Investigating Web-Based Approaches to Revealing Prototypical Music Artists in Genre Taxonomies

Markus Schedl[1]
markus.schedl@jku.at

Peter Knees[1]
peter.knees@jku.at

Gerhard Widmer[1,2]
gerhard.widmer@jku.at

[1]Department of Computational Perception
Johannes Kepler University
Linz, Austria

[2]Austrian Research Institute for Artificial Intelligence
Vienna, Austria

***Abstract*— We present three general approaches to detecting prototypical entities in a given taxonomy and apply them to a music information retrieval (MIR) problem. More precisely, we try to find prototypical music artists for each genre in a given real-world taxonomy. The three approaches rely on web-based data mining techniques and derive prototypicality rankings from properties based on the number of web pages found for given entity names.**

We illustrate the approaches using a genre taxonomy created by music experts and present results of extensive evaluations. In detail, three evaluation approaches have been applied. First, we model and evaluate a classification task to determine accuracies. Taking the ordinal character of the prototypicality rankings into account, we further calculate rank order correlation according to Spearman and to Kendall. Interesting insights concerning the performance of the respective approaches when confronting them to the expert rankings are given.

I. Introduction

Prototypical entities play an essential role in cognitive processes. Thus, detecting such entities in given taxonomies is of high interest for a wide variety of fields. Some examples are given in the following list.

- Biology: most prominent representative of a breed
- Science: most prestigious researchers in a research field
- Music: most typical artists for a genre

Prototypical entities are of vital importance for learning, e.g. [1]. Thus, information about them can be applied in various areas, especially in the context of information representation and visualization.

In this paper, we present three methods to compute prototypicality rankings. We apply them to the problem of determining music artists that are typical representatives of a genre. Such prototypical artists can be used, for example, in music information systems and online music stores to support users in finding music more efficiently than with conventional text-based search methods. Since prototypical artists are very well-known, they can also be used to enrich visualizations of, and user interfaces to, music repositories like those presented in [2], [3], [4]. In this context, prototypical artists may serve as reference points to discover similar but less known artists.

To measure the prototypicality of music artists in a given genre taxonomy, we make use of the world wide web. This offers the advantage of incorporating the knowledge and opinions of a large number of people. Thus, web-based data mining approaches reflect a kind of cultural knowledge that we extract and use for prototype detection. Nevertheless, web mining approaches also face some problems. The most obvious one is that they rely on the existence of web pages dealing with the topic under consideration. Therefore, our approaches can only be applied to areas for which enough information is available on the web. However, since the web is still growing rapidly, new areas of application arise every day.

Another issue is to find the requested information. For example, searching for web pages related to the music artist *Bush* will probably result in a large number of web pages not dealing with the band, but with politics and botany. We alleviate this problem by adding music-related terms to the search query. In addition, we will present an approach that corrects prototypicality rankings that are distorted by common speech words by penalizing exorbitant popularity.

Despite these challenges of web-based data mining, it has already been shown that exploiting the world wide web for MIR tasks yields promising results, e.g. [5], [6], [7]. In this paper, we investigate three different approaches to prototypical music artist detection. Two are based on co-occurrence analysis, the third one simply on the number of web pages found for the entity (the artist) under consideration.

The remainder of this paper is structured as follows. In Section II, related literature is briefly discussed. In Section III, the three approaches to prototype detection are presented. Hereafter, we describe in detail the setup of the evaluations performed as well as the obtained results (Section IV). Finally, we summarize our work and point out some future directions in Section V.

II. Related Work

Since the approaches we present in this paper are strongly related to co-occurrence analysis, we first give a short overview of this topic. In [8], playlists of radio stations and

databases of CD compilations were used to derive co-occurrences between tracks and between artists. In [5], [6], first attempts to web-based MIR were made. To this end, user collections of the music sharing service *OpenNap* were analyzed, co-occurrences were extracted and used to build a similarity measure based on community metadata. Co-occurrences of artist names on web pages were first investigated in [9], where the aim was to automatically retrieve related artists to a given seed artist. In [10], artist co-occurrences on web pages were used to create complete similarity matrices which were evaluated for genre classification.

As for the topic of automatic prototypical entity detection for music artists, in [11], an approach based on co-occurrence analysis is presented. Furthermore, a visualization method that illustrates similarities between artists using the most prototypical artist of every genre as reference point is elaborated. In [12], the approach of [11] is refined by downranking artist names that equal common speech terms.

