., Marks, M.A., , R.Webb., T.: Measuring structural similarity among Web documents: Preliminary results. In: Proceedings of the 7th International Conference on Electronic Publishing. (1998) 513–524
16. Schedl, M., Knees, P., Widmer, G.: Discovering and Visualizing Prototypical Artists by Web-based Co-Occurrence Analysis. In: Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR'05), London, UK (September 2005)