Unlike in [11], [12], where prototype detection approaches for music artists are demonstrated on a quite small set of 224 artists, we use a much larger set of 1 995 artists here. A further weakness of the test set used in [11], [12] is its high number of very popular artists. That makes a serious validation of the obtained prototypicality rankings very difficult.

In contrast, this paper presents the first quantitative evaluation of web-based prototypicality ranking approaches performed on a large test collection which comprises nearly 2 000 well-known as well as less popular music artists and checked against expert rankings.

III. Methods

We consider prototypicality as being strongly related to how often web pages related to the topic under consideration (music, in our case) refer to the entities (artists, in our case). Two of the approaches to prototypicality estimation we evaluate in this paper rely on co-occurrences of entity names (i.e. artist names) on web pages, the third one simply uses page counts.

Given a list of artist names, we use *Google* to obtain the URLs of the 100 top-ranked web pages containing each of the respective strings. *Google* was chosen since it is the most popular search engine and provides a Web API[1]. As for the number of retrieved URLs, preliminary experiments have shown that 100 web pages per artist seem to be a good trade-off between retrieval costs and quality of the results. Addressing the issue of finding only music-related web pages, we add additional keywords to the search query. More precisely, we use the scheme ***"artist name"+music+review*** since it was already successfully applied in [5], [7]. Subsequently, we crawl the top-ranked web pages of every artist and compute a co-occurrence matrix C. To this end, we successively analyze the textual content of each artist's web pages and count how many of them mention the names of the other artists. Performing this procedure for every artist yields a matrix C, where element c_{ij} gives the number of web pages returned for artist i that also mention artist j. The diagonal elements c_{ii} represent the total number of web pages retrieved for artist i, which does not necessarily equal 100 as some pages were not accessible.

This calculation method for co-occurrences differs from the one used in [11], [12] in that it restricts queries to *Google* to a minimum. Raising a query for every pair of artists would be unfeasible for a test collection of nearly 2 000 items.

In the following, we show how we obtain prototypicality rankings based on the co-occurrence matrix C.

A. Backlink/Forward Link (BL/FL) Ratio

The first method to infer prototypicality is based on an idea similar to the PageRank mechanism used by *Google*, where *backlinks* and *forward links* of a web page are used to measure relevancy, cf. [13]. Since we investigate co-occurrences rather than hyperlinks, we call any co-occurrence of artist a_i and artist a_j (unequal to a_i) on a web page that is known to contain artist a_j a *backlink* of a_i (from a_j). A *forward link* of an artist of interest a_i to another artist a_j, in contrast, is given by any occurrence of artist a_j on a web page that is known to mention artist a_i.

Using this interpretation of a backlink and a forward link, we obtain the prototypicality of an artist a_i^g for genre g by counting for how many of the artists $a_{j,j\neq i}^g$ the number of backlinks of a_i^g (from a_j^g) exceeds the number of forward links of a_i^g (to a_j^g). The larger this count, the higher the probability for artist a_i^g being mentioned in the context of other artists from the same genre g and thus, the higher the prototypicality of a_i^g for genre g.

Formally, the ranking function $r(a_i^g)$ that describes the prototypicality of an artist a_i^g for genre g is given by Formula 1, where n^g is the total number of artists in genre g and $bl(i,j)$ and $fl(i,j)$ are functions that return a boolean value according to Formulas 2 and 3 respectively.

$$r(a_i^g) = \frac{\sum_{j=1}^{n^g, j\neq i} bl(i,j)}{\sum_{j=1}^{n^g, j\neq i} fl(i,j)} \tag{1}$$

$$bl(i,j) = \begin{cases} 1 & \text{if } \frac{c_{ij}}{c_{ii}} < \frac{c_{ji}}{c_{jj}} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$fl(i,j) = \begin{cases} 1 & \text{if } \frac{c_{ij}}{c_{ii}} \geq \frac{c_{ji}}{c_{jj}} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$bl(i,j)$ returns the value 1 if artist a_i^g has more backlinks from artist a_j^g (relative to the total number of web pages retrieved for a_j^g) than forward links to artist a_j^g (relative to the total number of web pages retrieved for a_i^g). $fl(i,j)$ is defined analogously. We call $r(a_i^g)$ the *backlink/forward link (bl/fl) ratio* of artist $r(a_i^g)$ since it counts how often the relative frequency of backlinks for a_i^g exceeds the relative frequency of its forward links and relates these two counts.

[1] http://www.google.com/apis

B. BL/FL Ratio with Popularity Penalization

A drawback of the BL/FL approach is that artist names that equal common speech terms, e.g. *Kiss*, *Prince*, or *Hole*, are always top-ranked. The reason for this is that such words frequently occur on arbitrary web pages, regardless of their relatedness to the topic. Therefore, they create a lot of unjustified backlinks for artists with the respective names, what could distort the prototypicality ranking.

To avoid such distortions, we introduce a mechanism that basically pursues the idea of the commonly used information retrieval approach $tf \cdot idf$ (term frequency·inverse document frequency), cf. [14]. In this approach, the importance of a term is higher if it occurs frequently (high tf). On the other hand, a term is penalized if it occurs in many documents and hence, does not contain much relevant information (high df leads to low idf).

In the modified BL/FL approach, we adapt this principle to penalize the prototypicality of an artist if it is high over all genres (following the naming scheme of $tf \cdot idf$, we call this approach $gp \cdot iop$ for *genre prototypicality·inverse overall prototypicality*). This is reasonable since even very popular and important artists are unlikely to be prototypes for all genres.

To emphasize this, we take a look at the 224-artist-set used in [11]. Those artists whose names equal common speech words yield by far the highest overall bl/fl ratios, i.e. *Bush* (223/0), *Prince* (222/1), *Kiss* (221/2), *Madonna* (220/3), and *Nirvana* (218/5).

Incorporating information about overall prototypicality, the second ranking function we propose is shown in Formula 4. The used penalization term is given by Formula 5, where n is the total number of artists in the collection. The functions $bl(i,j)$ and $fl(i,j)$ are defined as in Formulas 2 and 3. $norm$ is a function that shifts all values in the positive range by subtracting the smallest (non negative infinite) value, replaces infinite numbers by 0, and normalizes the values by division by the maximum (in the order mentioned).

$$r(a_i^g) = \frac{\sum_{j=1}^{n^g, j\neq i} bl(i,j)}{\sum_{j=1}^{n^g, j\neq i} fl(i,j) + 1} \cdot penalty(a_i^g)^2 \qquad (4)$$

$$penalty(a_i) = norm\left(log\frac{\sum_{j=1}^{n, j\neq i} fl(i,j)}{\sum_{j=1}^{n, j\neq i} bl(i,j) + 1}\right) \qquad (5)$$

C. Simple Page Counts

The third approach we investigate is very straightforward. We simply query *Google* using the scheme ***"artist name"+"genre name"*** and retrieve the page count value, i.e. the number of found web pages returned for the query. Since in our test collection (cf. Section IV-A), every artist is assigned a single genre, we need to perform this step only once for every artist. For each genre, we then rank its artists according to the page counts to obtain a popularity ranking. Since prototypicality is strongly related to popularity, we simply use this as a prototypicality ranking.

TABLE I

THE DISTRIBUTION OF THE TEST SET AMONG THE TIERS GIVEN BY THE *AMG*. THE ABSOLUTE NUMBER OF ARTISTS ARE GIVEN FOR EVERY GENRE AS WELL AS THE RELATIVE FREQUENCIES AMONG THE *AMG* TIERS.

	absolute				*relative*		
	AMG tier				*AMG tier*		
Genre	*1*	*2*	*3*	$\sum$	*1*	*2*	*3*
Blues	37	95	56	**188**	0.20	0.51	0.30
Electronica	25	68	2	**95**	0.26	0.72	0.02
Reggae	28	32	0	**60**	0.47	0.53	0.00
Jazz	93	400	318	**811**	0.11	0.49	0.39
Folk	44	36	1	**81**	0.54	0.44	0.01
Heavy Metal	14	59	198	**271**	0.05	0.22	0.73
RnB	47	82	73	**202**	0.23	0.41	0.36
Country	39	132	75	**246**	0.16	0.54	0.30
Rap	33	8	0	**41**	0.80	0.20	0.00
Total	360	912	723	**1995**	0.18	0.46	0.36

IV. EVALUATION

Evaluating the quality of the prototypicality ranking approaches is a difficult task for various reasons. First, prototypicality is influenced by personal taste and cultural opinions. Thus, if we had asked a number of people which artists they considered prototypical for a certain genre, they might have named largely their favorites (maybe also those from their own country of origin). Another issue is that prototypical artists may also change over time. For example, formerly unknown artists may become very popular overnight. This raises the question in which way time should be considered in a prototypicality ranking. Should artists be downranked because they were very popular for a genre 30 years ago?

Since our aim was to perform evaluations on a large artist set, conducting a web survey to obtain a ground truth against which the approaches are evaluated was out of the question as this would have included ranking every artist with respect to all other artists of the respective genre. Alternatively, presenting only a subset of artists would have resulted in incomplete rankings.

A. Test Collection and Ground Truth

We finally decided to use a test collection of 1 995 artists from nine common genres, which were extracted from the popular music information system *All Music Guide (AMG)*[2]. The collection comprises very popular as well as less known artists. A list of the artists and their assigned genres can be downloaded from *http://www.cp.jku.at/people/schedl/music/C1995a_artists_genres.txt*.

As ground truth against which we evaluated the prototypicality ranking approaches, we used the "tiers" given by the *AMG*. The artists of each genre are usually clustered in three tiers according to their importance for the respective genre which is defined by experts:

> "The Tier value indicates a ranking of the choices in the list according to the *AMG* Editors' determination

[2] http://www.allmusic.com

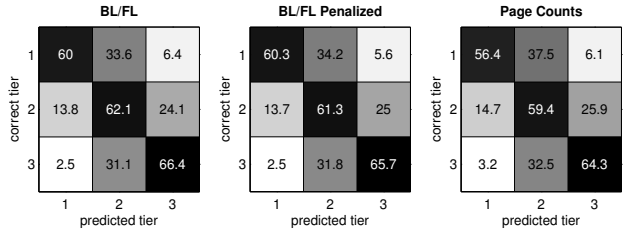

Fig. 1. Confusion matrices for the classification task for each of the three approaches. The columns indicate the tiers to which the approaches map their rankings, the rows indicate the actual *AMG* tiers. The values are given in percent.

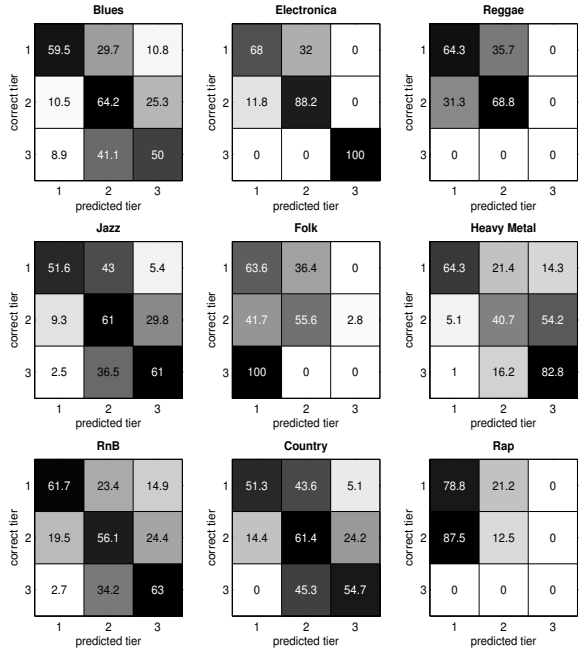

Fig. 2. Confusion matrices for the classification task shown for every genre. The BL/FL approach with penalization of exorbitant popularity was used. The values are given in percent.

> of importance, quality, and relevance to the selected category."[3]

The composition of the test collection can be seen in Table I, where for each genre and each tier, the absolute and relative numbers of included artists are shown.

B. Evaluation Methods

We investigated the quality of the prototypicality rankings using three different evaluation methods – simple accuracy estimation on a classification task, Spearman's rank-order correlation, and Kendall's tau.

1) Classification Accuracy: To gain an overall impression of the performance of the investigated approaches, we interpret the *AMG* tiers as classes and simulate a classification task using our prototypicality ratings as classifiers. To this end, we map the rankings obtained by the prototypicality detection approaches to the ones given by the *AMG* tiers and determine the concordances. More precisely, given that our prototypicality algorithm has produced a specific ranking R of the artists of a genre and assuming the three *AMG* tiers for this genre contain n_1, n_2, and n_3 artists, respectively, we assign the first n_1 elements of R to tier 1, the next n_2 to tier 2, and the last n_3 to tier 3. We can then view these assignments as classification decisions and calculate classification accuracy values.

2) Spearman's Rank-Order Correlation: To measure the correlation between the ground truth ranking of the *AMG* and the rankings obtained with our prototypicality detection approaches, we use the well-established Spearman's rank-order correlation coefficient, e.g. [15]. Since the rankings by the *AMG* are strongly tied, using the standard formula would spuriously inflate the correlation values. Therefore, we apply a tie-corrected version according to [16] as shown in Formulas 6–9, where r_{S_c} gives the rank-order correlation coefficient, n is the total number of ranked data items, X and Y represent the two rankings under consideration, s_X and s_Y are the numbers of sets of ties present in X and Y respectively, and t_{X_i} and t_{Y_i} are the numbers of X and Y scores that are tied for a given rank.

$$r_{S_c} = \frac{\sum x^2 + \sum y^2 - \sum d^2}{2 \cdot \sqrt{\sum x^2 \cdot \sum y^2}} \tag{6}$$

$$\sum x^2 = \frac{n^3 - n - T_X}{12} \qquad \sum y^2 = \frac{n^3 - n - T_Y}{12} \tag{7}$$

$$T_X = \sum_{i=1}^{s_X} (t_{X_i}^3 - t_{X_i}) \qquad T_Y = \sum_{i=1}^{s_Y} (t_{Y_i}^3 - t_{Y_i}) \tag{8}$$

$$\sum d^2 = \sum_{i=1}^{n} (X_i - Y_i)^2 \tag{9}$$

3) Kendall's Tau: We further calculated rank-order correlations according to Kendall's τ. Again, we used the tie-corrected version which is elaborated, for example, in [16]. However, since the Kendall's τ values yielded no new insights when compared to the Spearman's rank-order correlation values, we do not elaborate on them here.

C. Results and Discussion

The overall results of the classification task are depicted in Figure 1, where a confusion matrix for each of the three investigated approaches is shown. It can be seen that the BL/FL-based approaches, in general, perform better than the *Simple Page Counts* approach, especially for predicting first-tier-artists. Comparing the *BL/FL* to the *BL/FL Penalized* approach

[3] http://www.allmusic.com/cg/amg.dll?p=amg&sql=32:amg/info_pages/a_siteglossary.html

TABLE II

THE TEN TOP-RANKED ARTISTS FOR THE GENRES HEAVY METAL AND FOLK FOR EACH OF THE THREE APPROACHES.

Heavy Metal		
BL/FL	*BL/FL Penalized*	*Page Counts*
Death	Metallica	Metallica
Europe	AC/DC	Death
Tool	Black Sabbath	Kiss
Metallica	Death	Tool
Kiss	Led Zeppelin	Extreme
Filter	Riot	Europe
AC/DC	Iron Maiden	Trouble
Led Zeppelin	Judas Priest	Iron Maiden
Black Sabbath	Slayer	Filter
Alice Cooper	Marilyn Manson	Rainbow
Folk		
BL/FL	*BL/FL Penalized*	*Page Counts*
Woody Guthrie	Woody Guthrie	Woody Guthrie
Joan Baez	Joan Baez	Joan Baez
Lucinda Williams	Judy Collins	Pete Seeger
Pete Seeger	Pete Seeger	Lucinda Williams
Judy Collins	Lucinda Williams	Arlo Guthrie
Leadbelly	Doc Watson	Doc Watson
Doc Watson	Leadbelly	Judy Collins
Townes Van Zandt	Phil Ochs	Alan Lomax
Gordon Lightfoot	Gordon Lightfoot	Leadbelly
Phil Ochs	Townes Van Zandt	Gordon Lightfoot

reveals slightly significant better results for the version using penalization of exorbitant popularity when predicting first-tier-artists, but slightly worse results for predicting tiers two and three. This becomes particularly obvious when considering Table II, where the top-ranked artists for the genres Heavy Metal and Folk are shown. In this table, the penalization of artists whose names equal common speech terms can be seen very well when regarding the results for the genre Heavy Metal. In fact, the *BL/FL* approach (and also the *Simple Page Counts* approach) top-ranks artists like *Death*, *Europe*, *Tool*, *Kiss*, and *Filter*. The same artists are considerably downranked by the *BL/FL Penalized* approach. In contrast, the rankings for the genre Folk remain almost unmodified since the artists of this genre are usually known by their real name, cf. Table II.

To get an impression of the impact of the genre on the quality of the results, Figure 2 shows a confusion matrix for each of the nine genres for the best-performing *BL/FL Penalized* approach. It can be seen that the overall results for the genre Electronica are by far the best (weighted with the number of artists in every tier, we obtain an accuracy of 83%, which is 11% above the baseline, cf. Table I). The remarkable wrong confusion for correct tier 3 in genre Folk is due to only one single artist which is incorrectly classified as belonging to tier 1 instead of 3 and therefore, does not considerably influence the overall performance of the approach. Comparing Table III to Table I (for the baseline) reveals that the overall accuracies, except those for the genre Rap, considerably exceed the baseline. In the case of Electronica, Reggae, Jazz, and RnB they are even between 10% and more than 20% above the baseline. In contrast, the results for the genre Rap are very

TABLE III

OVERALL GENRE-SPECIFIC ACCURACIES FOR THE THREE APPROACHES, OBTAINED BY WEIGHTING THE GENRE-SPECIFIC ACCURACIES GIVEN BY FIGURE 2 WITH THE NUMBER OF ARTISTS IN EVERY TIER. acc_0 DENOTES THE ACCURACY THAT THE EVALUATED RANKING APPROACH MAPS AN ARTIST EXACTLY TO THE SAME AMG TIER IT SHOULD FALL INTO ACCORDING TO AMG'S RANKING. acc_1 DENOTES THE ACCURACY WHEN DEVIATIONS OF UP TO ONE TIER ARE ALLOWED.

	BL/FL		*BL/FL Pen*		*Page Counts*	
Genre	acc_0	acc_1	acc_0	acc_1	acc_0	acc_1
Blues	0.59	0.95	0.59	0.95	0.57	0.94
Electronica	0.81	1.00	0.83	1.00	0.73	0.98
Reggae	0.67	1.00	0.67	1.00	0.70	1.00
Jazz	0.60	0.98	0.60	0.98	0.60	0.99
Folk	0.62	0.99	0.60	0.99	0.62	0.99
Heavy Metal	0.73	0.98	0.73	0.99	0.67	0.98
RnB	0.62	0.95	0.60	0.96	0.50	0.93
Country	0.59	0.99	0.58	0.99	0.59	0.99
Rap	0.66	1.00	0.66	1.00	0.66	1.00

poor. Taking a closer look at the *AMG* tiers let us assume that this may be caused by subjective and time-dependent opinions of the experts at *AMG* since very popular Rap artists like *Eminem* and *Snoop Dogg* are assigned to the second tier, whereas many artists that were very popular some years ago are still assigned to tier 1.

As for the results of the correlation analysis, in Table IV, the Spearman's rank-order correlations between the ground truth ranking given by the *AMG* and the rankings obtained with our prototypicality detection approaches are shown for every genre. For all genres except Rap, the rank-order correlation coefficient is at least 0.3, for the genres Electronica, Jazz, Heavy Metal, and RnB it is about 0.5, and for Country it almost reaches 0.6. We also performed a significance test for the results of the Spearman's rank-order correlation according to [16]. Since we do not have any previous knowledge to predict the direction of the difference, we used a two-tailed test with a significance interval of 95%. We proved significance for all obtained correlations, despite those for the genre Rap. For this genre, we obtained a weak negative correlation, which was not stated significant.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we presented and investigated three web-based approaches to ranking entities according to their prototypicality and demonstrated them on the problem of finding prototypical music artists in a given genre taxonomy. Two of the three approaches rely on co-occurrence analysis of entity names on web pages, the third one simply uses page counts returned by *Google* when searching for the entity name (artist) together with the corresponding category (genre).

We used a test collection of nearly 2 000 artists from nine genres for evaluation. As ground truth, we relied on expert opinions taken from the music information system *All Music Guide*. We assessed the quality of the prototypicality rankings using three different evaluation methods – accuracy estimation

TABLE IV

SPEARMAN'S RANK-ORDER CORRELATIONS BETWEEN THE GROUND TRUTH RANKING BY *AMG* AND THE RANKINGS OBTAINED WITH THE PROTOTYPICALITY RANKING APPROACHES.

Genre	*BL/FL*	*BL/FL Pen*	*Page Counts*
Blues	0.40	0.39	0.35
Electronica	0.45	0.48	0.37
Reggae	0.31	0.30	0.31
Jazz	0.49	0.49	0.53
Folk	0.31	0.34	0.33
Heavy Metal	0.48	0.48	0.41
RnB	0.55	0.55	0.42
Country	0.58	0.59	0.58
Rap	-0.21	-0.21	-0.15
Mean	0.37	0.38	0.35

on a classification task, Spearman's rank-order correlation, and Kendall's tau.

To summarize the results, we have shown that the approaches based on *Backlink/Forward Link (BL/FL) Ratios* perform better than the *Simple Page Counts* approach. We further showed that penalization of exorbitant popularity improves results in some cases, e.g. for the genre Heavy Metal, where many artist names equal common speech terms. However, for genres like Folk or Jazz, where most artists use their real names, or at least pseudonyms that sound like real names, no significant improvements could be made out when using the *BL/FL* approach with penalization of exorbitant popularity.

As for future work, we plan to create a user interface that incorporates information about prototypical music artists. Our aim is to provide the user with reference points (the prototypical artists), so that he/she will be able to browse music repositories more efficiently than with conventional user interfaces. We further intend to apply our prototype detection approaches to domains other than music.

We are currently investigating web-based approaches to determining the period of activity of an artist. This information could help refining prototypicality by weighting an artist according to his/her principal years of musical activity.

ACKNOWLEDGMENTS

This research is supported by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung (FWF) under project number L112-N04 and by the Vienna Science and Technology Fund (WWTF) under project number CI010 (Interfaces to Music). The Austrian Research Institute for Artificial Intelligence acknowledges financial support by the Austrian ministries BMBWK and BMVIT.

REFERENCES

[1] R. T. Kellogg, *Cognitive Psychology*, 2nd ed. Thousand Oaks, California, USA: Sage Publications, Inc., 2003.

[2] E. Pampalk, S. Dixon, and G. Widmer, "Exploring Music Collections by Browsing Different Views," in *Proceedings of the Fourth International Conference on Music Information Retrieval (ISMIR'03)*, Washington, D.C., USA, October 2003.

[3] D. Gleich, M. Rasmussen, K. Lang, and L. Zhukov, "The World of Music: SDP Layout of High Dimensional Data," in *Proceedings of the IEEE Symposium on Information Visualization 2005*, Minneapolis, Minnesota, USA, October 2005.

[4] M. Goto and T. Goto, "Musicream: New Music Playback Interface for Streaming, Sticking, Sorting, and Recalling Musical Pieces," in *Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR'05)*, London, UK, September 2005.

[5] B. Whitman and S. Lawrence, "Inferring Descriptions and Similarity for Music from Community Metadata," in *Proceedings of the 2002 International Computer Music Conference*, Goeteborg, Sweden, September 2002, pp. 591–598.

[6] D. P. W. Ellis, B. Withman, A. Berenzweig, and S. Lawrence, "The Quest for Ground Truth in Musical Artist Similarity," in *Proceedings of the 3rd International Symposium on Music Information Retrieval (ISMIR'02)*, Paris, France, 2002.

[7] P. Knees, E. Pampalk, and G. Widmer, "Artist Classification with Web-based Data," in *Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR'04)*, Barcelona, Spain, October 2004, pp. 517–524.

[8] F. Pachet, G. Westerman, and D. Laigre, "Musical Data Mining for Electronic Music Distribution," in *Proceedings of the 1st WedelMusic Conference*, 2001.

[9] M. Zadel and I. Fujinaga, "Web Services for Music Information Retrieval," in *Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR'04)*, Barcelona, Spain, October 2004.

[10] M. Schedl, P. Knees, and G. Widmer, "A Web-Based Approach to Assessing Artist Similarity using Co-Occurrences," in *Proceedings of the Fourth International Workshop on Content-Based Multimedia Indexing (CBMI'05)*, Riga, Latvia, June 2005.

[11] ——, "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.

[12] ——, "Improving Prototypical Artist Detection by Penalizing Exorbitant Popularity," in *Proceedings of the Third International Symposium on Computer Music Modeling and Retrieval (CMMR'05)*, Pisa, Italy, September 2005.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," in *Proceedings of the Annual Meeting of the American Society for Information Science (ASIS'98)*, January 1998, pp. 161–172.

[14] G. Salton and C. Buckley, "Term-weighting Approaches in Automatic Text Retrieval," *Information Processing and Management*, vol. 24, no. 5, pp. 513–523, 1988.

[15] R. V. Hogg, A. Craig, and J. W. McKean, *Introduction to Mathematical Statistics*, 6th ed. Prentice Hall, June 2004.

[16] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, 3rd ed. Boca Raton, London, New York, Washington, D.C.: Chapman and Hall/CRC, 2004.